DESIGN AND EVALUATION OF REAL-TIME VOICE-OVER-IP (VOIP) SYSTEMS WITH HIGH PERCEPTUAL CONVERSATIONAL QUALITY

BY

BATU SAT

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Benjamin W. Wah, Chair
Professor Klara Nahrstedt
Associate Professor Jont Allen
Professor Nitin H. Vaidya

ABSTRACT

This research addresses real-time multimedia communication systems that can achieve high perceptual quality for their users. It focuses on the fundamental understanding of multiple quality aspects perceived and their trade-offs in the design of run-time control schemes that adapt to changing network conditions and expectations of the users of a system.

One of the main contributions of this thesis is the use of adaptively scheduled off-line subjective comparison tests to efficiently and accurately learn users' subjective preferences among the alternative trade-offs in order to guide the run-time control schemes to achieve high perceptual quality.

In order to illustrate the application of the general framework and our methodology, throughout this thesis we study the design of real-time VoIP (voice-over-IP) systems that can achieve high perceptual conversational quality. The trade-offs in the design and operation of a VoIP system involve the design of speech codecs and strategies for network control, playout scheduling, and loss concealment.

The perceptual quality of a conversation over a network connection depends on the quality of the one-way speech (listening-only speech quality or LOSQ) received and the delay incurred from the mouth of the speaker to the ear of the listener (mouth to ear delay or MED). In a conversation, each participant takes turns in speaking and listening, and both perceive a silence duration called mutual silence (MS) when the conversation switches from one party to another. When the connection has delays, the MSs perceived by a participant consist of alternating short and long silence periods between turns. As a result, listeners may experience lower perceptual quality when the MSs are highly asymmetric, some speakers appearing to be more distant than others, or some responding slower than others.

The evaluation of conversational quality is a largely unexplored area. There are many objective metrics for assessing the quality of a VoIP conversation, but there is no single objective metric whose results match well with subjective results. Indiscriminate subjective testing is not feasible because it is prohibitively expensive to carry out many such tests under various conversational and network conditions. Moreover, there is no systematic method to generalize the subjective test

results to unseen conditions. To this end, there are five key innovations in this research that will address the limitations of existing work and improve the perceptual quality of VoIP systems.

Firstly, we have developed a methodology and test-bed to objectively and subjectively evaluate VoIP systems under repeatable and fair conditions using a black-box approach. We have applied this methodology and test-bed on four commonly used VoIP systems: Skype, Google-Talk, Windows Live and Yahoo Messenger. The results show that different systems operate differently in coping with network imperfections such as jitter and loss. We also observe that systems do not change their operation as a function of the one-way delay of the connection, or of the turn-taking frequency of the conversation. The results show that Windows Live is preferred under a significant set of conditions, but none of the systems is dominant under all conditions.

We have also learned the mapping between easily obtainable objective measures and subjective preferences of users in order to allow any VoIP system to be comprehensively compared against others without expensive subjective comparisons. We later use the mapping learned to subjectively compare our newly designed system with the four systems already evaluated.

Secondly, we have developed a general model of pair-wise subjective comparisons, based on individually identified properties, axioms and lemmas, that models comparisons on a continuous operating curve with a single control parameter. The model provides a basis for developing the method to schedule adaptive off-line subjective tests and for identifying the optimal point by fusing the information obtained from separate subjective evaluations on the same operating curve. The model can be used for formulating and solving any type of pair-wise comparison problem that exhibits the same properties identified. The model is flexible to allow the existence of multiple optimal points on an operating curve and includes a belief function framework that can guide the search for optimal points efficiently. Furthermore, the model is built on a statistical framework that allows for the confidence of individual evaluation results to be represented in the conclusiveness of the combined belief function.

Thirdly, we have developed a method for tackling the control design problem of finding the optimal point in an N-dimensional space, which includes all the metrics that affect quality. The overall problem is transformed into two independent problems of finding the optimal point on a continuous but one-dimensional curve, and learning the mapping on a set of curves that adequately spans the K-dimensional ($K < N$) curve space, where K stands for the number of metrics characterizing the network and conversational conditions.

Fourthly, we have applied the methodology to the design of play-out scheduling control for VoIP systems by conducting subjective tests and by learning from them under a comprehensive set of network and conversational conditions. We have also verified the accuracy and efficiency of

our methodology using exhaustive subjective tests on a subset of the network and conversational conditions. The verification of our methodology on a real-life application also justified our model of pair-wise subjective comparisons and our optimal algorithm to adaptively choose upcoming comparisons using information learned from previously conducted tests.

Lastly, we have generalized the learned results in the design of play-out scheduling control to conditions that are unseen at design time and observed only at run-time. Along with design choices in other components of the architecture, this results in a VoIP system that can achieve higher and more consistent perceptual quality than other four VoIP systems analyzed. We have also shown that our system performs very close to a non-causal ideal system where the POS decisions are made optimally using future information.

Our model and methodologies are applicable to a wide variety of real-time multimedia communication system-control design problems which operate under constrained resources, communicating over non-stationary IP networks, and for which the overall quality perceived by users of the system has multiple counteracting aspects which cannot be represented by a single objective metric.

To my wife for her love and support. To my parents for their belief in me.

ACKNOWLEDGMENTS

I would like to thank my adviser, Prof. Benjamin W. Wah, for his support, dedication and attention to detail. I would like to especially acknowledge his idealism to identify the most important and challenging problems and provide guidance to solve those problems in the most elegant way possible.

I would like to thank my committee members, Prof. Nahrstedt, Prof. Allen and Prof. Vaidya for their valuable suggestions and support.

I would like to thank my wife Pinar, who supported me throughout my studies.

I also would like to thank my parents, Umit and Haluk, my sister Burcu and my entire family for their belief in me and their support.

I would like to thank my friends in Urbana-Champaign, especially Tansu and Ozlem Alpcan, Cigdem Sengul, Onur Pekcan, Tayfun Deniz, Arif Ozkan, Hakan Bagci, Tolga Memioglu, Serdar Yuksel, Serdar Kozat, Kivanc Mihcak, Nejan Huvaj, Suleyman Sarihan, Lale Ozhahya and Norma Linton.

I also would like to thank the members of our research group for their insightful comments.

This research was supported by National Science Foundation Grant CNS 08-41336 and by the Motorola Center for Communications at the University of Illinois at Urbana-Champaign.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ACR	Absolute Category Rating
CCR	Comparative Category Rating
CE	Conversational Efficiency
CDF	Cumulative Distribution Function
CMOS	Comparative Mean Opinion Score
CND	Complete Noticeable Difference
COD	Comparative Opinion Distribution
CS	Conversational Symmetry
DMOS	Degradation Mean Opinion Score
IETF	Internet Engineering Task Force
ITU	International Telecommunication Union
IP	Internet Protocol
JND	Just Noticeable Difference
LC	Loss Concealment
LO	Local Optima
LOSQ	Listening-Only Speech Quality
LR	Loss Rate
MED	Mouth-to-Ear Delay
MOS	Mean Opinion Score
MTU	Maximum Transmission Unit

PDF	Probability Distribution Function
PESQ	Perceptual Evaluation of Speech Quality
POS	Play-out Scheduling
PSTN	Public Switched Telephone Network
ROD	Region of Dominance
RTP	Real-time Transport Protocol
QoE	Quality of Experience
QoS	Quality of Service
UCFLP	Unconcealed Frame Loss Pattern
UCFLR	Unconcealed Frame Loss Rate
UDP	User Datagram Protocol
VBR	Variable Bit Rate

$\mathcal{O}$	Operating curve
A^*	Optimal point on an operating curve

CHAPTER 1

INTRODUCTION

1.1 Motivations

In today's global world, the ability to reliably communicate with people in distant locations is increasingly important. In the absence of face-to-face interactions, voice communication is considered one of the most effective forms of communication.

In the last century, the traditional approach to satisfy this demand has been the deployment of networks specialized to carry voice signals between specialized devices such as land-line and wireless phones.

With the proliferation of IP networks around the world that can deliver connectivity to a wider set of devices in a more cost effective manner, the possibility to extensively utilize IP networks for real-time communication needs became attainable.

In this context, VoIP technology can provide real-time speech communications between users connected by an IP network, public or private, in such a way that closely resembles a face-to-face conversation. The process involves the delivery of speech frames from one location to another with high quality and low latency. VoIP technology has a significant impact on the multi-billion dollar telecommunication industry: the promise of less expensive phone calls with comparable quality and better features than PSTNs has accelerated its adoption, both in business and home applications. As the number of households and businesses connecting to the Internet steadily increases around the world, more users utilize the public Internet for their communication needs.

Seamless interoperability with existing public switched telephone networks (PSTNs) has been one of the crucial accelerants of VoIP adoption. Furthermore, VoIP's better integration with various forms of collaborative communications, such as instant messaging, desktop sharing, voice-mails, and video calls, has made it a suitable communication solution for today and the future.

1.2 Overall Goals

This research addresses real-time VoIP (voice-over-IP) systems that can achieve high perceptual quality for their users. It focuses on the fundamental understanding of multiple quality aspects perceived by users of the system and their trade-offs in the design of run-time control schemes that adapt to changing network conditions and communication scenarios.

Our two goals in this thesis are to evaluate existing real-time VoIP (voice-over-IP) systems and to design new VoIP systems that can achieve high perceptual conversational quality. For this reason we study trade-offs in the design and operation of a VoIP system, which involve the design of strategies for network control, playout scheduling, and loss concealment as well as evaluating speech codecs to be employed.

Due to reasons discussed later in this chapter, off-line subjective evaluations are needed in both the evaluation and the design phases of this study. Thus, an important aspect of this research is the development of new methods for reducing the large number of subjective tests needed and for automated learning and generalization of the results of subjective evaluations.

In the last phase of the study, the methods developed are applied to the design and operation of a new VoIP system that outperforms existing systems in terms of delivering high conversational quality.

In the remainder of this chapter we first present the characteristics of real-time VoIP communication systems as they relate to the evaluation and design of such systems. Secondly, we discuss the suitability of evaluation methodologies for conversational quality. Lastly, we present the problems studied, the contributions of the thesis and the outline of the thesis.

1.3 Real-Time VoIP Communication Systems

In this section, we characterize the properties of real-time VoIP systems. This characterization is used as the basis for choosing appropriate methodologies in our approach to evaluate and design such systems with superior perceptual quality. As the overall goal is to evaluate and guide the design of VoIP communication systems, each characteristic is presented in relation to this goal.

a) *Multiple objective quality metrics.* A common approach currently used to evaluate a multimedia communication system is to use some objective metrics recommended by a standardization body, such as the International Telecommunication Union (ITU) or the Internet Engineering Task Force (IETF), as well as metrics that can be computed easily. Examples include the delay incurred and the quality of the received media. However, similar to many multimedia applications, VoIP is characterized by multiple objective metrics, which cannot be collapsed into a single metric that captures all aspects of quality. A detailed discussion on this is presented later in this chapter in Section 1.4.

When there are multiple metrics, quality can be denoted by a point in a multi-dimensional space, whose axes correspond to the individual metrics.

b) *Constrained resources.* The control schemes in these systems usually operate under limited network resources, such as constraints on bandwidth, packet rates and computational resources. These constraints limit the quality that can be achieved under ideal conditions as well as possible adaptations to operation under non-ideal conditions. Thus, before evaluating or designing a system, it is crucial to understand the constraints imposed on the system's operation to correctly identify the feasible options for operation.

c) *Best-effort IP network.* Real-time communication systems commonly utilize IP networks, which may consist of several partitions, such as the core, the access and the last-mile networks, and may include wired, wireless or satellite media of transmission. Thus, IP networks may exhibit dynamic non-stationary delay and loss behaviors that differ from connection to connection, and can possibly change due to packet rate and bit-rate that each end-point transmits.

d) *Communication scenario among participants.* As mentioned before, this affects the subjective quality perceived by participants. For example, delay degradations may be more important when participants have frequent interactions; thus, users may be less tolerant of delays. In some cases excessive delays may result in double-talk conditions in a conversation, where both participants speak at the same time.

e) *System control.* To mitigate network imperfections, the control schemes employed in these systems have adjustable parameters, such as the transmission rate and the playout schedule. For a control scheme under given constraints and conditions, the set of *operating points* in the multi-dimensional quality-metric space correspond to its feasible control values. This set of points forms an *operating curve*. The control scheme can operate based on a pre-determined scheme or can adapt its operation based on observable parameters at run-time. The control scheme should be designed with dynamic system constraints, network conditions, and communication scenarios in mind.

f) *Trade-offs among objective metrics on subjective preferences.* Due to system constraints and network imperfections, trade-offs must be made among the multiple counteracting quality metrics. Since their effect on subjective user preferences is not defined, it is difficult to select the proper control parameter values in order to arrive at an operating point with the highest subjective quality.

g) *Multiple locally optimal operating points.* Due to the counteracting effects of the multiple quality metrics, there may be more than one locally optimal operating point of preferred subjective quality. Each such point is the most preferred point among the alternatives in its neighborhood on the operating curve.

In the next section we present a summary of evaluation methodologies that may be used in the identification of the optimal operating point mentioned above.

1.4 Evaluations of Conversational Quality

The goal of this thesis requires the ability to evaluate conversational quality of VoIP conversations. Thus, it is crucial to understand the methodologies for the evaluation of conversational quality and their suitability to perform the required evaluations. The evaluation is important both for the comparative evaluation of existing systems and in the design of control schemes for new systems.

Evaluations in general can be conducted in two ways: objectively or subjectively. The International Telecommunication Union (ITU) has several recommendations for the objective and subjective evaluations of the end-to-end quality of a voice transmission system. Objective metrics, which will be discussed in detail in Chapter 3, include PESQ (ITU P.862), E-Model (ITU G.107), and Call Clarity Index (ITU P.561 and P.562) [20, 21]. Although there are many objective metrics for assessing a VoIP conversation, as is discussed in Chapter 3, there is no single objective metric whose results match well with subjective results. None of these metrics can capture the trade-offs among user-perceptible attributes that affect subjective conversational quality, as they either make simplifying assumptions that are not necessarily true, or omit the effect of some attributes altogether.

There are three difficulties of reducing all aspects of quality to a single dimension in an objective metric. The first is due to the fact that perception of quality is closely influenced by the communication scenario. Most standard metrics mentioned above that attempt to collapse quality into a single dimension, either completely ignore this relation or assume a typical communication scenario. This approach simplifies the construction and calculation of the single-dimensional quality expression, but reduces the accuracy of the metric under differing conditions.

Secondly, the influence of a quality aspect on the overall perception of quality is non-linear, even if other quality aspects are constant. For example, if a quality aspect is close to perfect, any further improvement may not be perceived. Similarly, if a quality aspect is severely degraded, beyond any usable level, any further degradation is usually not perceived.

Thirdly, the overall perception of quality is influenced differently by each quality aspect in a way that depends on the other quality aspects. For example, if one aspect of quality is severely degraded, the influence of another quality aspect on the overall perception is smaller than usual. Thus, single-dimensional quality metrics that are overly simplified do not lead to accurate evaluations of quality.

On the other hand, the ITU P.800 [21] methodology describes how to obtain MOS_{CQS} (subjective conversational quality) that provides subjective evaluations of conversations. Because such subjective evaluations cannot be performed at run time, offline tests have to be conducted during which the information learned is used to guide the operation of the control scheme(s) at run time. In general, subjective evaluations are time-consuming and expensive and will require multiple subjects in order to arrive at some statistically significant results. In general, subjective evaluations are not preferred because they are expensive to conduct and hard to repeat. Furthermore, since there may be prohibitively many network conditions and communication scenarios that can be observed at run time, it is infeasible to conduct exhaustive subjective tests in order to cover all possibilities.

A standard method for conducting subjective evaluations is to ask subjects to rank the quality by an *absolute category rating* (ACR) and to take an algebraic mean of the opinions of the subjects in response to the same stimuli. The result obtained is the *mean opinion score* (MOS) [21].

There are two reasons why MOS is only useful for verifying a system's performance but not suitable for designing new control schemes. Firstly, absolute scores obtained for two points on an operating curve can be used to deduce their relative positions. If all alternatives are mutually related under pair-wise comparisons, then a total ordering can be established under ACR. In practice, two operating points may not be comparable when they involve multiple quality metrics. In this case, the perceived effects on the difference of one metric may not be consistently translated into the differences of the other metrics. Consequently, the feasible operating points of an operating curve lie on a Pareto-optimal boundary. Secondly, although each MOS score can be determined with some statistical confidence, no statistical significance can be associated with the difference of two MOS scores. For instance, if the variances in the scores are large relative to their difference, then the conclusion reached on the difference is not statistically meaningful. As is stated in ITU P.800 [21] for evaluating telephone communication quality, absolute ratings are not accurate for evaluating quality when samples have high quality or their difference is barely perceptible. Hence, the number of samples required to obtain MOS with a certain level of statistical significance can be inadequate for some pair-wise comparisons but excessive for other cases.

Evaluation of conversational quality for the design of a new VoIP system is a largely unexplored area. Not only is it prohibitively expensive to carry out many subjective tests under various conversational and network conditions, but also the generalization of the results to unseen conditions has not been systematically studied.

Problems studied to address this issue are presented next in this chapter, and our detailed approach to address the issues identified is presented in Chapter 3 along with the relevant previous work.

1.5 Problems Studied

The goal of VoIP systems is to provide voice communication that closely resembles a face-to-face conversation (also called orthophonic) across remote locations. To fulfill this need, specialized network infrastructure has already been deployed around the world which is simply referred to as PSTN (*public switched telephone network*). However, since the network delays in VoIP can be long and time-varying, its design is different from those for PSTN with short and consistent delays [22, 26].

Our study is focused on the evaluation and design of real-time VoIP communication systems, where there are controllable parameters which affect the system's performance in terms of the quality of experience (QoE) the user of the system perceives. However, due to the conditions under which the system operates, such as network conditions, there are some constraints on the controllable parameters in achieving perceptually favorable results.

The design of the on-line operation of a VoIP system consists of the design of *play-out scheduling* (POS) and *loss-concealment* (LC) strategies that involve delay-quality trade-offs that optimize user-perceptible attributes. These algorithms dynamically adapt to changing network and conversational conditions. In addition to directly designing system control components, design choices are made based on system constraints and analysis of other components, such as the speech codec and the collection of network and conversational conditions. LC and POS components take into account the choices made on the other components to control the overall quality perceived by a user of the system.

Furthermore, in such systems an overall scalar quality of experience metric that is consistent across all operating conditions has not been or cannot be objectively defined. Thus, it is imperative to utilize subjective evaluations to guide the design of such systems. However, due to the expensive and time-consuming nature of the subjective evaluations, the inhibitively large number of operating conditions and the multitude of feasible control possibilities under each set of conditions, it is not trivial to conduct subjective tests to learn the preference of subjects and to generalize to unseen conditions at run-time.

Figure 1.1 depicts the outline of the thesis, and Table 1.1 lists the problems studied. The thesis can be divided into three stages, where the first stage provides background to the other two stages.

The first stage of our study, denoted by P0 (*Problem 0*) in Table 1.1, involves the analysis of

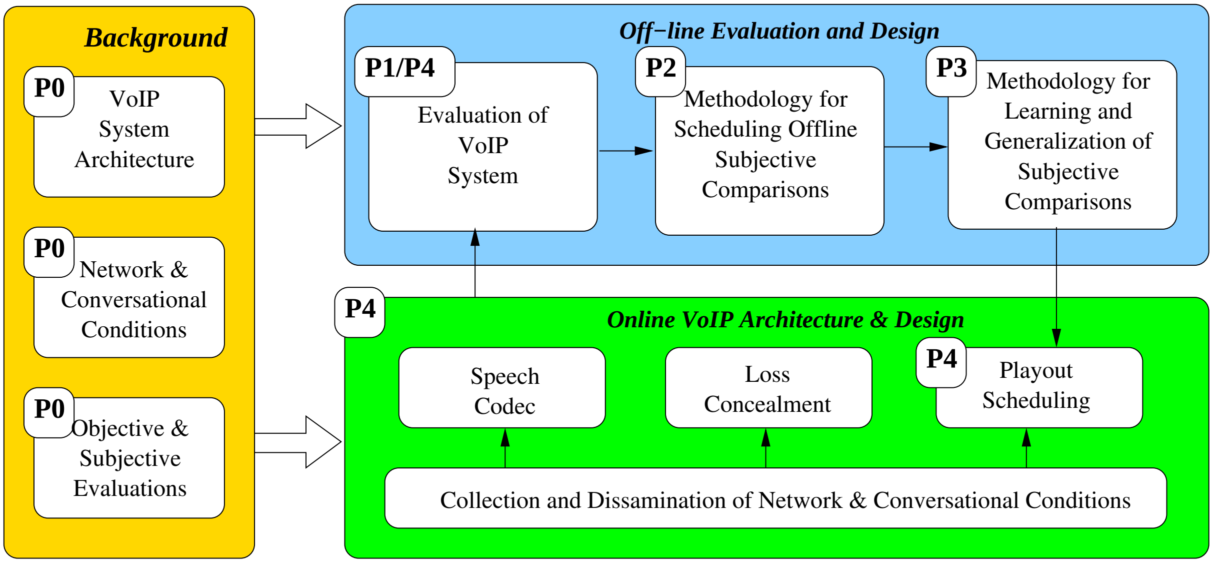

Background
P0
VoIP System Architecture
P0
Network & Conversational Conditions
P0
Objective & Subjective Evaluations
Off−line Evaluation and Design
P1/P4
Evaluation of VoIP System
P2
Methodology for Scheduling Offline Subjective Comparisons
P3
Methodology for Learning and Generalization of Subjective Comparisons
P4
Online VoIP Architecture & Design
Speech Codec
Loss Concealment
P4
Playout Scheduling
Collection and Dissamination of Network & Conversational Conditions

Figure 1.1: Overview of the problems studied in the thesis that consist of background (stage 1), off-line design and evaluation (stage 2) and online VoIP architecture and design (stage 3).

a VoIP system architecture, the study of network and conversational conditions that affect quality perception, and the study of methodologies for objective and subjective evaluation of quality.

The study of VoIP architecture involves the analysis of network and endpoint components in an end-to-end VoIP communication solution. Based on this analysis we identify two environmental factors that are critical in the operation and performance of a VoIP client. The first is the IP network(s) that the VoIP call utilizes; thus, later in this chapter we study the characteristics of IP networks. The second is the users of the VoIP client, who perceive the quality of the conversation and are the ultimate judge of performance of a communication system. The study of the *network* and *conversational environments* entails the identification of objective metrics for characterizing network and conversational conditions and the collection and dissemination of this information at run time. Later, we discuss different methodologies to evaluate the conversational quality. Lastly, in the first stage, we study the tradeoffs in the design of control schemes in the VoIP client, which include play-out scheduling and loss concealment.

The second stage of our study includes P1, P2 and P3, and encompasses the off-line tasks related to the evaluation and design VoIP systems. This portion of the thesis is critical as it utilizes the background analysis and develops a series of methodologies to evaluate both VoIP systems as a whole and individual control alternatives under specific sets of conditions. It also tackles the problem of designing a POS control scheme which makes real-time decisions among a continuous set of alternatives under infinitely many sets of conditions that the system may encounter during a

Table 1.1: Research issues addressed in this thesis.

P0: Background on VoIP systems
- VoIP system architecture
- Network and conversational conditions
- Objective and subjective evaluation

P1: Evaluation of Conversational Quality in VoIP Systems
- Method for studying trade-offs under repeatable network and conversational conditions
- Application of the methodology on four commonly used VoIP systems
- Predicting subjective preference between two VoIP systems using objective measures

P2: Efficient Methodologies for Off-line Subjective Evaluation of Control Schemes
- Statistical model for comparative subjective tests between points over an operating curve
- Efficient algorithms for scheduling of off-line comparative subjective evaluations
- Evaluation of the scheduling algorithms designed for their accuracy and efficiency

P3: Learning of Subjective Evaluations for guiding design of real-time control
- Application of methodology in P2 for conducting off-line subjective tests
- Exhaustive subjective evaluations for verifying accuracy of limited subjective tests
- Identification of objective measures available at run-time that can predict desirable points
- Learning the mapping between objective metrics identified and subjective preferences

P4: Design and Evaluation of VoIP systems with High Perceptual Quality
- Generalization of mapping learned to unseen operating conditions observed at run-time
- Performance evaluation of designed VoIP system against other systems using P1
- Evaluation under unseen conditions against an ideal system with future information

VoIP call.

In P1, we study a systematic methodology to evaluate existing VoIP systems both objectively and subjectively. Later we learn the mapping between objective measures characterizing conversational quality and conditions and the subjective preferences, allowing us to predict subjective preference between any two systems under any conditions based on the easily obtainable objective metrics.

In P2, we construct a model for comparative subjective tests, which provides the basis for an off-line scheduling algorithm to adaptively choose comparison pairs on a feasible set of choices under a given set of conditions. The algorithm also identifies the optimal operating point for a given set of conditions based on the sequence of subjective comparison tests.

In P3, we apply the methodology developed in P2 to carefully identified set of conditions to conduct subjective comparison tests that would be utilized to guide the design of the POS control scheme for VoIP systems. This part of the thesis utilizes all the knowledge accumulated so far and results in a mapping between objective measures that can be obtained at run-time and the optimal alternative identified by the subjective evaluations.

In the third stage of our study, denoted by P4, we bring everything together, utilizing the mapping learned in P3 to design a POS control scheme. We combine the newly designed POS with other components of a system, such as loss concealment and speech codecs, and create a new VoIP system with high perceptual quality. Lastly, we thoroughly evaluate the newly designed VoIP system under a variety of conditions against existing VoIP systems and a system with ideal POS decisions using the methodologies developed in P1. The comparative evaluation with existing systems is depicted in Figure 1.1 with an arrow from the *On-line VoIP architecture* to the *Evaluation of VoIP systems*, completing the loop that started with that task.

1.6 Contributions of this Research

The first contribution of this thesis, which provides a basis for all the other contributions, is the identification of a comprehensive set of objective measures that are related with the perception of conversational quality of VoIP calls. In addition to measures that have been used before in characterizing network conditions, new measures are defined that can characterize the human perception of conversational dynamics, and can be obtained at run-time during a VoIP call. The close relation of these objective metrics and the subjective preferences of subjects are evident by the high self-prediction accuracy of the mappings learned in P1 and P3.

The second contribution of this thesis is the methodology developed to comparatively evaluate VoIP systems. This contribution entails two parts. Firstly, we have improved the ITU recommendations for comparative evaluations (the MOS and CMOS framework) to allow comparative evaluation methodology to output comparisons that are statistically meaningful. Secondly, we design and implement a testbed that completely isolates and captures the differences among VoIP systems, by exactly simulating interactive conversations and network conditions, to the finest granularity possible.

The third contribution is the SVM model learned in P1 that successfully predicts subjective preference between two unseen VoIP systems under unseen conditions, based on easily obtainable objective measures for characterizing the multi-dimensional aspects of conversational quality of each system, along with the network and conversational conditions under which the comparison is conducted.

The fourth contribution is the development of a *model of pairwise subjective comparisons* based on individually identified properties, axioms and lemmas. The model provides a basis for developing the methodology to schedule adaptive off-line subjective tests and for identifying the optimal point by fusing the information obtained from separate subjective evaluations on the same operating curve. Aside from its use in this thesis, the model can be used in formulating and solving any type of pair-wise comparison problem that exhibits the same properties identified. The model is flexible to allow for the existence of multiple optimal points on an operating curve, and includes a belief function framework that can guide the search for optimal points with efficiency. Furthermore, the model is built on a statistical framework which allows for the confidence of individual evaluation results to be represented in the conclusiveness of the combined belief function.

The fifth contribution is the methodology developed to tackle the control design problem of finding the optimal point in an N-dimensional space to two independent problems of finding the optimal point on a continuous, but one-dimensional curve and learning the mapping on a set of curves that adequately spans the K-dimensional ($K < N$) curve space. In this framework, K stands for the number of metrics characterizing the network and conversational conditions, where N stands for all the metrics that affect quality, which include quality metrics characterizing the VoIP conversation as well as the K metrics mentioned above.

The last contribution of this thesis is the application of all the methodologies developed to the design of POS control for a VoIP system, which includes conducting extensive subjective comparison tests that lead to the development of a new VoIP system that outperforms existing systems and performs very close to an ideal system where the POS decision is made optimally using future information.

1.7 Outline of the Thesis

As mentioned above, Figure 1.1 and Table 1.1 summarize the problems studied in this thesis. The organization of the chapters corresponding to the block diagram and the list of problems are as follows: The background on VoIP systems (P0) is presented in Chapter 2, followed by the previous work and our approach in Chapter 3. In Chapter 4 we present the evaluation of conversational quality in VoIP systems (P1). In Chapters 5 and 6 we present efficient methodologies for off-line subjective evaluation of control schemes (P2). Learning of subjective evaluations for guiding design of real-time control (P3) is presented in Chapter 7, followed by the design and evaluation of VoIP systems with high perceptual quality (P4) in Chapter 8. Lastly, we present our conclusions and our future work in Chapter 9.

In order to maintain the flow of the thesis, we have placed some of the derivations and other related information in the appendixes. Appendix A presents the derivations related to the acceptable range of values around the optimal point on an operating curve. Appendix B describes the details of the Monte Carlo simulations we have conducted in solving P2.

CHAPTER 2

BACKGROUND

In this chapter, we present background information related to the problem studied. Firstly, we present the VoIP system architecture, usage of the VoIP technology, and the components of a VoIP client.

Secondly, we present the IP network environment in which a VoIP system operates and discuss our observations on the network conditions observed in the Internet and how these conditions relate to the design of VoIP systems.

Lastly, we present the conversational dynamics of an interactive voice communication scenario and illustrate the effects of delays on a VoIP conversation in comparison to a face-to-face conversation. We also discuss how users of the system perceive delay through conversational dynamics and present a set of objective metrics that can be used to characterize such dynamics.

2.1 VoIP System Architecture

VoIP network technologies can be classified according to the IP network in which the speech data is transmitted. They can run on privately owned IP networks, such as enterprise networks and leased lines, or the public Internet, which can be accessed via Internet service providers (ISPs). Since VoIP calls may traverse a combination of networks that are not owned by a single entity, it may not be feasible to regulate and enforce their quality from a network provisioning point of view.

The physical interfaces for VoIP clients include general purpose hardware (e.g. PCs), PDAs, smart-phones (e.g. iPhone), dedicated VoIP boxes that usually come with a subscription to a VoIP service, and any communication devices with access to the Internet.

- VoIP clients running on general purpose hardware have the benefit over dedicated hardware when adding new features like unified messaging and video conferencing. They can simply utilize existing interfaces, such as keyboard, monitor, and camera. These soft clients (such as Skype) also provide free VoIP calls to users, which help grow their user base and eventually

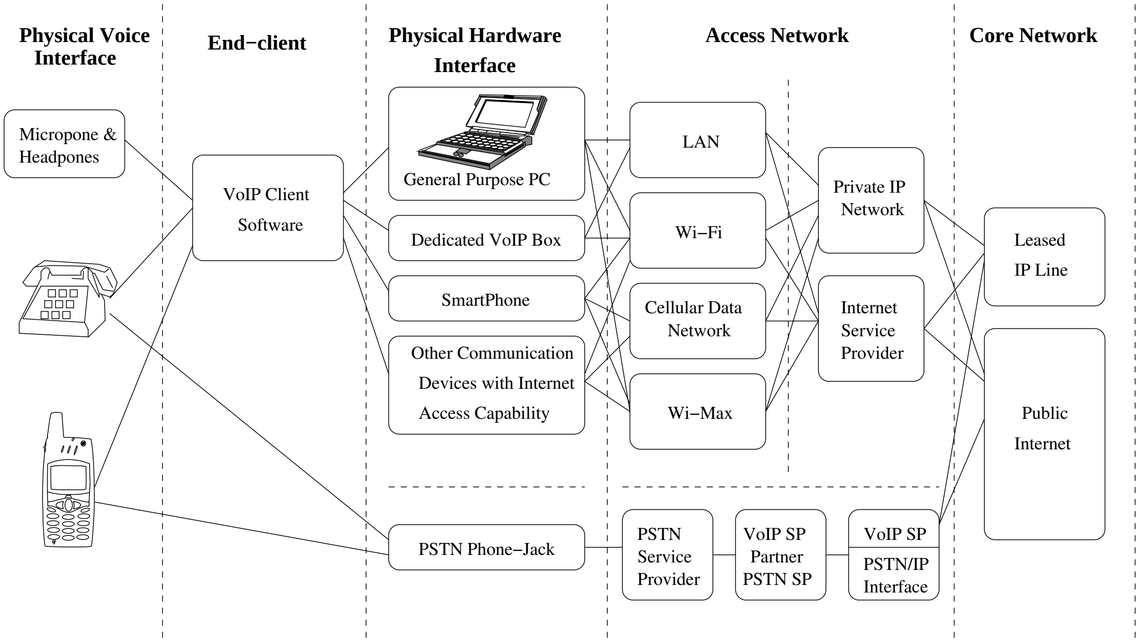


Figure 2.1: VoIP system infrastructure depicting different physical interfaces, hardware used and networks traversed.

increase their usage, which generates revenue for the company (e.g. SkypeIn and SkypeOut).

- Systems using dedicated hardware (such as Vonage) promote the subscription of their service by giving away free hardware that can be easily interfaced to regular phone adapters. However, the requirement of broadband access limits potential subscribers to people who own computers with broadband connections. These people also have the option of using software-based systems for free.
- Both need to enable calling to or receiving calls from PSTN phones in order to ensure their wide spread adoption. Further, they need large-scale partnership with local telecommunication companies in order to interface to PSTN networks. Such requirements limit the scalability of the services and increase the barrier to enter the market, favoring a small number of VoIP companies to succeed in the long run.

VoIP nodes can utilize a variety of hardware interfaces, such as laptop computers, PDAs, smartphones, and dedicated VoIP handsets (see Figure 2.1). Independent of the interface, there is a software client that communicates with a counterpart client over a network with a best-effort end-to-end service. The clients support an interactive conversation by processing speech packets and by shielding users from network imperfections. The popularity of these clients that run on general purpose computers has grown in recent years, even more than VoIP services that run on dedicated boxes. Furthermore, third-party vendors have developed phone-like handsets (like Skype phone and Google phone) that can connect directly to Wi-Fi networks without a PC. These devices improve the ease of use and make the technology more transparent to users.

In general, two parties in an interactive conversation take turns speaking and listening. A conversation, therefore, consists of alternating one-way speech segments and silence periods. From a user's perspective, the conversational dynamic of a face-to-face conversation is different from that over a communication system with delays. In a face-to-face conversation, users have a common reality in the perception of the sequence and the timing of speech segments. A speech segment of one user is separated from the other user's segment by a silence period that is identically perceived by both users.

When a conversation is conducted over the Internet, speech segments experience delays, jitter, and packet losses. In Chapter 2.2 we present our observations on Internet traffic behavior based on extensive experiments conducted on the PlanetLab [44]. Based on the network behavior, in Chapter 2.3 we present the dynamics of a conversation over a channel with delays. The quality of a speech conversation depends on two factors that are directly or indirectly perceived by users: the quality and the latency of the one-way speech segments received. The delays incurred in the reception of speech segments also lead to asymmetry in silence durations in between turns and cause inefficiency in communication as compared to a face-to-face setting. In this case, each user will experience speech segments that are separated by silence periods of alternating long and short durations. This asymmetry may lead to a perception that the other user is responding slowly to the conversation.

Due to path-dependent, non-deterministic, and non-stationary network behavior of the Internet, the factors that affect conversational quality may vary over time and are counteracting to each other. For example, the one-way quality and the delay incurred in the transmission of speech segments from the mouth of a speaker to the ear of a listener (MED) are counteracting to each other in their effects on conversational quality. On one hand, speech segments will have a higher chance to be received and consequently better one-way quality if the receiver waits longer. However, this additional delay will result in a longer MED, which worsens the asymmetry of delay durations and leads to lower perceptual quality. Another effect is the lower efficiency in completing the same conversation when compared to a face-to-face setting. The impact of delays on conversational quality also depends on the turn-switching frequency. For instance, an MED of 300 msec can significantly degrade the symmetry and efficiency of a conversation if participants take frequent turns, but will be virtually imperceptible if users take a long time (say 10 sec) in each turn. Variations in one-way quality and latency may cause double-talks and interruptions that further degrade conversational quality. The trade-offs between delay and one-way quality must also be dynamic and respond to changing conditions. The receiver may either adapt its MED in order to achieve a consistent speech quality, or keep a consistent MED but allow the speech quality to vary.

The Architecture of a VoIP Client commonly includes POS and LC components to conceal losses and jitter in packets received by the transport protocol. Unconcealed losses in this layer are further handled by the speech codec.

Existing VoIP systems usually employ redundancy-based LC algorithms for recovering losses when using the RTP/UDP protocols. However, as discussed in Chapter 3, none of the previous approaches considers delay-quality trade-offs for delivering VoIP of high perceptual quality to users. Previous LC algorithms based on analytic loss models [40, 45] do not always perform well, as these models may not fully capture the dynamic network behavior and do not take into account the LC strategies in codecs. Existing POS algorithms based on open-loop heuristic functions [45] may not be robust under all network conditions, whereas closed-loop approaches [40] are difficult to optimize without a good intermediate metric. Some recent approaches [6, 59] have employed an end-to-end objective metric, such as the E-model [22], as their intermediate metric.

In this section, we describe the architecture of a general VoIP client and its interactions with the network and human users (Figure 2.2). Its main components include the playout scheduler that controls the MED, and the loss-concealment scheme and the speech codec that affect the quality of the speech signals received. A detailed presentation of the previous work and our approach in the design of these components is presented in Chapter 3.

Playout scheduling (POS). To buffer irregular packet arrivals (jitter) and to achieve smooth playback of speech frames, VoIP systems commonly employ a *playout scheduler* at the receiver. POS maintains a consistent MED by controlling the time waited by each packet received in such a way that the utterance is played at the same pace it was spoken. Depending on the network-delay and jitter conditions, some packets may arrive later than their scheduled times, and this information is unavailable for the decoder in generating the speech waveform. In response to changes in network conditions, it is possible to delay the playout schedule and adjust the MED in order to allow more packets to have a chance to arrive in time.

There have been several studies on the design of adaptive POS algorithms, most of which use previously observed network conditions to decide on the MED at the beginning of a talk-spurt (portions of a speech segment separated by a short duration of silence). There are also algorithms that adjust POS within a talk-spurt by using techniques commonly called time-scale modification [36]. POS algorithms can be as simple as an open-loop heuristic or as complex as optimizing an end-

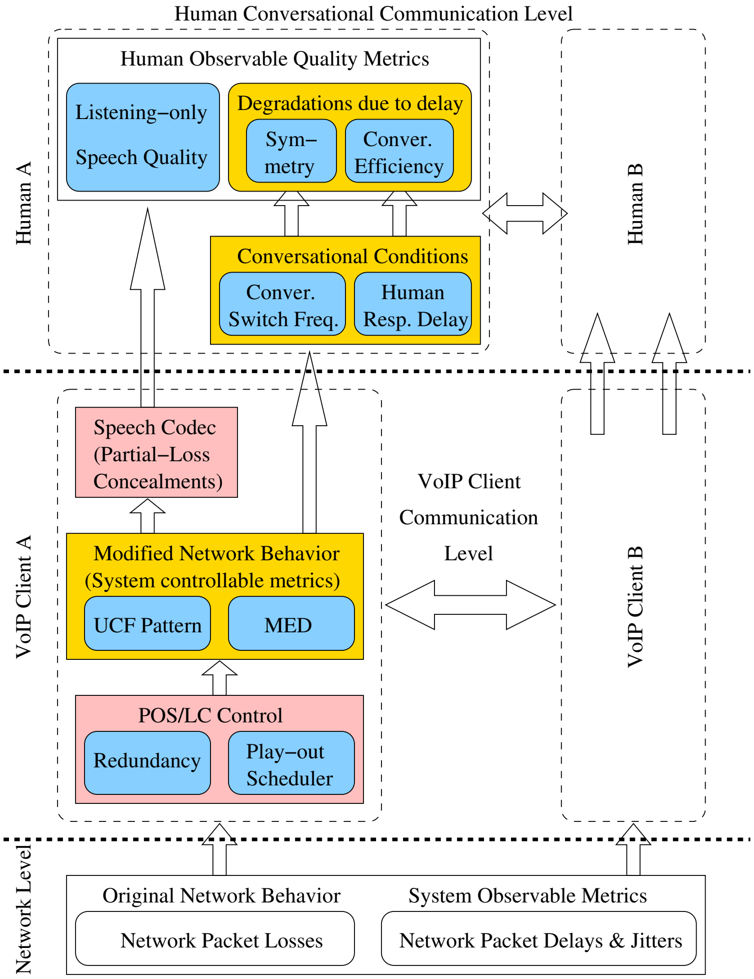

Human Conversational Communication Level
Human Observable Quality Metrics
Listening−only Speech Quality
Degradations due to delay
Sym−metry
Conver. Efficiency
Conversational Conditions
Conver. Switch Freq.
Human Resp. Delay
Human A
Human B
Speech Codec (Partial−Loss Concealments)
Modified Network Behavior (System controllable metrics)
UCF Pattern
MED
POS/LC Control
Redundancy
Play−out Scheduler
VoIP Client A
VoIP Client Communication Level
VoIP Client B
Network Level
Original Network Behavior
Network Packet Losses
System Observable Metrics
Network Packet Delays & Jitters

Figure 2.2: Architecture showing the interactions among the VoIP clients, the network, and the communicating humans.

to-end objective metric that estimates the conversational quality (e.g. E-Model [22]). However, it is difficult to deduce the POS algorithm of the four VoIP clients studied in this thesis, since we cannot decode the content of the packets received in each client.

Loss concealment (LC). Since fluctuations in the loss rate strongly affect the consistency of speech quality, VoIP clients commonly employ a redundancy-based loss concealment scheme for limiting unconcealable losses. These usually require the coordination of the sender and the receiver clients in order to control losses at the receiver in a closed-loop fashion. They may require sending auxiliary information in every speech packet in case previous packet(s) were lost.

The redundancy can be as simple as duplicating previously transmitted packets or as complex as speech-dependent FEC, multi-description, or layered coding. One simple scheme is to piggyback a redundant copy of one or more of the past packets transmitted when sending the current packet. This is feasible in VoIP because most speech frames are small, and the MTU (maximum transmission unit) in the Internet is large enough to allow multiple frames to be transmitted in the same packet. Piggybacking-based loss concealments will incur additional playout delays because the POS must wait for those redundant copies to arrive before declaring that the original packet is lost.

The trade-offs among MED, redundancy degrees, and the *unconcealable frame rate* (UCFR) under various network conditions are presented in the next section. As is discussed in Section 2.2, the MED and the redundancy needed for achieving a given UCFR are connection dependent. In particular, piggybacking-based LC is not always effective in high-jitter connections.

Some of the VoIP systems we have tested exhibit increases in their packet payload in response to network losses. However, we cannot deduce the specific LC scheme used or evaluate their effectiveness because these systems may use proprietary codecs of variable bit rate, which naturally change their payload according to the speech input. Moreover, the contents of their packets are unavailable because they are encrypted.

Speech codecs. These are designed to reduce the bit rate needed when transmitting a speech waveform. They range from simple-waveform codecs that mimic the shape of a waveform, to complex CELP and hybrid codecs that model speech production in humans. They generally aim to maintain high speech quality, while reducing the bit rate by 5 to 20 times with respect to the original PCM representation.

Speech codecs were initially designed for wireless and trans-oceanic speech transmissions with scarce network resources. Their role in VoIP systems, however, is different because network utilization in the current Internet is less of a concern. The more important feature in the VoIP context is its robustness to lost or late packets at the decoder. It is difficult to test the codecs used in existing VoIP systems because they are either proprietary (such as iSAC developed by GIPS [58] and used in Skype and Google Talk) or unknown. iSAC is a variable bit-rate wide-band speech codec that is claimed by GIPS to deliver quality better than PSTN. Even though iLBC, which is also developed by GIPS, is standardized in IETF RFC 3951 [1], it is not used in any of the four systems we evaluate in Chapter 4; thus, it is not useful in our analysis.

2.2 Network Environment

Our experiments show that public IP networks exhibit path-dependent, unreliable, and time-varying characteristics. Table 2.1 summarizes the statistics of 11 Internet connections collected on the PlanetLab [44] in 2007.

PlanetLab is an overlay network formed by academic institutions and industrial research labs to provide an open platform for developing, deploying and accessing planetary-scale services. It currently consists of more than 1000 nodes in more than 500 sites around the world. The cites are mostly concentrated on the east and west coast of the United States, as well as Europe, South Korea and Japan. There are a very small number of nodes in Africa, middle East and Australia.

As the nodes reside in academic institutions and industrial research labs, they are, in most cases, connected to the Internet backbone with high bandwidth links. Furthermore, as the nodes form an overlay network, each node can only be accessed by another PlanetLab node. Access from outside the network is only possible for the purpose of controlling the experiments conducted by an authorized researcher using a secure connection. Thus, the communication between PlanetLab nodes do not suffer from the last mile bandwidth restrictions and congestion conditions that typical home users experience. For these reasons, the network measurements collected in PlanetLab is not typical of the conditions observed by home or mobile users. As conducting network measurement experiments between home users and other users are not practical and scalable, we utilize the measurements collected in PlanetLab in our studies, noting that the home or mobile users may experience higher packet loss, delay and jitter.

Using the Linux platform provided by PlanetLab, we have developed software to collect packet delay, loss and jitter information. The software is deployed on multiple nodes in the PlanetLab network and obtains this information by simultaneously transmitting and receiving UDP packets. The transmissions are configured to exhibit characteristics that are typical of transmissions of VoIP systems in terms of the packet payload size and packet period.

We have collected traces of packets by embedding the source timestamp along with a sequence number on each packet and by comparing against the receiver timestamp to identify losses and one-way delay of each packet. The nodes are synchronized with NTP, which provides accuracy in the order of miliseconds which is adequate for our purpose.

The statistics are based on 71 unique connections among 22 nodes, 6 of which are in North America, 8 in Europe, and 8 in Asia. Since 59 of these connections are inter-continental, the observations deduced should be interpreted accordingly. To allow the data to be used for both two-party and multi-party VoIP, we set up each as a broadcast connection that sends 7 messages to 7 destinations simultaneously. Our results show that, at the packet rate evaluated, there was little

Table 2.1: Internet traces collected on the PlanetLab in July and August, 2007.

Set	DL	JT	LR	Hour	Source		Dest.	Mean DL (ms)		JT60 (%)		LR (%)	
	(L/H/M)			(CST)	Location	IP Address	(S,A,U)	Min	Max	Min	Max	Min	Max
1*	L	L	L	20:00	CA,USA	169.229.50.14	(1,2,4)	42.2	94.6	0.00	0.15	0.00	0.00
2	H	L	L	18:00	China	219.243.201.77	(0,3,4)	107.3	190.4	0.00	3.5	0.00	0.01
3*	H	L	H	23:00	Hong Kong	137.189.97.18	(0,3,4)	101.2	204.3	0.00	1.64	14.7	22.7
4*	H	H	L	22:00	Taiwan	140.112.107.80	(1,3,3)	198.0	280.4	68.3	72.2	0.14	0.22
5	M	L	L	20:00	Czech	195.113.161.82	(2,3,2)	56.0	158.4	0.45	0.97	0.00	3.39
6*	M	H	L	17:00	CA,USA	171.66.3.181	(2,2,3)	74.9	170.9	5.2	6.2	0.00	4.33
7	M	L	H	1:00	Hong Kong	137.189.97.18	(1,3,3)	85.4	195.9	0.00	1.6	15.3	22.8
8*	M	L	M	11:00	Canada	198.163.152.229	(2,2,3)	52.4	147.3	0.00	0.83	0.00	16.9
9*	M	M	L	5:00	UK	128.232.103.203	(2,3,2)	26.5	139.9	0.00	8.10	0.00	3.2
10	H	M	M	1:00	China	211.94.143.61	(0,4,3)	103.7	198.9	1.2	6.6	1.9	8.6
11	M	M	M	8:00	Hungary	152.66.244.49	(3,2,2)	22.6	190.6	0.00	79.0	0.00	25.1

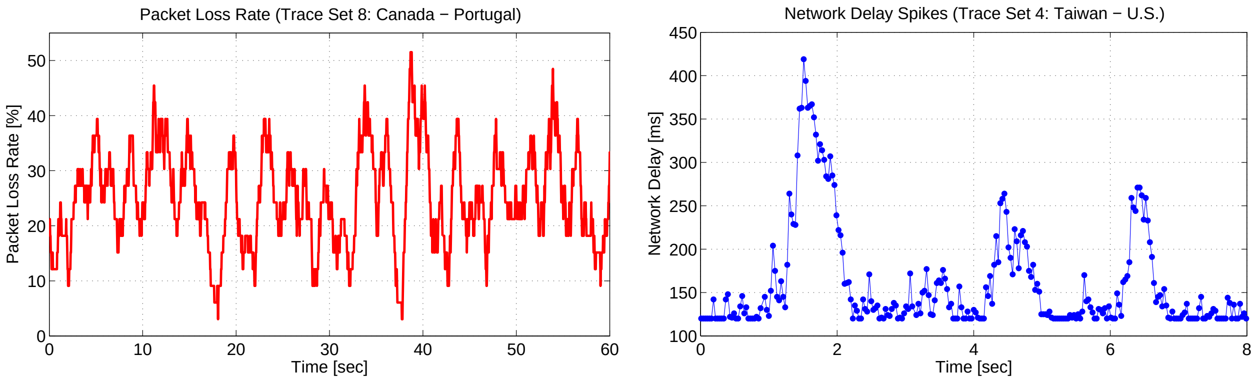
Keys: Each set is based on a broadcast connection from one source to 7 destinations (duration 10 min; packet period 30 ms; DL: delay; JT: jitter; JT60: jitters larger than 60 ms with respect to mean delay; and LR: loss rate). Delays are classified into low ($<$ 100 ms), high ($\geq$ 100 ms), and mixed (a combination of both). Similarly, jitters are classified into low ($<$ 5% in JT60), high ($\geq$ 5% in JT60), and mixed; and losses into low ($<$ 5%), high ($\geq$ 5%) and mixed. Each destination is listed by a triplet of three numbers (# in aSia, # in America, # in eUrope). '*' indicates a connection used in subjective tests.

effect on packet losses and delays by broadcasting to multiple destinations. We have classified each connection by a triplet (delay, jitter, loss). About one third of these connections are low delay, low jitter, and low loss (L,L,L). Except for (L,H,H) and (H,H,H), there are at least four connections in each of the remaining six classes.[1] For subjective testing, we have chosen a representative connection in each of the six classes (indicated by '*' in Table 2.1). The tests were conducted using one trace out of each of the 6 classed of traces in Table 2.1 with distinct network conditions. In addition, we include a new trace on an ideal network with no loss and no delay.

a) There are two events that cause the quality of received speech frames to be degraded. In some cases, packets carrying speech frames may be lost in the network, either in single packets or in multiple consecutive packets. It is also possible for packets to be delayed beyond a point when they are too late for playback. In both cases, the receiver will not be able to recover these packets without redundant transmissions.

b) The loss behavior of the Internet can change in a matter of seconds, and stationary models [4] are not helpful for tracking these fast-changing conditions. Figure 2.3a depicts the temporal changes in packet losses for a connection with medium loss rate, where loss rates are calculated

[1] We cannot find any connections in the (L,H,H) and (H,H,H) classes because high jitter and high losses are unlikely to happen in low-delay or highly congested connections.



a) Trace set 8: Canada ⟶ Portugal b) Trace set 4: Taiwan ⟶ US

Figure 2.3: Traffic behavior (delay and loss rate) of two PlanetLab connections in Table 2.1.

over a sliding window of one second. The data shows that the loss rates fluctuate between 3% and 51% and are unpredictable. The use of a one-second averaging window is meaningful because several phonemes can be uttered within this interval, and the words can be unintelligible if several consecutive packets are lost.

c) The loss behavior of a connection also varies depending on the hour of the day. Figure 2.4 depicts the average loss rate of three connections between US, Europe and Asia where 10-minute experiments have been conducted at the beginning of each hour for 22 hours. The figure shows that the connection between Europe and Asia exhibits a consistent loss rate of 5% for most hours, where there is a more than twofold increase in loss rate for 3 hours. The connection between US and Europe exhibits almost no losses for 7 hours of the experiment, but can reach up to 5.3% depending on the time of day. The connection between US and Asia exhibits the most variation in loss rate depending on the hour of the experiment, where the loss rate averaged over a 10-minute period can range from 0.6% to 30.3%.

d) Packets transmitted in the Internet experience path-dependent delays before reaching their destinations. Most intra-continental connections in North America and Europe have mean propagation delays of less than 75 ms, whereas most inter-continental connections and some intra-continental connections within Asia exhibit delays in excess of 150 ms.

e) The delays experienced by IP packets can change quickly in a short interval (called jitter), increasing by hundreds of milliseconds from the delay of the previous packet in a packet period of 30 ms. These conditions are commonly referred to as delay spikes, which indicate sudden congestion in an intermediate router on the path of the packets. Figure 2.5 depicts the network delay behavior for a connection between US and China on 3 time scales, where network delay spikes are observed.

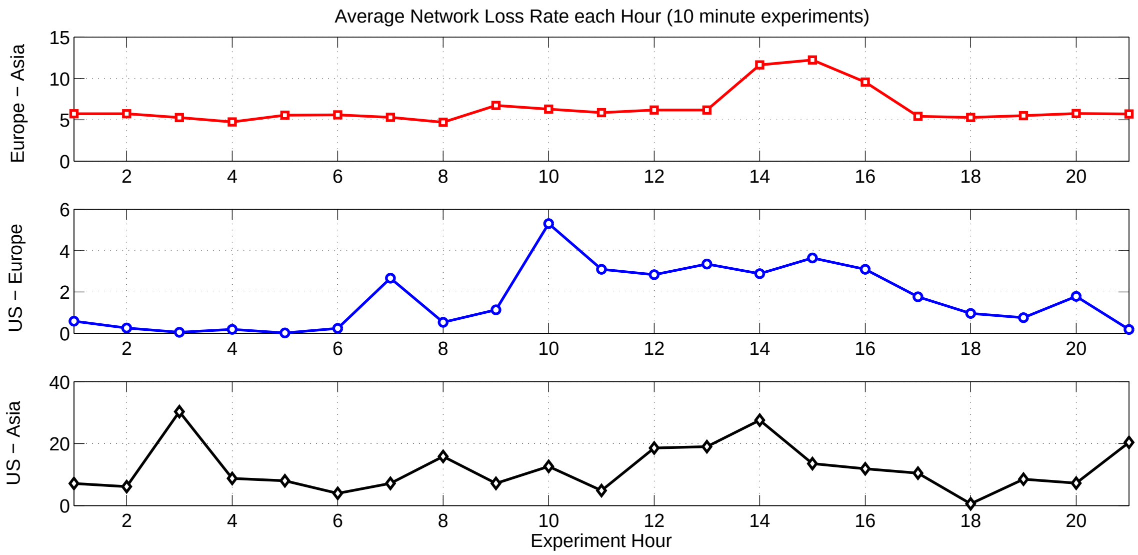


Figure 2.4: Time-dependent network loss behavior for three PlanetLab connections between US, Europe and Asia. The figure depicts the average packet loss percentage as a function of the experiment hour.

When congestion is resolved, multiple consecutive packets can be received almost instantaneously. These consecutive packets experience less and less delay until the delay value reaches the level before the spike. This behavior indicates that the congested router has emptied its buffers quickly after the congestion.

Figure 2.3b depicts the temporal changes in network delays for an international connection between Taiwan and US with high jitter. It shows that several spikes can occur within one second, either in an individual or in a coupled fashion.

f) To limit degradations caused by jitter, VoIP clients commonly employ playout schedulers (POS) that adjust the time waited before playing out the received speech frames. These schemes will incur additional playout delays and extend the MED. Figure 2.6 depicts the trade-off between the Unconcealed Frame Loss Rate (UCFLR) and MED as controlled by the POS for the two network connections depicted in Figure 2.3.

Figure 2.6a shows that in network conditions where there is a significant amount of individual or consecutive packet loss in the network, increasing the redundancy degree and increasing the MED correspondingly results in a significant reduction in UCFLR. It should be noted that increases in MED without increasing redundancy or increases in redundancy without increasing MED does not result in a significant improvement in UCFLR. Thus, it is clear that the redundancy degree and MED should be controlled either jointly or in a coordinated fashion.

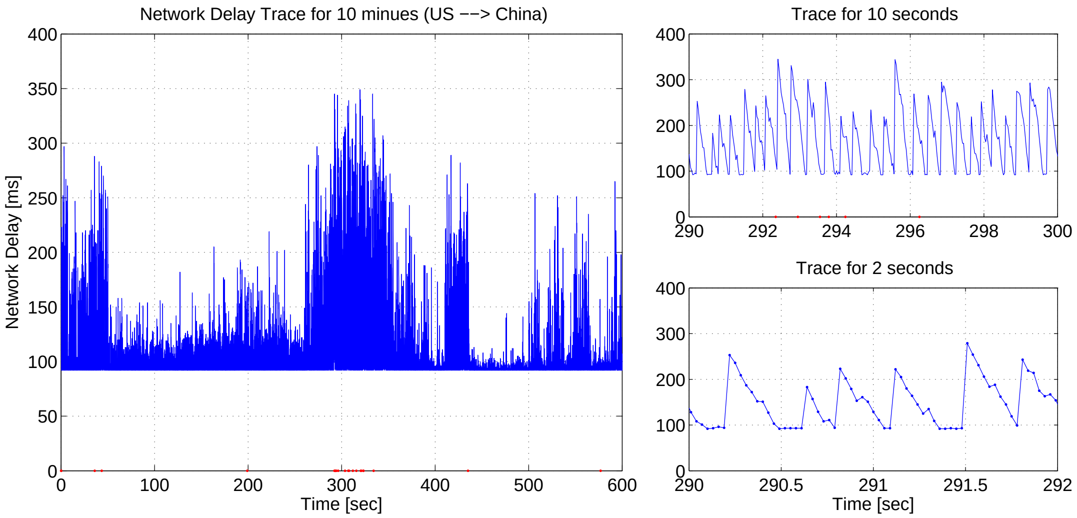
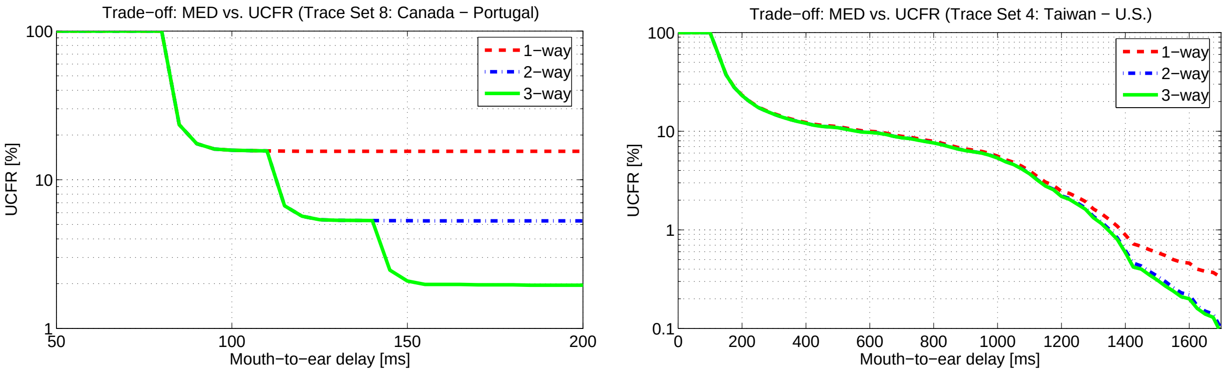


Figure 2.5: Network delay behavior for a PlanetLab connection between US and China.



a) Trace set 8: Canada → Portugal b) Trace set 4: Taiwan → US

Figure 2.6: UCFR and MED trade-off of two PlanetLab connections in Table 2.1.

Figure 2.6b depicts that in network conditions where there is no significant packet loss observed, increasing redundancy does not affect UCFLR significantly despite the increase in MED. We also observe that in connections exhibiting network jitter, increasing MED up to the maximum network delay observed reduces UCFLR. However, as we have discussed in Chapter 1, when making a MED decision, the perceptual benefits of reducing UCFLR that is already small should be considered against further degradations due to increases of MED.

In the next section, we investigate the effects of delays on conversational dynamics.

2.3 Effects of Delays on Interactive VoIP Conversations

In a two-party conversation, each participant takes turns in speaking and listening [47, 66, 5], and both perceive a silence duration (called *mutual silence* or *MS*) between turns when the current speaker ceases the floor and the listener takes over. A conversation, therefore, consists of alternating speech segments and silence periods.

In a face-to-face setting, both participants have a common reality of the conversation: one speech segment is separated from another by a silence period that is identically perceived by both. However, when the same conversation is conducted over the Internet, the participants' perception of the conversation is different due to delays, jitter, and losses incurred on the segments during their transmission [53, 52]. In this section, we analyze the effects of delays on the conversational dynamics.

As described in Chapter 1, the quality of a conversation over a network connection depends on two factors that are directly or indirectly perceived by users: the quality of the one-way speech received and the delay incurred from the mouth of the speaker to the ear of the listener. LOSQ can be measured by measures that model the perception of speech such as PSQM, PSQM+ and PESQ [27]. There are also measures that are variations of signal-to-noise ratio (SNR) that measure the differences in original and degraded signals. One such measure is articulation index which is used by audiologists to predict the amount of speech that is audible to a patient with a specific hearing loss. However, such measures are not commonly used in the evaluation of one-way speech in the context of speech coders utilized in VoIP systems. In this study we utilize PESQ as the main measure for LOSQ as it is shown to exhibit high correlation with subjective quality evaluations for VoIP applications and is indicated to superceed PSQM and PSQM+ by ITU.

When the connection has delays, the MSs perceived by a participant consist of alternating short and long silence periods between turns [53].

Human response delay and mutual silence. We first define the silence durations observed during turn-taking. Since there are two perspectives, we start from the perspective of the current speaker. We define *human response delay* from B's perspective (HRD_B) as the period after B perceives that A has stopped talking and before B starts talking, during which B thinks about a response to A's speech. However, the same event is perceived to be longer from A's perspective, which we define as MS_A^j, the *mutual silence* before the j^{th} single-talk speech segment (ST_j) is spoken/heard. Let $MED_{A,B}^j$ be the MED between A's mouth and B's ear for transmitting ST_j

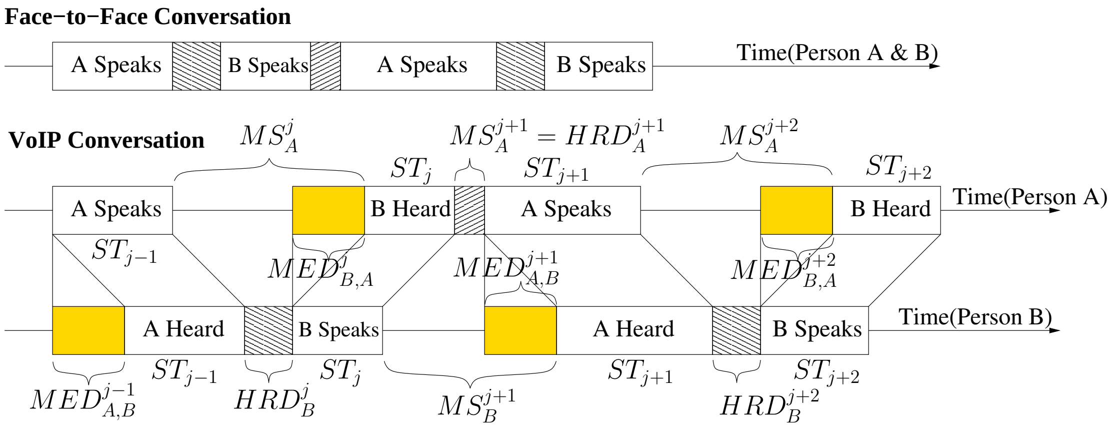


Figure 2.7: Conversational dynamics in a face-to-face and two-party VoIP setting.

from A to B; the relation among MS, HRD, and MEDs is as follows (see Figure 2.7):

$$\begin{aligned} MS_A^j &= MED_{A,B}^{j-1} + HRD_B^j + MED_{B,A}^j, \\ MS_A^{j+1} &= HRD_A^{j+1}, \\ MS_B^j &= HRD_B^j, \\ MS_B^{j+1} &= MED_{B,A}^j + HRD_A^{j+1} + MED_{A,B}^{j+1}. \end{aligned} \tag{2.1}$$

During a VoIP session, a user does not have an absolute perception of MED because the user does not know when the other person starts talking. However, by perceiving the indirect effects of MED, such as MS, the user can deduce the existence of MED. This asymmetry leads to a perception that each user is responding slowly to the other, and consequently results in degraded efficiency and perceptual quality [53].

Conversational quality cannot be improved by simultaneously improving LOSQ and reducing MED. A longer MED will improve LOSQ because segments will have a higher chance to be received, but will worsen the symmetry of MSs. Figure 2.8 shows the delay-quality trade-off and a suitable MED with the best quality. This trade-off also depends on the turn-switching frequency [32, 53] and on changes in network and conversational conditions [4].

In the rest of this section we present other metrics that capture the effects of delay on conversational dynamics and that can be perceived by users.

Conversational symmetry. Symmetry is related to the activities of the entities that affect each other. In the context of speech communication, turn-taking is the interaction between the participants. For this reason, we define conversational symmetry (CS) based on the user perceptible MS

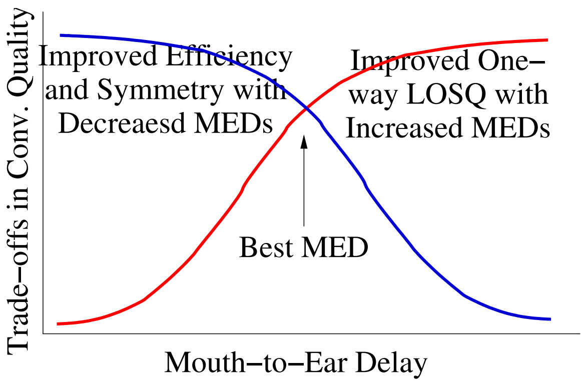


Figure 2.8: Trade-off considerations.

between turn-taking. Since the perception of temporal events is user dependent, we define *conversational symmetry (CS_A)* of A to be the ratio of the maximum and the minimum MSs experienced by A recently (say in a past window of time):

$$CS_A = \frac{\max_j MS_A^j}{\min_j MS_A^j}, \qquad\qquad CS_B = \frac{\max_j MS_B^j}{\min_j MS_B^j}. \tag{2.2}$$

In a face-to-face conversation, MS and HRD are perceived to be equal; thus, CS_A and CS_B are approximately 1. However, as the round-trip delay increases, the silence periods perceived during turn taking are no longer symmetric. If the asymmetry in the perceived response times increases, humans tend to have a degraded perception of symmetry that will result in the degradation of the quality of the conversation. One possible effect is that, if A perceives that B is responding slowly, then A tends to respond slowly as well.

Conversational efficiency. Another effect of communicating over a channel with delays is that it takes longer to accomplish a task with respect to the same conversation in a face-to-face setting. Since users are charged according to the duration of the conversation, a task will cost more for a channel with longer delays. This effect is especially pronounced in international and mobile phone calls, in which both the network delay and the per-minute charge are higher. We define *conversational efficiency* (CE) as the ratio of the duration the participants actively speak or listen to the total duration of the call:

$$CE = \frac{\text{Total Speaking Time} + \text{Total Listening Time}}{\text{Total Time including Silence}}. \tag{2.3}$$

Table 2.2 shows the statistics for five face-to-face conversations of different average ST dura-

Table 2.2: Statistics of five face-to-face conversations.

Conversation Type	Avg. single-talk duration, $\overline{ST}$	Avg. HRD duration, $\overline{HRD}$	# of switches	Total Time	Switching Frequency
1 - Counting numbers	311 ms	220 ms	11	6.2 sec	106.4/min
2 - Confirming numbers	1,334 ms	450 ms	9	17.4 sec	31.0/min
3 - Order Lunch	1,706 ms	552 ms	7	17.5 sec	24.0/min
4 - Dental Appointment	3,055 ms	710 ms	5	21.9 sec	13.7/min
5 - Social Conversation	5,502 ms	827 ms	3	24.5 sec	7.4/min

tions. Note that CS depends on the value of HRD; that is, if HRD is shorter, then the loss of symmetry due to MED is perceived to be more. Likewise, when ST is longer, the loss of efficiency is perceived to be less.

In summary, MS, CS and CE are user-perceived metrics that can be calculated objectively, whereas MED is a system-controlled metric that intimately affects those user-perceived metrics. CS and CE are used in later chapters as candidate objective metrics in our training to predict comparative subjective quality.

Double talk. In case of a large spike in network delays, if the system does not detect the spike and adapt its MED in time, a considerable number of consecutive frames can be lost for a duration that is perceived by the listener. Depending on the size and frequency of the spikes, an utterance, a word, or even a sentence can be inaudible or unintelligible at the receiver due to the unavailability of frames for playout. If this scenario occurs during a speech utterance, the listener can either assume that the speaker has stopped and start uttering his/her response, or ask the speaker to repeat the last words or sentence. In either case, the speaker, unaware of the difficulty of the listener, would most likely continue speaking and cause a collision of speeches or unintentional interruptions (double-talk) from one or both parties' perspective. Depending on the situation, the person observing the collision struggles to resolve the problem by waiting longer for the other to respond or repeat the previously spoken utterances. Further, as depicted in Figure 2.9, the point where the interruption happens may be perceived differently by the two parties, disrupting the rhythm of a natural interactive conversation and causing confusion and degradation in perceived quality. These degradations due to double-talks were observed by Brady [7] and Richards [46] in the 1970s, who conducted subjective experiments and concluded that double-talks and confusion increases with increased channel delays.

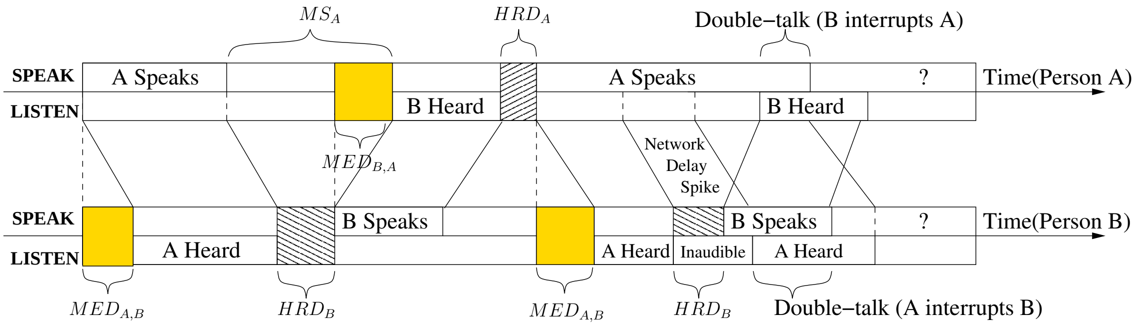


Figure 2.9: The occurrence of a double-talk due to a lack of adequate system reaction to network-delay spikes.

Adaptation of human behavior. In case of extreme difficulties in communication, such as extreme delays in getting a response or extremely low listening quality, users can either hang up and re-dial or change their talking style in order to ease the efforts needed. The style change usually involves talking slowly, or talking in longer batches, or deserting the wait for acknowledgment gestures. Users who are forced to take these behavioral-adaptation measures feel that their additional effort significantly decreases their satisfaction of the call. Further, this behavioral change might not be acceptable in some languages, cultures, and business-related or mission-critical communication tasks. We are, however, not proposing objective measures to capture the effects of delay on double-talk and the adaptation of human behavior, as these effects are too subjective and are heavily dependent on the users and the conversation.

2.4 Summary

In this chapter, we have presented our VoIP system architecture, including the usages of the VoIP technology and the speech processing and control components of a VoIP client.

We have also summarized our observations on the network conditions found on the Internet and how these conditions relate to the design of VoIP systems. The observation that the connections exhibit connection-dependent and time-varying loss and delays leads us to the conclusion that adaptive loss concealment and play-out scheduling schemes are needed in the design of VoIP systems in order to achieve robust performance against changing conditions. Furthermore, it is established that the coordination of LC and POS is needed to reduce the overall UCFLR, which represents the effects of both network loss and delay spikes, to an acceptable level.

Thirdly, we have formalized the definitions related to conversational dynamics and illustrated the effects of delay on a VoIP conversation in comparison to a face-to-face conversation. We have presented two metrics that characterize the conversational conditions that are perceptible by users of the VoIP system. We have also presented a list of conversations with varying temporal characteristics that are used in later parts of the thesis. Lastly, we have presented the effects of delay on a conversation that are hard to characterize as a measurable quantity. One of these conditions is the occurrence of double-talk where both parties in the conversation speak at the same time, from one or both persons' perspective.

CHAPTER 3

PREVIOUS WORK AND OUR APPROACH

In this chapter we present the relevant previous work and our approach for the problems studied in this thesis.

The focus of this chapter is mainly on the evaluation of conversational quality, as it is an integral part of our work in both the evaluation of VoIP systems as well as the design of VoIP system components that control run-time parameters to achieve high conversational quality.

In the first section of this chapter, we present the previous work on the evaluation of conversational quality. Even though there are differences in the evaluation of conversational quality at a system-level after the design and at a component level during the design, we present the related previous work together, as this distinction is rarely made in the previous work. However, in the presentation of our approach for this problem, we clearly separate the two cases and formulate our approach in different sections in this chapter.

In the first section, we also survey previous work on the evaluation of Quality of Experience (QoE) on problems other than the design of POS for a real-time VoIP system.

In Section 3.2, we present our systematic approach for evaluating conversational quality of VoIP systems, where we incorporate objective as well as subjective evaluations and a method to predict subjective preferences from objective quality measures.

In Section 3.3, we present the relevant previous work on the design of VoIP system components, such as play-out scheduling (POS) and loss concealment (LC). Some of this work utilizes the evaluation methods we have surveyed in Section 3.1.

In Section 3.4, we present our approach for evaluating conversational quality in the design of VoIP system components that control run-time parameters in order to achieve high conversational quality. We mainly focus on the design of a POS control scheme to achieve high conversational quality. Our approach includes the development of a comparative subjective model, an efficient scheduling algorithm to conduct subjective tests, and a method to learn and generalize such tests over a multitude of operating curves under different operating conditions. We leave the presentation of our design choices in loss concealment and speech codec for the newly designed VoIP system to Chapter 8.

3.1 Previous Work on the Evaluations of Conversational Quality

As is discussed in Chapter 1, a conversation consists of alternating speech segments and silence periods. In this context, Richards [46] has identified three factors that influence the quality of service in telephone systems: difficulty in listening to one-way speech, difficulty in talking, and difficulty in conversing during turn-taking. Hence, we evaluate the quality of VoIP by the quality of the one-way speech and that of the interactions [53].

3.1.1 Effects of Mouth-to-Ear Delay (MED) on conversational quality

As initially summarized in Chapter 2, MED is an important element that affects conversational speech quality [53, 48]. It has various effects on human perception through conversational symmetry and efficiency. MED consists of the delays incurred in speech encoding, packing speech frames into packets at the sender, the network, the playout buffers at the receiver, and decoding. Of these delays, the encoding, decoding and packing delays are fixed and usually negligible. The component that can be controlled by the VoIP system is the playout delay that includes the amount of jitter-buffer delay and the delay in waiting for redundant information to arrive for loss concealment.

Due to the dependency of MEDs on network conditions, users may experience different MEDs to different connections and at different times, even throughout a conversation. Thus, the effects of MED on conversational quality need to be considered as connection dependent and dynamic.

Subjective tests by Brady [7] and Richards [46] have led to the conclusions that MED affects the user perception of conversational quality, and that longer MEDs increase the dissatisfaction rate. However, their conclusions are limited when used for evaluating VoIP systems, since only a few constant delays were experimented. Subjective tests by Kiatawaki and Itoh [32] at NTT show that one-way delays are detectable, by users of a communication system, with a detectability threshold of 100-700 ms for trained crew and of 350-1100 ms for untrained subjects. The variation in detectability largely depends on the conversational task. ITU G.114 [20] prescribes that a *one-way delay* of less than 150 ms is desirable in voice communication, and that a delay of more than 400 ms is unacceptable.

The collective summary of the previous work presented above suggests that MED is an integral part of user perceived conversational quality and its importance may change with differing communication scenarios. However, none of these studies specifies a favorable trade-off between listening-only speech quality (LOSQ) and MED, when system constraints and controls require the

Table 3.1: ITU P.800.1 terminology on telephone transmission quality in terms of MOS_{aQb} (*mean opinion score*), where a can be one of $\{$**L**istening-only, **C**onversational$\}$, and b can be one of $\{$**O**bjective, **S**ubjective, **E**estimate$\}$.

Methodology	Listening-Only Conditions Tested	Conversational Conditions Tested
Subjective	MOS_{LQS}: P.800 Listening-only Tests	MOS_{CQS}: P.800 Conversational Tests
Objective	MOS_{LQO}: P.862 PESQ	MOS_{CQO}: P.562 for PSTN, not defined for VoIP
Estimated	MOS_{LQE}: Not defined	MOS_{CQE}: G.107 E-model

VoIP system to balance these counteracting components of quality.

This observation leads us to search for other standard and non-standard metrics and methodologies to evaluate the conversational quality completely.

ITU P.800.1 defines terminologies for *mean opinion score* (MOS) in order to discriminate between different conditions and different evaluation methods used to arrive at a particular type of MOS score. Table 3.1 shows the naming standard established by ITU for the evaluation of the telephone transmission quality [21]. Note that the methods for calculating some MOS types are not defined. These are placeholders for future standardization efforts that fit into the current terminology.

3.1.2 Objective measures on conversational quality

There are several recommendations for evaluating the objective conversational quality of a system in ACR (*absolute category rating*).

a) *Perceptual Evaluation of Speech Quality (PESQ)* (ITU P.862) is an objective measure for evaluating listening-only speech quality based on the original and the degraded waveforms. It has been shown [27] to have high correlations to subjective MOS results for a variety of land-line, mobile and VoIP applications as well as evaluation of speech codecs, effects of packets losses, and loss-concealment schemes.

In compliance with ITU P.800.1 [21], a conversion from PESQ in ITU P.862.1 [21] to MOS_{LQO} is defined as follows:

$$MOS_{LQO} = 0.999 + \frac{4}{1 + e^{(-1.4945 * PESQ + 4.6607)}}.$$

The benefits of using PESQ include its accuracy in predicting LOSQ in applications relevant to our studies and the repeatability of its results. However, since the calculation of PESQ requires the original as well as the degraded waveform for comparison, it can only be used in an intrusive testing where a signal known to both sender and receiver is transmitted for evaluation. Thus, for most practical scenarios, PESQ is not useful in real-time evaluations of LOSQ. Furthermore, since it only assesses the LOSQ but not the effects of delay, PESQ must be used in conjunction with other metrics when evaluating conversational quality.

Thus, in our study, we utilize PESQ as an important component of conversational quality in our off-line evaluations, knowing full well the need for other components characterizing the degradations of delay to be used in conjunction.

b) **ITU G.114** declares that a one-way delay of less than 150 ms is desirable in a speech communication applications and more than 400 ms is unacceptable for such applications. However, this recommendation only provides a ternary (desirable, acceptable, unacceptable) evaluation of the delay over a communication system, and does not provide any method to combine such preference information with the listening-only quality of speech heard by users of the system. Furthermore, G.114 does not consider the communication scenarios under which a system is evaluated.

Thus, in our study, we do not utilize this recommendation in the evaluation of real-time VoIP systems.

c) The **E-Model (ITU G.107)** was originally designed for estimating conversational quality in network planning and considers the effects of the speech encoder, packet losses, one-way delay, and echo. The E-Model uses the transmission factor R to represent conversational quality on a psycho-acoustical scale, where the effects due to different degradations are additive and defined as follows:

$$\begin{aligned} R &= R_o - I_s - I_d - I_e + A; \qquad I_d = I_{dte} + I_{dle} + I_{dd}, \\ I_{dd} &= \begin{cases} 0 & \text{if } T_a \leq 100 \text{ ms}, \\ 25\left[(1+\left(\log_2 \frac{T_a}{100}\right)^6)^{\frac{1}{6}} - 3\left[1+\left(\frac{\log_2 \frac{T_a}{100}}{3}\right)^6\right]^{\frac{1}{6}} + 2\right] & \text{if } T_a > 100 \text{ ms}, \end{cases} \end{aligned}$$

where R_0 is the basic SNR, and I_s (*resp.*, I_d, I_e, and A) is the simultaneous impairment (*resp.*, delay impairment, equipment impairment, and advantage) factor. The delay impairment factor is further divided into I_{dte} and I_{dle} that, respectively, estimate the impairment due to the talker and listener echoes, and I_{dd} that estimates the degradation caused by too-long absolute delay even with perfect echo cancellation. Based on R, MOS_{CQE} of the E-model is defined as:

$$MOS_{CQE} = 1 + \frac{35R}{10^3} + \frac{7R(R-60)(100-R)}{10^6}.$$

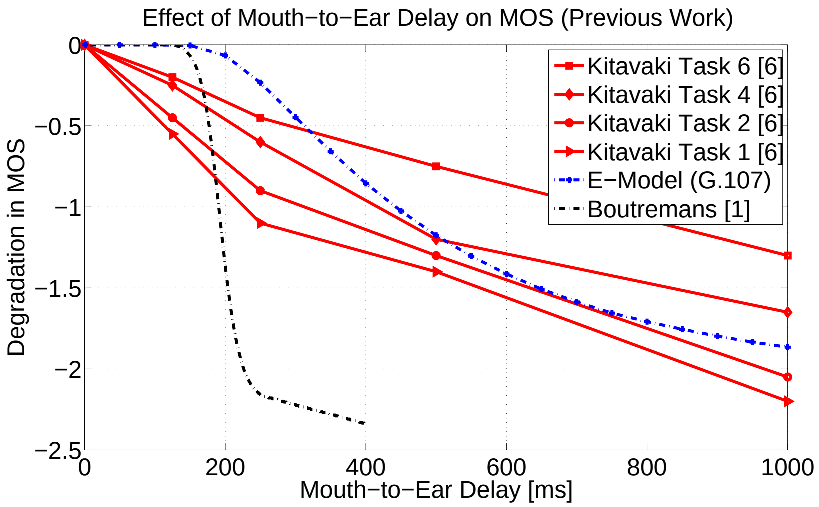


Figure 3.1: Effects of MED on MOS [32], E-model [22], and a conversion by Boutremans and Boudec [6].

Figure 3.1 depicts the effect of MED on MOS in the E-model for a perfect listening-only speech. The E-model oversimplifies the evaluation of conversational quality because it assumes the independence and additivity of degradations due to LOSQ and delay. For example, the E-model makes the implicit assumption that the same MED affects quality in the same way for conversations with slow or fast turn switching frequency and for conversations with low or high LOSQ. This oversimplifies the situation because, according to subjective evaluations, there is less emphasis on delay when LOSQ is low, but more emphasis on the asymmetry and inefficiency of the conversation when quality is high. Likewise, the effect of MED is more pronounced in a conversation with a high turn-taking frequency.

Conversational quality calculated by the E-model is speech-independent and is based on tabulated values of the effects of the codec used and packet losses in the average sense. Furthermore, the E-model also does not take into account the variations in speech quality and delay, which are known to be important in subjective evaluations. Thus, the E-model is not adequate for capturing conversational quality in a real-time conversation.

A number of extended models [59, 60, 6, 62, 39] were developed that inherit similar limitations as the E-model in evaluating conversational quality. A *combined E-model and PESQ* [59, 60] was proposed to incorporate PESQ in the E-model in order to represent the impairments due to the codec and losses. However, the model still assumes the additivity of the degradations and does not address other issues of the E-model. To overcome the need to use both the original and the degraded waveforms in calculating PESQ, a subsequent study uses regression models to predict PESQ on-line. Boutremans and Boudec [6] proposed a utility function to represent the effects of MED when choosing FEC in which a conversation is perceived to be half-duplex and quality degrades suddenly after some MED threshold. Their goal was to incorporate the effect of MED on the choice of FEC, rather than study its effects on conversational quality. Based on the NTT study [32], Markopoulou *et al.* [39] proposed a similar approach that incorporates into the E-model the degradation of conversational quality due to delays and conversational conditions. A modified E-model was proposed [62] to model conversational quality as a quadratic function of the degradations in one-way speech and those due to delay. However, that model too does not consider effects of communication scenarios (e.g. turn-taking frequency) or the variations in LOSQ and MED on overall perception of quality.

In summary, despite a number of extensions [59, 60, 6, 62, 39] that try to address E-model's limitations, it is difficult to extend its role beyond its original intended role of network planning and use it for evaluating conversational quality VoIP systems at run-time.

Furthermore, due to the proprietary nature of commercial VoIP systems, their evaluation using the E-model or Call Clarity Index (CCI) (described later in this section) is not even possible. Some of the numbers required in these metrics are unavailable because either the codecs are proprietary or the amount of late or lost packets at the decoder is unavailable.

Thus, in our study we do not utilize E-model as a quality metric, as it has many implicit assumptions that we aim to relax in our approach.

d) **ITU P.562** specifies the objective parameters to be collected via INMD (in-service non-intrusive measurement device) for analyzing and interpreting INMD voice service measurements in conjunction with P.561. There are four classes of systems for which different models are used for calculating quality. Class A is limited to short-delay circuit-switched routes containing analog and 64 kbps PCM components only (i.e, no low bit rate codecs) and no echo-control devices. Class B is limited to moderate-delay circuit-switched networks that include echo-control devices. Class C is for use in long-delay circuit-switched networks that may include signal processing devices, such as echo control and speech compression (e.g. ADPCM), but no speech encoders (e.g. LPC). Class D is for use in long-delay packet-switched networks that may include signal processing devices, such as echo control and speech compression, possibly non-linear and time-variant.

For the first three classes, the *Call Clarity Index* (CCI) was developed for estimating the customer opinion of a voice communication system in a way similar to the E-Model. The CCI model relates the objective parameters to the customer-opinion predictions. For Class D networks, there is currently no customer opinion model that considers all aspects required by P.561. Only when the IP impairments are negligible can CCI be used for class D. A parametric model that considers IP impairments is under study by the ITU.

Although CCI provides models for PSTN systems, it does not have a user opinion model for packet switched networks with long delays and with non-linear and time-variant signal processing devices, such as echo control and speech compression. As a result, it is unsuitable for evaluating the conversational quality in our study.

At this time, it is unlikely that a single objective metric can adequately capture the trade-offs among the factors that affect subjective conversational quality under all network and conversational conditions.

3.1.3 Subjective measures on conversational quality

A user's perception of a speech segment mainly depends on the intelligibility of the speech heard because the user lacks a reference to the original segment. Intelligibility, in turn, depends on factors other than signal degradations, such as the topic of the conversation, the commonality of the words used, and the familiarity of the speakers. To assess subjective conversational quality, formal mean-opinion-score (MOS_{LSQ}) tests (ITU P.800) [21] are conducted by a panel of listeners who do not participate in the conversation but only listen to pre-recorded speech segments.

On the other hand, the methodology for obtaining MOS_{CSQ} score involves two subjects conversing over a communication system in order to complete a specific task, such as arranging a meeting or describing a picture to each other. Subjects then rank quality using an absolute category rating (ACR) scale, and the opinions of multiple subjects are averaged. There are several shortcomings of this approach for evaluating VoIP. Firstly, when completing a task and evaluating the quality of a conversation simultaneously, the cognitive attention required for both may interfere with each other. Secondly, the type and complexity of the task affects the quality perception. Tasks requiring faster turn taking can be more adversely affected by transmission delays than others. Thirdly, there is no reference in subjective evaluations, and ACR highly depends on the expertise and the expectation of the subjects. Moreover, the results of current subjective tests are not useful as a measure for relative comparisons. If system A's absolute rating is better than B's rating, it does not lead to the conclusion that if subjects were asked to compare the two systems, they would have found A to be of better quality. Likewise, the difference in absolute ratings does not translate into the relative difference in quality of the two systems. Lastly, the results are hard to repeat, even for the same subjects and the same task.

In the NTT study [32] discussed earlier, subjective conversational experiments were conducted between two parties using a voice system with adjustable delays. The tasks studied range from reading random numbers, to verifying city names, and to free conversation with varying average single-talk duration. The results revealed that the degradation in MOS is more pronounced when a task requires shorter single-talk durations. Since the study did not consider the effect of losses and variations in delay, it is not directly applicable to the evaluation of VoIP systems.

ITU-T Study Group 12 has identified a lack of methods for evaluating conversational speech quality in networks and is currently conducting a study, called the *Objective Assessment of Conversational Speech Quality in Networks* [28]. Even though it is listed in the study period 2005-2008, no results have been published as of May 2010. Furthermore, it is not clear if the study will lead to an objective metric that can help design better VoIP systems. Below is a summary of the motivations, questions under study, and tasks as they appear in ITU website [28].

The study group acknowledges that the listening quality is not adequate in evaluating interactive conversation and that the probability of double-talk increases with increased delay. The study group also indicates an immediate need for a real-time, or near real-time, method for assessing overall conversational quality perceived by users, which combines conversational impairments such as delay and listening quality. The tasks listed in the document include the identification of already available measures, finding new measures to be used in combination and collecting subjective data for training.

3.1.4 Previous work on the evaluation of Quality of Experience (QoE)

In this subsection we survey previous work on the evaluation of Quality of Experience on problems other than the design of POS for real-time VoIP system. A comprehensive framework on QoE is presented in [67]. The study tries to answer what is QoE, how it can be modeled and how it is related to QoS in a measurable way. The study identifies different QoS measured at the network, system and application levels, respectively, as well as the QoE perceived at the user level. This approach is in line with our framework for evaluating quality in VoIP systems in Figure 2.2.

The study also defines QoE as a multi-dimensional construct of user perceptions and behaviors and shows the relation between QoE and QoS as a causal chain (or loop) as opposed to a one directional dependency. In this framework the environmental influences in QoS affect cognitive perception in the QoE domain, which has behavioral consequences, and in turn affects the environmental influences.

In our study, we use a similar multi-dimensional representation of user perceptible aspects (see Figure 3.5), and we acknowledge the existence of such causal chain of relationships in Chapter 2.3, while describing the adaptation of human behavior. However, it is extremely hard to measure the behavioral changes due to double-talk and excessive delay in a quantitative and repeatable way, during off-line experiments or live conversations. Thus, in our study, we assume that the system imperfections are kept as small as possible, to prevent any behavioral change in users of the system. This assumption is in line with our design goal which is to deliver VoIP conversations that closely resemble face-to-face conversations. This allows us to model the relation between system controllable metrics and the user perception of conversational quality as a unidirectional dependency relation.

There have also been studies on the design of subjective tests to evaluate QoE in problems other than VoIP system design. One such example is [9], where the goal is to evaluate the QoE of multimedia content in a way that is economical, yet statistically meaningful. The study utilizes the *crowdsourcing* methodology which essentially assigns tasks to an undefined crowd over the Internet in a distributed fashion. The study uses pair-wise comparison in their evaluation and a consistency checking scheme to weed out unreliable subjects, both of which have similar counterparts in our overall methodology. However, since it is infeasible to conduct batches of locked-step subjective tests over an anonimous crowd in the Internet, the study does not learn from previous tests to adaptively choose upcoming tests to minimize the tests conducted.

In another study [12], the goal is to choose a unique subset of pairs for different subjects that are presented in a random order to reduce the number of comparisons conducted while maintaining the accuracy of the assessment. The results presented show threefold reduction with the methodology proposed while maintaining a strong correlation with the evaluations obtained by exhaustive (or factorial) tests. The study considers different pair-wise comparison response alternatives, and considers the possibility of subjects mastering simple cognitive tasks, similar to our approaches in Chapters 6 and 7. However, in the study, the number of users responding to each comparison is allowed to be different. This causes discrepancy between the confidence of the results obtained for different comparisons and does not allow for the information obtained from different comparisons to be easily combined to reveal information on the global search space.

As the comparison pairs are randomly chosen, it is not clear if the result of the comparisons would lead to the prescribed accuracy when all subjects are complete. This issue can be circumvented in cases where exhaustive tests are conducted to verify the accuracy of the random tests; however, since there is no running estimate of the confidence of the overall result, there is no well-defined stopping criteria for the tests. Thus, when exhaustive tests are not conducted, the accuracy of the random pair comparison cannot be confidently predicted. Similar to the previous study mentioned, since the comparisons tested are not adaptively chosen based on previously completed comparison results, the scheme does not lead to the optimal sequence of comparisons that minimizes the total number of comparisons conducted.

In [16], a new QoE model and evaluation method for broadcast audio contribution over IP is proposed. This application has some similarities to the design of POS scheme in VoIP systems. However, the study does not consider the effects of jitter in the decision of the mouth-to-ear delay in the context of losses perceived at the decoder. Furthermore as the study relies on the E-model for measuring the effect of delay on the conversational quality, it also inherits the shortcomings of the E-model discussed extensively in Chapter 3.1.

3.2 Our Approach for Evaluating Conversational Quality of VoIP Systems

Our survey shows that new methods developed for comparing conversational quality in VoIP systems need to be aided by subjective tests because there is no suitable model of interactive VoIP conversations.

In our research we have identified two different levels of evaluation. The first is the evaluation of existing and newly developed VoIP systems as a whole, and the second is the evaluation of control algorithms employed in VoIP systems. The latter include the dynamic POS and LC algorithms in VoIP systems. This level of evaluation considers the possibly infinite number of alternatives in the operation of such control algorithms and is crucial in the design of VoIP systems with high conversational quality. Each level of evaluation requires some similar and some different tasks to be completed. In this section we present our approach for system-level evaluation. In Section 3.4, we present our approach for component-level evaluation, after surveying previous work on the design of VoIP components.

Our approach involves the following tasks:

- We develop a testbed for conducting subjective tests. This entails the collection of Internet packet traces and interactive conversations and the design of a system to replay these traces and conversations. The system allows subjective tests to be conducted under different VoIP systems as well as comparisons under identical network and conversational conditions.
- To address the issue that there are infinitely many possible network and conversational conditions, we develop a classifier [35] that learns from training examples generated under limited conditions and that generalizes to unseen conditions. Based on the property that humans cannot distinguish differences in conversational quality under slightly different operating conditions, we develop statistical approaches that can significantly prune the number of subjective tests required.

Testbed for evaluating VoIP systems. Our approach is to develop a testbed for evaluating [53, 48] VoIP systems. The testbed allows subjective tests to be repeated for different VoIP systems under identical network and conversational conditions. The testbed ensures that variations in quality are only due to differences in the algorithms or systems compared. It consists of multiple computers, each running the VoIP client software, and a Linux router for emulating the real-time network traffic [53, 48]. We have modified the kernel of the router in order to intercept all UDP packets carrying encoded speech packets between any two clients. The router runs a troll program that drops or delays intercepted packets in each direction according to packet traces collected in the PlanetLab.

Our approach is to develop a human-response-simulator (HRS) that runs on each end-client. The HRSs simulate a conversation with pre-recorded speech segments by taking turns speaking their respective segments. We use a software interface to digitally transfer the waveforms to and from the VoIP clients without quality loss. The setup can be thought of as a pair of voice response systems conversing with each other by following a script in such a way that the conversation can be repeated almost exactly for the same VoIP system under the same conditions. This enables the opinion ratings to be directly related to the conversational quality of the systems tested under the same conditions.

Subjective evaluations of four VoIP systems. Our methodology involves comparing four VoIP clients: Skype (3.6), Google-Talk (beta), Windows Live Messenger (8.1), and Yahoo Messenger (8.1) [55]. Using conversations recorded by our testbed under some network and conversational conditions, human subjects will be asked to comparatively evaluate two conversations by the CCR scale. The recordings will be presented in a random order to the human subjects who do not know the system or the network conditions used for each recording. The tests are conducted using several Internet traces that represent different network conditions and an additional trace representing an ideal condition with no loss and delay. We use three distinct conversations of different single-talk durations, HRD, and switching frequencies in Chapter 4, while evaluating the four systems.

Classifiers for generalizing subjective evaluation results. Based on pairwise comparisons of the conversations recorded on the four VoIP systems mentioned above, our approach is to generate training patterns, each consisting of a number of objective measures characterizing the two systems, network and conversational conditions and a subjective preference measure. We learn these mappings by a classifier implemented as a support vector machine (SVM) [8] with a radial basis kernel function in Chapter 4.2.

To simplify learning, our approach is to map the average of the user CCR opinions of A against

B into three classes: *A better than B*, *B better than A*, and *A about the same as B*. To verify that the classifier can generalize to unseen network and conversational conditions, our approach is to use cross-validation techniques commonly employed in statistics [33]. Since the classifier does not learn from the samples in the testing set, a high cross-validation score is interpreted as the ability of the classifier to generalize to samples with conditions not in the training set. The application of our methodology is presented in Chapter 4.2.

3.3 Previous Work on the Design of VoIP Systems

Most VoIP systems employ LC and POS algorithms to mitigate delays, losses, and jitter at the packet level. In this section, we identify the limitations of existing LC and POS algorithms. As is outlined in Chapter 1, these algorithms must be designed in conjunction with the speech codec and with an understanding of conversational quality. However, most studies approach the design of LC and POS individually, while some consider delay-aware LC design or redundancy-aware POS design.

The detailed analysis of network conditions is presented in Chapter 2.2. In this section, we utilize these observations on the network conditions along with other observations to provide a basis for our discussion of previous work on LC and POS schemes employed by VoIP systems.

Internet Loss and Delay Conditions. Studies on Internet traces show that connections exhibit consecutive packet losses (bursts) rather than random losses, meaning that if the n^{th} packet is lost, the likelihood of $n+1^{th}$ packet being lost is higher than the average loss rate [31]. This dependency is more pronounced in applications that transmit frequent packets, since when a packet is lost due to overflow of a router buffer, there is usually not enough time till the next packet for the buffer occupancy to decrease. However, the number of consecutive losses are usually small [68].

Traces collected on the Internet exhibit non-stationary loss behavior [68, 53, 48] in time-scales that are relevant to the VoIP applications. Packet-loss conditions may change in a matter of seconds, and stationary models [4, 31] are not capable of tracking fast-changing conditions or controlling transmission parameters in real-time [53].

Internet traces also exhibit sudden and dramatic changes in packet delays, called delay spikes [3, 31, 53], which are caused by a sudden decrease in the buffers in one of the routers on the path of the packets. After the spike, multiple consecutive packets may be received almost instantaneously when the congested router empties its buffers quickly. This causes those consecutive packets after the spike to experience less and less delay until the delay value reaches the level before the spike. As presented in Chapter 2 and in our previous studies [51, 53], we observe that within a second,

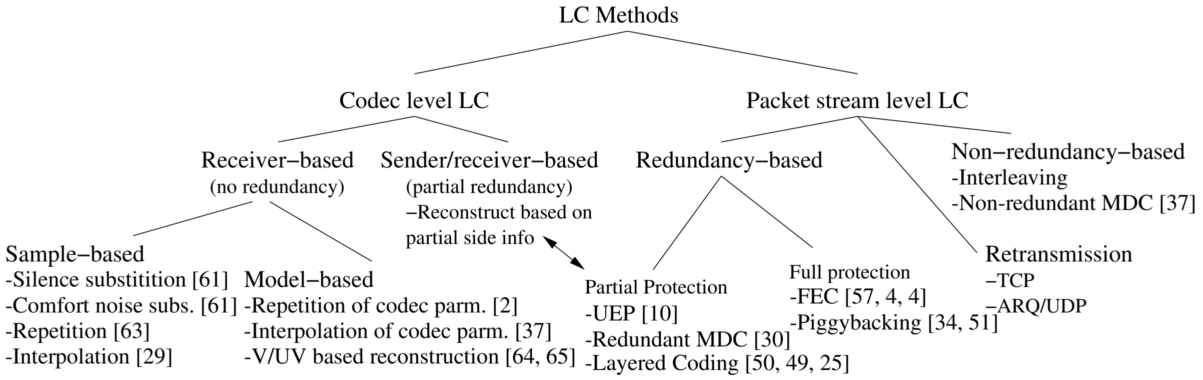


Figure 3.2: A classification of existing LC methods at the codec and the packet-stream layers.

several spikes can occur, either in an individual or in a coupled fashion. The behavior of delay spikes and their effects on those inaudible segments of speech in real-time VoIP transmissions can be evaluated by the magnitude and the frequency of the spikes.

Loss concealment. Figure 3.2 summarizes some existing LC techniques at the packet-stream layer. These techniques aim to either directly reduce the amount of unconcealable frames experienced by the decoder or provide partial redundancy for helping the decoder reduce perceptual degradations due to losses.

Retransmission of speech frames after the detection of a network loss is infeasible in real-time VoIP, due to the excessive delays involved and their effects on MED.

Non-redundant LC schemes are generally based on the interleaving of frames during packetization [43]. One way is to exploit the fact that shorter distortions are less likely to be perceived, and to break an otherwise long segment into several shorter segments that are close by, but not consecutive. This is not strictly an LC technique because it does not actually recover losses. Another way is MDC [14, 15, 38] that generates two or more descriptions with correlated information from the original speech data. This may be hard in low bit-rate streams whose correlated information has been largely removed during coding [38].

By sending these descriptions in different packets, the information in the descriptions received can be used to reconstruct those lost descriptions. However, because of less efficient coding of each description, the quality of the decoded stream is usually lower than that without MDC when all the descriptions are received. Another disadvantage is that the receiver will incur a longer MED when waiting for all the descriptions to arrive.

Since non-redundant schemes do not provide an adequate level of loss concealment while maintaining a high intrinsic quality, in our study, we direct our focus to redundancy-based schemes.

Redundant LC schemes exploit trade-offs among the level of redundancies, the delay required to recover losses from the redundant information, and the quality of the reconstructed speech in response to network loss, jitter, and delay behavior. They work in the Internet because increases in packet size, as long as they are less than the MTU [19], do not lead to noticeable increases in loss rate [51]. They consist of schemes that use partial and full redundancies. Examples that employ partial redundancies include layered coding [50, 49, 25], UEP (unequal error protection) [10], and redundant MDC [30]. The partial redundancies contain information that works in conjunction with the speech codec in reconstructing speech frames in case of loss. The recently designed ITU G.729.1 [25] uses enhancement layers to help in reconstructing lost frames. Examples that employ full redundancies include FEC (forward error correction) [57, 4, 4] and redundant piggybacking [34, 51]. An FEC-based LC scheme [6] has been proposed for VoIP that incorporates into its optimization metric the additional delay incurred due to redundancy. FEC can be implemented using block-based Reed-Solomon coding or parity coding, where $n + 1$ packets are sent to represent n original packets, and full recovery can be achieved if at most one of the $n + 1$ packets is lost. The value of n determines the level of concealment available against frequent losses. In our previous work, we have used piggybacking as a simple yet effective technique for sending copies of previously sent frames together with new frames in the same packet, without increasing the packet rate [51, 53, 48]. By including information on one or more previously sent packets, it provides robustness against single or multiple consecutive losses.

The main difficulty of using redundant LC schemes is that it is hard to know a suitable redundancy level. Its dynamic adaption to network conditions may either be too slow, as in Skype [51], or too conservative [53]. Another consideration is that the redundancy level is application-dependent. Lastly, it is important to design LC schemes by utilizing the information about the robustness of the speech codec used in the VoIP system in order to achieve the desired level of protection with the minimum delay and within the bit budget.

Play-out scheduling. Figure 3.3 summarizes the various POS methods. Simple schemes with fixed MEDs either hard-coded at design time or decided during the establishment of a call do not provide consistent protection against late losses because delays and losses are non-stationary and path-dependent. Adaptive POS schemes that adjust the playout schedule at the talk-spurt or the packet level are more prevalent.

At the talk-spurt level, silence segments can be added or omitted at the beginning of a talk spurt in order to make the changes virtually imperceptible to the listener. Adjustments can also be made for each frame using *time-scale modification* (TSM) [36]. The scheme stretches or compresses

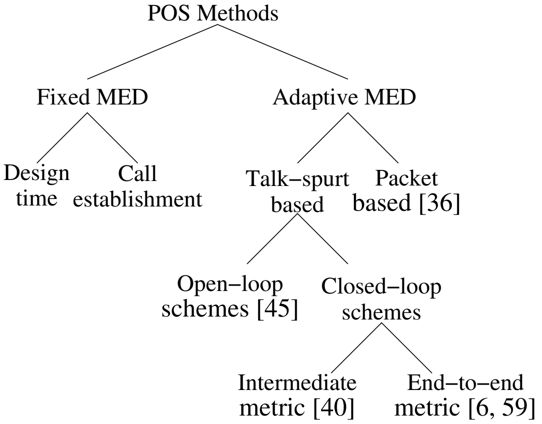


Figure 3.3: A classification of existing fixed and adaptive POS methods.

speech frames without changing its pitch period. However, it requires additional computational resources, has small effects on MEDs, and is generally perceptible to listeners.

At the packet level, there have been several studies that aim to balance between the number of packets late for playout and the jitter-buffer delay that packets wait before their scheduled playout times.

Open-loop schemes use heuristics for picking some system-controllable metrics (such as MED), based on network statistics available [45]. For example, Algorithms 1-3 [45] calculate running estimates of the mean (d) and the variations (v) in network delays and choose a playout delay $p = d + 4v$ at the beginning of each talk-spurt. They aim to set the playout delay far enough from the mean delay in order to conceal most of the late frames. Algorithm 4 [45] improves the estimations by tracking delay spikes in order to avoid long sequences of unconcealed lost frames. They are less robust because they do not explicitly optimize a target objective. Moreover, they do not consider the effects of the codec on speech quality, although their performance depends on the codec used.

Closed-loop schemes with intermediate quality metrics [40] control an intermediate metric based on the late-loss rate collected in a window. Their difficulty lies in choosing a good intermediate metric. The metric must be easy to compute at run time and be tied to a target objective.

Closed-loop schemes with end-to-end quality metrics generally use the E-model [22] for estimating conversational quality as a function of some objective metrics. One study uses this estimate in a closed-loop framework to jointly optimize the POS and FEC-based LC [6]. It makes some limiting assumptions, such as a quasi-static Gilbert packet-loss process, a stationary with an observable delay distribution, and mutually independent delays and losses. Another study [59] proposes to use the E-model but separately trains a regression model for modeling the effects of the loss rate and the codec on PESQ. In that study, a POS algorithm, called *p-optimum*, was proposed to adapt the playout delay on each talk spurt in order to optimize LOSQ. For simplicity, the model was trained by a Bernoulli loss model neither employs, nor is designed to work in conjunction with, a redundancy-based LC scheme. Because unconcealed lost frames can be bursty, such a model under estimates the degradations due to unconcealed frames. These models are limited because, without a redundancy-based LC scheme, lost frames cannot be recovered by adjusting the playout delays alone.

Most of the existing POS algorithms do not consider the redundancy-based LC, or base their decisions in a redundancy-aware way. Moreover, since none of their objectives for measuring quality captures the effects of user-perceptible attributes, using them as the objective does not lead to the best perceptual conversational quality.

3.4 Our Approach for Evaluating Conversational Quality in the Design of Control Components

In this section we present our approach for evaluating conversational quality in the design of control components of a VoIP system. Our main focus is on the play-out scheduling scheme, and the design and choice of other components are secondary. The reasons for this choice are presented at the beginning of the section.

Secondly, we describe our multi-step approach that eventually leads to the design of a POS scheme that achieves high and consistent perceptual quality. These steps correspond to P2, P3 and P4 in the problems studied in the thesis (Table 1.1).

3.4.1 Importance of play-out scheduling control in achieving high conversational quality

Play-out scheduling control is a crucial component in the design and operation of a VoIP system. Here we discuss the reasons why POS is important in the operation of real-time VoIP and why it would benefit from the subjective learning methods developed in this thesis.

- **Adaptation due to time-varying network conditions:** Even though POS, LC and bit-rate adaptation of variable bit rate (VBR) speech codecs may benefit from adapting its operating parameters dynamically in response to changing network conditions, only POS requires this change to be much faster than the other adaptations. For example POS changes in response to dynamic network jitter conditions may occur in a matter of tens of milliseconds whereas changes in loss conditions may require the receiving side to coordinate with the transmitting side for adaptation. Similarly, the coordination of the two sides is needed when making changes to the speech codec used, which reduces the ability to adapt.

- **Affecting multiple counteracting and user perceptible quality aspects:** POS affects both delay and LOSQ of a conversation in a counteracting way. Thus, the effects of changes in POS parameters on subjective preferences are not known and need to be learned. On the other hand, changes in LC and VBR codec parameters only affect LOSQ that is user perceptible, and this effect can be learned easily with off-line tests. Furthermore, the relationship between the changes in LOSQ (while keeping all other quality aspects the same) and subjective preference is monotonic and can also be learned with relative ease with off-line tests.

- **Run-time evaluation of component-level performance:** POS's ability to conceal late and lost packets can be evaluated at run-time without any intrusive tests affecting the operation of the system. Furthermore, this run-time evaluation can be used to adjust the control parameters. LC performance can also be evaluated indirectly as well. However, the evaluation of the performance of VBR speech codec (its robustness to packet losses) is not feasible, as the original waveform is not available at the decoder, where the degraded waveform is generated. Even if the effective packet losses were relayed to the encoder, it is computationally expensive to conduct an analysis-by-synthesis to evaluate a speech codec's performance.

- **Number of possible operating points:** Another reason why POS control design is a more challenging problem is that it takes continuous control values. This makes the off-line learning and on-line decision making processes much more difficult than in control systems with a few or finite number of choices. In this framework, LC control offers only a few choices of types of redundancies and, in the case of a redundant piggybacking algorithm [53], only four control alternatives, namely sending no, 1, 2 or 3 redundant copies of previously transmitted frames. Similarly, a VBR speech codec has a finite number of alternatives, which dramatically reduces the complexity of the offline learning efforts.

3.4.2 Multi-dimensional representation of quality

To achieve high conversational quality, it is necessary to select suitable operating points for the POS and LC algorithms when given the network and conversational conditions at run time. Assuming that the playout schedule is adjusted on a talk-spurt basis, we use the statistics at run time in order

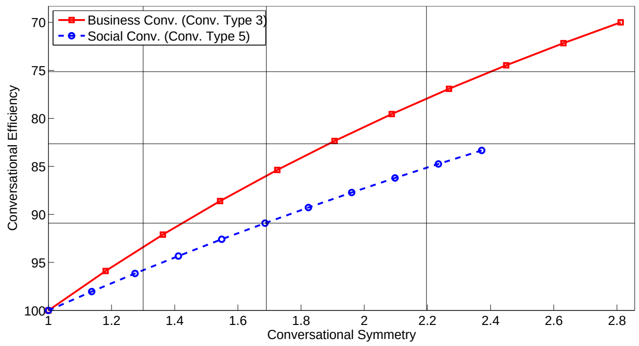


Figure 3.4: Effect on CS and CE when MED changes for the two conversations in Table 2.2.

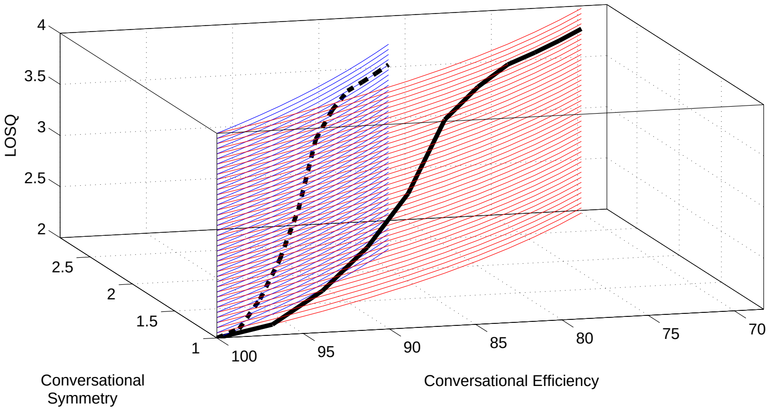


Figure 3.5: The planes representing the conditions of the conversational type. The curve on each plane represents the conditions imposed by the network.

to evaluate the objective metrics of alternative operating points of these control algorithms. To have efficient use of resources, we have designed the POS algorithms to depend on the current LC decision made and the robustness of the speech codec used.

Figure 3.4 depicts the trade-off between CE and CS as a function of MED for the conversations in Table 2.2. We see that the degradations in CS and CE are less pronounced for the social conversation with a lower switching frequency.

Figure 3.5 illustrates the trade-offs among CS, CE, LOSQ, and the system-controllable MED in conversational quality. The red and blue planes parallel to the LOSQ (PESQ) axis represent the conditions imposed by the conversational type (type3 - business versus type 5 - social in Table 2.2). For given network and conversational conditions, the black curve on each plane represents the trade-offs among CS, CE, and LOSQ, when parameterized by the system-controllable MED and subject to the constraints imposed by the network and the conversation.

3.4.3 Subjective evaluations to guide POS design

As we initially mentioned in Chapter 1, subjective evaluations can be conducted to evaluate the quality of a control scheme. Because such evaluations cannot be performed at run time, offline tests have to be conducted during which the information learned is used to guide the operation of the control scheme(s) at run time. In general, subjective evaluations are time-consuming and expensive and will require multiple subjects in order to arrive at some statistically significant results. Further, since there may be prohibitively many network conditions and communication scenarios that can be observed at run time, it is infeasible to conduct exhaustive subjective tests in order to cover all possibilities.

As discussed in Section 1.4, a standard method for conducting subjective evaluations is to ask subjects to rank the quality by an ACR and to take an algebraic mean of the opinions of the subjects in response to the same stimuli. The result obtained is the *mean opinion score* (MOS) [21].

Even though MOS may be useful for verifying a system's performance, there are several reasons why it is not suitable for designing new control schemes. Firstly, absolute scores obtained for two points on an operating curve can be used to deduce their relative positions. If all alternatives are mutually related under pairwise comparisons, then a total ordering can be established under ACR. However, in practice, two operating points may not be comparable when they involve multiple quality metrics. In this case, the perceived effects on the difference of one metric may not be consistently translated into the differences of the other metrics. Consequently, the feasible operating points of an operating curve lie on a Pareto-optimal boundary.

Secondly, although each MOS score can be determined with some statistical confidence, no statistical significance can be associated with the difference of two MOS scores. For instance, if the variances in the scores are large relative to their difference, then the conclusion reached on the difference is not statistically meaningful. As is stated in ITU P.800 [21] for evaluating telephone communication quality, absolute ratings are not accurate for evaluating quality when samples have high quality or their difference is barely perceptible. Figure 3.6 illustrates the case where two comparisons with the same MOS difference (Δ) lead to different pairwise comparison results. Hence, the number of samples required to obtain MOS with a certain level of statistical significance can be inadequate for some pairwise comparisons but excessive for other cases.

To address the issues described above, our approach is to develop a statistical scheduling of offline comparative subjective tests for evaluating alternative operating points on an operating curve of a real-time multimedia system. Without loss of generality, we only consider an operating curve due to a single control scheme, although the approach can be easily extended to multiple control schemes. Our goal is to minimize the number of subjective tests needed in order to determine a lo-

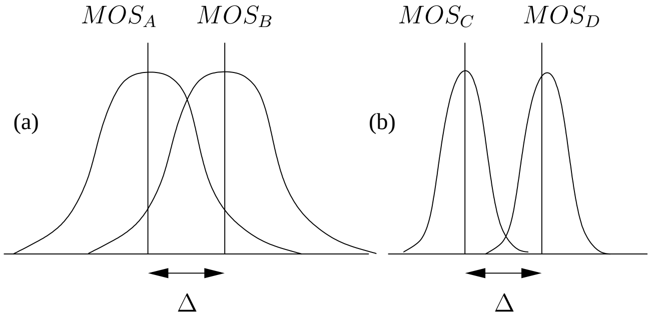


Figure 3.6: Two scenarios where user opinions take continuous values. The difference in MOS values for (A,B) and (C,D) are equal to Δ, where A and B are not significantly different, but C and D are significantly different.

cally optimal operating point to within some prescribed level of statistical confidence. A secondary goal is to efficiently schedule the subjective tests of multiple operating curves in a multimedia application. Our approach consists of the following steps:

a) *Comparative ranking.* To determine the preferred operating point among a set of alternatives, a partial order that requires pairwise comparisons suffices. The partial order can be assessed by a measure that evaluates the relative quality of two alternatives in a *comparison category rating* (CCR) (similar to that described in Annex E of ITU P.800 [21]). By presenting two alternatives to each subject, one after another, the approach allows the incomparability of some alternatives to be identified and small differences between two to be more accurately evaluated. The disadvantage, however, is a significant increase in the number of tests because such tests will need to be conducted for each pair of alternatives instead of each alternative.

b) *Stochastic evaluations under given conditions.* To identify the best operating point at run time, we first consider the problem of determining the best operating point offline under a given set of network conditions and communication scenarios. Our approach is to develop a versatile VoIP system simulator to generate conversations on which to conduct a limited number of subjective evaluations. The property of the simulator is such that it can repeat the network and communication conditions exactly in order to eliminate variations other than the differences in the control schemes tested.

Our approach is to then collect the comparative subjective opinions and represent them as discrete distributions. Based on hypothesis testing, we deduce the dominant opinion (opinion that is likely true more than 50% of the time), if it exists, with a specified level of statistical significance. By eliminating other sources of variations, the approach allows us to significantly reduce the number of tests conducted per alternative pair while maintaining the significance level.

c) *Logistic limitations of conducting subjective tests to multiple subjects.* The idea is to system-

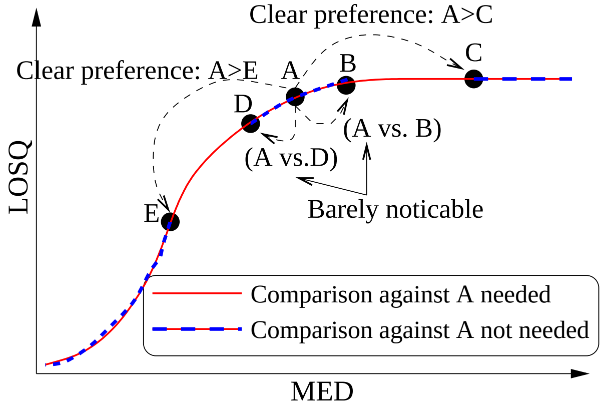


Figure 3.7: Illustration of how JND may help reduce the number of subjective evaluations.

atically use the observations from past subjective tests to prune tests that have not been conducted. Thus, this requires subjects to be coordinated in their evaluations in a locked-step fashion. However, if all subjects cannot conduct the evaluation simultaneously, or in a pre-determined finite amount of time, a statistically confident deduction cannot be established; thus, pruning comparisons at test-time would not be feasible. Hence, for practical reasons, it is hard to conduct simultaneous subjective tests and prune those upcoming comparisons during test-time. We present the application of our approach in detail in Chapter 6.1.1.

On the other hand, presenting a pre-defined set of evaluations to each subject for a single-session evaluation would result in a significant number of unnecessary evaluations. For these reasons, we adopt a batch-by-batch approach, where a subset of alternatives are compared by all subjects within a few days. After that we process the pair-wise relations for determining statistical significance and prune some of the alternatives scheduled for later batches (sessions). Since a comprehensive set of network and conversational conditions are tested, the subjects can be presented with a diverse set of conversational samples in each session, which is needed for avoiding anticipation and bias.

d) *Pruning of search space.* As mentioned above, our idea is to systematically use the observations from past subjective tests to prune tests that have not been conducted. Our approach is based on a statistical model of subjective evaluations that utilizes the following two principles: (a) the subjective quality induced by small changes in the control scheme cannot be perceived by subjects, and (b) subjective preferences between points that are in a contiguous subset of the operating curve generally point toward the locally optimal point in that subset. Figure 3.7 simply illustrates the pruning of the search space. We present the application of our approach in Chapter 6.1.4.

e) *Learning the mapping between parameters characterizing operating curves and their optimal point.* Based on the subjective preferences under a comprehensive set of test conditions, we learn and cross-validate an SVM (support-vector-machine) classifier. The classifier learns the mapping between the parameters characterizing the operating curves and the target MED. To evaluate the MEDs predicted by the classifier, we develop a statistical method for estimating the accuracy predicted using limited subjective results. We present the application of our approach in Chapter 7.

f) *Run-time evaluation of POS.* Based on conditions measured periodically at run time and unseen in learning, we present in Chapter 8 the application of our classifiers for identifying the best MED. We also present experimental results to demonstrate the high subjective quality of the resulting conversations.

3.5 Summary

In this chapter we have presented the relevant previous work and our approach for the evaluation of conversational quality on a system level. We have also presented the control-component-level evaluation of conversational quality for the design of VoIP systems with high perceptual quality.

The analysis of the previous work and its shortcomings have guided our focus areas, where new understanding of the material and further study are needed. Even though the previous work did not distinguish the evaluation of quality between system level evaluation and control scheme level design, we have identified the differences and have presented the approach appropriately. The approach presented in this chapter defines the tasks conducted in our study for the evaluation and design of VoIP systems with high perceptual quality

CHAPTER 4

EVALUATING THE CONVERSATIONAL SPEECH QUALITY OF VOIP SYSTEMS

In this chapter we present a general method for comparing the conversational quality of speech communication systems over networks with delays and its application in studying four VoIP systems. When applied on four commonly used VoIP systems, our results show that each achieves some trade-offs, but none attains the best quality under all conditions. Lastly, we discuss a systematic approach for predicting user preference between two VoIP systems in terms of objective measures that can be easily acquired.

The evaluation of speech communication systems has been an important field for both academia and industry for decades. With the introduction of VoIP systems that use new speech codecs and schemes for handling network imperfections, there is a need to develop new methods for evaluating these systems. There has been only a small number of comprehensive evaluations of commonly used VoIP systems.

The perceptual evaluation of an interactive conversation depends on the quality and the latency of the one-way speech segments received. Due to path-dependent and non-stationary conditions in the Internet, the delays incurred when a conversation switches from one client to another are asymmetric and may lead to a degraded perceptual quality as compared to that of a face-to-face setting. The conversational dynamics can further be disrupted by variations in the one-way speech quality and latency. As those factors that affect conversational quality may counteract to each other, trade-offs must be made among them. The most straightforward approach for the evaluation of conversational quality is through subjective tests. However, currently available objective and subjective standards do not adequately capture these trade-offs, and there are no good methods for evaluating the conversational quality of VoIP systems. Furthermore, subjective tests cannot be used in large-scale experiments due to their large overhead, the high costs of listening experts, and their unrepeatable nature.

In this chapter, we present a method for studying these trade-offs under repeatable network and conversational conditions. In this method, we evaluate VoIP systems by objective measures that are augmented by subjective ones. We also examine several ITU (International Telecommunication Union) recommendations on the objective evaluations of a system.

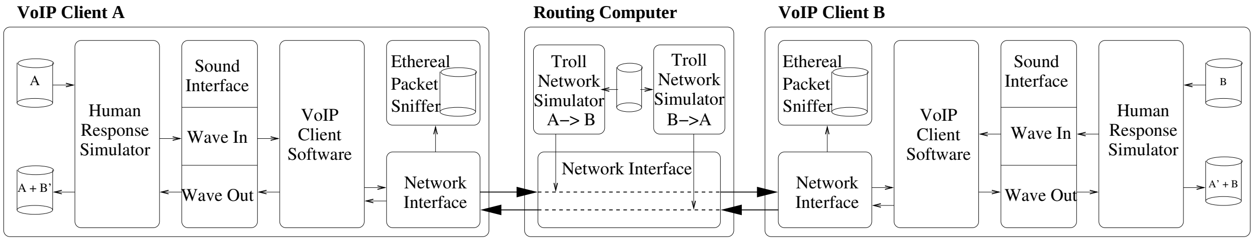


Figure 4.1: Our testbed for emulating a two-way interactive speech communication.

The evaluation of some commercial VoIP systems is also hampered by their proprietary nature. Most use codecs and algorithms that are not freely available for testing. As a result, it is impossible to obtain some of the critical parameters, such as the amount of packets unavailable at the decoder due to network losses or delays. To this end, the evaluation of current systems must be done by treating them as black boxes whose input and output waveforms are the only information available. Both subjective and objective metrics are important in the evaluations because each alone is inadequate. Objective metrics, including the *listening-only speech quality* (LOSQ) and MED, can be calculated based on the information collected on both sides.

Our goals in this study are threefold. Firstly, we present an evaluation method that captures the trade-offs in user perception of conversational quality in a repeatable way under given network and conversational conditions. Secondly, using our testbed, we present our results on the comprehensive evaluations of four VoIP clients: Skype (3.6), Google-Talk (beta), Windows Live Messenger (8.1), and Yahoo Messenger (8.1). Lastly, we discuss a comparative evaluation method for generalizing the subjective evaluation results using objective metrics that can be measured by our testbed. We show the performance of our predictor and the generalizability of our method using standard cross-validation techniques.

4.1 A New Method for Evaluating Conversational Quality

In this section, we first describe our testbed for recording conversations that are repeatable under identical network and conversational conditions. We then present a comparative subjective evaluation method for measuring the relative conversational quality between two VoIP systems.

Testbed. Figure 4.1 depicts our testbed for emulating a two-way interactive speech communication, where there are two computers running the VoIP client software and a router for emulating the network condition.

The testbed aims to address the repeatability of standard subjective tests described in ITU P.800. To address the repeatability of network conditions, we have modified the Linux kernel of the router in order to intercept UDP packets carrying encoded speech packets between two VoIP clients. In conjunction with the kernel, we use a troll program to drop or delay intercepted packets in each direction according to the traces collected on the PlanetLab (Section 2.2). To address the repeatability of conversational conditions, we have developed a human-response-simulator (HRS) software that runs on the two end-client computers. HRS is capable of simulating any conversation with prerecorded speech segments, each using the segments that belong to one side of the conversation. One of the HRSs is configured to start the conversation; then both take turns speaking their respective segments. Each HRS listens to the speech played by the VoIP client and responds after waiting appropriately. The HRS interfaces with the VoIP client software via the Virtual Audio Cable [41] that allows the transfer of waveforms to and from the VoIP client digitally without quality loss. Each HRS records the waveforms spoken and heard by the respective client it is interfaced to.

The setup can be thought of as a pair of speech response systems that converse with each other by following a script in such a way that the conversation can be repeated almost exactly for the same VoIP system under the same condition. Since the speech coding, LC, and POS algorithms differ for the VoIP systems tested, the speech quality and the latency observed vary from one system to another under the same network and conversational conditions. When comparing two VoIP systems under the same condition, it ensures that variations in quality are only due to differences in the systems. Hence, it enables the relative quality of the two systems to be compared.

After the recordings have been collected, human subjects are asked to comparatively evaluate two conversations recorded under the same condition using the Comparative Category Rating (CCR) scale in ITU P.800. We also extract objective measures, such as PESQ, MED, CS, and CE, from each recording. Finally, we present the recordings on two systems instrumented under the same conditions in a random order to the human subjects, who do not know the system or the network conditions used in each recording.

Evaluation of four VoIP systems. Due to the proprietary nature of commercial VoIP systems, it is hard to understand the internal structures of these software clients, which makes it difficult to find shortcomings and improvement opportunities. We apply our tested and comparative subjective-evaluation method on four VoIP systems.

Since there are prohibitively many network and communication scenarios at run time, it is infeasible to conduct exhaustive subjective tests to cover all possibilities. To this end, we have identified a representative set of network conditions and two-way conversational recordings that span a wide range of conditions to be evaluated.

In this chapter we utilize the 6 connections indicated by '*' in Table 2.1 in our evaluations of the 4 systems. Similarly, we utilize 3 of the conversations listed in Table 2.2 in our subjective tests. These cover a wide range of conversational conditions observed in phone conversations. Together with the three conversations there are altogether 21 distinct combinations to be tested for each system.

In our comparative subjective tests, we compare each pair of systems (six comparisons for four systems) under each combination of network and conversational conditions, which result in 126 comparisons. With six human subjects, we carry out a total of 756 subjective tests. We observe that the opinions from different users under the same pair of systems and test conditions lie within 3 CCR values of each other in 83% of the cases. These indicate a high confidence in the results.

Performance analysis: objective metrics. Table 4.1 summarizes the objective measures collected for the four systems evaluated under seven network conditions and three conversations. We observe that their performance is comparable under the ideal network condition, but that their performance starts to deviate as more delay, jitter, and loss is introduced. These differences indicate that different trade-offs are made by each system to overcome network impairments. We further observe that Windows Live is superior in terms of listening-only speech quality (measured by PESQ), that Skype has consistently larger MEDs, and that Google Talk generally has shorter MEDs and better CS and CE.

The PESQ quality of the systems under the ideal network condition is a strong indication of the intrinsic performance of their respective speech codecs without LC. When the network has delays and jitters, the PESQ observed reflects the combined performance of the POS scheme for concealing late packets and the robustness of the speech codec to unconcealed late packets. Similarly, under lossy conditions, the combined performance of LC and the robustness of the codec under unconcealed network losses is reflected.

We observe that the systems do not have widely different MEDs for the three conversations of different turn-taking frequencies and the same network conditions. This result indicates that the POS schemes of these systems do not optimize their MEDs in response to different turn-taking conditions.

Performance analysis: comparative subjective evaluations. We have carried out extensive comparative subjective evaluations of the six system pairs under the same network and conversational conditions. (The complete results are not shown here.)

Table 4.1: Objective evaluations of four VoIP systems tested under six Internet and one ideal connections. The best quality for each of the four systems is indicated with '*'.

Trace Class	VoIP	Conv. 3				Conv. 4				Conv. 5			
(Delay,Jitter,Loss)	System	PESQ	MED	CS	CE	PESQ	MED	CS	CE	PESQ	MED	CS	CE
(No,No,No)	Skype	3.192	286	2.04	67	3.244	338	1.95	74	3.418	290	1.70	83
	GTalk	3.557	130*	1.47*	71*	3.506	147	1.42	78*	3.536	160	1.39	85*
	Yahoo	3.553	140	1.51	71*	3.676	139*	1.39*	78*	3.785	151	1.37	85*
	WinLive	3.562*	171	1.62	70	3.856*	154	1.43	78*	3.928*	133*	1.32*	85*
(L,L,L)	Skype	3.328	319	2.15	66	3.119	541	2.52	71	3.254	392	1.95	82
	GTalk	3.371	203*	1.74*	69*	3.525*	368	2.04	74	3.092	201*	1.49*	84*
	Yahoo	3.534	205	1.74*	69*	3.492	203*	1.57*	77*	3.354	298	1.72	83
	WinLive	3.675*	222	1.81	69*	3.492	218	1.61	77*	3.746*	393	1.95	82
(L,L,H)	Skype	2.339	442	2.60	63	2.461	416	2.17	73	2.565	424	2.02	81
	GTalk	2.484	230	1.83	69*	2.501	265*	1.75*	76*	2.305	275	1.67	83
	Yahoo	2.502	217*	1.79*	69*	2.755	276	1.78	76*	2.485	239*	1.58*	84*
	WinLive	3.306*	336	2.22	66	3.309*	340	1.96	74	3.257*	321	1.78	83
(L,H,L)	Skype	2.693	408	2.48	64	2.882	487	2.37	72	3.083*	420	2.02	82
	GTalk	3.145	216*	1.78*	69*	3.145	227*	1.64*	77*	2.854	261*	1.63*	83*
	Yahoo	3.085	274	1.99	67	3.097	240	1.68	76	2.987	274	1.66	83*
	WinLive	3.454*	404	2.47	64	3.512*	432	2.22	73	2.953	420	2.02	82
(H,L,L)	Skype	3.096	550	2.99	61	3.325	462	2.30	72	3.444	420	2.02	82
	GTalk	3.466	281*	2.02*	67*	3.517	279*	1.79*	76*	3.435	287*	1.69*	83*
	Yahoo	3.531	283	2.03	67*	3.464	305	1.86	75	3.687*	301	1.73	83*
	WinLive	3.792*	313	2.13	66	3.803*	315	1.89	75	3.647	309	1.75	83*
(H,L,H)	Skype	2.619	535	2.94	61	2.564	504	2.42	72	2.564	503	2.22	81
	GTalk	2.639	273*	1.99*	67*	2.666	283*	1.80*	75*	2.469	300*	1.73*	83*
	Yahoo	2.749	281	2.02	67*	2.472	365	2.03	74	2.617	314	1.76	83*
	WinLive	3.060*	440	2.60	63	3.251*	421	2.19	73	3.286*	363	1.88	82
(H,H,L)	Skype	2.985	612	3.22	59	2.983	574	2.62	70	2.652	648	2.57	79
	GTalk	3.296	399*	2.45*	64*	3.151*	410*	2.15*	73*	2.729	397*	1.96*	82*
	Yahoo	3.022	544	2.97	61	3.068	487	2.37	72	2.841	573	2.39	80
	WinLive	3.327*	595	3.15	60	2.937	589	2.66	70	2.930*	748	2.81	78

Table 4.2 defines the *comparison MOS* between two conversations (B compared to A) in our listening tests (similar to ITU P.800 Annex E, Comparison Category Rating method):

$$CMOS(A \rightarrow B) \in \{-3, -2, -1, 0, 1, 2, 3\}. \tag{4.1}$$

Figure 4.2 illustrates the distribution of the opinions for each of the six system pairs. Figures 4.3 and 4.4 further present the opinions under different network and conversational conditions respectively for each of the six system pairs seperately. To get a reasonable number of samples in each distribution, we have combined the results for some of the network conditions that are relatively similar in terms of their effects on performance: (N,N,N), (L,L,L), (H,L,L) for good conditions; (L,H,L), (H,H,L) for jittery conditions; and (L,L,H), (H,L,H) for lossy conditions.

Table 4.2: The seven possible opinions in a subjective test comparing System A and System B under the same conditions.

User Response	CMOS Score
A is strongly preferred over B	-3
A is preferred over B	-2
A is slightly preferred over B	-1
A and B are preferred equally	0
B is slightly preferred over A	1
B is preferred over A	2
B is strongly preferred over A	3

Figure 4.3a shows that Skype performs similarly to Google-Talk under lossy conditions; however, Skype is preferred under jittery conditions, whereas Google-Talk is preferred under good network conditions. Figure 4.3b shows that Skype performs similarly to Yahoo Messenger under good and lossy conditions; however, Skype is slightly preferred under jittery conditions. Figure 4.3c shows that Windows-Live is preferred to Skype under all network conditions. Figure 4.3d shows that Google-Talk performs similarly to Yahoo Messenger under good and lossy conditions; however, Google-Talk is slightly preferred under jittery conditions. Figure 4.3e and f show that Windows-Live is preferred to Google-Talk and Yahoo Messenger under all conditions and the preference is even stronger for lossy conditions.

These results suggest that different systems employ different trade-offs in addressing losses and jitters. They also suggest that the Windows-Live VoIP system employs either a better loss concealment scheme or a more robust speech codec, or both, in comparison to the other systems. Furthermore, the observation that Windows-Live is preferred against other systems even under good network conditions suggests that the intrinsic quality of the speech codec it employs is superior to the codecs other systems use.

Figure 4.4a shows that Skype is slightly preferred to Google-Talk under all conversational conditions. Figure 4.4b shows that Skype performs similarly to Yahoo Messenger under all conversational conditions. Figure 4.4c shows that Windows-Live is preferred to Skype under all conversational conditions. Figure 4.4d shows that Google-Talk is slightly preferred to Yahoo Messenger under all conversational conditions. Figure 4.4e and f show that Windows-Live is preferred to Google-Talk and Yahoo Messenger under all conversational conditions.

These results suggest that there is not much, if any, effect of the conversational conditions in the operation of the four systems; thus, each system pair preference is about the same for all the three conversational conditions tested.

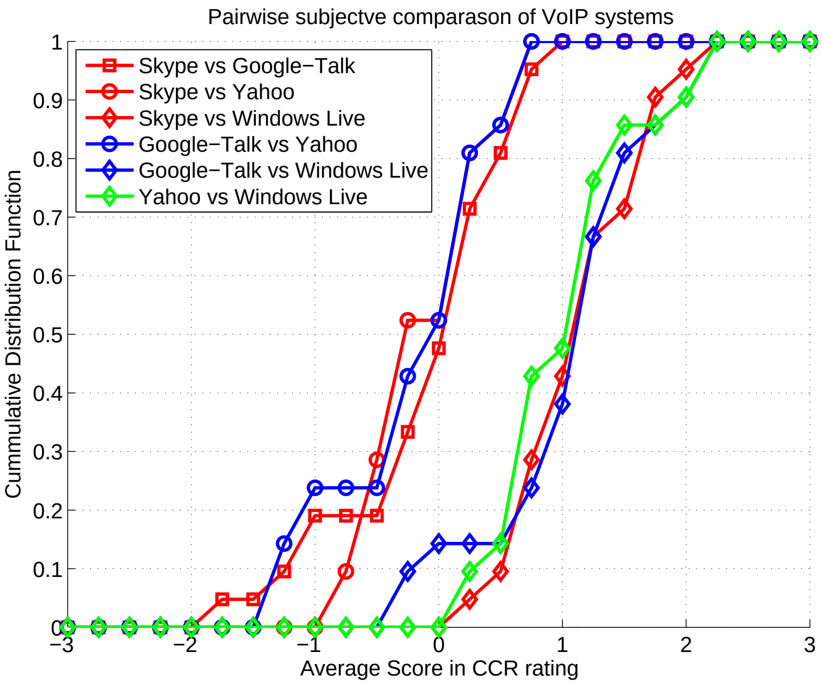


Figure 4.2: Distribution of pairwise subjective comparison scores of four VoIP systems under all network conditions. Positive value indicates 2nd system is preferred over 1st system.

4.2 Generalization of Comparative Subjective Evaluation Results

In this section, we present a classifier for predicting the comparative subjective evaluation results using objective measures that can be easily collected by our testbed. A classifier has the advantage over the time-consuming subjective tests, which are expensive to conduct and do not scale well with the number of systems compared. Note that subjective results cannot be fully predicted by objective means because the subjective results are not totally consistent themselves.

There are several statistical tools that are commonly used for multi-class classifications. In this study, we employ a Support Vector Machine (SVM) due to its speed and accuracy [8]. We use 22 inputs (features) that can be objectively obtained from the conversation recordings collected by our testbed. For each of the two systems compared, we input their CS, CE, the average PESQ and the MED of the conversation, and their variances across the speech segments and turns. To characterize the relative performance of the system, we also input their ratio and difference of the average PESQs and MEDs. To characterize the conversational condition, we input the average single-talk duration ($\overline{ST}$) and average HRD ($\overline{HRD}$), as well as the turn-switching frequency (SF). To characterize the network condition, we use the average network delay (ND), the percentage of packets that exhibit jitter of more than 60 msec (NJ) and the average loss rate (NL).

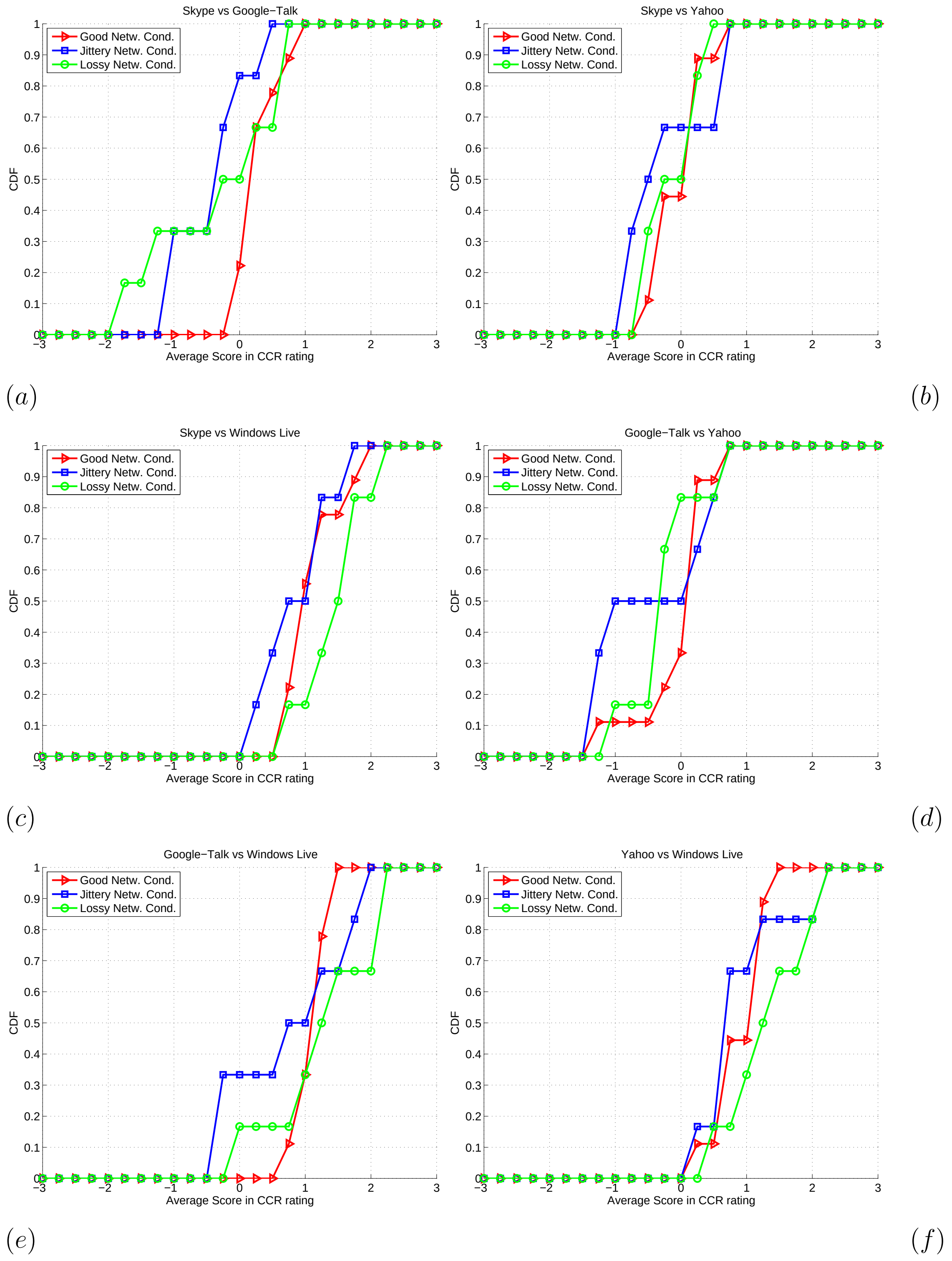
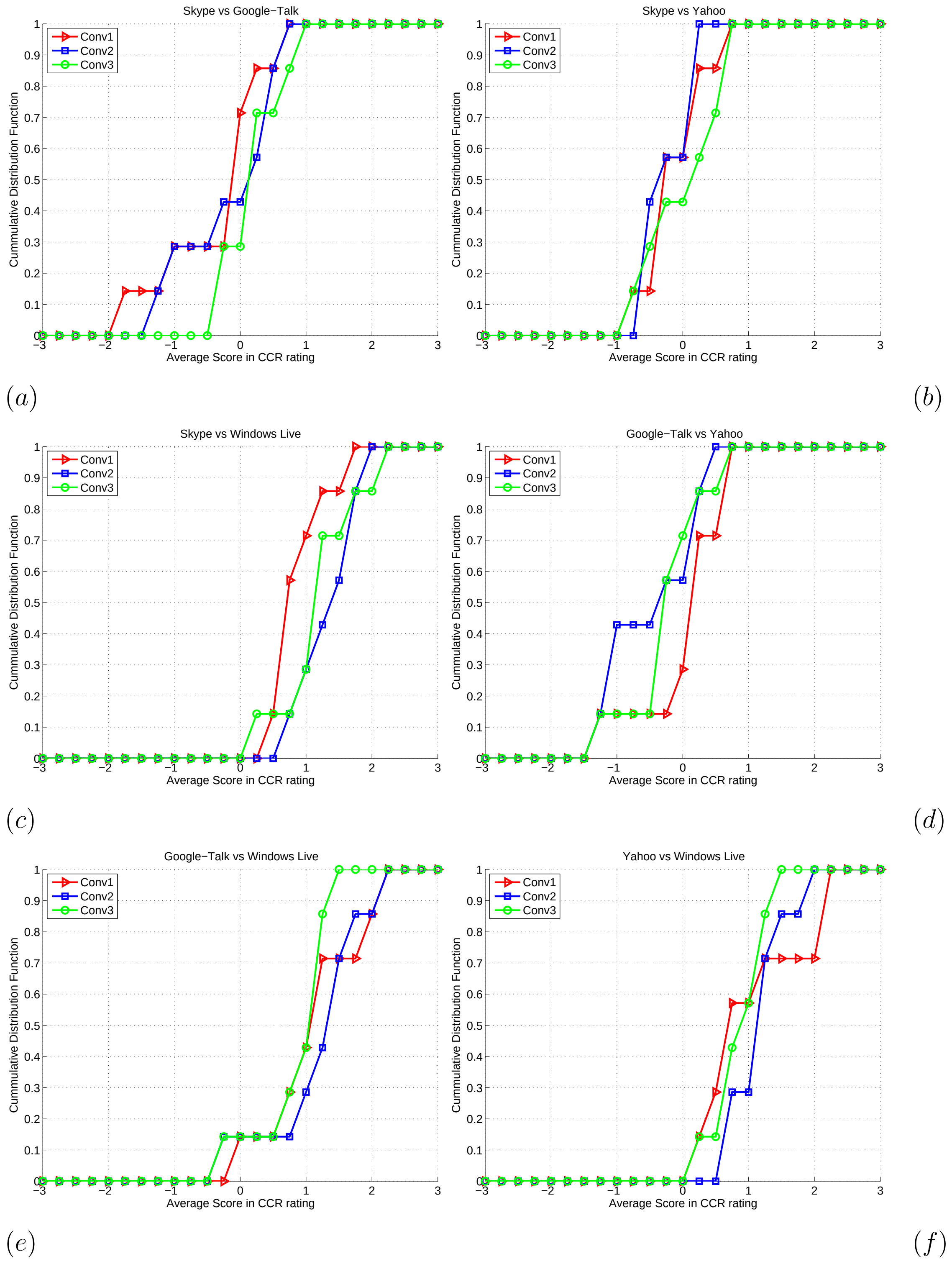


Figure 4.3: Distribution of pairwise subjective comparison scores of four VoIP systems conditioned on good, jittery, and lossy network conditions. Positive value indicates 2nd system is preferred over 1st system.



(a) (b) (c) (d) (e) (f)

Figure 4.4: Distribution of pairwise subjective comparison scores of four VoIP systems conditioned on conversational conditions. Positive value indicates 2nd system is preferred over 1st system.

To ensure that training results can be generalized, we reduce the number of target classes from 7 in CCR scale to 3 classes: (A better than B), (B better than A), and (About the same). The reduction in the number of classes does not significantly affect our ability to compare the systems, since it still answers the fundamental question on whether the difference in quality between two systems is perceptible. Table 4.3 lists the raw subjective preferences of the panel with respect to the two systems compared under good, jittery and lossy conditions.

Once we obtain the distribution of the subjects' opinions, we apply hypothesis testing to determine the dominant opinion with statistical significance. Given K subjects have evaluated each comparison pair, the distribution obtained is denoted by the triplet (p_{-1}, p_0, p_1) that corresponds, respectively, to the three opinions: A is better than B, A is about the same as B, and A is worse than B. In our hypothesis tests to determine if one opinion is dominant with statistical significance, the number of responses choosing opinion $i \in \{-1, 0, 1\}$ is compared against a binomial distribution with non-dominant opinion. Specifically, the null hypothesis (H_0) is defined to be p_i drawn from $binomial(K, p \leq 0.5)$. If the null hypothesis can be rejected with 90% statistical significance, then opinion i is called *dominant*. By construction, no two opinions can be dominant at the same time. However, it is possible that no opinion is dominant with 90% statistical significance. In that case, the comparison between A and B is *inconclusive*. We use the dominance information as the target value for the classifier.

Once the input features and the target labels have been obtained, we use the radial-basis function as the kernel function to project the 22 dimensions to higher dimensions, where we search for a set of hyperplanes to separate the classes. We use a dynamic search tool in LIBSVM to find the optimal kernel parameters.

To ensure that the results can be generalized, we turn to cross-validation techniques commonly used in statistics. There are four general approaches for evaluating the performance of prediction and generalization.

- The same set of samples are used for training and validation. This is an upper bound on validation accuracy.

- Leave-one-out cross validation is used to train $N - 1$ samples and test the sample left out. The training/testing process is conducted N times by enumerating the sample left out for each sample. The result is deterministic since all combinations are enumerated.

- K-fold ($K = 10$) cross validation is used to randomly divide the sample set into 10 equal-sized subsets. In each evaluation, 9 subsets are used for training, and one used for testing. The process is repeated 10 times by enumerating the subset left out. To get a reliable average

Table 4.3: Comparative subjective evaluations of pairs of VoIP systems and the prediction results of our SVM model. In comparing A and B, the dominant opinion with 90% statistical significance is shown: $<$ (*resp.*, $\approx$, $>$, and ?): A is better than (*resp.*, about the same as, worse than, and inconclusive with respect to) B. In the inconclusive case, no dominance relation with 90% significance is found. Boxes marked with '*' represent those training or prediction results that differ from the subjective results.

System Pairs A vs. B	Conv. Type	Subjective Test Results							Prediction Results (Training Data)							Prediction Results (Unseen Data)						
		Trace							Trace							Trace						
		NNN	LLL	LLH	LHL	HLL	HLH	HHL	NNN	LLL	LLH	LHL	HLL	HLH	HHL	NNN	LLL	LLH	LHL	HLL	HLH	HHL
Skype	3	?	?	<	?	?	>	>	?	?	<	?	?	>	>	?	?	>*	?	?	>	>
vs.	4	?	<	>	>	?	?	<	?	<	>	>	?	?	<	?	<	>	>	?	>*	?
GTalk	5	≈	?	?	?	<	<	≈	≈	?	?	?	<	<	≈	≈	?	?	?	<	?*	?*
Skype	3	?	?	<	?	?	?	<	?	?	<	?	?	?	<	?	?	>*	?	?	>*	<
vs.	4	?	?	<	?	?	≈	>	?	?	<	?	?	≈	>	?	?	<	?	?	≈	<
Yahoo	5	?	≈	>	≈	<	<	>	?	≈	>	≈	<	<	>	<*	?*	>	?*	<	<	?*
Skype	3	<	<	<	≈	?	<	<	<	?*	<	≈	?	<	<	<	<	<	<*	?	<	<
vs.	4	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<
WinLive	5	<	<	<	<	<	<	?	<	<	<	<	<	<	?	<	<	<	<	<	<	?
GTalk	3	<	≈	<	>	?	<	≈	<	≈	<	>	?	<	≈	<	≈	<	>	?	<	≈
vs.	4	≈	≈	?	<	≈	>	>	≈	≈	?	<	≈	>	>	<*	?*	?	?*	≈	?*	≈*
Yahoo	5	≈	≈	?	≈	≈	?	>	≈	≈	?	≈	<	?	>	≈	≈	?	≈	≈	?	?
GTalk	3	<	≈	?	<	?	<	<	<	≈	?	<	?	<	<	<	≈	<*	<	?	<	<
vs.	4	<	<	<	<	<	<	?	<	<	<	<	<	<	?	<	<	<	<	<	<	?
WinLive	5	<	<	<	?	<	<	?	<	<	<	?	<	<	?	<	<	<	?	<	<	?
Yahoo	3	?	<	<	<	?	<	<	<*	<	<	<	?	<	<	<*	<	<	<	?	?*	<
vs.	4	<	<	<	<	<	<	?	<	<	<	<	<	<	?	<	<	<	<	<	<	?
WinLive	5	<	<	<	?	<	<	<	<	<	<	?	<	<	<	<	<	<	?	<	<	<

over random divisions of training and testing sets, the entire process is repeated over 10 different random divisions.

- K-fold ($K = 2$) cross validation is used to randomly choose half of the samples for training and the other half for testing. The training and testing sets are then swapped and the process is repeated. To get a reliable average, the entire process is repeated over 50 different random divisions.

The average classification rate is referred to as the cross-validation score. Since the classifier does not learn from the samples in the testing set, a high cross-validation score is interpreted as the ability of the classifier to generalize to samples with conditions not in the training set.

With our SVM model, we are able to successfully predict 97.6% of the samples in our training set and 64.3% when using 10-fold cross validations. To further validate our results, we use new conversations and packet traces and apply our classifier to predict the subjective results.

Table 4.3 shows the dominant comparative opinions for the subjective experiments, SVM predictions of the training set, and SVM predictions of the unseen data. We observe that all systems operate well under good network conditions, as the difference in performance among the systems is too small to be perceived. However, as network imperfections are introduced, there are clear user preferences in terms of conversational quality. Windows Live is strongly preferred over the other systems under lossy conditions. In contrast, Skype is slightly preferred over Google-Talk, and Windows Live is slightly preferred over Yahoo Messenger under jittery conditions. Further, the distributions of predicted comparative opinions between system pairs match closely to those obtained through subjective tests, even for unseen data. The results indicate that our SVM classifier can be used to comparatively evaluate the conversational quality of VoIP systems under a variety of network and conversational conditions.

4.3 Summary

In this chapter, we have presented our objective and comparative subjective evaluations of four popular VoIP systems under a representative set of network and conversational conditions. In order to isolate individual system designs for comparative evaluations, we have conducted this evaluation using a testbed we specifically designed for this purpose. Such a comprehensive analysis has not been conducted before.

In this chapter we have also presented a new methodology and testbed to conduct objective and subjective evaluations of VoIP systems and a mapping to predict the subjective preference between any two VoIP systems by only using the objective measures characterizing the VoIP conversation and the conditions under which the evaluation is conducted.

The significance of this work is that it allows us to comparatively evaluate the conversational quality of any two systems, under any network and conversational conditions, using the objective measures easily obtainable using our testbed. Later in this thesis, in Chapter 8, we utilize the mapping learned here to evaluate our newly designed VoIP system against the four VoIP systems evaluated in this chapter.

CHAPTER 5

MODEL OF PAIRWISE SUBJECTIVE COMPARISONS

Real-time multimedia communication systems are characterized by multiple counteracting objective quality metrics (such as delay and signal quality) that can be affected by various control schemes. However, the trade-offs among these metrics with respect to the subjective preferences of users are not defined.

In the previous chapter, we have presented the evaluation of VoIP systems in terms of conversational quality. Since the architecture and components of commercial VoIP systems are usually not available for analysis, we have to employ a black-box approach, which relies on controlling the inputs to the systems, and capture its outputs in a consistent method to evaluate the system itself. Thus, our analysis of the systems is limited to observations derived from the comprehensive objective and subjective evaluations under controlled conditions.

However, in the design of a new system with high perceptual quality, it is imperative to be able to understand the relationship between the conditions under which the system operates and the resulting performance of the system components and the system as a whole. Thus, we need to learn the trade-offs among the quality metrics characterizing a system with respect to the subjective preferences to select the proper control schemes that lead to the best subjective quality at run time. Since subjective tests are expensive to conduct and the number of possible control schemes and run-time conditions is prohibitively large, it is important that a minimum number of such tests be conducted offline, and that the results learned can be generalized to unseen conditions with statistical confidence.

To this end, Chapters 5-7 report how we devised a set of methodologies to evaluate perceptual quality in relation to the control parameters used and the conditions under which the system operates. These methodologies include the development of a model for comparative subjective tests, the study of efficient algorithms for scheduling a sequence of subjective tests, and the learning and generalization of limited offline subjective tests to guide the operation of the control schemes at run time.

Without loss of generality, we only consider an operating curve due to a single control scheme, although the approach can be easily extended to multiple control schemes. Our goal is to minimize the number of subjective tests needed in order to determine a locally optimal operating point to within some prescribed level of statistical confidence. A secondary goal is to efficiently schedule the subjective tests of multiple operating curves in a multimedia application. In our previous work [54], we have initially developed a model with a single global optimum. However, in real-life problems, it is possible to have more than one local optimum on an operating curve. Thus, to improve the applicability of our model, in this chapter we present the extended work which allows for multiple local optima on an operating curve. The method to identify one of the local optima using adaptive subjective comparisons is presented in Chapter 6.

In this chapter, we present our model of pair-wise subjective comparisons over an operating curve representing the feasible set of points of a control scheme employed by a real-time multimedia communication system. This model is based on the characteristics of real-time multimedia communication systems presented in Chapter 1.3 and the POS control design problem for VoIP systems presented in Chapter 3. In the next chapter (Chapter 6), we utilize this model extensively to develop efficient scheduling methodology to conduct subjective comparisons offline. Lastly, in Chapter 7, we apply this methodology on the design of POS control in VoIP systems.

One of the most important characteristics of the quality metrics is that each one of them is either monotonically non-decreasing or non-increasing with respect to the control parameter. These characterizations provide a basis for the formal definitions, properties and axioms that will be presented later in this chapter. We formally define notations, properties, axioms and lemmas about subjective comparison of two operating points, leading to the development of the model. To better illustrate the concepts, POS control design is used as a running example in our development of the model as well as in the development of the efficient scheduling methodology. Lastly, in this chapter we provide a concise but complete representation of the pair-wise comparison model on a two-dimensional figure (which is simply referred as the *model*) and describe the meaning of points, lines and regions on it.

Comparative Ranking. Similar to the subjective evaluations in Chapter 4.1, we utilize comparative rankings in the evaluation of two operating alternatives, but modify the possible responses. To determine the preferred operating point among a set of alternatives, a partial order that requires pairwise comparisons suffices. The partial order can be assessed by a measure that evaluates the relative quality of two alternatives in a *comparison category rating* (CCR) (similar to that described in Annex E of ITU P.800 [21]). By presenting two alternatives to each subject, one after another, the approach allows the incomparability of some alternatives to be identified and small differences between two to be more accurately evaluated. The disadvantage, however, is a significant increase in the number of tests because such tests will need to be conducted for each pair of alternatives instead of each alternative.

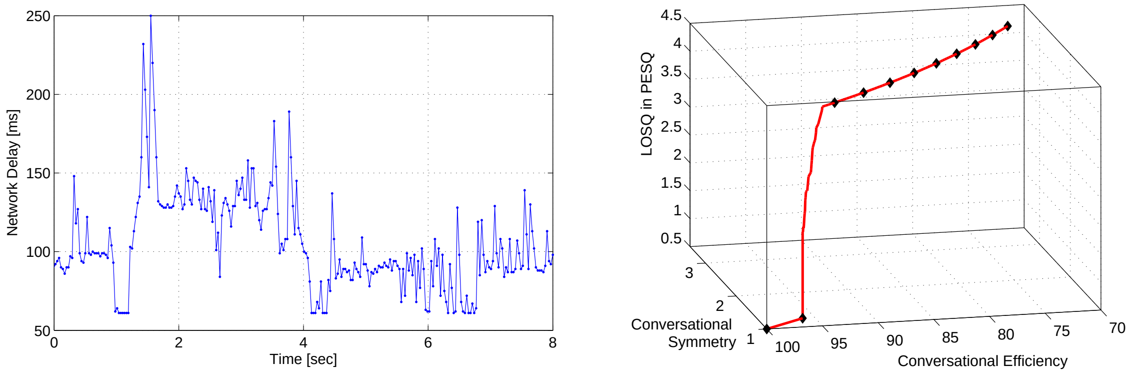
Testbed. In parallel to the testbed concept in Chapter 4.1, we have developed a specialized testbed for POS control scheme level simulation of a VoIP call. The testbed utilizes pre-recorded speech segments of a face-to-face conversation and the network traces collected from PlanetLab to replicate the conditions of an operating curve exactly on a packet-by-packet and sample-by-sample basis. Similar to the usage of the testbed in Chapter 4.1, this allows us to evaluate the effects of the changes in the control parameter by eliminating any changes in the other factors.

5.1 Running Example: Design of POS Control in VoIP Systems

Our approach can be used in many real-time multimedia communication applications. In this section, we describe an example application on the design of a *playout scheduler* (POS) algorithm for a real-time two-party VoIP system. This application demonstrates that subjective tests are needed to achieve high perceptual quality. It is also used as a running example to illustrate the algorithms developed for scheduling subjective tests.

Figure 3.5b depicts the quality of a conversational segment as a point in a 3-D space whose axes correspond to three objective metrics. Under a given conversational condition, the possible operating points are restricted to a curved plane perpendicular to the CS-CE plane, where two such conversational conditions are depicted in Figure 3.5b. Each plane illustrates the relation between CS and CE, parameterized with MED and conditioned on the conversational scenario. By using MED as the control parameter and by using LOSQ to characterize the quality of the speech segments received, the possible operating points are further restricted to a curve on one of these planes, where the operating point shifts towards the upper right-hand corner (higher LOSQ and CS and lower CE) as MED is increased.

The quality of the conversation perceived by the two clients is controlled by a POS algorithm, whose goal is to find the MED under some operating condition that leads to the best subjective conversational quality. A longer MED allows more packets to arrive in time for playout and improves LOSQ, but will degrade the interactivity and the efficiency of the conversation. The MED that optimizes subjective quality is not at the extremes of an operating curve, but at a point where the counteracting effects on subjective quality are relatively balanced. Further, the optimal MED depends on the given network and conversational conditions. For example, for a connection with high delays and jitter, the optimal MED may need to be higher in order to improve the poor LOSQ. In contrast, for a conversation in which clients take frequent turns, a lower MED may be preferred in order to reduce the annoying delay degradations. Since the network and conversational conditions



(a) Network delays from California to Maryland

(b) Example operating curve as a function of MED

Figure 5.1: Network condition and the corresponding operating curve for a conversation with medium switching frequency.

may change during a conversation, MED will need to be dynamically adjusted in a closed-loop fashion in order to continually maintain high perceived conversational quality.

Finding the best MED under a given condition is challenging because multiple subjective evaluations are needed in each comparison in order to arrive at some statistically significant conclusions. Moreover, MED can take continuous values, which result in infinitely many realizations of the POS algorithm. As a result, it is infeasible to conduct indiscriminate subjective tests in order to evaluate all possible pairs of conditions that can exist at run time.

Example 1 *The operating curve presented in this running example is used to illustrate the various concepts in this paper. We use a network condition with low delays, high jitter and low losses. Figure 5.1a depicts the packet delays observed as a function of sent times for a Maryland-California connection. (A comprehensive set of network conditions can be found in Chapter 2.2.) We use a conversational scenario with an average 24 switches/minute under no delay and whose average speech-segment and HRD durations are 1706 ms and 552 ms, respectively.*

Figure 5.1b shows a 3-D representation of the operating curve, where the curve is parameterized by MED. Starting from the lower left-hand corner, each diamond represents an increase of 100 ms in MED.

Table 5.1: The four possible opinions in a subjective test of A and B.

Condition	Notation	Probability	Notation
A better than B	$A >_s B$	$Pr(A >_s B)$	$p_1(A, B)$
A about the same as B	$A \approx_s B$	$Pr(A \approx_s B)$	$p_0(A, B)$
A worse than B	$A <_s B$	$Pr(A <_s B)$	$p_{-1}(A, B)$
A incomparable to B	$A?_sB$	$Pr(A?_sB)$	$p_2(A, B)$

5.2 Model of Subjective Comparisons

In this section, we present the properties of comparative subjective tests, which lead to a general model for evaluating points on an operating curve. We first present the notations and basic axioms on comparative subjective evaluations. Next, we define the local optimality of an operating point and the possibility of multiple local optima on an operating curve. Based on the region of dominance of a local optimum, we present stronger axioms that are valid within the region. This is followed by a stochastic model on pairwise subjective comparison tests.

5.2.1 Comparative subjective tests

Comparative subjective tests are conducted by comparing two points A and B on an operating curve. Each test is conducted by asking a subject to compare A and B in a random order (to avoid any perceived bias). The alternatives are generated under the same operating conditions but under different control parameter values.

Table 5.1 shows the comparison results in one of the four opinions, where p_i is obtained by normalizing the number of subjects who responded with opinion i with respect to K, the total number of subjects. Note that the complement to the opinion "$A \approx_s B$" is "A is not about the same as B" (or $A \not\approx_s B$), which consists of $A >_s B$, $A <_s B$, and $A?_sB$.

The results of the subjective tests can be combined into a stationary probability distribution (or a sample distribution when K is finite) in the form of a vector called the *comparative opinion distribution* (COD):

$$COD(A, B) = \overline{p} = (p_{-1}, p_0, p_1, p_2) \tag{5.1}$$

$$\text{where } \textstyle\sum_i p_i(A, B) = 1 \text{ for each } (A, B) \text{ pair.} \tag{5.2}$$

To avoid confusion, p_i is assumed to be stationary (when K is large), and we represent a sample probability by $\hat{p}_i$ where necessary.

Example 1 (cont'd). *Using the testbed we have developed, under the given network and conversational condition in Figure 5.1b, each point on the operating curve can be simulated to result in a VoIP conversation. For example,* $A = 0.111$ *and* $B = 0.198$ *result in two conversations generated using, respectively, MEDs of 111 ms and 198 ms, and represented by* $(PESQ, CS, CE) = (2.60, 1.27, 0.96)$ *and* $(4.18, 1.48, 0.93)$*. After subjective tests,* $COD(A, B) = (p_{-1}, p_0, p_1, p_2) = (0.75, 0.125, 0.125, 0)$*, which means that* B *is preferred over* A*.*

5.2.2 Monotonicity of objective metrics

An *operating curve* is denoted by $\mathcal{O}$, which is mapped to a real number in $[0, 1]$ with extreme points $A^{\min} = 0$ and $A^{\max} = 1$. Each point on the curve is denoted by a capital letter (such as A and B) and has a one-to-one correspondence to the value of the associated control scheme that realizes the communication application. Thus, changes to the scalar control value are mapped to changes in one of the two directions along the operating curve.

Property 1 Monotonicity. *Each objective metric is either monotonically non-increasing or monotonically non-decreasing with respect to increases in the corresponding control value.*

Example 1 (cont'd). *Referring to Figure 5.1b,* $A^{\min}$ *represents the degenerate case in which POS plays each speech frame at the instant it was spoken by the remote client, whereas* $A^{\max}$ *represents the case in which POS waits 1 sec after each speech frame is spoken before playing it.*

In general, changes in MED may improve, degrade or have no effect on the corresponding objective metrics. In this example, new packets can arrive in time for playback when MED is increased (due to either a redundant loss-concealment scheme or a higher chance for late packets arriving), which means that there is a non-increasing relationship between MED and the rate of late packets. Because the arrivals of packets happen at discrete times, the objective metrics may have finite discontinuities when MED is increased. The relation between LOSQ and MED depends on the robustness of the speech codec on losses and the network jitter, although it is always monotonically non-decreasing. Similarly, the relation between CS (resp. CE) and MED depends on HRD (resp. ST and HRD) and is monotonically increasing (resp. decreasing) with respect to MED by definition [55].

It is possible for the monotonicity property to be violated in such a way that there are multiple local optima with respect to an objective metric when the control variable is increased. In this case, the operating curve can be divided into multiple non-overlapping segments called *regions of dominance* (Section 5.2.4), where the objective metrics in each region satisfy the monotonicity property.

Implications on subjective preferences. The monotonicity property ensures that when perturbing from A to B, a subset of the objective metrics exhibit a non-decreasing trend, whereas the remaining metrics exhibit a non-increasing trend. However, the trade-offs among the metrics do not necessarily result in a bell-shaped subjective preference curve. These trade-offs can change at different operating points because they depend on the perceived degradation of each quality metric as well as the relative change in that perception. As a result, there can be multiple locally optimal subjectively preferred points within a region where monotonicity of the objective metrics is satisfied. For example, if a dominant degradation is common to both A and B, then a subject may more likely prefer the point that exhibits improvement in the dominant degradation. In contrast, if A exhibits a significant improvement on the less perceived degradation, while no perceptible difference is observed between A and B with respect to the dominant degradation, then A is more likely to be preferred.

5.2.3 Basic axioms

Axiom 1 Reflectivity. *Comparing a point with itself results in the $A \approx_s A$ opinion from an individual perspective and $p_0(A, A) = 1$ from a collective perspective.*

Since there is no difference in the objective metrics between A and itself, subjects should not perceive them to be different except for mistaken evaluations.

Axiom 2 IID. *Each subject has the same level of expertise, and their responses to comparing any two points on an operating curve are independent and identically distributed (IID).*

This axiom allows us to model the sample COD in (5.1) by a multi-nomial distribution. In particular, the order of a comparison does not affect $COD(A, B)$, as stated in the following axiom.

Axiom 3 Symmetry/anti-symmetry. *Indistinguishable ($\approx_s$) and incomparable ($?_s$) opinions are* symmetric*: $p_i(A, B) = p_i(B, A)$ for $i \in \{0, 2\}$. Preference opinions ($>_s$ and $<_s$) are* anti-symmetric*: $p_{-1}(A, B) = p_1(B, A)$.*

Let $B - A$ be the perturbation in the control value from a fixed A to a variable B. Since each objective metric is monotonic with respect to B and a small change in B may result in a possibly discrete but finite change in the objective metric, there will be a small fraction of subjects perceiving a difference in quality. Hence, a small change in B will result in a possibly discrete change in the probability of perceiving such a difference when the number of subjects is large. As the difference between A and B increases, the perception of the difference in their subjective quality increases. This noticeable difference is commonly used in psychophysics and is defined as follows.

Definition 1 Just Noticeable Difference (JND) of A. *When comparing a fixed A and a variable B on an operating curve $\mathcal{O}$, $JND(A)$ is the $B - A$ value for which 50% of the subjects perceive a difference in their quality.*

In statistical inference from a finite number of subjective evaluations, we define $JND(A)$ to be the minimum value of $|B - A|$ such that the hypothesis, $\{H_0 : p_0(A, B) < 0.5\}$, is rejected with a given statistical significance. If B is inside the JND of A ($|B - A| \leq JND(A)$), then A and B are *indistinguishable*; otherwise, they are *distinguishable*.

Definition 2 Complete Noticeable Difference (CND) of A. *When comparing a fixed A and a variable B, $CND(A)$ is the minimum $|B - A|$ value such that $p_0(A, B) = 0$.*

Axiom 4 Indistinguishability. *The probability of an indistinguishable opinion, $p_0(A, B)$, is monotonically non-increasing with respect to $|B - A|$ for fixed A and variable B.*

When $B = A$ (thus, $|B - A| = 0$), $p_0(A, A)$ is equal to 1 due to Axiom 1. As $|B - A|$ increases, there are larger differences in their objective metrics, resulting in a non-decreasing number of subjects perceiving a difference in their quality. Eventually, all subjects perceive that A is not the same as B.

As a result of Axiom 4, Figure 5.2 shows that the $JND(A_0)$ and $CND(A_0)$ regions are single contiguous regions around A_0. $JND(A)$ and $CND(A)$ can vary as a function of A. For some A, a small perturbation in the control may result in the perception of a difference in subjective quality. For another A, it may require a large perturbation. Our subjective tests in two-party VoIP conversations confirm the variations in JND and CND as a function of A.

Example 1 (cont'd). *The example VoIP application satisfies Axioms 1-3. To illustrate JND, we have conducted pair-wise subjective evaluations between alternatives on the operating curve in Figure 5.1b. Firstly, we compare $A = 0.2$ with B that has larger MED with respect to that of A. As the difference in MEDs increases, the fraction of responses indicating that the two are about the same decreases. This happens because the differences in LOSQ, CS, and CE increase at the same time, and a larger fraction of the subjects can perceive the difference between the conversations. We then repeat the experiments using $A' = 0.3$. We observe that the JND observed tends to be*

Comparison of fixed A_0 and variable $B \in \mathcal{O}$

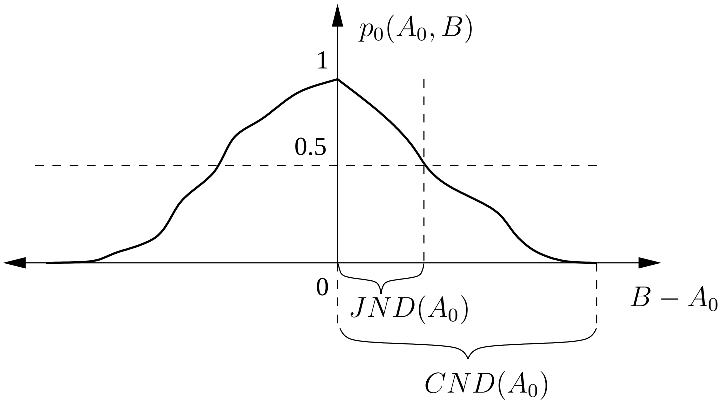


Figure 5.2: Comparing a fixed A_0 with a variable B: $p_{_0}$ is non-increasing as a function of $B - A_0$.

Table 5.2: $p_{_0}$ of subjective comparisons between various B when compared to A and A' in Figure 5.1b.

	B					
	0.25	0.30	0.35	0.40	0.50	0.60
$A = 0.20$	0.95	0.85	0.65	0.45	–	–
$A' = 0.30$	–	–	1.00	0.90	0.60	0.40

larger than that in the first experiment, which means that subjects are less sensitive to perceiving the changes. This is due to the fact that, as the baseline degradations due to MED are larger, the noticeability threshold, which is related to the baseline MED, is also larger. This behavior is illustrated in Table 5.2, which lists the values of $p_{_0}$ *for various* B *when compared to* A *and* A'*.*

5.2.4 Locally optimal points on an operating curve

Intuitively, an optimum is a point that is preferred when compared to every other feasible point on an operating curve. It is preferred because it achieves the optimal trade-off among the various objective metrics and cannot perform better by operating at another point. Hence, identifying such a point is paramount in the design of adaptive system-control schemes.

Definition 3 Local optimum. *Point* A_i^* *is locally optimal over points in a subset of the operating curve* $\mathcal{O}_i \subseteq \mathcal{O}$*:*

$$A_i^* = \{A \mid p_{_1}(A, B) > 0.5\ \forall B \in \mathcal{O}_i \text{ s.t. } |B - A| > JND(A)\}. \tag{5.3}$$

There can be multiple local optima on an operating curve, since changes in the multiple objective metrics along the operating curve may lead to locally optimal trade-offs.

Definition 4 Region of Dominance (ROD**).** *The* ROD *of a local optimum* A_i^**,* $ROD(A_i^*)$*, is the largest contiguous region* $\mathcal{O}_i$ *of an operating curve* $\mathcal{O}$ *in which (5.3) is satisfied.*

A local optimum is dominant (or preferred more than 50% of the time) against any point within its ROD, except for points in its JND region. However, when A_i^* is compared against B outside its ROD ($B \notin ROD(A_i^*)$), we cannot conclusively say whether A_i^* is preferred over B; that is, the hypotheses $p_0(A_i^*, B) > 0.5$, $p_1(A_i^*, B) > 0.5$, and $p_{-1}(A_i^*, B) > 0.5$ are all rejected with some statistical significance. Similarly, nothing can be concluded when comparing a point in the ROD of one local optimum with a point in the ROD of another local optimum.

Example 1 (cont'd). *The operating curve in Figure 5.1b has a single local optimum, although it cannot be proved unless infinitely many subjective tests are conducted. In the next chapter, we illustrate the existence of the local optimum by conducting a finite number of tests. With one local optimum, the operating curve has one ROD.*

Lemma 1 *There cannot be multiple local optima that are within the ROD of each other.*

Proof 1 *Assume two local optima that are within the ROD of each other (e.g.* $A_1^* \in ROD(A_2^*)$ *and* $A_2^* \in ROD(A_1^*)$*). From (5.3),* $p_1(A_1^*, A_2^*) > 0.5$*, and* $p_1(A_2^*, A_1^*) > 0.5$*. Due to Axiom 3,* $p_{-1}(A_1^*, A_2^*) > 0.5$*; thus* $\sum_i p_i(A_1^*, A_2^*) > 1$*. Contradictions!*

Definition 5 Global optimum, A^**, is a point that dominates all points on an operating curve (*$ROD(A^*) = \mathcal{O}$*).*

Each operating curve is not guaranteed to have a global optimum. However, if one exists, it is unique. That is, there cannot be another point outside the JND of the global optimum that satisfies the property of global optimality. This is stated formally as follows.

Lemma 2 *If a global optimum exists, then it is unique.*

Proof 2 *Assume that two global optima* A_1 *and* A_2 *satisfy (5.3) for* $\mathcal{O}_i = \mathcal{O}$ *and that* $|A_1 - A_2| > \max\{JND(A_1), JND(A_2)\}$*. Then* $p_1(A_1, A_2) > 0.5$ *and* $p_1(A_2, A_1) > 0.5$*. Thus,* $\sum_i p_i(A_1, A_2) > 1$*. Contradiction!*

It is, however, possible that multiple points, all within the JND of each other, satisfy the definition of global optimality. This does not cause any inconsistencies because any of the candidate points can be chosen as the global optimum and no candidate is distinguishable from another.

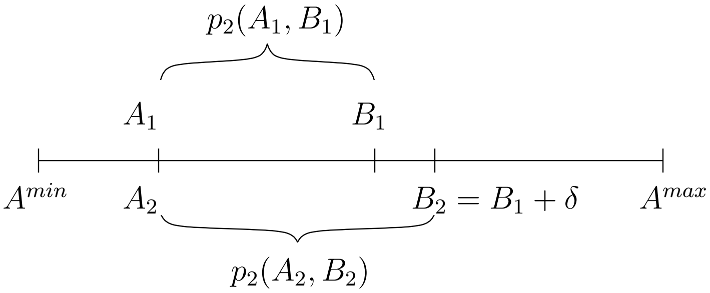


Figure 5.3: Comparison of A_1 with B_1 and A_1 with B_2. p_2 increases when $\delta > 0$.

5.2.5 Incomparability within the ROD of a local optimum

At a local optimum A_i^*, the quality metrics have an optimal trade-off. However, due to Property 1, as the point is perturbed away from A_i^* in one direction (say towards $A^{\max}$) but still within the ROD of A_i^*, a subset of the metrics exhibit more perceptible degradations that dominate the other metrics. On the other hand, if the point is perturbed in the other direction (say towards $A^{\min}$) but within the ROD of A_i^*, then a different subset of the metrics exhibit more perceptible degradations. Thus, when a subject is asked to compare the two points on different sides of A_i^*, the subject may indicate that the pair is incomparable, since different subsets of quality aspects dominate the degradation. As the distance between these points increases, the overlap between the subsets of dominant quality aspects is reduced, which causes more subjects to indicate that the pair is incomparable. As an example, in two-party VoIP, perturbations from the local optimum in one direction cause degradations due to delay to be dominant, while such perturbations in the other direction cause degradations due to speech quality to be dominant.

Axiom 5 Incomparability of A and B. *In the stationary case when there are a large number of subjects* $(K \to \infty)$*,* $\lim_{\delta\to 0^+} p_2(A, B + \delta) \geq p_2(A, B)$*, and* $\lim_{\delta\to 0^+} p_2(A - \delta, B) \geq p_2(A, B)$*.*

Figure 5.3 illustrates the axiom. Assume that metrics 1 and 2 are monotonically non-increasing and that metrics 3 and 4 are monotonically non-decreasing. Given the comparison of A_1 and B_1, a second comparison between A_1 and B_2 is conducted, where B_2 is perturbed by an infinitesimal amount from B_1 ($B_2 = B_1 + \delta$, $\delta \to 0$). Due to the monotonicity of the metrics, a perturbation from B_1 to B_2 causes in some of the perceptible objective metrics to be less perceptible and some less perceptible ones to be more perceptible. This causes the subject to depend on a slightly different set of quality aspects in evaluating the quality trade-offs. This change results in a reduction in the overlap of the important metrics between A_1 and B_2 with respect to A_1 and B_1. Hence, the probability that subjects perceive A_1 and B_2 to be incomparable is monotonically non-decreasing with respect to A_1 and B_1.

Table 5.3: $p_2(A, B)$ for the operating curve in Figure 5.1.

(A, B)	A				B				p_2
	MED	PESQ	CS	CE	MED	PESQ	CS	CE	
(0.11,1.00)	110	2.39	1.27	0.96	1000	4.18	3.42	0.72	0.6
(0.11,0.50)	110	2.39	1.27	0.96	500	4.18	2.21	0.84	0.2
(0.25,0.50)	250	4.18	1.60	0.91	500	4.18	2.21	0.84	0.0

Lemma 3 *For $K \to \infty$ and any finite $\Delta > 0$, $p_2(A, B + \Delta) \geq p_2(A, B)$, and $p_2(A - \Delta, B) \geq p_2(A, B)$, where A, B, $B + \Delta$, $A - \Delta \in ROD(A_i^*)$.*

Proof 3 *The proof follows directly from Axiom 5 after cascading together infinitesimal changes.*

Corollary 1 $p_2(A_2, B_2) \geq p_2(A_1, B_1)$ *if* $[A_1, B_1] \subseteq [A_2, B_2]$*, where* A_1*,* B_1*,* A_2*,* $B_2 \in ROD(A_i^*)$*.*

Corollary 2 $p_2(A_i^{\min}, A_i^{\max}) \geq p_2(A, B)$ *for all* $A, B \in ROD(A_i^*)$*.*

Example 1 (cont'd). *As the difference between the MEDs of two operating points is decreased by moving one or both of the points closer to the other, the incomparability rate among the subjects decreases as well. Table 5.3 summarizes the results of the subjective tests. Three comparisons were conducted, where the $[A, B]$ segment of each subsequent pair is a subset of the previous segments (e.g. $[0.25, 0.50] \subset [0.11, 0.50] \subset [A^{\min}, A^{\max}]$).*

5.2.6 Subjective preference within the ROD of a local optimum

Consider a pair of operating points in the ROD of a local optimum A_i^*. In contrast to the indistinguishable (p_0) and incomparable (p_2) opinions, p_1 and p_{-1} contain information on the location of A_i^*. In this section, we present our observations and basic axiom on the preference of one point over another. These results are used later to represent the information deduced on the location of A_i^*.

Axiom 6 Subjective preference. *For $K \to \infty$ and A and B on the same side of A_i^* where $A < B$,*

$$|p_1(A, B) - p_{-1}(A, B)| \leq \begin{cases} \lim_{\delta \to 0^+} |p_1(A - \delta, B) - p_{-1}(A - \delta, B)| \\ \lim_{\delta \to 0^+} |p_1(A, B + \delta) - p_{-1}(A, B + \delta)|. \end{cases} \tag{5.4}$$

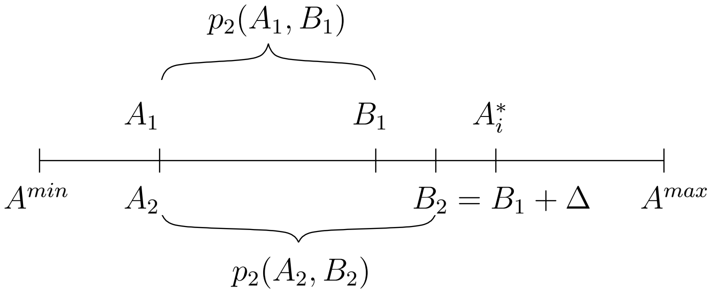


Figure 5.4: Comparing A and B in the same side of the local optimum A_i^*: $|p_1 - p_{-1}|$ increases when B is perturbed towards A_i^*.

Figure 5.4 explains the axiom intuitively. As B_1 moves to B_2 towards A_i^*, it will have more balance in its objective metrics and better perceived quality when compared to A_1. The difference between p_1 and p_{-1} indicates the conclusiveness of this perceptual comparison because it represents the improvement of the preferred opinion with respect to the non-preferred opinion. As B moves towards A_i^*, the conclusiveness of the comparison improves.

The POS-design problem described in Section 5.1 generally exhibits this property, where the preference towards the alternative closer to the optimal point increases as the other alternative moves away from the optimum. This perturbation makes either LOSQ or delay degradation more dominant and, thus, more perceptible.

Definition 6 Control symmetry. *For A and B on opposite sides of A_i^*, A and B are objectively symmetric, denoted by $A\|_0 B$, if they are equi-distant from A_i^* in terms of their control value; that is, $|A - A_i^*| = |B - A_i^*|$.*

Definition 7 Subjective symmetry. *For A and B on opposite sides of A_i^*, A and B are subjectively symmetric, denoted by $A\|_s B$, if $p_1(A, B) \leq p_{-1}(A, B - \delta)$ and $p_1(A, B) \geq p_{-1}(A, B + \delta)$, where $\delta \rightarrow 0$. This means that the probabilities of one point being preferred over another are equal from both directions.*

In the special case in which the objective metrics are continuous at A and B with respect to the control value, subjective symmetry results in $p_1(A, B) = p_{-1}(A, B)$. To account for possibly finite discontinuities in the objective metrics at A and B, we need to define subjective symmetry with respect to $\delta \rightarrow 0$.

Lemma 4 *A subjectively symmetric point B on the opposite side of A_i^* with respect to A exists if $p_1(A, A_i^*) \geq p_{-1}(A, A_i^*)$ and $p_1(A, A_i^{\max}) \leq p_{-1}(A, A_i^{\max})$, where $A < A_i^* < B < A_i^{\max}$. Such a point B, if it exists, is unique.*

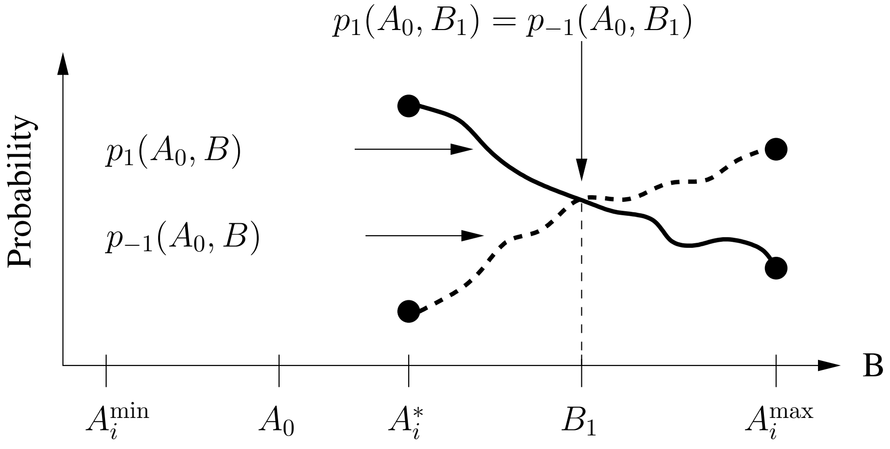


Figure 5.5: Subjective symmetry of A_0 and B.

Proof 4 Existence. *We know that p_1 and p_{-1} are either non-increasing or non-decreasing between A_i^* and $A_i^{\max}$ when A is fixed and B is between A_i^* and $A_i^{\max}$. Assuming $p_1(A_0, A_i^*) \geq p_{-1}(A_0, A_i^*)$ and $p_1(A_0, A_i^{\max}) \leq p_{-1}(A_0, A_i^{\max})$, then there exist at least one B_1 at the cross-over point in Figure 5.5 between the two curves that satisfy the condition in Definition 7 with respect to A_0 and B, namely, $A_0 \|_s B$.*

Uniqueness. *Since the functions $p_1(A_0, B)$ and $p_{-1}(A_0, B)$ are monotonic (non-increasing or non-decreasing), point B that satisfies the condition must be unique. In cases where both functions are constant, namely, $p_1(A_0, B) = p_{-1}(A_0, B)$, then there is a region in which the condition is satisfied. Since all points in such a region satisfy the condition, the region is unique.*

The comparison of subjectively symmetric points does not result in any new information on the location of A_i^*. However, when comparing A with any point that is larger than B where $A \|_s B$, then A is more preferred. This observation will be useful for deducing the location of A_i^* from the result of subjective tests.

5.2.7 General model of subjective comparisons

In this section, we use the axioms presented to develop a general model of subjective comparisons. Figure 5.6 depicts the general case with multiple local optima, where the axes represent the two points compared. Due to Axiom 3, it suffices to assume $B \geq A$. We focus on a detailed description of the COD for $ROD(A_i^*)$.

A more restricted model describes the probabilities of occurrence of the four possible opinions when comparing A and B in one ROD. Figure 5.7 depicts the model and the eight regions, whose

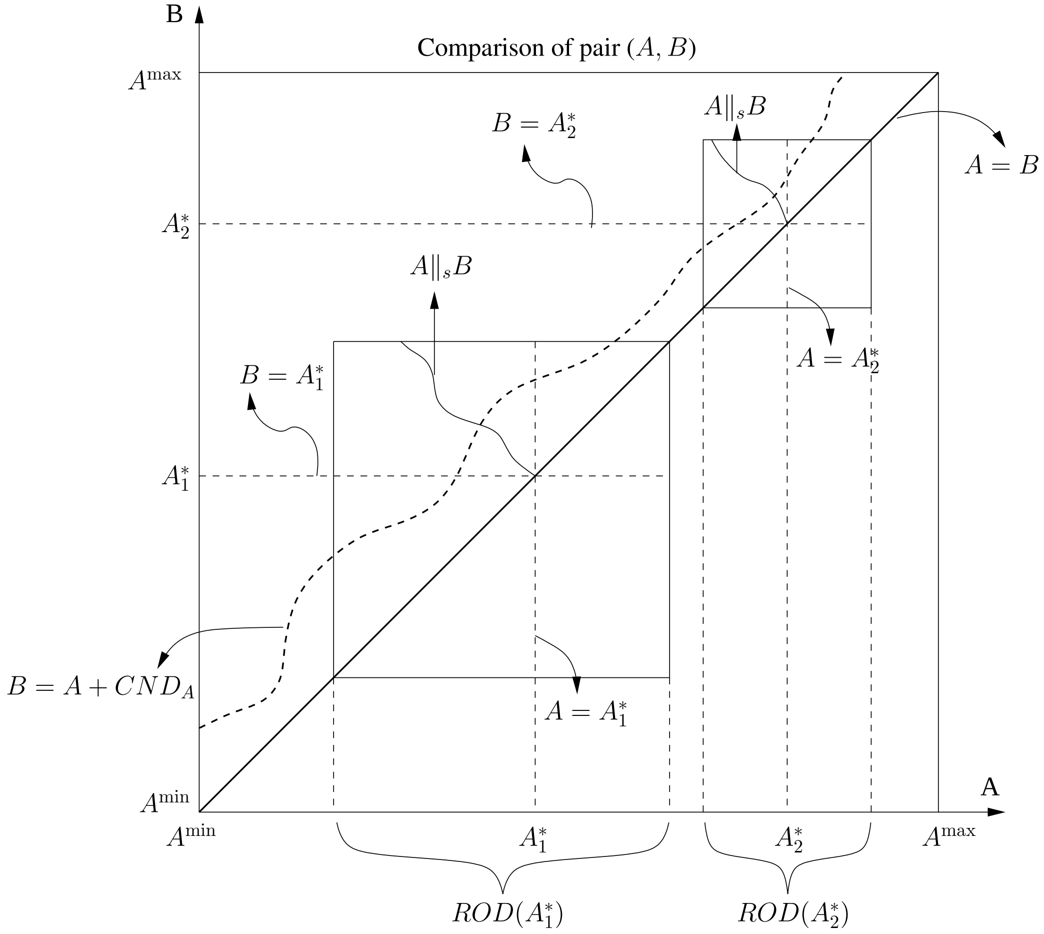


Figure 5.6: Model of subjective comparison of two operating points A and B on an operating curve. Model with multiple local optima and their RODs.

properties on COD are summarized in Table 5.4. The eight regions defined with respect to the four boundary lines have the following properties.

- *Regions R1 and R5:* A and B are on the same side of A_i^*, where $A < B < A_i^*$. If a pair compared belongs to these regions, then the result satisfies $p_{-1} < p_1$, according to Axiom 6. Without knowing A_i^*, such a result indicates that A_i^* is more likely to be larger than B. This is consistent with the actual location of A_i^* and guides the search in the right direction.

- *Regions R4 and R8:* A and B are on the same side of A_i^*, where $A_i^* < A < B$. If a pair compared belongs to these regions, then the result satisfies $p_1 < p_{-1}$, according to Axiom 6. Without knowing A_i^*, such a result indicates that A_i^* is more likely to be smaller than A. This is consistent with the actual location of A_i^* and guides the search in the right direction.

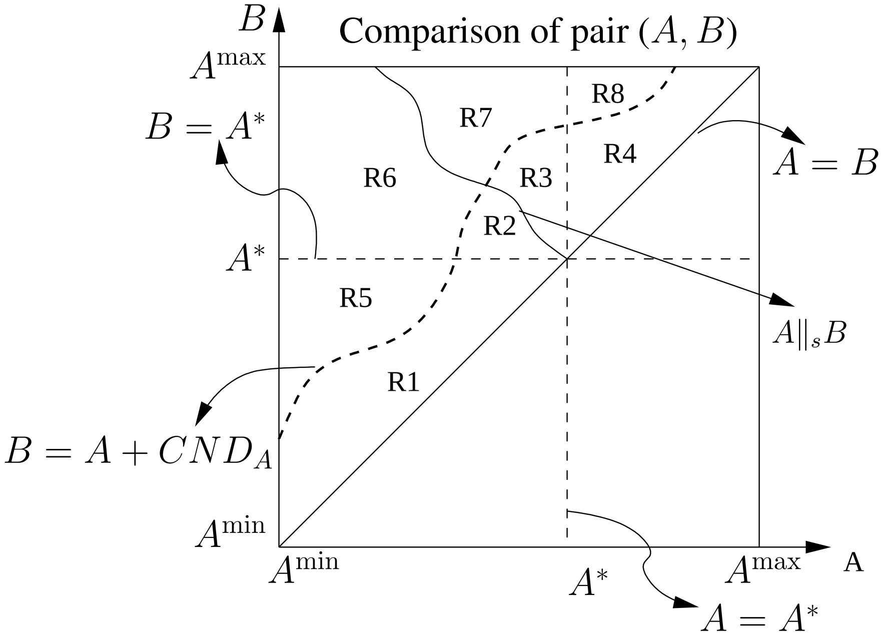


Figure 5.7: Model of subjective comparison of two operating points A and B on an operating curve. Regions identified in one ROD.

Table 5.4: Properties on COD of the eight regions in Figure 5.7 defined with respect to the four boundary lines with $B > A$: $B - A - CND(A)$, $A - A^*$, $A\|_s B$, and $B - A^*$.

	Regions in a ROD							
	R1	R2	R3	R4	R5	R6	R7	R8
p_0	> 0	> 0	> 0	> 0	0	0	0	0
p_{-1} vs. p_1	$p_{-1} > p_1$	?	?	$p_{-1} < p_1$	$p_{-1} > p_1$	?	?	$p_{-1} < p_1$
p_2	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0

- *Regions R2, R3, R6, and R7:* A and B are on opposite sides of A_i^*, where $A < A_i^* < B$. If a pair compared belongs to these regions, then p_1 or p_{-1} can be larger. Such a result is inconclusive for guiding the search.

The model does not specify the result when comparing one point in a ROD against another point outside of the ROD. Such comparisons do not provide information for identifying a local optimum and should be avoided.

Lemma 5 *The RODs corresponding to different local optima do not overlap.*

Proof 5 *Assume that $ROD(A_1^*)$ and $ROD(A_2^*)$ overlap and that A and B are chosen in the overlapped region. Without loss of generality, assume $A_1^* < A_2^*$. By applying Axiom 6 on A_1^*,*

$p_1(A,B) > p_{-1}(A,B)$. *On the other hand, when applying Axiom 6 on* A_2^**,* $p_1(A,B) < p_{-1}(A,B)$. *Contradiction!*

5.3 Summary

In this chapter, we have presented our model of pair-wise subjective comparisons over an operating curve representing the feasible set of points of a control scheme employed by a real-time multimedia communication system. The model is constructed using only the most basic properties and axioms, and thus would fit of a wide variety of comparison domains. The model allows a concise but complete representation of all the possible comparisons between any two points on the operating curve. The model allows for multiple local optima, and thus for the formulation of hierarchical search algorithms to find one or all of the local optima as the problem requires. The model also does not make any implicit assumptions on the multiple underlying quality metrics that affect the trade-off on the overall quality. The only assumption is that each individual quality metric is either monotonically non-decreasing or non-increasing with respect to the control parameter and that at least one metric belongs to each of these groups such that there is a non-trivial trade-off between the quality metrics. This assumption is reasonable and quite simple to satisfy, as only the metrics that satisfy these criteria individually are utilized in the prediction of the overall quality.

In the following chapter (Chapter 6), we utilize this model extensively to develop an efficient scheduling method to conduct subjective comparisons offline. The reason for separately presenting the two chapters is that the model itself is general and can be utilized in solving other problems involving multiple metrics and a previously undefined overall metric that defines the overall performance. On the other hand, the next chapter is more specific in terms of the type of problem solved and what is expected as the result. To be more specific, the next chapter considers efficient and accurate algorithms to find local optima of the model. However, in other problems the goal may be to find all local optima, and there may be some constraints on the efficiency or the accuracy of the search. Thus, the general model presented in this chapter can be used directly to formulate such problems, thus deserving a separate chapter for its presentation.

CHAPTER 6

STATISTICAL SCHEDULING OF OFFLINE COMPARATIVE SUBJECTIVE EVALUATIONS

In this chapter, we present our methodology to schedule subjective comparisons based on the model developed in Chapter 5.

One of the main goals of this thesis is the design of POS schemes that achieve high subjective quality. This goal requires the identification of the optimal point at run time under previously unseen conditions. To this end, we consider in this chapter the problem of determining the best operating point offline under a given set of network conditions and communication scenarios.

This requires conducting subjective comparisons between some pairs of alternatives. When conducting a limited number of subjective evaluations, we use a simulator to generate the alternatives compared. The goal is to repeat the network and communication conditions in order to eliminate variations other than the differences in the control schemes tested. We then collect the comparative subjective opinions and represent them as discrete distributions. Thus, the first part of the study is on the stochastic comparative evaluations between alternatives generated under a given set of conditions.

Secondly, in this chapter, we present our study on the *pruning of search space*. The idea is to systematically use the observations from past subjective tests to prune tests that have not been conducted. Our approach is based on a statistical model of subjective evaluations that utilizes the following two principles: (a) the subjective quality induced by small changes in the control scheme cannot be perceived by subjects, and (b) subjective preferences between points that are in a contiguous subset of the operating curve generally point towards the locally optimal point in that subset. This portion of the study develops a framework to combine the information obtained from separate comparisons and guide the choice of alternatives for future comparisons.

In the next chapter (Chapter 7), we apply the method developed in this chapter to the design of POS scheduling for VoIP systems, which extends the method to multiple operating curves.

6.1 Efficient and Accurate Search Algorithms for Finding Local Optima

Based on the fundamental understanding of subjective tests in Chapter 5, we develop in this section a systematic approach for conducting pair-wise comparative subjective tests among points on one or more operating curves. There are two counteracting metrics of success for this task, the most important being the accuracy of the local optimum estimated. Since there are infinitely many points on a continuous operating curve, it is impossible to identify the optimum via a finite number of tests. However, it suffices to estimate the local optimum to within the JND of its actual location, since both are indistinguishable in this region. The second metric of success is the number of subjective comparisons conducted. Although more comparisons would lead to a better estimate, it is important to develop a method that achieves the desired level of accuracy using the minimum number of tests.

The development of an efficient search strategy is based on the following observations. Firstly, the COD obtained by comparing two points on an operating curve provides information on the preferred trade-offs among the associated objective metrics, which indicates a direction in which the local optimum is likely to be located. As more evidence is collected, the collective information leads to an estimate of the ROD and its local optimum with higher statistical confidence. Using the information on the estimated location of a local optimum, our strategy chooses the next pair of points to be compared in order to minimize the total number of comparisons to achieve a given level of accuracy.

6.1.1 Conducting subjective evaluations of a single operating curve

There are several alternatives for conducting subjective evaluations of points on an operating curve. A general approach is to divide the sequence of tests into *batches* with M tests each, ask all subjects to conduct the tests in a batch, update the estimation of the local-optimum candidates, and adaptively choose a new sequence of tests in the next batch. Because the test results in one batch are used to optimize the tests in the next batch, the tests must be synchronized so that those in the current batch are completed before beginning those in the next batch.

Sequential evaluations ($M = 1$). In one extreme, each batch consists of one pair of points to be tested. This approach results in the least number of tests before updating the estimate of the local-optimum candidate. Hence, it leads to a better choice of the next test to be conducted and a lower bound on the total number of tests. However, it also results in an upper bound on the number of batches, making it inconvenient for subjects because they have to combine their results at the end of each test before the next test can be determined.

Batch-parallel evaluations ($M > 1$). To avoid frequently synchronizing the subjects in their tests, M pairs of tests can be evaluated by all subjects in each batch, before updating the local-optimum estimate. Its disadvantage is that, when M is large, most of the test results do not provide new information on the estimate.

Fully parallel evaluations ($M \gg 1$). In the other extreme, all evaluations are conducted in a single batch. In this case, the estimate on the local optimum can only be obtained after all the subjects have completed a predefined set of comparisons. A trivial solution is to select $N \doteq \frac{A_i^{\max} - A_i^{\min}}{JND}$, which represents a finite number of operating points that are JND from each other. A complete evaluation of the $N(N-1)/2$ pairs allows us to estimate A_i^* to within JND of the actual A_i^*. This approach gives an upper bound on the number of tests. However, since $A_i^{\min}$, $A_i^{\max}$ and JND are unknown, a separate set of tests is needed to first find these values. Such tests can be as expensive as conducting tests to find the local optima.

Because sequential evaluations are more effective for reducing the total number of tests in identifying a local optimum of an operating curve, we study this method in detail in this section.

Figure 6.1 depicts the three steps in our method, which involve estimating the values of $A_i^{\min}$, $A_i^{\max}$ and JND, as well as the local optimum in a ROD.

- Step 1: Given the evidence collected, estimate the ROD of each local optimum.
- Step 2: Given the ROD of a local optimum and the evidence collected so far, estimate the local optimum.
- Step 3: Given an estimate of the local optimum, choose the next pair of points to be evaluated.

In the rest of this section, we first present the second step for estimating the local optimum in a ROD, since the first and the last steps utilize this step. Based on our proposed method, we present at the end of the section a heuristic for batch-parallel evaluations and the approach for the subjective evaluation of multiple operating curves.

6.1.2 Step 2: Finding a local optimum in a given ROD

In this section, we develop the second step of our method, using an estimate of the ROD and the previous comparison results. Our goal is to refine the estimate of the local optimum in order to get a better confidence.

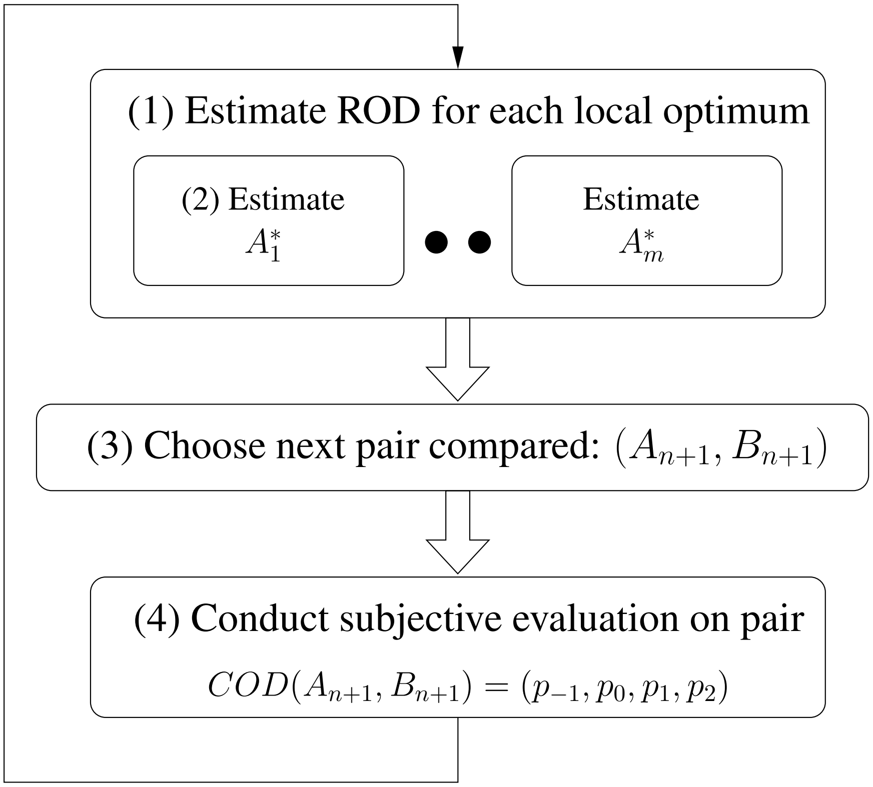

(1) Estimate ROD for each local optimum
(2) Estimate A_1^*
Estimate A_m^*
(3) Choose next pair compared: (A_{n+1}, B_{n+1})
(4) Conduct subjective evaluation on pair
$COD(A_{n+1}, B_{n+1}) = (p_{-1}, p_0, p_1, p_2)$

Figure 6.1: Method for identifying a local optimum through subjective comparisons.

The model in Section 5.2.7 for comparing points in $ROD(A_i^*)$ allows us to determine a likely direction on the location of A_i^*. However, its non-parametric nature makes it difficult to combine the result of a test with the prior information obtained. Hence, we cannot calculate the statistical likelihood on the probable locations of A_i^*.

To address this issue, we develop in this section a parametric model of subjective comparisons in $ROD(A_i^*)$ after simplifying the general model. The simple model allows a probabilistic representation of our knowledge on the location of A_i^* and a way to statistically combine the deductions from multiple comparisons. It also allows us to develop an adaptive search algorithm (Section 6.1.4) that significantly reduces the number of comparisons needed for identifying A_i^*. In addition, an estimate on the confidence of the result provides a consistent stopping condition for our algorithm. We evaluate the effect of our simplifications using Monte Carlo simulations in Section 6.2.

Our simplified parametric model on $ROD(A_i^*)$ is derived with the following assumptions.

Assumption 1 *$CND(A_i)$ and $JND(A_i)$ are constant in $ROD(A_i^*)$. Further, p_0 is linear with respect to $B - A$.*

We know intuitively and from subjective experiments that $JND(A_i)$ and $CND(A_i)$ depend on A_i and can vary in $ROD(A_i^*)$. However, the task of estimating a continuous function is as hard as estimating the optimum itself. For tractability, we make a simplification that $JND(A_i)$ and $CND(A_i)$ do not change with A_i.

Assumption 2 *The boundary line representing subjectively symmetric pairs, $A\|_s B$, is a straight line of the form $B = mA + n$ on the A-B plane, where $m = \frac{-\gamma}{\Delta-\gamma}$ and $n = \frac{\Delta}{\Delta-\gamma}A_i^*$.*

This approximation is justified because the preferred trade-offs among objective metrics are slowly changing around a point. Hence, it is reasonable within the ROD of a local optimum.

Assumption 3 *For tractability in derivations, we specify the parameters m and n of the $A\|_s B$ line as a probability distribution. By symmetry, $A_i^*\|_s A_i^*$; thus, (A_i^*, A_i^*) is on the $A\|_s B$ line. It suffices to specify another point on the $A\|_s B$ line to uniquely identify it. Since control symmetry is defined for points that satisfy $A < A_i^* < B$, the line has to pass through $B - A = \Delta$ (where $\Delta > 0$) between $(A_i^* - \Delta, A_i^*)$ and $(A_i^*, A_i^* + \Delta)$. For simplicity, we assume that the cross-over point is uniformly distributed on this line segment. The cross-over point is represented by $(A_i^* - \Delta + \gamma, A_i^* + \gamma)$, where γ is a random variable uniformly distributed in $[0, \Delta]$.*

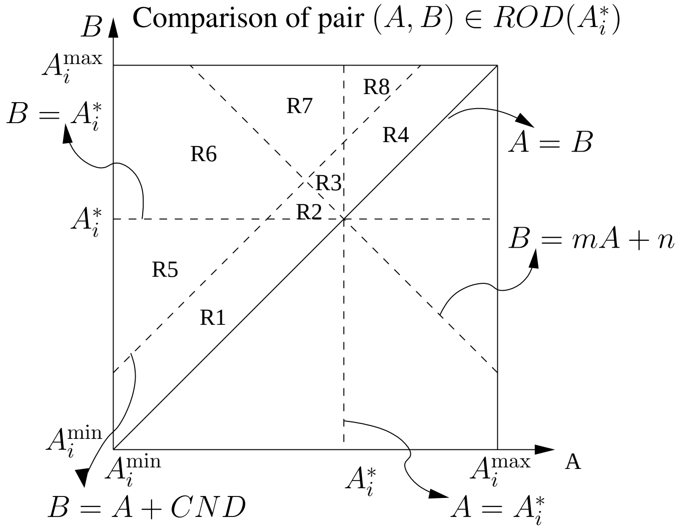
This assumption results in a piecewise linear likelihood function derived later to represent the information learned on the location of A_i^*.

Assumption 4 *In the general model, A is more preferred than B ($p_1(A,B) > p_{-1}(A,B)$) if $A < A_i^* < B$ and $B > B'$ where $A\|_s B'$. In the simplified model, we assume that $p_{-1}(A,B) = 0$ when deducing the likely direction of A_i^* after obtaining an $A >_s B$ opinion. Similarly, when $B < B'$ where $A\|_s B'$ or when subjective symmetry with respect to A does not exist, we assume $p_1 = 0$ and use this property in our derivations when an $A <_s B$ opinion is obtained.*

Our model describes the probabilities of occurrence of the four possible opinions when comparing two operating points in $ROD(A_i^*)$. For a constant $CND(A_i)$ independent of A_i, Figure 6.2 depicts the 2-D model and the eight regions with respect to the four boundary lines.

Bayesian Formulation. We assume an estimate of the ROD of a local optimum in which we are fairly certain that a local optimum exists. Since there may be evidence to suggest multiple local optima, we apply the following procedure to each ROD individually.

The information deduced on the location of A_i^* can be represented by a *belief function*, which is a probability density function (PDF) defined over the set of operating points in $ROD(A_i^*)$. It is denoted by $f_{A_i^*}(a)$ when the operating curve is continuous and by a probability mass function when



(a) The simplified parametric model

Probability	Regions in $ROD(A_i^*)$							
Densities	R1	R2	R3	R4	R5	R6	R7	R8
p_0	> 0	> 0	> 0	> 0	0	0	0	0
p_{-1}	> 0	> 0	0	0	> 0	> 0	0	0
p_1	0	0	> 0	> 0	0	0	> 0	> 0
p_2	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0

(b) COD for the simplified model (assuming $B > A$)

Figure 6.2: The eight regions corresponding to different pairwise comparisons on the A-B plane. The boundary lines separating the regions are similar to those in Table 5.4, except that the boundary line $A\|_s B$ becomes $B - mA - n$.

the curve is discrete. In the rest of this thesis, we use belief functions defined over a 1-D continuous space to represent the likelihood of each operating point being optimal. It is understood that, for a discrete operating curve, the notation can be converted by replacing PDF by its probability mass function and integration by summation.

Initial knowledge on the location of A_i^*. Before any subjective test is conducted, the location of A_i^* is assumed to be uniformly likely at any point on the operating curve. Thus, the initial belief function is

$$f^0_{A_i^*}(a) = 1; \quad a \in [A_i^{\min}, A_i^{\max}]. \tag{6.1}$$

Deductions from a single pairwise comparison. Based on the distribution of the opinions ob-

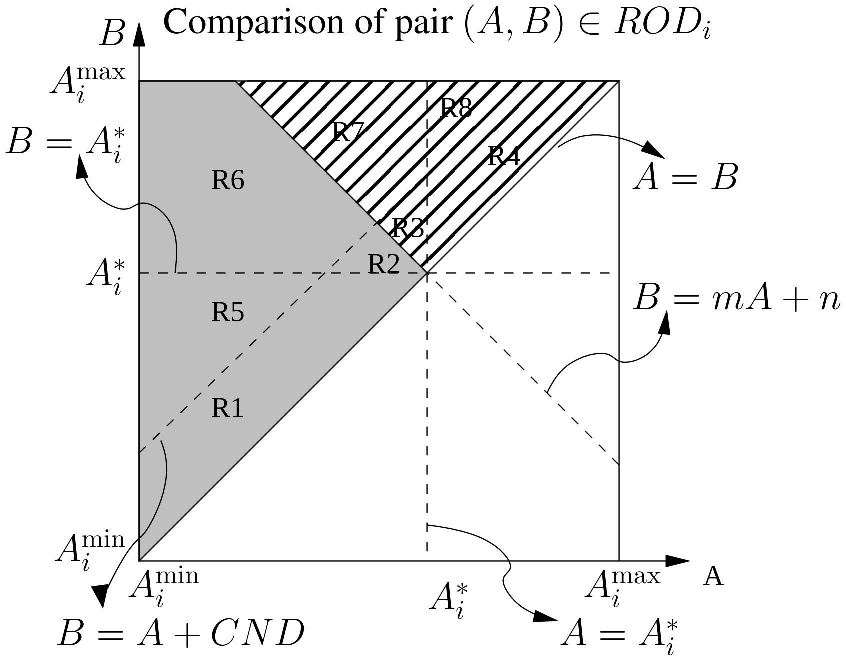


Figure 6.3: Deductions on which regions the (A, B) pair may be located, based on the $A >_s B$ and $A <_s B$ responses. Cross-shaded regions indicate that the subject perceives A_i^* to be in $[A + \gamma, A_i^{\max}]$. Solid-shaded regions indicate that the subject perceives A_i^* to be in $[A_i^{\min}, A + \gamma]$.

tained by comparing A and B, we can improve our knowledge of the location of A_i^* by a Bayesian formulation. The analysis allows us to obtain the posterior probability from the prior probability and the new evidence.

$$f_{A_i^*}(a|COD(A, B)) = \overline{p}) = \frac{L(a|COD(A, B) = \overline{p}) \times f_{A_i^*}(a)}{\int_0^1 L(\eta|COD(A, B) = \overline{p}) \times f_{A_i^*}(\eta) d\eta}. \tag{6.2}$$

The formulation requires the prior belief function on the location of A_i^* and the likelihood function $L(a|\overline{p})$. Before deriving the likelihood function, we first show the deductions on the subjects' responses.

Based on Assumptions 2 and 3, the $A \|_s B$ line satisfies $B = mA + n = \frac{-\gamma}{\Delta - \gamma} A + \frac{\Delta}{\Delta - \gamma} A^*$, where $B - A = \Delta$ and γ is uniform in $[0, \Delta]$. Next, we analyze the deductions of the four responses.

a) Implications of $A >_s B$. If a subject prefers A over B, this means $A_i^* \notin [A + \gamma, A_i^{\max}]$, since $p_1 = 0$ in Regions 1, 2, 5, and 6 (Figure 6.2b). Thus, $A_i^* \in [A_i^{\min}, A + \gamma]$ (solid-shaded regions in Figure 6.3).

b) Implications of $A <_s B$. If a subject prefers B over A, this means $A_i^* \notin [A_i^{\min}, A + \gamma]$, since $p_{-1} = 0$ in Regions 3, 4, 7, and 8 (Figure 6.2b). Thus, $A_i^* \in [A + \gamma, A_i^{\max}]$ (cross-shaded regions).

c) Implications of $A \approx_s B$. If a subject indicates that A is about the same as B, A_i^* can be in any of Regions 1, 2, 3, and 4. This does not provide any information on the location of A_i^*, and $A_i^* \in [A_i^{\min}, A_i^{\max}]$.

d) Implications of $A ?_s B$. If a subject indicates that A is incomparable to B, A_i^* can be in any of the 8 regions. This does not provide any information on the location of A_i^*, and $A_i^* \in [A_i^{\min}, A_i^{\max}]$.

The *likelihood function* $L(a|\overline{p})$ is a function of $a \in [A_i^{\min}, A_i^{\max}]$ and indicates the likelihood of obtaining $\overline{p}$ as a result of subjective comparison of A and B if $A_i^* = a$. Using Axiom 2, the likelihood of a being the optimum can be evaluated by the occurrence frequencies of the four outcomes analyzed above. Conditioned on the value of γ and the result of the subjective comparison, we can represent the likelihood function as

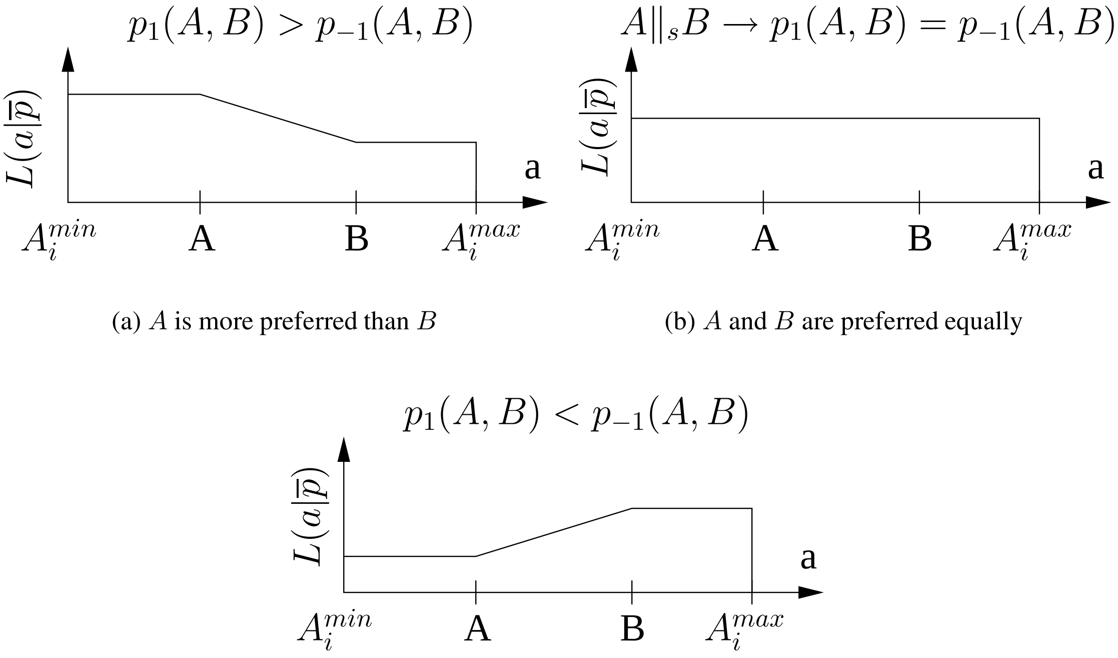
$$L(a|\overline{p}, \gamma) = \begin{cases} p_1 + p_0 + p_2 & \text{if } A_i^{\min} < a < A + \gamma \\ p_{-1} + p_0 + p_2 & \text{if } A + \gamma < a < A_i^{\max}. \end{cases} \tag{6.3}$$

Since γ is uniformly distributed over $[0, \Delta]$, where $\Delta = B - A$, the expectation taken over γ results in a likelihood function that is only conditioned on $COD(A, B) = \overline{p}$, the result of the subjective evaluation. $L(a|\overline{p})$ is defined as

$$\begin{aligned} L(a|\overline{p}) \; &= \; E_\gamma[L(a|\overline{p}, \gamma)] \; = \; \textstyle\int_0^\Delta L(a|\overline{p}, \gamma) Pr(\gamma) d\gamma \\ &= \begin{cases} p_0 + p_2 + p_1 & \text{if } A_i^{\min} < a < A \\ p_0 + p_2 + \frac{p_1(B-a) + p_{-1}(a-A)}{B-A} & \text{if } A \le a \le B \\ p_0 + p_2 + p_{-1} & \text{if } B < a < A_i^{\max}. \end{cases} \end{aligned} \tag{6.4}$$

Figure 6.4 depicts the three possible cases of the likelihood function defined in (6.4) for a subjective comparison.

Deductions on subsequent evaluations. The belief function (posterior density) obtained from the Bayesian formulation can be used as the prior knowledge in a subsequent application of the formulation. We assume that the COD results from comparing different pairs are independent in terms of the information on the location of A_i^*.



(a) A is more preferred than B

(b) A and B are preferred equally

(c) B is more preferred than A

Figure 6.4: Likelihood functions based on the subjective comparison of A and B.

For $a \in [A_i^{\max}, A_i^{\max}]$, the combined belief function after the n^{th}, $n \geq 1$, comparison is

$$\begin{aligned} f_{A_i^*}^n(a) &= \frac{f_{A_i^*}^{n-1}(a) \times L(a|COD(A_n,B_n)=\overline{p})}{\int_{A_i^{\min}}^{A_i^{\max}} f_{A_i^*}^{n-1}(\eta) \times L(\eta|COD(A_n,B_n)=\overline{p}) d\eta} \\ &= \frac{\prod_{i=1}^{n} L(a|COD(A_n,B_n)=\overline{p})}{\int_{A_i^{\min}}^{A_i^{\max}} \prod_{i=1}^{n} L(\eta|COD(A_n,B_n)=\overline{p}) d\eta}. \end{aligned} \tag{6.5}$$

The combination process is associative, meaning that the order of the combination does not affect the combined belief function. Further, based on the independence property, the combined belief function found by cascading the Bayesian formulation can be written in a closed form as is shown in the above equation.

Utility. The aim of the subjective tests is to obtain $\widehat{A_i^*}$, an estimate of A_i^*, with high confidence. Thus, the utility of a belief function is the confidence or the probability that $\widehat{A_i^*}$ is in $JND(A_i^*)$. The estimation error of less than $JND(A_i^*)$ is insignificant, since any point in $JND(A_i^*)$ is indistinguishable to A_i^*. Given f, $\widehat{A_i^*}$ is defined to be the point that maximizes the probability of a

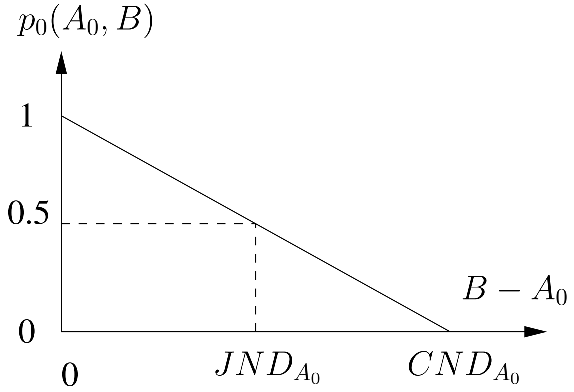


Figure 6.5: Percentage of subjects not perceiving any difference decreases linearly as a function of B minus A.

successful estimation:

$$\widehat{A_i^*}(f) \quad = \quad \arg\max_a \left\{ \int_{a-JND/2}^{a+JND/2} f(\xi) d\xi \right\}. \tag{6.6}$$

Given f and $\widehat{A_i^*}$, the utility is defined as

$$U(f) = Pr(|\widehat{A_i^*} - A_i^*| \leq JND) = \int_{\widehat{A_i^*}-JND/2}^{\widehat{A_i^*}+JND/2} f(\xi) d\xi. \tag{6.7}$$

However, since JND can only be estimated using the results of the comparisons already conducted, its estimation has an inherent error. The calculation of this error and the confidence bounds are discussed later in this section. Using the confidence bounds on the estimated JND (which increases as a function of the number of tests conducted), we calculate the confidence bounds for $\widehat{A_i^*}$.

Estimation of JND and CND. As is discussed above, JND is needed in (6.7) when evaluating the confidence of $\widehat{A_i^*}$. The estimated CND is also used in Step 3 when choosing the next pair of points to be compared. To this end, we use the p_0 and $B - A$ values of those previously conducted subjective results for estimating CND and JND. Figure 6.5 depicts the linear relation between p_0 and $B - A$, according to Assumption 1 of our simplified model. The model further assumes that JND and CND (which is twice of JND) are constant and do not vary with A:

$$p_0 = 1 - \frac{B - A}{CND}. \tag{6.8}$$

For K subjects in the evaluation, let $\widehat{p_0}$ be the empirical distribution of a binomial random variable with p_0:

$$Pr(\widehat{p_0} = x) = \binom{K}{Kx} (p_0)^{Kx} (1 - p_0)^{K(1-x)}. \tag{6.9}$$

Given the model that assumes a linear function of p_0 with respect to $B - A$ and that passes through $(B-A, p_0) = (0, 1)$ and $(CND, 0)$, CND is the only unknown to uniquely specify the line. Based on a single comparison, using the empirical $\widehat{p_0}$ value and $B - A$, the estimated CND is

$$\widehat{CND} = \frac{B - A}{1 - \widehat{p_0}}. \tag{6.10}$$

The error in $\widehat{CND}$ is due to variations in the empirical distribution $\widehat{p_0}$ around the actual p_0. It can be calculated using the binomial distribution and the likelihood that $\widehat{p_0}$ can be obtained when p_0 is equal to a particular value.

$$L(p_0 = y | \widehat{p_0} = x) = \binom{K}{Kx} y^{Kx} (1 - y)^{K(1-x)}. \tag{6.11}$$

Furthermore, using (6.10), the likelihood can be defined for cases when CND is equal to a particular value.

$$L(CND = z \mid \widehat{p_0} = x, B - A = \Delta) = \frac{\Delta}{z^2} \binom{K}{Kx} \left(1 - \frac{\Delta}{z}\right)^{Kx} \left(\frac{\Delta}{z}\right)^{K(1-x)}. \tag{6.12}$$

Similar to the estimation of the belief function in Section 6.1.2, $\widehat{CND}$ can be obtained by a Bayesian formulation. Since the actual p_0 needs to be in $[0, 1]$, we normalize the belief function as follows:

$$Pr^{post}(CND = z) = \frac{Pr^{prior}(CND = z) \times L(CND = z | \widehat{p_0} = x, B - A = \Delta)}{\int_0^1 Pr^{prior}(CND = \eta) \times L(CND = \eta | \widehat{p_0} = x, B - A = \Delta) d\eta}. \tag{6.13}$$

The distribution of $\widehat{CND}$ can then be calculated using (6.10), whose confidence improves with the number of comparisons conducted. Once the distribution is obtained, the 90 percentile confidence intervals are calculated and used in estimating A_i^* and its utility in (6.7).

Initially, no evidence suggests that there will be more than one local optimum. Since the optimum is equally likely to be anywhere on the operating curve, we choose the initial $\widehat{A_i^*}$ to be 0.5. Further, the initial $\widehat{CND}$ is arbitrarily chosen to be 0.1.

6.1.3 Step 1: Estimating the ROD of an unknown local optimum

In this section, we develop a method for estimating the boundaries of the ROD for one or more local optima. An *evidence* of comparing A and B is denoted by e and consists of the tuple $(A, B, COD(A, B))$. An *evidence set* is denoted by $\mathcal{E}$ and is a collection of evidences obtained by subjective evaluations. The *complete evidence set*, containing the results of all the past n comparisons conducted so far, is denoted by $\mathcal{E}_{all}$:

$$\mathcal{E}_{all} = \{\underbrace{(A_1, B_1, COD(A_1, B_1))}_{e_1}, \ldots, \underbrace{(A_n, B_n, COD(A_n, B_n))}_{e_n}\}. \tag{6.14}$$

As is discussed above, there may be multiple local optima on an operating curve, where each local optimum dominates over its corresponding ROD. These multiple RODs on an operating curve, if they exist, do not overlap with each other (Lemma 5). Further, only evidence for which both comparison points are within the ROD of a local optimum provides a reliable direction on the location of that local optimum (Section 5.2.7).

In case of multiple local optima on an operating curve, the result obtained by comparing a pair of points in one ROD cannot be combined with that of comparing a pair in another ROD. As a result, when comparing points on an operating curve with multiple local optima, some of the evidences would give conflicting (or inconsistent) directions on the location of the local optima. If this situation cannot be explained by the noise in the finite number of subjective tests, then it indicates the existence of multiple local optima (and multiple ROD regions), where one evidence belongs to one ROD and another to the other ROD.

Let $\mathcal{E}_i \subseteq \mathcal{E}_{all}$ be the subset of evidences that correspond to $ROD(A_i^*)$. Based on the possibly inconsistent evidences found, we discriminate them into different subsets that correspond to different local-optimum candidates.

Definition 8 Inconsistent evidence. *For $\widehat{A_i^*} < A$, an evidence is inconsistent if the hypothesis $\{H_0 : p_1 \geq p_{-1}\}$ can be rejected with some statistical significance. Similarly, for $B < \widehat{A_i^*}$, an evidence is* inconsistent *if the hypothesis $\{H_0 : p_{-1} \geq p_1\}$ can be rejected with the same statistical significance.*

When two sets of evidences are inconsistent, it means that each is pointing to a different local optimum. As a result, the operating curve should be divided into two RODs, each corresponding to one local optimum.

Procedure for identifying multiple RODs on an operating curve. This consists of three steps.

a) Initially, all evidences that are mutually consistent with other evidences in the set are used to determine a ROD. Although this step provides a superset of the actual ROD region that can potentially overlap with each other, it ensures that the RODs are not over-pruned due to noisy evidences. This step is described in detail as follows.

Initial condition. After the first comparison, since it has no inconsistent evidence, we assume that there is one local optimum, and its ROD is the entire operating curve. As subsequent comparisons are done, we determine whether the new evidence on the current estimate of the local optimum is consistent with existing evidences.

Existence of multiple ROD regions. If the new evidence is found to be inconsistent, a new set of evidences is formed, say $\mathcal{E}_2$, that corresponds to a new local-optimum candidate. The new evidence will be taken from $\mathcal{E}_1$ and placed in $\mathcal{E}_2$. Further, all evidences that are consistent with the new evidence will be duplicated from existing sets to $\mathcal{E}_2$. The procedure will be repeated for all existing sets until each set has evidences that are mutually consistent with each other. The procedure results in the largest set of mutually consistent evidences in each set.

Initial ROD estimation. Next, we map each evidence set to its corresponding ROD. We identify the minimum and the maximum of the points compared in each evidence set in order to determine its bounds.

$$\begin{aligned} \widehat{A_i^{\min}} &= \min\{A_j \mid (A_j, \bullet, \bullet) \in \mathcal{E}_i\}; & (6.15) \\ \widehat{A_i^{\max}} &= \max\{B_j \mid (\bullet, B_j, \bullet) \in \mathcal{E}_i\}. & (6.16) \end{aligned}$$

We repeat the estimation of the ROD for each local-optimum candidate. At this point, the RODs estimated may overlap, since some evidences can be members of multiple sets. As RODs do not overlap (Lemma 5), it is necessary to update the initial RODs estimated in order to arrive to non-overlapping RODs. Although we can simply construct evidence sets that do not overlap, this condition is not sufficient. For example, one of the evidences in the first set can have one of its points compared in the ROD of the second set, which causes the two RODs to overlap.

b) In the second step, the local optima are estimated based on the initial RODs found. It uses the set of evidences for which both points compared are within a single ROD to obtain a belief function and a corresponding estimate of a local optimum. To obtain a subset of the initial ROD estimates that do not overlap, we first estimate the belief functions of different local-optimum candidates individually over their possibly overlapping RODs. We then estimate the local optimum using the procedure in Section 6.1.2.

c) Lastly, the RODs are refined again to ensure that each is a contiguous region with a local optimum and that they do not overlap with each other.

Updated estimation of RODs. Starting from the local optimum estimate, the corresponding ROD is a contiguous region until an inconsistent evidence is found. This step ensures that the RODs are non-overlapping. In case there is no inconsistent evidence, the entire operating curve is taken to be a single ROD:

$$\widehat{A_i^{\min}} = \max\left\{ \min\{A_j \mid (A_j, \bullet, \bullet) \in \mathcal{E}_i\},\ \max\{B_j \mid (\bullet, B_j, \bullet) \notin \mathcal{E}_i\} \right\} \quad (6.17)$$

$$\widehat{A_i^{\max}} = \min\left\{ \max\{B_j \mid (\bullet, B_j, \bullet) \in \mathcal{E}_i\},\ \min\{A_j \mid (A_j, \bullet, \bullet) \notin \mathcal{E}_i\} \right\}. \quad (6.18)$$

Eliminating noisy evidence and merging RODs. Due to noise in the subjective evaluations, it is possible to have inversions of preference directions in some comparisons (such as $p_{1} > p_{-1}$ instead of $p_{1} < p_{-1}$). This may cause the locations of the inconsistent evidences to interleave with each other and result in overlapping RODs. In this case, most of the evidences would point to one direction, whereas a few in the same vicinity would point to another. Eliminating such noisy evidences requires merging the divided ROD regions into one contiguous region. This task can be achieved by increasing the statistical significance level when identifying inconsistent evidence pairs.

Once the RODs are estimated, the information is passed to Step 2 (Section 6.1.2), and a local optimum is identified in one of the RODs.

6.1.4 Step 3: Identifying the next pair of points to be compared

Based on the procedures in Sections 6.1.2 and 6.1.3, we present our search algorithm for choosing the next pair of points to be compared. Our goal is to minimize the number of comparisons before A_i^* is identified with high confidence. We first describe our observations on the optimal sequence of comparisons and reduce the problem to choosing the optimal pair in each step. We then derive the optimal pair of points to be compared in the next step.

Sequence of comparisons. As more pairwise evaluations are conducted, the combined belief function evolves from uniform to a shape that is centered around $\widehat{A_i^*}$. Since it is not feasible to exactly identify A_i^* in a continuous search space, we stop the search once a certain level of confidence is reached. The confidence level chosen will affect the efficiency of the algorithm and the accuracy of the result.

At the beginning of the n^{th} comparison, when given the utility $U(f^{n-1})$, the expected number of comparisons left to reach the stopping condition if the optimal pair is chosen in the n^{th} test is denoted by

$$\begin{aligned} S(U(f^{n-1})) &= 1 + S(U(f^n)) \\ &= 1 + \min_{A_n, B_n} S(U(f^n \mid A_n, B_n)). \end{aligned} \tag{6.19}$$

The following are the arguments leading to the evaluation of $\min_{A_n,B_n} S(U(f^n \mid A_n, B_n))$. For any A and B, $L(a \mid A, B)$ is uni-modal. Let $mode(L)$ be the set of points satisfying the modality. It is clear that $A_i^* \in mode(L(a \mid A, B))$. Since any comparison conducted over the same ROD is consistent and A_i^* is common to all the comparisons, it is clear that $mode(L(a \mid A_1, B_1)) \cap mode(L(a \mid A_2, B_2)) \neq \emptyset$. For any sequence of A-B pairs, $f^n(a)$ is uni-modal and $A_i^* \in mode(f^n)$. Hence, $U(f^n)$ is monotonically non-decreasing with respect to n for any sequence of comparisons, and $S(U)$ is a non-increasing function of U. Thus, minimizing the expected number of steps left is equivalent to maximizing the expected utility of the current belief function.

Individual comparisons. Based on Section 6.1.2, when both points compared are in a ROD of a local-optimum candidate, their subjective comparison would provide a correct direction on the location of A_i^* with respect to the points compared. However, if one or both points are not in the same ROD or they are in RODs of different local optima, then we cannot guarantee that a correct direction can be found. The comparisons that do not point to the intended local optimum introduce inconsistencies and reduce the confidence of the A_i^* estimated. As is described in Section 6.1.3, such inconsistencies are eliminated from the set of evidences during the estimation of a particular local optimum. Of course, such comparisons are wasted and do not improve our knowledge of A_i^*. In short, based on the updated estimation of each ROD, we should identify points to be compared that are in the same ROD where the local optimum is to be located.

The following are the observations for identifying the optimal pair of points to be compared next.

1. Indistinguishable and incomparable opinions do not lead to any deductions on A_i^*. Further, when $p_1 = p_{-1}$, the two points compared are subjectively symmetric and do not lead to new information on A_i^*. Hence, for a comparison to be useful, p_0 and p_2 should be small and the difference between p_1 and p_{-1} should be large. Depending on which of p_1 or p_{-1} is larger, the likely direction of the search can then be determined.

2. Due to Axiom 4, p_0 is monotonically non-increasing with $B - A$ and reaches 0 at $B - A = CND(A)$. Hence choosing points that are very close to each other does not provide any evidence on A_i^* because p_0 is high and thus the difference between p_1 and p_{-1} is low. In

contrast, choosing A far away from B reduces p_0, although choosing them beyond $B - A = CND$ does not reduce p_0 further (since it is already equal to zero).

3. Due to Axiom 5, $p_2 = 0$ at $B - A = 0$ and is monotonically non-decreasing with respect to $B - A$ and reaches its maximum at $B - A = A^{\max} - A^{\min}$. Thus, choosing two points that are far apart increases p_2, which indirectly reduces the difference between p_1 and p_{-1} and the conclusiveness of the comparison.

4. Given an unknown A_i^* and any pair A and B, p_1 and p_{-1} are uncorrelated. Thus, maximizing $p_1 + p_{-1}$ (which minimizes $p_0 + p_2$) is equivalent to maximizing the expected value of $|p_1 - p_{-1}|$. Note that $\arg\min\{p_0 + p_2\}$ is achieved at $B - A = CND$.

5. The utility is maximized when the disparity between the two horizontal levels of the likelihood function in Figure 6.4 (or the first and last cases in (6.4) and represented by $|p_1 - p_{-1}|$) is maximized. Due to Axiom 6, the difference between p_1 and p_{-1} increases when one of the points is close to or equal to A_i^*.

6. Thus, given $\widehat{A_i^*}$ and $\widehat{CND}$ after the n^{th} comparison, the optimal choice of the $n + 1^{\text{st}}$ comparison should include $\widehat{A_i^*}$ as one of the points and the other CND away from it in either direction on the operating curve.

$$(A_{n+1}, B_{n+1}) = \begin{cases} (\widehat{A_i^*} - \widehat{CND}, \widehat{A_i^*}) & \text{if } n \text{ is odd.} \\ (\widehat{A_i^*}, \widehat{A_i^*} + \widehat{CND}) & \text{if } n \text{ is even.} \end{cases} \tag{6.20}$$

   Note that since A and $B \in \mathcal{O}_i$, the selection made by (6.20) needs to be augmented to keep A_{n+1} and B_{n+1} within their corresponding RODs.

As more comparisons are conducted, the estimated local optimum $\widehat{A_i^*}$ improves, which increases the disparity between p_1 and p_{-1} (Axiom 6). Since the region with a higher likelihood contains $\widehat{A_i^*}$, the corresponding utility increases as well. Note that utility improves at a rate related to n.

Example 1 (cont'd). *Based on the method in this section, we have conducted a sequence of subjective comparisons between pairs of points on the operating curve in Figure 5.1b. Table 6.1 shows the objective metrics of each pair of points compared and the COD of the subjective comparisons of the four comparisons made.*

As described above, the belief function, $\widehat{A^*}$, and $\widehat{CND}$ are updated based on the latest result after each comparison. Initially, A^* is equally likely to be anywhere on the operating curve. Thus, the initial $\widehat{A^*}$ is 500 ms (or 0.5), and the initial $\widehat{CND}$ is 0.25. Since there is no inconsistent

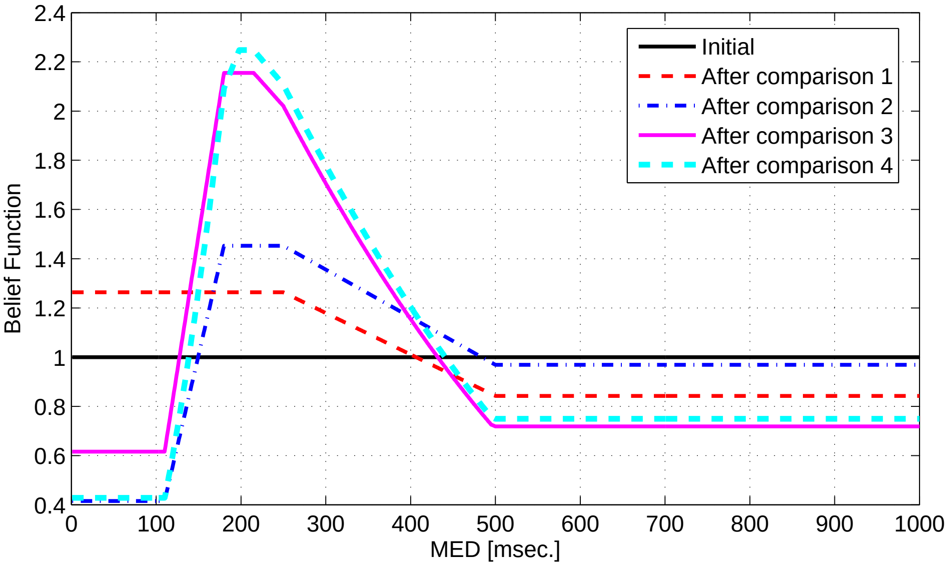


Figure 6.6: Initial belief function and its evolution after each subjective comparison.

Table 6.1: COD of the subjective comparisons conducted on pairs of points on the operating curve in Figure 5.1.

	A				B				COD(A,B)			
#	MED	PESQ	CS	CE	MED	PESQ	CS	CE	p_{-1}	p_0	p_1	p_2
1	250	4.18	1.60	0.91	500	4.18	2.21	0.84	0.125	0.500	0.375	0
2	110	2.39	1.27	0.96	180	4.18	1.44	0.93	0.750	0.250	0.000	0
3	215	4.18	1.52	0.92	539	4.18	2.30	0.82	0.125	0.625	0.250	0
4	111	2.60	1.27	0.96	198	4.18	1.48	0.93	0.750	0.125	0.125	0

evidence, ROD is equal to the entire operating curve, which happens to be valid throughout the comparisons for this example. After the first comparison, the belief function (Figure 6.6) indicates that A^* is more likely to be less than 250 ms.

However, for this operating curve, any MED less than 110 ms results in PESQ less than 2.0. As our previous experience shows that such conversations are not likely to be preferred over conversations with higher PESQ, we prune the operating curve below 110 ms in order to avoid unnecessary comparisons with conversations of inferior quality. Thus, after the first comparison, our $\widehat{A^*}$ is 180 ms (midway between 110 ms and 250 ms).

The second comparison selects $A = 0.110$ and $B = 0.180$, based on (6.20), since any A less than 0.110 would not provide information on the optimal point. The result indicates that $B = 0.180$ is strongly preferred over A. The updated belief function (Figure 6.6) leads to the selection of the *third pair compared as* $(A, B) = (0.215, 0.539)$*, since* $\widehat{A^*} = 0.215$ *(midway between 0.180 and 0.250) and* $\widehat{CND} = 0.324$*.*

In the third comparison, subjects slightly prefer the 215-ms alternative over the 539-ms alternative, since there is no difference in speech quality and the difference in degradation due to delay is not very perceptible (with a low switching frequency on this conversation).

Based on the previous results, the fourth comparison is selected to be $(0.111, 0.198)$*, and subjects prefer the 198-ms alternative significantly over the 111-ms alternative due to significant improvement in LOSQ.*

After four comparisons, $\widehat{A^*}$ *is 208 ms and the utility is 64%. The operating point identified has very high LOSQ and little delay degradation. We observe in Figure 6.6 that the belief function evolves with each comparison and centers around* $\widehat{A^*}$*.*

6.1.5 Batch-parallel evaluations

For batch-parallel evaluations ($M > 1$), the derivation of the optimal sequence of pairs is intractable, since $2M$ variables have to be optimized simultaneously. Further, a numeric solution is too expensive when the number of operating points or M is large. Thus, we use a heuristic to find the set of pairs compared in the next batch, based on the current belief function. We identify $M-1$ equally spaced points $\{C^j, j = 1, \ldots, M-1\}$ in the search space for which one of them is $\widehat{A^*_i}$.

$$C^j \quad = \quad \bmod\left(\frac{j-1}{M-1} + \widehat{A^*_i}, 1\right), \; j = 1, \ldots, M-1. \tag{6.21}$$

For equal spacing, points are wrapped around the operating curve via the modulo operation. We conduct two comparisons involving $\widehat{A^*_i}$, with points CND away from it in either direction, which correspond to the optimal pair for the even and odd cases in (6.20). For each of the remaining $M-2$ points identified, it is compared with a point CND away in the direction opposite to that of $\widehat{A^*_i}$:

$$(A^j_n, B^j_n) \quad = \quad \begin{cases} (C^j - \widehat{CND}, C^j) & \text{if } C^j < \widehat{A^*_i} \\ (C^j, C^j + \widehat{CND}) & \text{if } C^j > \widehat{A^*_i} \\ \text{Both pairs above} & \text{if } C^j = \widehat{A^*_i}. \end{cases} \tag{6.22}$$

As expected, as M increases, the information obtained in each batch also increases. However, for large M, most of the comparisons convey the same or very similar information. Only the comparisons that are close to the optimal point are beneficial in shaping the belief function significantly. Thus, the marginal benefit obtained by increasing M diminishes as M increases. Even though the number of batches needed may decrease, the total number of comparisons conducted increases to arrive to the same level of significance. The performances of batch-parallel tests are evaluated later in this chapter in Section 6.2.

6.1.6 Putting everything together: Conducting subjective evaluations of multiple operating curves

Recall that the goal of our subjective tests is to identify the most preferred point for each of a set of operating curves that model a comprehensive set of operating conditions in a multimedia system.

In this section we present how the subjective tests conducted over a single operating curve under a given set of conditions fit into the overall design of adaptive control algorithms that work under all conditions observable at run-rime. Secondly, we present some practical issues that are associated with conducting subjective tests over multiple operating curves corresponding to different conditions.

Design of adaptive control algorithms. To design control algorithms that work well under a variety of conditions that can be observed at run-time, collecting subjective preference information under a representative set of conditions is needed. The generalizability of the algorithm depends on how representative the conditions tested are to the unseen conditions that can arise at run-time. However, the set of conditions tested needs to be finite in order for the subjective tests to be feasible.

The results of the subjective tests lead to a mapping between the objective metrics representing the two alternatives on an operating curve and the pair-wise subjective preference among those alternatives. We then learn these mappings for the multitude of operating curves using a pairwise-preference SVM classifier. We utilize this classifier and the Bayesian formulation described above to combine individual preferences in order to identify the appropriate control at run time in response to an unseen operating condition.

Our approach is to learn a Support Vector Machine (SVM) classifier to use the results of the subjective tests and the conditions under which the tests are conducted as training data. At run-time, the parameters representing the current conditions are estimated and inputted to the SVM. For example, for the two-party VoIP POS design, loss, delay and jitter parameter can be used to represent network conditions, and switching frequency and singe talk duration parameters can be used for representing conversational conditions.

Since the relations between the multiple objective metrics, the control value and the conditions are deterministic and known, based on the currently observed conditions, the points on the operating curve corresponding to each of the control values can be estimated.

Later, two points on the operating curve are chosen and inputted to the SVM, which leads to the estimation of the user preferences based on the patterns learned via off-line subjective tests. It is analogous to asking the subjects to compare the two points at run-time. However, since the SVM is already trained off-line, many pairs of points can be compared quickly, within the time constraints of making a control decision. The pairwise comparisons can then be used to identify the optimal control value to be used under the currently observed conditions.

The control decision can be updated as frequently as needed by the application. For two-party VoIP systems, the MED value can be updated at the beginning of every talk-spurt.

Also recall from Section 6.1.1 that sequential evaluations of a single operating curve are the most effective in terms of minimizing the number of tests performed for that curve, when identifying a local optimum to within some statistical confidence. However, they are inconvenient because subjects have to synchronize their test results with each other in order to estimate the local-optimum candidate before the next test can be carried out.

The major inefficiency in conducting subjective tests is due to the synchronization of subjects before the local optima estimate can be updated and the new comparison pairs are generated. Secondly, as more comparisons are conducted on a single operating curve, the estimation of the local optima and CND converges (changes slowly); thus, the comparison points that are chosen based on the estimates do not change much from one comparison to the next. Furthermore, the network and communication scenario on a single operating curve is fixed.

Due to the two reasons mentioned, the subjects may anticipate the result of a comparison, before listening/viewing to the entire communication for both pairs carefully. For example if the subject consistently perceives that there is a degradation at a particular part in the conversation, then the subject may only concentrate on that part and not perceive subtle differences in other parts.

Another example is that if the second point consistently corresponds to the control with a higher value (e.g. higher MED), than the subject can anticipate this and respond without listening carefully, based on some preconceived notions of which alternative has better quality. All these biases cause noise in the COD, and require a higher number of subjects to conduct the tests to arrive at the same statistical confidence in our estimations.

Based on these observations, the optimal strategy to minimize the total number of subjective tests for a set of operating curves is to test each curve sequentially and all the curves in parallel. In this approach, each subject is presented with a set of operating points to be compared, one from each operating curve to be tested. The tests in each set can be performed in any order and independent of other subjects because the result of comparisons from one operating curve does not depend on that of another curve. At the end of the tests, the results from all the subjects are combined in order to generate a local optimum estimate and identify the next pair of operating points to be compared for each of the operating curves. As the number of operating curves to be tested is large, this approach allows subjects to independently carry out a batch of independent tests, without having to synchronize their results in a locked-step fashion with other subjects. The number of iterations is bounded by the typically small number of iterations to identify a local-optimum candidate of an operating curve.

6.2 Performance Analysis by Monte Carlo Simulations

In evaluating our approach, we use Monte Carlo simulations to generate the probabilities of the four opinions in Table 5.1 for our general model in Figure 5.7. Since these probabilities are functions of A and B, each will exist as a surface in the A-B plane. We then apply our search algorithm in Section 6.1, which is initialized by $\widehat{A_i^*} = 0.5$ and $\widehat{CND} = 0.1$ (Section 6.1.2). Based on the A and B selected and a multi-nomial distribution, we generate the corresponding sample $COD(A, B)$ for K subjects. We then update the estimates of the ROD and the local optimum using our Bayesian procedures and repeat the search until a local optimum is found. By verifying the accuracy of our estimate with respect to the local optimum in the reference general model, we can verify the robustness of our simplified parametric model used in deriving the algorithm.

Since the generation of the general model in Figure 5.6 is rather involved, we summarize its details as follows and leave the detailed description to Appendix B. Given the number of local optima on an operating curve, we first randomly determine the boundaries of each ROD and the position of the local optimum in it. We then generate the CND line as a continuous random walk around a given average CND value. Similarly, we generate the subjective symmetry line as a continuous random walk, when given the standard deviation of the subjective symmetry line with respect to a straight line. Finally, we generate the p_i values for a finite number of A-B pairs, specifically, 100 steps in A and 100 steps in B, and using cubic interpolations in between. In particular, p_2 is monotonically non-decreasing with $B - A$ ($p_2^{\max}$ at $B - A = A^{\max} - A^{\min}$ to 0 at $B = A$); and p_0 is monotonically non-increasing with $B - A$ (1 at $B = A$ to 0 at $B - A = CND(A)$). Since $p_1 > p_{-1}$ when A and B are on the same side of A_i^* and within $ROD(A_i^*)$, we set p_1 to be proportional to $|B - A_i^*|/(|B - A_i^*| + |A - A_i^*|)$ and p_{-1} proportional to $|A - A_i^*|/(|B - A_i^*| + |A - A_i^*|)$. We also normalize p_1 and p_{-1} so that their sum is $1 - (p_0 + p_2)$. A similar procedure is applied when A and B are on different sides of A_i^*. Figure 6.7 illustrates the boundary lines with two local optima generated.

We compare our method with two other procedures, one randomly choosing the next pair on an

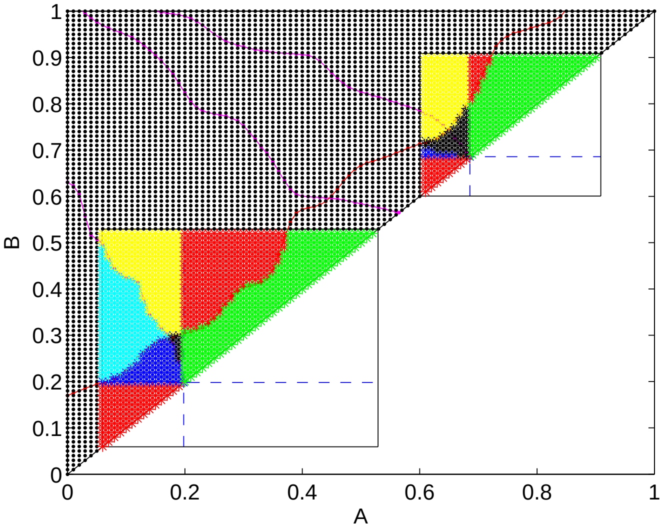


Figure 6.7: An example of the regions for two local optima on an operating curve.

Table 6.2: Expected number of comparisons for an operating curve with a single local optimum when JND is known. All comparisons result in a successful estimation of A^* (to within the JND of the actual value).

Algorithm	Absolute JND			
	0.1	0.03	0.01	0.003
1. Fully Parallel	45	≈ 500	≈ 5000	≈ 50000
2. Random (any M)	31.1	192	> 300	> 300
3. Sequential ($M = 1$)	6.4	9.9	18.3	49.6
4. Batch-Parallel ($M = 2$)	6.7	11.3	21.4	56.5
Batch-Parallel ($M = 3$)	9.6	15.6	30.4	78.7
Batch-Parallel ($M = 4$)	14.0	19.6	34.2	81.2

operating curve and the other based on the fully parallel approach that chooses $N(N-1)/2$ pairs of points (Section 6.1.1). Assuming a known JND, Table 6.2 compares the average performance of the four algorithms. It shows that conducting fully parallel evaluations and random comparisons is very expensive, and that choosing pairs sequentially based on our procedure reduces the number of comparisons by fivefold for simpler problems ($JND = 0.1$) and by 1000-fold for harder problems ($JND = 0.003$). Figure 6.8 illustrates the same results on a plot using logarithmic scale. As we show in Chapter 7, a typical absolute JND value for POS design problem is around 0.1 (or 100 ms), thus in that case the reduction in the number of comparisons needed to identify A^* is around 7-fold.

Also shown are the results of batch-parallel comparisons, which give the trade-offs between the

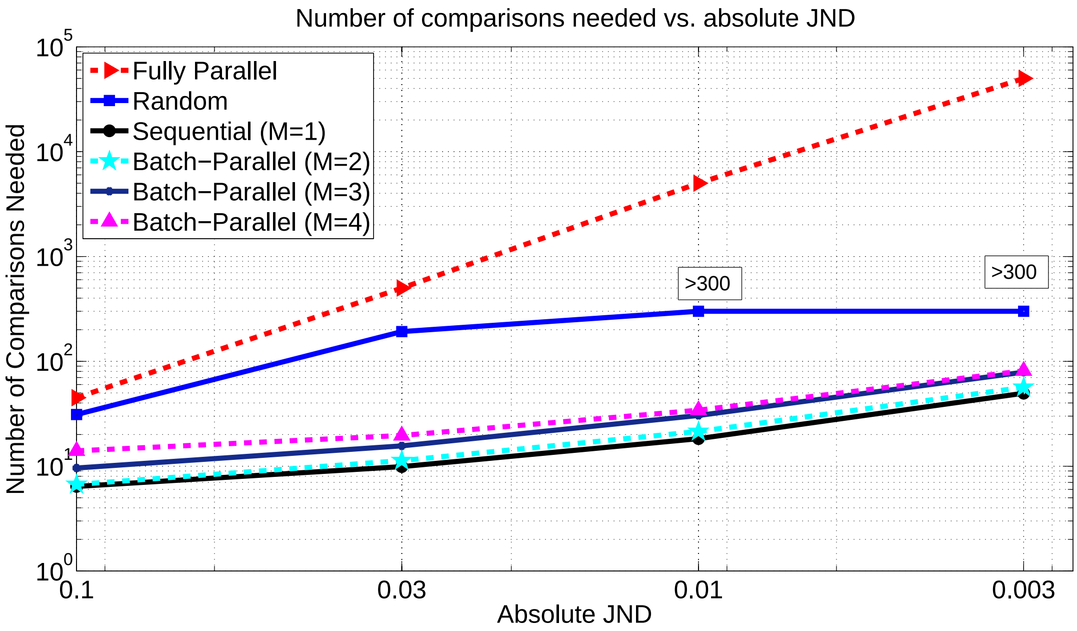


Figure 6.8: Visual representation of the results presented in Table 6.2. Expected number of comparisons for an operating curve with a single local optimum when JND is known. All comparisons result in a successful estimation of A^* (to within the JND of the actual value).

number of batches and the number of tests in each. For instance, with $JND = 0.03$, it takes an average of 4.9 batches, each with $M = 4$ tests, in a batch-parallel algorithm, as compared to an average of 9.9 batches, each with one test, in a sequential algorithm. Hence, tests should be designed to balance the overhead of synchronizing test results in each batch and the benefit of sequential algorithms that minimize the number of comparisons.

Table 6.3 summarizes the performance of our scheme for operating curves with single and multiple local optima. It also shows the trade-off between the expected number of comparisons and the accuracy of estimating the local optima (Acc %), using a stopping criterion defined by the utility (Utility Stop %) in (6.7). For all the cases studied, it suffices to stop the search when the utility reaches 50%, which leads to at least 95% success rate in predicting one of the local optima and requires approximately half of the number of comparisons. For those 5% of cases in which the algorithm fails to find a point within the JND of the local optimum, the estimation errors are very small. In short, there is a substantial reduction in the number of comparisons by using a stopping criterion based on a smaller utility value, while incurring a negligible error in the estimation.

Figure 6.9 illustrates a typical application of our sequential algorithm using a simulated model with two local optima. We observe that the belief function focuses around one of the local optima, and the utility of the estimation increases with each comparison. Since our algorithm does not

Table 6.3: Sequential scheme: expected number of comparisons and the percentage of successful estimation for single, double and triple local optima, where JND is unknown and estimated after each comparison.

JND	Utility Stop %	Single Optimum		2 Optima		3 Optima	
		Acc %	E[n]	Acc %	E[n]	Acc %	E[n]
0.1	15	85	2.7	80	2.8	90	2.1
	20	95	3.9	90	3.8	90	3.5
	30	95	5.7	90	5.5	100	6.2
	40	95	7.4	95	7.1	100	8.3
	50	100	8.7	100	8.9	100	10.6
	60	100	10.3	100	9.9	100	12.4
	90	100	17.5	100	18.8	100	20.4
0.03	15	75	5.8	75	5.8	45	3.3
	20	95	7.4	85	7.6	85	6.1
	30	95	8.5	85	8.6	100	8.2
	40	95	9.7	90	10.3	100	9.8
	50	95	10.5	95	12.2	100	11.1
	60	95	11.6	100	13.9	100	12.7
	90	100	16.7	100	19.1	100	20.7

know the number of local optima, their ROD boundaries, and the JND values, it estimates them after each comparison. Figure 6.9 also depicts the convergence of the optimum estimate in the simulations. When compared to the results in Table 6.2, the unavailability of JND causes an increase in the expected number of comparisons needed in order to find an operating point within the JND of a local optimum.

6.3 Summary

In this chapter, we have presented our method to schedule subjective comparisons based on the model developed in Chapter 5, We have presented our framework to obtain stochastic information about the comparison between two alternatives and a probabilistic derivation to combine individual pair-wise comparisons between different points on the same operating curve. The combined information is represented by belief function, which is a proper probability distribution over the region of dominance, that represents the likelihood that the optimal point is at a given location on the operating curve, when given the collective set of comparisons conducted. This representation of the combined information allows the development of a search algorithm to efficiently and accurately

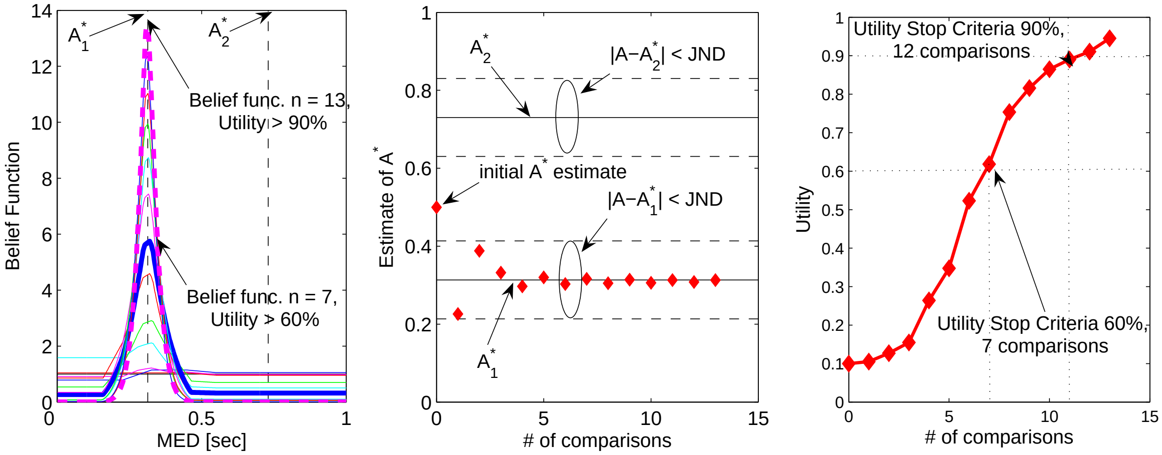


Figure 6.9: An illustration of the application of our sequential search algorithm on a simulated model with two local optima. (a) Evolution of belief function, (b) Convergence of the optimum estimated around one local optimum, (c) Improvement of utility (confidence) with number of comparisons.

identify the optimal point and its confidence.

Since it is costly to conduct actual subjective tests, we have conducted extensive Monte Carlo simulations to evaluate the accuracy and efficiency of the method against brute force and non-adaptive schemes. We have also evaluated the performance of our algorithm under conditions where there are multiple local optima. The results indicate that there are significant (up to 1000 times) gains in the reduction of the number of subjective evaluations needed to achieve a predefined level of accuracy in identifying the optimal point on a given operating curve.

In Chapter 7, we apply the method developed in this chapter over a multitude of operating curves for the design of POS scheduling for VoIP systems.

CHAPTER 7

LEARNING AND GENERALIZATION OF OFF-LINE SUBJECTIVE COMPARISONS

As discussed in Chapters 1 and 3, the quality of a VoIP conversation is characterized by multiple counteracting objective metrics (such as delay and signal quality) that are controlled by the playout scheduler (POS) of a VoIP system. However, their trade-offs leading to high subjective quality perceived by users at run time are not well defined.

In this chapter we apply the method developed in Chapter 6 to the design of a POS control scheme for a VoIP system with high subjective quality. We achieve this goal by conducting subjective tests offline, learning the mapping between the objective metrics measured and the corresponding subjective preferences, and generalizing the results to the online control of POS.

We apply our methodology for adaptive scheduling of offline subjective evaluations on a representative set of operating curves under a variety of network and conversational conditions. We verify the desirable operating points found using limited subjective tests against exhaustive evaluations on a subset of the operating curves. Finally, we learn the mapping between parameters characterizing the operating curves and the desirable operating points using support vector machines (SVMs). In Chapter 8 we utilize the classifier learned here to identify the best operating points on unseen operating curves at run time. Our approach consists of the following steps.

a) *Selecting a representative set of network and conversational conditions.* Since there are prohibitively many network and conversational conditions in a VoIP system, it is infeasible to conduct exhaustive subjective tests on all possibilities. To this end, we first identify a set of operating curves that span a wide range of network and conversational conditions for learning pairwise preferences of subjects.

b) *Scheduling offline comparative subjective tests on the multiple operating curves.* We extend the methodology for the adaptive scheduling of off-line subjective evaluations on a single operating curve developed in Chapter 6 to multiple operating curves where the evaluation is conducted sequentially on each curve and concurrently across multiple curves. Our approach uses Bayesian analysis to combine the individual subjective comparisons in order to identify the optimal operating points for each operating curve. In Section 7.1.3 we further validate our model and method using exhaustive subjective tests on a subset of operating curves.

c) *Learning the mapping between parameters characterizing operating curves and their optimal* *point.* Using SVM classifiers, in Section 7.2 we present the learning and the cross-validation of the mapping between the parameters characterizing the operating curves and the target MED. To evaluate the MEDs predicted by the classifier, we present in Appendix A, a statistical method for estimating the accuracy predicted using limited subjective results.

7.1 Experimental Results on Offline Subjective Tests of Interactive VoIP

7.1.1 Scheduling subjective evaluations of multiple operating curves

In subjective testing of a VoIP application, there are multiple independent operating curves to be evaluated. Since sequential evaluations of a single operating curve are the most effective in terms of minimizing the number of tests performed, the optimal strategy to minimize the total number of subjective tests for a set of operating curves is to test each curve sequentially and all the curves in parallel. In this approach, each subject is presented with a set of pairs of points to be compared, one pair from each operating curve. The tests in each set can be performed in any order and independent of other subjects. At the end of each set of tests, the results from all the subjects are combined to generate for each operating curve a local optimum estimate and the next pair of points to be compared. In order to conduct our tests on a comprehensive set of conditions, we utilize the same 6 network conditions used in Chapter 4.1 and expended the conversational conditions used from 3 conditions to 5 conditions, which are listed in Table 2.2.

7.1.2 Conducting subjective tests on operating curves

Based on the six 2-party connections indicated in Table 2.1 and the 5 conversations in Table 2.2, we have created 30 operating curves for our subjective tests. To generate an operating point corresponding to a two-party conversation under given MED, we have developed a simulator that generates the conversation on demand between the two parties using the appropriate MED between the two clients (assuming symmetric MEDs). Our simulator is designed to ensure the repeatability of subjective tests. As a result, when comparing two operating points under the same network and conversational conditions, variations in quality are only due to the difference in MED.

Since there are infinitely many operating points on an operating curve, our simulator generates the corresponding conversations on demand when a test is performed. To limit the amount of tests performed, we have eliminated those obviously suboptimal operating points from each curve based on earlier subjective evaluations, namely, those with PESQ less than 2.0 as well as those with MED greater than 1 sec.

Table 7.1: Pair-wise subjective preferences under LLL network and five conversational conditions.

Netw.	Conv.	A				B				Subjective Preference [%]			
Cond.	Cond.	MED	PESQ	CS	CE	MED	PESQ	CS	CE	$A <_s B$	$A \approx B$	$A >_s B$	$A?B$
LLL	1	250	4.13	1.91	0.78	500	4.13	2.81	0.63	12.5	25.0	62.5	0.0
		60	2.00	1.22	0.93	155	4.13	1.56	0.85	25.0	37.5	37.5	0.0
		60	2.00	1.22	0.93	62	4.13	1.22	0.93	75.0	25.0	0.0	0.0
		62	4.13	1.22	0.93	106	4.13	1.38	0.89	0.0	75.0	25.0	0.0
	2	250	4.15	1.70	0.89	500	4.15	2.41	0.80	0.0	50.0	50.0	0.0
		60	2.00	1.17	0.97	155	4.15	1.44	0.93	75.0	12.5	12.5	0.0
		203	4.15	1.57	0.91	522	4.15	2.47	0.80	12.5	0.0	87.5	0.0
		60	2.00	1.17	0.97	179	4.15	1.50	0.92	62.5	25.0	12.5	0.0
	3	250	4.18	1.60	0.91	500	4.18	2.21	0.84	12.5	25.0	50.0	12.5
		60	2.00	1.15	0.98	155	4.18	1.37	0.94	87.5	12.5	0.0	0.0
		203	4.18	1.49	0.93	483	4.18	2.17	0.84	0.0	50.0	50.0	0.0
		60	2.00	1.15	0.98	179	4.18	1.43	0.93	87.5	12.5	0.0	0.0
	4	250	4.01	3.27	0.93	500	4.01	5.55	0.87	12.5	62.5	25.0	0.0
		60	2.00	1.55	0.98	155	4.01	2.41	0.95	75.0	12.5	12.5	0.0
		203	4.01	2.85	0.94	545	4.01	5.95	0.86	0.0	75.0	25.0	0.0
		60	2.00	1.55	0.98	179	4.01	2.63	0.95	87.5	12.5	0.0	0.0
	5	250	3.94	2.11	0.96	500	3.94	3.22	0.92	0.0	50.0	50.0	0.0
		60	2.00	1.27	0.99	155	3.94	1.69	0.97	62.5	25.0	12.5	0.0
		203	3.94	1.90	0.97	529	3.94	3.35	0.92	0.0	62.5	37.5	0.0
		60	2.00	1.27	0.99	179	3.94	1.80	0.97	62.5	25.0	12.5	0.0

Based on the method described in Section 6.1, we first identified the initial pair of operating points on each operating curve to be compared. We then asked 8 subjects to evaluate pairs of operating points, one from each of the 30 operating curves, and combined the results after the comparisons were completed. To allow subjects to conduct each set of tests in a given window of time (say a day), while allowing them to appropriately rest in between portions of a test, we have developed a graphical user interface for subjects to listen to the alternative conversations for comparison and record their answers at their convenience during the window. After each set was completed, the estimated belief function, $\widehat{CND}$, and $\widehat{A}^*$ were updated for each operating curve. The next pair of operating points to be compared were then selected, and the process was repeated. The process stopped once the utility of the A^* predicted reached the 50% threshold identified in [56]. Such a threshold was found to be adequate for predicting more than 95% of the optimal points in a Monte Carlo simulation of thousands of randomly generated operating curves [56]. For all the 30 operating curves evaluated, the process stopped after 4 pairs of comparisons because the utilities had reached above 50%. In total, subjective tests were conducted on 120 pairs of operating points by 8 subjects over 6 network and 5 conversational conditions.

Tables 7.1- 7.3 present all of the subjective-test results for 30 operating curves by showing the

MEDs for the two operating points compared and the corresponding objective quality metrics.

Table 7.2: Pair-wise subjective preferences under LLH network and five conversational conditions.

Netw.	Conv.	A				B				Subjective Preference [%]			
Cond.	Cond.	MED	PESQ	CS	CE	MED	PESQ	CS	CE	$A <_s B$	$A \approx B$	$A >_s B$	$A?B$
LLH	1	250	4.13	1.91	0.78	500	4.13	2.81	0.63	12.5	25.0	62.5	0.0
		85	3.54	1.31	0.91	170	4.13	1.62	0.84	0.0	50.0	50.0	0.0
		83	2.00	1.30	0.91	88	3.74	1.32	0.91	37.5	62.5	0.0	0.0
		88	3.74	1.32	0.91	129	3.91	1.47	0.87	0.0	87.5	12.5	0.0
	2	250	4.15	1.70	0.89	500	4.15	2.41	0.80	0.0	37.5	62.5	0.0
		85	2.00	1.24	0.96	170	4.10	1.48	0.92	62.5	12.5	25.0	0.0
		210	4.15	1.59	0.91	508	4.15	2.43	0.80	0.0	75.0	25.0	0.0
		83	2.00	1.23	0.96	190	4.10	1.54	0.91	75.0	0.0	25.0	0.0
	3	250	4.18	1.60	0.91	500	4.18	2.21	0.84	12.5	50.0	37.5	0.0
		85	2.00	1.21	0.97	170	3.40	1.41	0.94	75.0	12.5	12.5	0.0
		210	3.99	1.51	0.92	528	4.18	2.28	0.83	12.5	50.0	37.5	0.0
		83	2.00	1.20	0.97	190	3.75	1.46	0.93	62.5	12.5	25.0	0.0
	4	250	4.01	3.27	0.93	500	4.01	5.55	0.87	0.0	50.0	50.0	0.0
		85	2.00	1.77	0.97	170	3.34	2.55	0.95	100.0	0.0	0.0	0.0
		210	3.73	2.91	0.94	485	4.01	5.41	0.87	0.0	75.0	25.0	0.0
		83	2.00	1.75	0.98	190	3.66	2.73	0.95	87.5	12.5	0.0	0.0
	5	250	3.88	2.11	0.96	500	3.88	3.22	0.92	12.5	50.0	37.5	0.0
		88	2.00	1.39	0.99	170	3.19	1.76	0.97	87.5	12.5	0.0	0.0
		210	3.46	1.93	0.97	528	3.88	3.35	0.92	12.5	50.0	37.5	0.0
		83	2.00	1.37	0.99	190	3.34	1.84	0.97	87.5	12.5	0.0	0.0

Under a high-delay (>100 ms), low-jitter, and high-loss (> 5%) condition, Tables 7.1, 7.2 and 7.3 show that A is consistently preferred over B when B has the same PESQ as A but with degraded CS and CE. However, B might be preferred when it has significantly better PESQ but with degraded CS and CE and the degradation in CS and CE is small enough. Tables 7.1-7.6 also show that, as the network conditions, such as jitter and loss, degrade, the optimal MED increases as well.

7.1.3 Validation of model parameters based on limited subjective tests

In this section, we validate our model by conducting exhaustive subjective tests and compare the results against those in Section 7.1.2 obtained using our adaptive method. In Chapter 6.2, we have validated the model and the adaptive off-line subjective method using extensive Monte Carlo simulations for a general domain that matches the characteristics of VoIP applications studied in this thesis. In this section, we further evaluate the model and the method using actual subjective

Table 7.3: Pair-wise subjective preferences under LHL network and five conversational conditions.

Netw.	Conv.	A				B				Subjective Preference [%]			
Cond.	Cond.	MED	PESQ	CS	CE	MED	PESQ	CS	CE	$A <_s B$	$A \approx B$	$A >_s B$	$A?B$
LHL	1	250	4.13	1.91	0.78	500	4.13	2.81	0.63	25.0	12.5	62.5	0.0
		110	2.00	1.40	0.89	180	4.13	1.65	0.83	37.5	50.0	12.5	0.0
		215	4.13	1.78	0.80	483	4.13	2.75	0.64	12.5	37.5	50.0	0.0
		111	2.00	1.40	0.89	198	4.13	1.72	0.81	50.0	25.0	25.0	0.0
	2	250	4.15	1.70	0.89	500	4.15	2.41	0.80	25.0	25.0	50.0	0.0
		110	2.00	1.31	0.95	180	4.15	1.51	0.92	75.0	12.5	12.5	0.0
		215	4.15	1.61	0.90	495	4.15	2.39	0.81	0.0	50.0	50.0	0.0
		111	2.00	1.31	0.95	198	4.15	1.56	0.91	50.0	25.0	25.0	0.0
	3	250	4.18	1.60	0.91	500	4.18	2.21	0.84	12.5	50.0	37.5	0.0
		110	2.39	1.27	0.96	180	4.18	1.44	0.93	75.0	25.0	0.0	0.0
		215	4.18	1.52	0.92	539	4.18	2.30	0.82	12.5	62.5	25.0	0.0
		111	2.60	1.27	0.96	198	4.18	1.48	0.93	75.0	12.5	12.5	0.0
	4	250	4.01	3.27	0.93	500	4.01	5.55	0.87	0.0	75.0	25.0	0.0
		110	1.73	2.00	0.97	180	4.01	2.64	0.95	87.5	12.5	0.0	0.0
		215	4.01	2.95	0.94	583	4.01	6.30	0.85	0.0	37.5	62.5	0.0
		111	1.74	2.01	0.97	198	4.01	2.80	0.94	87.5	12.5	0.0	0.0
	5	250	3.94	2.11	0.96	500	3.94	3.22	0.92	0.0	62.5	37.5	0.0
		110	2.00	1.49	0.98	180	3.91	1.80	0.97	75.0	12.5	12.5	0.0
		215	3.94	1.96	0.97	556	3.94	3.47	0.91	12.5	50.0	37.5	0.0
		111	1.42	1.49	0.98	198	3.94	1.88	0.97	87.5	12.5	0.0	0.0

tests. Our goal is to obtain a statistical estimation of the parameters of the model by conducting subjective tests between each pair of points on an operating curve. The process is exhaustive and does not involve the adaptive method in Section 6.1.

Since conducting exhaustive subjective tests is very costly, only a subset of the operating curves can be tested and verified. Further, as there are infinitely many feasible points on each operating curve, we also need to limit the points evaluated while ensuring that the results obtained are statistically accurate. In this subsection, we describe our systematic approach to select the operating curves to be validated and the operating points on each curve.

Operating curve selection. In our study we select 10 operating curves to conduct exhaustive tests out of 30 operating curves for which we have conducted limited adaptive subjective tests. Equation (7.1) lists the operating curves used, each represented as a tuple of network condition and conversational condition.

$$\mathcal{O}_{exhaustive} = \{(1,1),(1,4),(2,3),(3,2),(3,5),(4,1),(4,4),(5,3),(6,2),(6,5)\} \tag{7.1}$$

Table 7.4: Pair-wise subjective preferences under HLL network and five conversational conditions.

Netw.	Conv.	A				B				Subjective Preference [%]			
Cond.	Cond.	MED	PESQ	CS	CE	MED	PESQ	CS	CE	$A <_s B$	$A \approx B$	$A >_s B$	$A?B$
HLL	1	250	4.13	1.91	0.78	500	4.13	2.81	0.63	12.5	25.0	62.5	0.0
		140	3.82	1.51	0.86	195	4.13	1.71	0.82	0.0	62.5	37.5	0.0
		139	2.00	1.50	0.86	142	4.13	1.51	0.86	62.5	37.5	0.0	0.0
		142	4.13	1.51	0.86	172	4.13	1.62	0.83	25.0	37.5	37.5	0.0
	2	250	4.15	1.70	0.89	500	4.15	2.41	0.80	12.5	37.5	50.0	0.0
		140	2.00	1.39	0.94	195	4.15	1.55	0.91	75.0	25.0	0.0	0.0
		223	4.15	1.63	0.90	525	4.15	2.48	0.80	0.0	50.0	50.0	0.0
		139	2.00	1.39	0.94	209	4.15	1.59	0.91	50.0	12.5	25.0	12.5
	3	250	4.18	1.60	0.91	500	4.18	2.21	0.84	12.5	50.0	37.5	0.0
		140	2.00	1.34	0.95	195	4.18	1.47	0.93	75.0	0.0	25.0	0.0
		223	4.18	1.54	0.92	498	4.18	2.20	0.84	12.5	50.0	37.5	0.0
		139	2.00	1.34	0.95	209	4.18	1.51	0.92	62.5	12.5	25.0	0.0
	4	250	4.01	3.27	0.93	500	4.01	5.55	0.87	12.5	62.5	25.0	0.0
		140	2.00	2.27	0.96	195	4.01	2.77	0.94	87.5	12.5	0.0	0.0
		223	4.01	3.03	0.94	564	4.01	6.13	0.85	0.0	75.0	25.0	0.0
		139	2.00	2.26	0.96	209	4.01	2.90	0.94	100.0	0.0	0.0	0.0
	5	250	3.94	2.11	0.96	500	3.94	3.22	0.92	0.0	75.0	25.0	0.0
		140	2.00	1.62	0.98	195	3.94	1.87	0.97	50.0	37.5	12.5	0.0
		223	3.94	1.99	0.96	616	3.94	3.74	0.91	12.5	37.5	50.0	0.0
		139	2.00	1.62	0.98	209	3.94	1.93	0.97	75.0	12.5	12.5	0.0

We have selected these combinations in such a way that allows the selected sets to span across all conditions and avoid any bias towards one or a subset of conditions. This is done by selecting two samples from each conversational condition and at least one sample from each network condition.

Operating point selection. Due to infinitely many feasible choices of operating point and limited resources for conducting tests, we need to develop an approach to select a set of points that is small enough, yet can ensure that information on the location of the optimal point is not missed by not conducting more comparisons.

The principle that guides our approach is that the distance between two adjacent operating points selected for exhaustive tests should be so small that, if the actual optimal point is in between, it should not be perceptibly differentiable (within some statistical significance) from at least one of the sample points. The implies that the result of our "limited" exhaustive tests will be identical to that when infinitely many pairwise comparisons are conducted. This is exactly the same concept as just noticeable difference (JND) introduced in Section 5.2.2. However, the identification of JND for an operating curve is almost as hard as identifying the optimal point itself and requires subjective tests. But since the exhaustive tests need to be conducted in one batch, the selection

Table 7.5: Pair-wise subjective preferences under HLH network and five conversational conditions.

Netw.	Conv.	A				B				Subjective Preference [%]			
Cond.	Cond.	MED	PESQ	CS	CE	MED	PESQ	CS	CE	$A <_s B$	$A \approx B$	$A >_s B$	$A?B$
HLH	1	250	4.13	1.91	0.78	500	4.13	2.81	0.63	12.5	12.5	75.0	0.0
		160	3.45	1.58	0.84	190	4.13	1.69	0.82	0.0	50.0	50.0	0.0
		130	3.02	1.47	0.87	181	4.13	1.66	0.83	50.0	25.0	25.0	0.0
		160	3.45	1.58	0.84	181	4.13	1.66	0.83	37.5	37.5	25.0	0.0
	2	250	4.15	1.70	0.89	500	4.15	2.41	0.80	0.0	50.0	37.5	12.5
		130	2.00	1.37	0.94	190	3.94	1.54	0.91	87.5	12.5	0.0	0.0
		220	4.15	1.62	0.90	538	4.15	2.52	0.79	12.5	37.5	50.0	0.0
		130	2.00	1.37	0.94	208	4.15	1.59	0.91	100.0	0.0	0.0	0.0
	3	250	4.18	1.60	0.91	500	4.18	2.21	0.84	25.0	25.0	50.0	0.0
		140	2.00	1.34	0.95	195	3.33	1.47	0.93	75.0	12.5	12.5	0.0
		223	4.18	1.54	0.92	503	4.18	2.22	0.83	12.5	50.0	37.5	0.0
		130	2.00	1.31	0.95	209	4.18	1.51	0.92	75.0	12.5	12.5	0.0
	4	250	4.01	3.27	0.93	500	4.01	5.55	0.87	0.0	62.5	37.5	0.0
		140	2.06	2.27	0.96	195	3.81	2.77	0.94	100.0	0.0	0.0	0.0
		223	4.01	3.03	0.94	553	4.01	6.03	0.86	0.0	62.5	37.5	0.0
		130	2.00	2.18	0.96	209	4.01	2.90	0.94	87.5	12.5	0.0	0.0
	5	250	3.94	2.11	0.96	500	3.94	3.22	0.92	0.0	75.0	25.0	0.0
		130	1.48	1.58	0.98	190	3.35	1.84	0.97	75.0	12.5	12.5	0.0
		220	3.94	1.98	0.96	587	3.94	3.61	0.91	0.0	75.0	25.0	0.0
		130	1.48	1.58	0.98	208	3.94	1.92	0.97	75.0	12.5	12.5	0.0

of the points to be tested needs to rely on the information at hand. To this end, we utilize the CND (thus JND) approximations based on the limited adaptive subjective tests we have already conducted (described in the last subsection) on each operating curve in order to select the operating points for exhaustive tests.

The second principle that guides our approach is based on Weber's law on human perception of physical attributes [11, 13, 17]. It has been shown for several attributes that the ratio of JND of an attribute and its value roughly follow a constant. In our previous subjective experiments, we have seen that JND increases with MED. Our aim here is not to argue whether the ratio is a constant or find its value, but to utilize this observation to better select the sample points. To this end, we use a geometric separation between the sample operating points selected, where the ratio of MED between consecutive points is constant. Since the exact JND around each point is unknown, this is the best approach with the information at hand at the time of selection.

Thirdly, to avoid wasting limited tests on points that are guaranteed not to be optimal, we prune points on each operating curve based on the network condition. We do not select points smaller than the minimum MED, which is chosen to be the minimum delay observed on the network trace.

Table 7.6: Pair-wise subjective preferences under HHL network and five conversational conditions.

Netw.	Conv.	A				B				Subjective Preference [%]			
Cond.	Cond.	MED	PESQ	CS	CE	MED	PESQ	CS	CE	$A <_s B$	$A \approx B$	$A >_s B$	$A?B$
HHL	1	250	3.29	1.91	0.78	500	4.13	2.81	0.63	25.0	25.0	50.0	0.0
		130	2.63	1.47	0.87	360	3.28	2.30	0.71	0.0	25.0	75.0	0.0
		130	2.63	1.47	0.87	181	3.18	1.66	0.83	50.0	50.0	0.0	0.0
		181	3.18	1.66	0.83	469	4.13	2.70	0.65	25.0	25.0	50.0	0.0
	2	250	2.91	1.70	0.89	500	4.15	2.41	0.80	62.5	37.5	0.0	0.0
		750	4.15	3.11	0.73	1000	4.15	3.82	0.67	0.0	50.0	50.0	0.0
		500	4.15	2.41	0.80	625	4.15	2.76	0.77	0.0	37.5	62.5	0.0
	3	250	2.63	1.60	0.91	500	4.18	2.21	0.84	62.5	12.5	12.5	12.5
		750	4.18	2.81	0.77	1000	4.18	3.42	0.72	0.0	37.5	62.5	0.0
		500	4.18	2.21	0.84	625	4.18	2.51	0.80	0.0	50.0	50.0	0.0
	4	250	2.34	3.27	0.93	500	4.01	5.55	0.87	62.5	12.5	25.0	0.0
		750	4.01	7.82	0.81	1000	4.01	10.09	0.77	12.5	25.0	62.5	0.0
		500	4.01	5.55	0.87	625	4.01	6.68	0.84	12.5	37.5	50.0	0.0
	5	250	2.27	2.11	0.96	500	3.94	3.22	0.92	75.0	12.5	12.5	0.0
		750	3.94	4.33	0.89	1000	3.94	5.44	0.86	12.5	25.0	62.5	0.0
		500	3.94	3.22	0.92	625	3.94	3.78	0.90	12.5	50.0	37.5	0.0

This is reasonable because any MED smaller than the minimum will result in all packets being late. We have further chosen the maximum MED to be 750 ms. This is based on the observation that among all the limited subjective tests conducted, none of the subjects prefer points with MED larger than 750 msec with respect to any point with a smaller MED.

Based on the three principles, we select, for each operating curve, eight operating points, where the i^{th} point is

$$A_i(m,n) = D_{\min}(m)\left(\frac{750}{D_{\min}(m)}\right)^{(i-1)/7} \quad \text{where } i = 1, \ldots, 8 \text{ and } (m,n) \in \mathcal{O}_{exhaustive}. \quad (7.2)$$

Here, $D_{\min}(m)$ is the minimum network delay observed on the operating curve with network condition m.

Validation of model parameters We have recruited eight subjects to conduct the all pairwise comparisons in one batch. Based on 8 operating points for each of the 10 operating curves, these lead to 560 pairs compared by each subject and 4,480 individual results. The results can be repre-

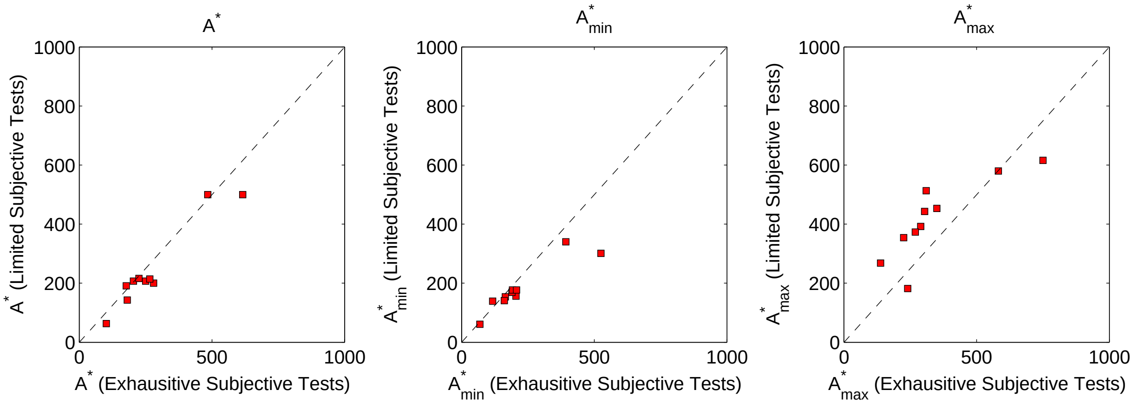


Figure 7.1: Values obtained from limited and exhaustive subjective tests on (a) A^*, (b) $A^* - JND^-$, (c) $A^* + JND^+$.

sented in 560 triplets as follows.

$$\{A_i(m,n), A_j(m,n), COD(A_i(m,n), A_j(m,n))\} \tag{7.3}$$
$$\text{for } i,j \in \{1,\ldots,8\},\ i \neq j,\ \text{and } (m,n) \in \mathcal{O}_{exhaustive}.$$

We then use Equation (6.5) and (6.6) to estimate the optimal point A^* for each operating curve.

In addition to the optimal point, there are two other parameters that are relevant to the operating curves in the context of learning algorithms described in Section 7.2. These parameters are JND estimations around the optimal point and are represented by $JND^-(A^*) < A^* < JND^+(A^*)$. Their meaning has been described in the context of learning algorithms in Section 7.2, and the method for obtaining them is described in the next section.

Figure 7.1 depicts the values obtained by the limited and exhaustive subjective tests on the 10 operating curves selected. The results show that the parameter values obtained are well aligned. To assess the closeness of the results, we evaluate the correlation coefficient, which indicates the ability to predict one value from another using a linear mapping. The correlation coefficient between the values obtained from the exhaustive and the limited subjective tests for A^* (*resp.* $A^* - JND^-$ *and* $A^* + JND^+$) is 0.977 (*resp. 0.977 and 0.942*).

In summary, the limited subjective tests are adequate for identifying the optimal point and other related parameters of the comparison model. Assuming two minutes to conduct one subjective test, our approach can complete the tests of one operating curve in 8 minutes, in contrast to the 112 minutes (not including resting time) when using exhaustive tests. Such savings can be more than 100-fold in cases where JND estimates (based on the results of our adaptive tests) are not available for selecting sample points of the exhaustive tests.

7.2 Learning and Generalization of the Optimal Operating Point

In this section, we present the learning and generalization of a classifier for predicting the optimal operating points at run time. This task requires learning a mapping between an operating curve and its optimal point, based on adaptive subjective tests conducted offline on limited operating curves.

There are two main approaches to tackle this learning problem. The *direct approach* combines the offline pairwise comparison results on multiple operating curves in order to identify their optimal points. It then directly learns the mapping between the objective attributes representing the operating curves and their optimal points. Once learned, it is straightforward to apply the classifier at run time. In contrast, the *indirect approach* learns the mapping between the pairwise preferences of subjects as a function of the objective attributes characterizing the two operating points compared. At run time, the classifier learned is used to predict the preferences of an adequate number of pairs of points on the operating curve, before combining the results to predict the optimal point.

The advantage of the indirect approach is that the mapping learned is more intuitive and corresponds to subjects' perception and trade-off among quality measures. Because there are more training samples and mis-prediction in some pairwise preferences can be corrected if the majority of the predictions are correct, the approach is expected to generalize better to unseen pairs of points on different operating curves. However, given very limited (four) subjective tests conducted on each operating curve, the learning of this classifier in order to generalize well to every pair of points on every operating curve is debatable. As a result, we focus on the direct approach.

7.2.1 Operating curve parameters: Inputs of the learning algorithm

In learning a classifier, it is difficult to directly model an operating curve as a continues multi-dimensional curve (where the number of dimensions equals the number of objective metrics representing the quality of the operating point) because it will lead to infinite tuples of continuous values as inputs to the learning algorithm. In our approach, we exploit known relationships among MED (our control variable) and the three objective quality metrics (CE, CS, and $LOSQ$), which uniquely defines the operating curve. The parameters identified as inputs to the learning algorithm are those that are changing as a function of the network and conversational conditions.

CE is a deterministic function of MED, parameterized by $HRD+ST$ value of the conversation.

$$CE = \frac{ST + HRD}{ST + HRD + MED} = 1 - \frac{1}{1 + \frac{ST+HRD}{MED}}. \tag{7.4}$$

Thus, we define the first parameter characterizing the operating curve as $X_1 = HRD + ST$.

CS is a deterministic function of MED, parameterized by the HRD value of the conversation.

$$CS = \frac{HRD + MED}{HRD} = 1 + \frac{MED}{HRD} \tag{7.5}$$

Thus, we define the second parameter characterizing the operating curve as $X_2 = HRD$.

LOSQ, measured in PESQ, is a function of MED and parameterized by the network-delay and jitter conditions as well as the robustness of the speech codec to unconcealed frames. It can be decoupled into two cascade functions, where $g()$ characterizes the network conditions and $h()$ characterizes the speech codec:

$$LOSQ = f(UCFR) = h(g(MED)). \tag{7.6}$$

Function $g()$ that models the relation between $UCFR$ and MED for a given network trace can be represented by a general exponential equation parameterized by X_3, X_4 and X_5.

$$UCFR = g(MED) = \begin{cases} 1 & \text{if } MED < D_{min} \\ 1 - X_3.(MED - X_4)^{X_5} & \text{if } D_{max} \geq MED \geq D_{min} \\ 0 & \text{if } D_{max} < MED \end{cases} \tag{7.7}$$

To satisfy the left boundary condition, X_4 is usually very close to D_{min}.

Since the relation between $UCFR$ and $LOSQ$ represented by function $h()$ does not depend on the network or conversational condition, it is not modeled or used as input to the learning algorithm.

We obtain X_3, X_4 and X_5 by standard curve fitting techniques that result in 50%, 90%, 95% and 98% concealed frame rates for the network traces used in the experiments. For a majority of the curves, the fitting results have exceptional accuracy in representing UCFR as a function of MED. However, in a few degenerate cases where the network jitter is very low (less than 5 ms), a pre-defined set of parameters are used to represent the trivial function: $X_3 = 1$, $X_4 = D_{min}$ and $X_5 = 0$, which result in a step function of UCFR as a function of MED. Figure 7.2 depicts the relation between MED and $UCFR$ for two different network conditions.

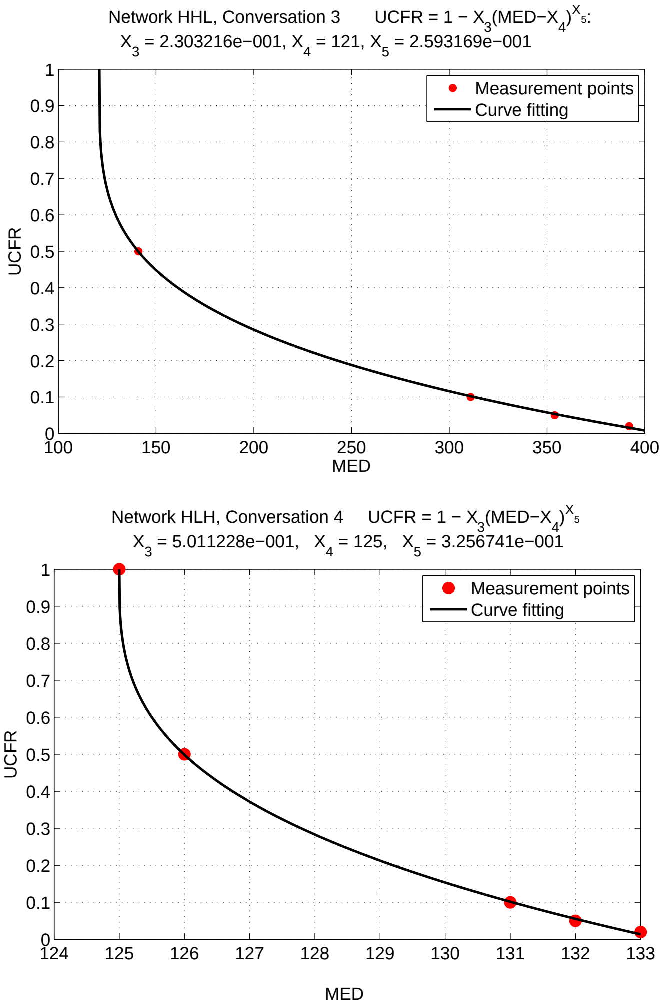


Figure 7.2: $UCFR$ vs. MED for (a) (H,H,L) (high-jitter) and (b) (H,L,H) (low-jitter) network conditions.

Loss Burstiness parameters indicate the percentage of losses that occur individually or in bursts of two or three consecutive packets. The burstiness of the losses are related with the choice of optimal MED through the redundancy degree [53] of the speech packets. Assuming enough redundant copies of a speech frame are transmitted, in order to conceal such a frame in case the original packet containing it is lost, the MED should be adequately long to allow for the arrival of redundant packets. In case of bursty losses, this additional wait time is a multiple of the packet period of the transmission.

In order to formally define the loss burstiness parameters, we utilize the definition of uncon-cealement in [53]. We define UC_i, the *unconcealment indicator* of packet i, to be zero when the loss of packet i can be concealed because either the original packet or a redundant copy is received before its scheduled playout time p_i; that is,

$$UC_i(\bar{p}, R) = \begin{cases} 0 & \text{if } (n_i + (R-1)T) \le p_i \textit{ or ... } \text{or } n_{i+R-1} \le p_{i+R-1} \\ 1 & \text{otherwise,} \end{cases} \tag{7.8}$$

where R is the redundancy degree (or the number of copies transmitted in the original and subsequent packets), T is the packet period and $\bar{p}$ is the play-out time for R packets starting with packet i in a vector form ($\bar{p} = [p_i, \ldots, p_{i+R-1}]$). In practice, R is an integer between 1 and 4.

We further define the *loss burstiness rate* (LBR_i^W) as the percentage of those unconcealed frames among a window of W frames ending with frame i, where the play-out time is infinitely long:

$$LBR_i^W(\bar{p}, R) = \frac{100}{W} \sum_{j=i-W+1}^{i} UC_j(\bar{p}, R),$$

where elements of $\bar{p}$ are all equal to infinity. In this definition, R also indicates the level of burstiness of the losses since a redundancy level of R can only conceal a burst of $R-1$ packet losses. Thus, simply denoted, $LBR(R)$ is the fraction of packets not concealable with redundancy rate of R. In other words, the reduction in unconcealment rate due to increase of redundancy rate from R to $R+1$ equals $LBR(R-1) - LBR(R)$.

Thus, we utilize the recent statistics of packet loss burstiness to map to the optimal MED for the operating conditions, where $X_6 = LBR(1)$, $X_7 = LBR(2)$ and $X_8 = LBR(3)$ for the window of last 30 seconds.

Based on subjective tests on the 30 operating curves for the combination of 6 network conditions and 5 conversational conditions in, respectively, Tables 2.1 and 2.2, the following matrices tabulate the values of the 8 parameters, where $\bar{X}_i(m, n)$ represents the i^{th} parameter for the m^{th} network and n^{th} conversational conditions.

$$\bar{X}_1 = \begin{bmatrix} 531 & 1784 & 2258 & 3765 & 6329 \\ 531 & 1784 & 2258 & 3765 & 6329 \\ 531 & 1784 & 2258 & 3765 & 6329 \\ 531 & 1784 & 2258 & 3765 & 6329 \\ 531 & 1784 & 2258 & 3765 & 6329 \\ 531 & 1784 & 2258 & 3765 & 6329 \end{bmatrix} \qquad \bar{X}_2 = \begin{bmatrix} 220 & 450 & 552 & 710 & 827 \\ 220 & 450 & 552 & 710 & 827 \\ 220 & 450 & 552 & 710 & 827 \\ 220 & 450 & 552 & 710 & 827 \\ 220 & 450 & 552 & 710 & 827 \\ 220 & 450 & 552 & 710 & 827 \end{bmatrix}$$

$$\bar{X}_3 = \begin{bmatrix} 60 & 60 & 60 & 60 & 60 \\ 83 & 83 & 60 & 60 & 83 \\ 78 & 61 & 61 & 61 & 61 \\ 139 & 139 & 75 & 75 & 139 \\ 125 & 125 & 125 & 125 & 125 \\ 121 & 121 & 121 & 121 & 121 \end{bmatrix} \qquad \bar{X}_4 = \begin{bmatrix} 1.0000 & 1.0000 & 0.8212 & 0.8266 & 0.8063 \\ 0.7747 & 1.0000 & 0.4176 & 1.0000 & 0.8408 \\ 0.0241 & 0.0483 & 0.0656 & 0.0371 & 0.0466 \\ 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ 0.5011 & 1.0000 & 0.5525 & 0.5143 & 0.5853 \\ 0.1939 & 0.4357 & 0.3610 & 0.3555 & 0.2303 \end{bmatrix}$$

$$\bar{X}_5 = \begin{bmatrix} 0 & 0 & 0.0865 & 0.0776 & 0.1008 \\ 0.1365 & 0 & 0.2599 & 0 & 0.0584 \\ 0.8867 & 0.7664 & 0.6155 & 0.7426 & 0.6768 \\ 0 & 0 & 0 & 0 & 0 \\ 0.3257 & 0 & 0.3420 & 0.3761 & 0.2002 \\ 0.2945 & 0.1923 & 0.2531 & 0.2245 & 0.2593 \end{bmatrix} \qquad \bar{X}_6 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 18.6 & 9.7 & 20.5 & 20.7 & 22.4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 14.0 & 7.6 & 18.5 & 17.2 & 17.1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\bar{X}_7 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 2.3 & 4.8 & 5.9 & 4.8 & 5.4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2.1 & 2.0 & 1.8 & 1.7 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \bar{X}_8 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1.1 & 1.5 & 1.8 & 2.0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{7.9}$$

At run-time, recent speech and silence durations (say within the last 30 sec) can be used to obtain the average ST and HRD locally at each VoIP client. Similarly, recent network delay, jitter, and loss conditions can be used to obtain the MEDs to achieve 50, 90, 95 and 98% concealment for a recent window of time.

7.2.2 Comparison model parameters: Targets for the learning algorithm

Based on the parametric model in Figure 6.2a, there are three parameters that uniquely identify the model: A^*, CND, and α. In the context of choosing the best MED at run time, only A^* is relevant. However, it is still important to estimate JND (or CND) at design time in order to evaluate the accuracy of the learning algorithm.

In practice, it is extremely hard to find JND for the entire operating curve using only the limited (in this case 4) pairwise-comparison results in our adaptive tests. Instead of conducting more tests, if suffices to estimate JND around A^* using the belief function from the pairwise comparisons.

To differentiate JND at a point where MED can be decreased (*resp.* increased), we extend $JND(A)$ in Definition 1 (Chapter 5) into $JND^-(A)$ (*resp.* $JND^+(A)$). This extension allows us to better evaluate the trade-offs among CS, CE and $LOSQ$ at A^*. When increasing MED above A^*, subjects perceive the differences in CE and CS more with respect to the difference perceived in $LOSQ$. In contrast, when decreasing MED below A^*, subjects perceive the difference in $LOSQ$ more than they do for CE and CS. Since the set of dominant quality metrics is different for the two directions, the corresponding JND may be different as well.

In short, the belief function obtained from subjective comparisons of an operating curve is used to estimate the results of two comparisons, namely, $(A^* - JND^-, A^*)$ and $(A^*, A^* + JND^+)$. Since the JNDs are unknown, a variety of pairs are estimated until the JNDs are found. Let (A, B), where $A < B$, be a pair of operating points whose COD is to be estimated. For the first (*resp.* second) type of comparison, $B = A^*$ (*resp.* $A = A^*$). The boundaries of this range $[A^* - JND^-(A^*), A^* + JND^-(A^*)]$ are identified in such a way that when comparing any point outside of this range with A^*, a hypothesis test that K_1 (number of responses indicating $A >_s B$) and K_{-1} (number of responses indicating $A <_s B$) are drawn from a binomial distribution with K subjects and $p = 0.5$ can be rejected with 85% significance. Here p represents the probability of $A >_s B$ in a binomial distribution.

Appendix A presents the derivation of $JND^-(A)$ and $JND^+(A)$. Note that the derivation is not intended to extend the general model in [56] and the limited model in Section 5.2.2. It is meant as an approximation when estimating the accuracy of A^*, using the hypothesis tests described last.

The matrices below tabulate the optimal operating point $A^*(m, n)$ and its minimum and maximum values for the m^{th} network and the n^{th} conversational conditions (a total of 30 conditions).

$$A^* = \begin{bmatrix} 92 & 196 & 202 & 206 & 202 \\ 103 & 216 & 214 & 216 & 216 \\ 216 & 215 & 225 & 222 & 224 \\ 170 & 226 & 232 & 237 & 234 \\ 156 & 233 & 233 & 233 & 234 \\ 186 & 490 & 492 & 492 & 503 \end{bmatrix} \tag{7.10}$$

$$A^* - JND^-(A^*) = \begin{bmatrix} 61 & 128 & 139 & 138 & 125 \\ 83 & 144 & 143 & 156 & 156 \\ 132 & 154 & 165 & 169 & 168 \\ 141 & 175 & 177 & 188 & 179 \\ 130 & 183 & 181 & 185 & 177 \\ 131 & 340 & 326 & 301 & 361 \end{bmatrix} \tag{7.11}$$

$$A^* + JND^+(A^*) = \begin{bmatrix} 268 & 317 & 354 & 545 & 375 \\ 140 & 372 & 453 & 390 & 453 \\ 359 & 373 & 538 & 392 & 429 \\ 182 & 370 & 443 & 564 & 451 \\ 297 & 394 & 429 & 409 & 513 \\ 263 & 580 & 600 & 616 & 820 \end{bmatrix} \tag{7.12}$$

7.2.3 Learning and generalization of the mapping between operating curve and A^*

There are a variety of methods for learning the mapping between the five parameters in (7.9) and A^*. In this section, we present our approach by support vector machines using the libSVM implementation [8].

Quantization of target values Since SVM is designed for classifying finite and discrete targets, we need to quantize our target values before learning. The same quantization mapping is used to convert the class predicted by SVM into an A^* value. In general, increasing the number of classes leads to less quantization error but lower accuracy of the SVM because the resulting SVM needs to identify more boundaries to partition the training vectors into classes, each with a smaller number of samples.

In our application, we use JND to determine the quantization mapping. In particular, the quantization bins are chosen to be so small that no two perceptually distinguishable points are mapped to the same bin. That is, the JND of a point mapped to a bin should be larger than its bin size.

Similar to what is done in Section 7.1.3, we use a geometric separation of quantization boundaries based on Weber's law and our observations on previous subjective tests. To this end, we

Table 7.7: Prediction accuracy of SVM based on four validation techniques.

	Average Accuracy			
Validation Type	Self-validation	Leave-one-out	10-fold	2-fold
Dependent inputs	100%	83.3%	90.0%	82.7%
Independent inputs	100%	86.7%	86.7%	84.7%

first determine the extreme network delays from the 30 operating curves in our experiments. The smallest delay is 61 ms ($D_{\min}$), and the largest $A^* + JND^+(A^*)$ estimated is 871 ms ($D_{\max}$). We then determine the quantization boundaries in this range, with 10 geometrically separated points as mid-points of the ranges they represent. The quantization level of A is

$$Q(A) = c, \text{ if } D_{\min}R^{c-1} \leq A \leq D_{\min}R^c, \ \text{ where } \ R = \left(\frac{D_{\max}}{D_{\min}}\right)^{\frac{1}{11}} \text{ (geometric ratio).} \tag{7.13}$$

Eleven quantization levels are chosen geometrically to convert the continuous control space into a finite alphabet, in order to simplify the learning of the mapping without loss of accuracy. In addition to the considerations in Section 7.1.3, we also considered the distribution of the quantized operating points for the 30 operating curves. Eleven quantization levels result in a relatively well distributed training set, which is considered to be a desired characteristic for unbiased prediction results.

For unseen operating curves, the quantized values in extreme cases with less than 61 ms (*resp.* greater than 871 ms) can be rounded up (*resp.* rounded down) without too much difference in perceptual quality.

Using the inverse function, the output class predicted by SVM (quantization index c of the optimal point) can be converted into the predicted value in ms by $\widehat{A^*} = Q^{-1}(c) = D_{\min}R^{(c-0.5)}$.

Accuracy of the learning results $\widehat{A^*}$ is deemed to be accurate if it lies in the range of A^* in (7.10), (7.11) and (7.12). Hence, the aggregate accuracy of learning is simply the algebraic average of the individual accuracies:

$$\text{Accuracy}(\widehat{A^*}) = \frac{1}{N_{\text{net}}N_{\text{conv}}} \sum_{m=1}^{N_{\text{net}}} \sum_{n=1}^{N_{\text{conv}}} 1_{\{\widehat{A^*}(m,n) \in [A^*(m,n) - JND^-(m,n), A^*(m,n) + JND^+(m,n)]\}} \tag{7.14}$$

where $1_{\{\text{condition}\}}$ is an indicator function which equals 1 if the condition is true and 0 otherwise.

We apply the four cross-validation techniques described in Chapter 4.1 to evaluate the performance of the trained SVM.

Table 7.7 summarizes the results of two learning experiments. In the first experiment, the SVM uses dependent inputs $X_1 = HRD + ST$ and $X_2 = HRD$. This dependency may pose problems when the learning algorithm pre-scales their inputs, which may cause dependent numbers to become independent. To address this issue, we decouple the two inputs in our second experiment before pre-scaling and use $X_1 = ST$ and $X_2 = HRD$. The results indicate comparable performance for self-validation and 2-fold cross validation, whereas leave-one-out and 10-fold cross validations perform better for dependent inputs.

7.3 Summary

In this chapter we have applied our methodology for adaptive scheduling of off-line subjective evaluations by conducting subjective tests on a representative set of conditions for a real-life control design problem. Using the domain knowledge about the POS control design problem, we have pruned the search space to a finite length and have observed that a significantly small number of comparisons are adequate in identifying an optimal operating point under each of the 30 conditions evaluated.

We have further verified the accuracy and efficiency of our model by conducting exhaustive subjective evaluations on 10 of the operating curves. We have observed that our adaptive scheduling scheme took 4 comparisons whereas the exhaustive evaluation took 56 comparisons to identify the optimal operating point. Furthermore, the correlation of the identified optimal points between the two methods is 0.977, which indicates that the 14 times reduction in the number of tests needed does not result in any reduction in the accuracy of the identification.

In the second part of the chapter, we have presented the learning of the optimal operating points identified by limited subjective tests using the network and conversational conditions obtained at run-time.

We have first identified the objective measures that can be obtained at run time and are related to the perception of quality. Secondly, we have identified the target optimal value obtained by subjective tests. In order to have a metric of success for the prediction of the optimal point on a continuous set of alternatives, we have developed a method to obtain an acceptable range of values around the optimal point with statistical significance. In essence, we have extended the model formulation to allow for the calculation of the JND around A^*, which is indicative of the sensitivity of human perception to changes of the control value around the optimal point. This relaxes the assumption used in the simplified parametric model, which is used in Chapter 6 for the derivation of our adaptive scheduling scheme, that JND is constant for all points on the operating curve. To maintain the flow of the presentation, we have presented this derivation in Appendix A.

The evaluation of the SVM classifier performance shows that the self-validation accuracy is 100%. This indicates that the outputs can be perfectly predicted by the inputs when all the samples are available for training, and provides an upper-bound for the other types of validation performance. Secondly, the leave-one-out, 10-fold and 2-fold cross-validation performance is between 83% and 90%, which is a pretty high prediction rate given the limited amount of subjective evaluations conducted. A very high cross-validation accuracy would have indicated that the conditions tested are repetitive or redundant. This result would have indicated that some of the subjective evaluations have been unnecessary and non-beneficial. On the other hand, a lower cross-validation accuracy would have indicated that the SVM classifier would not generalize well with unseen network and conversational conditions. Thus, in summary, we have achieved an ideal result in the cross-validation of the classifier in terms of the efficiency of the subjective tests conducted and its suitability to be used in the design of a new VoIP system that achieves high perceptual quality under unseen run-time conditions.

In the next chapter, we bring all the components of the VoIP architecture together to design a new VoIP system and compare against previously evaluated systems.

CHAPTER 8

DESIGN AND EVALUATION OF A VOIP SYSTEM WITH HIGH PERCEPTUAL QUALITY

In this chapter we first present the design of a new VoIP system based on our analysis of the overall VoIP system architecture. We present the design choices for the VoIP system components in detail based on our previous studies on those individual components and discuss how all the components operate in concert to achieve the overall goal.

We give particular attention to the design of the POS component due to its instrumental role in achieving high perceptual-quality conversations via adaptations to changing network and conversational conditions. The POS design utilizes the mapping learned in the previous section that relates the objective measures obtained at run time and the subjectively preferred operating point.

Secondly, we present objective and subjective evaluation of the newly designed VoIP system against other VoIP systems we have evaluated in Chapter 4. We conclude this chapter by presenting further evaluations of our newly designed system under unseen conditions.

8.1 VoIP System Design

As discussed in Chapters 2 and 3, VoIP systems commonly contain three main components which include the playout scheduler that controls the MED, the loss-concealment scheme that provides robustness against network losses, and the speech codec that encodes and decodes the speech to low bit-rate stream. These affect the quality of the speech signals received under ideal and imperfect network conditions. A separate component is responsible for collection and dissemination of network and conversational conditions.

Speech codec. In our system we use ITU G722.2 [24] wide-band codec for compression of the speech waveform and packetize two 20 ms frames into a one 40 ms packet. As described in Chapter 3, we pack redundant copies of previously transmitted packets to allow for loss concealment at the receiver. Since even the highest redundancy packets are much smaller than MTU sizes, IP packets do not fragment and thus the increase in the packet size does not affect the loss rate of the packet stream [50].

Dynamic collection and dissemination of network and conversational conditions. There are variations in the network and conversational conditions across talk-spurts. In order to adjust system controllable metrics in accordance to these variations, efficient collection and dissemination of network and conversational conditions are needed. There are trade-offs between the overhead incurred in collecting the statistics and the accuracy of the statistics. The adaptation decisions require current and relevant information about the changes in the network conditions in order to be effective in combating network imperfections.

MED adjustments during a conversation in response to changes in the network conditions need to be frequent enough to accommodate delay spikes in order to avoid long durations of late packets which would result in annoying gaps within a speech segment. Network delay conditions need to be collected to make such a decision. Loss concealment decision, on the other hand, needs less frequent adaptations but requires the cooperation of the two VoIP clients. The cooperation of VoIP clients incurs delays more than the round trip delay of the connection.

Loss Concealment. In our design, a redundancy-degree decision is made at the receiver and is fed to the transmitter, similar to the one in [53], where the redundancy degree is chosen to be the minimum redundancy needed to achieve at most 2% unconcealment, under the assumption that no packets are considered late for play-out. By using this criterion, the contribution of loss is isolated from the decisions of the play-out scheduler; thus, redundancy decisions can be made independent of the POS and are reported to POS. The scheme uses a network condition corresponding to a window of the last 30 seconds. Only changes in the redundancy degree are relayed, which allow the reaction time to be quick and the network overhead to be small. Thus, for the purpose of MED decisions, we can safely assume that an adequate level of redundancy is available in the received packet stream to conceal packet loss or losses when adequate MED values are chosen by the POS.

Play-out Scheduling. We utilize the SVM trained using limited subjective evaluations in Chapter 7 to make run-time decisions on MED for each talk-spurt in a VoIP conversation. Thus, the performance of the POS control scheme is an indication of the generalization of the mapping learned between parameters characterizing the operating curve and the optimal point (A^*) identified by the pair-wise subjective evaluations.

The system maintains a running estimate of the 8 parameters used as SVM inputs, based on the network and conversational conditions. The window of time used for this estimation is chosen to balance the accuracy of a longer observation window and the agility of a shorter observation window. Thus, the parameters $\bar{X}_4$ through $\bar{X}_8$ are estimated over a window of 30 seconds preceding the decision time. On the other hand, the parameters $\bar{X}_1$ and $\bar{X}_2$ are estimated using the last 4 conversational turns.

At decision time, the input features are fed into the SVM prediction subroutine, which quickly generates a quantization level based on the SVM model trained off-line. The prediction process is quite fast, as it only involves the evaluation of K (number of support vectors) linear equations, each with 8 inputs and comparing the sum against 10 ranked constants (representing 10 boundary values) to distinguish among 11 quantization levels.

Simulation of VoIP system. In generating the two-way VoIP conversations that our system realizes, we utilize the testbed we have implemented to exactly replay any network and conversational condition and allow the simulation of any POS scheme in MED decisions on a talk-spurt basis. We also utilize this tool to simulate other POS schemes under the same conditions which we compare with our system in the next section. The simulator is also capable of obtaining all relevant quality metrics, including the PESQ value, for off-line analysis of the performance of these systems.

8.2 Evaluation of Newly Designed VoIP System

8.2.1 Evaluation of our VoIP system

We have evaluated our VoIP system in three parts.

a) *Objective evaluations.* In the first part, we evaluate our VoIP system using objective metrics and compare it with objective evaluations we have previously conducted on four commonly used VoIP systems (Skype 3.6, Google-Talk (beta), Yahoo Messenger 8.1 and Windows Live 8.1) in Chapter 4. To ensure a fair comparison, we use the same network traces and conversational conditions. In addition to the 6 network conditions described earlier —(LLL), (LLH), (LHL), (HLL), (HLH), and (HHL)— we also use the perfect network condition labeled (No,No,No) with 50 ms constant network delay, no jitter, and no packet loss. Similarly, we use the same three conversational conditions used in the system evaluation, namely, conversation types 3, 4 and 5 in Table 2.2. In Table 8.1, we extend Table 4.3 to include our newly designed VoIP system (labeled *SubjOpt* for Subjectively Optimal).

The results show that *SubjOpt* achieves significantly better LOSQ than the other systems under almost all conditions. This is attributed to the fact that the optimal MED predicted at run time leads to a better trade-off between LOSQ and MED. For conversation types 3, 4 and 5 in Table 2.2, our system increases the MED in order to achieve low jitter and losses. Our system does not necessarily choose the lowest possible MED because the previous network conditions may change in the next talk-spurt. By choosing the optimal MED slightly above the minimum MED, the additional delay is usually not perceptible, whereas operating at the minimum MED may have degradations because losses cannot be concealed by the G.722.2 codec used.

Table 8.1: Objective evaluations of five VoIP systems tested under six Internet and one ideal connections. The best quality for each of the four systems is indicated indicated by '*'.

Trace Class	VoIP	Conv. Type 3				Conv. Type 4				Conv. Type 5			
(Del,Jit,Lo)	System	PESQ	MED	CS	CE	PESQ	MED	CS	CE	PESQ	MED	CS	CE
(No,No,No)	SubjOpt	4.100*	212	1.77	69	4.240*	263	1.74	76	4.100*	258	1.62	84
	Skype	3.192	286	2.04	67	3.244	338	1.95	74	3.418	290	1.70	83
	GTalk	3.557	130*	1.47*	71*	3.506	147	1.42	78*	3.536	160	1.39	85*
	Yahoo	3.553	140	1.51	71*	3.676	139*	1.39*	78*	3.785	151	1.37	85*
	WinLive	3.562	171	1.62	70	3.856	154	1.43	78*	3.928	133*	1.32*	85*
(L,L,L)	SubjOpt	4.100*	274	1.99	67	4.240*	274	1.77	76	4.100*	274	1.66	83
	Skype	3.328	319	2.15	66	3.119	541	2.52	71	3.254	392	1.95	82
	GTalk	3.371	203*	1.74*	69*	3.525	368	2.04	74	3.092	201*	1.49*	84*
	Yahoo	3.534	205	1.74*	69*	3.492	203*	1.57*	77*	3.354	298	1.72	83
	WinLive	3.675	222	1.81	69*	3.492	218	1.61	77*	3.746	393	1.95	82
(L,L,H)	SubjOpt	3.080	247	1.89	68	3.460*	256*	1.72*	76*	3.290*	247	1.60	84*
	Skype	2.339	442	2.60	63	2.461	416	2.17	73	2.565	424	2.02	81
	GTalk	2.484	230	1.83	69*	2.501	265	1.75	76*	2.305	275	1.67	83
	Yahoo	2.502	217*	1.79*	69*	2.755	276	1.78	76*	2.485	239*	1.58*	84*
	WinLive	3.306*	336	2.22	66	3.309	340	1.96	74	3.257	321	1.78	83
(L,H,L)	SubjOpt	3.810*	288	2.04	67	4.240*	293	1.83	75	4.100*	293	1.71	83*
	Skype	2.693	408	2.48	64	2.882	487	2.37	72	3.083	420	2.02	82
	GTalk	3.145	216*	1.78*	69*	3.145	227*	1.64*	77*	2.854	261*	1.63*	83*
	Yahoo	3.085	274	1.99	67	3.097	240	1.68	76	2.987	274	1.66	83*
	WinLive	3.454	404	2.47	64	3.512	432	2.22	73	2.953	420	2.02	82
(H,L,L)	SubjOpt	4.100*	313	2.13	66	4.240*	313	1.88	75	4.100*	313	1.76	83*
	Skype	3.096	550	2.99	61	3.325	462	2.30	72	3.444	420	2.02	82
	GTalk	3.466	281*	2.02*	67*	3.517	279*	1.79*	76*	3.435	287	1.69	83*
	Yahoo	3.531	283	2.03	67*	3.464	305	1.86	75	3.687	301*	1.73*	83*
	WinLive	3.792	313	2.13	66	3.803	315	1.89	75	3.647	309	1.75	83*
(H,L,H)	SubjOpt	4.110*	308	2.12	66	4.180*	293*	1.83*	75*	4.030*	293*	1.71*	83*
	Skype	2.619	535	2.94	61	2.564	504	2.42	72	2.564	503	2.22	81
	GTalk	2.639	273*	1.99*	67*	2.666	283	1.80	75	2.469	300	1.73	83*
	Yahoo	2.749	281	2.02	67*	2.472	365	2.03	74	2.617	314	1.76	83*
	WinLive	3.060	440	2.60	63	3.251	421	2.19	73	3.286	363	1.88	82
(H,H,L)	SubjOpt	3.760*	489	2.77	62	3.390*	467	2.32	72	3.380*	484	2.17	81
	Skype	2.985	612	3.22	59	2.983	574	2.62	70	2.652	648	2.57	79
	GTalk	3.296	399*	2.45*	64*	3.151	410*	2.15*	73*	2.729	397*	1.96*	82*
	Yahoo	3.022	544	2.97	61	3.068	487	2.37	72	2.841	573	2.39	80
	WinLive	3.327	595	3.15	60	2.937	589	2.66	70	2.930	748	2.81	78

It is important to note that the above observations are not based on heuristics. Rather, they are the culmination of subjective-comparison test results in which the preference degradations due to shifts in MED in positive and negative directions are systematically captured in the belief function, which in turn lead to the subjectively optimal MEDs with the highest expected subjective preference against other feasible operating points.

Table 8.2: Comparative subjective evaluations of pairs of VoIP systems and the prediction results using the SVM learned in Chapter 4. In comparing A and B, the dominant opinion with 90% statistical significance is shown: $<$ (*resp.*, $\approx$, $>$, and ?): A is better than (*resp.*, about the same as, worse than, and inconclusive with respect to) B. In the inconclusive case, no dominance relation with 90% significance is found.

System Pairs A vs. B	Conv. Type	Subjective Prediction Results						
		Trace						
		NNN	LLL	LLH	LHL	HLL	HLH	HHL
SubjOpt	3	>	?	>	>	?	>	>
vs.	4	>	>	>	>	>	>	>
Skype	5	>	>	>	>	>	>	>
SubjOpt	3	>	?	>	>	?	>	>
vs.	4	>	>	>	>	>	>	>
GTalk	5	>	>	>	>	>	>	>
SubjOpt	3	>	?	>	>	?	>	>
vs.	4	>	>	>	>	>	>	>
Yahoo	5	>	>	>	>	>	>	>
SubjOpt	3	>	?	>	>	?	>	>
vs.	4	>	>	>	>	>	?	>
WinLive	5	>	>	>	>	>	>	>

b) *Subjective evaluations.* Next, we evaluate our VoIP system against the four other systems by subjective comparisons. Our analysis in Chapter 4 shows that Windows Live is predominantly preferred over the other three systems due to the consistently higher LOSQ value, despite the slightly higher MEDs with respect to Yahoo and Google-Talk. Skype does not perform well in both objective and subjective evaluations due to generally lower LOSQ and generally higher MED than most of the other systems. Instead of conducting actual subjective tests, we utilize the SVM we have learned in Chapter 4 for subjective comparison of two systems. The inputs to this SVM consist of 22 metrics that include the objective measures for characterizing the common conversational and network condition of the two systems compared. The output is the predicted subjective preference trained by the results of human listening tests; namely, whether A is preferred over B, or B is preferred over A, or both have indistinguishable conversational quality, or no statistical conclusion can be deduced. In contrast to the SVM learned in Chapter 7 for the design of POS control, the SVM used for comparisons in Chapter 4 is at a system level and samples used represent the entire conversation.

Table 8.2 tabulates the relative preference of *SubjOpt* with respect to the other systems under the 7 network and 3 conversational conditions. We observe that our system is always preferred in 75 out of 84 cases in which the preference is conclusive with 90% statistical significance. For those few remaining cases, although the objective metrics indicate that *SubjOpt* achieves better results, the subjective comparison does not lead to a dominant preference with 90% statistical significance.

c) *Objective and subjective evaluations under unseen conditions.* Lastly, we evaluate *SubjOpt* using objective metrics and subjective preferences under unseen network and conversational conditions. To show that *SubjOpt* generalizes well to unseen conditions, we compare *SubjOpt* against an ideal VoIP conversation generated by subjectively optimal decisions. In this ideal conversation, we handpick the MED for each talk spurt based on future network conditions and our experience learned from subjective tests. Such MED allows the ideal conversation to conceal all lost frames in each upcoming talk spurt, except for those rare cases in which the delay of a particular packet is so large that it would be subjectively better to increase CE and CS rather than to conceal that frame.

Note that in order to handpick the MED values for each talk-spurt to generate the ideal conversation, complete knowledge of the future network and conversational conditions and the speech uttered in each talk spurt is needed. Furthermore, this information about future conditions needs to be processed to obtain the PESQ-MED relation. The processing involves the encoding of speech, injecting unconcealed frames based on a given MED, decoding of speech and finally evaluation of PESQ for each operating point on the operating curve for each of the talks-spurts in the conversations. The processing is very computationally expensive and, depending on the number of MED alternatives on an operating curve, takes 100-300 times longer than the duration of the talkspurt even for multi-core processor computers.

Thus, obtaining the ideal conversation through brute-force evaluation of the operating curve is impossible in real-time for two reasons; the first reason is the unavailability of the future network and speech information and the second reason is the intense computation needed to obtain the operating curve for each decision point.

For these reasons, it is not possible for a causal system like *SubjOpt* to achieve the same level of performance as the ideal conversation. Thus, our aim is to have *SubjOpt* achieve similar or slightly degraded performance under most unseen conditions, when compared to the ideal conversation.

We choose unseen network conditions in Table 2.1, other than those used in earlier evaluations. Similarly, we randomly order conversational segments from the 5 conversations used earlier to construct a conversational condition unseen by SVM. Further, each system uses the same G722.2 codec and redundancy packing scheme.

Table 8.3 summarizes the performance of both systems under the 6 combinations of unseen conditions. Table 8.3 also lists the subjective preference predicted by the SVM trained in [55]. For four of the combinations, *SubjOpt* is perceived equivalent to the ideal conversation with 90% statistical significance, whereas no conclusion is reached for the rest. The objective metrics further support the subjective results.

Table 8.3: Objective comparisons of *SubjOpt* and the ideal conversation.

Network	Conv. Cond.	*SubjOpt*				Ideal Conversation				Subjective Preference
Cond.	similar to	PESQ	MED	CS	CE	PESQ	MED	CS	CE	*SubjOpt* vs Ideal
9	2	4.239	300	1.85	75	4.239	198	1.56	77	$\approx_s$
9	3	4.051	293	1.71	83	4.097	235	1.57	84	$\approx_s$
10	4	3.806	277	3.52	38	3.831	176	2.60	44	$?_s$
13	1	3.916	274	1.99	67	4.100	232	1.84	69	$\approx_s$
13	2	4.071	274	1.77	76	4.238	238	1.67	76	$\approx_s$
19	4	3.653	249	3.26	40	3.831	161	2.46	45	$?_s$

Figures 8.1- 8.3 further depicts the LOSQ-MED trade-offs for each talk spurt, where talk-spurt based PESQ-MED trade-offs are represented by solid curves. The operating points predicted by *SubjOpt* and by the ideal conversation are indicated by circles and stars, respectively, on each trade-off curve. We observe that the trade-off curves vary significantly from one talk spurt to the next, indicating significant temporal variations in the network conditions. Consequently, the hand-picked optimal MEDs also vary significantly over time. On the other hand, the MEDs predicted by our SVM have smaller variations and yet have similar LOSQs with respect to the non-causal alternative.

Depending on the segment, the best achievable LOSQ varies by speech segments. This is an upper-limit of LOSQ performance of the speech codec used for the speech segment. However, since both systems compared use the same codec, the variations in the upper-limit of LOSQ do not effect the comparison.

The first plot in Figure 8.1 depicts the trade-offs and system operating points under network type 9 and conversation type similar to type 2, briefly referred as condition $(9, 2)$. The 6 talk-spurts in this conversation exhibit widely changing network conditions where the minimum and maximum network delay observed for each talk-spurt vary by as much as 100 ms within seconds. In this condition, *SubjOpt* achieves the higest LOSQ possible for each of the talk-spurts. On the other hand, due to the vastly changing network conditions, *SubjOpt*, which bases its prediction on previously observed conditions, overestimates the optimal point in some talk-spurts, up to a maximum of 70 ms. However, this increase in MED is not subjectively perceived; thus, the two systems are found to be subjectively indistinguishable with 90% statistical significance. (See Table 8.3).

Under condition (9,3), depicted in Figure 8.1, *SubjOpt* prediction for one of the talk-spurts achieves slightly degraded LOSQs compared to the ideal conversation. However, as the difference in PESQ was less than 0.250, the overall conversational quality is again indistinguishable from that of the ideal conversation. *SubjOpt* under conditions (13,2) and (13,1) exhibits small degradations in PESQ with respect to the ideal conversation similar to that of condition (9,3), and similarly is indistinguishable when compared to the ideal conversation under the corresponding condition.

Under condition (10,4) *(resp. (19,4))*, depicted in Figure 8.1, *SubjOpt* prediction for one of the talk-spurts achieves 140 ms higher MED *(resp. 1.500 lower PESQ)* with respect to the ideal conversation. These differences, even though only for one talk-spurt, are significant enough for the subjective comparison results not to conclude that the two systems are indistinguishable with 90% statistical significance. On the other hand, the differences are not significant enough to result in a subjective opinion that the ideal conversation is preferred over *SubjOpt* with any statistical significance.

In summary, under most of the conditions, *SubjOpt*'s performance is similar or slightly degraded with respect to the ideal conversation in terms of MED and LOSQ. However, the subjective predictions indicate that these slight degradations in most cases are insignificant; thus, the two systems are subjectively indistinguishable with 90% statistical significance.

8.3 Summary

In this chapter we have presented the design of a new VoIP system based on our analysis of the overall VoIP system architecture, giving particular attention to the design of the POS component due to its instrumental role in achieving high perceptual quality. We have also presented a systematic evaluation of the newly designed VoIP system against other VoIP systems we evaluated in Chapter 4. Lastly, we have presented objective and predicted subjective comparisons against other POS algorithms under unseen conditions.

The significance of this chapter is that our newly designed system is shown to perform better than all the other VoIP systems we evaluated in Chapter 4 under all 21 conditions tested, both in terms of objective measures and the subjective preferences with statistical significance. Only under a few of the conditions can the dominant preference of our system not be established with a 90% significance, but under no conditions does any other system outperform our system.

Furthermore, we have compared our system's performance with an ideal conversation under unseen conditions. The ideal conversation utilizes information about the future, such as the network and conversational conditions, and is assumed to have complete knowledge about the multitude of objective metrics for each point on the operating curve for the upcoming talk spurt. The comparison indicates that for a majority of the previously unseen conditions, the subjective preference between the ideal system and our system is perceptually indistinguishable with 90% statistical significance.

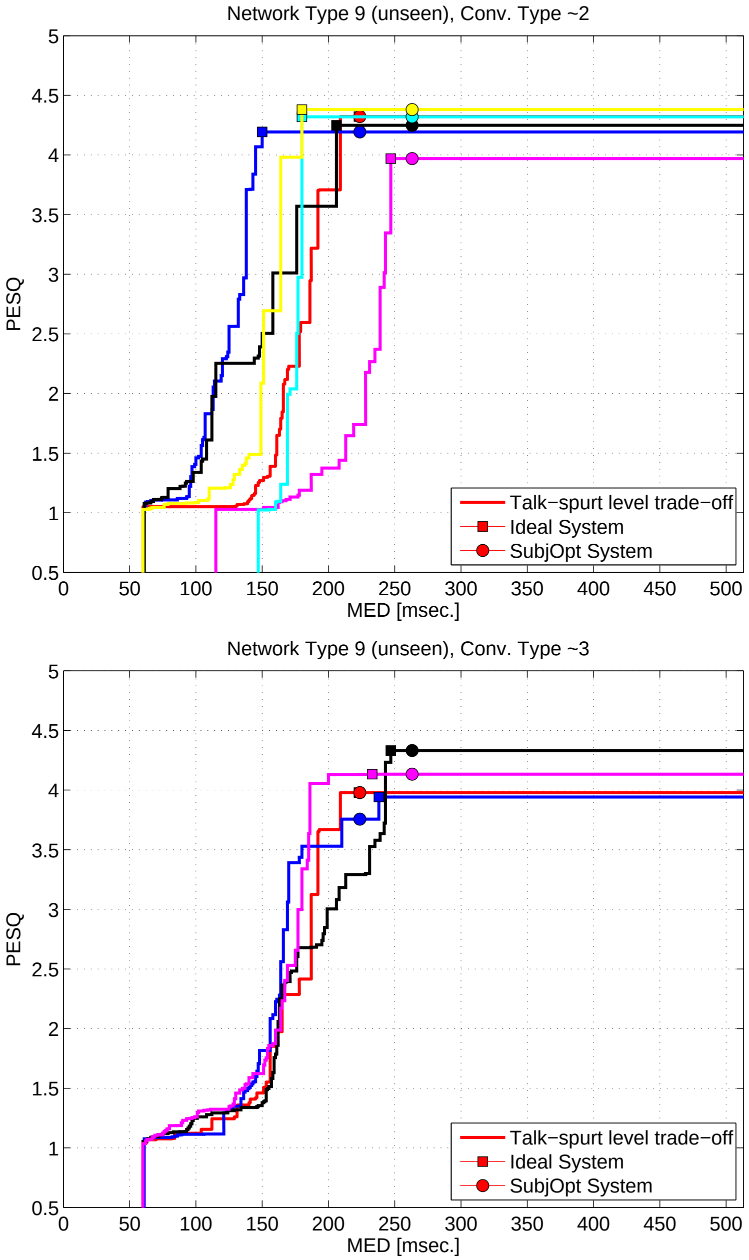


Figure 8.1: PESQ-MED trade-offs (simplified 2-D curve) of individual talk spurts of a conversation under 6 unseen conditions. Each color represents a trade-off curve for a talk-spurt, where the first talk-spurt is red, followed by blue, black, magenta, cyan, yellow and green. Depending on the type of conversation, there are 4 to 10 talk-spurts in each conversation. The operating points of *SubjOpt* and the ideal conversation are indicated by circles and squares, respectively, on each curve.

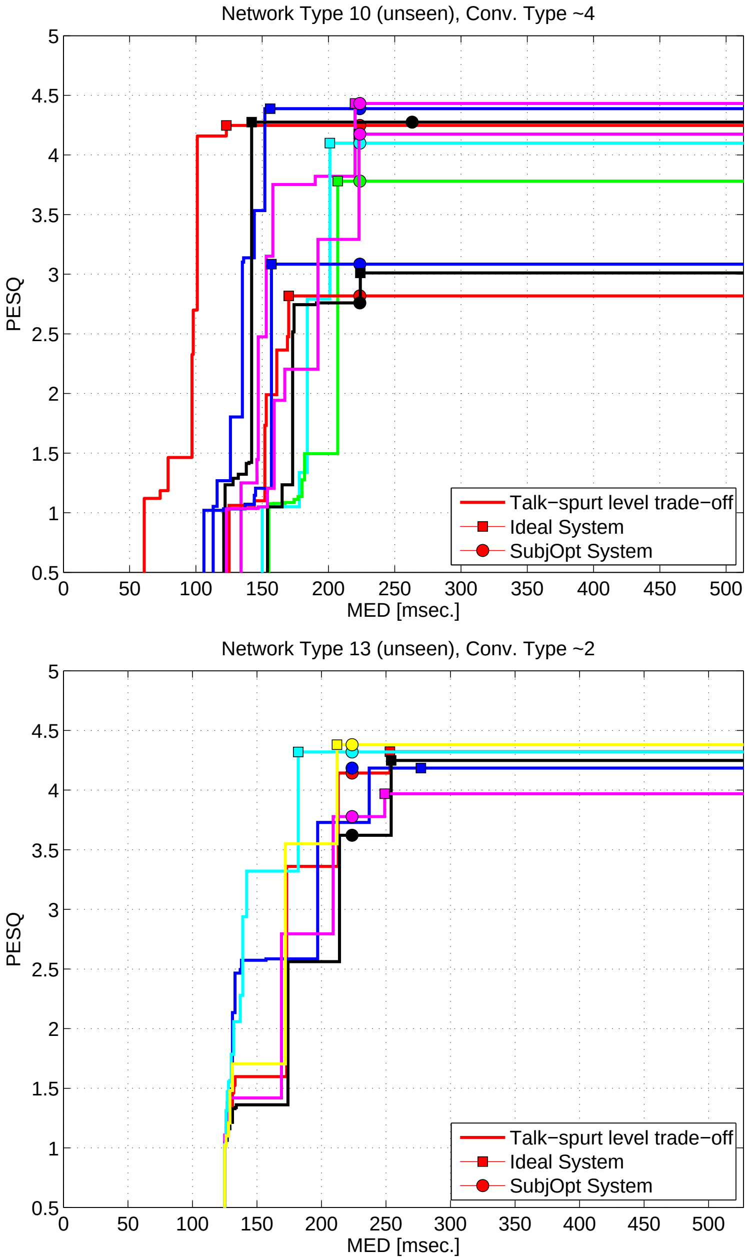


Figure 8.2: PESQ-MED trade-offs (simplified 2-D curve) of individual talk spurts of a conversation under 6 unseen conditions. Each color represents a trade-off curve for a talk-spurt, where the first talk-spurt is red, followed by blue, black, magenta, cyan, yellow and green. Depending on the type of conversation, there are 4 to 10 talk-spurts in each conversation. The operating points of *SubjOpt* and the ideal conversation are indicated by circles and squares, respectively, on each curve.

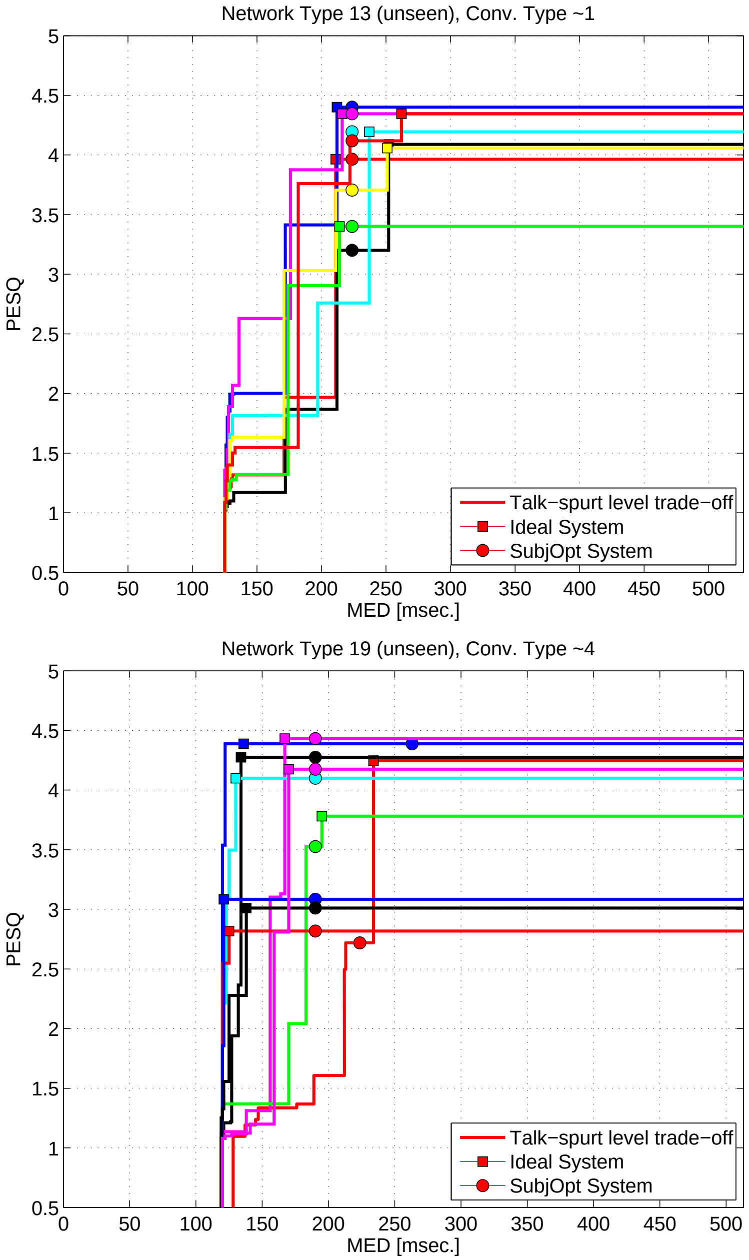


Figure 8.3: PESQ-MED trade-offs (simplified 2-D curve) of individual talk spurts of a conversation under 6 unseen conditions. Each color represents a trade-off curve for a talk-spurt, where the first talk-spurt is red, followed by blue, black, magenta, cyan, yellow and green. Depending on the type of conversation, there are 4 to 10 talk-spurts in each conversation. The operating points of *SubjOpt* and the ideal conversation are indicated by circles and squares, respectively, on each curve.

CHAPTER 9

CONCLUSIONS AND FUTURE WORK

In this chapter we summarize our conclusions for the thesis and present some of the future work opportunities to apply the methodology to other system control design problems and to extend the methodology further to solve a wider range of problems.

9.1 Summary of Accomplished Research

The following are our conclusions of the thesis:

- Firstly, we conclude that the evaluation of perceptual quality of real-time VoIP systems cannot be achieved by simple objective measures in an absolute rating. Perceptual quality needs to be evaluated on a pair-wise basis using a set of objective measures for characterizing the two alternatives and multiple parameters for characterizing network and conversational conditions under which the comparison is done.

- Secondly, the mapping between this set of objective metrics and the subjective preference between two alternatives can be learned. The learning of this mapping allows for any two alternative systems that are evaluated under the same set of conditions to be subjectively compared, without the need of subjective tests.

- Thirdly, we conclude that our model of pair-wise subjective comparisons provides a solid framework for any subjective evaluation problem, where absolute category rating is not adequate and pair-wise subjective comparisons are prohibitively expensive to conduct.

- Fourthly, we conclude that our methodology to adaptively schedule pair-wise subjective comparisons is quite efficient for reducing the number of subjective comparisons dramatically, both in extensive Monte Carlo simulations and in a real-life VoIP POS design problem. We further show that an absolute ranking is not needed to identify the best alternative among a set of feasible points on an operating curve, and that a few sufficiently obtained relative ranks would suffice.

- Lastly, we conclude that our overall design method can achieve high perceptual quality with minimal subjective comparisons conducted on a representative set of conditions identified. By utilizing the mapping learned in Chapter 4, we can compare our newly designed system against commercial systems, both objectively and subjectively, and observe that our system achieves superior perceptual quality under all the conditions evaluated.

The following are our contributions of the thesis:

- The first contribution of this thesis is the identification of a comprehensive set of objective measures that are related to the perception of conversational quality of VoIP calls and that can also be obtained at run-time during a call. The close relation of these objective metrics and the subjective preferences of subjects are evident by the very high self-prediction accuracies of the mappings learned in Chapters 4 and 7.

- The second contribution of this thesis is the method developed to comparatively evaluate VoIP systems in Chapter 4, along with the SVM model learned. The latter can successfully predict the subjective preference between two unseen VoIP systems under unseen conditions, based on objective measures obtainable using our VoIP system evaluation testbed.

- The third contribution is the development of the *model of pairwise subjective comparisons* based on individually identified properties, axioms and lemmas. The model provides a basis for developing a method to schedule adaptive off-line subjective tests and for identifying the optimal point by fusing the information obtained from separate subjective evaluations on the same operating curve. The model can be used in formulating and solving any type of pairwise comparison problem that exhibits the same properties identified. The model is flexible to allow the existence of multiple optimal points on an operating curve and includes a belief function framework that can guide the search for optimal points efficiently. Furthermore, the model is built on a statistical framework that allows for the confidence of individual evaluation results to be represented in the conclusiveness of the combined belief function.

- The fourth contribution is a method for tackling the control design problem of finding the optimal point in an N-dimensional space. This is transformed into two orthogonal problems of finding the optimal point on a continuous but one-dimensional curve, and learning the mapping on a set of curves that adequately spans the K-dimensional ($K < N$) curve space. In this framework, K stands for the number of metrics characterizing the network and conversational conditions, where N stands for all the metrics that affect quality. The latter include quality metrics characterizing the VoIP conversation as well as the K metrics mentioned above.

- The last contribution of this thesis is the application of all the methods developed to the design of POS control for a VoIP system. This includes conducting extensive subjective comparison tests that lead to the development of a new VoIP system. Our system outperforms existing systems and performs very close to an ideal system where the POS decision is made optimally using future information.

9.2 Future Work

In this section we provide two examples of the application of our method on the design of control schemes for real-time multimedia communication systems. We then describe the limitations of our model and methods and discuss possible extensions to allow for a wider set of problems to be solved using our method.

9.2.1 Multi-party VoIP system

Problem description. We consider the design of a mutual-silence equalization scheme for a VoIP system with multiple participants [48, 18] as a possible application. This problem exhibits the characteristics we have identified in real-time multimedia communication systems that would benefit from subjective-evaluation guided design of its control schemes. The system has a fundamental trade-off between two (or more) objective quality metrics that users of the system can perceive. The objective quality metrics are affected by the network conditions and communication scenario, thus requiring run-time adaptation to achieve robust and high perceptual quality. The objective quality metrics can also be affected by a system control with counteracting objectives, meaning that when one is improving, the other is degrading.

In this application, there are three quality metrics — LOSQ, CE and CS — which are all functions of MED chosen by the play-out scheduler of the system. However, since in this case there are multiple VoIP clients that are participating in the conference, for any one instance, there are multiple listeners. Thus, this poses a distributed control-scheme problem, where there is a trade-off

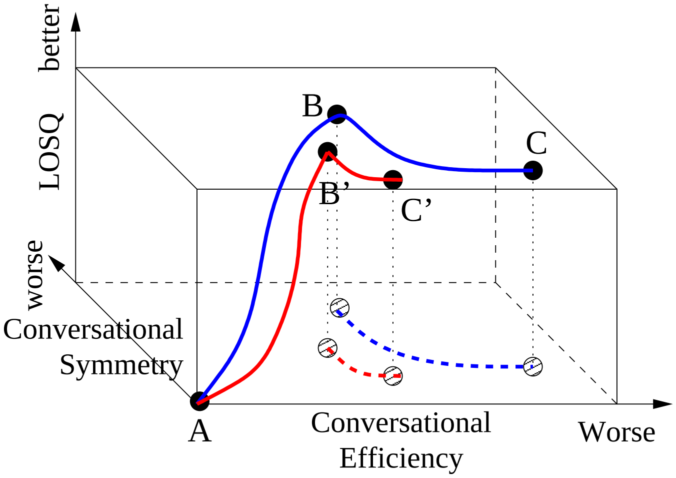


Figure 9.1: Trade-offs under different multi-party conditions.

between the balance of perceived mutual silences by different listeners and the efficiency of the conversation.

Figure 9.1 extends the trade-off curve for the two-party VoIP to the multi-party case. It depicts two trade-off curves: the blue curve containing A, B, and C corresponds to a network condition with high disparities in delays among the connections; and the red curve containing A, B' and C' corresponds to a condition with similar average delay but considerably less disparity. The control from A to B (*resp.*, A to B') is similar to the two-party case: increasing the MED towards B (*resp.*, B') conceals more packets and improves LOSQ but degrades CS and CE. In the multi-party case, further increasing the MED from B to C (*resp.*, B' to C') to achieve full equalization will lead to a high LOSQ with improved CS but degraded CE. Hence, C will result in a highly inefficient conversation with low conversational efficiency, whereas C' is relatively more efficient than C.

We envision a joint POS and mutual-silence equilization algorithm with a control parameter operating in a continuous spectrum of equalization levels. The algorithm controls the operating point on the operating curve for the given conditions.

Our method can be applied to this problem, where CS and CE are the counteracting quality metrics, and a single parameter for adjusting the balance of the mutual silences is the control metric. Limited subjective tests need to be conducted under a comprehensive set of network and conversational conditions identified based on domain knowledge. Once the limited subjective comparisons are completed, a mapping between objective metrics characterizing the network and conversational conditions and the subjectively preferred operating point under that condition is learned. The results can be generalized by incorporating the mapping in the design of a distributed POS / MS equalization control module.

9.2.2 Real-time video conferencing system

Problem description. We consider the design of a joint POS algorithm for audio and video for the real-time video conferencing as a possible application. This problem also exhibits the characteristics we have identified in real-time multimedia communication systems that would benefit from subjective-evaluation guided design of its control schemes. The system has a fundamental trade-off between two (or more) objective quality metrics that users of the system can perceive. The objective quality metrics are affected by the network conditions and communication scenario, thus requiring run-time adaptations to achieve robust and high perceptual quality. The objective quality metrics can also be affected by a system control and are counteracting, meaning that when one is improving, the other is degrading.

In this application, there are four quality metrics: one-way video quality (such as ITU P.910 [42] or ITU G.1070 [23]), LOSQ, CE and CS, which are all either non-decreasing or non-increasing monotonic functions of MED. Users of the system perceive the quality of video and speech, which improves with MED, and perceive the degradations due to delay such as CE and CS, which degrade with increasing MED. Furthermore, the non-increasing and non-decreasing sets of quality metrics counteract with each other. Similar to VoIP play-out scheduling, the control space is single dimensional and continuous.

As studied for VoIP systems, the network conditions affect the operating curve and the preferred operating point. In addition to that, for this application, the amount of movement in the video affects the robustness of the decoder at concealing errors. Thus, under the same network impairment and POS policy, the perceptual video quality may be higher if the movement in the video is less. This affects the operating curve which is represented in the multi-dimensional space of user perceived monotonic quality metrics.

Furthermore, the relative amount of speech with respect to the movement in video also affects the operating curve representing the overall quality tradeoff of a video conferencing system. If there is less movement and more speech, the optimal operating point may favor shorter MED values as speech decoders are usually far more robust to unconcealed frames than video decoders. On the other hand, if there is a lot of movement, a longer MED would be preferred to avoid significant degradation in video. This is depicted in Figure 9.2, where the optimal MED for speech only and video only may be different due to differing robustness of speech and video decoders to frame losses. Thus, the overall optimal joint MED for video and speech may be somewhere in between the two optimal MEDs depending on the relative amount of movement in the video and the speech content.

Our method can be applied to this problem, where limited subjective tests need to be conducted

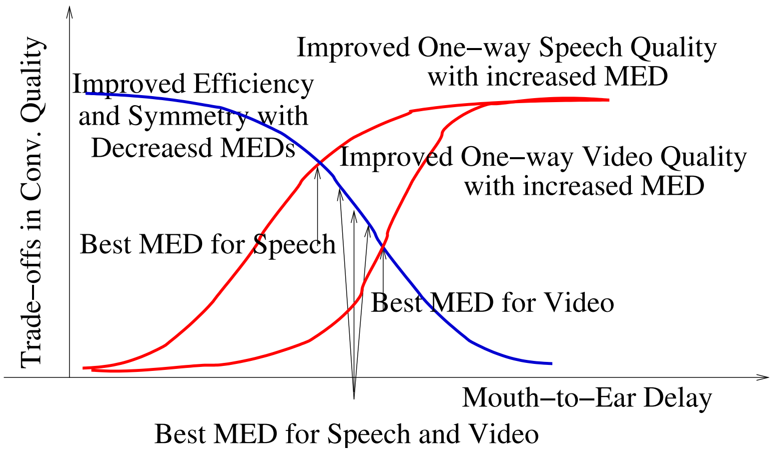


Figure 9.2: Trade-off considerations for a real-time video conferencing system.

under a comprehensive set of network and conversational conditions identified based on domain knowledge. Once the limited subjective comparisons are completed, a mapping between objective metrics characterizing the network and conversational conditions and the subjectively preferred operating point under that condition is learned. The results can be generalized by incorporating the mapping in the design of a joint POS control module for audio and video.

9.2.3 Limitations of our methodology and possible extensions

Our model of subjective comparison tests and the method to schedule efficient evaluations assumes that the control parameter can take continuous values in a single dimension. Thus, it is capable of solving control problems with infinitely many alternatives. However, in its current state, the methodology is not designed to guide the subjective optimization of multiple dimensional control parameters. However, it is expected that with some re-modeling to relax this assumption, even multiple dimensional control problems may be tackled.

It is also possible to approach the multi-dimensional problem from a different formulation, where a subset of control combinations in a multi-dimensional space is projected into a single dimension. This may be thought of as reduction of independent dimensions due to system constraints. For example, there may be multiple control components that affect the same set (two or more counteracting metrics) of quality metrics. Furthermore, the operation of these control components may be inter-related based on a system constraint, such as bit-rate or play-out scheduling time. Thus, in reality, these two control components cannot operate in total independence of each other; thus, we may restrict the joint operation of these components with a policy which is controlled by a single control parameter. In this case, we need to verify that the assumptions of our method hold true. The most important assumption is that each quality metric is either monotonically non-increasing or non-decreasing with the control parameter. This ensures that quality metrics used in the analysis do not have any maxima or minima other than the boundary points.

APPENDIX A

ESTIMATION OF THE JND RANGE OF A^*

In this appendix, we apply the Bayesian analysis in Section 6.1 along with some simplifying assumptions to derive approximate JND values around A^* using the limited subjective tests conducted.

For each operating curve, we are given the limited pairwise comparison tests conducted adaptively among pairs of points on that curve. In this case, there are 4 pairs of comparison for each of the 30 curves.

$$\text{Given: } \{A_i, B_i, COD(A_i, B_i)\}, \text{ for } i = 1, \ldots, 4. \tag{A.1}$$

Based on the 4 comparisons, the belief function is made up of 4 likelihood functions and a scaling constant to make it a proper probability distribution. All likelihood function have a three-piece linear shape depicted in Figure 6.4, where the function is constant below A and above B and linear between A and B.

To find the JND of A^*, we need to obtain $COD(A^*, B)$ for any arbitrary B on the operating curve. However, such comparison between A^* and B has not been conducted and is costly to conduct due to the unlimited number of possible pairs. Thus, our approach is to approximate $COD(A^*, B)$ based on the belief function already obtained by making a simplifying assumption.

As presented for the general case in Eq. (6.5) in Chapter 6.1.2, the combined belief function after the 4th comparison is

$$f^4_{A^*}(a) = \frac{\prod_{i=1}^4 L(a|COD(A_n, B_n) = \overline{p})}{\int_0^1 \prod_{i=1}^4 L(\eta|COD(A_n, B_n) = \overline{p})d\eta}. \tag{A.2}$$

This belief function is constructed from individual likelihood expressions based on subjective comparisons of points on an operating curve. In general, belief functions can be used to characterize an underlying operating curve. This is illustrated by the fact that the combined belief functions obtained by testing different pairs of points on the same operating curve would lead to a very similar curve encompassing the characteristics of the operating curve (see Figure 6.6). This property allows us to obtain A^* by multiple comparisons. Note that A^* obtained based on this belief function is not exact, but rather an *estimate* of the optimal point based on *limited* pairwise comparisons.

In estimating the JND of A^*, we like to estimate a two-comparison result between (A^*, Y_2) and (Y_1, A^*) on the same operating curve, where A^* is the estimate of the optimal point and Y_1 and Y_2 are variable points. The approximation is to obtain $COD(A^*, Y_2)$ and $COD(Y_1, A^*)$ from the belief function that consists of 4 comparisons, none of which are particularly between (A^*, Y_2) and (Y_1, A^*).

To use the belief function to estimate such a comparison result, we analyze the empirical data obtained from the subjective comparison tests we have already conducted.

Corollary A.1 *Given that values of the belief function are different at Y_1 and $Y_2 \neq Y_1$, $f(Y_1) \neq f(Y_2)$, there must exist at least one comparison (A_i, B_i), $i = 1, \ldots, 4$, for which A_i and B_i do not satisfy: $Y_1 < Y_2 < A_i < B_i$ or $A_i < B_i < Y_1 < Y_2$.*

Proof A.1 *If all comparisons (A_i, B_i), $i = 1, \ldots, 4$, satisfy $Y_1 < Y_2 < A_i < B_i$ or $A_i < B_i < Y_1 < Y_2$, then $L(Y_1|COD(A_i, B_i)) = L(Y_2|COD(A_i, B_i))$ for all comparisons. Thus, $f(Y_1) = f(Y_2)$. Contradiction!*

Definition A.1 Type 1 comparison. *Comparison between (A^*, Y_2), where Y_2 is variable.*

Definition A.2 Type 2 comparison. *Comparison between (Y_1, A^*), where Y_1 is variable.*

Application of corollary A.1: We apply the corollary to the problem at hand. Consider Type 1 comparison. Given $Y_2 \neq A^*$ and $f(Y_2) \neq f(A^*)$, there exists i such that $A^* < A_i < Y_2 < B_i$, or $A_i < A^* < Y_2 < B_i$, or $A_i < A^* < B_i < Y_2$, or $A^* < A_i < B_i < Y_2$ is true. In other words, there must be a comparison already conducted (among $((A_1, B_1), (A_2, B_2), (A_3, B_3), (A_4, B_4))$ that explains the reason why the belief function has different values at A^* and Y_2. The four cases indicate all the possible cases where the belief function at A^* and Y_2 would be different because of the contribution of the likelihood function corresponding to comparison i. Conversely, if $A^* < Y_2 < A_i < B_i$ or $A_i < B_i < A^* < Y_2$ was true for all comparison already conducted, than the belief function at A^* and Y_2 would have been equal, which would have lead to a contradiction. The index i for which this condition is true is called the *critical comparison* (or cc) and is specific to an operating curve and the Type 1 or 2 comparison. The corollary can be similarly applied to a Type 2 comparison between (Y_1, A^*), where Y_1 is variable.

Lastly, we approximate $COD(A^*, Y_2)$ and $COD(Y_1, A^*)$ by the COD of the corresponding critical comparison.

$$COD(A^*, Y_2) \approx COD(A_i, B_i), \text{ where } i \text{ is the critical comp. for Type 1 comp.} \quad (A.3)$$
$$COD(Y_1, A^*) \approx COD(A_i, B_i), \text{ where } i \text{ is the critical comp. for Type 2 comp.} \quad (A.4)$$

This approximation divides the likelihood expressions in Eq. (A.2) into two subsets: the first containing the *critical comparisons* (cc) where the value $L(A^*|\overline{p}) \neq L(Y_2|\overline{p})$, and the second including only the non-critical comparisons where the value $L(A^*|\overline{p}) = L(Y_2|\overline{p})$.

$$\prod_{i=1}^{n} L(a|COD(A_i, B_i) = \overline{p}) = \prod_{i \in cc} L(a|COD(A_i, B_i) = \overline{p}) \prod_{i \notin cc} L(a|COD(A_i, B_i) = \overline{p}) \quad (A.5)$$

We continue the derivation of Type 1 comparison between (A^*, Y_2), where there is a single critical comparison. The derivation for Type 2 comparison and multiple critical comparisons can be extended from the derivation below. By taking the ratio of the belief function values at A^* and Y_2, we obtain

$$\frac{f^n(A^*)}{f^n(Y_2)} \approx \frac{L(A^*|COD(A^*, Y_2))}{L(Y_2|COD(A^*, Y_2))} * \underbrace{\frac{\prod_{i \notin cc} L(A^*|COD(A_i, B_i))}{\prod_{i \notin cc} L(Y_2|COD(A_i, B_i))}}_{=1} * \underbrace{\frac{\int_0^1 \prod_{i=1}^n L(\eta|COD(A_n, B_n))d\eta}{\int_0^1 \prod_{i=1}^n L(\eta|COD(A_n, B_n))d\eta}}_{=1}$$

Since the denominator of the belief function is a constant and is common to both expressions, they cancel each other. Further, the likelihood expressions in the numerator corresponding to the non-critical comparisons, where $L(Y_1|\overline{p}) = L(Y_2|\overline{p})$, also cancel each other. Thus, the equation reduces to a simple ratio R_f of the likelihood expression of the comparison between A^* and Y_2, evaluated at A^* and Y_2.

$$R_f = \frac{f^n(A^*)}{f^n(Y_2)} = \frac{L(A^*|COD(A^*, Y_2) = \overline{p})}{L(Y_2|COD(A^*, Y_2) = \overline{p})} = \frac{p_0 + p_2 + p_1}{p_0 + p_2 + p_{-1}}. \quad (A.6)$$

Worst case analysis: There are 3 degrees of freedom when choosing the 4 elements of the COD. However, since in this case, we are interested in the relation between p_1 and p_{-1}, we can reduce the complexity of the analysis by considering a single degree of freedom. This can be done by assuming the worst case for the conclusiveness of the comparison, in terms of identifying one of the alternatives as preferred. In hypothesis testing, we use the most conservative (smallest) ratio

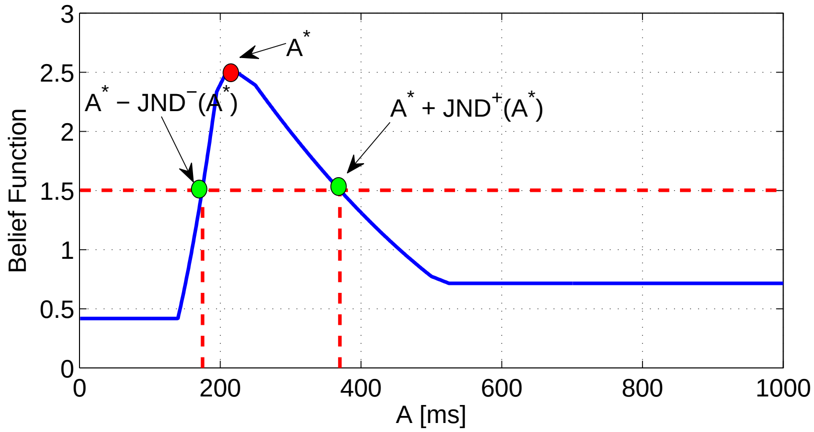


Figure A.1: A^* and the range of values that are indistinguishable when compared to A^*.

between p_1 and p_{-1}, which is achieved when p_0 and p_2 are assumed to be 0:

$$\frac{p_1}{p_{-1}} \geq \frac{p_0 + p_2 + p_1}{p_0 + p_2 + p_{-1}} = R_f. \tag{A.7}$$

We obtain the most conservative p_1 and p_{-1} values via normalizing by $p_1 + p_{-1} = 1$.

$$p_1 = \frac{R_f}{1 + R_f}, \quad p_{-1} = \frac{1}{1 + R_f}. \tag{A.8}$$

Given that there were 8 subjects conducting each test, we can make a hypothesis test on the estimated p_1 values against a binomial distribution with equal probability ($p = 0.5$). Formally, given that $COD(A^*, Y_2) = (\frac{1}{1+R}, 0, \frac{R}{1+R}, 0)$, and $K = 8$ subjects conducted the test, the hypothesis that $\frac{KR}{R+1}$ is drawn from $binom(K, p = 0.5)$ can be rejected with 85% significance if $R \geq \frac{5}{3}$ (obtained using CDF of $binom(8, p = 0.5)$).

For Type 1 comparison, we identify the minimum Y_2 that satisfies the rejection criteria (A.9). Similarly, for Type 2 comparison we identify the maximum Y_1 value that satisfies the rejection criteria (A.10).

$$JND^-(A^*) = \arg\min_{Y_2} \left\{ \frac{f^n(Y_1 = A^*)}{f^n(Y_2)} \geq \frac{5}{3} \right\} - A^* \tag{A.9}$$

$$JND^+(A^*) = A^* - \arg\max_{Y_1} \left\{ \frac{f^n(Y_1)}{f^n(Y_2 = A^*)} \geq \frac{5}{3} \right\} \tag{A.10}$$

Figure A.1 depicts the A^* range identified for Network Condition 4 (H,L,L) and Conversational Condition 5, where $A^* = 216$ ms and $A^* - JND^-$ and $A^* + JND^-$ are, respectively, 175 ms and

370 ms. In this case, the peak of the belief function at A^* is about 2.5, and the point that is $\frac{3}{5}$ times the peak value is 1.5.

In general, as the number of subjects increases, the confidence of the belief function increases and the ratio R_f to obtain the same statistical significance decreases. This results in a smaller range of A^* values that cannot be distinguished from A^*, which takes the prediction accuracy definition to a higher standard.

APPENDIX B

MODEL SIMULATION FOR MONTE CARLO EVALUATION

In this appendix, we describe the procedure to simulate pair-wise comparison models that are used in Chapter 6.2 to evaluate our method for identifying local optima on unknown operating curves.

A pair-wise comparison model, as described in Chapter 5 can be uniquely defined using the values of $p_i(A, B)$, where $i = \{-1, 0, 1, 2\}$ and $(A, B) \in [0, 1]^2$. This information can also be represented as 4 surfaces over the comparison plane, each corresponding to $p_i, i = \{-1, 0, 1, 2\}$.

Thus, the goal of the procedure is to generate these 4 surfaces in a way that is consistent with the axioms of the general model. The values representing the 4 surfaces are generated for the stationary case ($K \to \infty$) and stored as a file. When the algorithm schedules a comparison for a particular (A, B) pair, then the surface is read (or interpolated) for that point and K (a finite number of) responses are generated based on the multi-nomial distribution specified by the p_i values.

Figure B.1 depicts the procedures in the off-line model generation and the simulation-time application of the model.

Since the generation of the general model in Figure 5.6 is rather involved, we summarize its details as follows. Given the number of local optima on an operating curve, we first randomly determine the boundaries of each ROD and the position of the local optimum in it. We then generate the CND line as a continuous random walk around a given average CND value. Similarly, we generate the subjective symmetry line as a continuous random walk, when given the standard deviation of the subjective symmetry line with respect to a straight line.

In our computer generated model, we generate p_i values for a finite number of (A,B) pairs, 100 steps in A and 100 steps in B. When evaluating (A, B) pairs that do not match the 10000 points generated during the Monte-Cralo simulation, we use cubic interpolation to get smooth values.

We utilize the following principles, based on our model, when generating values for (p_{-1}, p_0, p_1, p_2) for each of the 10000 pairs of points on a given operating curve simulated.

- p_2 is monotonically non-decreasing with $B - A$ ($p_2^{\max}$ at $B - A = A^{\max} - A^{\min}$ to 0 at $B = A$).

- p_0 is monotonically non-increasing with $B - A$ (1 at $B = A$ to 0 at $B - A = CND(A)$.

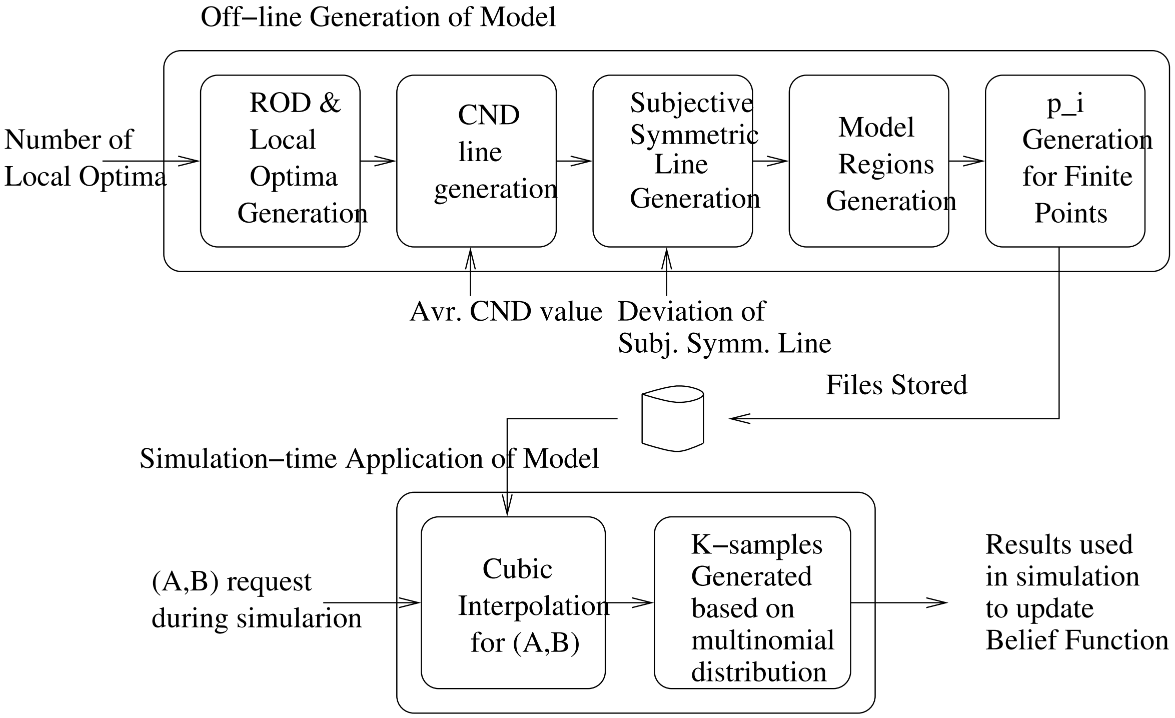

Off−line Generation of Model
Number of Local Optima
ROD & Local Optima Generation
CND line generation
Subjective Symmetric Line Generation
Model Regions Generation
p_i Generation for Finite Points
Avr. CND value
Deviation of Subj. Symm. Line
Files Stored
Simulation−time Application of Model
(A,B) request during simularion
Cubic Interpolation for (A,B)
K−samples Generated based on multinomial distribution
Results used in simulation to update Belief Function

Figure B.1: Procedures in the off-line generation and the simulation-time application of the model.

- A and B are on the same side of A_i^* and within $ROD(A_i^*)$, then $p_1 > p_{-1}$ $p_1 > p_{-1}$. p_1 is proportional to $|B - A_i^*|/(|B - A_i^*| + |A - A_i^*|)$ and p_1 is proportional to $|A - A_i^*|/(|B - A_i^*| + |A - A_i^*|)$. The ratio is relative to the distance of the other point to the local optimum. The sum of p_1 and p_{-1} equals to $1 - (p_0 + p_2)$, so they are normalized accordingly. The four probabilities are generated randomly by following a continuous trajectory.

- p_2^{max}: The maximum value of incomparability is achieved when $A^{\min}$and $A^{\max}$of the operating curve is compared.

We apply the following procedure to generate 4 surfaces to be used in the Monte-Carlo simulations:

1. One input to the general model is L, the number of local optima (LO).

2. Firstly, ROD boundaries for each of the L local optima are generated. (a) $2L$ points are generated uniformly on the operating curve [0,1], and sorted in ascending order $z_1, \ldots, z_{2L}$. $(z_i < z_{i+1})$ and there are no equalities. (b) The odd indexed z points correspond to $A_i^{\min}$ and

even indexed z points correspond to $A_i^{\max}$; e.g., for $L = 2$, $A_1^{\min} = z_1$, $A_1^{\max} = z_2$, $A_2^{\min} = z_3$ and $A_2^{\max} = z_4$.

3. The LO within each ROD is generated randomly. In order to avoid an LO being too close to the ROD boundary, we allow 10% buffer in each boundary. This is important since if there is no ROD on one side of the LO, then there are no comparisons possible to direct to the LO from that direction. $A_i^* = uniform([A_{min}^*, A_{max}^*])$ where $A_m^*in = 0.9 * A^{\min} + 0.1 * A^{\max}$ and $A_m^*ax = 0.1 * A^{\min} + 0.9 * A^{\max}$.

4. The CND(A) which is a function of A is generated along the entire operating curve (for each $A \in \mathcal{O}$). The second input to the general model is the average CND, denoted by $\overline{CND}$. It indicates the expected value for CND. Separate surfaces are generated for different CND values (e.g. 0.1 and 0.03), and results are tabulated separately.

   - Starting with $A = A^{\min} = 0$, for each step, A is increased by 0.05 and the $CND(A)$ is generated (total of 20 steps). $CND(A) = uniform([0, 2 * \overline{CND}])$. Generated CND values are verified to ensure that $CND(A_2) \geq CND(A_1) + (A_2 - A_1) - 0.05$ which means that the $B = CND(A) + A$ curve is monotonically non-decreasing. Otherwise inconsistency would happen; close-by points will be more distinguishable than farther ones (which should not happen due to Axiom 4).
   - For smaller granularities of A where CND(A) is not defined (other than the 20 points), cubic interpolation is done on 20 points to have a smooth curve.

5. The subjectively symmetric curves are generated. The third input is the variance in the subjective symmetry curve with respect to a straight line, denoted by $Var(SubjSym)$. The surfaces are generated using $Var(SubjSym) = 0.1$.

   - The $SUBJSYM(A) = SUBJSYM(A + 0.05) + Uniform([0, Var(SubjSym)])$ defined as a random walk starting at $A = A^*$ and going in reverse direction (A decreasing), iteratively defining the line for each A (total of 20 steps).
   - For smaller granularities of A where SUBJSYM(A) is not defined (other than the 20 points), cubic interpolation is done on 20 points to have a smooth curve.

6. Next, p_2 and p_0 are defined for each (A,B) since they do not depend on the location of A_i^*.

   - p_2 is non-decreasing as a function of $B - A$. Thus, a similar random walk is used to define the non-decreasing p_2 values for each $B - A$ value. $p_2 = 0$ at $B - A = 0$

and reaches $p_2 = p_2^{max}$ at $B - A = A^{\max} - A^{\min}$. The surfaces are generated using $p_2^{max} = 1$.

- Similarly, cubic interpolation smoothing is applied.
- p_0 is defined using non-increasing random walk from $p_0 = 1$ at $B - A = 0$ to $p_0 = 0$ at $B-A = CND(A)$. One random walk is generated in 20 steps from 1 to 0 and the same shape is stretched or shrunk on the $B-A$ axis to achieve $p_0 = 0$ at $B-A = CND(A)$, since $CND(A)$ is varying with A.
- Similarly, cubic-interpolation smoothing is applied.

7. Next p_1 and p_{-1} are generated within each ROD.
   - If the two points compared (A,B) are on the same side of A_i^* and within the ROD_i, then the point that is closer to A_i^* is preferred more. For example if $A_i^* < A < B$, then A is more preferred; thus, $p_1 > p_{-1}$. p_1 is proportional to $|B - A_i^*|/(|B - A_i^*| + |A - A_i^*|)$ and p_1 is proportional to $|A - A_i^*|/(|B - A_i^*| + |A - A_i^*|)$. The ratio is relative to the distance of the other point to the LO. The sum of p_1 and p_{-1} equals to $1 - (p_0 + p_2)$, so they are normalized accordingly.
   - The same is applied for points on different sides of A_i^*.
8. Finally p_1 and p_{-1} are generated outside of an ROD and across RODs. In this case multiple LO affect the relative preference of A vs. B. However, since both points are not in one ROD, there is no dominance between a particular A_i^* and the point tested outside of ROD. Thus, again the relative distance ratio is used to get the preference effects of multiple LO. The ratio presented above is multiplied for each LO. For example, p_1 is proportional to $\prod_{i=1}^{L} |A - A_i^*|/(|B - A_i^*| + |A - A_i^*|)$. The sum of p_1 and p_{-1} are again normalized so that $\sum p_i = 1$.
9. The above procedure (Steps 6,7 and 8) generates p_i values for a finite number of (A,B) pairs. 100 steps in A and 100 steps in B, 4 values for 10000 points are stored. Whenever COD needs to be generated from the model surface, cubic interpolations are done to get the smooth values for points that do not correspond to a stored point.

The algorithm is only supplied with initial values for the estimation of A^* and CND. As discussed in the procedure, at the beginning, it is assumed that there is only one LO and ROD is [0,1]. Thus, the initial estimate of A^* is 0.5. Furthermore, the initial estimate of CND is arbitrarily chosen to be 0.1. No other parameters are supplied to the algorithm, and the algorithm estimates the

ROD using Step 1, and A^* and CND using Step 2 via the Bayesian formulation. The comparison results in each batch are used to update the estimates. Only when a particular (A,B) is scheduled for comparison are the sampled probability values returned to the algorithm. The stopping criterion of the algorithm is also based on its own estimation of the belief function and CND.

Thus, the ability of the algorithm that was developed based on the principles of the simplified model to accurately estimate A^* under single and multiple local optima conditions with small number of comparisons indicate that the algorithm is indeed robust to randomness of the general model. Thus, utilizing such an algorithm in real-life subjective tests would most likely cope with the uncertainties (unknown shape of 4 surfaces) of the subjective preferences of humans on a given operating curve.

In simulating the behavior of subjects when comparing a pair of operating points, our algorithms do not know the stationary probabilities of the four opinions, similar to the real subjective tests. However, when our algorithm chooses a pair for comparison, the opinion of K subjects for that pair is returned to the algorithm. Note that in addition to the variations on the surfaces that are generated randomly, the empirical distribution exhibits noise due to the fact that a finite number of subjects would be able to evaluate the pair.

REFERENCES

[1] S. Andersen, A. Duric, H. Astrom, R. Hagen, W. Kleijn, and J. Linden, “Internet low bit rate codec (iLBC),” Dec. 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3951.txt

[2] S. A. Atungsiri, A. M. Kondoz, and B. G. Evans, “Error control for low-bit-rate speech communication systems,” *IEE Proc. I: Communications, Speech and Vision*, vol. 140, no. 2, pp. 97–103, Apr. 1993.

[3] J. C. Bolot, “Characterizing end-to-end packet delay and loss in the Internet,” *High-Speed Networks*, vol. 2, no. 3, pp. 305–323, Dec. 1993.

[4] J.-C. Bolot, S. Fosse-Parisis, and D. Towsley, “Adaptive FEC-based error control for Internet telephony,” in *Proc. IEEE INFOCOM*, vol. 3, 1999, pp. 1453–1460.

[5] L. T. Bosch, N. Oostdijk, and J. P. de Ruiter, “Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues,” in *Proceedings 7th International Conference on Text, Speech, and Dialogue*, 2004, pp. 563–570.

[6] C. Boutremans and J.-Y. L. Boudec, “Adaptive joint playout buffer and FEC adjustment for Internet telephony,” in *Proc. IEEE INFOCOM*, vol. 1, 2003, pp. 652–662.

[7] P. T. Brady, “Effects of transmission delay on conversational behaviour on echo-free telephone circuits,” *Bell System Technical Journal*, vol. 50, no. 1, pp. 115–134, Jan. 1971.

[8] C.-C. Chang and C.-J. Lin, “A library for support vector machines.” [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[9] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, “A crowdsourceable QoE evaluation framework for multimedia content,” in *MM ’09: Proceedings of the Seventeen ACM International Conference on Multimedia*, 2009, pp. 491–500.

[10] M. Chen and M. N. Murthi, “Optimized unequal error protection for voice over IP,” in *ICASSP*, 2004, pp. 865–868.

[11] S. Coren, L. M. Ward, and J. T. Enns, *Sensation and Perception, 4th ed.* New York, NY: Harcourt Brace Jovanovich, 1994.

[12] A. Eichhorn, P. Ni, and R. Eg, "Randomised pair comparison: an economic and robust method for audiovisual quality assessment," in *NOSSDAV '10: Proceedings of the 20th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2010, pp. 63–68.

[13] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. New York, NY: Springer, 1972.

[14] A. A. E. Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Transactions on Information Theory*, vol. 28, pp. 851–857, Nov. 1982.

[15] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Processing Magazine*, pp. 74–93, Sep. 2001.

[16] M. Graubner, P. S. Mogre, R. Steinmetz, and T. Lorenzen, "A new qoe model and evaluation method for broadcast audio contribution over ip," in *NOSSDAV '10: Proceedings of the 20th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2010, pp. 57–62.

[17] K. Haberlandt, *Cognitive Psychology*. Needham Heights, MA: Allyn and Bacon, 1994.

[18] Z. X. Huang, B. Sat, and B. W. Wah, "Automated learning of play-out scheduling algorithms for improving the perceptual conversational quality in multi-party VoIP," in *Proc. IEEE Int'l Conf. on Multimedia and Expo*, Jul. 2008, pp. 493–496.

[19] IETF, "RFC 791, Internet Protocol: DARPA Internet program protocol specification," Sep. 1981, http://www.ietf.org/rfc/rfc791.txt.

[20] International Telecommunication Union, "ITU-T G-Series recommendations." [Online]. Available: http://www.itu.int/rec/T-REC-G/en

[21] ——, "ITU-T P-Series recommendations." [Online]. Available: http://www.itu.int/rec/T-REC-P/en

[22] ITU-G.107, "The E-model, a computational model for use in transmission planning." [Online]. Available: http://www.itu.int/rec/T-REC-G.107/en

[23] ITU-G.1070, "Opinion model for video-telephony applications." [Online]. Available: http://www.itu.int/rec/T-REC-G.1070-200704-I

[24] ITU-G.722.2, "Wideband coding of speech at around 16 kbit/s using adaptive multi-rate wideband (AMR-WB)." [Online]. Available: http://www.itu.int/rec/T-REC-G.722.2/en

[25] ITU-G.729.1, "G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729." [Online]. Available: http://www.itu.int/rec/T-REC-G.729.1/en

[26] ITU-P.561, "In-service non-intrusive measurement device: Voice service measurements." [Online]. Available: http://www.itu.int/rec/T-REC-P.561/en

[27] ITU-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.” [Online]. Available: http://www.itu.int/rec/T-REC-P.862/en

[28] ITU-T, “Study group 12, question 20: Objective assessment of conversational speech quality in networks.” [Online]. Available: http://www.itu.int/ITU-T/2005-2008/com12/sg12-q20.html

[29] N. S. Jayant and S. W. Christensen, “Effects of packet losses in waveform coded speech and improvements due to odd-even sample-interpolation procedure,” *IEEE Transactions on Communications*, vol. 29, no. 2, pp. 101–110, Feb. 1981.

[30] W. Jiang and A. Ortega, “Multiple description coding via polyphase transform and selective quantization,” in *Proc. Visual Communications and Image Processing*, vol. 3653, Dec. 1998, pp. 998–1008. [Online]. Available: http://novel.crhc.uiuc.edu/papers.db/j/JiaOrt98.ps.gz

[31] W. Jiang and H. Schulzrinne, “Modelling of packet loss and delay and their effect on real-time multimedia service quality,” in *Proc. Int. Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, 2000.

[32] N. Kiatawaki and K. Itoh, “Pure delay effect on speech quality in telecommunications,” *IEEE Journal on Selected Areas of Communication*, vol. 9, no. 4, pp. 586–593, May 1991.

[33] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. Int’l Joint Conf. on Artificial Intelligence*, 1995, pp. 1137–1145.

[34] T. J. Kostas, M. S. Borella, I. Sidhu, G. M. Schuster, J. Grabiec, and J. Mahler, “Real-time voice over packet-switched networks,” *IEEE Network*, vol. 12, no. 1, pp. 18–27, January-February 1998.

[35] P. L. Lanzi, “Learning classifier systems: Then and now,” *Evolutionary Intelligence*, vol. 1, pp. 63–82, 2008.

[36] Y. J. Liang, N. Faber, and B. Girod, “Adaptive playout scheduling and loss concealment for voice communication over IP networks,” *IEEE Trans. on Multimedia*, vol. 5, no. 4, pp. 532–543, Dec. 2003.

[37] D. Lin, *Loss Concealments for Low Bit-Rate Packet Voice*. Urbana, IL: Ph.D. Thesis, Dept. of Electrical and Computer Engineering, Univ. of Illinois, Aug. 2002.

[38] D. Lin and B. W. Wah, “LSP-based multiple-description coding for real-time low bit-rate voice over IP,” *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 167–178, Feb. 2005.

[39] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, “Assessing the quality of voice communications over Internet backbones,” *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 747–760, 2003.

[40] S. B. Moon, J. Kurose, and D. Towsley, “Packet audio playout delay adjustment: Performance bounds and algorithms,” *Multimedia Systems*, vol. 6, no. 1, pp. 17–28, Jan. 1998.

[41] E. Muzychenko, “Virtual audio cable.” [Online]. Available: http://nrcde.ru/music/software/eng/vac.html

[42] I.-T. P.910, “Subjective video quality assessment methods for multimedia applications,” 1999.

[43] C. Perkins, O. Hodson, and V. Hardman, “A survey of packet loss recovery techniques for streaming audio,” *IEEE Network*, vol. 12, pp. 40–48, Sept.-Oct. 1998.

[44] d. PlanetLab: An open platform for developing and accessing planetary-scale services, “http://www.planet-lab.org/.” [Online]. Available: http://www.planet-lab.org/

[45] R. Ramjee, J. Kurose, D. Towsley, and H. Schulzrinne, “Adaptive playout mechanisms for packetized audio applications in wide-area networks,” in *Proc. 13th Annual Joint Conf. IEEE Computer and Communications Societies on Networking for Global Commmunication*, vol. 2, 1994, pp. 680–688.

[46] D. L. Richards, *Telecommunication by Speech*. London, UK: Butterworths, 1973.

[47] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” *Language*, vol. 50, pp. 696–735, 1974. [Online]. Available: http://novel.crhc.uiuc.edu/papers.db/h/HSEASGJ74.pdf

[48] B. Sat, Z. X. Huang, and B. W. Wah, “The design of a multi-party VoIP conferencing system over the Internet,” in *Proc. IEEE Int’l Symposium on Multimedia*, Taichung, Taiwan, Dec. 2007, pp. 3–10.

[49] B. Sat and B. W. Wah, “Speech- and network-adaptive layered G.729 coder for loss concealments of real-time voice over IP,” in *IEEE Int’l Workshop on Multimedia Signal Processing*, Oct. 2005.

[50] ——, “Speech-adaptive layered G.729 coder for loss concealments of real-time voice over IP,” in *Proc. IEEE Int’l Conf. on Multimedia and Expo*, Jul. 2005.

[51] ——, “Analysis and evaluation of the Skype and Google-Talk VoIP systems,” in *Proc. IEEE Int’l Conf. on Multimedia and Expo*, Jul. 2006.

[52] ——, “Evaluation of conversational voice quality of the Skype, Google-Talk, Windows Live, and Yahoo Messenger VoIP systems,” in *IEEE Int’l Workshop on Multimedia Signal Processing*, Oct. 2007.

[53] ——, “Playout scheduling and loss-concealments in VoIP for optimizing conversational voice communication quality,” in *Proc. ACM Multimedia*, Augsburg, Germany, Sep. 2007, pp. 137–146.

[54] ——, “Statistical testing of off-line comparative subjective evaluations for optimizing perceptual conversational quality in voip,” in *Proc. IEEE Int’l Symposium on Multimedia*, Dec. 2008, pp. 424–431.

[55] ——, “Analyzing voice quality in popular voip applications,” *IEEE Multimedia*, vol. 16, pp. 46–58, Jan-Mar 2009.

[56] ——, “Statistical scheduling of offline comparative subjective evaluations for real-time multimedia,” *IEEE Trans. on Multimedia*, vol. 11, no. 6, pp. 1114–1130, Oct. 2009.

[57] N. Shacham and P. McKenney, “Packet recovery in high-speed networks using coding and buffer management,” in *Proc. of IEEE INFOCOM*, May 1990, pp. 124–131.

[58] G. I. Solutions, “Global ip solutions codecs, iLBC, iSAC, iPCM-wb, enhanced G.711.” [Online]. Available: http://www.gipscorp.com/files/english/datasheets/Codecs.pdf

[59] L. Sun and E. Ifeachor, “New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks,” in *Proc. IEEE Communications*, vol. 3, 2004, pp. 1478–1483.

[60] ——, “Voice quality prediction models and their applications in VoIP networks,” *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 809–820, 2006.

[61] J. Suzuki and M. Taka, “Missing packet recovery techniques for low-bit-rate coded speech,” *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 5, pp. 707–717, Jun. 1989.

[62] A. Takahashi, H. Yoshino, and N. Kitawaki, “Perceptual QoS assessment technologies for VoIP,” *IEEE Communications Magazine*, vol. 42, no. 7, pp. 28–34, Jul. 2004.

[63] R. C. F. Tucker and J. E. Flood, “Optimizing the performance of packet-switch speech,” in *IEEE Conf. on Digital Processing of Signals in Communications*, Loughborough University, Apr. 1985, pp. 227–234.

[64] J. Wang and J. D. Gibson, “Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks,” in *IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, pp. 745–748.

[65] J. F. Wang, J. C. Wang, J. F. Yang, and J. J. Wang, “A voicing-driven packet loss recovery algorithm for analysis-by-synthesis predictive speech coders over Internet,” *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 98–107, Mar. 2001.

[66] K. Weilhammer and S. Rabold, “Durational aspects in turn taking,” in *Proc. Int. Conf. on Phonetic Sciences*, 2003. [Online]. Available: http://novel.crhc.uiuc.edu/papers.db/k/KWSR03.pdf

[67] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang, "Quality of experience in distributed interactive multimedia environments: toward a theoretical framework," in *MM '09: Proceedings of the Seventeen ACM International Conference on Multimedia*, 2009, pp. 481–490.

[68] M. Yajnik, S. Moon, J. Kurose, and D. Towsley, "Measurement and modelling of temporal dependence in packet loss," in *Proc. IEEE INFOCOM*, march 1999, pp. 345–352.