Robust Multimedia Communications over Packet Networks

Guoqiang Zhang  张国强

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)

Stockholm 2010
Abstract

Multimedia communications over packet networks, and in particular the voice over IP (VoIP) application, have become an integral part of society. However, the unreliable and heterogeneous nature of packet networks has led to a best-effort delivery of services. Delay, limitation of bandwidth, and packet-loss rate all affect the quality of service (QoS). In this thesis, we address two important network impairments in the design of robust multimedia communication systems: packet delay-variation and packet-loss.

Paper A considers the mitigation of the effect of packet delay-variation for audio communications by introducing a buffer at the receiver side. A new adaptive playout scheduling approach is proposed to control the buffering length, or, equivalently, the packet playout deadlines, in response to varying network conditions. A Wiener process is used to model the fluctuation of the buffering length without any playout adjustment. The playout scheduling problem is then reformulated as a stochastic impulse control problem by taking the playout adjustment as the control signal. The proposed approach is shown to be the optimal solution to the new control problem. It is demonstrated experimentally that the proposed approach provides improved perceived conversional quality.

Papers B, C and D address the packet-loss issue. Paper B focuses on the design of a low-complexity packet-loss concealment (PLC) method that is compatible with existing speech codecs for VoIP application. The new method is rigorously motivated based on the autoregressive (AR) speech model and the minimum mean squared error (MMSE) criterion. The effect of model estimation error on the prediction of the missing speech segment is also considered and an upper bound for the prediction error is derived. Both the theoretical and experimental results provide insight in the performance of the heuristically designed PLC methods. On the other hand, Paper C and D consider an active packet-loss-resilient coding scheme, namely multiple description coding (MDC). In general, MDC can be used for the transmission of any media data. Paper C derives a simple and accurate approximation of the rate-distortion lower bound of a particular multiple-description scenario and then demonstrates that the performance loss of some practical MD systems can be evaluated easily with the new approximation. Paper D studies the performance limit of a vector Gaussian multiple description scenario. An outer bound to the rate-distortion region is derived, and the outer bound is tight when the problem specializes to the scalar Gaussian case.

Keywords: playout scheduling, VoIP, autoregressive model, packet-loss concealment, multiple description coding, vector quantization, rate-distortion region.
List of Papers

The thesis is based on the following papers:


In addition to papers A-D, the following papers have also been produced in part by the author of the thesis:


Acknowledgements

During my Ph.D. studies I received a lot of help from many people. First I would like to take this opportunity to thank my supervisor Prof. Bastiaan Kleijn. You have always been supportive to my work and have done your utmost to guide me, which I deeply appreciate. Your creativity, dedication and professionalism inspired me in both work and life.

I devote special thanks to Prof. Gyorgy Dán, Dr. Jan Østergaard, Haopeng Li, Janusz Klejsa and Henrik Lundin for work resulting in joint publications. It is an enjoyable experience to work with you all.

I would also like to thank all my colleges at the Sound and Image Processing lab. Thank you for creating a pleasant working environment. In particular, I am thankful to: Anders Ekman for helping me in preparing tutorials in Information theory and Source Coding course; Prof. Markus Flierl for valuable research discussions; Prof. Arne Leijon for proof-reading my thesis; my office-mate Janusz Klejsa for so many interesting conversations about culture, history, life, and so on; Jan Plasberg and Konrad Hofbauer for teaching me to play table football.

Thanks to my Chinese friends at KTH, particularly David, Zhanyu, Zhongwei, Xi Zhang, Jing Fu, Minyue, Jingfen and Lei Bao. I am glad to get to know you. I will always remember the fun we have had together.

I would like to express my deep gratitude to my family, my parents for your endless support and your sacrifice; my two sisters for the joy we have. Last but certainly not least, thank you Xiaoying for your understanding and encouragement.

Guoqiang Zhang, Stockholm, 2010
Contents

Abstract i
List of Papers iii
Acknowledgements v
Contents vii
Acronyms xi

I Introduction 1
1 Motivation ........................................ 1
1.1 Background ...................................... 1
1.2 Research Challenges ............................ 3
2 Approaches Addressing Delay-Jitter .......... 6
  2.1 Overview of Playout Scheduling for Audio Communication ............... 6
  2.2 Overview of Playout Scheduling for Video Communication ............... 8
3 Approaches Addressing Packet-Loss ......... 8
  3.1 Packet-Loss Concealment ...................... 9
  3.2 Multiple Description Coding ................... 10
4 Summary of Contributions ..................... 24
References ........................................ 26

II Included papers 39

A Adaptive Playout Scheduling for Voice over IP: Event-
  Triggered Control Policy A1
  1 Introduction .................................. A1
  2 Dejitter Buffer Modelling .................... A3
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Wiener Process Modelling</td>
<td>A4</td>
</tr>
<tr>
<td>2.2</td>
<td>Impulse Control Adjustment</td>
<td>A4</td>
</tr>
<tr>
<td>3</td>
<td>Impulse Control of the Dejitter Buffer</td>
<td>A6</td>
</tr>
<tr>
<td>3.1</td>
<td>Problem Formulation</td>
<td>A6</td>
</tr>
<tr>
<td>3.2</td>
<td>A Band Control Policy and its Performance</td>
<td>A8</td>
</tr>
<tr>
<td>4</td>
<td>Optimal Control Policy</td>
<td>A11</td>
</tr>
<tr>
<td>4.1</td>
<td>Lower Bound over All Impulse Control Policies</td>
<td>A11</td>
</tr>
<tr>
<td>4.2</td>
<td>Optimality of a Band Control Policy</td>
<td>A11</td>
</tr>
<tr>
<td>5</td>
<td>Implementing the Band Control Policy</td>
<td>A13</td>
</tr>
<tr>
<td>5.1</td>
<td>Negative Jump of the Buffer Length</td>
<td>A14</td>
</tr>
<tr>
<td>5.2</td>
<td>Positive Jump of the Buffer Length</td>
<td>A14</td>
</tr>
<tr>
<td>5.3</td>
<td>Packet loss and out-of-order arrivals</td>
<td>A15</td>
</tr>
<tr>
<td>5.4</td>
<td>Summarizing The Band-Control Algorithm</td>
<td>A16</td>
</tr>
<tr>
<td>5.5</td>
<td>Selection of the Upper Bound</td>
<td>A17</td>
</tr>
<tr>
<td>6</td>
<td>Experimental Results</td>
<td>A18</td>
</tr>
<tr>
<td>6.1</td>
<td>Performance Evaluation using Synthetic Delay Traces</td>
<td>A20</td>
</tr>
<tr>
<td>6.2</td>
<td>Performance Evaluation using Measured Delay Traces</td>
<td>A22</td>
</tr>
<tr>
<td>7</td>
<td>Conclusion</td>
<td>A25</td>
</tr>
<tr>
<td>Appendix I.</td>
<td>Proof of Proposition 2</td>
<td>A25</td>
</tr>
<tr>
<td>Appendix II.</td>
<td>Proof of Proposition 3</td>
<td>A26</td>
</tr>
<tr>
<td>Appendix III.</td>
<td>Proof of Theorem 1</td>
<td>A26</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**B Autoregressive Model-Based Speech Packet-Loss Conceal-ment**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>B1</td>
</tr>
<tr>
<td>2</td>
<td>Signal Model</td>
<td>B3</td>
</tr>
<tr>
<td>2.1</td>
<td>Basic Signal Model</td>
<td>B3</td>
</tr>
<tr>
<td>2.2</td>
<td>Extension to Long Term Correlations</td>
<td>B4</td>
</tr>
<tr>
<td>3</td>
<td>MMSE Missing Segment Prediction</td>
<td>B4</td>
</tr>
<tr>
<td>3.1</td>
<td>MMSE Prediction</td>
<td>B4</td>
</tr>
<tr>
<td>3.2</td>
<td>Effect of Model Estimation Error on Prediction</td>
<td>B5</td>
</tr>
<tr>
<td>3.3</td>
<td>Why perceptual weighting does not affect Prediction</td>
<td>B7</td>
</tr>
<tr>
<td>4</td>
<td>Results</td>
<td>B8</td>
</tr>
<tr>
<td>4.1</td>
<td>Experimental MSE Bound Verification</td>
<td>B8</td>
</tr>
<tr>
<td>4.2</td>
<td>Subjective Quality Comparison</td>
<td>B8</td>
</tr>
<tr>
<td>5</td>
<td>Conclusion</td>
<td>B10</td>
</tr>
<tr>
<td>Appendix I.</td>
<td>Proof of Proposition 1</td>
<td>B10</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### C High-Rate Analysis of Symmetric L-Channel Multiple Description Coding

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>C1</td>
</tr>
<tr>
<td>2. Analysis of Multiple Description Lower Bound</td>
<td>C4</td>
</tr>
<tr>
<td>2.1 Preliminaries</td>
<td>C5</td>
</tr>
<tr>
<td>2.2 Duality of the Inner-Tight and Outer-Tight expressions</td>
<td>C6</td>
</tr>
<tr>
<td>2.3 Asymptotic Tight Lower Bound to the Rate-Distortion Function</td>
<td>C10</td>
</tr>
<tr>
<td>3. Evaluation of MDLVQ Systems</td>
<td>C15</td>
</tr>
<tr>
<td>3.1 System Settings</td>
<td>C16</td>
</tr>
<tr>
<td>3.2 Index Assignment of a MDLVQ</td>
<td>C18</td>
</tr>
<tr>
<td>3.3 The Geometry of $L$-tuples</td>
<td>C20</td>
</tr>
<tr>
<td>3.4 Asymptotic Analytic Performance</td>
<td>C22</td>
</tr>
<tr>
<td>4. Conclusion</td>
<td>C27</td>
</tr>
</tbody>
</table>

### D Bounding the Rate Region of Vector Gaussian Multiple Descriptions with Individual and Central Receivers

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>D1</td>
</tr>
<tr>
<td>2. Problem Formulation and Main Results</td>
<td>D3</td>
</tr>
<tr>
<td>2.1 Problem Formulation</td>
<td>D3</td>
</tr>
<tr>
<td>2.2 Gaussian Description Scheme</td>
<td>D4</td>
</tr>
<tr>
<td>2.3 Outer Bound to the Rate Region</td>
<td>D6</td>
</tr>
<tr>
<td>2.4 Optimal Weighted Sum Rate</td>
<td>D7</td>
</tr>
<tr>
<td>2.5 Rate Region for the Scalar Gaussian Source</td>
<td>D10</td>
</tr>
<tr>
<td>3. Proof of Theorem 1</td>
<td>D11</td>
</tr>
<tr>
<td>4. Proof of Theorem 2</td>
<td>D13</td>
</tr>
<tr>
<td>5. Rate region for Scalar Gaussian Source</td>
<td>D17</td>
</tr>
<tr>
<td>6. Conclusion</td>
<td>D22</td>
</tr>
<tr>
<td>Appendix I. Useful Matrix Lemmas</td>
<td>D22</td>
</tr>
<tr>
<td>Appendix II. Proof of Lemma 4</td>
<td>D23</td>
</tr>
<tr>
<td>Appendix III. Proof of Lemma 2</td>
<td>D25</td>
</tr>
<tr>
<td>Appendix IV. Proof of Lemma 5</td>
<td>D27</td>
</tr>
<tr>
<td>Appendix V. Proof of Lemma 3</td>
<td>D31</td>
</tr>
<tr>
<td>References</td>
<td>D33</td>
</tr>
</tbody>
</table>

### References

ix
## Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>Auto-Regressive</td>
</tr>
<tr>
<td>ARQ</td>
<td>Automatic Repeat reQuest</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
</tr>
<tr>
<td>DCR</td>
<td>Degradation Category Rating</td>
</tr>
<tr>
<td>FEC</td>
<td>Forward Error Correction</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile Communications</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>independent and identically-distributed</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>ISDN</td>
<td>Integrated Services Digital Network</td>
</tr>
<tr>
<td>LP</td>
<td>Linear Prediction</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Prediction Coefficient</td>
</tr>
<tr>
<td>MDC</td>
<td>Multiple Description Coding</td>
</tr>
<tr>
<td>MDSQ</td>
<td>Multiple Description Scalar Quantization</td>
</tr>
<tr>
<td>MDLVQ</td>
<td>Multiple Description Lattice Vector Quantization</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MMSE</td>
<td>Minimum Mean Square Error</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODE</td>
<td>Ordinary Differential Equation</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse-Code Modulation</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>PESQ</td>
<td>Perceptual Evaluation of Speech Quality</td>
</tr>
<tr>
<td>PLC</td>
<td>Packet Loss Concealment</td>
</tr>
<tr>
<td>PSTN</td>
<td>Public Switched Telephone Network</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RTP</td>
<td>Real-time Transport Protocol</td>
</tr>
<tr>
<td>SCN</td>
<td>Switched Communication Network</td>
</tr>
<tr>
<td>UDP</td>
<td>User Datagram Protocol</td>
</tr>
<tr>
<td>VoIP</td>
<td>Voice over Internet Protocol</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantizer</td>
</tr>
<tr>
<td>WGN</td>
<td>White Gaussian Noise</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission Control Protocol</td>
</tr>
</tbody>
</table>
Part I

Introduction
Introduction

1 Motivation

1.1 Background

With the expansion of broadband infrastructure, multimedia communications over packet-switched networks such as the Internet have become increasingly popular. Multimedia communications can be either one-way from one transmitter to one or more receivers (e.g., audio/video streaming), or interactive multi-way (including two-way) between multiple communicating parties (e.g., VoIP, video conferencing). Fig. 1 shows a demonstration of multimedia communication over IP networks. The Switched Communication Network (SCN) can be a wired or wireless network, such as PSTN, ISDN or GSM. Some particular applications experience a phenomenal growth of popularity, and are likely to replace traditional communication systems such as landline phones. One example is that the new generation of mobile phones (e.g., the iphone) is built with VoIP compatible systems. A driving impetus is the user-demand for easily accessible and inexpensive services with a satisfactory quality. However, the transmission bandwidth is limited while the traffic load and user-demand for communication quality are increasing. This is a fundamental challenge in multimedia communications.

Figure 1: Multimedia communication environment.
Conceptually, a multimedia communication link consists of an encoding system, a network and a decoding system, as shown in Fig. 2. The media data is first compressed and packetized by the encoding system, then delivered over the network, and finally estimated by the decoding system based on what was received from the network. There are two causes of degradation of the communication quality. The first is the coding distortion and the second is damaged or lost data due to network impairments. The coding distortion is determined by the coding system given the available bandwidth. On the other hand, the technological challenges stemming from the network impairments are a major barrier for providing a guaranteed performance with respect to packet-loss, delay and bandwidth [95], [104], [131], [48], [135], [21]. Thus, it is of great interest to understand the characteristics of packet-switched networks and thereafter design robust multimedia communication systems.

The unreliability and heterogeneity of packet-switched networks result in a best-effort service. The packets transmitted over packet-switched networks suffer from network delay, network packet-loss, network delay variation (also known as delay-jitter) and various connection speeds (or bandwidths). We provide an overview of these impairments in more detail in the following.

Network packet-loss occurs due to the congestion in network nodes (wired) and/or the corrupted packet checksum (wireless). The effect of packet-loss is a major issue for multimedia communications. When media packets are lost, the playout is interrupted, leading to a poor user experience. The automatic repeat request (ARQ) technique is in general not suitable for multimedia communications due to a low-latency service demand. Special treatment is required to mitigate the effect of packet-loss. Besides the packets being dropped or corrupted by the network, packet-loss could be a result of the compensation of delay-jitter, since the packets received after the intended playout time are not useful.

Delay-jitter is caused by dynamic data traffic, or possibly the congestion of the network. Delay-jitter can be compensated at the cost of increased communication-delay. The additional delay introduced by combating delay-jitter is irrelevant for most one-way transmissions but is crucial for interactive communications [87].
This quality of service (QoS) limitation poses a challenge for multimedia communications. Robust multimedia communications generally demand low delay, high bandwidth and low packet-loss rates in order to provide acceptable quality for the end users. The research goal of this thesis is to develop methods that address network impairments, leading to robust multimedia communications.

1.2 Research Challenges

This thesis focuses on two network impairments: delay-jitter and packet-loss. These impairments are recognized to be central to the degradation of the communication quality. This subsection introduces and identifies the research questions that must be answered to resolve the problems resulting from the delay-jitter and the packet-loss.

Delay-Jitter Impairment

Packet delay-jitter refers to the variation in packet network delay. Due to the variable nature of dynamic data traffic over a network, the statistics of delay-jitter in general change over time. An example of this behavior is the existence of delay spikes [92]. A delay spike represents the phenomena that a set of consecutive packets arrives in a very short time. A segment of network delay trace is shown in Fig. 3.

In general, delay-jitter has a large effect on the perceived quality of the interactive communication and a relatively small effect on the perceived
quality of the one-way transmission. We now discuss the effect of delay-jitter in detail. To compensate for delay-jitter, a typical multimedia communication system introduces a buffer (known as dejitter buffer) at the receiver side. The dejitter buffer temporally stores the incoming packets before playing them out, allowing slower packets to arrive in time. The basic idea behind the operation is to introduce additional buffering-delay to even out the delay-jitter. The packet end-to-end playout delay is thus the summation of the network delay and the buffering-delay. The packets that miss the scheduled playout deadline are considered lost. The total packet-loss consists of the packets being dropped or corrupted by the network and the packets missing their delivery deadline. The length of the introduced buffering-delay determines the packet-loss rate due to late arrival. As the buffering-delay increases, more media packets arrive before their scheduled playout time, resulting in a decreasing packet-loss rate.

Generally, for one-way multimedia transmissions, a long buffering-delay can be introduced since the transmission involves little interactivity between the communicating parties. Thus, the delay-jitter affects the quality little in these applications. On the other hand, interactive communications are sensitive to delay. In this case, a short buffering-delay is desirable to enhance the communication interactivity. Another special application which also requires low delay is live streaming (e.g., transmission of a realtime football match over a network). The end users generally expect the immediate playout of media data after transmission [52]. Compared with interactive communications, live streaming is relatively flexible with regard to delay. To briefly summarize, in designing a dejitter buffer for interactive communications and live streaming an inevitable trade-off must be made between correcting the effect of delay-jitter and maintaining a desirable low-delay. The research goal is to design an efficient buffer control method that attempts at achieving the best possible trade-off between delay and packet-loss.

Packet-Loss Impairment

Packet-loss causes information loss, which results in, for instance, non-fluency in audio or freezing frames in video. The philosophy for combating packet-loss in the scenario of multimedia communications is that end-users are more willing to accept a degraded quality than a long service delay. This philosophy is different from that in the scenario of traditional data transmissions (e.g., FTP, email), where no distortion or error is allowed. Packet-delay is generally not critical in traditional data transmissions, and packet-retransmission can efficiently cope with packet-losses.

One strategy to combat packet-loss is to apply the techniques of packet-loss concealment at the receiver side. In general, the media content is temporarily (e.g., audio and video) and/or spatially (e.g., image and video) correlated. As a result, neighboring packets share redundant information.
This facilitates the partial recovery of information in the lost packets by exploiting the redundant information that is embedded in neighboring packets. Packet-loss concealment methods are designed to be compatible with existing coding systems. Specifically, a module of packet-loss concealment (PLC) is built and integrated with a decoding system, facilitating reasonable performance of the coding system in an environment with packet-loss. The research goal is to design an effective and compatible PLC method to optimize the perceived quality of media presentation.

The other strategy is to design packet-loss-resilient coding systems. This strategy requires the encoder to play a primary role while in the first strategy only the decoder is involved. To address the packet-loss, the encoder produces a higher bit-rate than required for reconstructing the media data. In other words, the encoder introduces redundant information to the media data with the intention to mitigate the effect of packet-loss. This strategy is able to provide stronger robustness than the first strategy towards packet-loss for multimedia communications on packet switched networks.

One natural scheme is to adopt the forward error correction (FEC) across neighboring packets (see [10], [96], [22] for audio transmission and [34], [69], [105] for video transmission), referred to as the FEC scheme. The recovery of lost packets is performed at the cost of high-latency and high bandwidth.

A more advanced scheme is to generate a set of packets of a media data and then send them separately through a network. By applying this scheme, the probability of losing all the information of the data is low, leading to a relatively reliable transmission. The set of packets can be delivered through either one transmission link via time-shifting or different transmission links in a parallel manner. Conceptually, this scheme exploits path-diversity in the design of packet-loss-resilient coding systems. The transmission of each packet can be virtually viewed as that the packet is delivered through a separate path (or channel).

A simple way to achieve robustness is to duplicate every media packet and then send them through a network. If at least one packet is received, the media quality is maintained. If no packet is successfully delivered, the quality degradation is the same as that of the one-packet transmission. In this situation, the robustness via diversity is achieved at the cost of doubled bandwidth. A natural question one may ask is if the media quality can be improved upon receiving two packets through advanced coding. This is a source coding problem. Formally stated, the idea of compression efficiency w.r.t. bit-rate and quality between a number of packets (or descriptions) of a source is referred to as the multiple-description (MD) problem. Multiple description coding (MDC) allows a graceful quality degradation with decreasing number of received descriptions, which is highly desirable for multimedia communications.

Recent studies showed that the MDC scheme performs better than the FEC scheme through some particular applications [22], [20], [56]. In this
The dejitter buffer plays an important role in multimedia communications. It reduces the number of playout interruptions due to delay-jitter. In particular, the dejitter buffer reduces the sensitivity of the communication systems to fluctuations in the packet network-delay by waiting for packets that have long network delays. The protection of the playout continuity comes at the cost of prolonged latency which may negatively affect the performance of the applications with strict delay-constraints. The technical challenge sits in the adaptation of the buffer length in response to varying network conditions so as to find the best compromise between playout interruption and the communication delay for the given scenario, referred to as adaptive playout scheduling. In the following, we provide a literature review of adaptive playout scheduling approaches for both audio and video communications, respectively.

2.1 Overview of Playout Scheduling for Audio Communication

Adaptive playout scheduling for audio is the subject of a significant ongoing research effort because of the popularity of the IP telephony. Many methods have been proposed in the last two decades. The basic principle of the reported methods is to estimate the statistics of recent packet-network delays and then determine the buffer-length or equivalently the packet playout deadline for the next communication unit (e.g., a talkspurt or a block of packets). The research objective is to minimize the packet-loss rate due to late-arrivals while keeping a short buffering delay.

Depending on the particular statistics being estimated, the methods can be classified into different groups. Some of the methods estimate the mean and the variation in the network delay to calculate the playout deadlines [92], [80]. Some methods estimate the cumulative distribution function (CDF) of the packet network-delay from a recent history in the calculation of the deadlines, which are either histogram-based [73], [78], [88], [100] or parametric-PDF based [9], [31], [103]. In particular, the Weibull distribution [103], the Pareto distribution [31] and the Exponential distribution [9] are three widely exploited distributions in parametric-PDF modelling. However, the above methods fail to model network delay statistics when packet delay-spikes happen [103]. Special treatments are usually designed to deal with the delay-spikes to make the methods reliable. Some methods exploit stochastic processes to model the packet-arrivals. Specifically, in [70] the packet-arrivals are modelled as a combination of several Poisson pro-
cesses. However, the modelling accuracy becomes low when the delay-jitter increases.

Given a statistical model of the network packet-delay, different approaches for playout adjustment have been proposed in the literature. A first approach imposes the buffer-size adjustment without regarding the media signal (e.g., the talk spurt character of human speech) [75]. This approach may result in unnaturalness of the output audio due to, for example, the drop of consecutive voiced packets. A second approach adjusts the playout deadline at the beginning every talk spurt ([92], [103], [72]). The approach artificially elongates or shortens the silence duration before active speech. In general, this approach does not affect the listening quality, but it has limited effectiveness when talkspurts are long. A third approach exploits the pitch correlations of active speech. It stretches or compresses the active speech segment on a per-packet basis (e.g., [76], [73]). The difference between the processed speech segment duration and the original segment duration is constrained to be an integer multiple of the pitch period. This approach in general introduces degradation of the listening quality, but it gains flexibility in playout adjustment.

Informally, every adaptive playout scheduling method can produce a trade-off curve w.r.t. the total packet-loss rate and the packet playout-delay. From a QoS perspective, the choice of the optimal point on the trade-off curve should be determined by the perceived communication quality. This poses another question as how to estimate the communication quality.

Recently many studies have been conducted on measuring the overall communication quality between two parties [103], [79], [49] [72], [3]. In particular, a so-called E-model of the overall communication quality was proposed by the ITU-T union [49]. The model (in terms of a quality rating factor $R$) captures various impairments in estimating the speech-transmission quality in a form of linear combination of individual impairments. The $R$ factor ranges from 100 down to 0, where 100 is excellent and 0 is poor, of which the expression is given as

$$R = R_0 - I_s - I_d - I_e + A.$$  \hspace{1cm} (1)

In the above expression, $R_0$ represents the basic signal-to-noise ratio. $I_s$ represents a combination of impairments that occur simultaneously with the voice signal, which could be the error in the loudness level, a non-optimum sidetone, and distortion resulting from quantization. $I_d$ represents the impairments caused by delay. $I_e$, the effective equipment factor, represents the impairments caused by low bit-rate codecs and packet-loss. Finally, the advantage factor $A$ compensates for impairment factors when access advantages are available to the end user. The $R$ factor can be converted to a MOS score of conversational quality [103].

To apply the E-model in voice playout scheduling, one only needs to focus on the two factors $I_d$ and $I_e$. The other factors are unaffected by
the dejitter buffer. It is worth noting that the two factors $I_d$ and $I_e$ are nonlinear functions of delay and packet-loss, respectively (e.g., [103]). In principle, if a method can provide a trade-off curve lower than another one produced by another method, then the method should outperform the other method also in terms of the overall quality criteria.

2.2 Overview of Playout Scheduling for Video Communication

Playout scheduling for video communication is required not only for interactive communication but also for live streaming. For the application of live streaming, the end users generally expect the immediate playout of media data after transmission [52]. Thus, similar to interactive communication, live streaming also requires low playout latency. As no interactivity is essentially present in live streaming, the playout delay constraint is not as strict as that of interactive communication.

A number of playout scheduling methods have been developed to compensate for jitter in video communications. Some methods adjust the playout time by either slowing down or speeding up the playout speed [53, 54, 67, 68, 101]. The control of playout speed is realized by scaling the duration that each video frame is shown. Recently a new method has been proposed, which performs content-aware playout adjustment [71]. The idea is to vary the playout speed of frames, based on the frame content. For instance, frames with low or no motion might be less affected by playout adjustment than frames with high motion.

In the application of video communication, the audio/video synchronization is another important technique for providing high perceived quality by end users. If the synchronization is not properly addressed, the playout of the media data can be annoying, leading to a poor end-user experience. Several methods on synchronizing audio and video content have been proposed in the literature [25, 33, 63, 74, 132].

3 Approaches Addressing Packet-Loss

Packet-loss is another major issue in the design of robust multimedia communications. In the past years considerable attention has been devoted to mitigate the effect of packet-loss. Some approaches focus on recovering missing packets only at the receiver side. They can be viewed as passive approaches since only the receiver is involved. We refer to them as packet-loss concealment (PLC) approaches. Other approaches attempt to design packet-loss-resilient systems that involve both the transmitter and the receiver. Accordingly, they can be viewed as active approaches as compared to the receiver-oriented approaches.
As is mentioned in subsection 1.2, there are two popular schemes in the design of packet-loss-resilient systems: one is the FEC scheme and the other is the MDC scheme. The MDC scheme exploits path-diversity to achieve robust transmission. Our attention will be restricted to the MDC scheme as our work on the design of packet-loss-resilient systems in the thesis falls inside this category.

3.1 Packet-Loss Concealment

PLC for Audio Communication

With the widespread expansion of the Internet, voice over IP has become increasingly popular. To combat the unreliable delivery of voice packets over the internet, various packet-loss concealment (PLC) approaches have been proposed. PLC approaches are designed to work with existing decoders and with the corresponding encoders untouched for combating packet-loss. With properly designed PLC approaches, some traditional audio codecs can be directly utilized in an unreliable communication environment. The basic principle behind PLC is to take advantage of the neighboring received packets in the reconstruction of the missing signal segment by exploiting the redundant information. Many of the approaches on PLC are formulated in a heuristic manner. However, the procedures can roughly be divided into methods that can be motivated with a maximum likelihood (ML) based criterion and methods that can be motivated with the minimum mean squared error-based (MMSE) criterion.

A large number of PLC procedures replace the missing segment with a signal that is similar to that of the neighboring received signal segment. These approaches can be interpreted as ML methods. The typical set theorem from information theory states that asymptotically with increasing length any sequence generated by the model is equally likely [16]. In the literature, some approaches intend to reconstruct missing packets in the signal domain. The work in [35], [127] simply repeats the signal from the packets prior to the lost one to cover the gap due to packet-loss. A more advanced PLC approach using time-scale modification is described in [97, 102]. Compared with the waveform substitution approach, the second approach keeps the periodicity of the audio signal, resulting in an improved perceived quality. Other approaches intend to reconstruct missing packets in the transform domain. Specifically, the audio or speech signal is assumed to be generated by feeding an excitation signal into a statistical model. When a packet is lost, an excitation signal is constructed and then fed to the model in reconstruction. A commonly used audio or speech model is the autoregressive (AR) model, which is estimated using linear prediction (LP). Gündüzhan and Monttahan proposed to construct the excitation signal for the autoregressive model of the missing segment by repeating the excitation signal of
the previous received speech with a periodicity that equals the pitch period [43]. In [120], Wong et al. propose to classify the missing segment into voiced, unvoiced or partially voiced types and construct the excitation signal correspondingly.

The literature on usage of the MMSE criterion in the PLC context is relatively limited. In [94], Rødbro et al. employ a hidden Markov model (HMM) to track the evolution of speech features such as the pitch frequency. Benefiting from the sophisticated HMM model, the feature vector of a missing packet is estimated using the MMSE criterion and the harmonic sinusoidal parameters for synthesizing the speech are then constructed from the feature vector. In [59], Kondo and Nakagawa propose the use of a high-order AR model to capture both the short-term and long-term speech correlations. The missing speech samples are then predicted recursively by running the model with zero input.

**PLC for Video Communication**

Video communication typically requires higher bandwidth than audio communication. The coding efficiency has received considerable attention in the past. The video compression is usually a combination of temporal motion compensation to reduce temporal redundancy and spatial image compression to reduce spatial redundancy. The spatial image compression is in general achieved by transform coding on a block basis. The resulting parameters are then entropy-coded to produce the compressed bitstream which is further packetized. In the coding process, some frames are coded independently (referred to as *intraframe compression*) and other frames are coded by using earlier and/or later frames in a sequence (referred to as *interframe compression*). If the intraframe compression is dominant, the communication is much reliable as packet-loss would not result in serious error propagation.

A survey on approaches to recover the missing areas based on adjacent information is provided in [125]. Depending on the domain where the signal recovery is performed, spatial-domain scheme, temporal-domain scheme can distinguished. A spatial-domain scheme recovers impaired macroblocks based on neighboring ones (see [1, 45, 61, 126]). A temporal-domain scheme estimates the missing information by making use of temporal correlations embedded in adjacent frames. Examples include motion-compensated interpolation [2] [15], and motion vector recovery [44], [81], [62].

**3.2 Multiple Description Coding**

Multiple description (MD) coding deals with lossy coding of an information source for transmission with $L$ unreliable channels connecting the source and the destination. Each channel is assumed to provide an independent
means of transmission, facilitating the design of MD systems. In practice, the \( L \) channels could be created by using one transmission link via time-shifting. To make use of all the available channels, an encoder generates \( L \) descriptions of the source and then sends them over the \( L \) channels, respectively. Each channel either delivers the description successfully to the receiver or loses it during transmission, resulting in \( 2^L \) possible outcomes of received descriptions. The receiver reconstructs the best possible approximation of the source based on the received descriptions. The status of the \( L \) channels are known to the receiver but not to the sender. In the design of a MD system, the problem is to minimize the \( 2^L - 1 \) distortions due to reconstruction using information from any subset of the \( L \) descriptions for given channel rates.

For the special case that \( L = 1 \), i.e., only one channel is available for communication, the MD problem reduces to a single-description lossy source-coding problem. In this case, the design of a coding system concerns with only one distortion and one transmission rate. This single-description problem has received a lot of attention since the pioneering work by Shannon in his two landmark papers [98], [99]. For an overview of the literature to the problem, one can refer to the papers [7], [55], [42] and the books [6], [17], [16]. Single-description coding provides a basis for understanding the performance limits of multi-description cases.

The progress on the general MD problem in the literature is partly related to the theoretical MD performance limits (e.g., [85], [32]) and partly related to the design of efficient practical MD schemes that provide near optimal performance (e.g., [13, 30, 37, 38]). Practical MD schemes are in general content-independent. Specifically, the MD schemes can carry different types of media data for transmission, e.g., speech, image, video and so on.

Roughly speaking, two groups of practical MD schemes have been proposed in the literature: quantization-based and transform-based. Quantization-based schemes can be further classified into scalar MD quantization [5, 8, 106, 108, 114, 115], trellis coded MD quantization [51, 121, 130, 138] and vector MD quantization [11, 13, 18, 19, 28, 29, 84, 107, 110]. Transform based schemes include correlation transforms [36, 38, 39, 82, 122–124], over-complete expansions and filterbanks [4, 14, 23, 40, 41, 60].

Our work on MDC in this thesis is devoted to both the MD theoretical performance limits and practical MD schemes. For the practical MD schemes, we consider the vector MD quantization. In the literature review below, we will restrict our attention to those MD quantization-based schemes instead of transform-based schemes.

In the following, we first present a background for the MD problem. Then we provide an overview of the information theoretical results for single-description, two-description and multi-description scenarios, respectively. Finally we provide an overview of MD quantization-based schemes.
Introduction

Background

Suppose $X^n = \{X[i]\}, i = 1, \ldots, n$, is a sequence of source data that originates for instance from sampling a continuous signal. Let the $L$ encoding function be $f^n_l(X^n), l = 1, \ldots, L$, in the generation of $L$ descriptions, e.g., the $l$'th description is $f^n_l(X^n)$. In principle, there are $2^L - 1$ (the case that no description is received is trivial and often ignored) reconstruction functions $g_S(\cdot)$, each corresponds to a particular subset $S \subseteq \{1, \ldots, L\}$ of the $L$ descriptions. We denote the reconstruction output as $\hat{X}^n_S = g_S(f^n_l(X^n), l \in S)$ for a subset $S \subseteq \{1, \ldots, L\}$. Each reconstruction output introduces a quality degradation, or equivalently, distortion. We refer to the distortion resulting from $\hat{X}^n_{\{1,\ldots,L\}}$ as the central distortion and to distortions resulting from other reconstruction outputs as side distortions. In particular, the distortions resulting from $\hat{X}^n_{\{l\}}, l = 1, \ldots, L$ are referred as individual side distortions. Without ambiguity, $\hat{X}^n_{\{l\}}$ can also be denoted as $\hat{X}^n_l$. Fig. 4 shows the transmission architecture of an $L$-description system.

In the design of a MD system, a fidelity criterion $\rho^n(X^n, \hat{X}^n)$ has to be determined to measure the quality degradation of a reconstruction output $\hat{X}^n$ with regard to $X^n$. If $\rho^n(X^n, \hat{X}^n)$ takes the form

$$\rho^n(X^n, \hat{X}^n) = \frac{1}{n} \sum_{i=1}^{n} \rho(X[i], \hat{X}[i]),$$

(2)

$\rho^n(\cdot, \cdot)$ is referred to as a single-letter criterion. Distortion criteria of the form $\rho(X[i] - \hat{X}[i])$ are called difference distortion criteria. Once $\rho^n(X^n, \hat{X}^n)$ is specified, the average distortion is given by

$$d = \mathbb{E} \left[ \rho^n(X^n, \hat{X}^n) \right],$$

(3)
where \( \mathbb{E}[\cdot] \) represents the statistical expectation operation. We use subscripts to distinguish between the distortions resulting from different decoding functions, i.e., \( d_S \) is due to the reconstruction \( \hat{X}_S^n \), \( S \subseteq \{1, \ldots, L\} \). We denote the central distortion by \( d_c \) for simplicity.

In principle, the distortion criterion should capture the perceived quality degradation by the end users. See [26] for a number of distortion criteria applied to gray scale image coding in the evaluation of their relations with the perceived quality degradation. In practice the mean squared error (MSE) criterion is widely applied in the design of source coding systems. The dominance of MSE criterion stems from the fact that system performance w.r.t. the criterion is often traceable and closed form expressions can be derived. The MSE criterion is a difference distortion criterion. The measurement \( \rho(X[i] - \hat{X}[i]) \) takes the form: \( \rho(X[i] - \hat{X}[i]) = (X[i] - \hat{X}[i])^2 \). Correspondingly, the MSE due to the reconstruction \( \hat{X}^n \) is

\[
d = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ (X[i] - \hat{X}[i])^2 \right].
\]

In the multiple description scenario, the mean squared error with regard to \( \hat{X}_S^n \) takes the form:

\[
d_S = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ (X[i] - \hat{X}_S[i])^2 \right],
\]

where \( S \subseteq \{1, \ldots, L\} \).

The derivation of the MD rate-distortion region deals with the determination of a set of simultaneously achievable rates satisfying a prescribed fidelity criterion (e.g., a set of distortion constraints). Next we introduce a few definitions that help in describing the known MD information theoretical results.

**Definition 1** (e.g., [83]) An inner bound to the MD problem is a set of achievable rate-distortion points for a specific source and fidelity criterion.

**Definition 2** (e.g., [83]) An outer bound to the MD problem is a set of rate-distortion points, for a specific source and fidelity criterion, for which it is known that no points outside this bound can be reached.

For a given source and fidelity criterion, the MD rate-distortion region is not always tractable and closed form expressions may even be unavailable. In this situation, it is of great interest to find reasonable inner and/or outer bounds. An inner and an outer bound provide information about the position of the MD rate-distortion region. If the inner bound and outer bound coincide they are called tight. In this case, the MD rate-distortion region is fully specified by the inner or outer bound.
Single-Description Theoretical Results

Single-description coding can be viewed as a special case of the MD problem by letting $L = 1$. Since the single-description coding only involves one transmission rate and one distortion, the determination of the corresponding rate-distortion region is equivalent to specifying the optimal trade-off between the transmission rate and the distortion, which is referred to as the rate-distortion function. The rate-distortion function specifies a lowest rate (expressed in bits per sample) for a given distortion constraint. The distortion criterion is usually limited to be a single-letter criterion.

Formally stated, the rate-distortion function $R(d)$ for a stationary source $\{X[i]\}_{i=1}^{\infty}$ with memory is defined as (e.g., [6])

$$R(d) = \lim_{n \to \infty} R_n(d),$$

(6)

where the order-$n$ rate-distortion function $R_n(d)$ takes the form

$$R_n(d) = \inf \left\{ \frac{1}{n} I(X^n; \hat{X}^n) : \mathbb{E} \rho^n(X^n, \hat{X}^n) \leq d \right\}.$$  

(7)

The term $I(X^n; \hat{X}^n)$ denotes the mutual information between the two vectors (e.g., [6]). The infimum in (7) is over all conditional distributions $f_{\hat{X}^n|X^n}(\hat{x}^n|x^n)$ satisfying the distortion constraint. The function $R(d)$ specifies the minimum achievable rate (on a per sample basis) for transmitting an infinite sequence $\{X^n, n \to \infty\}$ with a distortion constraint $d$.

In general, it is difficult to obtain an analytic expression for the rate-distortion function for an arbitrary source. However, for a memoryless Gaussian source with variance $\sigma_X^2$ and MSE criterion, the rate-distortion function takes a simple form [6]

$$R(d) = \frac{1}{2} \log \left( \frac{\sigma_X^2}{d} \right),$$

(8)

where $d \leq \sigma_X^2$. Fig. 5 shows the $R(d)$ function for a unit-variance Gaussian source.

With the $R(d)$ expression in (8) as a basis, the rate-distortion function of an arbitrary memoryless scalar source and MSE criterion can be upper and lower bounded. Before presenting the two bounds for encoding a general source, we first introduce two measurements: entropy and differential entropy. Suppose $Y$ is a discrete random variable with a probability mass function $P_Y(y)$. The entropy is defined as (e.g., [57])

$$H(Y) = -\mathbb{E} [\log(P_Y(y))]$$

$$= - \sum_{y \in \mathcal{Y}} P_Y(y) \log(P_Y(y)).$$

(9)
Figure 5: Rate-distortion function for a unit-variance Gaussian source and the mean squared-error distortion criterion.

The entropy acts as a lower bound on the transmission rate in lossless coding of a discrete source [47, 64–66, 93, 129]. However, for a continuous random variable, the entropy is infinite, which is not useful. Instead, the differential entropy is introduced for a continuous random variable. Compared with the entropy, the differential entropy is abstract and has no physical meaning. Suppose $X$ is a continuous random variable with a probability density function $f_X(x)$. The differential entropy is defined as [6]

$$h(X) = -\mathbb{E}[\log(f_X(x))] = -\int f_X(x) \log(f_X(x)) \, dx.$$  

(10)

For a Gaussian random variable with variance $\sigma_X^2$, its differential entropy is $\frac{1}{2} \log(2\pi e \sigma_X^2)$.

Upon introducing the differential entropy, we present a well-known lower and upper bound of the rate-distortion function for a general memoryless scalar source and MSE criterion. Suppose a memoryless source $X$ has variance $\sigma_X^2$ and finite differential entropy $h(X)$. The rate-distortion function $R(d)$ satisfies the following inequality [6]:

$$\frac{1}{2} \log \left( \frac{\sigma_X^2}{d} \right) \geq R(d) \geq \frac{1}{2} \log \left( \frac{P_X}{d} \right),$$  

(11)

where $P_X = (2\pi e)^{-1} e^{2h(X)}$ is the entropy power which is the variance of a Gaussian density that has the same differential entropy as $X$. The lower bound of $R(d)$ in (11) is often referred to as the Shannon lower bound. The inequality (11) shows that, of all sources, the Gaussian source is the most difficult to compress.
The rate-distortion function can also be extended to stationary vector sources. Similar to the MSE criterion for scalar sources, the covariance distortion measure is often exploited for vector sources in various source coding problems (e.g., [77, 86, 91, 118, 119]). Suppose the information source is a memoryless vector source $X^n = \{X[i]\}, i = 1, \ldots, n$, with a $N \times N$ marginal covariance matrix $K_x$. The covariance distortion of a reconstruction $\hat{X}^n$ with regard to $X^n$ is defined as
\begin{equation}
\rho^n(\hat{X}^n, X^n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ (\hat{X}[i] - X[i])(\hat{X}[i] - X[i])^\top \right].
\end{equation}
In (12), the quantity $\rho^n(\hat{X}^n, X^n)$ is an $N \times N$ semi-definite matrix. The rate-distortion function for a memoryless vector Gaussian source and covariance distortion measure is known. Suppose the covariance distortion is $D$ that satisfies $0 \prec D \preceq K_x$. The rate-distortion function for the vector Gaussian source takes the form (e.g., [118]):
\begin{equation}
R(D) = \frac{1}{2} \log \det \left( \frac{K_x}{D} \right).
\end{equation}
If instead the MSE criterion is used for encoding a memoryless vector Gaussian source, i.e., $\text{tr} \left[ \rho^n(\hat{X}^n, X^n) \right]$ is considered, the well-known reverse waterfilling technique comes in, which we will not present here.

Two-Description Theoretical Results

Two-description coding consists of two encoding functions and three decoding functions. Fig. 6 shows the two-description transmission diagram. The total sum rate $R_T$ is split between the two descriptions: that is $R_T = R_1 + R_2$. When only one channel works, either Decoder 1 or Decoder 2 is triggered, which produces $\hat{X}_1^n$ or $\hat{X}_2^n$. The output results in a side distortion $d_1$ or $d_2$. When both channels work, Decoder c is triggered, which produces an output $\hat{X}_{1,2}^n$. Correspondingly, the output results a central distortion $d_c$.

The MD performance limit for the two-description case has been explored extensively over three decades (e.g., [32, 85, 133]). Formally, the two-description performance limit concerns the determination of smallest possible transmission rates $(R_1, R_2)$ with the distortions satisfying the distortion constraints $(d_1, d_2, d_c)$ respectively for encoding an infinite source sequence.

The pioneering work on the multiple description problem is the two-description achievable rate region by El Gamal and Cover [32] (known as the EGC region). Ozarow showed that the EGC region is tight for a memoryless scalar Gaussian source under the MSE criterion [85]. On the other hand, Zhang and Berger found that the EGC region was not tight in general when there is excess rate (redundancy) among the descriptions [136].
Later, Wang and Viswanath showed that the EGC region is also tight for a memoryless vector Gaussian source and covariance distortion measure [118]. Vaishampayan and Batllo in [112] further investigated the EGC region for a memoryless scalar Gaussian source under the MSE criterion. The authors considered a symmetric description scenario that is all the transmission rates (i.e., $R_1 = R_2$) are equal and the distortion depends only upon the number of received descriptions (i.e., $d_1 = d_2$). They provide a useful high-rate result that the product of the central distortion and the side distortion is approximately a function of the rate $R_1$. The advantage of this property is that it gives an asymptotically accurate approximation of the theoretical lower bound with high transmission rate.

In general, it is difficult to obtain the MD rate-distortion region for arbitrary sources and distortion criteria. Instead, the research focus has been on deriving the rate-distortion inner or outer bounds. An MD inner bound and an MD outer bound for general memoryless sources with finite differential entropy and MSE criterion have been derived by Zamir [133,134]. The derivation of the bounds was inspired by the single-description bound (11). Later an inner bound for arbitrary memoryless sources and MSE criterion larger than that derived in [133] was obtained by Feng and Effros [27]. Outer bounds for the binary symmetric source and Hamming distortion have been obtained by Wolf et al. [50], Witsenhausen [128] and Zhang and Berger [137].

Next we briefly review some of the two-description theoretical results. As was mentioned above, the two-description rate-distortion region is known for a memoryless Gaussian source and MSE criterion [85,118]. For a memoryless scalar Gaussian source with marginal variance $\sigma_X^2$, the region consists of the convex hull (which can be achieved by time-sharing) of the set of...
achievable tuples \((R_1, R_2, d_1, d_2, d_c)\) of which the rates satisfy [13, 85]

\[
R_1 \geq \frac{1}{2} \log \left( \frac{\sigma_X^2}{d_1} \right) \quad (14)
\]

\[
R_2 \geq \frac{1}{2} \log \left( \frac{\sigma_X^2}{d_2} \right) \quad (15)
\]

\[
R_1 + R_2 \geq \frac{1}{2} \log \left( \frac{\sigma_X^2}{d_c} \right) + \frac{1}{2} \delta(d_1, d_2, d_c)
= \frac{1}{2} \log \left( \frac{\sigma_X^2}{d_c} \right) + \frac{1}{2} \delta(d_1, d_2, d_c), \quad (16)
\]

where \(\delta(\cdot)\) is given by

\[
\delta(d_1, d_2, d_c) = \begin{cases} 1, & d_c < d_1 + d_2 - \sigma_X^2 \\ \frac{\sigma_X^2 d_c}{d_1 d_2}, & d_c > \left( \frac{d_1 + d_2 - \sigma_X^2}{\sigma_X^2} \right)^{-1} \\ \frac{(\sigma_X^2 - d_c)^2 - (\sqrt{\sigma_X^2 - d_1})(\sqrt{\sigma_X^2 - d_2} - \sqrt{d_1 - d_c})(d_2 - d_c))}{(\sigma_X^2 - d_c)^2 - (\sqrt{\sigma_X^2 - d_1})(\sqrt{\sigma_X^2 - d_2} - \sqrt{d_1 - d_c})(d_2 - d_c))^{-1}, & \text{otherwise} \end{cases}
\]

(14)-(17) describes all the achievable rates given the distortion constraints. For simplicity, we denote the MD rate distortion region (14)-(17) by
\( \mathcal{R}^*(\sigma_X^2, d_c, d_1, d_2) \). Fig. 7 shows the achievable rate region given a distortion triplet \((d_1, d_2, d_c)\), where \(d_c\) is not loose (i.e., \(d_c \leq 1/(\frac{1}{d_1} + \frac{1}{d_2} - \frac{1}{\sigma_X^2})\)). Equivalently, the rate-distortion region (14)-(17) can also be represented in terms of the achievable distortions given the rates:

\[
\begin{align*}
    d_1 &\geq \sigma_X^2 e^{-2R_1} \\
    d_2 &\geq \sigma_X^2 e^{-2R_2} \\
    d_c &\geq \begin{cases} 
        \sigma_X^2 e^{-2(R_1 + R_2)} \left(1 - \sqrt\Pi - \sqrt\Delta\right)^2 & \text{if } \Pi \geq \Delta \\
        \sigma_X^2 e^{-2(R_1 + R_2)} & \text{otherwise}
    \end{cases}
\end{align*}
\]

where

\[
\begin{align*}
    \Pi &= (1 - d_1/\sigma_X^2)(1 - d_2/\sigma_X^2) \\
    \Delta &= d_1d_2/\sigma_X^2 - e^{-2(R_1 + R_2)}.
\end{align*}
\]

Vaishampayan et al. studied the expressions (14)-(17) for a symmetric descriptions scenario (i.e., \(d_1 = d_2\) and \(R_1 = R_2\)) at high resolution [112, 113]. They showed that if the side distortions satisfy

\[
d_1 = \sigma_X^2 be^{-2R_1(1-a)},
\]

for \(0 < a < 1\) and \(b \geq 1\), the distortion product \(d_c d_1\) is lower-bounded by a simple bound:

\[
d_c d_1 \geq \frac{\sigma_X^4}{4} e^{-4R_1}.
\]

**Figure 8:** Graphic depiction of the relationship of \(d_c\) and \(d_1\) for different rates. The per-channel rate \(R\) is measured in bits.
The lower bound in (22) serves as a simple means of relating the performance of practical MD schemes to the real theoretical bound (14)-(17) (or equivalently (18)-(20)). The asymptotic behavior of the distortion product has been utilized widely as a tool to assess the efficiency of practical two-channel MD systems [106, 107, 110]. It was shown that an optimal two-description scheme behaves similarly to the lower bound approximation (22) at high resolution and when \(d_c \ll d_1\). Fig. 8 shows the real rate-distortion bound and the approximation (22) for different transmission rates. It is seen the approximation is quite close to the real lower bound when the transmission rate increases. A more general but less used lower bound approximation was also obtained by Vashampayan [113]:

\[
d_c d_1 = \frac{\sigma_X^2}{4} \frac{1}{1 - d_c/d_1} e^{-4R_1},
\]

which meets the approximation (22) if \(d_c/d_1 \to 0\). Although the approximation (22) is less accurate than (23), it has a simple form, which leads to a wide usage of the expression (22).

Based on (11) and the derivation procedure in [85], Zamir derived an inner and an outer bound of the MD rate distortion region for a general memoryless source and MSE criterion. Suppose a memoryless source has a variance \(\sigma_X^2\) and finite differential entropy \(h(X)\). Its MD rate distortion region \(R_X(d_c, d_1, d_c)\) is upper and lower bounded by [133]

\[
R_X^*(\sigma_X^2, d_c, d_1, d_2) \subseteq R_X(d_c, d_1, d_c) \subseteq R_X^*(P_X, d_c, d_1, d_2),
\]

(24)

where \(P_X\) takes the form as that in (11). \(R_X^*(A, d_c, d_1, d_2)\) is the rate distortion region of a memoryless Gaussian source with variance \(A\). The two bounds (24) parallel Shannon’s lower and upper bounds (11) for the single description coding.

In [118], Wang and Viswanath studied the performance limit of an \(L\)-description vector Gaussian scenario. The covariance distortion measure was considered. For the special two-description case, the authors obtained the rate distortion region. Suppose the data source is an i.i.d. vector Gaussian source with a marginal covariance matrix \(K_x\). The covariance distortion constraints are \((D_c, D_1, D_2)\), where \(0 < D_c < D_1 < K_x\), \(l = 1, 2\). The rate distortion region is given as [118]

\[
R^*(K_x, D_c, D_1, D_2) = \left\{ \begin{array}{l}
(R_1, R_2) : \\
R_l \geq \frac{1}{2} \log \frac{|K_x|}{|D_l|}, \quad l = 1, 2 \\
R_1 + R_2 \geq \sup_{K_x > 0} \frac{1}{2} \log \frac{|K_x||K_x + K_l||D_l + K_l|}{|D_l||D_l + K_l||D_2 + K_2|}
\end{array} \right\}.
\]

(25)

When the source is specialized to a scalar Gaussian source, (25) coincides with (14)-(17).
Generally speaking, opening problems for two-description coding would be to consider performance limits for general sources and distortion criteria. One can focus on deriving better inner or outer MD rate distortion bounds.

**L-Description Theoretical Results**

$L$-description coding consists of $L$ encoding functions and $2^L - 1$ decoding functions. This suggests that in the derivation of theoretical results one has to consider the trade-off between $L$ transmission rates and $2^L - 1$ distortions. Due to the problem complexity, the full characterization of the MD rate-distortion region for a source and a general distortion criterion is still an open problem.

Significant effort has been spent on deriving inner and/or outer rate distortion bounds for $L$-descriptions coding. Venkaramani et al. first obtained an achievable $L$-description rate-distortion region [116, 117] for arbitrary memoryless sources and a single-letter distortion criterion. In general the region is described by a complicated expression but for the quadratic Gaussian case the region ends up with a simple form. Later, the work in [89], [90], provided an enlarged achievable rate region by using the random binning ideas from distributed source coding. The recent work by Tian [109] provided both an inner and an outer rate distortion bounds for a scalar Gaussian source under a symmetric MSE distortion constraint.

Other research has been devoted to characterizing the optimal sum rate or deriving the rate-distortion region for some special multiple description scenarios. Wang and Viswanath obtained the optimal sum rate when only a subset of distortion constraints are of concern [118, 119]. In particular in [119], the authors studied the symmetric descriptions scenario with two levels of receivers. Each of the first-level receivers obtains $\kappa$ of the $L$ descriptions, ($\kappa < L$). The second-level receiver obtains all the $L$ descriptions. For the considered MD scenario, the optimal sum rate has been derived for a memoryless vector Gaussian source and the quadratic distortion criterion. The work by Chen in [12] characterized the rate-distortion region when only the central distortion and the $L$ individual side distortion constraints are imposed. However, the work is limited to a memoryless scalar Gaussian source and MSE criterion.

As Paper D in the thesis also studies the MD problem addressed in [12] but for a vector Gaussian source, we briefly present the result of [12] in the following for reference. Suppose a scalar Gaussian source has variance $\sigma_X^2$. The central and individual side distortion constraints are $d_{\{1,\ldots,L\}}$ (or $d_c$) and $d_l$, $l = 1, \ldots, L$, respectively. The rate distortion region for this MD scenario is described by characterizing every optimal weighted sum rate in [12]. Let $\mathcal{R}^\star(d_1, \ldots, d_L, d_{\{1,\ldots,L\}})$ denote the convex closure of the set of all achievable rate vectors satisfying the distortion constraints. The optimal weighted sum rate for a weighting vector $(\alpha_1, \ldots, \alpha_L)$, $\alpha_1 \geq \ldots \geq \alpha_L > 0$, is
is given as

$$R_{wei} = \min_{(R_1, \ldots, R_L) \in \mathbb{R}^* (d_1, \ldots, d_L, d_{1(\ldots,L)})} \sum_{l=1}^{L} \alpha_l R_l. \quad (26)$$

In the calculation of $R_{wei}$ in (26), the author first derived the optimal weighted sum rate with respect to individual distortion constraints $d_l$, $l = 1, \ldots, L$, auxiliary distortion constraints $d_{(1,\ldots,l)}$, $l = 2, \ldots, L - 1$, and central distortion constraint $d_{1(\ldots,L)}$. The new optimal weighted sum rate $\eta(d_1, \ldots, d_L, d_{(1,\ldots,2)}, \ldots, d_{(1,\ldots,L-1)}, d_{1(\ldots,L)})$ takes the form

$$\eta(d_1, \ldots, d_L, d_{(1,\ldots,2)}, \ldots, d_{(1,\ldots,L-1)}, d_{1(\ldots,L)}) = \min_{d_{1(\ldots,l)} \in [0, d_{(1,\ldots,l)}]} \max_{\sigma_j \in [0, \sigma^2_{X}]} \psi(d_1, \ldots, d_L, d_{(1,\ldots,2)}, \ldots, d_{(1,\ldots,L)}, \sigma_j^2, \ldots, \sigma_{L-1}^2)(27)$$

where

$$\psi(d_1, \ldots, d_L, d_{(1,\ldots,2)}, \ldots, d_{(1,\ldots,L)}, \sigma_j^2, \ldots, \sigma_{L-1}^2) = \sum_{i=1}^{L-1} \left( \frac{\alpha_{i+1}}{2} \log \left( \frac{\sigma_j^2 d_{(1,\ldots,i)}}{\sigma^2_i d_{(1,\ldots,i+1)} + \sigma^2_j d_{(i,\ldots,i+1)}} \right) \right) + \frac{\alpha_i - \alpha_{i+1}}{2} \log \left( \frac{\sigma^2_j}{d_{(1,\ldots,l)}} \right) + \frac{\alpha_L}{2} \log \left( \frac{\sigma^2_j}{d_{1(\ldots,L)}} \right).$$

The optimal weighted sum rate $R_{wei}$ is obtained by varying the auxiliary distortion constraints $d_{(1,\ldots,l)}$, $l = 2, \ldots, L - 1$, in (27) and searching for the minimum value. For fixed auxiliary distortion constraints, (27) involves a min-max optimization.

**Multiple Description Quantization**

After introducing the MD theoretical results above, we now focus on practical MD schemes that are discussed in the literature. In particular, we restrict our attention to the multiple description quantization schemes.

Since the pioneering work by Vaishampayan [111] where a practical multiple-description scalar quantization (MDSQ) system was first proposed, many researches have focused on the design of efficient quantization-based MD schemes. The design of MDSQ is essentially converted to construct good index assignment matrices for two description case (see [106], [110] and [111]) or index assignment arrangements for multi-description case (see [8]). Fig. 9 shows the linear index assignment proposed in [111] as an example. The two-description MDSQ design problem is particularly well understood [58], [112]. Suppose a symmetric multiple-description encoder sends information over each channel at a rate of $R$ bits per sample. The performance of the system is measured by a three-tuple $(R, d_1, d_c)$ where $d_1$
is the side distortion and $d_c$ is the central distortion. Suppose the source to be transmitted has a finite differential entropy $h(X)$. The work of [112] showed that for an entropy-constrained multiple-description quantizer, the distortions satisfy

$$
d_1d_c = \lim_{R \to \infty} \frac{1}{4}\left(\frac{e^{2h(X)}}{12}\right)^2 e^{-4R}, \quad (28)$$

which has the same structure as (22). To make a fair comparison with (22), we consider a scalar Gaussian source with variance $\sigma^2_X$. In this case, we have

$$
d_1d_c = \lim_{R \to \infty} \frac{\sigma^4_X}{4}\left(\frac{2\pi e}{12}\right)^2 e^{-4R}. \quad (29)$$

The gap between (29) and (22) is a constant, which is $(2\pi e/12)^2$. The performance gap is mainly due to the low dimensionality of quantization space.

![Linear index assignment with five diagonals.](image)

It is well known that vector quantization has a space filling advantage over scalar quantization [57]. This is because one has the freedom to construct cell shapes that are more "spherical" than a hypercube in higher dimensional space. Specifically, for the scenario of single-description entropy-coded at $R$ bits per sample, when using an $n$-dimensional lattice $\Lambda$ as a codebook the distortion $d_V(R)$ related with the distortion of $d_S(R)$ of scalar quantizer by

$$
\lim_{R \to \infty} \frac{d_V(R)}{d_S(R)} = \frac{G(\Lambda)}{1/12},
$$
where $G(\Lambda)$ is the normalized second moment of a Voronoi cell of the lattice \( \Lambda \). It has been shown that good lattices exist that satisfy $G(\Lambda) \rightarrow \frac{1}{\pi^2}$ as $n \rightarrow \infty$ [24]. Lattices are commonly used in the design of MD quantization systems to gain quantization efficiency over MDSQ [110]. The MD quantizers that exploits lattice structure are referred to as multiple-description lattice vector quantization (MDLVQ) systems. As in MDSQ, the main design task in MDLVQ is to construct a good index assignment. The performance of a MDLVQ system for the two-description case was analyzed in [110], culminating in the relation

\[
d_1 d_c = \lim_{R \rightarrow \infty} \frac{1}{4} G(\Lambda)G(S_n)e^{4h(X)}e^{-4R},
\]

where $G(S_n)$ is the normalized second moment of a sphere in $n$ dimensions. As compared to (28), the distortion product exhibits a reduction due to the use of lattice structure. As $n$ goes to infinity, (30) approaches the lower bound approximation (22) for a Gaussian source by exploiting good lattices.

The design of practical $L$-channel MDLVQ systems was addressed in [46, 84]. The index assignment was recognized to play a critical role in the system design. However no simple form of the distortion product like the expression (30) for the two-description case has yet been developed.

4 Summary of Contributions

The focus of this thesis is on robust multimedia communications over packet switched networks. The main contributions of the thesis can be summarized as follows:

- it proposes a more efficient adaptive playout scheduling approach for VoIP application,
- it proposes a low-complexity and effective packet-loss concealment approach for VoIP application and
- it provides insight in the performance limits for some scenarios of multiple description coding.

The thesis consists of four research papers. I formulated the approaches in the papers and did all of the fundamental mathematics: the co-authors mainly provided advise and experimental data. Short summaries of the papers are presented below.

Paper A: Adaptive Playout Scheduling for VoIP: Event Triggered Control Policy

In paper A, we propose a new adaptive-playout scheduling approach for the application of Voice over IP. Differently from the literature, we use Wiener
process to model the fluctuation of the buffer length in the absence of any adjustment. The advantage of the Wiener process modelling is that it can model various packet-network delay statistics reliably, including the situation when delay-spikes happen. The problem of adaptive-playout scheduling is then formulated as a stochastic impulse control problem. The adjustment of the buffer length is taken to be the control signal in the new control problem. Specifically, the control signal consists of length units that correspond to inserting or dropping a pitch cycle from the buffer. The control problem penalizes both the buffer length and the control signal (or the adjustment) in terms of a Lagrange cost function. A so-called band control policy is rigorously shown to be optimal for this control problem. The band control policy maintains the buffer length within a band region by imposing impulse control (inserted or dropped pitch cycles) whenever the bounds of the band are reached. Experiments performed on both synthetic and real network-delay traces show that the proposed playout scheduling scheme outperforms two recent algorithms in most cases.

**Paper B: Autoregressive Model-Based Speech Packet-Loss Concealment**

In paper B, we study packet-loss concealment for speech based on autoregressive modelling using a rigorous minimum mean square error (MMSE) approach. Addressing the fact that most reported methods are designed in a heuristic manner, the main aim of the paper is to consider the PLC problem rigorously. We expect the obtained result to provide a foundation for understanding the functionality of the heuristic methods. In particular, we investigate the effect of the model estimation error on predicting the missing segment and derive an upper bound on the mean square error of the missing segment on a sample basis. Our experiments show that the upper bound is tight when the estimation error is less than the signal variance. We also consider the usage of perceptual weighting on prediction to improve the perceived speech quality. A rigorous argument is presented, showing that perceptual weighting is not useful in this context. Subjective quality comparison tests show that the proposed MMSE-based system provides state-of-the-art performance.

**Paper C: High-Rate Analysis of Symmetric L-Channel Multiple Description Coding**

In paper C, we study the tight rate-distortion bound for symmetric $L$-description coding of a vector Gaussian source with two levels of receivers. Each of the first-level receivers receives $\kappa$ ($\kappa < L$) out of the $L$ descriptions. The second-level receiver obtains all $L$ descriptions. The rate-distortion bound takes a complex form that, in general, is inconvenient as a refer-
ence to calculate the performance loss of practical MD schemes. The main aim of this work is to derive a simple but good approximation of the rate-distortion bound, which can serve as a useful tool to evaluate the efficiency of practical \( L \)-description schemes. We find that when the theory is applied to the scalar Gaussian source, the product of a function of the side distortions (corresponding to the first-level receivers) and the central distortion (corresponding to the second-level receiver) is asymptotically independent of the redundancy among the descriptions. Using this property, we assess the performance loss of a practical multiple-description lattice vector quantizer (MDLQVQ) as an example. Another contribution of the work is that we provide a new geometric analysis of the performance of the considered MDLQVQ, which results in an expression for the side distortions using the normalized second moment of a sphere of higher dimensionality than the quantization space.

**Paper D: Bounding the Rate Region of Vector Gaussian Multiple Descriptions with Individual and Central Receivers**

In paper D, we study the rate region of the vector Gaussian multiple description problem with individual and central quadratic distortion constraints. Although the rate region for the considered scenario is known for a scalar Gaussian source, the expression involves a min-max optimization problem, which is complicated. For a vector Gaussian source, the rate region is unknown. In this work, we derive an outer bound to the rate region of the vector Gaussian \( L \)-description problem. The bound is obtained by lower bounding a weighted sum rate for each supporting hyperplane of the rate region, of which the expression only involves a maximization problem. The outer bound is tight when the data source is a scalar Gaussian source. In this case, the optimal weighted sum rate for each supporting hyperplane is obtained by solving a single maximization problem. This contrasts with existing results, which require solving a min-max optimization problem.

**References**


[79] M. Narbutt, A. Kelly, L. Murphy, and P. Perry. Adaptive VoIP Play- 
out Scheduling: Assessing User Satisfaction. *IEEE Internet Comput-

Adaptive Estimation of Network Delays. In *Proc. of International Tel-

[81] A. Narula and J. Lim. Error concealment techniques for an all-digital 
high-definition television system. In *Proceedings of the SPIE*, volume 

Redundancy Rate-Distortion Analysis of Multiple Description Coding 
Using Pairwise Correlating Transforms. In *Proc. IEEE Conf. on Im-

[83] J. Østergaard. *Multiple-Description Lattice Vector Quantization*. PhD 

[84] J. Østergaard, J. Jensen, and R. Heusdens. n-Channel Entropy- 
Constrained Multiple-Description Lattice Vector Quantization. *IEEE 

[85] L. Ozarow. On a Source-Coding Problem with Two Channels and 

[86] D. P. Palomar and S. Verdú. Gradient of Mutual Information in Linear 
154, 2006.


231, 1999.

[89] S. S. Pradhan, R. Puri, and K. Ramchandran. n-Channel Symmetric 
Multiple Descriptions-part I: (n,k) Source-Channel Erasure Codes. 

[90] R. Puri, S. S. Pradhan, and K. Ramchandran. n-Channel Symmetric 
Multiple Descriptions- part II: An Achievable Rate-Distortion Region. 


