LOSS CONCEALMENTS FOR LOW BIT-RATE PACKET VOICE

BY

DONG LIN

B.E., University of Science and Technology of China, 1996
M.S., University of Illinois at Urbana Champaign, 1999

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2002

Urbana, Illinois

ABSTRACT

In recent years, packet voice over the Internet has become an attractive alternative to conventional public telephony. However, the Internet is a best-effort network, with no guarantee on its quality of service. A fundamental issue in real-time interactive voice transmissions over an unreliable Internet Protocol (IP) network is the loss, or late arrival of packets. This problem is especially serious in transmitting low bit-rate coded speech when pervasive dependencies are introduced in a bit stream, leading to errors propagated to subsequent frames when a packet is lost. As a result, the study of loss-concealment schemes is important in ensuring high-quality playback.

In this research, we choose an end-to-end loss-concealment approach that requires no special support from the underlying network. In particular, we focus on developing a nonredundant sender-receiver collaborated multiple-description coding (MDC) scheme.

We propose a new coder-dependent parameter-based MDC. Based on the correlations of coding parameters, we generate multiple descriptions systematically. In particular, we have observed high inter-frame correlations in linear predictors of low bit-rate coders, but low correlations in excitation parameters. Hence, we generate multiple descriptions at the sender side by interleaving linear predictors, and reconstruct lost ones at the receiver side by linear interpolations. In addition, we replicate hard-to-reconstruct excitation parameters to multiple descriptions. We also require our design to be done in such a way that does not require extra transmission bandwidth and that can adapt its number of descriptions to network-loss conditions dynamically. We have compared three linear predictor representations, Reflection Coefficient, Log Area Ratio, and Line Spectral Pair (LSP), by their reconstruction qualities, and have found that the LSP representation performs the best. Finally, we have studied LSP reconstructions and enhancements on excitation quality. Extensive tests on FS CELP, ITU G.723.1, and FS MELP under different loss scenarios have shown good quality and reliability of our proposed scheme.

To my dear husband and parents

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my thesis adviser, Dr. Benjamin W. Wah, for all the fruitful and stimulating discussions and the guidance he provided me during the course of this research. I have consistently benefited from his professionalism and unwaveringly high standards.

I would also like to thank Professor Stephen Levinson, Professor Klara Nahrstedt, and Professor William H. Sanders for their willingness to serve in my Ph.D. committee, sparing for me their precious time and help. I am especially grateful to Professor Levinson for helping me gain an understanding of speech processing.

Special thanks go to fellow members in my research group, Xiao Su, Minglun Qian, Yixin Chen, and Hang Yu, for their valuable discussions.

I would also like to thank my husband, Zhe Wu, for his continuous encouragement and support. His unfailing faith in me helps make this thesis a reality.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Motivations

1.1.1 Packet voice

Traditionally, voice communication has been done through the public switched telephone network (PSTN), which is also referred to as the "plain old telephone service" (POTS). PSTN is a time-division-multiplexing (TDM) circuit-switched network [1]. In establishing a session between two end-users, a virtual circuit is allocated. In this way, bandwidth is guaranteed and so is reliability. In a PSTN design, end points, such as telephones, are dumb devices with minimum functionalities, while its network is intelligent and handles features like access control, scheduling, and signaling.

With the exponential growth of the Internet in recent years, packet voice becomes an attractive alternative to conventional public telephony [2]. Generally, the basic steps to transmit voice signals over a packet-switched network at the sender side include the

conversion of analog voice signals to a digital format, processing/coding/compressing the digital signals, forming packets of the resulting signals, and transmitting the packets over the network. At the other end, a receiver receives the packets, unpacks them, and decompresses/decodes/processes the signals back to their original format.

The Internet is very different from PSTN in the sense that it moves the processing power of a network to its end points. The end points of the Internet are computers with powerful computational capabilities and an ever-growing number of applications. The network itself is designed to be simple. The Internet Protocol (IP) delivers a packet by checking its destination address and by forwarding it to the next hop along the way, without knowing the contents. All packets contending for one outgoing link are statistically multiplexed. Functions, such as access control, scheduling, and signaling are not included. This simplicity leads to its widespread deployment and its low access cost. The constant growth of the Internet opens up great opportunities for a variety of voice applications, such as Internet telephony, teleconferencing, and wireless voice communication. The telecommunication industry is developing a next generation network (NGN) based on a common packet-based architecture for voice, data, and multimedia services [3].

One of the main challenges in using the Internet for real-time voice applications is to deliver similar levels of quality and reliability comparable to those of the traditional telephone network.

1.1.2 Challenges

Interactive voice communications usually have strict delay requirements. International Telecommunication Union (ITU) Recommendation G.114 specifies a "one-way transmission delay" of up to 400 ms to be acceptable [1]. The standard explains such a delay requirement as follows:

a) 0 to 150 ms is acceptable for most applications;

b) 150 to 400 ms is acceptable, provided that the network designer is well aware of its impact on some user applications; and

c) 400 ms or above is unacceptable.

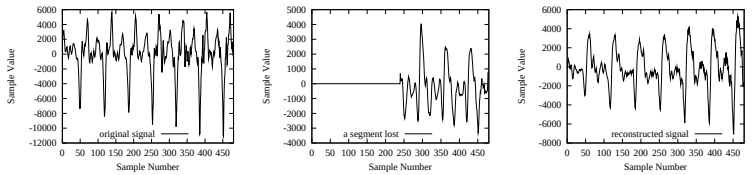
In a packet network, this requirement implies that packets delayed over a certain time limit are considered lost and cannot be used by the receiver. Consequently, from the application's point of view, those delayed packets are equivalent to lost packets.

Voice communications also have loss requirements. Comparing packet-switched networks to circuit-switched networks, loss happens in an entirely different form. In circuit-switched networks, losses happen at the bit or sample level, while in packet networks, losses happen at the packet level. Losing a packet corresponds to losing an entire interval of speech, with the duration depending on the inter-packet transmission time. Hence, packet losses are generally perceptible, and frequent losses make playback intermittent and annoying. Packet losses are common in the Internet, and loss rates sometimes can be quite high. Delayed packets will further increase loss rates. Measurements carried

out by others [4] and us have shown that packet-loss rates of some connections can be as high as 50%.

In addition, coding further complicates ways to maintain quality. In order to reduce transmission bandwidth, speech coding is employed to compress voice signals. Compression is achieved by exploiting temporal redundancies among speech signals and by realizing actual compression through quantization. Hence, speech coding is lossy. Current compression algorithms are not robust enough to transmission errors. Their sole objective is to maximize coding gain, assuming error-free channels. In particular, low bit-rate speech coding algorithms generally incorporate complex mechanisms to remove as much redundancy as possible, which in turn introduces a great deal of time dependencies in a coded sequence. Although error resilience is not an objective in coder designs, it is not a severe problem for PSTN in which infrequent occurrences of bit errors can be corrected using error-correction codes. However, for packet networks, the loss of a speech packet degrades playback quality not only of the lost packet itself but also of subsequent packets [5].

An advantage of voice communications over data communications is that 100 percent accuracy is not a must. This relaxation is indeed very useful because a receiver can tolerate a certain level of signal distortions without significant performance degradation [6]. Therefore, if we can convert packet losses to voice-sample distortions, the playback quality may become tolerable. To illustrate this idea, we compare the impact of a lost packet to the impact of voice-sample distortions in Figure 1.1. Figure 1.1a plots a segment of a speech stream transmitted by our transmission system, described in



(a) Original voice segment (b) Without loss recovery (c) With loss recovery

Figure 1.1: Illustration of the impact of packet losses vs. that of sample distortions.

Chapter 4. Using a segment size of 480 samples, the samples are coded using the Federal Standard Code-Excited Linear Prediction (FS CELP) coder and put into two packets. After transmitting the two packets over the UIUC-Slovakia Internet connection (details described in Chapter 3), we plot in Figure 1.1b the corresponding segment received and decoded. In this example, the first packet is lost during transmission, and the second packet arrives successfully at the receiver. Figure 1.1b illustrates clearly that the signals in the first half of the segment are wiped out and that the second half of the signals played back differ very much from the original ones.

In the next illustration, we replace the FS CELP coder with our proposed loss recovery-enabled FS CELP coder, redo the delivery, and plot the reconstructed transmission-system output in Figure 1.1c. (The details of the loss recovery-enabled FS CELP coder and the reconstruction process are described in Chapter 4.) The graph shows that the loss-recovery scheme can reconstruct the missing samples, although there are distortions in each sample.

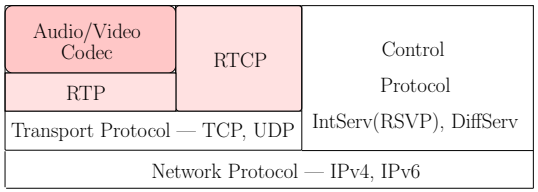


Figure 1.2: The current Internet protocol stack for supporting real-time applications. (Shaded areas are within the scope of H.323.)

The examples demonstrate that by converting packet losses to sample distortions, we can improve playback quality and, thus, achieve loss concealments.

1.1.3 Protocol and network hierarchy for supporting real-time voice applications

Since we are interested in voice transmissions over packet networks, we discuss in this section the network and protocol environments for real-time voice applications. Figure 1.2 shows the current Internet protocol stack for supporting real-time applications.

Consider the network-layer protocol. Currently, it is IP version 4 (IPv4) that only supports best-effort service. Its fundamental idea is to do everything possible to deliver each packet, without guaranteeing its actual delivery or its delivery time. As said before, losses are common, and variations in delay further worsen the packet-loss problem.

The next generation network-layer protocol is IP version 6 (IPv6) [7] with the following major features:

- expanded addressing capabilities;

- header format simplification;
- improved support for extensions and options;
- flow labeling capability; and
- authentication and privacy capabilities.

Among these features, the flow-label capability is pertinent to real-time applications. The specification states that the flow-label field in an IPv6 header may be used by a source to label sequences of packets that require special handling by IPv6 routers, such as non-default quality of service (QoS) or real-time service. It also says that this aspect of IPv6 is still experimental and subject to change. Another field in an IPv6 header that is closely related to multimedia applications is the traffic-class field. The traffic-class field can be used by originating nodes and/or forwarding routers to identify and distinguish among different classes or priorities of IPv6 packets. It serves a similar purpose as the type-of-service (ToS) field in an IPv4 header.

The IPv6 specification is, however, rather vague with regard to how packets of a given flow are treated, stating that routers may or may not honor flow labels in treating packets. A source using a flow label may, therefore, have no guarantee that its packets will receive special attention from the network. In other words, it cannot be assumed that the concept of flows in IPv6 will have significant performance impact on the network [8]. As for the traffic-class or ToS field, the specification states that there are no common agreements on the type-of-traffic classifications that are most useful for IP packets. Regarding IPv6 implementation status, the implementation notes published by the Internet Engineering Task Force (IETF) in 1998 [9] show that almost no vendor supports the flow and traffic-

class features. Therefore, these innovations in IPv6 alone are not enough to support QoS, unless all routers and gateways implement these features and have special mechanisms to meet service requirements.

There are two major service proposals in the literature to assist IPv6 in enhancing real-time support in the Internet (the "control protocol" block in Figure 1.2): integrated service (IntServ) [10] and the more recent differentiated service (DiffServ) [11].

In IntServ, three classes of service have been proposed: guaranteed service, controlled-load service, and best-effort service. Both the guaranteed and controlled-load services are based on quantitative service requirements and require signaling and admission control in network nodes. The Resource ReServation Protocol (RSVP) [12], often regarded as the protocol for IntServ, allows hosts to request a specific QoS for an application, as well as for routers to pass QoS requests along a routed path and to maintain state information about flows in routers. The flow label of IPv6 is, therefore, useful here. An accepted resource-reservation request is guaranteed that the associated service quality meets the desired requirements, leading to satisfactory operations of the application. The major advantage of IntServ is that it provides special service support for real-time traffic, while its disadvantage is that end-to-end service guarantees cannot be supported unless all nodes along the route support IntServ [8]. More importantly, it is well recognized that the support of per-flow guarantees in the core Internet will pose severe scalability problems.

In contrast to the flow service provided by IntServ, traffic-class service is provided by DiffServ, using the traffic-class field in IPv6. DiffServ specifies a hierarchical model for network resource management:

- *Interdomain resource management*: unidirectional service levels and, hence, traffic contracts are agreed at each boundary point between a customer and a provider for traffic entering the provider network;
- *Intradomain resource management*: a service provider is solely responsible for the configuration and provisioning of resources within its domain; furthermore, service policies are also left to the provider.

DiffServ does not impose either the number of traffic classes or their characteristics on a service provider. The advantage of DiffServ over IntServ is that per-flow states are avoided since flows are aggregated in classes. Also, no *a priori* resource reservation is necessary, and no extra setup delay is involved.

Although DiffServ seems attractive, it still has some serious shortcomings [8]. First, it is very difficult to provide simultaneously several services with different qualities within the same network. More research is required to study the added complexity due to possibly adverse interactions among different classes of service. Another disadvantage of DiffServ is that it is based on local service agreements at customer/provider boundaries. Therefore, end-to-end services will be built by concatenating such local agreements at each domain boundary along the route to the final destination. The concatenation of such local services to provide meaningful end-to-end services is still an open research issue. The last drawback is that it is impossible to totally avoid overloading resources unless resources are massively over-provisioned.

For both IntServ and DiffServ, further research and development is required to make them practical. The above discussions lead us to conclude that strict QoS cannot be

guaranteed, and packet loss is inevitable in the network layer. This is especially true for wireless networks. Although voice communication over wireless networks today is still circuit based, it is going to be packet based in the future. For example, the fourth-generation wireless networks will be all IP networks [13]. The packet-loss problem is unavoidable for wireless networks because wireless links are inherently unreliable.

Next, consider the transport-layer protocols, Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) (Figure 1.2). TCP is connection oriented, or circuit switched, and is responsible for the correct delivery of data by detecting lost data and by triggering end-to-end retransmissions. Although TCP is reliable, it may result in unacceptable delays when transmitting real-time data. On the other hand, UDP is connectionless, with no guarantee of delivering packets to their destinations. With no concept of connection, packets arriving at the destination may be out of order. Hence, UDP requires the application to reorder packets and take care of losses.

These two transport protocols are two extremes when reliability is concerned. TCP guarantees total reliability without considering delays, and is normally not a good choice for real-time applications. In contrast, UDP guarantees no reliability and, clearly, adds little extra delay. It is well recognized that UDP is preferred for handling real-time traffic; hence, the packet-loss problem will need to be solved.

On top of the transport layer, there are application-level standards developed to support end-to-end real-time transmissions and ensure interoperability among different products. The international standards that are closely related to packet-voice transmissions are H.323, H.225.0, and H.245 [14, 15, 16]. All three standards belong to a family

of ITU-T recommendations and are covered under the umbrella standard H.323 (shaded areas in Figure 1.2).

H.323 is a standard that specifies the components, protocols, and procedures for providing multimedia communication services, including audio, video, and data communications, over packet networks, such as IP networks. The standard describes H.323 system components, or entities, including terminals, gateways, gatekeepers, multipoint controllers, multipoint processors, and multipoint control units.

An H.323 terminal can be either a personal computer or a stand-alone device. It must support audio communications and can optionally support video or data communications. H.225.0, the standard for the call-signaling protocol and media-stream packetization for packet-based multimedia communication systems, defines how to use the Real-time Transport Protocol (RTP) [17] and RTP control protocol (RTCP) to packetize audio and video data and achieve synchronization. RTP is an application-layer protocol that provides functionalities like resequencing of packets, loss detection, delay adaptation by buffering, and playback-point determination. RTP uses RTCP to monitor and convey statistics about an ongoing session. Although it is called the Real-time Transport Protocol, RTP does not contain any mechanism to support timely delivery of data, nor does it provide any recovery scheme for packet losses. Such losses will need to be handled by the applications themselves.

In short, none of the network-layer protocols, such as IPv4 and IPv6, transport-layer protocols, such as TCP and UDP, and application-level protocols, such as RTP and H.323, truly provides practical QoS support and/or can solve the fundamental packet-loss

problem for real-time voice transmission. To achieve better voice transmission quality, it is, therefore, both necessary and natural to develop end-to-end loss-concealment schemes.

1.2 Problem Statement

Based on the shortcomings identified in the last section, the objective of this dissertation is to *design, analyze, and evaluate robust end-to-end coding and loss-concealment schemes in order to support reliable and real-time low bit-rate voice transmissions over unreliable IP networks, such as the Internet and wireless wide-area networks (WWAN).*

1.3 Outline of This Dissertation

This dissertation is organized as follows. In Chapter 2, we survey existing work on methods to improve the quality of multimedia data transmitted over the Internet. There are generally two end-to-end approaches: coder-independent and coder-dependent loss-concealments. We analyze these two approaches in detail.

Before we can design loss-concealment schemes, we need to know the kind of losses (bursty or random) and the amount of losses in a typical voice transmission over UDP. To answer these questions, Chapter 3 investigates in detail the traffic patterns for real-time voice transmissions for a variety of connections.

Throughout this dissertation, we perform tests on different loss-concealment and coding strategies using eight speech streams. These streams are listed in Table 1.1. We have chosen these streams to include a variety of speakers and syllables in order for our de-

Table 1.1: Voice streams used in our experiments.

Index	Length (ms)	Speakers	Index	Length (ms)	Speakers
1	21 432	2 male, 1 female	5	4160	1 male
2	22 560	2 male, 1 female	6	4082	1 male
3	4424	1 female	7	4867	1 male, 1 female
4	5091	1 female	8	73 615	1 male, 1 female

Streams 1 and 2 were obtained from John Hopkins University, *http://www.apl.hju.edu/Classes/Notes/Campbell/Joe/celp_3.2a_tar.Z*; Streams 3 through 7 were obtained from Cybernetics InfoTech. Inc., *http://www.cybit.com*; and Streams 8 was obtained from VoiceAge Corp., *http://www.voiceage.com*.

signed schemes to be general enough for different voice characteristics. For each stream, the table shows its features, such as its length and speakers. All streams are of 16-bit Pulse Code Modulation (PCM) format sampled at 8 kHz.

In Chapter 4, we study both coder-independent and coder-dependent MDC. Experimental results show that coder-independent MDC, or sample interleaving, performs rather poorly when applied in low bit-rate coders. By studying parameter correlations in low bit-rate coders, we propose a new parameter-based MDC scheme by tightly coupling the processes of coding and loss concealment. We demonstrate the effectiveness of our proposed scheme on three low bit-rate speech-coding standards. Chapter 4 also studies our proposed MDC scheme by comparing three common representations of the most important component of low bit-rate speech coders — the vocal-tract model. The comparisons are conducted from different perspectives: spectral distortion, correlation, and objective quality. Although different representations seem to give similar results, one of them — Line Spectral Pair (LSP) — gives slightly better reconstruction quality.

Two quality measures need to be considered for MDC: reconstruction quality in case of loss, and decoding quality that is loss independent. In order to improve the quality of our proposed MDC strategy for both quality measures, Chapter 5 investigates parameter reconstruction in the first case and parameter generation in the second case for low bit-rate coded speech.

Finally, Chapter 6 summarizes the work of this dissertation.

1.4 Significance of This Work

We summarize in this section the main contributions of this dissertation.

a) *A comprehensive study of voice-traffic behavior over the Internet.* Based on statistics collected on connections with different characteristics, from low loss to high loss and from domestic connections to international connections, we have found that traffic behaviors are time varying and connection dependent. Even for a normally low-loss connection, losses may be high sometimes. The fact that packet losses generally occur in small bursts makes interleaving a particular attractive loss-concealment scheme.

b) *A novel MDC scheme for loss concealment of low bit-rate coded speech.* We have focused our study on the latest coding standards that operate on low/very-low bit rates. Loss concealments of low bit-rate coded speech are difficult because the loss of one packet usually causes many subsequent packets not decodable. As there are no good MDC schemes designed for low bit-rate coded speech that does not require extra bandwidth, we start by studying the correlations of coding parameters. This leads to a simple yet

effective scheme that interleaves in the parameter domain instead of the sample domain. Without requiring extra bandwidth, our scheme achieves good transmission qualities on both synthetic tests and real Internet tests.

c) *Reconstruction-quality comparisons of various representations of the linear prediction model for low bit-rate coded speech.* Various representations studied extensively in the literature [18, 19, 20, 21, 22, 23] aim to find the best representation that approximates transition regions inside a speech stream. In contrast, our goal is to find a good representation that gives the best reconstruction quality in case of loss. Our results show that the LSP representation gives the best performance.

d) *LSP reconstruction.* Our proposed scheme reconstructs lost LSP vectors through interpolations in case of loss. We have studied the design of various interpolation schemes based on spectral distortion. They are done by representing spectral distortion in terms of LSP, and by designing interpolation schemes that minimize the second-order approximation of spectral distortion. Our experimental results show that interpolations by simple averaging perform similarly to optimal first-order and second-order interpolations.

e) *Improving the quality of MDC parameter generation.* Our proposed MDC scheme achieves high reliability by sacrificing some quality. Specific to low bit-rate coders, parameters other than LSP have lower quality, which in turn cause lower decoding quality even in case of no loss. By classifying distortions in the frequency domain, we have developed a new spectral shaping scheme for MDC. Extensive tests on the new scheme have shown improved performance of MDC.

f) *Real-time voice-transmission prototype.* We have built an end-to-end software transmission prototype. Further, we have designed a trace-driven approach that reads collected traffic traces in our experiments and delays and drops packets just as a real network would. Such an approach allows various strategies to be compared under the same Internet environment.

CHAPTER 2

PREVIOUS WORK

Voice transmissions over packet networks have attracted a lot of attention recently with the availability of fast processors and ever increasing demands. However, their delivery quality is not satisfactory due to frequent packet losses. An active research direction is, therefore, to develop simple and robust loss-concealment and coding strategies at connection end points. These end-to-end approaches generally exploit redundancies in voice data and reconstruct lost data from that received. In this chapter we summarize the major work in this area.

Based on interactions with source coders, we classify existing techniques into source coder-independent schemes that treat underlying source coders as black boxes, and source coder-dependent schemes that exploit coder-specific characteristics to perform reconstruction. Here, "source coder," same as "codec," refers to a combination of encoding and decoding modules. As shown in Figure 2.1, both categories can be divided into receiver-based and sender/receiver-based subcategories, according to where loss concealment is

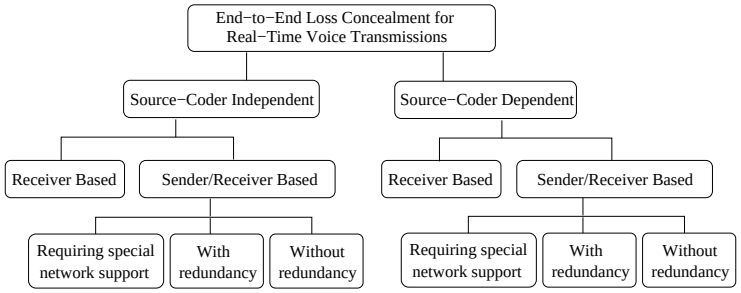


Figure 2.1: Classifications of loss-concealment strategies for real-time voice transmissions.

carried out. For sender/receiver-based schemes, we further categorize them according to their requirement for special network supports and that for extra bandwidth.

2.1 Coder-Independent Loss-Concealment Schemes

2.1.1 Receiver-based schemes

These schemes perform loss-concealment actions only at receivers by trying to guess the lost speech waveform. A naive way is to recreate lost packets by padding silence or white noise [24], whereas a better way is to repeat the last received packet [25], although it may cause echoes.

Wasem *et al.* [26] have tested more sophisticated methods. One is *pattern matching* that extracts packet-length segments from received speech and uses them to replace the missing packets. Another method, *pitch waveform replication*, estimates the pitch of received speech and replicates prior pitch waveforms for the duration of a missing packet.

Both methods can be used to maintain phase continuity at the boundaries of substituted packets and prior received packets. Tests have shown that pitch waveform replication is more effective.

The problem with the above methods is that they may introduce discontinuities at the end of reconstructed packets, although they take care of continuity at the beginning. Valenzuela and Animalu [27] have suggested eliminating such discontinuities by modifying the time scale of the reconstructed waveform so that the phase at the end of the missing packet matches that at the beginning of the following packet. Still, only prior received packets are effectively used in the reconstruction process.

Tang [28] has gone a step further by recognizing that the missing speech waveform has resemblance to both the past and future waveforms of a missing interval. Therefore, speech waveforms in the past and future are used to reconstruct waveform in the missing interval. This is called *double sided waveform substitution* (DSWS). *Double sided periodic substitution* (DSPS) enhances DSWS by iteratively substituting waveforms in the missing interval if this interval is greater than the pitch periods of the preceding and following received waveforms.

Sanneck *et al.* [29] have presented another waveform substitution technique, namely, *waveform similarity overlap-add*. Distortions due to discontinuities at the boundary of the substituted and the received packets are reduced by overlapping them using a Hanning window.

Two major problems encountered by the above methods are that periodical reconstruction is inappropriate for partially or unvoiced speech, and good results only occur

for frame sizes up to 10 ms. To address these problems, Cluver and Noll [30] have developed a finer waveform substitution technique that generates substitution signals for partially voiced speech by performing separate voicing decisions in different sub-bands. At a receiver, for each received frame, speech is passed through an analysis linear predictor, and the resulting linear predictor residues are then split into eight sub-bands. Each sub-band is classified as either voiced or unvoiced. Note that voicing decisions are made on received frames. For lost frames, voicing decisions of the previous received frames give guidelines for substitution: for voiced sub-bands, the previous sub-band signals are repeated periodically; for unvoiced sub-bands, white noise signals are used. The sum of the sub-band signals forms the substitute excitation signals, and the synthesis filter coefficients of the last frame are retained.

Because detailed information of lost frames is not available, these strategies work well when losses are infrequent and when packet sizes are small, generally covering a duration of 16 ms [31]. They are not promising for use over the Internet that may have a high probability of loss.

2.1.2 Sender/receiver-based schemes

In the following schemes, senders and receivers cooperate in loss concealment. These schemes are usually more effective because senders can convey knowledge about lost packets to receivers. Hence, receivers can better estimate lost packets than when using just receiver-based schemes.

Schemes requiring special network support. These schemes require packets to be demarcated from underlying networks and are considered as having very limited use over the Internet because such service is not available. These schemes assign different priorities to different audio packets, and require an underlying network to drop packets according to their priorities when congestion happens. Yong [32] has proposed several methods. In one of the methods that prioritizes speech frames, all very-low-energy frames, mostly containing silence, are given low priority. Onset speech frames, *i.e.*, the beginning of a talk spurt, and highly transitional frames are assigned high priority because they cannot be reconstructed using previously received frames. Frames having little change from their preceding ones are given low priority, but no more than two consecutive frames can all have low priority. For all other frames, their priority alternates between high and low. Using this method, speech frames are divided roughly into 40% of low-priority frames and 60% of high-priority frames. Another method uses the analysis-and-synthesis approach. Reconstruction performance is first tested at the sender side for every frame using the previous frame. Those frames showing large reconstruction errors are assigned high priority. Similar methods were proposed in [33, 34].

Retransmission schemes. By controlling the playback time for the first packet in each talk spurt, schemes in this category manage to perform timely retransmissions of lost packets [35, 36, 37]. However, they are designed for local-area networks with short delays and, therefore, are not suitable for the Internet.

The above priority-based and retransmission-based schemes are not designed for transmissions over the Internet that does not provide QoS support. For schemes targeting Internet transmission, we further classify them into those requiring redundancies and those that do not.

Schemes adding redundant information. Some methods add redundancy uniformly at the packet level, leading to unnecessarily high bandwidth usage. The most naive method uses two copies of each packet and forces them to be sent from a source to a destination along two disjoint paths in order to minimize loss [38]. Besides requiring double bandwidth, it is hard to force packets to traverse different paths in the Internet. Another variation of this method adds copies of the previous k frames in the packet containing the current frame [39]. For example, when $k = 2$, packet n will contain frames $n, n-1$, and $n-2$. Obviously, in this case, if packet n is lost, we can still reconstruct it from packet $n+1$ or $n+2$. Nevertheless, this method also increases bandwidth usage by about k times.

The *forward error correction* (FEC) approach directly extends the traditional bit-error correction algorithm to packet-loss recovery. Shacham [40] presented a technique in which the sender adds an error-control packet consisting of parity bits for each block of k data packets. If a receiver identifies missing packets by gaps in the arriving sequence, it can recover any single missing packet in a block by using the parity bits of the k remaining packets. As the author has pointed out, an erroneous or missing packet needs to be delayed until the entire block of packets arrive. Also, adding a parity packet to

each block of k packets increases the rate of packets in the network by $1/k$, which in turn increases the packet-loss probability. As a matter of fact, reducing delays and reducing bandwidth used are contradictory requirements. Lowering k will decrease delay but will increase bandwidth. Further, bursty losses of length greater or equal to two cannot be recovered. An extension of this method to generate parity to cover two or more lost packets is discussed in [41] by adding more parity packets.

Bolot and Garcia [42, 43] have developed another redundancy scheme in which packet n includes, in addition to its encoded samples, a redundant version of packet $n-1$ obtained by a low bit-rate coder. This method requires less overhead than the previous method, but also has a lower quality in case of loss because only a coarser version of the lost packet can be obtained. Clearly, this method can only recover from isolated losses.

The above methods do not exploit any special characteristics in voice streams and treat them as ordinary bit sequences. Hence, they need to add redundant information uniformly and require considerable increases in bandwidth.

Another class of schemes add excerpted speech information as redundancy. These schemes attempt to improve the blind reconstruction of receiver-based waveform substitution. There are three methods in this class:

a) *Classified waveform substitution* fragments speech samples into smaller segments and classifies each frame into background noise, voiced speech, fricatives, and "others." The classification is sent redundantly to receivers. The reconstruction of a lost frame at a receiver can, thus, be based on the lost frame's class and not on its previous frame's class [44]. Although the added redundancy here is minimal, the reconstruction quality

is limited because padding is still blind. Another difficulty with this method is the high packet rate required, typically, about 125 packets per second.

b) *Redundant pitch-guided waveform substitution* sends extra pitch structure as part of its side information. For example, Choi and Constantinides [45] have used this extra information as an indication of how a missing segment can be filled. The output speech segment corresponding to the lost packet is generated by repeating the previously decoded signal pitch period. As the authors have pointed out, this technique is based on the assumption that speech properties do not change drastically over the period of packet loss. Significant changes in the characteristics of speech in the packet lost would obviously cause failures in compensation. Since such a problem exists for all waveform substitution methods, it restricts the packet length to be around 16 ms, leading to a packet rate of 63 packets per second that is still too high for the Internet.

c) *Short-time energy or zero-crossing guided waveform substitution* puts additional information, such as short-time energy and zero-crossings, in the current packet into the next packet [46]. These methods divide a packet into small segments and generate a short-time energy value for each segment. Additionally, for each sample, one bit is used to indicate the absence or presence of a zero-crossing at that point. If a packet is lost, the short-time energy values are interpolated and extended to the length of the packet lost, followed by the generation of a constant amplitude sinusoid that passes through all the zero-crossings. The reconstructed signals are the product of the two signals generated. To save some bandwidth, the authors have suggested using the total number of zero-crossings plus one extra bit to indicate whether the packet contains voiced

or unvoiced speech, instead of one extra bit for each sample. In case of loss, if the lost packet contains voiced speech, a constant frequency sinusoidal waveform is generated; otherwise, the sinusoid is generated to pass through some randomly generated zero-crossings. The overhead reported is about 12.5%. This method differs from the above waveform substitution methods in that it does not use other received speech frames to aid reconstruction. However, it is questionable whether short-time energy and zero-crossings are good abstraction of speech fragments.

The above methods try to find good trade-offs between redundancy and reconstruction quality, but do not add redundancy uniformly in order to take advantage of speech properties. Still, they rely on a strong continuity of speech streams and require the packet size, or loss duration, to be small. In general, good quality is hard to achieve unless a considerable amount of redundant information is sent.

Schemes without adding redundant information. These schemes exploit implicit redundancies in voice streams more effectively and allow senders to take more responsibility in loss concealment.

Adaptive packetization and concealment [47] was proposed by Sanneck. A sender estimates pitch periods and puts two pitch periods of audio chunks into one packet. This results in small packets sent for voiced speech and large ones sent for unvoiced speech. When loss is detected at a receiver, adjacent speech chunks are reused. One problem with this method is that accurate pitch and voicing information for lost packets are not available. Also, improvements are limited because it cannot recover other important

information in lost packets, such as amplitudes and spectral changes. Further, it only applies to compression strategies that allow variable frame lengths.

Multiple-description coding (MDC) is a simple yet effective scheme for concealing loss. It divides data into equally important substreams in such a way that the decoding quality using any subset is acceptable, and that better quality is obtained by more descriptions. It is assumed in MDC that losses to different descriptions are uncorrelated and that the probability of losing all the descriptions is small. According to this definition, all the methods discussed above belong to the class of *single-description coding* (SDC).

As early as 1981, Jayant and Christensen [48] have proposed to use *interleaving*, or sample-based MDC, to conceal packet losses. Their procedure is based on arranging a 2B-block of samples into two B-sample packets, an odd-sample packet and an even-sample packet. If the even (resp. odd) packet is lost, the odd (resp. even) samples of the 2B-block are estimated from the even (resp. odd) samples by means of first-order Wiener interpolations. This method effectively converts bursty losses in the sample domain to random losses. Suzuki and Taka [24] have investigated odd-even sample interpolations and have concluded that the performance of optimal and average interpolations are similar. To further improve reconstruction quality, the following sample-interpolation schemes have been proposed:

a) Yuito and Matsuo [49] have presented a pattern-matching sample-interpolation method based on interleaving. The receiver uses a template consisting of a segment of samples right before the lost sample and a segment of samples right after the lost sample. The waveform that best matches the template is searched in a window of samples just

before the template. Next, the amplitude of the best matched waveform is adjusted to match that of the template by multiplying a scale factor. Finally, the lost sample is calculated using second-order divided differences in order to achieve a smooth waveform. Note that all samples in the template are assumed to be received.

b) *Kalman-based sample-interpolation* interpolates received samples using a Kalman filter in order to approximate the lost ones [50]. To avoid adding redundancy, Kalman filter coefficients are computed from the received incomplete speech packets. As the author has pointed out, the gain of this method is mainly for large interleaving factors, and packet-loss rates have more effect on reconstruction quality because only partial statistical information is used to generate interpolation coefficients.

Although coder-independent approaches do not require modifying the underlying compression scheme, they may have inferior reconstruction quality when combined with speech coding, because they are unaware of the requirements of the underlying coding scheme. For example, they cannot solve the error propagation problem in low bit-rate coded speech. Therefore, in the following, we examine existing coder-dependent loss-concealment schemes.

2.2 Coder-Dependent Loss-Concealment Schemes

All the methods discussed in this section exploit source-coder properties for loss concealment. They fall into either receiver-based or sender/receiver-based schemes.

2.2.1 Receiver-based schemes

The following are some schemes designed specifically for linear predictive coders.

a) Cox *et al.* [51] have demonstrated that using relatively simple measures, such as repeating the coder parameters from the most recent error-free frame, can effectively conceal error frames. It is easy to see that this method can be applied to conceal lost frames. The full-rate Global System for Mobile (GSM) telecommunication coder takes a similar approach and reconstructs lost frames by repeating the parameters of the previous frame with scaled-down gains [52]. In case of loss of multiple contiguous frames, this has the effect of muting the output. Likewise, ITU G.723.1 pads a lost frame by simulating the vocal characteristics of the previous frame and by slowly damping the signals [53]. These schemes can be used to conceal random losses up to 10% [39].

b) Like coder-independent receiver-based schemes, the above simple replication schemes can be improved by classifying previously received packets as voiced, unvoiced, or partially voiced speech. Excitations take the form of random noise if the previous packet is unvoiced. If the previous packet is partially voiced, speech signals are divided into sub-bands and recovered separately for each sub-band. Otherwise, for a voiced packet, excitations are replaced by the peaks of the previous frame's excitations [54, 55]. Such methods still have difficulties in dealing with high loss rates, as generally losses greater than 10% cannot be handled.

The receiver-based schemes described above are not scalable under high probabilities of loss that may happen in the Internet. Further, they do not perform well under bursty losses, losses occurring during voice transitions, and large packets.

2.2.2 Sender/receiver-based schemes

In this class of schemes, a sender will manipulate speech data before sending it, with a goal of helping receivers to better reconstruct lost packets. Methods in this category can be further divided into a) schemes requiring special network support, b) schemes with redundant information, and c) schemes without redundant information.

Schemes requiring special network support. When speech signals are coded by waveform coding and the underlying network has priority support, the most significant bits (MSB) and the least significant bits (LSB) can be interleaved into different packets, and the MSB packets are sent with higher priorities. In case of congestion, the network can drop some or all LSB packets [24, 56].

For linear predictive coders, Yong [32] has suggested splitting information bits in a frame into two groups of different levels of importance. One possible method gives high priority to linear prediction coefficients and pitch parameters, and low priority to excitation signals. When packets containing excitation signals are lost, a receiver uses zero excitation or white noise to excite its speech synthesis filter. Clearly, this method cannot restore decoding states properly. Another method is based on a two-stage coder that codes input speech in its first stage and the resulting quantization residual signals

in its second stage. Since the second stage reuses parameters, such as linear prediction coefficients, pitch period, and gain, generated in the first stage, it can spend more bits on excitation codewords. If no loss happens, the decoder receives codes for both the first and second stages, decodes each of them separately, and adds the two resulting signals to produce an output signal of high quality. If loss happens, the decoding quality of the first stage can be maintained since its packets are assigned high priority and assumed always delivered successfully. This scheme not only needs priority supports from the network, but also requires a special coding scheme.

In general, schemes in this class suffer from the same drawback as their coder-independent counterparts because priority support is not available in the Internet. In the following, we discuss methods that do not require such special network supports.

Schemes adding redundant information. Some methods extend traditional channel-coding methods for bit errors to packet losses and modify them to suit different source coders. *Unequal error protection* codes the most sensitive coding parameters, such as the MSBs of waveform-coded streams and the linear-prediction coefficients in linear-predictive coders, using FEC [57].

Without using FEC, there are two ways of adding redundant protection in waveform coders:

a) *Redundant MDC* aims to improve the reconstruction quality of two-way sample-based MDC. It pre-computes the reconstruction errors of even (resp. odd) samples in the odd (resp. even) description at the sender side and inserts this extra information into

the odd (resp. even) description [58]. Its drawback is that this extra information needs considerable bandwidth of about 15% to 45%.

b) *Unbalanced MDC* assumes that it is possible to decompose waveform-coded speech samples in such a way that one description is more important than the other. It then codes the less important description, that can be used to correct coding errors of the more important description at low resolution if both were received [59]. The difference between unbalanced MDC and priority-based schemes is that the less important description can be decoded independently and is of acceptable quality to a receiver. The recovery algorithm uses a deterministic distance measure to find the most likely waveform for lost data, given the received data and the side information. The authors did not mention the bandwidth allocation between two descriptions and the extra bandwidth needed. Further, the scheme is hard to extend to more than two descriptions.

Because loss has a worse effect on linear predictive coders than on waveform coders, considerable attention has been paid to loss concealments for linear predictive coders. The following are some typical schemes of sending extra information to aid recovery at receivers.

a) *Recovery by re-initialization* is an approach proposed to solve the problem of error propagation of linear predictive coders when a packet is lost. The method synchronizes the coder and the decoder by re-initializing their states periodically at both ends. However, the quality of the method is sensitive to whether the frame following re-initialization is received. The authors have suggested sending redundant information about re-initialized frames in order to guarantee the availability of state information [60].

Still, this method cannot avoid error propagations because packets may be lost in between synchronizations. Moreover, synchronizations cannot be too frequent because they degrade coding quality significantly.

b) *Diversity* schemes generate several redundancy levels with decreasing bandwidth requirements for linear predictive coders [61]. The first level of redundancy consists of linear predictive coefficients, pitch lags, adaptive codebook gains, fixed codebook gains, and the first few or all of the fixed codebook pulses. In this level, two descriptions are almost exactly same. Unlike the first level, the second level interleaves fixed codebook pulses and replicates all other information in both descriptions. The third level further interleaves both fixed codebook pulses and gains, but still replicates all remaining parameters. Obviously, reconstruction quality is best for the first level. However, its bandwidth requirement is almost doubled when compared to schemes that require no redundant information. For the second and third levels, fixed codebook information cannot be reconstructed when loss happens.

c) *Redundant MDC* sends the base or important information and a subset of enhancement information in one packet, and the same base or important information and a complimentary subset of enhancement information in the next packet [61]. Obviously, the base or important information is duplicated. The base information always consists of linear prediction coefficients, pitch lags, adaptive codebook gains, and fixed codebook gains. It may or may not have a subset of fixed codebook pulses. In contrast, the enhancement information consists of the remaining pulses of the fixed codebook that are interleaved. The method differs from diversity schemes only in that it always interleaves

the fixed codebook pulses. Consequently, this method still requires considerable increases in bandwidth over non-redundant schemes. The authors have reported bandwidth overheads similar to those of the diversity schemes.

Schemes not requiring redundant information. One group of methods first code speech signals using waveform coders and then interleave them. To be able to reconstruct lost samples, one method designs reconstruction filters, assuming that input voice streams can be modeled by first-order Gauss-Markov distributions [62]. However, the generation of the reconstruction filter assumes that previous samples are available, which may not be possible in a real network environment. Further, the assumption that input voice streams can be modeled by first-order Gauss-Markov model is not general. Another method performs a tree search to find the best-matched quantized samples in case of loss [63]. It can only be applied in cases in which coded samples have just a few discrete values. Moreover, it is computationally expensive even to get suboptimal assignments.

In general, no existing nonredundant-MDC scheme has been developed for linear-predictive coders. Since most of the current low/very-low bit-rate speech-coding standards are based on linear prediction [64], one objective in this dissertation is to design MDC methods specifically for low bit-rate speech coders in order to recover lost information.

2.3 Summary

The end-to-end loss concealments of real-time voice transmissions can be either coder independent or coder dependent. The coder-independent approach is easy to implement because it does not require any modification of the coding process and can be standard compliant. On the other hand, the coder-dependent approach utilizes some special properties of underlying coders and will be able to achieve better reconstruction quality. In both approaches, MDC without added redundancy is attractive because it converts bursty losses to isolated losses and in turn ease reconstruction.

Previous work, however, has been focused more on loss concealments for waveform-coded or uncompressed voice transmissions that require higher bandwidth. Results for contemporary low bit-rate coded speech are limited and not satisfactory. Especially, there is no existing nonredundant coder-dependent MDC for low bit-rate coders.

In this research, we target at concealing packet losses for low bit-rate coded speech. We evaluate the performance of coder-independent MDC and design coder-dependent MDC to achieve better reconstruction quality.

CHAPTER 3

IP VOICE TRAFFIC LOSS CHARACTERISTICS

This chapter presents a series of Internet traffic experiments in order to reveal the underlying loss patterns for voice traffic over the Internet and to guide us through the design of loss-concealment strategies. As discussed before, IP networks were built to support non-real-time applications such as file transfer or e-mail. These applications are inherently different from voice transmissions. For example, transmission losses that pose little threat to non-real-time data traffic can introduce severe problems to real-time packetized voice traffic. Further, voice traffic is periodic in nature, while data traffic is usually bursty in nature. Another purpose of our experiments is to design an environment for evaluating our proposed loss-concealment schemes in later chapters.

3.1 Traffic Collection Setup

In order to cover a variety of voice traffic behaviors over the Internet, we have chosen sites with loss rates ranging from low to high and including both domestic and international destinations. Table 3.1 lists the six remote hosts involved in our experiments.

Table 3.1: Hosts used in our Internet traffic experiments.

Host Name	IP Address	Location	Loss characteristic
`s700.di.uminho.pt`	193.136.20.1	Portugal	low
`daedalus.cs.berkeley.edu`	169.229.62.38	Western U.S.	low to medium
`potato.claes.sci.eg`	193.227.5.37	Egypt	low to medium
`dns.gdcc.edu.cn`	202.116.48.8	Southern China	medium to high
`www.lzu.edu.cn`	202.201.0.152	Western China	medium to high
`us.svf.stuba.sk`	147.175.16.9	Slovakia	high

Among the six connections, the UIUC-Portugal connection has low loss rates, the UIUC-Berkeley and UIUC-Egypt connections have low to medium loss rates, the UIUC-S.China and UIUC-W.China connections have medium to high loss rates, and the UIUC-Slovakia has almost constant and high loss rates. Their physical locations cover domestic site, South America, Africa, Europe, and Asia.

Figure 3.1 shows the configuration of our voice traffic trace collector. Because we have no control over the remote computers, we take advantage of the well-known echo service to collect traffic traces. For each remote computer, a computer at UIUC periodically sent 2000 UDP probe packets, at a rate of 30 packets per second and 500 bytes per packet, to the echo port of a remote computer in order to simulate a typical voice application. The sender, at the same time, recorded sending time of each packet. (The sending rate and packet size were chosen to reflect the upper bound on traffic in voice communications over the Internet. The packet size was chosen to be smaller than the Maximum Transfer Unit (MTU) of the Internet in order to avoid fragmentation. Results on other packet transmission rates can be found elsewhere [65, 66].) Another process in the same local computer received the packets echoed back, and recorded the receiving time and sequence

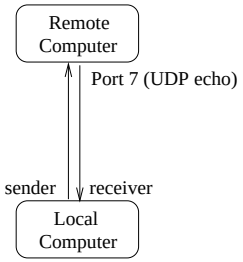


Figure 3.1: Voice-traffic trace collector.

number of each echoed packet. Because the sending and receiving processes were in the same local computer, there was no need for clock synchronization when calculating round-trip times (RTT). The process was repeated at the beginning of each hour over a 24-hour period during the first week of November 2001 for each remote computer. All the statistics in the following sections are drawn from these collected traces.

3.2 Loss Statistics

In this section, we first report the overall loss rates of the six connections. However, loss rates alone are not helpful for designing loss-concealment strategies. Therefore, we further study the bursty behavior of packet losses and examine the effectiveness of interleaving to convert bursty losses to isolated losses.

3.2.1 Loss rates

We first compute the loss rates for all the test sites. As shown in Figure 3.2, all connections exhibit packet losses, and losses may approach 50% for some connections. The loss behavior depends not only on the network connection and the remote location, but also on the time of the experiments.

The UIUC-Portugal connection (Figure 3.2a) exhibits small amount of losses (below 10%). Under such low-loss conditions, voice transmissions without loss recovery may be practical. However, Figure 3.2b shows that, even for U.S. domestic connections, loss rates sometimes can go up to around 25%. The loss percentages for other connections (Figure 3.2c - 3.2f) are even higher. For example, the UIUC-Slovakia connection incurred losses on almost half of its total packets. This high loss rate will be very challenging for real-time system designers. Without loss recovery, we can expect the transmission quality to be unpredictable and poor.

The results also reveal the relationship between the loss percentage and the time of day. For all our test sites except Portugal and Slovakia, there were time periods with loss rates much higher than average. For both sites in China, the loss rate could sometimes be 20% higher than normal. These results demonstrate that, even for a specific connection, its loss fraction may fluctuate, depending on how busy a server is at a particular time and how congested the network is.

In summary, loss percentage is connection dependent and time varying. This means that both senders and receivers must have the ability to adapt to different loss scenarios.

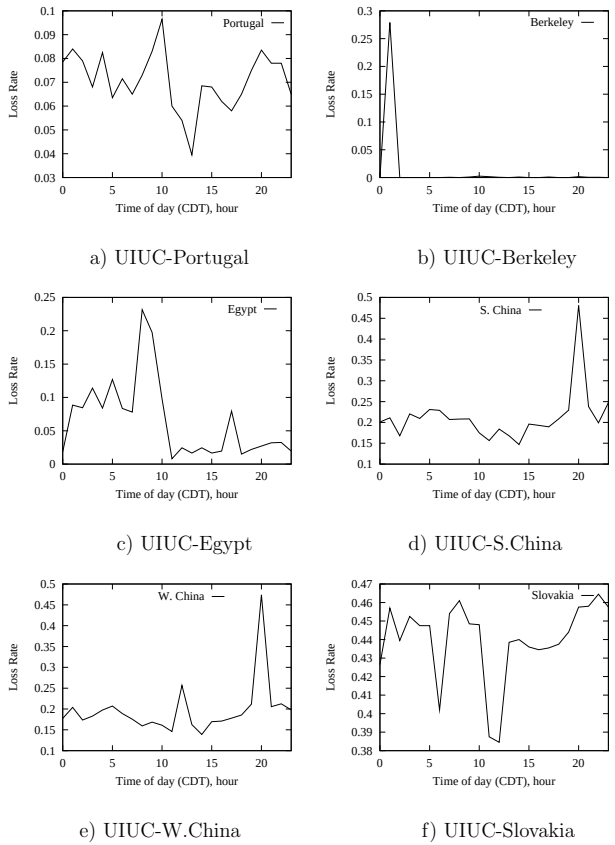


Figure 3.2: Loss rates in round-trip paths between UIUC and six remote locations.

To achieve this, the system needs to know the current network behavior and incorporate statistics collected at run-time to adjust its operations accordingly.

3.2.2 Burst-length distribution

Figure 3.3 plots the distributions of consecutive packet losses, or burst length. For all the six destinations, burst lengths of 1 and 2 were predominant. For sites in Portugal, Berkeley, and even Slovakia, the highest-loss site, bursty losses with length one took more than 90% of all bursty losses. For sites in Egypt, Southern China, and Western China, more than 80% of the losses were of burst length one or two.

Tables 3.2a thru 3.7a list, for the six connections, the conditional probability of the next burst length, given the current burst length. For all these connections, losses with burst length longer than three happened very infrequently, and one long burst did not imply that the next burst would also be long. For example, for the UIUC-Slovakia connection, the unconditional probability for the current burst length to be four and the next burst length to be larger than or equal to four was only $0.004 \cdot (1 - 0.667) = 0.001$. The unconditional probabilities of the next burst length for the six connections are listed in Tables 3.2b thru 3.7b. The highest unconditional probability for the current burst length to be four and the next burst length to be larger than or equal to four was only 0.004.

The fact that burst lengths were usually small (similar results have been shown in [67, 43]) leads us to conclude that interleaving can be a good method to ease reconstruction. When the burst length is less than the interleaving factor, there are always parts of the

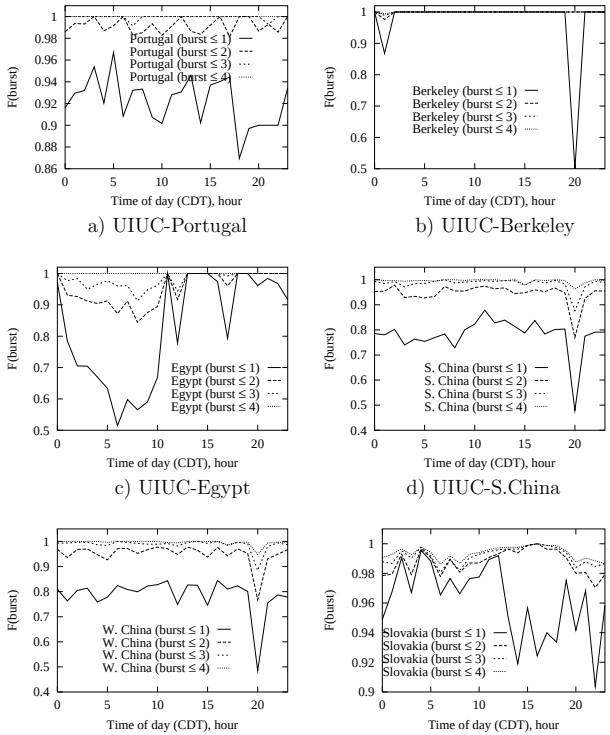


Figure 3.3: Cumulative distributions of raw bursty losses ($F(burst)$) in round-trip paths between UIUC and six remote locations.

Table 3.2: Relation between the current and adjacent burst lengths for $n_p = 2000$ packets sent from UIUC to Portugal and back with 9.7% loss rate at 10 a.m CDT on November 3rd, 2001.

a) The conditional probability distribution of the next burst length, given the current burst length in the first element in that row. The second element in each row represents the probability of occurrences of the burst length listed in the first element.

Current burst length	Fraction of occurrence	Conditional distribution of next burst length		
		1	2	3
1	0.902	0.897	0.987	1.000
2	0.081	0.929	0.929	1.000
3	0.017	1.000		

b) The unconditional probability distribution of the next burst length.

Current burst length	Uncond. prob. of next burst length		
	≥ 1	≥ 2	≥ 3
1	0.902	0.093	0.012
2	0.081	0.006	0.006
3	0.017		

Table 3.3: Relation between the current and adjacent burst length for $n_p = 2000$ packets sent from UIUC to Berkeley and back with 27.9% loss rate at 1 a.m CDT on November 7th, 2001.

a) The conditional distribution of the next burst length, given the current burst length in the first element in that row. The second element in each row represents the probability of occurrences of the burst length listed in the first element.

Current burst length	Frac. of occ.	Conditional prob. dist. of next burst length						
		1	2	3	4	5	9	18
1	0.867	0.872	0.975	0.987	0.992	0.995	0.997	1.000
2	0.108	0.860	0.980	1.000				
3	0.013	0.833	1.000					
4	0.004	0.000	1.000					
5	0.002	1.000						
9	0.002	1.000						
18	0.002	1.000						

b) The unconditional probability distribution of the next burst length.

Current burst length	Uncond. prob. of next burst length						
	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5	≥ 9	≥ 18
1	0.868	0.111	0.022	0.011	0.007	0.004	0.002
2	0.108	0.015	0.002				
3	0.013	0.002					
4	0.004	0.004					
5	0.002						
9	0.002						
18	0.002						

Table 3.4: Relation between the current and adjacent burst lengths for $n_p = 2000$ packets sent from UIUC to Egypt and back with 23.1% loss rate at 8 a.m CDT on November 5th, 2001.

a) The conditional distribution of the next burst length, given the current burst length in the first element in that row. The second element in each row represents the probability of occurrences of the burst length listed in the first element.

Current burst length	Fraction of occurrence	Conditional distribution of next burst length			
		1	2	3	4
1	0.566	0.558	0.827	0.891	1.000
2	0.279	0.566	0.868	0.961	1.000
3	0.069	0.684	0.895	0.895	1.000
4	0.087	0.500	0.833	0.917	1.000

b) The unconditional probability distribution of the next burst length.

Current burst length	Uncond. prob. of next burst length			
	≥1	≥2	≥3	≥4
1	0.565	0.250	0.098	0.062
2	0.279	0.121	0.037	0.011
3	0.069	0.022	0.007	0.007
4	0.087	0.043	0.014	0.007

Table 3.5: Relation between the current and adjacent burst length for $n_p = 2000$ packets sent from UIUC to Southern China and back with 48% loss rate at 8 p.m CDT on November 6th, 2001.

a) The conditional distribution of the next burst length, given the current burst length in the first element in that row. The second element in each row represents the probability of occurrences of the burst length listed in the first element.

Current burst length	Frac. of occ.	Conditional prob. dist. of next burst length							
		1	2	3	4	5	6	7	9
1	0.479	0.477	0.743	0.861	0.954	0.979	0.992	0.996	1.000
2	0.291	0.465	0.778	0.875	0.972	0.993	0.993	1.000	
3	0.107	0.462	0.904	0.981	0.981	1.000			
4	0.087	0.512	0.791	0.930	0.977	0.977	0.977	1.000	
5	0.020	0.400	0.500	0.600	1.000				
6	0.008	1.000							
7	0.006	0.333	0.333	0.333	0.667	0.667	1.000		
9	0.002	1.000							

b) The unconditional probability distribution of the next burst length.

Current burst length	Uncond. prob. of next burst length							
	≥1	≥2	≥3	≥4	≥5	≥6	≥7	≥9
1	0.479	0.251	0.123	0.067	0.022	0.010	0.004	0.002
2	0.291	0.156	0.065	0.036	0.008	0.002	0.002	
3	0.107	0.058	0.010	0.002	0.002			
4	0.087	0.042	0.018	0.006	0.002	0.002	0.002	
5	0.020	0.012	0.010	0.008				
6	0.008							
7	0.006	0.004	0.004	0.004	0.002	0.002		
9	0.002							

Table 3.6: Relation between the current and adjacent burst length for $n_p = 2000$ packets sent from UIUC to Western China and back with 47.3% loss rate at 8 p.m CDT on November 4th, 2001.

a) The conditional distribution of the next burst length, given the current burst length in the first element in that row. The second element in each row represents the probability of occurrences of the burst length listed in the first element.

Current burst length	Frac. of occ.	Conditional prob. dist. of next burst length									
		1	2	3	4	5	6	7	8	9	11
1	0.485	0.506	0.788	0.879	0.939	0.961	0.983	0.987	0.991	0.996	1.000
2	0.282	0.474	0.733	0.867	0.933	0.963	0.970	0.993	1.000		
3	0.117	0.500	0.768	0.911	0.982	0.982	1.000				
4	0.063	0.400	0.767	0.933	0.967	0.967	1.000				
5	0.019	0.444	0.667	0.889	1.000						
6	0.017	0.375	0.625	0.875	1.000						
7	0.008	0.750	1.000								
8	0.004	0.500	1.000								
9	0.002	0.000	1.000								
11	0.002	0.000	1.000								

b) The unconditional probability distribution of the next burst length.

Current burst length	Uncond. prob. of next burst length									
	≥1	≥2	≥3	≥4	≥5	≥6	≥7	≥8	≥9	≥11
1	0.485	0.240	0.103	0.059	0.029	0.019	0.008	0.006	0.004	0.002
2	0.282	0.149	0.075	0.038	0.019	0.010	0.008	0.002		
3	0.117	0.059	0.027	0.010	0.002	0.002				
4	0.063	0.038	0.015	0.004	0.002	0.002				
5	0.019	0.010	0.006	0.002						
6	0.017	0.010	0.006	0.002						
7	0.008	0.002								
8	0.004	0.002								
9	0.002	0.002								
11	0.002	0.002								

Table 3.7: Relation of the current and adjacent burst length for $n_p = 2000$ packets sent from UIUC to Slovakia and back with 45.7% loss rate at 8 p.m CDT on November 2nd, 2001.

a) The conditional distribution of the next burst length, given the current burst length in the first element in that row. The second element in each row represents the probability of occurrences of the burst length listed in the first element.

Current burst length	Frac. of occ.	Conditional prob. dist. of next burst length									
		1	2	3	4	5	6	7	8	10	13
1	0.942	0.962	0.992	0.992	0.993	0.996	0.997	0.999	1.000		
2	0.039	0.710	0.839	0.839	0.903	0.935	0.968	0.968	0.968	1.000	
3	0.004	0.333	0.667	1.000							
4	0.004	0.333	0.667	0.667	0.667	0.667	0.667	0.667	1.000		
5	0.004	0.667	0.667	1.000							
6	0.002	0.500	1.000								
7	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
8	0.002	1.000									
10	0.001	0.000	0.000	1.000							
13	0.001	0.000	1.000								

b) The unconditional probability distribution of the next burst length.

Current burst length	Uncond. prob. of next burst length									
	≥1	≥2	≥3	≥4	≥5	≥6	≥7	≥8	≥10	≥13
1	0.942	0.036	0.007	0.007	0.006	0.004	0.002	0.001		
2	0.039	0.011	0.006	0.006	0.004	0.002	0.001	0.001	0.001	
3	0.004	0.002	0.001							
4	0.004	0.002	0.001	0.001	0.001	0.001	0.001	0.001		
5	0.004	0.001	0.001							
6	0.002	0.001								
7	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
8	0.002									
10	0.001	0.001	0.001							
13	0.001	0.001								

information received that can be used to recover the lost parts. For instance, with an interleaving factor of two, a bursty loss of length one and a bursty loss of length two with samples belonging to different interleaving pairs can be recovered approximately. With an interleaving factor of four, a bursty loss of length less than or equal to three and a burst length of four, five, and six with lost packets belonging to different interleaving sets can be recovered. In general, with an interleaving factor of i, it is possible to recover a bursty loss of length less than or equal to $i-1$ and some of the bursty losses of length in the range $[i,\ (2i-2)]$.

Let the total number of packets sent be n_p and the interleaving factor be i. Over all the interleaving sets, assuming that consecutive losses of length j, $j \leq i$[1], happen m_j^i times, the total number of packets lost is n_s (independent of i):

$$n_s = \sum_{j=1}^{i} j \times m_j^i. \tag{3.1}$$

We can derive $Pr(fail \mid loss, i)$, the conditional probability that a packet cannot be recovered for interleaving factor i. This happens when all the packets in an interleaving set are lost. From (3.1),

$$Pr(fail \mid loss, i) = \frac{i \times m_i^i}{n_s}. \tag{3.2}$$

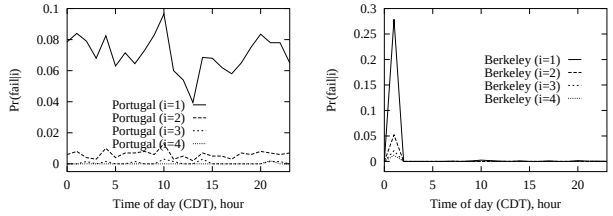
$Pr(fail \mid i)$, the unconditional probability that a packet cannot be recovered for interleaving factor i, can be computed as follows:

$$Pr(fail \mid i) = Pr(fail \mid loss, i) \times Pr(loss) = Pr(fail \mid loss, i) \times \frac{n_s}{n_p} = \frac{i \times m_i^i}{n_p}. \tag{3.3}$$

[1]Because we count consecutive losses in each interleaving set, j cannot be larger than i.

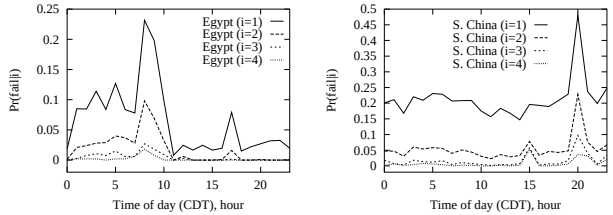
Figure 3.4 plots $Pr(fail \mid i)$ for various interleaving factor i and connections. $Pr(fail \mid i)$ drops quickly when i increases. For all times and all six connections, $Pr(fail \mid i)$ is negligible when $i \geq 4$. Moreover, except for the UIUC-Egypt and the two UIUC-China connections, an interleaving factor of two works well in general, achieving $Pr(fail \mid i)$ well below 5%. For the UIUC-Egypt and the two UIUC-China connections (see Figures 3.4c, 3.4d, and 3.4e), an interleaving factor of two is not always enough because about 10%-20% of the total losses will not be recoverable.

The above experimental results suggest that a small interleaving factor (between two to four) is adequate. In most cases, an interleaving factor of two leads to good recovery. Moreover, these interleaving factors do not increase end-to-end delays significantly. Assuming an interleaving factor of i and that each packet covers a sampling period of T ms, the additional end-to-end delay introduced by interleaving is $(i-1) \times T$ at both ends without considering the effect of jitters. With jitter buffers included at the receiver, there might be no or very little additional delay at the receiver site. The additional delay at the sender site will be $(i-1)T$ that is tolerable for small i. For example, if T is 30 ms and jitter buffer size is 200 ms, the additional end-to-end delay caused by interleaving alone is 30 (*resp.* 90) ms for interleaving factor of two (*resp.* four). Including the delay caused by jitter buffer, the worst-case additional end-to-end delay is $30 + 200 = 230$ (*resp.* $90 + 200 = 290$) ms for interleaving factor of two (*resp.* four). Figure 3.5 plots the average network end-to-end delays for the six-connections, assuming that one-way delay to be half of round-trip delay. As discussed before, interleaving factor of two can be used for UIUC-Portugal, UIUC-Berkeley, and UIUC-Slovakia connections.



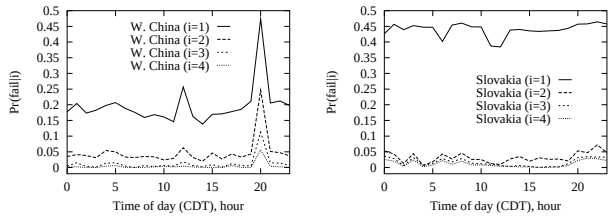
a) UIUC-Portugal

b) UIUC-Berkeley



c) UIUC-Egypt

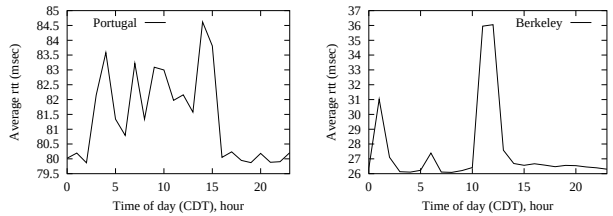
d) UIUC-S.China



e) UIUC-W.China

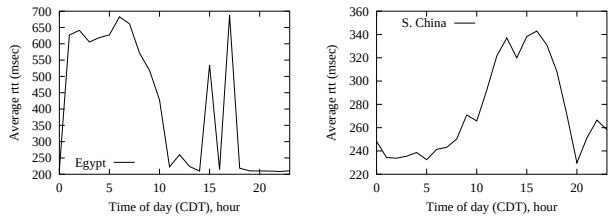
f) UIUC-Slovakia

Figure 3.4: $P(fail|i)$, probabilities of bursty losses that cannot be recovered under interleaving factor i, in round-trip paths between UIUC and six remote locations.



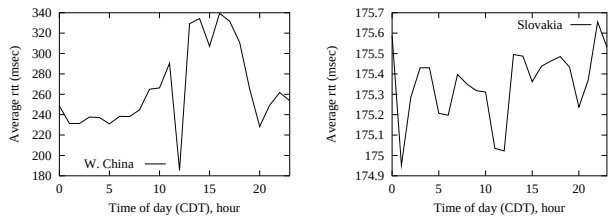
a) UIUC-Portugal

b) UIUC-Berkeley



c) UIUC-Egypt

d) UIUC-S.China



e) UIUC-W.China

f) UIUC-Slovakia

Figure 3.5: Average network end-to-end delays between UIUC and six remote locations.

For these three connections, the 400 ms delay requirement can be roughly met. For the other three connections, their network end-to-end delays are already long and the 400 ms delay requirement cannot be met even with only jitter buffer and without interleaving. Interleaving itself does not add much to the end-to-end delay.

The above experiments were performed on wired networks. For a wireless IP network, a mobile user typically connects over a WWAN to a dedicated proxy in the backbone, which then acts as the service point for all requests from the mobile user. Also, a WWAN connection typically experiences $1\%-10\%$ packet loss unrelated to congestion [68]. Therefore, we conclude that the end-to-end loss behavior of WWAN is similar to that of wired IP networks.

3.3 Summary

In this chapter, we have focused on studying the traffic patterns for voice transmissions over the Internet. We have picked connections of different loss rates, ranging from low loss with less than 10%, to high loss with up to 50%. Interleaving is a very attractive scheme to help overcome loss because network losses happen frequently and the length of consecutive packet losses is relatively small. From our traffic study, we have found that the interleaving factor does not have to be large: two is enough for low- or medium-loss connections, and four is enough for high-loss connections. Although interleaving will increase end-to-end delays because more data is buffered at both ends, it can overcome loss by utilizing the inherent redundancy in voice signals, without overloading the network.

By keeping both the packet size and interleaving factor to be small, end-to-end delays and losses can be controlled in a reasonable manner. The application end-to-end delay depends heavily on network end-to-end delay and the 400 ms delay requirement may be exceeded for some international connections with long delays.

CHAPTER 4

LOSS CONCEALMENTS FOR LOW BIT-RATE CODERS

Modern low bit-rate speech coders are developed based on the concept of linear prediction. Loss concealments for these coders are important because all recent standardized speech coders for multimedia communications belong to this class [69]. Yet loss-concealments for these coders are difficult to achieve because these coders achieve their high compression by embedding a great deal of dependencies within a coded sequence, resulting in propagated errors when loss happens.

In this chapter, our major focus is to design multiple-description coding to enable better loss concealments for linear-predictive coders without extra bandwidth requirements. First, we review briefly the basic concepts of low bit-rate speech coding and some background on the test coders used in this thesis. Second, we propose to design MDC systematically by evaluating the correlations of coding parameters. Using voice-sample correlations as a baseline, we start with the linear predictor, the fundamental parameter for all coders, followed by excitation parameters that are specific to each coder. Third, we apply sample-based MDC to the test coders and explain the results. Fourth, we present

a detailed design of our new parameter-based MDC scheme guided by correlation results. Last, we test our new scheme extensively under both synthetic loss scenarios and real Internet environments. Additionally, we compare the reconstruction qualities of different linear-predictor representations from several perspectives.

4.1 Overview of Low Bit-Rate Speech Coding

4.1.1 Foundation of low bit-rate speech coding

Waveform coding attempts to reconstruct accurate representations of a time-domain waveform. Low bit-rate coding, on the other hand, only aims to reproduce perceptually similar signals and, hence, can achieve a higher compression rate. Generally, low bit-rate coding models speech signals by a set of parameters and compresses these parameters efficiently. The process usually operates on a segment of speech signals, called a *frame*, because speech signals are considered as quasi-stationary over a short period of time, but are non-stationary over a long duration. From this point of view, low bit-rate coders are sometimes called *frame-based coders*.

The success of modern low bit-rate coding is based on the key concept that speech is produced by the human vocal system. Because sound waves are created by vibration and are propagated in air or other media by vibrations of the particles of the media [70], we examine how the vocal system produces a phoneme by following air propagation.

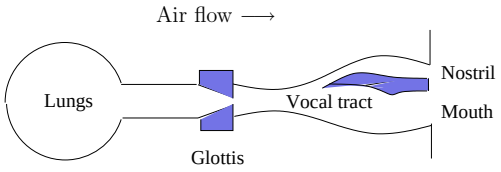


Figure 4.1: An illustration of the human vocal system.

Figure 4.1 shows how air flow starts from lungs that expel air. When air arrives at the glottis, the entrance of the vocal tract, it is adjusted to form the phoneme source, or excitation, and then continues to travel through the vocal tract.

At the glottis, if the vocal cord vibrates in a relaxed oscillation mode when air flow passes, then a voiced-sound excitation is generated. The vibration makes the vocal cord open and close in a periodic fashion, and results in a sequence of quasi-periodic flows of air. The vibration period is defined as the *pitch period*, and its corresponding frequency, the *fundamental frequency*. The shape and the tension of the vocal cord control the vibration frequency. This periodicity is an important characteristic of voiced sounds, such as vowels.

If air flow goes through the glottis at high velocity, is restricted by a constriction on its way, and in turn produces a turbulence, then an unvoiced-sound excitation is generated. The turbulence is observed to have a broad spectrum and is, thus, similar to noise. Therefore, conventionally, random signals, or white noise, are used to model unvoiced-sound excitations. The location of the constriction determines what unvoiced

sound is produced. Compared to voiced sounds, the energies of unvoiced sounds are generally lower, and unvoiced sounds are less important to speech quality.

After the formation of excitations, the wave continues to travel though the remaining path of the vocal tract. The vocal tract is shaped in such a way that produces the desired phoneme at its end point — the lips. The vocal tract looks like a tube and has resonance frequencies that are determined by its shape. These resonance frequencies are referred to as *formant frequencies*, or simply *formants* [70]. They have the function of frequency selectivity, and produce peaks at the formant frequencies of the spectrum. Different phonemes have different shapes of the vocal tract.

The excitations and the vocal tract are the primary determinants of speech sounds. They are the foundation of low bit-rate speech coding, which makes a further simplification that excitation generation and vocal-tract shaping are decoupled. In the following, we examine the basic concepts of low bit-rate coding. Generally, the encoder estimates the excitations and the shape (or parameters) of the vocal tract from an input speech, before compressing the parameters. After receiving the parameters, the receiver regenerates speech by simulating the vocal production system. Because speech is quasi-stationary, the above parameters are always generated frame-by-frame.

4.1.2 Linear predictive coding

The encoder first extracts vocal-tract parameters from an input speech stream. In most current low/very-low bit-rate speech-coding algorithms, a physical vocal tract is usually modeled as concatenated lossless tubes with different cross-sectional areas (see

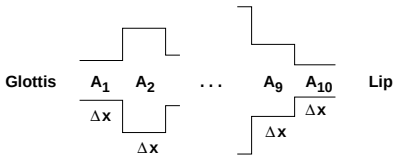


Figure 4.2: Acoustic-tube model of a vocal tract.

Figure 4.2). The descriptions in the following are excerpted from the book of Rabiner and Schafer [70]. In this figure, subscript k, $k = 1, \ldots, 10$, refers to the tube number with ordering from glottis to lips. Every tube has the same length Δx and constant cross-sectional area A_k. The number of tubes is usually 10.

To characterize the shaping effect of a vocal tract, acoustic theory uses the laws of conservation of mass, momentum, and energy to model air movements in the tubes. Basically, air velocity $u(x,t)$ and pressure $p(x,t)$ at position x and time t inside a lossless tube satisfy [70]:

$$\begin{aligned} -\frac{\partial p}{\partial x} &= \frac{\rho}{A}\frac{\partial u}{\partial t}, \\ -\frac{\partial u}{\partial x} &= \frac{A}{\rho c^2}\frac{\partial p}{\partial t}, \end{aligned} \tag{4.1}$$

where ρ is the air density in the tube, c is the sound-wave velocity, and A is the value of the tube's cross-sectional area. The solution of the above differential equations is:

$$u(x,t) = u^+(t - x/c) - u^-(t + x/c), \tag{4.2}$$

$$p(x,t) = \frac{\rho c}{A}[u^+(t - x/c) + u^+(t + x/c)], \tag{4.3}$$

where u^+ and u^- can be interpreted as traveling waves in the forward and backward directions, respectively. The "forward" direction refers to the direction from glottis to

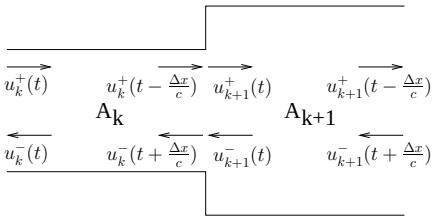


Figure 4.3: Wave flows in k^{th} and $(k+1)^{st}$ tubes.

lips, and the "backward" is the reverse direction. We can see that at any position x and instant t, the velocity or pressure consists of two components: one coming from the forward wave at instant $t - \frac{x}{c}$ and the other coming from the backward wave at instant $t + \frac{x}{c}$. As an example, the wave flows in the k^{th} and $(k+1)^{st}$ tubes are shown in Figure 4.3.

By forcing the velocity and pressure to be continuous at tube boundaries, we have:

$$u_k(\Delta x, t) = u_{k+1}(0,t), \tag{4.4}$$

$$p_k(\Delta x, t) = p_{k+1}(0,t), \tag{4.5}$$

for every k. After converting (4.2)-(4.5) to the frequency domain and combining the results of every tube, we get the transfer function of the vocal tract. It has the form of an infinite-impulse response (IIR) filter:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - a_1 z^{-1} - \cdots - a_{10} z^{-10}}. \tag{4.6}$$

An example of this filter is shown in Figure 4.4. It looks like a low-pass filter, where "bumps" in the graph represent formants.

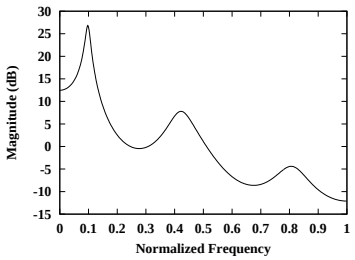


Figure 4.4: Example of vocal-tract frequency response.

Let $E(z)$ be the frequency response of the input to the vocal tract and $S(z)$ be the frequency response of speech at the lips. Then the following relation holds:

$$S(z) = E(z)H(z). \tag{4.7}$$

Here, $H(z)$ is also referred as the *synthesis filter*, and the corresponding inverse $A(z)$ is called the *analysis filter*. Rewriting this equation into the time domain, we have:

$$e(n) = s(n) - \sum_{k=1}^{10} a_k s(n-k) = s(n) - s'(n), \tag{4.8}$$

where $e(n) = Z^{-1}\{E(z)\}$ denotes the excitations, or residues, to the vocal tract, $s(n) = Z^{-1}\{S(z)\}$ denotes the speech, and $s'(n)$ is the predicted speech sample using past speech samples. It shows that excitations are the errors between $s(n)$ and $s'(n)$, which are predicted from past history $s(n-10), \cdots, s(n-1)$. Accordingly, the IIR filter in (4.6) is a linear-prediction filter, and the corresponding low bit-rate coders are often called *linear-predictive coders*. It follows that the estimation of the vocal tract is reduced to the computation of filter coefficients a_k. The most commonly used approach to estimate

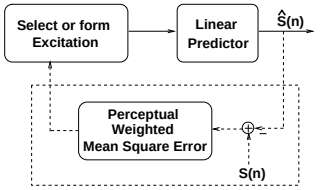


Figure 4.5: A typical linear predictive encoder.

these coefficients is to minimize the sum of $e^2(n)$ with respect to a_k for known $s(n)$. The result $\vec{a} = [1, a_1, \cdots, a_{10}]^T$ is:

$$\vec{a} = \begin{pmatrix} r(0) & r(1) & \cdots & r(10) \\ r(1) & r(0) & \cdots & r(9) \\ & & \vdots & \\ r(10) & r(9) & \cdots & r(0) \end{pmatrix}^{-1} \times \begin{pmatrix} r(0) \\ r(1) \\ \vdots \\ r(10) \end{pmatrix}, \tag{4.9}$$

where $r(k) = E[s(n)s(n-k)]$ for $k = 0, 1, \cdots, 10$.

Figure 4.5 shows the basic blocks of a linear-predictive encoder. The encoder sends the parameters of the top two blocks to the decoder that synthesizes the speech. In addition to estimations of the vocal tract, the encoder generates excitation parameters. This process is very specific to each coder. Roughly speaking, periods and amplitudes are two important attributes, because voiced sounds can be modeled by exciting the synthesis filter with quasi-periodic signals that simulate the oscillation of the vocal cords. Spectrum energy is also an important attribute, since unvoiced sounds result from noise-like excitations.

Depending on how excitations are formed, there are two major categories of linear-predictive coders. If the coder does not have the feedback block bounded by the dashed

rectangle in Figure 4.5, it is open-looped and is frequently called a *vocoder*. On the contrary, if the feedback block is present and excitations are chosen by minimizing the perceptual weighted mean-square error between synthesized speech $\hat{s}(n)$ and original speech $s(n)$, then the coder is close-looped and is generally called a *hybrid coder* or a *linear-prediction analysis-by-synthesis coder* (LPAS) [71]. Vocoders operate at a bit rate of around 2.4 kbits/s or lower. They simply extract important excitation parameters, such as pitch periods and signal energy. In contrast, hybrid coders not only extract these basic parameters, but also make an effort to match the waveform with the help of the feedback block. Therefore, hybrid coders can deliver better speech quality than vocoders at the cost of higher bit rates.

These low bit-rate coders are very attractive for encoding speech for transmissions over the Internet because they provide good-quality speech at much lower bit rates than conventional PCM, leading to significantly reduced network resource requirements in a large-scale deployment. Their high complexity is not of great concern because speech encoding and decoding can now be performed inexpensively in real-time using either software implementation in fast workstations or in dedicated embedded signal processing systems.

4.1.3 Test coders

To thoroughly investigate our loss-concealment schemes for low bit-rate coders, we have tested several representative linear-predictive coders. The linear-prediction (LP) coefficients in these coders are represented as LSPs. Table 4.1 lists the bit rate, quanti-

Table 4.1: Comparison of bit rates and major techniques in four LP coders (AC: adaptive code, VQ: vector quantization).

Standard	bps	LSP Quant.	Excitation
FS 1016 CELP	4800	scalar	stochastic code with AC
ITU G.723.1 (I)	5300	predictive-split VQ	algebraic code with AC
ITU G.723.1 (II)	6300	predictive-split VQ	multi-pulse with AC
FS MELP	2400	multi-stage VQ	mixed pulse-like and noise-like with pitch

zation method of LSPs, and excitation codebook structure for each coder, respectively. The top three are LPAS coders, and the last, a vocoder.

FS CELP uses a codeword to specify a segment of the time-domain excitation signals, and quantizes the codeword using a fixed stochastic codebook. It then produces pitch information by an adaptive codebook search and vector quantizes the result. The linear-prediction coefficient (LPC) vector is represented as a ten-element LSP vector, each of which is then scalar quantized. FS CELP operates at 4800 bps.

ITU standard G.723.1 has two operating modes, 5300 and 6300 bps. In both modes, the coder divides the ten elements of the LSP prediction-error vector into three subgroups, and vector quantizes each group separately. This method is indicated as the predictive-split method in Table 4.1. Similar to CELP, pitch information is coded using an adaptive codebook. Each fixed codebook vector in the 5300 bps Algebraic-Code-Excited Line Prediction (ACELP) mode contains at most four non-zero pulses. The possible positions of each pulse are restricted by the algebraic codebook. Each fixed codebook vector in the 6300 bps Multi-Pulse Maximum-Likelihood-Quantization (MP-MLQ) mode contains more pulses, and each pulse has more freedom in the choices of its position. Therefore, the 6300 bps mode can achieve better quality by finer excitation specification.

For the Federal Standard 2400-bps Mixed Excitation Linear-Prediction (MELP) vocoder, a 10-element LSP vector is multi-stage vector quantized, and pitch value is logarithmic quantized. To improve speech quality over the classic two-state, voiced/unvoiced LP vocoder, MELP includes both harmonic and noise-like components simultaneously in the modeling and regeneration of excitation signals. The relative contributions of harmonic and noise-like components are based on voicing strength in separate bands across the frequency spectrum.

4.2 Correlation Analysis of Interleaving Candidates

In this section, we design multi-description coding methods guided by correlation analysis. It is well known that sample-based MDC two-way performs fairly well on uncompressed or waveform-coded speech streams because sequential voice samples are highly correlated [66, 65]. This prompted us to determine interleaving candidates according to their correlations.

4.2.1 MDC design

Generally, when signals are compressed before transmission, multi-description generation can be performed either on voice samples before compression, as in a coder-independent approach, or on parameters generated by the coder, as in a coder-dependent approach. Specifically, for linear-predictive coders, two types of MDC are considered. In Type one or two-way sample-based MDC (Figure 4.6), a speech stream is interleaved

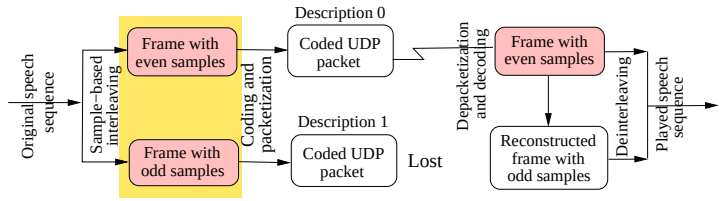


Figure 4.6: Sample-based MDC and reconstruction of a lost description at a receiver (shown with two descriptions).

into two streams, one containing even samples and the other containing odd ones, before coding each using an LP coder. After receiving the packets and decoding them, the receiver reconstructs missing samples by sample-based interpolation if not all packets in an interleaving sets are received.

Type two is the parameter-based MDC shown in Figure 4.7. The sender generates coding parameters and constructs each description by including a subset of the parameters. If the parameter sets of different descriptions are overlapping, the coding algorithm has to be modified in such a way that the total bandwidth taken by all descriptions is the same as the bandwidth taken by SDC. Figure 4.7 illustrates such a framework, although many variations can be derived from this framework, depending on how parameters are partitioned. Each coding parameter can belong to S_1, S_2, or both. When loss happens, the receiver will be able to reconstruct lost parameters before decoding if related parameters in another description are received.

For convenience of discussion, we call both a voice sample in sample-based MDC and a parameter in parameter-based MDC an *interleaving candidate*. The design of parameter-based MDC involves choosing which description an interleaving candidate be-

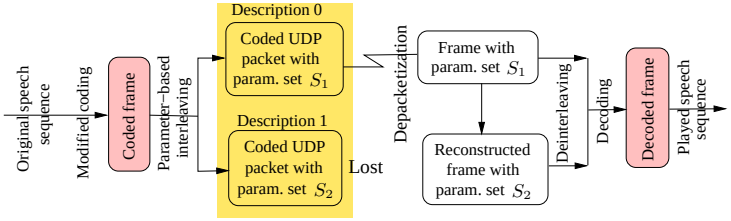


Figure 4.7: Parameter-based MDC and reconstruction of a lost description at a receiver (shown with two descriptions).

longs to. Here, correlation analysis can help make the decision. Intuitively, if a parameter is highly correlated with its neighboring ones, then they can be placed in different descriptions because the parameter can be reconstructed from its neighboring ones when the description containing the parameter is lost and that containing the neighboring ones are received. Therefore, it is straightforward for us to interleave highly correlated parameters, rather than low-correlated ones. For those low-correlated parameters, we can replicate them in all the descriptions, as they are difficult to reconstruct.

4.2.2 Correlations of voice samples and linear-predictor representations

Since sample-based MDC in the literature is well established for uncompressed or waveform-coded voice streams, correlations of voice samples can be used as a reference for picking interleaving candidates. Taking into account that normally a small number of descriptions (between two to four) is adequate, in all following correlation analysis, we only list correlation coefficients up to distance three. Table 4.2 lists the voice-sample

Table 4.2: Correlation coefficients of voice samples for the eight test streams in Table 1.1 (8000 Hz sampling frequency and 8061 frames of 240 samples each).

Sample Dist.	1	2	3
Correlation	0.83	0.60	0.35

correlations for the test streams in Table 1.1. It shows that voice samples have good correlations to their neighbors.

As said earlier, the parameters of linear-predictive coders generally include linear predictors and excitation parameters. A linear predictor can take several forms, such as Prediction Coefficient (PC), Reflection Coefficient (RF), Log Area Ratio (LAR), and Line Spectral Pair (LSP). The PC form is the direct IIR form described in the last section. On the other hand, the RF form is derived by starting from the same physical equations, followed by manipulating the signal-flow graph to a lattice-filter implementation. If we substitute (4.2)-(4.3) to the boundary conditions (4.4)-(4.5), we get [70]:

$$\begin{aligned} u^+_{k+1}(t) &= (1+r_k)u^+_k\left(t-\frac{\Delta x}{c}\right) + r_k u^-_{k+1}(t) \\ u^-_k\left(t+\frac{\Delta x}{c}\right) &= -r_k u^+_k\left(t-\frac{\Delta x}{c}\right) + (1-r_k)u^-_{k+1}(t), \end{aligned} \tag{4.10}$$

where

$$r_k = \frac{A_{k+1}-A_k}{A_{k+1}+A_k} \tag{4.11}$$

is called the *reflection coefficient* for the k^{th} junction. When air travels from the glottis to the lips, r_k signifies how much air is reflected and how much air continues to propagate at each junction. In this way, these reflections shape the air flow. Eq. (4.11) shows that r_k is completely determined by the adjacent area function ratio $\frac{A_k}{A_{k+1}}$. Figure 4.8 further shows

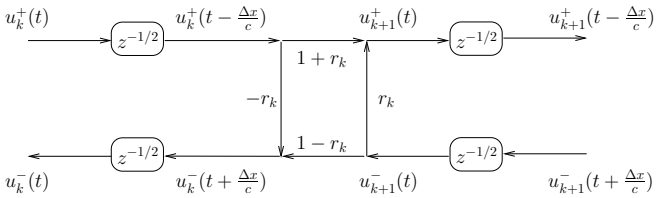


Figure 4.8: Signal-flow graph of the k^{th} and $(k+1)^{st}$ tubes.

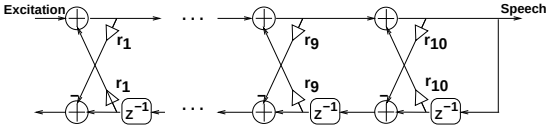


Figure 4.9: Lattice-filter model of the vocal tract.

the corresponding signal-flow graph of Figure 4.3, where $z^{-1/2}$ denotes $\dfrac{\Delta x}{c}$, the one-way delay of each tube. This flow graph looks like a digital-filter implementation that can be transformed into a lattice-filter structure (see Figure 4.9) after some manipulations, in which the reflection coefficients are just the lattice-filter coefficients. The direction from excitation to speech in the graph is the direction of forward wave propagation, whereas the bottom part of Figure 4.9 represents the direction of backward wave propagation [70].

Since both the IIR and lattice filters were derived from the acoustics equations, they are equivalent in the sense that a prediction coefficient vector $\vec{a}$ has a one-to-one corresponding RF vector $\vec{r}$. Figure 4.10 shows the procedures to convert $\vec{a}$ to $\vec{r}$ and vice versa [70].

// $\vec{a} \longrightarrow \vec{r}$
for $i = 1$ **to** 10
  $a_i^{(10)} = a_i$
end
for $i = 10$ **to** 2 **do**
  $r_i = -a_i^{(i)}$
  for $m = 1$ **to** $i-1$ **do**
    $a_m^{(i-1)} = \dfrac{a_m^{(i)} - r_i a_{i-m}^{(i)}}{1 - r_i^2}$
  end
end
$r_1 = -a_1^{(1)}$

// $\vec{r} \longrightarrow \vec{a}$
for $i = 1$ **to** 9 **do**
  $a_i^{(i)} = -r_i$
  for $m = 1$ **to** i **do**
    $a_m^{(i+1)} = a_m^{(i)} + r_{i+1} a_{i-m+1}^{(i)}$
  end
end
$a_{10}^{(10)} = -r_{10}$
for $i = 1$ **to** 10
  $a_i = a_i^{(10)}$
end

Figure 4.10: Conversion procedures for prediction coefficients and reflection coefficients.

An obvious deduction from the definition of r_k is $|r_k| \leq 1$. This is useful when we need to check whether a vocal-tract model is valid. It is much easier to check whether $|r_k| \leq 1$ is satisfied than to check whether an IIR filter is stable. This is the major advantage of RF over PC.

However, r_k is very sensitive to errors when its magnitude approaches 1. A little perturbation may cause the model to be unstable. Therefore, LAR, another representation, is often used. LAR is a non-uniform scaling of r_k with the following relation with RF [70]:

$$l_k = \log \frac{1 + r_k}{1 - r_k} = \log \frac{A_{k+1}}{A_k}, \qquad 1 \leq k \leq 10, \tag{4.12}$$

where l_k is the k^{th} LAR. We can directly infer from the above definition that a linear interpolation of LARs is equivalent to a geometric interpolation of the ratio of area

functions. Further, RF can be converted from LAR by:

$$r_k = \frac{e^{l_k} - 1}{e^{l_k} + 1}, \quad 1 \le k \le 10. \tag{4.13}$$

Here, the stability of the corresponding linear predictor is naturally satisfied because the magnitude of the derived r_k is always less or equal to one.

In recent coding standards, LSP is the preferred format for the linear-predictor representation. LSPs are the angles $x_1, x_2, \cdots, x_{10}$ of the zeros $z_1, z_2, \cdots, z_{10}$ on the upper unit circle of the following two functions [72]:

$$\begin{aligned} P(z) &= A(z) + z^{-11}A(z^{-1}) = (1 + z^{-1}) \prod_{k=1}^{5} (1 - 2\cos x_{2k-1} z^{-1} + z^{-2}) \\ Q(z) &= A(z) - z^{-11}A(z^{-1}) = (1 - z^{-1}) \prod_{k=1}^{5} (1 - 2\cos x_{2k} z^{-1} + z^{-2}), \end{aligned} \tag{4.14}$$

with $z_k = e^{jx_k}$. It is easy to convert the LSP polynomials back to the LP analysis filter $A(z)$ as follows:

$$A(z) = \frac{P(z) + Q(z)}{2}. \tag{4.15}$$

Frequencies $f_1, f_2, \cdots, f_{10}$ of LSP are called *line spectral frequencies* (LSF), where $f_k = \frac{x_k}{2\pi}$. The LSP indices in a frame are always monotonically increasing ($0 < x_1 < x_2 < \cdots < x_{10} < \pi$) for a stable linear predictor [72]. Similar to the case of RF, it is easy to determine whether an LSP vector corresponds to a valid vocal-tract model.

Figure 4.11 shows an example of an LSP vector, where $f_1, \cdots, f_{10}$ denote the normalized LSFs with respect to the Nyquist frequency, and the solid curve is the corresponding frequency response of the linear predictor.

Another observation of Figure 4.11 is that the LSFs cluster around the speech formants, an important property of LSPs. Goldberg and Riek [73] have found experimen-

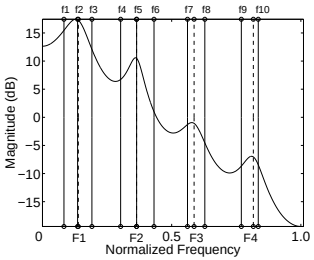


Figure 4.11: An example of an LSP vector. ($f1, \cdots, f10$ denote the normalized LSFs and $F1, \cdots, F4$ denote the four normalized formant frequencies.)

tally that the differences between adjacent LSP indices are closely related to the formant bandwidths of speech, which suffice to specify the entire spectral envelope for vowels. This suggests that a linear interpolation of LSP vectors, which is equivalent to the interpolation of the differences of adjacent LSP indices, is closely related to the generation of smoothly changing formant information.

The above four representations, namely, PC, RF, LAR, and LSP, are equivalent in the sense that there is a one-to-one mapping between any two. However, not all of them are suitable for use in the MDC scheme. The PC representation is not a good choice for interleaving, because the reconstruction of a lost PC vector by interpolating its neighboring received vectors is not guaranteed to be a stable linear predictor. Accordingly hereafter, we only compute the correlations of RF, LAR, and LSP.

Table 4.3 shows the inter-frame correlations of the three representations, RF, LAR and LSP. The tests were done on the combined eight streams of 8061 frames that include significant variations in speaker characteristics. All three representations show good inter-

Table 4.3: Correlation coefficients of inter-frame RF, LAR, and LSP for the eight test streams in Table 1.1 (8000 Hz sampling frequency, 30 msec frame period, 45 msec Hamming window, 10^{th} analysis order, and 8061 frames).

a) Correlations of RF

Frame Distance	RF									
	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
1	0.73	0.69	0.69	0.69	0.66	0.69	0.68	0.63	0.64	0.54
2	0.50	0.36	0.44	0.42	0.41	0.41	0.45	0.41	0.42	0.26
3	0.31	0.15	0.30	0.26	0.26	0.23	0.31	0.31	0.30	0.11

b) Correlations of LAR

Frame Distance	LAR									
	l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	l_{10}
1	0.80	0.71	0.69	0.70	0.66	0.68	0.69	0.63	0.64	0.54
2	0.63	0.38	0.43	0.43	0.41	0.41	0.45	0.41	0.42	0.26
3	0.53	0.16	0.29	0.27	0.26	0.23	0.32	0.31	0.30	0.11

c) Correlations of LSP

Frame Distance	LSP									
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	0.82	0.81	0.75	0.72	0.81	0.76	0.74	0.73	0.73	0.74
2	0.61	0.64	0.50	0.45	0.59	0.46	0.43	0.43	0.45	0.55
3	0.46	0.52	0.35	0.26	0.40	0.24	0.21	0.24	0.26	0.42

frame correlations that are comparable to those of sample correlations (Table 4.2). The results show that, for most indices and frame distances, RF has the worst correlations and LSP has the best, although the differences among the three representations are not significant. The correlations demonstrate that all three representations vary slowly from one frame to the next and that the linear predictor in a lost frame can be reconstructed from those received in neighboring frames by any of the representations. A natural

conclusion is that linear predictors are good candidates for two-way interleaving in our parameter-based MDC scheme.

4.2.3 Correlations of excitation parameters

Besides the linear predictor, each coder has its unique representation for excitations. In the following, we investigate the correlations of excitations for each coder.

First, we present correlation results of FS CELP excitation parameters. In the encoder, a linear predictor and four groups of adaptive and stochastic codewords are generated for each 240-sample speech frame (see Figure 4.12). In each 60-element subframe, excitations are coded by both adaptive codeword and stochastic codeword. Each adaptive-codeword represents a segment of scaled past excitations and further has two components: delay and gain, where delay specifies the position of the past excitation segment and gain specifies the scaling factor. All parameters are packed into 144 bits. The decoder reverses the encoder process by recovering the parameters from the 144 received bits and by synthesizing speech using the combined excitations to excite the linear predictor.

Table 4.4 shows the adaptive codeword's correlations. As expected, delays have reasonable correlations, whereas adaptive codeword gains show no correlation. Consequently, for FS CELP, adaptive codewords are not appropriate for interleaving.

In FS CELP, a stochastic codeword corresponds to a 60-element vector. For stochastic codewords, we compute a correlation coefficient for each of the sixty element e_i and

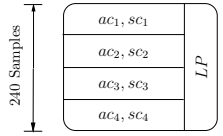


Figure 4.12: FS CELP SDC at sender (*ac*: adaptive codeword; *sc*: stochastic codeword; LP: linear-prediction vector).

Table 4.4: Correlation coefficients of FS CELP adaptive codewords for the eight test streams in Table 1.1 (*ac*: adaptive codeword).

Subframe Distance	Correlations of *ac* delay	Correlations of *ac* gain
1	0.57	0.004
2	0.22	0.007
3	0.21	0.006

subframe distance d as:

$$c_{i,d} = \frac{E[(e_i^n - m_{e_i})(e_i^{(n-d)} - m_{e_i})]}{\sigma_{e_i}^2}, \quad i = 0, \cdots, 59 \text{ and } d = 1, 2, 3 \tag{4.16}$$

where m denotes the mean, σ denotes the standard deviation, and n is the subframe index. Since the stochastic codebook is generated from random noise, we expect that there would be no correlations among stochastic codewords. Figure 4.13 plots the correlation coefficients for the stochastic codeword elements. The lack of correlations indicates that stochastic codewords should not be interleaved.

Second, consider the excitation parameters of the ITU G.723.1 coder. The frame organization of ITU G.723.1 is similar to that of FS CELP. For each 240-sample speech frame, a linear predictor and four groups of adaptive and fixed codewords are extracted. Both of its working modes, ACELP and MP-MLQ, share almost exactly the same module of adaptive codeword generation. Similar to FS CELP, an adaptive codeword consists

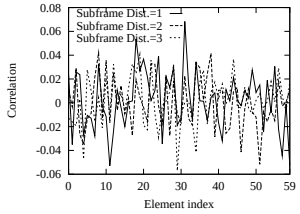


Figure 4.13: Correlation coefficients of FS CELP stochastic codewords for the eight test streams in Table 1.1.

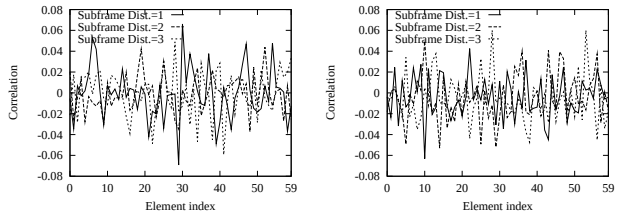
Table 4.5: Correlation coefficients of ITU G.723.1 adaptive codewords for the eight test streams in Table 1.1 (*ac*: adaptive codeword).

Subframe Distance	Correlations of *ac* delay	Correlations of *ac* gain				
1	0.64	0.03	0.07	0.26	0.10	0.04
2	0.27	0.04	0.06	0.23	0.08	0.04
3	0.25	0.01	0.03	0.22	0.04	0.02

of two parts: delay and a gain vector. Here, the replication of past excitations is more complex as discussed below. An excitation sample is no longer just a scaled past residue, but an interpolation of five past residues around the specified position. This leads to a gain vector of five elements.

Table 4.5 lists the correlations of the delay and the five gain elements. The delay shows fair correlations; the center gain element shows low correlations; whereas other gain elements have no correlations. Consequently, adaptive codewords in ITU G.723.1 are not suitable for interleaving.

In ITU G.723.1, fixed codewords also contribute to the excitations. A fixed codeword for both ACELP and MP-MLQ has 60 elements. Although ACELP and MP-MLQ



(a) ACELP (b) MP-MLQ

Figure 4.14: Correlation coefficients of the fixed codewords in ITU G.723.1 for the eight test streams in Table 1.1.

generate different fixed codewords, they both show little correlations (see Figures 4.14a and 4.14b), which are computed using (4.16) and, thus, should not be interleaved. Conceptually, in both FS CELP and ITU G.723.1, signal correlations inside a segment, or short-term correlations, are removed by passing the segment through a linear prediction analysis filter. The signal correlations to history data, or long-term correlations, are then removed by subtracting the excitations generated by adaptive codewords from the residues. The remaining residues, coded by the stochastic codewords, or fixed codewords, therefore, exhibit almost no correlations.

Third, we evaluate the correlations of excitations in FS MELP. As mentioned before, FS MELP is a vocoder and is very different from the other three coders. A MELP encoder generates a linear predictor vector and excitation information for each 180-sample frame. One of the excitation parameters is the pitch period. The MELP decoder synthesizes speech one pitch-period at a time. Specifically, the decoder regenerates excitations for each pitch period and passes them through a synthesis filter to generate the decoded

Table 4.6: Correlation coefficients of FS MELP pitch periods for the eight test streams in Table 1.1.

Frame Distance	1	2	3
Pitch Corr.	0.63	0.58	0.54

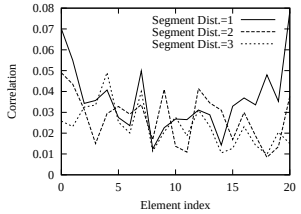


Figure 4.15: Correlation coefficients of FS MELP excitations for the eight test streams in Table 1.1.

speech. Table 4.6 shows the correlations of pitch periods. Pitches have good correlations to their neighbors, especially noticeable for frame distance three. As a result, they are possible candidates for interleaving.

Depending on the pitch period value, the number of excitations in each synthesis segment varies. The minimum number is twenty and excitations may not exist for some segments with element indices beyond twenty. Therefore, Figure 4.15 shows the correlations of the first twenty excitations of each segment. Apparently, these excitations have little correlation and should not be interleaved.

In conclusion, by evaluating parameter correlations and comparing them with sample correlations, we find that the linear predictor is the best candidate for interleaving and can be implemented using any coder. Further, excitation parameters are specific to each

coder and generally are not suitable for interleaving, except for the case of FS MELP pitches. Parameters that are not interleaved have to be replicated to all descriptions in order to enable recoverability. Further, the codec has to be modified in such a way that the total bandwidth of all the descriptions is the same as that of the original single-description case.

4.3 Sample-Based MDC

In this section, we code signals generated by conventional sample-based MDC using linear-predictive coders. Figure 4.6 shows the two-way sample-based MDC scheme in which a speech stream is interleaved into two streams, one containing the even samples and the other containing the odd ones, before coding each using an LP coder. Under the best conditions, both coded streams will be received, decoded separately, and de-interleaved to rebuild the original stream. Even in this case, we find that the playback quality is very poor, as shown in Figure 4.16 for the four coders in Table 4.1.

Here, we measure reconstruction qualities of low bit-rate coded speech by the Itakura-Saito likelihood ratio (LR) and the cepstral distance (CD) [74]. LR for the n^{th} speech frame is defined as:

$$LR_n = \frac{\vec{a}_{n,r}^T R_{n,o} \vec{a}_{n,r}}{\vec{a}_{n,o}^T R_{n,o} \vec{a}_{n,o}}, \tag{4.17}$$

where $\vec{a}_{n,o}$ and $\vec{a}_{n,r}$ are the vectors of linear-prediction coefficients of the n^{th} original and reconstructed speech frames, respectively, and $R_{n,o}$ is the correlation matrix derived from the n^{th} original speech frame. LR for a speech stream is the average LR_n over all frames.

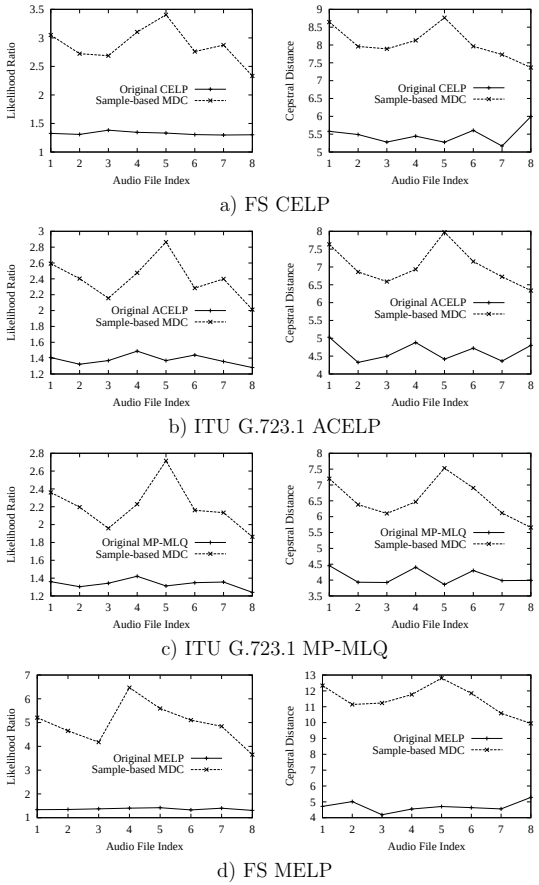


Figure 4.16: Quality comparison, in terms of LR and CD between SDC with no loss and two-way sample-based MDC with both streams received, for FS CELP, ITU G.723.1 ACELP, ITU G.723.1 MP-MLQ, and FS MELP.

We show in the next chapter that the numerator (*resp.* denominator) in (4.17) represents the prediction error when the linear predictor is $\vec{a}_{n,r}$ (*resp.* $\vec{a}_{o,r}$). If $\vec{a}_{n,r}$ is close to $\vec{a}_{o,r}$, then the ratio is close to one. Therefore, LR is good at measuring reconstruction qualities of linear predictors. CD for the n^{th} speech frame is defined as:

$$CD_n = 4.34[(c^n_{r,\,0} - c^n_{o,\,0})^2 + 2\sum_{i=1}^{\infty}(c^n_{r,\,i} - c^n_{o,\,i})^2]^{\frac{1}{2}} \text{ [dB]} \tag{4.18}$$

where $c^n_{o,\,i}$ and $c^n_{r,\,i}$ denote the cepstra of the n^{th} original and reconstructed speech frames, respectively. Here, cepstrum is defined as [74] the Inverse Discrete Fourier Transform (IDFT), with i in (4.18) denoting its index, of the log-scaled spectrum $\frac{G}{A(e^{jw})}$, where G is the square root of the excitation energy. CD measures log-scaled spectrum differences between the reconstructed and the original speech streams. The final CD for a stream is the averaged result for all the frames tested.

All tests were performed on the eight streams in Table 1.1. Figure 4.16 shows that both LR and CD of two-way sample-based MDC increase dramatically for the four coders under no loss. Table 4.7 summarizes the average degradations of sample-based MDC as compared to SDC:

$$\frac{1}{8}\sum_{i=1}^{8}\frac{Q_i^{MDC} - Q_i^{SDC}}{Q_i^{SDC}}, \tag{4.19}$$

where i is the test stream index and Q represents quality measured by LR or CD. The most significant degradation is for FS MELP, the very-low bit-rate coder. Subjective hearing tests of the decoded streams also indicate that sample-based MDC performs poorly. Obviously, sample-based MDC is not suitable for low bit-rate coding.

Table 4.7: Average degradations of sample-based MDC as compared to SDC on eight test streams.

Coder	LR	CD
FS CELP	116%	47%
ACELP	74%	52%
MP-MLQ	65%	60%
MELP	263%	145%

The quality degradations of sample-based MDC are due to two major factors: aliasing introduced when the original stream is down-sampled into even and odd streams, and the doubling of the time span of a coded frame in each interleaved stream. For example, if the frame size in the SDC coder is n samples, or t seconds assuming the sampling frequency is F_s Hz, the frame size of each interleaved stream has to be n samples as well in order to keep the total MDC's bandwidth equal to SDC's. However, since the sampling frequency of the interleaved stream is halved, the time span of each frame is now $n/\frac{F_s}{2} = 2n/F_s = 2t$ seconds. These two factors result in wrong speech spectra computed and, thus, a wrong speech-production model.

In summary, sample-based MDC is not a good choice for the loss concealments of low bit-rate coded speech. Further, low bit-rate speech coders cannot be tailored to work in a coder-independent approach.

4.4 Parameter-Based MDC

In this section, we study the coder-dependent MDC approach. Previous correlation results show that a linear predictor is closely related to its neighboring linear predictors.

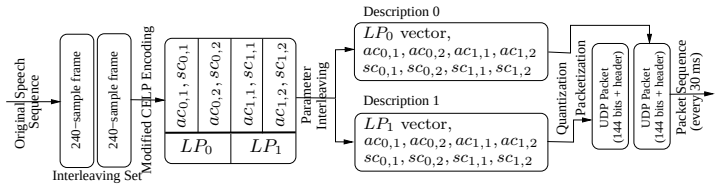


Figure 4.17: Parameter-based two-way MDC for FS CELP (*ac*: adaptive codeword; *sc*: stochastic codeword).

Therefore, if we interleave linear predictors, a lost linear predictor can be approximated using neighboring received ones. In contrast, excitation parameters have very low or no correlations and cannot be reconstructed from their neighbors. Hence, we should not interleave them, but rather replicate them to all descriptions.

4.4.1 Two-way and four-way MDC algorithms

In this section, we illustrate our proposed parameter-based MDC scheme, or LP-based MDC, on the FS CELP coder. Figure 4.17 shows the two-way MDC design. The sender groups each pair of 240-sample frames in the original speech sequence into an interleaving set, performs linear-prediction analysis, once for each frame, and distributes the two linear-predictor vectors to the two descriptions. However, instead of generating codewords for four 60-sample subframes, it extends the subframe size to 120 samples, generates codewords for four 120-sample subframes, and replicates all the codewords to both descriptions. With replicated codewords, we need to extend the subframe size in coding in order to keep the frame size after quantization in each description to be 144 bits, the same size as that of a coded frame in SDC. After coding and parameter interleaving,

Description i, $i = 0, 1$, contains the linear predictor of frame $2n + i$ and the excitations from all speech frames. Finally, the sender quantizes the parameters and encapsulates a frame in each description in a UDP packet and alternates between Descriptions 0 and 1 in sending packets to the destination. Note that we have maintained the same frame size of 240 in linear-prediction analysis and have overcome the aliasing problem, without down-sampling the original speech samples into odd-even ones.

At the receiver side, if all the frames in both descriptions are received, the receiver carries out the reverse process in Figure 4.17. It first deinterleaves the information received into a single coded stream by extracting the linear predictors from frames in both descriptions and the codewords from frames in either description, before decoding the coded stream. Obviously, the quality of the decoded stream is equivalent to a coder with a frame size of 240 and a subframe size of 120. As said already, since we have preserved the precision of linear-prediction analysis and have eliminated aliasing, the decoded stream can be guaranteed to have better quality than that of sample-based MDC. However, the decoded stream has worse quality than that of SDC because of its increased subframe size for excitation generation. Since we have to code twice as many excitations as SDC into the same amount of bandwidth as SDC, the excitation coding is less precise and, thus, decoding quality degrades. Results are shown in Section 4.5.2.

When some frames in one description are lost, the receiver only needs to reconstruct the lost linear predictors using the linear predictors in those frames received in the other description. It does not need to reconstruct the codewords because they are replicated in both descriptions.

For all representations of linear predictors, the reconstruction process is as follows. Assume $\vec{v}_n$ is the lost vector of linear predictors represented in RF, LAR, or LSP in the n^{th} frame of Description 1, and $\vec{v}_{n-1}$ and $\vec{v}_{n+1}$ are the immediately preceding and following vectors received in Description 0. We can approximate the missing vector $\vec{v}_n$ by $\vec{v'}_n$ computed as follows:

$$\vec{v'}_n = \frac{\vec{v}_{n-1} + \vec{v}_{n+1}}{2}. \tag{4.20}$$

For LAR, the interpolated vector always renders a stable linear predictor. For RF and LSP, such linear interpolations also result in stable linear predictors as discussed in the last section. Moreover, since the receiver reconstructs the coding parameters of lost frames before decoding, there is no need to estimate the decoding states of lost frames as in SDC.

When the loss of a burst of packets leads to the loss of all the frames corresponding to those of an interleaving set, the receiver will not be able to reconstruct the lost parameters and can generate silence in the decoded speech sequence, similar to the recovery process in SDC. Of course, more sophisticated schemes, like padding by scaled-down parameters in previously received interleaving set, can be employed to enhance quality. No matter what padding method is used, the chance for such bursty losses is much smaller than in SDC because all losses of burst length one and some losses of burst lengths two can be recovered in two-way MDC.

The above idea can be extended to four-way MDC. Figure 4.18 shows that a sender interleaves the linear predictors of four 240-sample frames in an interleaving set into

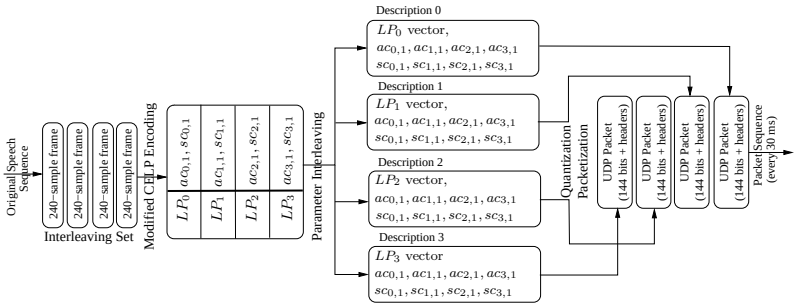


Figure 4.18: Parameter-based four-way MDC for FS CELP (*ac*: adaptive codeword; *sc*: stochastic codeword).

four descriptions. Here, the frame size in linear-prediction analysis is still 240, but the subframe size has been increased to 240 in order to maintain the same 144-bit coded frame size after the codewords are replicated. After coding and parameter interleaving, Description i, $i = 0, \cdots, 3$, contains the linear predictor of frame $4n+i$ and the codewords from all the coded frames. Note that the quality of four-way MDC is expected to be worse than that of two-way MDC due to its even longer subframe size, which means we have to code four times as many excitations as SDC into the same amount of bandwidth as SDC. Although the precision of linear prediction is managed to be the same as that of SDC, the final decoding quality degrades because excitations are coded in a coarser level. Results are shown in Section 4.5.2.

In four-way MDC, there are five loss patterns of frames in an interleaving set in which losses can be concealed at a receiver: one out of four frames received (0, 1, 2, or 3), two consecutive frames received (0-1, 1-2, 2-3, or 3-0), two disjoint frames received (0-2, or

1-3), three frames received (0-1-2, 1-2-3, 2-3-0, or 3-0-1), and all four frames received. If four frames are received, then the receiver deinterleaves the parameters before decoding. In all other cases, the receiver can recover the lost linear-predictor vectors by interpolating the vectors received in their immediately preceding and following frames. For example, in the case when two consecutive descriptions are received, assuming 0 and 1, the lost vectors v_{4n+2} and v_{4n+3} are approximated as follows:

$$\begin{aligned} \vec{v'}_{4n+2} &= \frac{2\vec{v}_{4n+1} + \vec{v}_{4n+4}}{3} \\ \vec{v'}_{4n+3} &= \frac{\vec{v}_{4n+1} + 2\vec{v}_{4n+4}}{3}. \end{aligned} \tag{4.21}$$

Similar to two-way MDC, interpolations in RF, LAR, and LSP domains result in stable linear predictors.

We do not study MDC beyond four ways because we have shown in Chapter 3 that four-way interleaving will be enough to conceal errors in most, if not all, loss scenarios. Further, an interleaving degree larger than four will result in even larger subframe sizes that will degrade quality further at a receiver, even when there are no losses.

For the other three coders in Table 4.1, we apply the same idea to construct two-way and four-way MDC, although there are some coder-specific variations.

4.4.2 Evaluation criteria

The LP-based MDC studied has three variations depending on the form of linear predictors. In this section, we compare the reconstruction quality of the three linear-predictor representations, RF, LAR, and LSP. Besides comparing the quality of the final

reconstructed streams for the three representations by LR and CD defined earlier, we evaluate the performance of our proposed MDC scheme using the following two measures.

First, we use spectral distortion (SD) [73] to compare the reconstruction quality of the three linear-predictor representations. SD is defined as follows:

$$SD = E\Big[\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big|10\log_{10}|H_n(\omega)|^2 - 10\log_{10}|H'_n(\omega)|^2\Big|d\omega\Big] \text{ dB}, \tag{4.22}$$

where $H_n(w)$ is the original linear predictor for the n^{th} frame, and $H'_n(w)$ is the reconstructed linear predictor for the n^{th} frame. SD has been used [73, 75] to compute spectral differences and we can use it directly to measure the differences between the reconstructed and original LP vectors in the frequency domain. In an actual computation, we use a 256-point Discrete Fourier Transform (DFT) to approximate the integral. For each representation, we convert the corresponding vectors to the direct form in order to obtain SD.

Another measure we used for comparing the three representations is the correlation between the target vector and the reconstructed vector in the corresponding representation domain. Assuming that the original parameter vector is $\vec{v} = [v_1, v_2, \cdots, v_{10}]^T$ and that the reconstructed parameter vector is $\vec{v'} = [v'_1, v'_2, \cdots, v'_{10}]^T$, the correlation can be computed for each element of the vector as follows:

$$c_i = \frac{E[(v_i - m_{v_i})(v'_i - m_{v'_i})]}{\sigma_{v_i}\sigma_{v'_i}}, \quad i = 1, \cdots, 10, \tag{4.23}$$

where m represents the mean and σ, the standard deviation. Correlation is commonly used to measure the similarity between two vectors [76, 77], since it is directly related to the cosine value of the angle between these two vectors.

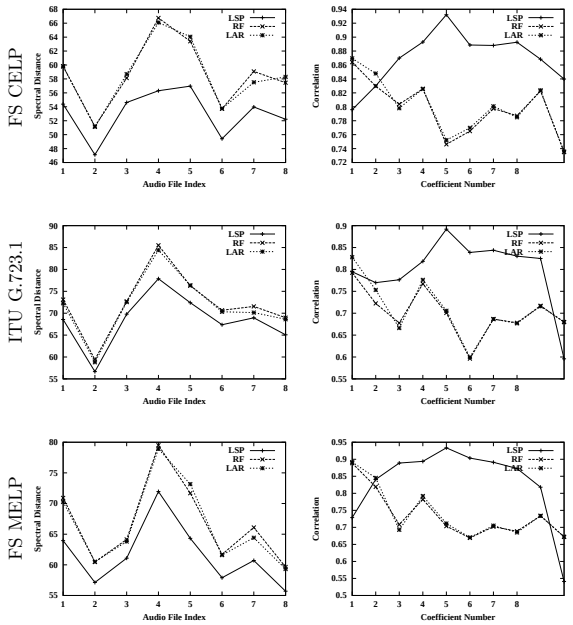
4.5 Synthetic Tests

In this section, we report our results on testing our proposed two-way and four-way MDC algorithms using the three linear-predictor representations under controlled losses or when all descriptions are received. The tests were performed on the four coders in Table 4.1 and the eight test streams in Table 1.1.

4.5.1 Reconstruction quality under controlled losses

We first compare the three representations by the reconstruction quality in terms of SD, and correlation under synthetic loss scenarios. Because the three representations are equivalent, the decoding qualities are the same if no loss happens. Therefore, in the following, we only compare them in case of partial loss.

Figure 4.19 plots the results in terms of SD and correlations for two-way MDC with only one description received. Because the two coding modes, ACELP and MP-MLQ, in ITU G.723.1 share the same linear-prediction module, we do not need to distinguish them in our comparisons here. In each graph, the solid (*resp.* dashed and dotted) line indicates the performance if reconstruction is performed in the LSP (*resp.* RF and LAR) domain. The results show that the reconstruction of linear predictors in the LSP domain gives the lowest spectral distortion for all files and all coders, whereas RF and LAR behave similarly. Likewise, for most of the indices, the reconstructed LSP vectors correlate the best to the target vectors. Moreover, the difference in correlations between LSP and both RF and LAR are significant.

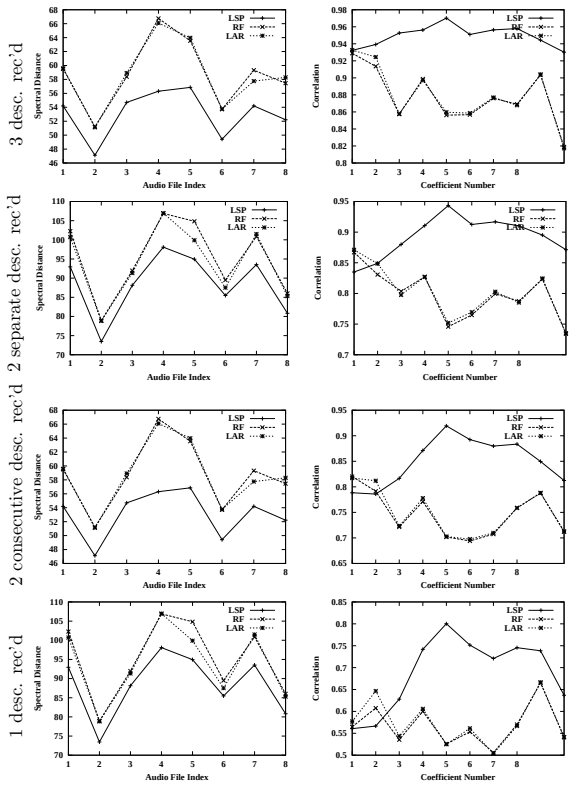


a) Spectral Distortion b) Correlation

Figure 4.19: Reconstruction qualities of RF, LAR, and LSP representations in FS CELP, ITU G.723.1 ACELP, and FS MELP for two-way MDC when only one description received.

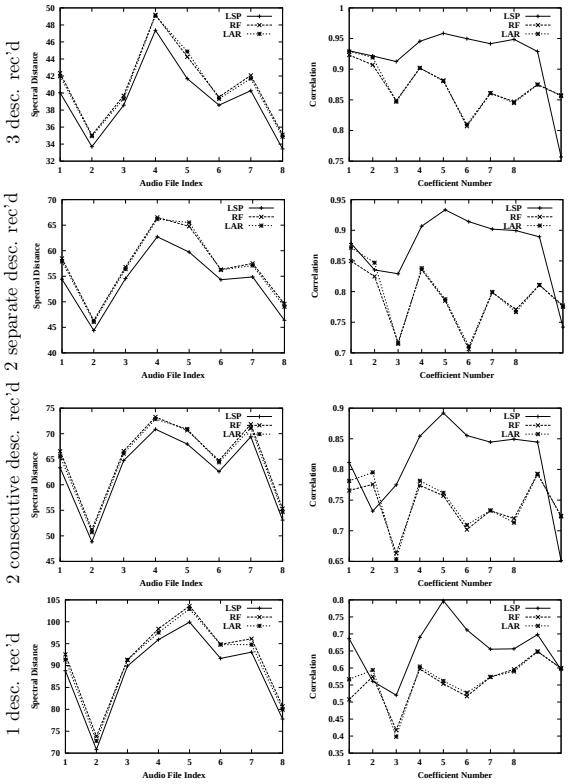
Figures 4.20-4.22 present the reconstruction qualities in SD and correlations for four-way MDC under the four loss scenarios for each coder, respectively. In terms of SD, LSP constantly reconstructs the best, although its advantage over RF and LAR is not very significant. In terms of correlation, we observe that LSP has higher correlations than those of RF or LAR for all element indices except for the first and the last. Since the conversions among these representations can only be performed on 10-element vectors as a whole and not on each individual element, we conclude that reconstructions using LSP give the highest overall correlations to the target vectors. Moreover, the advantage of reconstructions using LSP over RF or LAR is very obvious, since the correlations can be higher by 0.2 to 0.3. Even when only one description is received, the correlations between LSP reconstructed vectors and the target vectors are relatively high.

Further, for each representation, we compare the distortions of the reconstructed streams to the original streams. Figures 4.23-4.27 plot the reconstruction qualities measured in terms of LR and CD for the two-way MDC and four-way MDC, respectively. Here, we need to distinguish the two working modes of ITU G.723.1, because ACELP and MP-MLQ produce different reconstructed streams. Similar to the spectral distortion and correlation results, LSP performs the best among the three representations in terms of LR with noticeable differences. For example, for FS CELP, reconstructions using RF (*resp.* LAR) is 12% (*resp.* 9%) worse than LSP on the average for two-way MDC. For four-way MDC with only one description received, reconstructions using RF (*resp.* LAR) is 73% (*resp.* 31%) worse than LSP. The improvements of LSP in terms of CD is not as



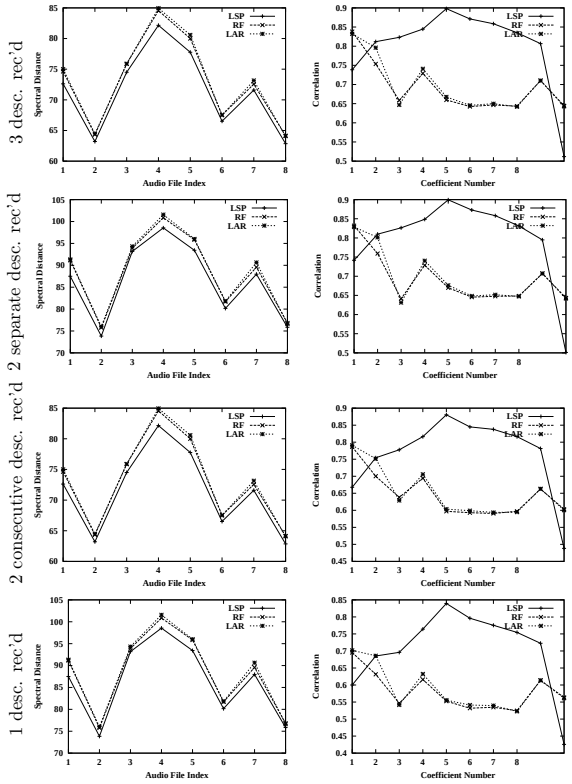
a) Spectral Distortion b) Correlation

Figure 4.20: Reconstruction qualities of LSP, RF, and LAR representations for four-way MDC in FS CELP under four different loss patterns: three descriptions received, two separate descriptions received, two consecutive descriptions received, and one description received.



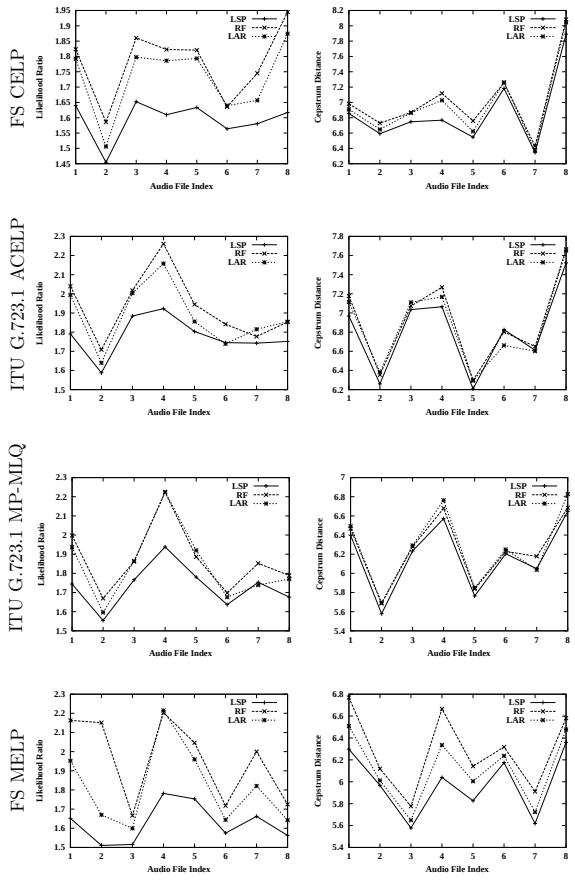
a) Spectral Distortion b) Correlation

Figure 4.21: Reconstruction qualities of LSP, RF, and LAR representations for four-way MDC in ITU G.723.1 under four different loss patterns plotted from top to bottom: three descriptions received, two separate descriptions received, two consecutive descriptions received, and one description received.



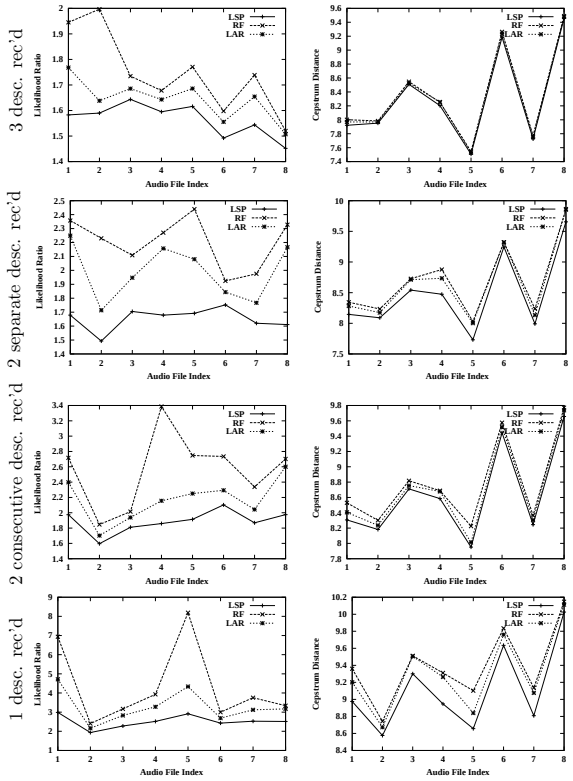
a) Spectral Distortion b) Correlation

Figure 4.22: Reconstruction qualities of LSP, RF, and LAR representations for four-way MDC in FS MELP under four different loss patterns plotted from top to bottom: three descriptions received, two separate descriptions received, two consecutive descriptions received, and one description received.



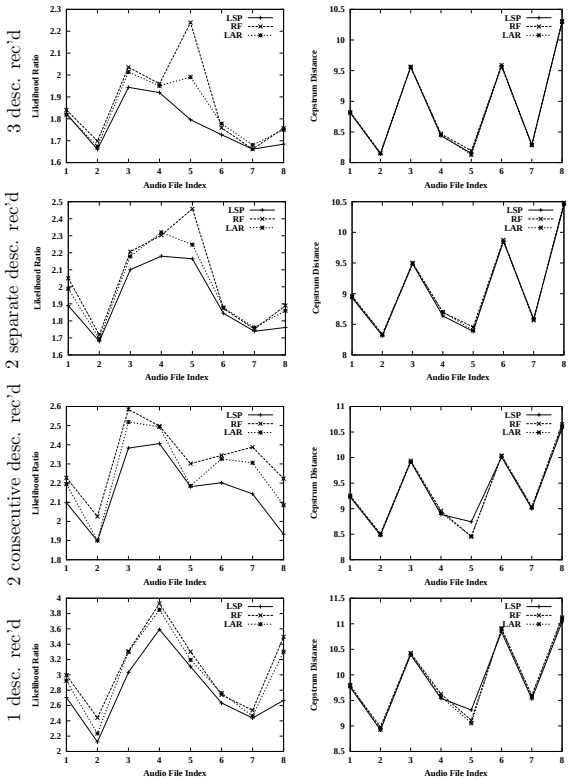
a) LR b) CD

Figure 4.23: Quality comparisons, in terms of LR and CD, of reconstructions using LSP, RF, and LAR for two-way MDC when only one description is received in FS CELP, ITU G.723.1 ACELP, ITU G.723.1 MP-MLQ, and FS MELP.



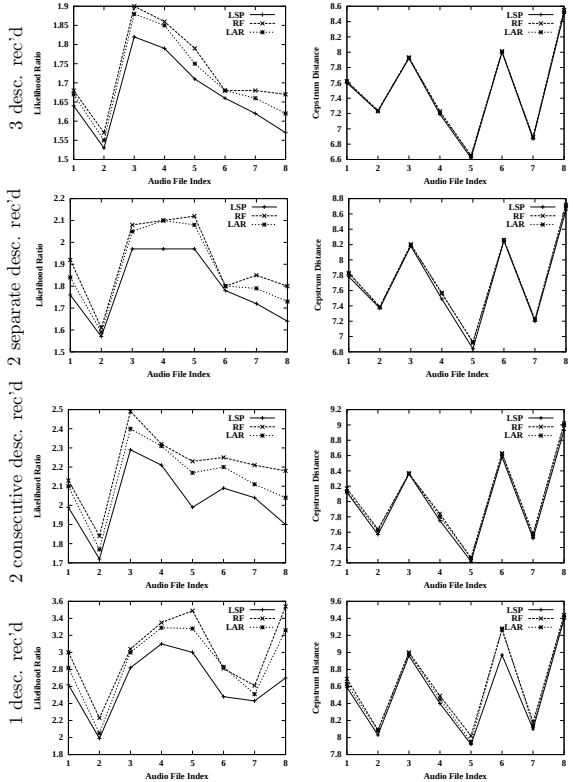
a) LR b) CD

Figure 4.24: Reconstruction qualities of LSP, RF, and LAR representations in terms of LR and CD for four-way MDC in FS CELP under four different loss patterns: three descriptions received, two separate descriptions received, two consecutive descriptions received, and one description received.



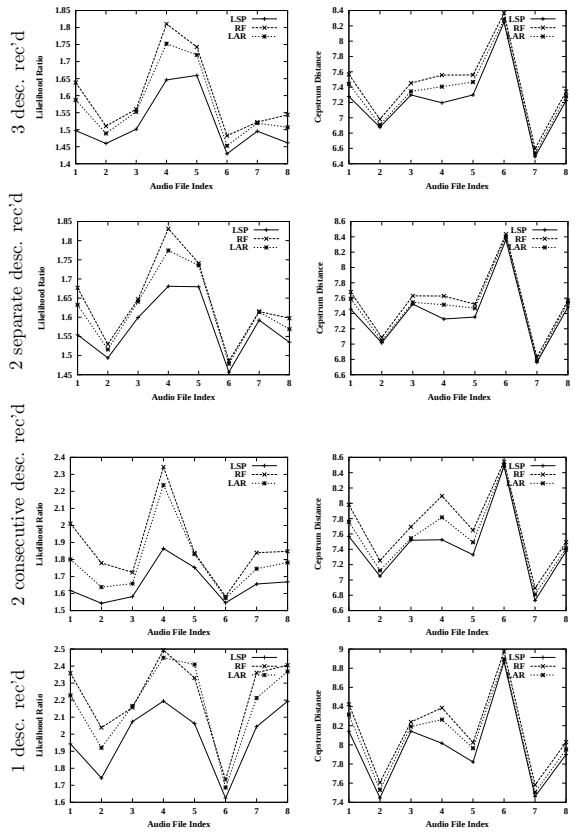
a) LR b) CD

Figure 4.25: Reconstruction qualities of LSP, RF, and LAR representations in terms of LR and CD for four-way MDC in ITU G.723.1 ACELP under four different loss patterns: three descriptions received, two separate descriptions received, two consecutive descriptions received, and one description received.



a) LR b) CD

Figure 4.26: Reconstruction qualities of LSP, RF, and LAR representations in terms of LR and CD for four-way MDC in ITU G.723.1 MP-MLQ under four different loss patterns: three descriptions received, two separate descriptions received, two consecutive descriptions received, and one description received.



a) LR b) CD

Figure 4.27: Reconstruction qualities of LSP, RF, and LAR representations in terms of LR and CD for four-way MDC in FS MELP under four different loss patterns: three descriptions received, two separate descriptions received, two consecutive descriptions received, and one description received.

obvious. This is due to the fact that CD reflects excitation qualities more, and excitation generation is the same for all LP representations.

Overall, our proposed MDC scheme, even under four-way MDC, gives acceptable quality under various loss patterns. Four-way MDC introduces larger degradations due to its larger segment size needed to extract excitation parameters. Such degradations are reflected more obviously by CD. The degree of degradations is also influenced by the degree of loss. The graphs show that qualities degrade more as more descriptions are lost. However, even in the worst case, with 75% packets lost, LP-based MDC has reasonable reconstruction qualities. In case of FS MELP, comparing Figure 4.27 to Figure 4.16, we find that LP-based MDC using LSP improves over sample-based MDC with no loss by 217% (resp. 74%) in terms of LR (resp. CD) on the average.

To give a more concrete idea of the effectiveness of our proposed parameter-based MDC scheme, we list in Table 4.8 the average improvements of our two-way MDC scheme with 50% packet loss over sample-based MDC with no loss. Here, we illustrate using LSP as the linear-predictor representation. The improvements are significant for all coders and especially profound for the very low bit-rate MELP coder, in which LR improves by more than twofold and CD improves by more than onefold.

4.5.2 Reconstruction quality with all descriptions received

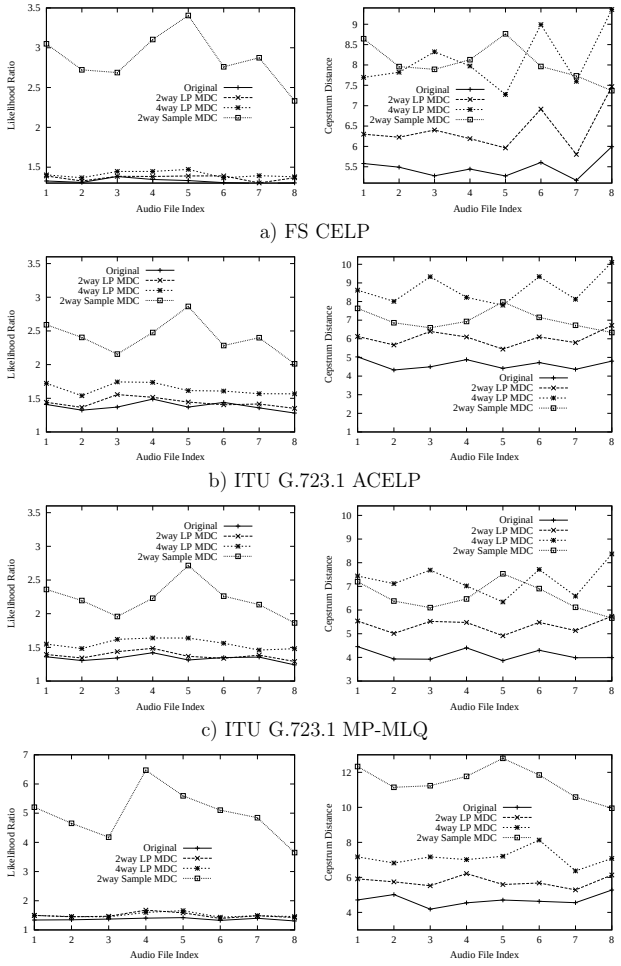
In contrast to the above results that show performance in cases of loss, Figure 4.28 shows the performance of our proposed two-way and four-way MDC with no loss. In particular, the decoding quality of SDC and two-way sample-based MDC with no loss

Table 4.8: Average improvements of our proposed two-way MDC with one description received over sample-based MDC with both descriptions received.

Coder	LR	CD
FS CELP	96%	22%
ACELP	45%	5%
MP-MLQ	36%	9%
MELP	243%	117%

is plotted for comparison purposes. Because all three representations in the case of no loss produce the same decoding streams, there is no need to distinguish them. The LR plots, shown in the left column in Figure 4.28, clearly show that the performance of two-way and four-way LP-based MDC is close to that of SDC and far better than that of two-way sample-based MDC. The results are consistent for all coders. In terms of CD, two-way LP-based MDC is still far better than two-way sample-based MDC, but four-way LP-based MDC may be worse than two-way sample-based MDC in some cases. Quantitatively, Table 4.9 summarizes the average improvements of two-way LP-based MDC over sample-based MDC. The advantage of LP-based MDC is quite obvious. To be complete, Table 4.10 summarizes the average degradations of two-way LP-based MDC as compared to SDC. The degradations in terms of LR are not significant, whereas the degradations in terms of CD reflect the compromise between quality and reliability.

In summary, the above synthetic test results demonstrate that our proposed MDC is effective in concealing loss and achieves a good trade-off between quality and reliability.



a) FS CELP

b) ITU G.723.1 ACELP

c) ITU G.723.1 MP-MLQ

d) ITU G.723.1 FS MELP

Figure 4.28: Quality comparison in terms of LR and CD among SDC with no loss, two-way sample-based MDC with both streams received, two-way LP-based MDC with both streams received, and four-way LP-based MDC with all streams received, for FS CELP, ITU G.723.1 ACELP, ITU G.723.1 MP-MLQ, and ITU G.723.1 MELP.

Table 4.9: Average improvements of two-way LP-based MDC with no loss over two-way sample-based MDC with no loss.

Coder	LR	CD
FS CELP	113%	31%
ACELP	70%	21%
MP-MLQ	62%	29%
MELP	253%	122%

Table 4.10: Average degradation of two-way LP-based MDC with no loss as compared to SDC with no loss.

Coder	LR	CD
FS CELP	3%	17%
ACELP	4%	31%
MP-MLQ	3%	30%
MELP	9%	23%

4.6 Internet Tests

In this section, we present trace-driven simulation results using Internet packet traces collected under the conditions specified in Chapter 3. We have built a real-time voice transmission system for this purpose.

4.6.1 Transmission system prototype

To further verify our algorithms, we evaluate and compare SDC and MDC schemes using traces collected in the Internet. To this end, we built a real-time voice transmission system and an Internet simulator. As shown in Figure 4.29, the test system consists of three processes: a sender, a receiver, and an Internet simulator.

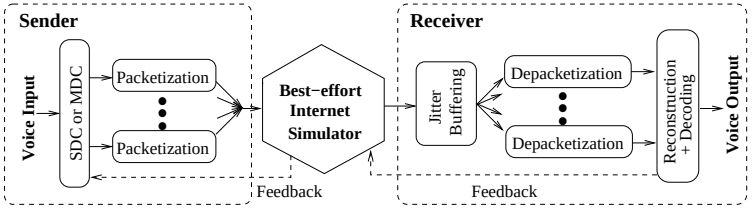


Figure 4.29: Voice transmission and network simulation system.

The *sender* in Figure 4.29 first records voice streams and codes the signals in real time using SDC or LP-based MDC. Next, it packetizes descriptions independently and transmits them concurrently.

After receiving a packet, the *receiver* puts it into the jitter buffer, depacketizes each description, reconstructs missing information if lost packets are detected, decodes the coded stream, and finally plays the regenerated voice streams. Since all traffic experiment results in Chapter 3 demonstrate that traffic patterns are connection dependent and time varying, our system needs to adapt to a specific traffic pattern on the fly. To this end, the receiver should collect and analyze statistics in real time, and feeds this information back to the sender in order for it to adapt its loss-concealment scheme to the dynamic network condition.

Specifically, in our MDC scheme, in order for the sender to adapt the number of descriptions to various loss conditions, the receiver collects loss statistics periodically and asks the sender to switch between two-way and four-way MDC adaptively depending on preset thresholds. In our current implementation, the receiver collects loss statistics every

second and sends a one-bit message in UDP to the sender, indicating whether two-way or four-way MDC should be used. Since the feedback sent by the receiver is subject to loss as well, the receiver will send a feedback packet every second, regardless of whether the degree of interleaving is changed.

In our implementation, the sender starts initially in two-way MDC. When the receiver detects the loss of over 10% of the interleaving sets, it asks the sender to switch to four-way MDC. As long as the decoding state at the receiver is maintained correctly, there will be no quality degradation in switching from two-way to four-way MDC. The sender continues to operate in four-way MDC until the receiver detects the loss of less than 5% of the interleaving sets, as well as the loss of less than 30% of the packets transmitted. Such a strategy is designed to avoid operating in four-way MDC as much as possible unless there are many long bursty losses, as two-way MDC performs better under low-loss conditions.

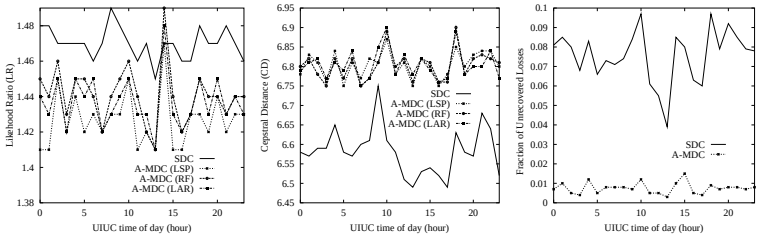
To make fair comparisons, we need to make sure that the different schemes are tested under the same network conditions. Hence, we design a network simulator to simulate the Internet behavior using the recorded traces discussed in Chapter 3. During the tests, the network simulator read in the Internet traffic traces, including round-trip times and received sequence numbers. When it received a packet from the sender, it checked the sequence number and compared it with the sequence numbers in the trace. If the trace indicated that this packet was delivered to the receiver, then the simulator put the packet into its queue and delayed it by the corresponding round-trip time in the trace. When

the time of delay expired, it sent the packet to the receiver. In this way, we make sure that the scheme is tested exactly as if the Internet were used.

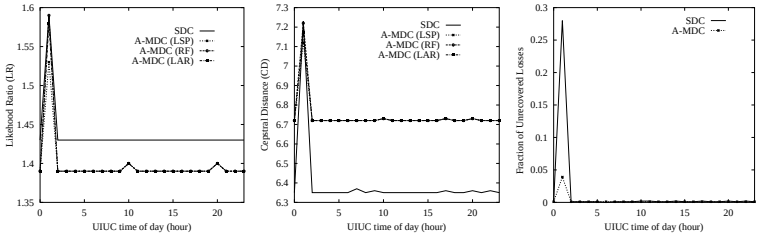
4.6.2 Internet test results

Figures 4.30-4.37 show the trace-driven results on SDC and adaptive two-way/four-way MDC for FS CELP, ITU G.723.1 ACELP, ITU G.723.1 MP-MLQ, and FS MELP, respectively, using the previously discussed transmission system. In the adaptive MDC scheme, the linear predictor may take the form of RF, LAR, or LSP. For each site tested, we show the playback qualities measured in LR and CD, averaged over all received and reconstructed frames for SDC and adaptive MDC with three linear-predictor representations. However, one must be careful in comparing the results of SDC and MDC because LR and CD are not computed for unrecovered frames. Here, the quality of adaptive MDC is averaged over all received and reconstructed frames that account for over 90% of all the frames transmitted, whereas the quality of SDC is averaged over all received frames that account for 60% to 80% of all the frames transmitted (for the UIUC-China and UIUC-Slovakia connections). To illustrate this difference, we also plot the fraction of frames that were lost or unrecovered at the receiver for both SDC and adaptive MDC. The fractions of unrecoverable loss are the same for all three LP representations.

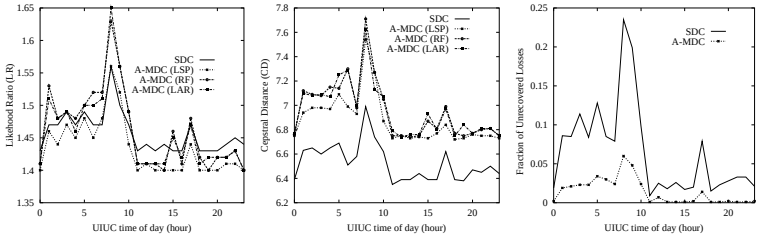
In general, adaptive MDC using LSP always has less distortions than SDC in terms of LR, but may sometimes have more distortions than SDC in terms of CD (to be explained later). In contrast, adaptive MDCs using RF and LAR may have more distortions than



a) UIUC-Portugal round-trip connection

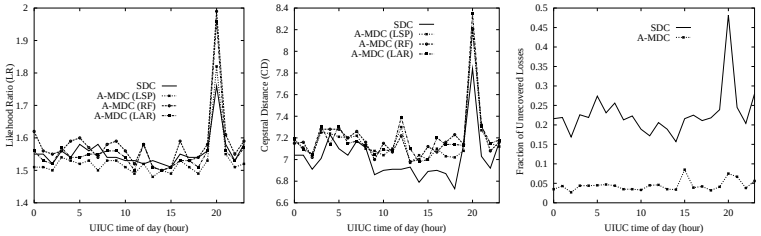


b) UIUC-Berkeley round-trip connection

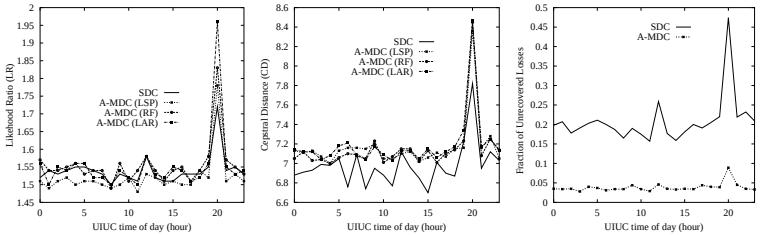


c) UIUC-Egypt round-trip connection

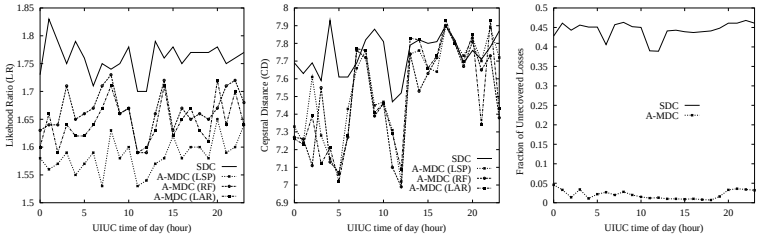
Figure 4.30: Comparison of reconstruction quality between SDC and adaptive two-way/four-way MDC for FS CELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three low-to-medium-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-S. China round-trip connection

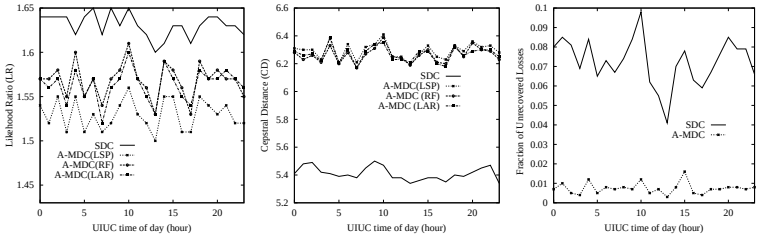


b) UIUC-W. China round-trip connection

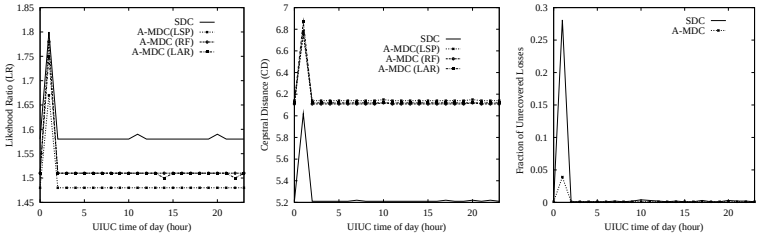


c) UIUC-Slovakia round-trip connection

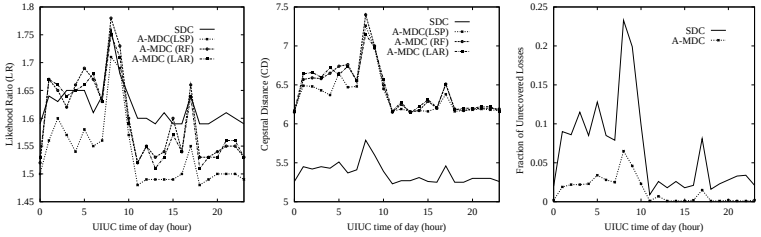
Figure 4.31: Comparison of reconstruction quality between SDC and adaptive two-way/four-way MDC for FS CELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three medium-to-high-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-Portugal round-trip connection

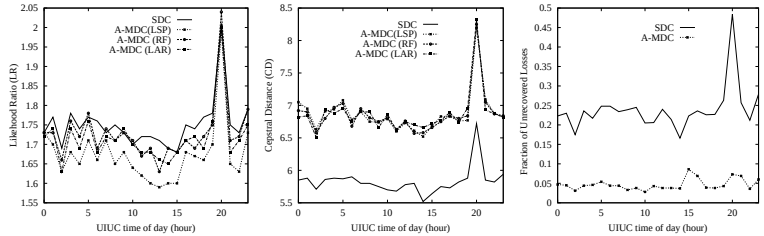


b) UIUC-Berkeley round-trip connection

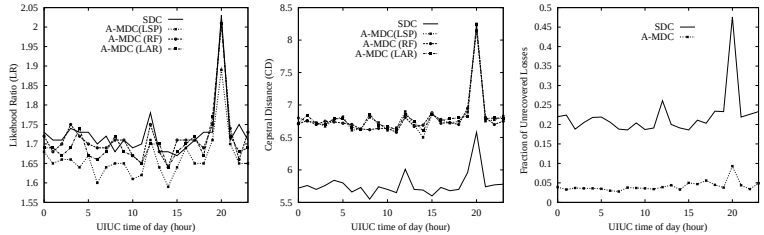


c) UIUC-Egypt round-trip connection

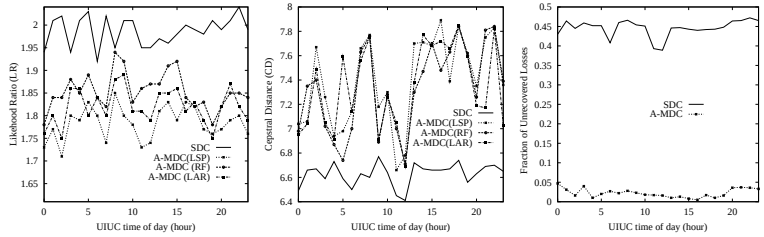
Figure 4.32: Comparison of reconstruction quality between SDC and adaptive two-way/four-way MDC for ITU G.723.1 ACELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three low-to-medium-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-S. China round-trip connection

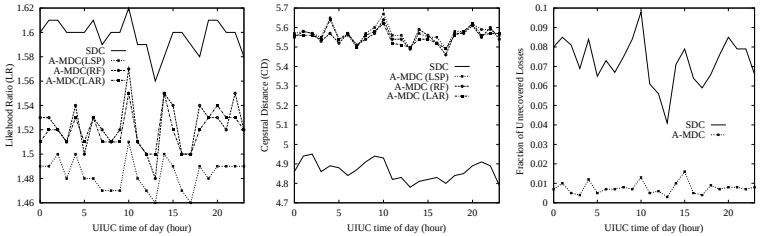


b) UIUC-W. China round-trip connection

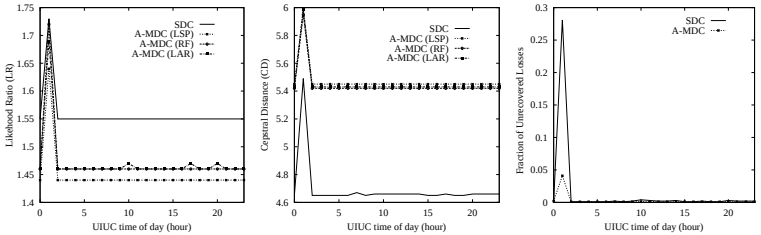


c) UIUC-Slovakia round-trip connection

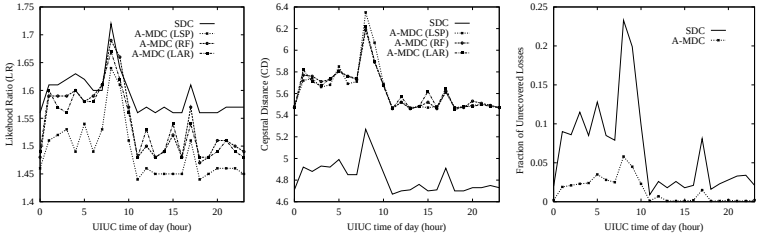
Figure 4.33: Comparison of reconstruction quality between SDC and adaptive two-way/four-way MDC for ITU G.723.1 ACELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three medium-to-high-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-Portugal round-trip connection

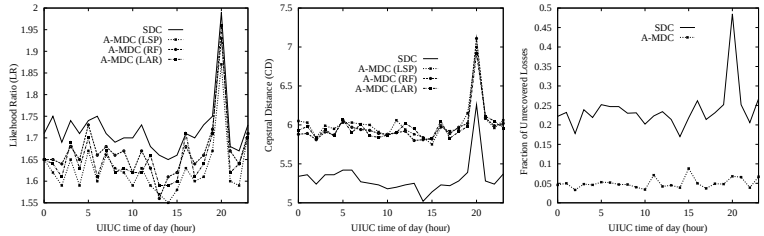


b) UIUC-Berkeley round-trip connection

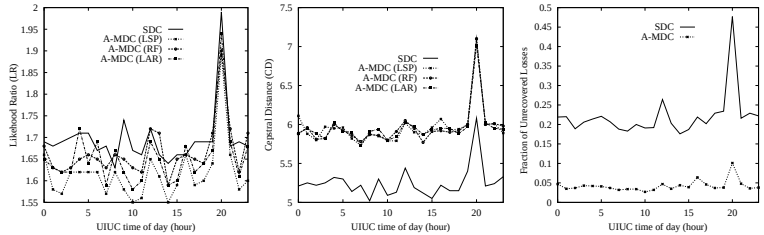


c) UIUC-Egypt round-trip connection

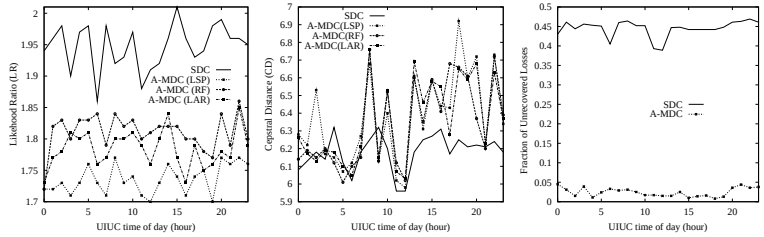
Figure 4.34: Comparison of reconstruction quality between SDC and adaptive two-way/four-way MDC for ITU G.723.1 MP-MLQ on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three low-to-medium-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-S. China round-trip connection

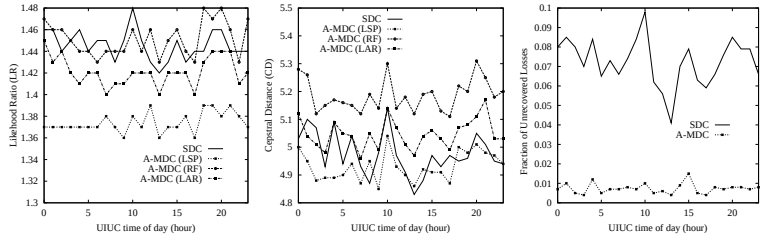


b) UIUC-W. China round-trip connection

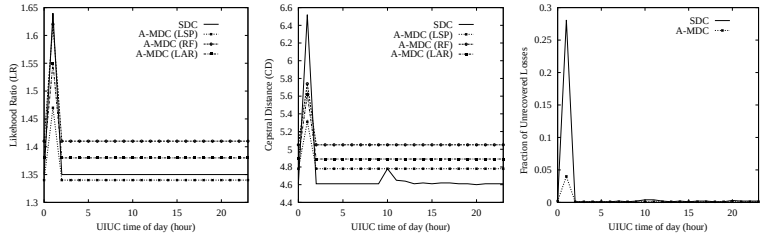


c) UIUC-Slovakia round-trip connection

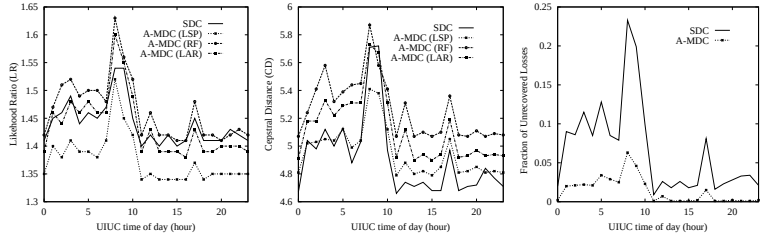
Figure 4.35: Comparison of reconstruction quality between SDC and adaptive two-way/four-way MDC for ITU G.723.1 MP-MLQ on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three medium-to-high-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-Portugal round-trip connection

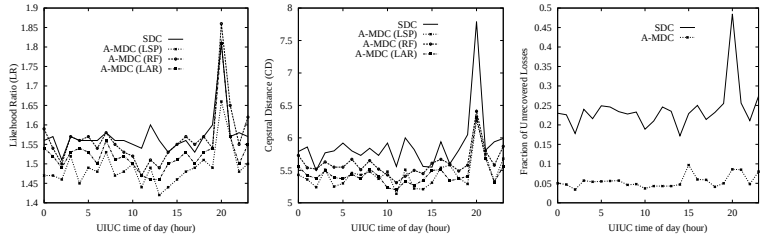


b) UIUC-Berkeley round-trip connection

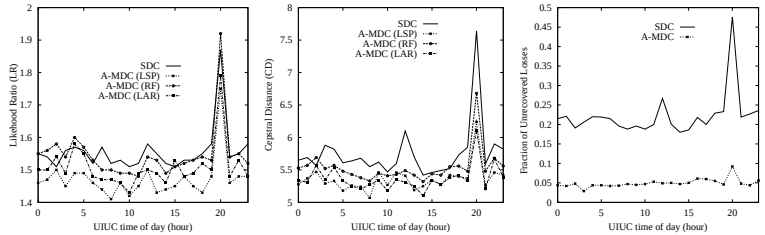


c) UIUC-Egypt round-trip connection

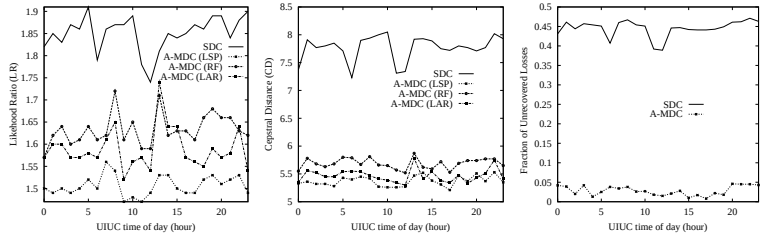
Figure 4.36: Comparison of reconstruction quality between SDC and adaptive two-way/four-way MDC for FS MELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three low-to-medium-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-S. China round-trip connection



b) UIUC-W. China round-trip connection



c) UIUC-Slovakia round-trip connection

Figure 4.37: Comparison of reconstruction quality between SDC and adaptive two-way/four-way MDC for FS MELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three medium-to-high-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.

SDC, even in terms of LR, and almost always have similar distortions as MDC using LSP in terms of CD.

To understand the difference between SDC and MDC, we need to consider two kinds of distortions. First, some distortions in adaptive MDC were introduced because excitations were extracted from larger subframes and were reflected more obviously in terms of CD. Other distortions, in terms of both LR and CD, were introduced when some frames were lost and unrecoverable, leading to incorrect decoding states for subsequent frames received. Such distortions happened to both SDC and adaptive MDC, but affected the quality of SDC more severely due to the large fraction of unrecoverable frames in SDC (see the rightmost graphs in Figures 4.30-4.37). Therefore, based on the combined effects, adaptive MDC almost always performs better than SDC in terms of LR, since adaptive MDC has the same precision in linear-prediction analysis but with far fewer unrecoverable frames. In terms of CD, adaptive MDC may perform better or worse than SDC on received and reconstructed frames, depending mostly on the fraction of unrecoverable losses. For example, for the UIUC-Slovakia connection and for all coders, adaptive MDC using LSP shows significant improvements over SDC in terms of LR. In terms of CD, adaptive MDC improves over SDC for FS CELP and FS MELP, and does not degrade much for ACELP and MP-MLQ.

As shown in Figures 4.36-4.37, the advantage of applying adaptive MDC to FS MELP becomes more apparent. Except for slightly higher distortions in CD in the UIUC-Berkeley and UIUC-Egypt connections, adaptive MDC using LSP performs al-

most consistently better than SDC. The performance gain is particularly noticeable for the high-loss UIUC-Slovakia connection.

From the perspective of end users, SDC will give discontinuous playbacks in high-loss connections, such as the UIUC-China and UIUC-Slovakia connections with over 40% losses. Also, SDC will not be able to reconstruct lost packets and will take time to restore its decoding states, even when a valid packet stream starts flowing again. In contrast, adaptive MDC will give a much smoother playback, despite slightly lower quality on all the frames received or reconstructed due to its increased subframe size.

4.7 Summary

In this chapter, we have proposed to generate multiple descriptions systematically by correlation analysis of voice samples and coding parameters. We have illustrated that sample-based MDC is not applicable to low bit-rate linear-predictive coders. Further, we have developed and tested an LP-based MDC scheme for selected coders, in which a linear predictor may be represented as RF, LAR, or LSP. The advantages of this scheme include the following:

- Aliasing is avoided by not performing down-sampling.
- The precision of linear-prediction analysis is the same as that of the original SDC coder.
- No extra bandwidth is required.
- There is little additional computation overhead of the MDC scheme.

Because we have traded coding quality for reliability by replicating excitation parameters, our proposed MDC scheme has lower coding quality even when all the descriptions are received. This is the major drawback of this scheme. Further, our LP-based MDC scheme applies only to linear-predictive low bit-rate coders, although the MDC design by correlation analysis can be applied to other coders. We have tested our algorithm extensively under various conditions, including both synthetic loss scenarios and real Internet losses, and by various performance measures. Our experimental results lead us to conclude that LP-based MDC has good performance and is effective in concealing losses for both close-looped and open-looped low bit-rate coders. In addition, we have found that the LSP representation has better reconstruction quality than that of RF and LAR.

CHAPTER 5

IMPROVING MDC QUALITY

In the previous chapter, we have found that LP-based MDC using the LSP representation is a good loss-concealment scheme for low bit-rate coded speech. For simplicity, we call it the LSP-based MDC from now on.

Our experimental results in the last chapter have demonstrated quality and reliability trade-offs of our proposed scheme. We replicate hard-to-reconstruct excitations to multiple descriptions for higher reliability, but sacrifice some quality by increasing the number of samples for generating excitation in order to meet the bandwidth constraint.

In this chapter, we investigate different components of the LSP-based MDC scheme in order to further enhance its quality. We first investigate LSP-interpolation algorithms by associating LSP-reconstruction errors with spectral distortions. The method can be applied to the case where packets are lost and the lost LSP vectors need to be reconstructed. We then investigate methods for improving the quality of excitation generation.

Our investigation is based on the observation that even when there is no network loss, quality still degrades due to the longer segment size for generating excitation.

5.1 LSP Reconstruction

In the LSP-based MDC scheme, when some descriptions are lost, we reconstruct lost LSP vectors by linear interpolations of the adjacent received vectors. This reconstruction is naive in the sense that it does not consider its impact on the resulting spectral distortion, which may result very dissimilar playback streams comparing to transmitted streams.

In this section, we study the use of spectral distortions in order to guide the design of LSP reconstructions. To accomplish this, we first need to relate LSP-reconstruction errors to spectral distortions, before we can design interpolation schemes with the objective of minimizing spectral distortions.

5.1.1 Relation between Itakura-Saito Likelihood Ratio and LSP

We have introduced two spectral distortion measures, LR and CD, for low bit-rate coded speech. For the purpose of measuring LSP reconstruction, we choose LR as our measure, because we show in the following that it can be represented solely in terms of LSP. Additionally, we evaluate reconstruction qualities using both LR and CD experimentally.

LR has been introduced in the last chapter. For the sake of clarity, we redefine it here:

$$LR = \frac{\vec{a}_r^T R_o \vec{a}_r}{\vec{a}_o^T R_o \vec{a}_o}, \tag{5.1}$$

where $\vec{a}_o$ and $\vec{a}_r$ are, respectively, the vectors of linear-prediction coefficients of the original and reconstructed speech frame, and R_o is the auto-correlation matrix derived from the original speech:

$$R_o = \begin{pmatrix} r_o(0) & r_o(1) & \cdots & r_o(10) \\ r_o(1) & r_o(0) & \cdots & r_o(9) \\ & & \vdots & \\ r_o(10) & r_o(9) & \cdots & r_o(0) \end{pmatrix}. \tag{5.2}$$

We can express the term $\vec{a}^T R_o \vec{a}$ as:

$$\begin{aligned} \vec{a}^T R_o \vec{a} &= \vec{a}^T \mathrm{E}[\vec{s}\vec{s}^T]\vec{a} \\ &= \mathrm{E}[(\vec{a}^T \vec{s})^2] \\ &= \mathrm{E}\left[(s(n) - \sum_{k=1}^{10} a_k s(n-k))^2\right] = PE, \end{aligned} \tag{5.3}$$

where $\vec{a} = [1, a_1, \cdots, a_{10}]^T$ can be either $\vec{a}_r$ or $\vec{a}_o$, and PE is the prediction error of predictor $\vec{a}$. Because $\vec{a}_o$ is derived by minimizing the prediction error, LR is always greater than or equal to 1. To relate LR to LSP, a spectral representation of $\vec{a}$, we transform PE into the spectral domain:

$$\begin{aligned} PE &= \sum_{i=0}^{10} \sum_{k=0}^{10} a_i r_o(i-k) a_k \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} PSD_o(w) |A(w)|^2 dw, \end{aligned} \tag{5.4}$$

where $PSD_o(w) = \mathcal{F}\{r_o(k)\}$ is the power spectral density of the original speech, $F\{r_o(k)\}$ is the Fourier transform of r_o, and $A(w) = \mathcal{F}\{a(k)\}$. The optimal predictor has the

property [78] that

$$PSD_o(w) = \frac{PE_o}{|A_o(w)|^2}. \tag{5.5}$$

Substituting (5.4)-(5.5) in the (5.1), we have the spectral representation of LR as:

$$\begin{aligned} LR &= \frac{PE_r}{PE_o} && (5.6)\\ &= \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{PE_o|A_r(w)|^2}{|A_o(w)|^2}dw}{PE_o} && (5.7)\\ &= \frac{1}{2\pi}\int_0^{2\pi}\frac{|A_r(w)|^2}{|A_o(w)|^2}dw. && (5.8)\end{aligned}$$

From (4.15) the analysis filter $A_r(w)$ can be represented using the LSP polynomials as [72]:

$$A_r(w) = \frac{1}{2}[P_r(w) + Q_r(w)], \tag{5.9}$$

where $P_r(w)$ and $Q_r(w)$ are derived from (4.14) by replacing z by e^{jw}:

$$\begin{aligned} P_r(w) &= e^{-jw\frac{11}{2}}2^6\cos\frac{w}{2}\prod_{i=1}^{5}(\cos w - \cos x_{r,\,2i-1}) && (5.10)\\ Q_r(w) &= je^{-jw\frac{11}{2}}2^6\sin\frac{w}{2}\prod_{i=1}^{5}(\cos w - \cos x_{r,\,2i}), && (5.11)\end{aligned}$$

and $\vec{x}_r = [x_{r,\,1}, \cdots, x_{r,\,10}]^T$ is the reconstructed LSP vector. Because the complex conjugate of A_r, A_r^* satisfies:

$$A_r^*(w) = A_r(-w), \tag{5.12}$$

and

$$\begin{aligned} P_r(-w) &= e^{jw11}P_r(w) && (5.13)\\ Q_r(-w) &= -e^{jw11}Q_r(w), && (5.14)\end{aligned}$$

$|A_r(w)|^2$ is now:

$$|A_r(w)|^2 = \frac{1}{4}e^{jw11}[P_r^2(w) - Q_r^2(w)]. \tag{5.15}$$

If $|A_r(w)|^2$ in (5.6)-(5.8) is substituted by the above term, LR becomes:

$$LR(\vec{x}_r) = \frac{1}{8\pi}\int_0^{2\pi} e^{jw11}[P_r^2(w) - Q_r^2(w)]\frac{dw}{|A_o(w)|^2}. \tag{5.16}$$

Since the above LR representation is still too complex for us to design LSP reconstructions, we further express it by Taylor expansions:

$$\begin{aligned} LR(\vec{x}_r) &= LR(\vec{x}_o) + \sum_{i=1}^{10}(x_{r,\,i} - x_{o,\,i})\frac{\partial LR(\vec{x}_o)}{\partial x_{r,\,i}} \\ &\quad + \frac{1}{2}\left[\sum_{i=1}^{10}(x_{r,\,i} - x_{o,\,i})^2\frac{\partial^2 LR(\vec{x}_o)}{\partial x_{r,\,i}^2} + \sum_{i=1}^{10}\sum_{j=1,j\neq i}^{10}(x_{r,\,i} - x_{o,\,i})(x_{r,\,j} - x_{o,\,j})\frac{\partial^2 LR(\vec{x_o})}{\partial x_{r,\,i}\partial x_{r,\,j}}\right] \\ &\quad + R_2(\vec{x}_r), \end{aligned} \tag{5.17}$$

where $R_2(\vec{x}_r)$ is the second-order remainder, which by definition is:

$$R_2(\vec{x}_r) = \frac{1}{6}\sum_{i=1}^{10}\sum_{j=1}^{10}\sum_{k=1}^{10}(x_i' - x_{o,\,i})(x_j' - x_{o,\,j})(x_k' - x_{o,\,k})\frac{\partial^3 LR(\vec{x'})}{\partial x_{r,\,i}\partial x_{r,\,j}\partial x_{r,\,k}}, \tag{5.18}$$

and $\vec{x'} = \vec{x}_o + \theta(\vec{x}_r - \vec{x}_o)$ with $0 < \theta < 1$.

Obviously, $LR(\vec{x}_o)$ equals 1. Furthermore, because $\vec{x}_o$ minimizes LR, $\frac{\partial LR(\vec{x}_o)}{\partial x_{r,\,i}}$ is zero. Now, consider the second-order term. Define the second-order coefficients as $W_{i,j}$ for $i \neq j$ and W_i for $i = j$. Starting from (5.19), we evaluate $W_{i,j}$ first:

$$\begin{aligned} W_{i,j} &= \frac{\partial^2 LR(\vec{x_o})}{\partial x_{r,\,i}\partial x_{r,\,j}} && (5.19)\\ &= \frac{1}{8\pi}\int_0^{2\pi} e^{jw11}\frac{\partial^2[P_r^2(w) - Q_r^2(w)]}{\partial x_{r,\,i}\partial x_{r,\,j}}\bigg|_{\vec{x}_o}\frac{dw}{|A_o(w)|^2}. && (5.20)\end{aligned}$$

If i is even and j is odd, or vise versa, $W_{i,j}$ equals zero because $P_r(w)$ only contains odd $x_{r,\,i}$ terms and $Q_r(w)$ only contains even $x_{r,\,i}$ terms. If both i and j are even or odd, without loss of generality, assume that i and j are both odd. After evaluating the second-order derivative of $P_r(w)$ with respect to $x_{r,\,i}$ and $x_{r,\,j}$, substitute the results into $W_{i,j}$:

$$\begin{aligned} W_{i,j} &= \frac{1}{2\pi}\int_0^{2\pi} e^{jw11}\frac{P_o^2(w)\sin x_{o,\,i}\sin x_{o,\,j}}{(\cos w-\cos x_{o,\,i})(\cos w-\cos x_{o,\,j})}\frac{dw}{|A_o(w)|^2} && (5.21)\\ &= \frac{\sin x_{o,\,i}\sin x_{o,\,j}}{2\pi(\cos x_{o,\,j}-\cos x_{o,\,i})} \\ &\quad \cdot\int_0^{2\pi}\frac{e^{jw11}}{|A_o(w)|^2}\left[\frac{P_o^2(w)}{(\cos w-\cos x_{o,\,i})}-\frac{P_o^2(w)}{(\cos w-\cos x_{o,\,j})}\right]dw. && (5.22)\end{aligned}$$

To compute the above expression, we use the property that the first-order derivative of LR evaluated at $\vec{x}_o$ is zero:

$$\begin{aligned} 0=\frac{\partial LR(\vec{x}_o)}{\partial x_{r,\,i}} &= \frac{1}{8\pi}\int_0^{2\pi} e^{jw11}\frac{\partial[P_r^2(w)-Q_r^2(w)]}{\partial x_{r,\,i}}\bigg|_{\vec{x}_o}\frac{dw}{|A_o(w)|^2} \\ &= \frac{\sin x_{o,\,i}}{4\pi}\int_0^{2\pi} e^{jw11}\frac{P_o^2(w)}{(\cos w-\cos x_{o,\,i})}\frac{dw}{|A_o(w)|^2}. \end{aligned} \quad (5.23)$$

Hence, $W_{i,j}$ equals zero.

Next, we compute W_i:

$$\begin{aligned} W_i &= \frac{\partial^2 LR(\vec{x_o})}{\partial x_{r,\,i}^2} && (5.24)\\ &= \frac{1}{4\pi}\int_0^{2\pi} e^{jw11}\frac{P_o^2(w)\sin^2 x_{o,\,i}}{(\cos w-\cos x_{o,\,i})^2}\frac{dw}{|A_o(w)|^2}. && (5.25)\end{aligned}$$

Substituting e^{jw} by z, and $A_o(z)$ by:

$$A_o(z)=\prod_{k=1}^{10}(1-z_{o,\,k}z^{-1}), \quad (5.26)$$

we have

$$\begin{aligned} W_i &= \frac{1}{\pi j}\int_{u.c.}\frac{z^8P_o^2(z)\sin^2 x_{o,\,i}}{(1-2\cos x_{o,\,i}z^{-1}+z^{-2})^2}\frac{dz}{|A_o(z)|^2} && (5.27)\\ &= \frac{1}{\pi j}\int_{u.c.}\frac{z^8P_o^2(z)\sin^2 x_{o,\,i}}{(1-2\cos x_{o,\,i}z^{-1}+z^{-2})^2}\frac{1}{\prod_{k=1}^{10}(1-z_{o,\,k}z^{-1})}\frac{1}{\prod_{k=1}^{10}(1-z_{o,\,k}z)}dz && (5.28)\\ &= \frac{\sin^2 x_{o,\,i}}{\pi j}\int_{u.c.}\frac{(1+z)^2\prod_{k=1,2k-1\neq i}^{5}(z^2-2\cos x_{o,\,2k-1}z+1)^2}{\prod_{k=1}^{10}(z-z_{o,\,k})\prod_{k=1}^{10}(1-z_{o,\,k}z)}dz, && (5.29)\end{aligned}$$

where $u.c.$ denotes the unit circle. There are 10 first-order poles, $z_{o,\,1},\cdots,z_{o,\,10}$, inside the unit circle. By applying the Residue Theorem [79] and by letting $D(z)$ denote the function inside the integral, we can convert the integration to summation:

$$W_i = 2\sin^2 x_{o,\,i}\sum_{l=1}^{10} res[D(z_{o,\,l})] \quad (5.30)$$

$$= 2\sin^2 x_{o,\,i}\sum_{l=1}^{10}\frac{(1+z_{o,\,l})^2\prod_{k=1}^{5}(z_{o,\,l}^2-2\cos x_{o,\,2k-1}z_{o,\,l}+1)^2}{(1-2\cos x_{o,\,i}z_{o,\,l}+z_{o,\,l}^2)^2\prod_{k=1,k\neq l}^{10}(z_{o,\,l}-z_{o,\,k})\prod_{k=1}^{10}(1-z_{o,\,k}z_{o,\,l})}. \quad (5.31)$$

Further, because the following equality holds,

$$\begin{aligned} (1+z_{o,\,l})\prod_{k=1}^{5}(z_{o,\,l}^2-2\cos x_{2k-1}z_{o,\,l}+1) &= P(z_{o,\,l}^{-1}) \\ &= A(z_{o,\,l}^{-1})+z^{-11}A(z_{o,\,l}) \\ &= A(z_{o,\,l}^{-1}) \\ &= \prod_{k=1}^{10}(1-z_{o,\,k}z_{o,\,l}), \end{aligned} \quad (5.32)$$

W_i can be simplified as:

$$W_i = 2\sin^2 x_{o,\,i}\sum_{l=1}^{10}\frac{(1+z_{o,\,l})\prod_{k=1,(2k-1)\neq i}^{5}(z_{o,\,l}^2-2\cos x_{o,\,2k-1}z_{o,\,l}+1)}{(1-2\cos x_{o,\,i}z_{o,\,l}+z_{o,\,l}^2)\prod_{k=1,k\neq l}^{10}(z_{o,\,l}-z_{o,\,k})}. \quad (5.33)$$

Applying the Residue Theorem again, we arrive at:

$$W_i = \frac{\sin^2 x_{o,\,i}}{\pi j} \int_c \frac{(1+z) \prod_{k=1,(2k-1)\neq i}^{5} (z^2 - 2\cos x_{o,\,2k-1} z + 1)}{(1 - 2\cos x_{o,\,i} z + z^2) \prod_{k=1}^{10} (z - z_{o,\,k})} dz, \tag{5.34}$$

where c is the integration curve inside the unit circle but outside the outermost $z_{o,\,k}$. After changing the integration variable z to z^{-1},

$$W_i = \frac{\sin^2 x_{o,\,i}}{\pi j} \int_{c'} \frac{z(1+z) \prod_{k=1,(2k-1)\neq i}^{5} (z^2 - 2\cos x_{o,\,2k-1} z + 1)}{(z^2 - 2\cos x_{o,\,i} z + 1) \prod_{k=1}^{10} (1 - z_{o,\,k} z)} dz, \tag{5.35}$$

where c' is the integration contour outside the unit circle but inside the innermost $z_{o,\,k}^{-1}$, we have only two poles, $e^{x_{o,\,i}}$ and $e^{-x_{o,\,i}}$, inside the integration curve and on the unit circle. By applying the Residue Theorem once more, we have:

$$\begin{aligned} W_i = 2\sin^2 x_{o,\,i} \Bigg[& \frac{z(1+z) \prod_{k=1,(2k-1)\neq i}^{5} (z^2 - 2\cos x_{o,\,2k-1} z + 1)}{(z - e^{-x_{o,\,i}}) \prod_{k=1}^{p} (1 - z_{o,\,k} z)} \Bigg|_{z=e^{x_{o,\,i}}} \\ & + \frac{z(1+z) \prod_{k=1,(2k-1)\neq i}^{5} (z^2 - 2\cos x_{o,\,2k-1} z + 1)}{(z - e^{x_{o,\,i}}) \prod_{k=1}^{10} (1 - z_{o,\,k} z)} \Bigg|_{z=e^{-x_{o,\,i}}} \Bigg]. \end{aligned} \tag{5.36}$$

Taking into account that $e^{x_{o,\,i}}$ and $e^{-x_{o,\,i}}$ are the roots of $P_o(z)$, we have the following equalities:

$$\begin{aligned} \prod_{k=1}^{10} (1 - e^{x_{o,\,i}} z_{o,\,k}) &= A(e^{-x_{o,\,i}}) \\ &= \frac{P_o(e^{-x_{o,\,i}}) + Q_o(e^{-x_{o,\,i}})}{2} \\ &= \frac{Q_o(e^{-x_{o,\,i}})}{2} \\ &= \frac{(1 - e^{x_{o,\,i}}) \prod_{k=1}^{5} (e^{2x_{o,\,i}} - 2\cos x_{o,\,2k} e^{x_{o,\,i}} + 1)}{2}, \end{aligned} \tag{5.37}$$

and

$$\begin{aligned} \prod_{k=1}^{10} (1 - e^{-x_{o,\,i}} z_{o,\,k}) &= A(e^{x_{o,\,i}}) \\ &= \frac{P_o(e^{x_{o,\,i}}) + Q_o(e^{x_{o,\,i}})}{2} \\ &= \frac{Q_o(e^{x_{o,\,i}})}{2} \\ &= \frac{(1 - e^{-x_{o,\,i}}) \prod_{k=1}^{5} (e^{-2x_{o,\,i}} - 2\cos x_{o,\,2k} e^{-x_{o,\,i}} + 1)}{2}. \end{aligned} \tag{5.38}$$

Also, we know that

$$\begin{aligned} e^{2x_{o,\,i}} - 2\cos x_{o,\,l} e^{x_{o,\,i}} + 1 &= 2e^{x_{o,\,i}} (\cos x_{o,\,i} - \cos x_{o,\,l}) \\ e^{-2x_{o,\,i}} - 2\cos x_{o,\,l} e^{-x_{o,\,i}} + 1 &= 2e^{-x_{o,\,i}} (\cos x_{o,\,i} - \cos x_{o,\,l}). \end{aligned} \tag{5.39}$$

Finally, substituting (5.37)-(5.39) into (5.36), we resolve W_i as

$$W_i = 2\sin^2 x_{o,\,i} \frac{\prod_{k=1,(2k-1)\neq i}^{5} (\cos x_{o,\,i} - \cos x_{o,\,2k-1})}{\prod_{k=1}^{5} (\cos x_{o,\,i} - \cos x_{o,\,2k})(1 - \cos x_{o,\,i})}. \tag{5.40}$$

Here, W_i can be computed from LSPs directly.

The above derivation is for the case when i is odd. Similarly, if i is even, the second-order Taylor expansion coefficient W_i is

$$W_i = -2\sin^2 x_{o,\,i} \frac{\prod_{k=1,2k\neq i}^{5} (\cos x_{o,\,i} - \cos x_{o,\,2k})}{\prod_{k=1}^{5} (\cos x_{o,\,i} - \cos x_{o,\,2k-1})(1 + \cos x_{o,\,i})}. \tag{5.41}$$

Now, consider the signs of W_i. For each i, there are $i-1$ LSPs less than $x_{o,\,i}$, and each contributes a negative sign in the above product terms of W_i (see (5.40) and (5.41)). If i is odd, then $i-1$ is even, and we have W_i positive. If i is even, then $i-1$ is odd, and we still have W_i positive due to the minus sign at the beginning of the right-hand side of (5.41). Hence, these second-order terms will not cancel out.

Another observation of W_i is that it is determined by the differences between x_i and other LSPs. Because of the ordering property, the smallest term among all $|\cos x_{o,\,i} - \cos x_{o,\,k}|$ for $k \neq i$ is either $|\cos x_{o,\,i} - \cos x_{o,\,i-1}|$ or $|\cos x_{o,\,i} - \cos x_{o,\,i+1}|$. Both terms are in the denominators of (5.40) and (5.41), with the smaller of these two terms being the dominant determinant of the value of W_i. In the previous chapter, we have mentioned that if $x_{o,\,i}$ is very close to its neighboring LSPs, then there will be a spectral formant. Here, we see that W_i increases as the distances between $x_{o,\,i}$ and its neighboring LSPs decrease. That is to say, the second-order error is treated more importantly if it were in the formant region. This result agrees with the quantization policy of LSP used in ITU G.723.1. In the coder, the quantization of an LSP is guided by its minimum distance to its neighbors.

For the remainder term R_2, we were not able to get a simple form similar to the second-order terms, because it is evaluated at $\vec{x'}$, not $\vec{x}_o$. However, experimental evaluations in the next section show that R_2 is also reduced by minimizing the second-order term.

In summary, we have built the relationship between the spectral distortion measure, LR, and LSP reconstruction errors. This relationship is used in the following sections to generate the optimal reconstructions.

5.1.2 Optimized two-point linear LSP reconstruction

In our LSP-based MDC scheme, if a frame is lost, we only need to reconstruct the lost LSP vector by interpolating adjacent received ones. In the following, we design LSP interpolations with the goal of minimizing the second-order approximation of the LR.

In the last section, we have derived the spectral distortion in terms of LSP for a single frame. In practice, we like to minimize the average distortion across all the frames tested; that is,

$$\min \mathcal{E} = \min \mathrm{E}\left[\sum_{i=1}^{10} W_i^n (x_{r,\,i}^n - x_i^n)^2\right], \tag{5.42}$$

where $[x_{r,\,1}^n, \cdots, x_{r,\,10}^n]^T = \vec{x}_r^n$ is the reconstructed LSP vector for the n^{th} frame, and $[x_1^n, \cdots, x_{10}^n]^T = \vec{x}^n$ is the original LSP vector for the n^{th} frame. We can decompose the above term as:

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^{10} W_i^n (x_{r,\,i}^n - x_i^n)^2\right] &= p_l \times \mathrm{E}\left[\vec{W^n}^T (\vec{x}_r^n - \vec{x}^n)^2 \mid \widetilde{f}_n\right] \\
&= \left\{ p_1 \mathrm{E}\left[\vec{W^n}^T (\vec{x}_r^n - \vec{x}^n)^2 \mid f_{n-1}\widetilde{f}_n f_{n+1}\right] \right. \\
&\qquad \left. + p_2 \mathrm{E}\left[\vec{W^n}^T (\vec{x}_r^n - \vec{x}^n)^2 \mid \text{else}\right] \right\} \\
&= p_1 \mathcal{E}_1 + p_2 \mathcal{E}_2,
\end{aligned} \tag{5.43}$$

where $\vec{W^n}$ equals $[W_1^n, \cdots, W_{10}^n]^T$, and p_l is the probability that a frame is lost. Here, p_1 is the probability that a frame is lost but its preceding and succeeding frames are received; and p_2 is the loss probability of all other cases, namely, $p_l - p_1$. Further, f denotes the event that a frame is received, and $\widetilde{f}$ denotes the event that a frame is lost. Clearly, p_1 and p_2 depend on the statistics of the on-going connection. To get connection-independent results, we minimize each $\mathcal{E}_i, i = 1, 2$ in (5.43) separately, as we do not need to be restricted to using the same interpolation for every loss scenario. Without loss of generality, in the following, we derive optimal interpolations targeted to minimize $\mathcal{E}_1$ in (5.43).

In this case, the interpolation used before to reconstruct $\vec{x}_r^n$ is:

$$\vec{x}_r^n = \frac{\vec{x}^{n-1} + \vec{x}^{n+1}}{2}. \tag{5.44}$$

It is a special form of the more general two-point first-order interpolation:

$$\vec{x}_r^n = \alpha\vec{x}^{n-1} + \beta\vec{x}^{n+1}, \tag{5.45}$$

where α and β are used to minimize $\mathcal{E}_1$ as follows:

$$\mathrm{E}\left[\sum_{i=1}^{10} W_i^n(\alpha x_i^{n-1} + \beta x_i^{n+1} - x_i^n)^2\right]. \tag{5.46}$$

$\mathcal{E}_1$ can be estimated by

$$\frac{1}{N}\sum_{n=1}^{N}\left[\sum_{i=1}^{10} W_i^n(\alpha x_i^{n-1} + \beta x_i^{n+1} - x_i^n)^2\right], \tag{5.47}$$

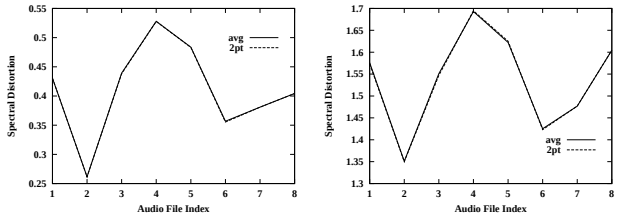
where N is the number of frames. Hence, α and β are given by:

$$\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \begin{bmatrix}\sum_n\sum_{i=1}^{10} W_i^n x_i^{n-1}x_i^{n-1}, & \sum_n\sum_{i=1}^{10} W_i^n x_i^{n+1}x_i^{n-1}\\ \sum_n\sum_{i=1}^{10} W_i^n x_i^{n-1}x_i^{n+1}, & \sum_n\sum_{i=1}^{10} W_i^n x_i^{n+1}x_i^{n+1}\end{bmatrix}^{-1}\begin{bmatrix}\sum_n\sum_{i=1}^{10} W_i^n x_i^n x_i^{n-1}\\ \sum_n\sum_{i=1}^{10} W_i^n x_i^n x_i^{n+1}\end{bmatrix}. \tag{5.48}$$

In our experiments, we compute α and β by going through the following steps. First, we extract linear-prediction vectors $\vec{a}_o$ and auto-correlation matrices R_o from the speech streams with a frame length of 240 samples. Second, after converting linear-prediction vectors to LSP vectors $\vec{x}$, we compute weighting vectors W. Last, we compute α and β according to (5.48). Once α and β are available, we can compute the reconstructed LSP vectors x_r, as well as $\mathcal{E}_1$ according to (5.47). Further, we can convert the reconstructed LSP vectors back to the linear-prediction vectors $\vec{a}_r$ and calculate the corresponding averaged LR using (5.1).

Table 5.1: Optimal two-point first-oder interpolation coefficients for the eight test streams in Table 1.1.

	Test stream							
	1	2	3	4	5	6	7	8
α	0.5115	0.5053	0.4966	0.5040	0.5052	0.5084	0.5032	0.5052
β	0.4887	0.4947	0.5030	0.4960	0.4950	0.4914	0.4970	0.4950

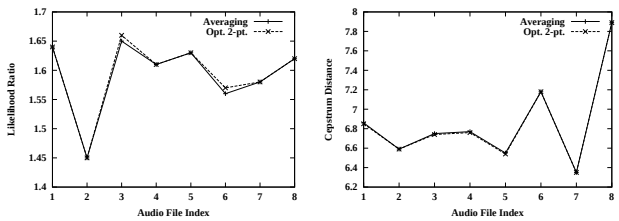


a) $\mathcal{E}1$ b) LR

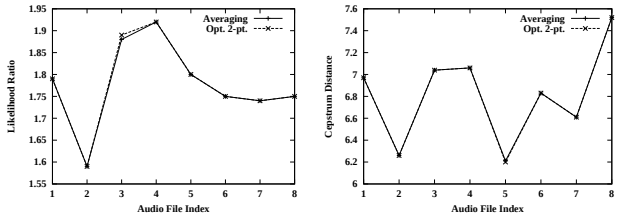
Figure 5.1: A comparison between optimal two-point first-order interpolation and averaging in terms of $\mathcal{E}_1$ and its corresponding LR evaluated on the LSP vectors extracted from the eight test streams.

For the eight test streams in Table 1.1, Table 5.1 shows the resulting α and β. We observe that α and β are both positive and are very close to 0.5. This means that all the reconstructed LSP vectors are valid linear predictors. Figure 5.1 plots the $\mathcal{E}_1$ and its corresponding LR for interpolations based on averaging and optimized two-point first-order interpolations. Since the results for the two interpolation methods are overlapping, optimized two-point interpolations do not improve over averaging.

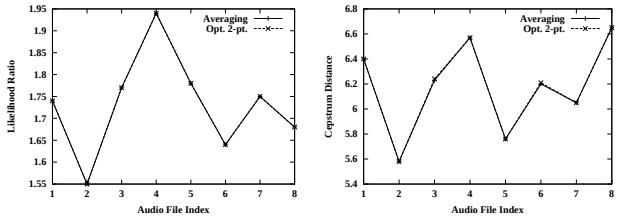
Further, Figure 5.2 shows the reconstruction quality of both interpolation methods for the two-way LSP-based MDC implemented on the test coders. Here, LR and CD mea-



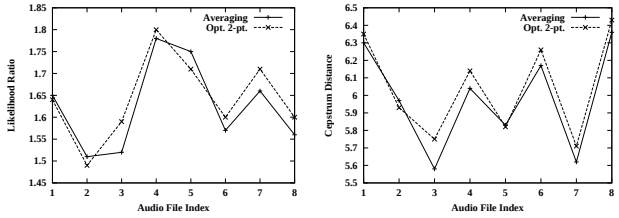
a) FS CELP



b) ITU G.723.1 ACELP



c) ITU G.723.1 MP-MLQ



d) ITU G.723.1 FS MELP

Figure 5.2: A comparison of optimal two-point first-order interpolation and averaging in terms of LR and CD for two-way MDC when only one description is received.

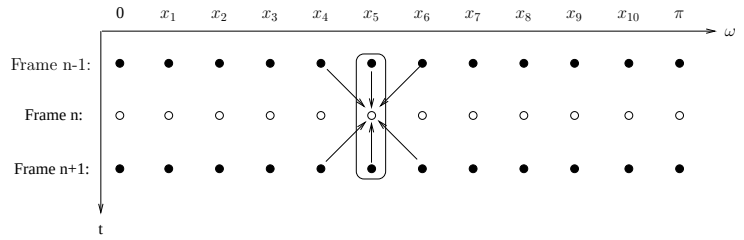


Figure 5.3: A two-dimensional view of LSP.

sure the distortions of the decoded streams with respect to those of the original speech. For the top three coders, the two interpolation methods give almost identical results, with minor degradations for the optimized interpolation method in several cases. This result is possibly due to the fact that the optimizations are carried out on unquantized LSP vectors and do not include quantization effects. Overall, for FS MELP, averaging outperforms optimized interpolations. Besides quantization effects, the differences may also be attributed to frame size, since the optimized coefficients are calculated using a frame size of 240 samples, whereas the frame size of MELP is 180 samples.

5.1.3 Optimized six-point linear LSP reconstruction

Optimal two-point first-order interpolations do not improve over averaging because both methods restrict reconstructions along the time axis. Figure 5.3 plots a two-dimensional view of LSP, with time in one dimension and spectrum along the other. For each time interval, an LSP vector is an ordered vector falling in the range of $[0,\ \pi]$ in

spectral space. This suggests that each x_i is related to its neighbors x_{i-1} and x_{i+1} along the spectral axis. The interpolations studied in the last section are along the time axis that is shown by the bounded box in the graph. However, based on the ordering property and the fact that spectral distortion is related to LSP differences, we know, for example, that x_4^n and x_6^n contain information about x_5^n. Also, x_4^{n-1} and x_4^{n+1} contain information about x_4^n, while x_6^{n-1} and x_6^{n+1} contain information about x_6^n. Therefore, it is reasonable to incorporate x_4^{n-1}, x_4^{n+1}, x_6^{n-1}, and x_6^{n+1} besides x_5^{n-1} and x_5^{n+1} in reconstructing x_5^n. In the following, we design interpolations using six LSPs around the lost LSP in both time and spectral domains.

We start with six-point linear interpolations. Similar to the two-point case, we minimizes $\mathcal{E}_1$ with respect to $\vec{\alpha}$ and $\vec{\beta}$:

$$\min_{\vec{\alpha},\vec{\beta}} \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{i=1}^{10} W_i^n (x_{r,\,i}^n - x_i^n)^2\right], \tag{5.49}$$

where

$$x_{r,\,i}^n \;=\; [x_{i-1}^{n-1}, x_i^{n-1}, x_{i+1}^{n-1}, x_{i-1}^{n+1}, x_i^{n+1}, x_{i+1}^{n+1}] \times [\alpha_{-1}, \alpha_0, \alpha_1, \beta_{-1}, \beta_0, \beta_1]^T, \tag{5.50}$$

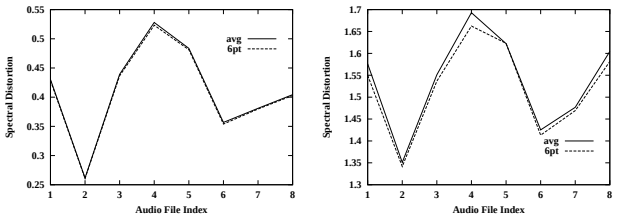
and $\vec{\alpha}$ and $\vec{\beta}$ are the solutions to the following simultaneous equations:

$$\sum_n \sum_{i=1}^{10} W_i^n (\hat{x}_i^n - x_i^n) x_{i+m}^{n+l} = 0, \quad l = \{-1, 1\}, m = \{-1, 0, 1\}. \tag{5.51}$$

Table 5.2 lists the optimized six-point linear-interpolation coefficients. Coefficients α_0 and β_0 are the dominant terms for all the test streams, and both are close to 0.5. They show that the added terms make little contribution to the final interpolation results. Figure 5.4 plots the resulting $\mathcal{E}_1$ and LR for optimized six-point linear interpolations and

Table 5.2: Optimal six-point first-order interpolation coefficients for the eight test streams in Table 1.1.

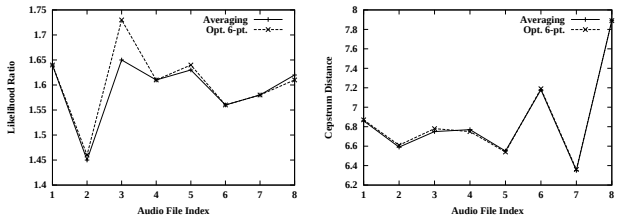
	Test stream							
	1	2	3	4	5	6	7	8
α_{-1}	-0.0147	0.0073	-0.0112	-0.0248	-0.0203	0.0012	0.0158	0.0166
α_0	0.5214	0.4981	0.4856	0.5179	0.5247	0.5438	0.5088	0.5074
α_1	0.0094	0.0087	0.0294	0.0141	0.0061	-0.0373	-0.0220	-0.0135
β_{-1}	0.0294	-0.0013	0.0276	0.0274	0.0066	0.0063	-0.0050	-0.0058
β_0	0.4920	0.5110	0.4947	0.5119	0.5119	0.4754	0.4972	0.5041
β_1	-0.0318	-0.0213	-0.0264	-0.0426	-0.0278	0.0140	0.0090	-0.0049



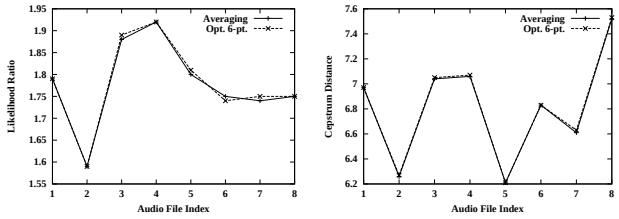
a) $\mathcal{E}1$ b) LR

Figure 5.4: A comparison of optimal six-point first-order interpolations and averaging in terms of $\mathcal{E}_1$ and its corresponding LR evaluated on the LSP vectors extracted from the eight test streams.

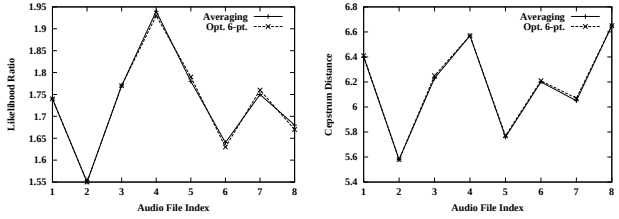
averaging. The values of $\mathcal{E}_1$ are almost identical for both methods, while the optimized interpolation method improves the overall LR slightly. Figure 5.5 shows the reconstruction qualities of the two interpolation methods when applied to the test coders. In short, six-point linear interpolations do not improve over averaging, although it is better than two-point linear interpolations when applied to FS MELP.



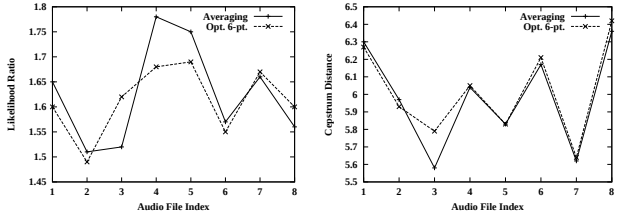
a) FS CELP



b) ITU G.723.1 ACELP



c) ITU G.723.1 MP-MLQ



d) ITU G.723.1 FS MELP

Figure 5.5: A comparison of optimal six-point first-order interpolations and averaging in terms of LR and CD for two-way MDC when only one description is received.

5.1.4 Optimized six-point second-order LSP reconstruction

Last, we test another alternative, six-point second-order interpolations by minimizing $\mathcal{E}_1$:

$$\min_{\vec{\alpha},\vec{\beta},\vec{\gamma},\vec{\eta},\vec{\zeta}} \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{i=1}^{10} W_i^n (x_{r,\,i}^n - x_i^n)^2\right], \tag{5.52}$$

where

$$\begin{aligned} x_{r,\,i}^n = &\sum_{k=-1}^{1} \alpha_k x_{i+k}^{n-1} + \sum_{k=-1}^{1} \beta_k x_{i+k}^{n+1} \\ &+ \sum_{j=-1}^{1}\sum_{k\geq j}^{1} \gamma_{jk} x_{i+j}^{n-1} x_{i+k}^{n-1} + \sum_{j=-1}^{1}\sum_{k\geq j}^{1} \eta_{jk} x_{i+j}^{n+1} x_{i+k}^{n+1} + \sum_{j=-1}^{1}\sum_{k=-1}^{1} \zeta_{jk} x_{i+j}^{n-1} x_{i+k}^{n+1}. \end{aligned} \tag{5.53}$$

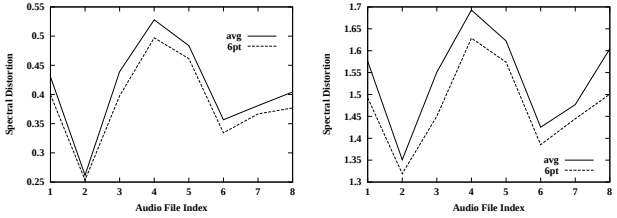
The coefficients are the solutions to the following simultaneous equations:

$$\begin{cases} \sum_n \sum_{i=1}^{10} W_i^n (\hat{x}_i^n - x_i^n) x_{i+m}^{n+l} = 0, \quad l = \{-1,1\}, m = \{-1,0,1\} \\ \sum_n \sum_{i=1}^{10} W_i^n (\hat{x}_i^n - x_i^n) x_{i+l}^{n-1} x_{i+m}^{n-1} = 0, \quad l = \{-1,0,1\}, l \leq m \leq 1 \\ \sum_n \sum_{i=1}^{10} W_i^n (\hat{x}_i^n - x_i^n) x_{i+l}^{n+1} x_{i+m}^{n+1} = 0, \quad l = \{-1,0,1\}, l \leq m \leq 1 \\ \sum_n \sum_{i=1}^{10} W_i^n (\hat{x}_i^n - x_i^n) x_{i+l}^{n-1} x_{i+m}^{n+1} = 0, \quad l = \{-1,0,1\}, m = \{-1,0,1\}. \end{cases} \tag{5.54}$$

Table 5.3 shows the coefficients for the eight test streams. The center linear coefficients α_0 and β_0 are still not far from 0.5. Figure 5.6 plots the resulting $\mathcal{E}_1$ and LR, showing some improvements in $\mathcal{E}_1$ and more obvious improvements in LR. Both the six-point first-order and second-order interpolations show that minimizing second-order LR errors has the good side effect of also reducing higher-order LR errors. This result indicates that second-order LR approximations provide good guidance for designing reconstructions. However, the improvements of second-order interpolations are undermined when applied to the test coders, as shown in Figure 5.7.

Table 5.3: Optimal six-point second-order interpolation coefficients for the eight test streams in Table 1.1.

	Test stream							
	1	2	3	4	5	6	7	8
α_{-1}	-0.0441	0.0491	0.0161	-0.2125	-0.0524	-0.0389	0.0265	0.0587
α_0	0.5094	0.4184	0.4194	0.5647	0.5289	0.5590	0.4867	0.4130
α_1	0.0438	0.0266	0.0781	0.0822	-0.0359	0.0063	-0.0058	0.0167
β_{-1}	0.0855	-0.0007	0.0109	0.2185	0.1387	0.0900	0.0576	-0.0933
β_0	0.4664	0.5328	0.5066	0.4986	0.3921	0.4300	0.4174	0.6728
β_1	-0.0478	-0.0177	-0.0279	-0.1403	0.0451	-0.0317	0.0314	-0.0628
$\gamma^0_{-1,-1}$	-0.4096	-0.1610	-1.0075	-0.6604	0.2327	-1.0116	-0.6646	-0.1853
$\gamma^0_{-1,0}$	4.0742	2.3605	5.9233	2.5066	2.9956	3.6147	1.6231	3.3036
$\gamma^0_{-1,1}$	0.2326	0.0979	0.8842	0.3155	-1.8789	0.5975	0.9296	-0.0450
$\gamma^0_{0,0}$	-1.4108	-1.3219	-2.2537	-0.6511	-3.4070	-1.2737	-0.3783	-0.4836
$\gamma^0_{0,1}$	-1.2868	0.2413	0.5219	-1.2907	2.8555	0.0174	0.3287	-1.3148
$\gamma^0_{1,1}$	-1.2504	-1.3083	-3.5003	-1.2622	-1.0153	-2.1899	-1.6058	-1.1913
$\eta^0_{-1,-1}$	0.3500	0.2402	-0.0046	1.1281	0.5161	0.6067	-0.0797	-0.3983
$\eta^0_{-1,0}$	2.5311	0.9224	0.9273	0.5050	0.9612	0.3960	1.5969	3.3455
$\eta^0_{-1,1}$	0.3940	1.3942	4.3338	-0.7057	0.2286	1.2385	1.2618	0.9947
$\eta^0_{0,0}$	-0.3273	0.1203	1.0704	0.5570	-0.5811	1.0539	-0.8627	-0.1697
$\eta^0_{0,1}$	-2.4009	-1.8035	-3.3181	-2.6927	-0.4455	-3.0464	-0.8592	-2.9642
$\eta^0_{1,1}$	-0.5929	-0.9130	-2.5057	0.1034	-0.8033	-0.4924	-0.7672	-0.7153
$\zeta^0_{-1,-1}$	0.4583	0.4390	1.4395	-0.0283	0.7429	1.5030	1.9861	0.4002
$\zeta^0_{-1,0}$	-2.8916	-1.3562	-2.6726	-1.3510	-4.2022	-2.2966	-2.5376	-2.1809
$\zeta^0_{-1,1}$	-0.8878	-1.2073	-3.4021	0.1955	2.1905	-1.2409	-0.7158	-1.0557
$\zeta^0_{0,-1}$	-3.7134	-2.4023	-4.4627	-1.9696	-1.1407	-3.0370	-1.9390	-5.0540
$\zeta^0_{0,0}$	1.5887	1.2139	0.9087	0.5990	4.1959	0.8784	-0.0270	2.4696
$\zeta^0_{0,1}$	2.0952	1.2665	1.7726	1.3850	-2.0549	0.9351	0.6197	1.5632
$\zeta^0_{1,-1}$	-0.4237	-0.7754	-2.2714	-0.2198	-1.7900	-1.1853	-2.4735	1.0151
$\zeta^0_{1,0}$	1.9315	0.7872	1.9019	1.8109	0.5691	2.0225	3.2615	-0.2687
$\zeta^0_{1,1}$	1.9361	2.1646	5.7207	1.7234	1.8472	2.9229	1.2792	2.9242



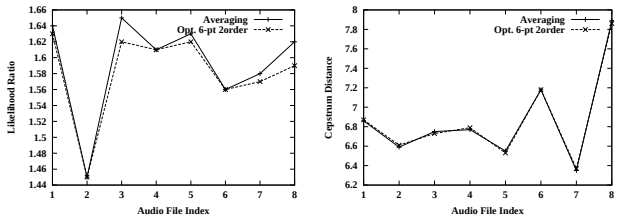
a) $\mathcal{E}1$ b) LR

Figure 5.6: A comparison of optimal six-point second-order interpolations and averaging in terms of $\mathcal{E}_1$ and its corresponding LR evaluated on the LSP vectors extracted from the eight test streams.

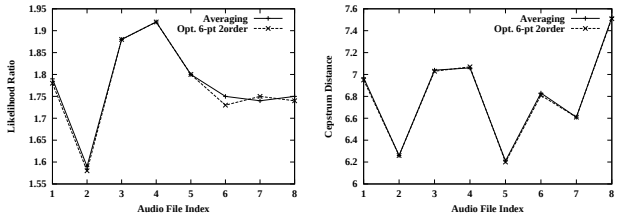
In summary, we have developed, in this section, the relationship between spectral distortion, LR, and LSP reconstruction errors. We have used the second-order approximation of LR as a guide in designing LSP interpolation methods. Experimental results show that the average interpolation method is good in the sense that it gives good reconstruction quality, is near-optimal, and performs well regardless of input streams and test coders. Additionally, experimental results demonstrate that second-order Taylor expansions of LR are good approximation to use.

5.2 Improving Excitation Quality

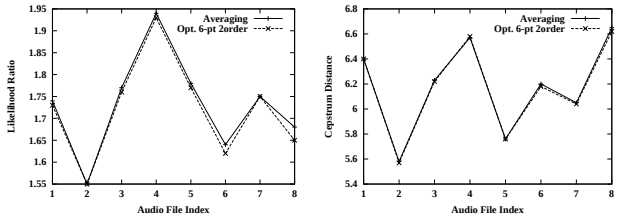
The excitation quality affects the overall transmission quality no matter whether loss happens or not. As we have to lengthen the segment size in extracting excitation information in order to avoid consuming additional bandwidth, the larger segments cause quality degradation in all loss scenarios. The most obvious degradation happens when



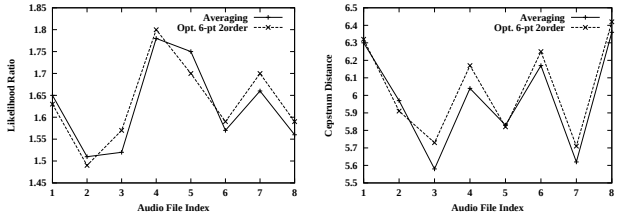
a) FS CELP



b) ITU G.723.1 ACELP



c) ITU G.723.1 MP-MLQ



d) ITU G.723.1 FS MELP

Figure 5.7: A comparison of optimal six-point second-order interpolations and averaging in terms of LR and CD for two-way MDC when only one description is received.

all the descriptions are received. In the following, we explore a new method to enhance the quality of excitation generation without increasing bandwidth.

5.2.1 Identifying the causes of degradation

To identify where excitation degrades the most, we compare coding noise $d(n) = s(n) - \hat{s}(n)$ in Figure 4.5 for both SDC and MDC. The following discussion refers to FS CELP.

For two-way MDC, the subframe size for excitation computation is 120 samples. The grouping of coding noises is also subframe-by-subframe. For each subframe of coding noises, we compute the energy-density spectrum $|D(jw)|^2$ of $d(n)$. Then we average the $|D(jw)|^2$ across all subframes for SDC and two-way MDC of the eight test streams from Table 1.1 in Figure 5.8. Also, Figure 5.8 plots the corresponding averaged $|S(jw)|^2$, the energy-density spectrum of $s(n)$, as a reference. The graph shows that speech energy concentrates mostly in the low-frequency part, and coding errors of both SDC and MDC are, hence, larger in this region. Further, there are significant differences in the energy-density spectra between SDC and MDC.

Because this uniform treatment of the whole spectrum does not give us much information, we propose to categorize noises by the frequency regions in which they occur. To have a meaningful classification, we first investigate the features of a short-time speech spectrum. Figure 5.9 plots the spectrum of a speech frame of 240 samples. Across the spectrum, there are peaks and valleys. Speech perception principles tell us that the noise-masking threshold is higher in the regions of spectral peaks than in the regions of

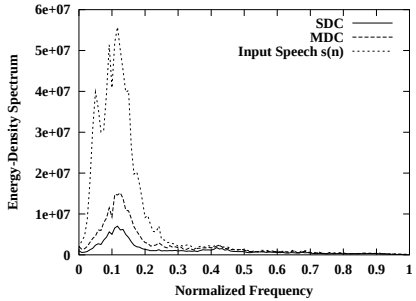


Figure 5.8: Average energy-density spectra of SDC and two-way MDC coding noises ($d(n) = s(n) - \hat{s}(n)$) for FS CELP.

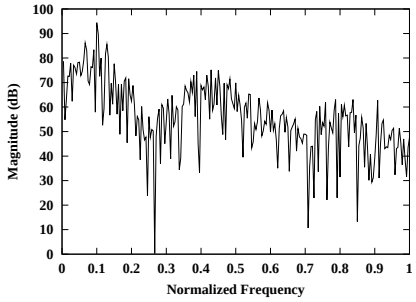


Figure 5.9: Spectrum of a typical voiced speech frame.

spectral valleys [80]. Speech perception also tells us that formants have greater perceptual importance than valleys, which suggests that the formant-region noise should not be too high. The above two points lead us to conclude that balancing noises in regions of peaks and valleys is essential for achieving good quality. Therefore, we decide to divide up the spectrum according to peaks and valleys.

Peaks generally correspond to formants, and can be singled out these regions by first finding formant frequencies. The formants are located by finding the roots of linear-prediction filters, because they model the envelopes of speech spectra [81]. For each subframe, the linear-prediction analysis filters can be expressed by their roots as:

$$\begin{aligned} A(z) &= \prod_{k=1}^{10}(1 - z_k z^{-1}) \\ &= \prod_{k=1}^{10}(1 - \rho_k e^{jw_k} z^{-1}), \end{aligned} \tag{5.55}$$

where z_k denotes the k^{th} complex root, and ρ_k and w_k denote the corresponding magnitude and angle of the root, respectively. All roots are inside the unit circle, with complex roots in conjugate pairs. In a rough sense, roots satisfy the following two conditions corresponding to formants: a) they are not on the real axis; and b) their bandwidths are less than 400 Hz [82]. The bandwidth of a root is defined as:

$$b_k = \frac{-\log \rho_k}{\pi} F_s, \tag{5.56}$$

where F_s is the sampling frequency, normally 8000 Hz. For those roots that satisfy the above two conditions, the formant frequency f_k is defined as:

$$f_k = \frac{|w_k|}{2\pi} F_s. \tag{5.57}$$

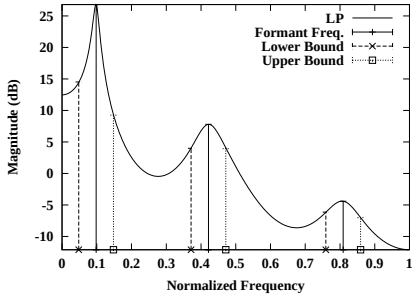


Figure 5.10: Example of formant region classification.

For convenience, denote I to be the set of all indices k corresponding to formants. After finding formants, we define heuristically the formant region Ω_k to be from $f_k - 200$ to f_k+200 for all k in I. An example classification of formant regions is shown in Figure 5.10. Here, the solid curve represents the frequency response of the linear-prediction filter; the solid vertical lines mark the positions of normalized formant frequencies; and the dashed and dotted lines mark the boundaries of formant regions.

After obtaining the formant regions, we group the energy densities by their frequency locations in each subframe and average over all test subframes. Let E_{R_F} represent the average energy inside formant regions, and let $E_{\overline{R_F}}$ represent the average energy outside formant regions. The computation is performed across all tested subframes:

$$\begin{aligned} E_{R_F} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{f\in\Omega_k^n, k\in I^n}|D_n(j2\pi f)|^2 \\ E_{\overline{R_F}} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{f\notin\Omega_k^n, k\in I^n}|D_n(j2\pi f)|^2, \end{aligned} \tag{5.58}$$

where n is the subframe index and N denotes the number of subframes.

Table 5.4: A comparison of noise energies inside and outside formant regions for SDC and two-way MDC of FS CELP.

	E_{R_F}	$E_{\overline{R_F}}$
SDC	1.1591e+8	2.3976e+7
Two-way MDC	2.3049e+8	4.4340e+7
Ratio	1.99	1.85

We are now ready to conduct a finer comparison between SDC- and MDC-coding noises. Table 5.4 lists E_{R_F} and $E_{\overline{R_F}}$ of SDC and two-way MDC for FS CELP, respectively. Here, we compute subframe coding noises of SDC in the same way as for MDC. For ease of comparison, we also compute the ratio between E_{R_F} (*resp.* $E_{\overline{R_F}}$) of two-way MDC and E_{R_F} (*resp.* $E_{\overline{R_F}}$) of SDC. For both SDC and two-way MDC, the fact that E_{R_F} is far greater than $E_{\overline{R_F}}$ reveals that the coding algorithm codes excitations more finely outside formant regions. This is a direct consequence of the speech perception principle. For two-way MDC, the third row in Table 5.4 shows a much higher increase of E_{R_F} than $E_{\overline{R_F}}$. This may signal an over-emphasis of noises outside formant regions in the MDC scheme. Under a fixed excitation bandwidth constraint, this over-emphasis means that more excitations can be extracted from outside formant regions and, thus, less from inside formant regions, leading to higher distortions inside formant regions. Therefore, in order to improve MDC quality, we can adjust the allocation of noise inside and outside formant regions.

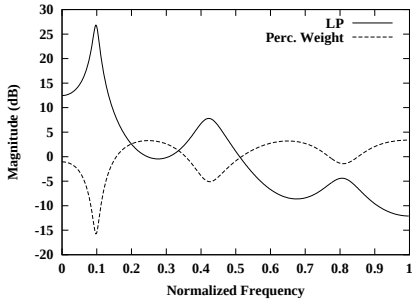


Figure 5.11: A perceptual-weighting filter example for FS CELP.

5.2.2 Adjustment of coding noise allocation by perceptual weighting

Foremost, we need to understand how noise allocation is achieved in LP coders. As shown in Figure 4.5, analysis-by-synthesis LP coders select excitations by minimizing their perceptually weighted mean-square errors, or noises, between $s(n)$ and $\hat{s}(n)$. This perceptual weighting is done through a filter that is responsible for shaping noises differently depending on whether they belong to formant regions or not.

Specifically for FS CELP, the perceptual-weighting filter $W(z)$ is defined as:

$$W(z) = \frac{A(z)}{A(z/\gamma)}, \tag{5.59}$$

where γ is fixed as 0.8, and $A(z)$ is the LP analysis filter. Figure 5.11 illustrates a perceptual-weighting filter, where the solid curve denotes the frequency response of the linear predictor $\frac{1}{A(z)}$, and the dashed curve denotes the frequency response of its corresponding perceptual-weighting filter $W(z)$. It is obvious that the peaks on the linear-predictor curve correspond to the valleys on the perceptual-weighting curve. By looking at the expression of the perceptual-weighting filter, we can see that a zero in the numerator is a zero in the denominator with a smaller radius ρ, leading to the combined effect of a spectral valley at the frequency of the peak position of $\frac{1}{A(z)}$. In this way, formant-region noises are de-emphasized. Parameter γ in (5.59) controls the degree of de-emphasis. As γ gets closer to 1, the concaves will be shallower, and the filter will be flatter. Consequently, formant-region noises are treated relatively more importantly. On the contrary, if γ gets further away from 1, formant-region noises will be suppressed more, and excitations outside formant regions will be more finely coded.

Table 5.5: A comparison of noise energies inside and outside formant regions for SDC and two-way MDC with different perceptual-weighting filters in FS CELP. The ratio column gives the ratio between E_{R_F} (*resp.* $E_{\overline{R_F}}$) of MDC and that of SDC.

	E_{R_F}		$E_{\overline{R_F}}$	
	Magnitude	Ratio	Magnitude	Ratio
SDC	1.1591e+8	1	2.3976e+7	1
Two-way MDC ($\gamma = 0.8$)	2.3049e+8	1.99	4.4340e+7	1.85
Two-way MDC ($\gamma = 0.82$)	2.2368e+8	1.93	4.3659e+7	1.82
Two-way MDC ($\gamma = 0.84$)	2.1501e+8	1.85	4.4372e+7	1.85
Two-way MDC ($\gamma = 0.86$)	2.1810e+8	1.88	4.5733e+7	1.91
Two-way MDC ($\gamma = 0.88$)	2.0096e+8	1.73	4.6161e+7	1.92
Two-way MDC ($\gamma = 0.9$)	2.0098e+8	1.73	4.7287e+7	1.97

To choose a suitable γ for two-way MDC, we compute E_{R_F} and $E_{\overline{R_F}}$ for different γ and list them in Table 5.5. The E_{R_F} and $E_{\overline{R_F}}$ for SDC are also shown as a reference. Here, we only increase γ because we want to decrease noises in formant regions. Table 5.5 shows that, as γ increases, E_{R_F} decreases, while $E_{\overline{R_F}}$ increases. At $\gamma = 0.84$, E_{R_F} and

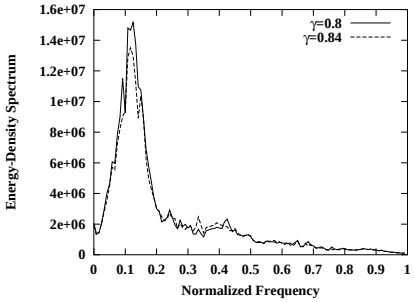


Figure 5.12: A comparison of energy-density spectra of two-way MDC coding noises before and after changing γ.

$E_{\overline{R_F}}$ of MDC have similar increases compared to those of SDC. Therefore, we change γ to 0.84 to improve the two-way MDC's quality for FS CELP. Figure 5.12 plots the energy-density spectra before and after changing γ. The plot demonstrates that changing γ also reduces the overall coding noises.

Likewise, we examine perceptual weighting for the four-way MDC in FS CELP. The first and second rows in Table 5.6 compare E_{R_F} and $E_{\overline{R_F}}$ between SDC and four-way MDC, which show an increase of formant-region noises of over 80% more than noises outside formant regions. Hence, the original noise shaping filter is not suitable for four-way MDC either. As in the two-way MDC case, we enumerate E_{R_F} and $E_{\overline{R_F}}$ for different γ in Table 5.6. In particular, when γ equals 0.92, E_{R_F} has a little higher percentage of increase than $E_{\overline{R_F}}$. Then, as γ increases to 0.94, $E_{\overline{R_F}}$ has a higher percentage of increase. Consequently, unlike the two-way MDC case, we will change γ to 0.94 in the improved four-way MDC. Because the excitation bandwidth for four-way MDC is very limited,

Table 5.6: A comparison of noise energies inside and outside formant regions for SDC and four-way MDC with different perceptual-weighting filters of FS CELP. The ratio column gives the ratio between E_{R_F} (*resp.* $E_{\overline{R_F}}$) of MDC and that of SDC.

	E_{R_F}		$E_{\overline{R_F}}$	
	Magnitude	Ratio	Magnitude	Ratio
SDC	2.3507e+8	1	4.4775e+7	1
Four-way MDC ($\gamma = 0.8$)	8.6771e+8	3.69	1.2685e+8	2.83
Four-way MDC ($\gamma = 0.82$)	8.5068e+8	3.62	1.2911e+8	2.88
Four-way MDC ($\gamma = 0.84$)	8.5003e+8	3.62	1.3041e+8	2.91
Four-way MDC ($\gamma = 0.86$)	8.2722e+8	3.52	1.3157e+8	2.94
Four-way MDC ($\gamma = 0.88$)	8.0007e+8	3.40	1.3387e+8	2.99
Four-way MDC ($\gamma = 0.9$)	7.8380e+8	3.33	1.3819e+8	3.09
Four-way MDC ($\gamma = 0.92$)	7.5565e+8	3.21	1.4193e+8	3.17
Four-way MDC ($\gamma = 0.94$)	7.3941e+8	3.14	1.4591e+8	3.26

noises both inside and outside formant regions cannot be pushed below noise-masking thresholds. Chen and Gersho [83] pointed out that it is a good strategy to sacrifice valley regions and preserve the formant regions when noises cannot be masked, because in speech perception the formants of speech are perceptually much more important than spectral valley regions.

Next, we apply the same reasoning to improve the excitation quality of ITU G.723.1. In this standard, the perceptual-weighting filter is defined as

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \tag{5.60}$$

where γ_1 is 0.9 and γ_2 is 0.5. This filter is common to ACELP and MP-MLQ. Here, the difference between γ_1 and γ_2 decides how much suppression we can get for the formant-region noises. In the following, we adjust γ_2 to find better noise shaping for two-way and

four-way MDC. Although ACELP and MP-MLQ share the same perceptual-weighting filter, their excitation-generation modules are different; hence, we discuss them separately.

First, for ITU G.723.1 ACELP, Table 5.7 summarizes E_{R_F} and $E_{\overline{R_F}}$ for different γ_2 under two-way and four-way MDC. The results show that the original noise shaping is not suitable for MDC as well (see the second row in Table 5.7a and 5.7b). To get better excitation quality, we increase γ_2. For both two-way and four-way MDC, as γ_2 increases, noise energy inside the formant regions first decreases and then increases. Therefore, we choose a different γ_2 corresponding to the lowest formant-region distortion for each case. Specifically, the improved MDC scheme uses 0.65 for two-way and 0.75 for four-way MDC in the following experiments.

Second, we repeat the analysis for ITU G.723.1 MP-MLQ. Table 5.8 lists the resulting E_{R_F} and $E_{\overline{R_F}}$ for various γ_2 in two-way and four-way MDC. The original MDC has the same problem as it suppresses formant-region noises too much (see the second row in Tables 5.8a and 5.8b). Like ACELP, two-way MP-MLQ has the behavior that its formant-region distortion first decreases and then increases as γ_2 increases. The minimum distortion happens at around $\gamma_2 = 0.8$. For four-way MP-MLQ, its formant-region distortion continues to decrease until γ_2 equals 0.9, which corresponds to no perceptual weighting at all. Although in the SDC case, ACELP and MP-MLQ share the same perceptual-weighting filter, the above results show that they need to use different γ_2 for MDC schemes.

The above analysis only applies to close-looped linear-predictive coders that generate excitations by perceptually weighting their coding noises and in turn deciding code words

Table 5.7: A comparison of noise energies inside and outside formant regions for SDC and MDC with different perceptual-weighting filters in ITU G.723.1 ACELP. The ratio column gives the ratio between E_{R_F} (*resp.* $E_{\overline{R_F}}$) of MDC and that of SDC.

a) Two-way MDC

	E_{R_F}		$E_{\overline{R_F}}$	
	Magnitude	Ratio	Magnitude	Ratio
SDC	3.3567e+7	1	1.3826e+7	1
Two-way MDC ($\gamma = 0.5$)	8.5002e+7	2.53	2.8066e+7	2.03
Two-way MDC ($\gamma = 0.55$)	8.3368e+7	2.48	2.7676e+7	2.00
Two-way MDC ($\gamma = 0.6$)	8.2505e+7	2.46	2.8244e+7	2.04
Two-way MDC ($\gamma = 0.65$)	8.0817e+7	2.41	2.7880e+7	2.02
Two-way MDC ($\gamma = 0.7$)	8.4900e+7	2.53	2.9065e+7	2.10
Two-way MDC ($\gamma = 0.75$)	8.5197e+7	2.54	2.9588e+7	2.14

b) Four-way MDC

	E_{R_F}		$E_{\overline{R_F}}$	
	Magnitude	Ratio	Magnitude	Ratio
SDC	6.7600e+7	1	2.7446e+7	1
Four-way MDC ($\gamma = 0.5$)	4.2910e+8	6.37	1.0552e+8	3.84
Four-way MDC ($\gamma = 0.55$)	4.2603e+8	6.32	1.0594e+8	3.86
Four-way MDC ($\gamma = 0.60$)	4.1271e+8	6.13	1.0243e+8	3.73
Four-way MDC ($\gamma = 0.65$)	4.1891e+8	6.22	1.0485e+8	3.82
Four-way MDC ($\gamma = 0.7$)	4.1124e+8	6.11	1.0420e+8	3.80
Four-way MDC ($\gamma = 0.75$)	4.0469e+8	6.00	1.0723e+8	3.91
Four-way MDC ($\gamma = 0.8$)	4.0536e+8	6.02	1.1019e+8	4.01
Four-way MDC ($\gamma = 0.85$)	4.0858e+8	6.07	1.1402e+8	4.15

Table 5.8: A comparison of noise energies inside and outside formant regions for SDC and MDC with different perceptual-weighting filters in ITU G.723.1 MP-MLQ. The ratio column gives the ratio between E_{R_F} (*resp.* $E_{\overline{R_F}}$) of MDC and that of SDC.

a) Two-way MDC

	E_{R_F}		$E_{\overline{R_F}}$	
	Magnitude	Ratio	Magnitude	Ratio
SDC	2.2409e+7	1	1.0297e+7	1
Two-way MDC ($\gamma = 0.5$)	5.8606e+7	2.61	2.1563e+7	2.09
Two-way MDC ($\gamma = 0.55$)	5.6044e+7	2.50	2.1345e+7	2.07
Two-way MDC ($\gamma = 0.6$)	5.5152e+7	2.46	2.0768e+7	2.02
Two-way MDC ($\gamma = 0.65$)	5.3583e+7	2.39	2.1014e+7	2.04
Two-way MDC ($\gamma = 0.7$)	5.2732e+7	2.35	2.1279e+7	2.07
Two-way MDC ($\gamma = 0.75$)	5.2065e+7	2.32	2.1578e+7	2.10
Two-way MDC ($\gamma = 0.8$)	5.1579e+7	2.30	2.2258e+7	2.16
Two-way MDC ($\gamma = 0.85$)	5.2978e+7	2.36	2.3484e+7	2.28
Two-way MDC ($\gamma = 0.9$)	5.2668e+7	2.35	2.6058e+7	2.53

b) Four-way MDC

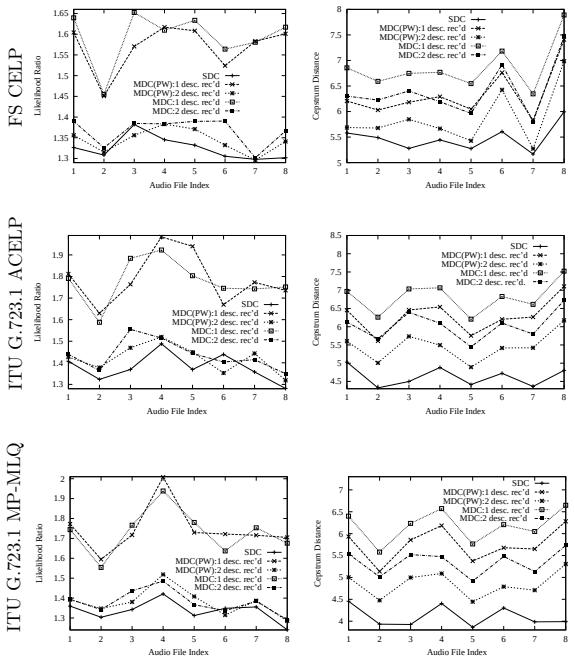
	E_{R_F}		$E_{\overline{R_F}}$	
	Magnitude	Ratio	Magnitude	Ratio
SDC	4.5095e+7	1	2.0331e+7	1
Four-way MDC ($\gamma = 0.5$)	2.9425e+8	6.53	7.6603e+7	3.77
Four-way MDC ($\gamma = 0.55$)	2.8652e+8	6.35	7.5818e+7	3.73
Four-way MDC ($\gamma = 0.60$)	2.8406e+8	6.30	7.3009e+7	3.59
Four-way MDC ($\gamma = 0.65$)	2.7839e+8	6.17	7.3381e+7	3.61
Four-way MDC ($\gamma = 0.7$)	2.7153e+8	6.02	7.3990e+7	3.64
Four-way MDC ($\gamma = 0.75$)	2.6833e+8	5.95	7.5144e+7	3.70
Four-way MDC ($\gamma = 0.8$)	2.6499e+8	5.88	7.7354e+7	3.80
Four-way MDC ($\gamma = 0.85$)	2.6401e+8	5.85	8.1866e+7	4.03
Four-way MDC ($\gamma = 0.9$)	2.6278e+8	5.83	9.0080e+7	4.43

for excitations. It does not apply to coders that do not incorporate this mechanism in coding excitations, such as FS MELP.

5.2.3 Experimental results

In this section, we test the performances of MDC by adopting the changes to perceptual-weighting filters discussed previously. The tests are performed on the eight test streams in Table 1.1 and under both synthetic and Internet losses. The improvements over the original MDC scheme are then evaluated.

We start with tests under synthetic loss scenarios. Figure 5.13 compares the qualities measured in terms of LR and CD for SDC with no loss, the original two-way MDC scheme, and the two-way MDC scheme with the improved perceptual-weighting filter. There are two loss scenarios for two-way MDC: a) both descriptions are received, and b) one description is received. For the three coders, FS CELP, ITU G.723.1 ACELP, and ITU G.723.1 MP-MLQ, both MDC schemes have similar distortions in terms of LR. When there is no loss, the LR values of both MDC schemes are close to SDC's. Unlike for LR, obvious improvements in measured distortions can be seen in terms of CD. The different behaviors of LR and CD conform to our previous observation that LR reflects linear-prediction quality more, whereas CD reflects excitation quality more. After modifying the noise shaping filter, the distortions in terms of CD decrease about 50% when both descriptions are received. To see the improvements more quantitatively for both loss scenarios, Table 5.9 shows the average improvements in terms of CD of the modified MDC over the original MDC. There are consistent and considerable improvements of



a) LR b) CD

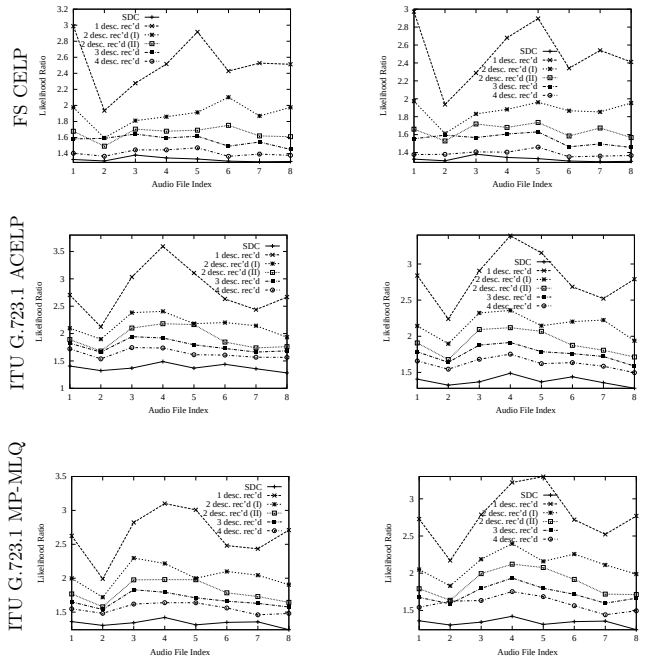
Figure 5.13: Quality comparisons in terms of LR and CD among SDC with no loss, two-way MDC under two loss scenarios: one description received or both description received, and two-way MDC with improved perceptual-weighting filter under the two loss scenarios. Results are for FS CELP, ITU G.723.1 ACELP, and ITU G.723.1 MP-MLQ.

Table 5.9: Average improvements of two-way MDC with improved perceptual-weighting filter versus the original MDC in terms of CD under two loss scenarios: one description received or both description received.

	2 desc. rec'd	1 desc. rec'd
FS CELP	0.53 dB	0.52 dB
ACELP	0.58 dB	0.51 dB
MP-MLQ	0.50 dB	0.42 dB

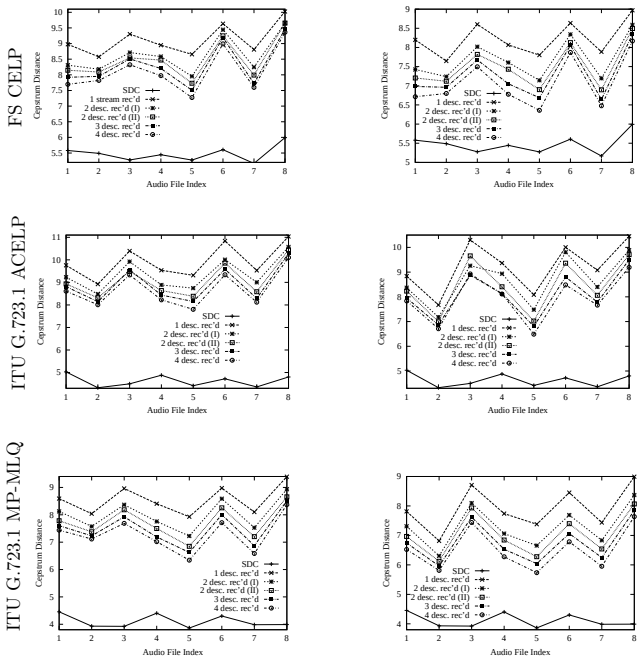
about 0.5 dB for all three coders. For example, for FS CELP in case of no loss, the average degradation after the improvements is only 0.40 dB when compared to SDC with no loss.

Next, we examine the four-way MDC with the improved perceptual-weighting filter for FS CELP, ITU G.723.1 ACELP, and ITU G.723.1 MP-MLQ. After adjusting the perceptual-weighting filter as computed in the last section, we gather the results for the improved MDC and compare them with the original MDC. Figures 5.14 and 5.15 show the comparisons in terms of LR and CD, respectively. In both figures, the graphs on the left show the distortions of the original MDC scheme, whereas the graphs on the right show the distortions of the improved MDC scheme. Similarly to the two-way MDC case, the two MDC algorithms do not differ much in terms of LR. In contrast, distortions measured in terms of CD demonstrate the apparent advantage of the improved MDC scheme. Further, Table 5.10 summarizes the average improvements under the five loss scenarios. The improvements validate that modifying perceptual weighting is a good strategy to enhance MDC's quality.



a) MDC b) MDC with improved PWF

Figure 5.14: Quality comparison in terms of LR among SDC with no loss, four-way MDC under five loss scenarios (left column), and four-way MDC with improved perceptual-weighting filter under five loss scenarios (right column). The five loss scenarios are: one description received, two consecutive descriptions received (I), two disjoint descriptions received (II), three descriptions received, or four descriptions received. Results are for FS CELP, ITU G.723.1 ACELP, and ITU G.723.1 MP-MLQ.



a) MDC b) MDC with improved PWF

Figure 5.15: Quality comparisons in terms of CD among SDC with no loss, four-way MDC under five loss scenarios (left column), and four-way MDC with improved perceptual-weighting filter under five loss scenarios (right column). The five loss scenarios are: one description received, two consecutive descriptions received (I), two disjoint descriptions received (II), three descriptions received, or four descriptions received. Results are for FS CELP, ITU G.723.1 ACELP, and ITU G.723.1 MP-MLQ.

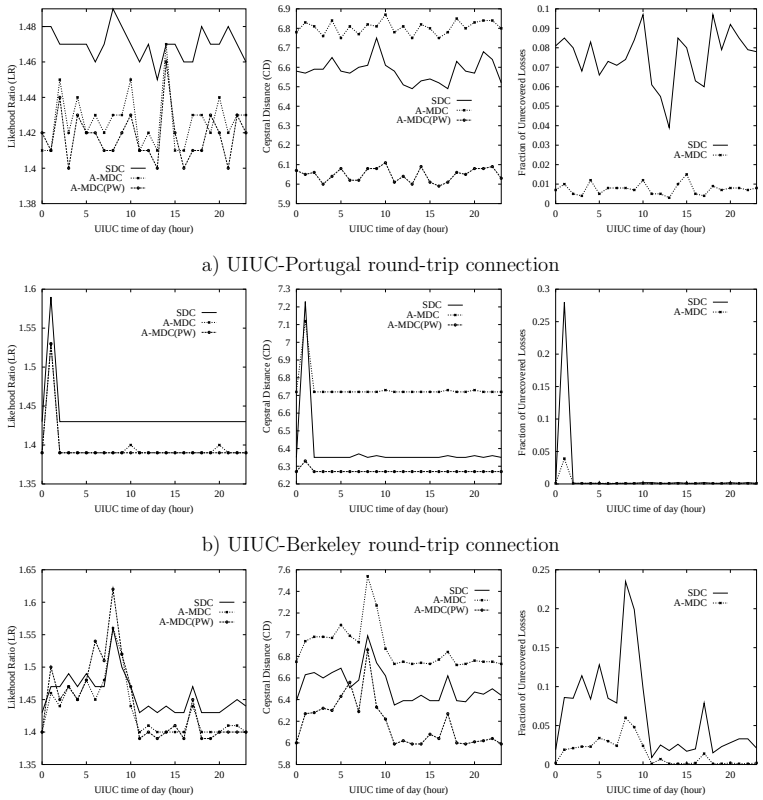
Table 5.10: Average improvements of four-way MDC with improved perceptual-weighting filter as compared to the original MDC in terms of CD under five loss scenarios.

	4 desc. rec'd	3 desc. rec'd	2 desc. rec'd (II)	2 desc. rec'd (I)	1 desc. rec'd
FS CELP	1.05	1.01	0.99	0.94	0.89
ACELP	0.76	0.82	0.65	0.69	0.69
MP-MLQ	0.76	0.74	0.71	0.72	0.63

Besides testing with synthetic losses, we also test MDC with the improved perceptual-weighting filter on real Internet traffic. The test system is configured as described in Chapter 4 using traffic traces explained in Chapter 3. Both MDC schemes are adaptive that dynamically adapt to two-way or four-way MDC depending on loss conditions, using the same adaptation mechanism described in Chapter 4. Similar to the Internet tests in Chapter 4, in addition to the distortions, we also plot the fraction of unrecovered losses for each connection in SDC and MDC in order to make fair comparisons. The two MDC schemes have the same fraction of unrecovered losses.

Figures 5.16-5.21 show our test results on SDC, the original adaptive MDC, and the improved adaptive MDC schemes for FS CELP, ITU G.723.1 ACELP, and ITU G.723.1 MP-MLQ, from top to bottom. In terms of LR, for all coders and all connections, both MDC schemes perform alike and have lower or similar distortions to SDC. We have already explained this behavior in Chapter 4.

In terms of CD, for FS CELP and all six connections, the improved MDC scheme has the lowest distortions among SDC, the original MDC, and the improved MDC. It even performs better than SDC because it reconstructs decoding states when it reconstructs

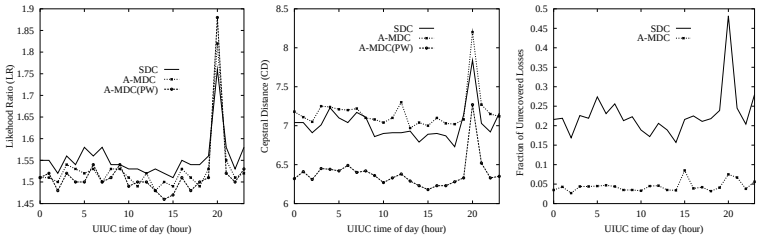


a) UIUC-Portugal round-trip connection

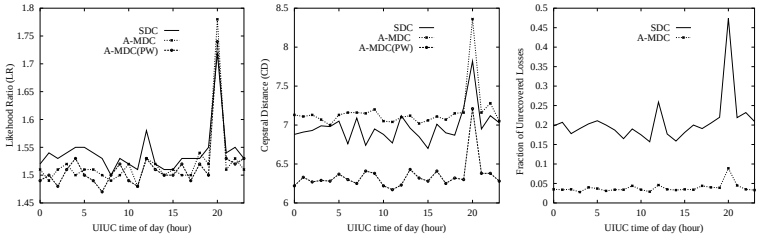
b) UIUC-Berkeley round-trip connection

c) UIUC-Egypt round-trip connection

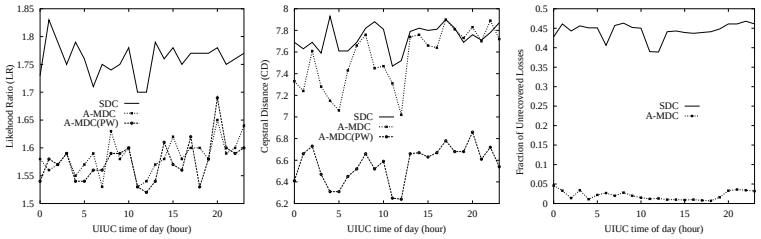
Figure 5.16: A comparison of reconstruction quality among SDC, adaptive two-way/four-way MDC, and adaptive two-way/four-way MDC with improved perceptual-weighting filter for FS CELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three low-to-medium-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-S. China round-trip connection

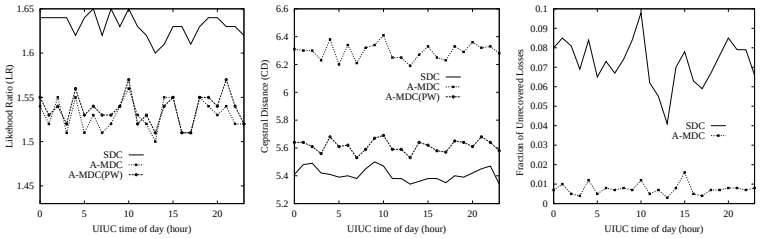


b) UIUC-W. China round-trip connection

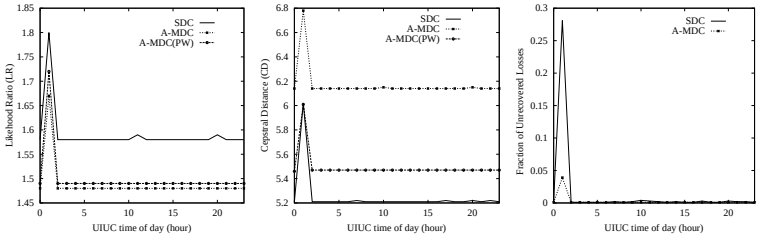


c) UIUC-Slovakia round-trip connection

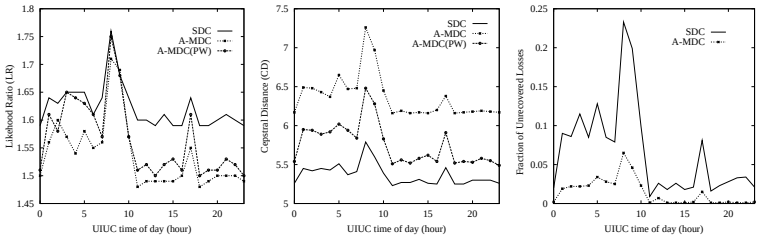
Figure 5.17: A comparison of reconstruction quality among SDC, adaptive two-way/four-way MDC, and adaptive two-way/four-way MDC with improved perceptual-weighting filter for FS CELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three medium-to-high-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-Portugal round-trip connection

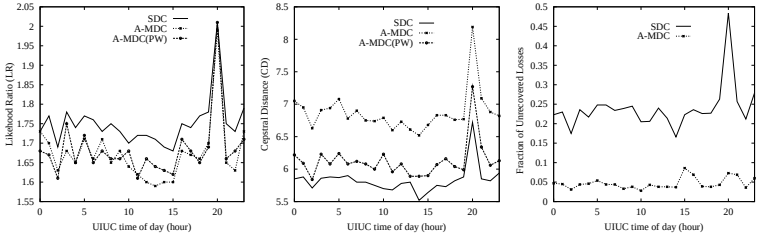


b) UIUC-Berkeley round-trip connection

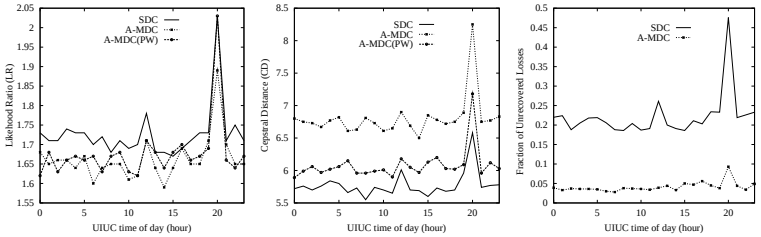


c) UIUC-Egypt round-trip connection

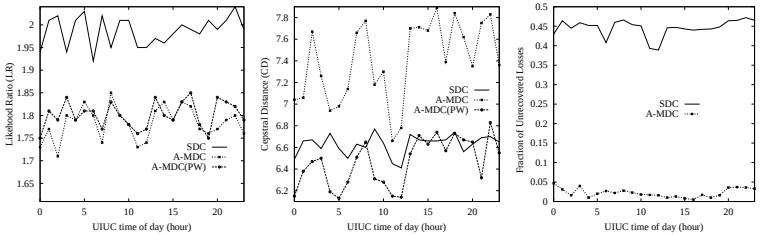
Figure 5.18: A comparison of reconstruction quality among SDC, adaptive two-way/four-way MDC, and adaptive two-way/four-way MDC with improved perceptual-weighting filter for ITU G.723.1 ACELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three low-to-medium-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-S. China round-trip connection

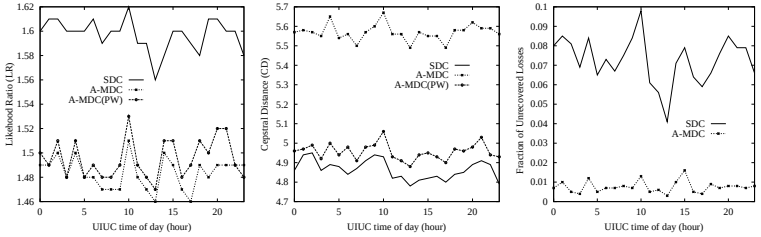


b) UIUC-W. China round-trip connection

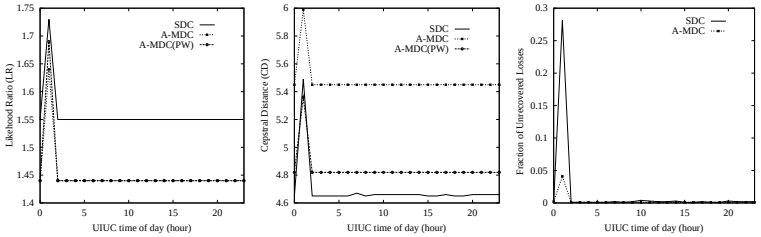


c) UIUC-Slovakia round-trip connection

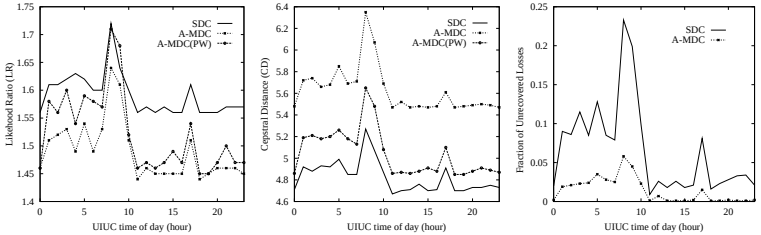
Figure 5.19: A comparison of reconstruction quality among SDC, adaptive two-way/four-way MDC, and adaptive two-way/four-way MDC with improved perceptual-weighting filter for ITU G.723.1 ACELP on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three medium-to-high-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-Portugal round-trip connection

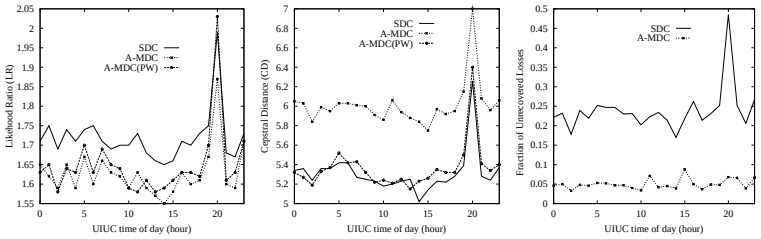


b) UIUC-Berkeley round-trip connection

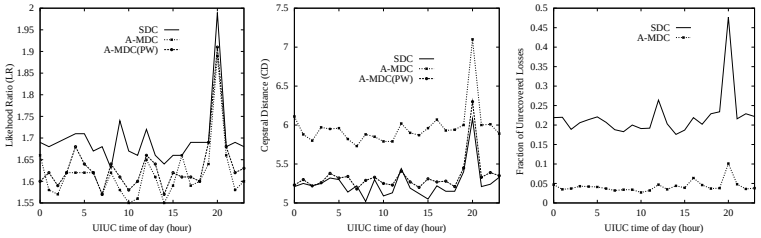


c) UIUC-Egypt round-trip connection

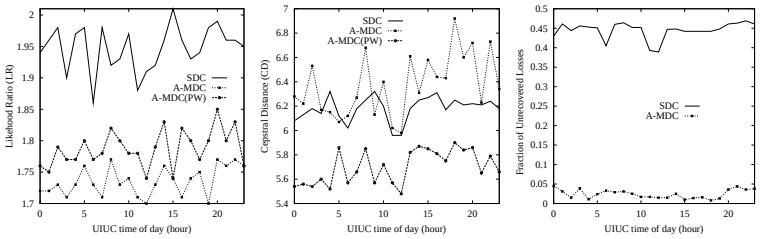
Figure 5.20: A comparison of reconstruction quality among SDC, adaptive two-way/four-way MDC, and adaptive two-way/four-way MDC with improved perceptual-weighting filter for ITU G.723.1 MP-MLQ on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three low-to-medium-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.



a) UIUC-S. China round-trip connection



b) UIUC-W. China round-trip connection



c) UIUC-Slovakia round-trip connection

Figure 5.21: A comparison of reconstruction quality among SDC, adaptive two-way/four-way MDC, and adaptive two-way/four-way MDC with improved perceptual-weighting filter for ITU G.723.1 MP-MLQ on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and three medium-to-high-loss destinations. The graphs on the right show the fraction of frames that were lost or could not be reconstructed.

lost packets, leading to better decoding quality. For ITU G.723.1 ACELP and MP-MLQ, the improved MDC has the lowest distortions among the three schemes for the UIUC-Slovakia connection. For the other five connections, the improved MDC performs significantly better than the original MDC and much closer to SDC.

5.3 Summary

In this chapter, we have focused on improving the decoding quality of LSP-based MDC, and, in addition, reconstruction quality when loss happens and when reconstructions are done through interpolating received LSP vectors in order to approximate lost ones. We have investigated various schemes to guide such interpolations by minimizing spectral distortions. We have first found the relationship between reconstruction errors of LSP vectors and the second-order LR approximation. We have further generated optimal interpolations in order to minimize second-order LR distortions. Our experimental results show that the scheme using second-order LR approximations is good and that the scheme using average interpolations is near optimal.

Decoding-quality degradations are caused by the expanded segment sizes in generating excitations under bandwidth constraints. This lengthening leads to fewer excitations to be coded per unit time interval. We have found that the original noise shaping schemes of test coders, FS CELP, ITU G.723.1 ACELP, and MP-MLQ, actually cause far larger distortions inside formant regions than outside formant regions. To alleviate this problem and improve excitation-generation quality, we have modified the perceptual-

weighting filters of the test coders in the LSP-based MDC scheme. For all coders, tests on both synthetic loss scenarios and real Internet environments have shown consistent improvements in the modified MDC scheme.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

In this research, we have designed and evaluated a novel end-to-end loss concealment scheme for reliable real-time low bit-rate coded voice transmissions over unreliable IP networks.

From our extensive traffic study of voice transmissions over the Internet, we have found that multiple-description coding is a very attractive scheme to overcome loss, as network loss happens frequently and the length of consecutive packet losses is relatively small.

Since conventional coder-independent multiple-description coding (sample interleaving) does not work well in low bit-rate coded speech, we have designed a parameter-based MDC, a coder-dependent scheme. We have proposed generating multiple descriptions systematically by a correlation analysis of coding parameters. For low bit-rate coders, FS CELP, ITU G.723.1, and FS MELP, we have developed and tested an LP-based MDC

scheme, in which a linear predictor may be represented as RF, LAR, or LSP. In this scheme, LP vectors are interleaved to different descriptions, while excitation parameters are replicated into all descriptions. This scheme was designed without increasing the transmission bandwidth. We have tested our algorithms extensively under various conditions, including both synthetic loss scenarios and real Internet losses, and have used various performance measures for fair comparisons. Based on improved performance of our LP-based MDC, we conclude that our proposed scheme is effective in concealing loss for both close-looped and open-looped low bit-rate coders. In addition, we have found that the LSP representation gives better reconstruction quality than RF and LAR.

Our scheme achieves a well-balanced trade-off between the quality of received packets and the ability to reconstruct lost ones. We have further investigated different alternatives to improve two types of MDC quality: reconstruction quality and decoding quality. To improve reconstruction quality, we have built the relationship among spectral distortion, LR, and LSP reconstruction errors, and have studied methods to improve LSP reconstruction quality by designing interpolations with the objective of minimizing the second-order LR approximation. Our experimental results have shown that the scheme based on average interpolations performs similarly to those with optimal first-order and second-order interpolations. To improve decoding quality, we have explored excitation-quality degradations. By classifying distortions in the frequency domain, we have identified the problem of inappropriate noise shaping for our proposed MDC schemes. After modifying perceptual-weighting filters in FS CELP, ITU G.723.1 ACELP, and MP-MLQ, we have made significant improvements in decoding quality.

6.2 Future Work

In this section, we propose some possible improvements and future work in this research.

First, the qualities of our MDC schemes have considerable room for improvement, especially in the case of the four-way MDC. Even after we have improved the perceptual-weighting filters for the MDC schemes, there is still a noticeable degradation in MDC's decoding quality as compared to SDC's. In addition, this modification only applies to close-looped linear-predictive coders that generate excitations by perceptually weighting their coding noises. Hence, it does not apply to FS MELP's MDC schemes. The fundamental issue here is that the excitation bandwidth in the MDC schemes is very limited. Further improvements should be investigated to extract excitations more efficiently.

Second, in this research, we have applied bandwidth constraints that restrict the bandwidth usage of our MDC schemes to be equal to that of SDC. This restriction may not always be a necessity. If we loosen this constraint, MDC can achieve better quality. To this end, it is useful to establish quality and bandwidth trade-offs for the MDC schemes and for different coders.

Third, we have fixed the transmission bandwidth in our proposed MDC schemes for a chosen speech coder. The drawback of this approach is that the transmission system will keep sending voice packets and ask for the same bandwidth even when congestion happens. Clearly, this will worsen network congestion and deprive bandwidth from other

TCP-friendly applications. Therefore, an enhancement worth considering is the design of multi-rate MDC schemes that adapt its bandwidth to be more TCP-friendly.

REFERENCES

[1] D. D. Chowdhury, *Unified IP Internet-Working*. Berlin, Germany: Springer-Verlag, 2001.

[2] U. Varshney, A. Snow, M. McGivern, and C. Howard, "Voice over IP," *Communications of the Acm*, vol. 45, pp. 89–96, January 2002.

[3] International Engineering Consortium, "Voice portal solutions: An introduction to Next-Generation Network services; the next big opportunity on the web," 2002, http://www.iec.org/online/tutorials/voice_portal/topic01.html.

[4] Z. G. Chen, "Coding and transmission of digital video on the Internet," Ph.D. dissertation, University of Illinois, Urbana-Champaign, 1997.

[5] A. Shah, S. Atungsiri, A. Kondoz, and B. Evans, "Lossy multiplexing of low bit rate speech in thin route telephony," *Electronics Letters*, vol. 32, pp. 95–97, January 1996.

[6] B. Dempsey, T. Strayer, and A. Weaver, "Adaptive error control for multimedia data transfers," in *Proc. of Int'l Workshop on Advanced Communications and Applications for High Speed Networks*, March 1992, pp. 279–289.

[7] S. Deering and R. Hinden, "Internet protocol, version 6 (IPv6) specification," Internet requests for comments 2460, December 1998, http://www.cis.ohio-state.edu/htbin/rfc/rfc2460.html.

[8] L. Mathy, C. Edwards, D. Hutchison, and L. University, "The Internet: A global telecommunications solution?" *IEEE Network*, vol. 14, pp. 46–57, July-Aug. 2000.

[9] "Implementation report: Internet protocol, version 6 (IPv6) specification," Internet Engineering Task Force, 1998, http://www.ietf.org/IESG/Implementations/ipv6-implementations.txt.

[10] R. Braden, D. Clark, and S. Shenker, "Integrated services in the Internet architecture: An overview," Internet requests for comments 1633, June 1994, http://www.cis.ohio-state.edu/htbin/rfc/rfc1633.html.

[11] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," Internet requests for comments 2475, December 1998, http://www.cis.ohio-state.edu/htbin/rfc/rfc2475.html.

[12] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation protocol (RSVP) – Version 1 functional specification," Internet requests for comments 2205, September 1997, http://www.cis.ohio-state.edu/htbin/rfc/rfc2205.html.

[13] J. Khan, "Introduction to 3G/4G wireless network architectures," in *IEEE International Symposium on Circuits and Systems: Tutorial Guide*, 2001, pp. 7.1.1–7.1.13.

[14] International Telecommunication Union, "ITU-T H.323 — Packet-based multimedia communications systems," 2000.

[15] International Telecommunication Union, "ITU-T H.225.0 — Call signalling protocol and media stream packetization for packet-based multimedia communication systems," 2001.

[16] International Telecommunication Union, "ITU-T H.245 — Control protocol for multimedia communications," 2000.

[17] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real time applications," Internet requests for comments 1889, January 1996, http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1889.txt.

[18] J. Erkelens and P. Broersen, "Analysis of spectral interpolation with weighting dependent on frame energy," in *1989 Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, 1994, pp. 481–484.

[19] B. Atal, R. Cox, and P. Kroon, "Spectral quantization and interpolation for CELP coders," in *1989 Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, 1989, pp. 69–72.

[20] J. Erkelens and P. Broersen, "LPC interpolation by approximation of the sample autocorrelation function," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 569–573, November 1998.

[21] T. Islam, "Interpolation of linear prediction coefficients for speech coding," M.S. Thesis, McGill University, Montreal, Canada, April 2000.

[22] K. Paliwal, "Interpolation properties of linear prediction parametric representations," in *4th European Conference on Speech Communication and Technology*, Madrid, 1995, pp. 1029–1032.

[23] H. Choi, W. Wong, B. Cheetham, and C. Goodyear, "Interpolation of spectral information for low bit rate speech coding," in *4th European Conference on Speech Communication and Technology*, Madrid, 1995, pp. 1033–1036.

[24] J. Suzuki and M. Taka, "Missing packet recovery techniques for low-bit-rate coded speech," *IEEE Journal on Selected Areas in Communications*, vol. 7, pp. 707–717, June 1989.

[25] 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, April 1985, pp. 227–234.

[26] O. J. Wasem, D. J. Goodman, C. A. Dvordak, and H. G. Page, "The effect of waveform substitution on the quality of PCM packet communications," *IEEE Trans. on Acoutics, Speech, and Signal Processing*, vol. 36, pp. 342–348, March 1988.

[27] R. A. Valenzuela and C. N. Animalu, "A new voice-packet reconstruction technique," in *1989 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, vol. 2, May 1989, pp. 1334–1336.

[28] J. Tang, "Evaluation of double sided periodic substitution (DSPS) method for recovering missing speech in packet voice communications," in *Proc. Tenth Annual Int'l Phoenix Conf. on Computers and Communications*, March 1991, pp. 454–458.

[29] H. Sanneck, A. Stenger, K. B. Younes, and B. Girod, "A new technique for audio packet loss concealment," in *Global Telecommunications Conf.*, November 1996, pp. 48–52.

[30] K. Cluver and P. Noll, "Reconstruction of missing speech frames using sub-band excitation," in *Proc. IEEE-SP Int'l Symposium on Time-Frequency and Time-Scale Analysis, 1996*, June 1996, pp. 277–280.

[31] V. Hardman, M. A. Sasse, M. Handley, and A. Watson, "Reliable audio for use over the Internet," in *Int'l Networking Conf.*, June 1995, pp. 171–178.

[32] M. Yong, "Study of voice packet reconstruction methods applied to CELP speech coding," in *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, vol. 2, March 1992, pp. 125–128.

[33] M. M. Lara-Barron and G. B. Lockhart, "Packet-based embedded encoding for transmission of low-bit-rate-encoded speech in packet networks," *IEE Proc.-I*, vol. 139, pp. 482–487, October 1992.

[34] G. B. Lockhart and M. M. Lara-Barron, "Implementation of packet-based encoding schemes for speech transmission," in *Proc. Sixth Int'l Conf. on Digital Processing of Signals in Communications*, 1991, pp. 326–330.

[35] Z. G. Chen, S. M. Tan, R. H. Campbell, and Y. Li, "Real time video and audio in the World Wide Web," *World Wide Web Journal*, vol. 1, January 1996.

[36] B. J. Dempsey, J. Liebeherr, and A. C. Weaver, "A new error control scheme for packetized voice over high-speed local area networks," in *Proc. 18th Conf. on Local Computer Networks*, 1993, pp. 91–100.

[37] B. J. Dempsey and Y. Zhang, "Destination buffering for low-bandwidth audio transmission using redundancy-based error control," in *Proc. of 21st IEEE Local Computer Networks Conf.*, October 1996, pp. 345–355.

[38] S. Chiou and V. Li, "An optimal two-copy routing scheme in a communication network," in *Proc. of the Seventh Annual Joint Conf. of the IEEE Computer and Communications Societies. Networks: Evolution or Revolution, INFOCOM '88*, 1988, pp. 288–297.

[39] 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, pp. 18–27, January-February 1998.

[40] N. Shacham, "Packet recovery and error correction in high-speed wide-area networks," in *1989 IEEE Military Communications Conf.*, vol. 2, October 1989, pp. 551–557.

[41] 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.

[42] J. Bolot and A. Garcia, "Control mechanisms for packet audio in the Internet," in *Proc. IEEE Infocom'96*, San Francisco, CA, April 1996, pp. 232–239.

[43] J. Bolot and A. Garcia, "The case for FEC-based error control for packet audio in the Internet," submitted to ACM Multimedia Sys., 1997.

[44] L. DaSilva, D. Petr, and V. Frost, "A class-oriented replacement technique for lost speech packets," in *Proc. of the Eighth Annual Joint Conf. of the IEEE Computer and Communications Societies. Technology: Emerging or Converging, INFOCOM '89*, vol. 3, 1989, pp. 1098–1105.

[45] A. Choi and A. Constantinides, "Effects of packet loss on 3 toll quality speech coders," in *IEE National Conf. on Telecommunications*, 1989, pp. 380–385.

[46] N. Erdol, C. Castelluccia, and A. Zilouchian, "Recovery of missing speech packets using the short-time energy and zero-crossing measurements," *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 295–303, July 1993.

[47] H. Sanneck, "Concealment of lost speech packets using adaptive packetization," in *Proc. IEEE Int'l Conf. on Multimedia Computing and Systems*, 1998, pp. 140–149.

[48] 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 Trans. on Communications*, vol. 29, pp. 101–110, February 1981.

[49] M. Yuito and N. Matsuo, "A new sample-interpolation method for recovering missing speech samples in packet voice communications," in *1989 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, vol. 1, May 1989, pp. 381–384.

[50] Y.-L. Chen and B.-S. Chen, "Model-based multirate representation of speech signals and its application to recovery of missing speech packets," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 220–231, May 1997.

[51] R. V. Cox, W. B. Kleijn, and P. Kroon, "Robust CELP coders for noisy backgrounds and noisy channels," in *1989 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, vol. 2, May 1989, pp. 739–742.

[52] S. Atungsiri, A. Kondoz, and B. Evans, "Error control for low-bit-rate speech communication systems," *IEE Proc. I: Communications, Speech and Vision*, vol. 140, pp. 97–103, April 1993.

[53] International Telecommunication Union, "ITU-T G.723.1 — Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," 1996.

[54] J. F. Wang, J. C. Wang, J. Yang, and J. J. Wang, "A voicing-driven packet loss recovery algorithm for analysis-by-synthesis predictive speech coders over Internet," *IEEE Trans. on Multimedia*, vol. 3, pp. 98–107, March 2001.

[55] J. Wang and J. Gibson, "Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001*, vol. 2, 2001, pp. 745–748.

[56] K. Saito, H. Fujiya, H. Kohmura, and S. Kanno, "Voice packet communication system for private networks," in *IEEE Global Telecommunications Conf., 1989, and Exhibition. Communications Technology for the 1990s and Beyond. GLOBECOM '89*, vol. 3, 1989, pp. 1874–1878.

[57] S. Atungsiri, A. Kondoz, and B. Evans, "Multi-rate coding: A strategy for error control in mobile communication systems," in *1993 IEE Colloquium on Low Bit-Rate Speech Coding for Future Applications*, 1993, pp. 3/1–3/4.

[58] W. Jiang and A. Ortega, "Multiple description speech coding for robust communication over lossy packet networks," in *IEEE International Conference on Multimedia and Expo, 2000*, vol. 1, 2000, pp. 444–447.

[59] R. Singh and A. Ortega, "Erasure recovery in predictive coding environments using multiple description coding," in *1999 IEEE 3rd Workshop on Multimedia Signal Processing*, 1999, pp. 333–338.

[60] C. Montminy and T. Aboulnasr, "Improving the performance of ITU-T G.729A for VoIP," in *IEEE International Conference on Multimedia and Expo, 2000*, vol. 1, 2000, pp. 433–436.

[61] A. Anandakumar, A. McCree, and V. Viswanathan, "Efficient CELP-based diversity schemes for VoIP," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP '00*, vol. 6, 2000, pp. 3682–3685.

[62] A. Ingle and V. A. Vaishampayan, "DPCM system design for diversity systems with applications to packetized speech," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 48–58, January 1995.

[63] W. Erhart and J. Gibson, "A speech packet recovery technique using a model based tree search interpolator," in *IEEE Workshop on Speech Coding for Telecommunications*, 1993, pp. 77–78.

[64] R. V. Cox and P. Kroon, "Low bit-rate speech coders for multimedia communication," *IEEE Communication Magazine*, vol. 34, pp. 34–41, December 1996.

[65] B. Wah and D. Lin, "Real-time voice transmissions over the Internet," *IEEE Trans. on Multimedia*, vol. 1, pp. 342–351, December 1999.

[66] D. Lin, "Real-time voice transmissions over the Internet," M.S. Thesis, University of Illinois at Urbana Champaign, December 1998.

[67] J. C. Bolot, "Characterizing end-to-end packet delay and loss in the Internet," *High-Speed Networks*, vol. 2, pp. 305–323, December 1993.

[68] P. Sinha, N. Venkitaraman, R. Sivakumar, and V. Bharghavan, "WTCP: A reliable transport protocol for wireless wide-area networks," in *Proc. of the Fifth Annual ACM/IEEE International Conference on Mobile Computing and Networking*, August 1999, pp. 231–241.

[69] B. H. Juang, "The past, present, and future of speech processing," *IEEE Signal Processing Magazine*, vol. 15, pp. 24–48, May 1998.

[70] L. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[71] A. S. Spanias, "Speech coding: A tutorial review," in *Proc. of the IEEE*, vol. 82, October 1994, pp. 1441–1582.

[72] F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," in *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, 1984, pp. 1.10.1–1.10.4.

[73] R. Goldberg and L. Riek, *A Practical Handbook of Speech Coders*. Boca Raton, Florida: CRC Press, 2000.

[74] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. on Acoutics, Speech, and Signal Processing*, vol. 24, October 1976.

[75] K. H. Lam, O. C. Au, C. C. Chan, K. F. Hui, and S. F. Lau, "Objective speech quality measure for cellular phone," in *1996 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, vol. 1, May 1996, pp. 487–490.

[76] H. Stark and J. Woods, *Probability, Random Processes and Estimation Teory for Engineers*. Upper Saddle River, NJ: Prentice-Hall, 1994.

[77] K. Choukri, G. Chollet, and Y. Grenier, "Spectral transformations through canonical correlation analysis for speaker adaptation in asr," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1986. ICASSP '86*, 1986, pp. 2659–2662.

[78] K. S. Shanmugan and A. M. Breipohl, *Random Signals, Detection, Estimation, and Data Analysis*. Chichester: John Wiley & Sons, 1988.

[79] E. B. Saff and A. D. Snider, *Fundamentals of Complex Analysis for Mathematics, Science and Engineering, 2/e*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[80] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. of the IEEE*, vol. 81, pp. 1385–1422, October 1993.

[81] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 313–327, July 1998.

[82] P. Ladefoged, "Phonetic data analysis: An introduction to phonetic fieldwork and instrumental techniques," 2002, http://www.jladefoged.com.

[83] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 59–71, January 1995.

VITA

Dong Lin received the B.E. degree in electrical engineering and information science from the University of Science and Technology of China in 1996, and the M.S. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 1999.

Her interests include speech, audio, and video coding, computer networking, signal processing, communication, cryptography, and security.

Following the completion of her Ph.D, she will begin working for Oracle Corporation in its *inter*Media group.