LSP-Based Multiple-Description Coding for Real-Time Low Bit-Rate Voice Over IP

Benjamin W. Wah, *Fellow, IEEE,* and Dong Lin

***Abstract*—A fundamental issue in real-time interactive voice transmissions over unreliable IP networks is the loss or late arrival of packets for playback. This problem is especially serious when transmitting low bit rate-coded speech with pervasive dependencies introduced. In this case, the loss or late arrival of a single packet will lead to the loss of subsequent dependent frames. In this paper, we study end-to-end loss-concealment schemes for ensuring high quality in playback. We propose a novel multiple description-coding method for concealing packet losses in transmitting low bit rate-coded speech. Based on high correlations observed in linear predictor parameters–in the form of Line Spectral Pairs (LSPs)–of adjacent frames, we generate multiple descriptions in senders by interleaving LSPs, and reconstruct lost LSPs in receivers by linear interpolations. As excitation codewords have low correlations, we further enlarge the segment size for excitation generation and replicate excitation codewords in all the descriptions in order to maintain the same transmission bandwidth. Our proposed scheme can be extended easily to more than two descriptions and can adapt its number of descriptions dynamically to network-loss conditions. Experimental results on FS-1016 CELP, ITU G.723.1, and FS MELP coders show good performance of our scheme.**

***Index Terms*—Internet, Line Spectral Pairs, loss concealment, low bit-rate speech coding, multiple-description coding, single-description coding, UDP, voice over IP.**

I. Introduction

INCREASES in network bandwidth and computational speed have led to growing interests in real-time interactive voice applications, such as Internet telephony, teleconferencing, and cellular communications. However, the quality of such applications is usually subpar to toll quality provided by telephone companies because the Internet is a packet-switched, best-effort delivery service, with no quality-of-service (QoS) guarantees. Hence, packets carrying real-time data may be dropped or arrive too late to be useful.

There are two distinct ways to improve the quality of voice data transmitted in a network: enhancing QoS supports in the network and developing better end-to-end protocols and coding algorithms.

The design of new protocols with QoS supports will help enhance the quality of real-time voice communications, although these protocols still cannot guarantee delivery within bounded time because they may use an over-subscribed network path or an unreliable network medium (such as a wireless link with interference). Such is the case in IPv6 that allows flows and priorities to be handled differently for real-time traffic but does not guarantee link reliability. Unless strict admission control is imposed with guaranteed bandwidth, losses are, therefore, inevitable when periodic data has to be received under time constraints.

An alternative is to design new end-to-end protocols that retransmit lost or delayed packets, either automatically or at the request of a receiver. This approach is not feasible in interactive voice communications when one-way delays have to be limited to 400 ms or less (as specified in ITU-T G.114) and when bandwidth is limited. Further, indiscriminate retransmissions may compete unfairly with other ongoing transmissions. A better end-to-end approach that works well with the first approach is to design loss-concealment schemes that exploit redundancies in voice data and that reconstruct lost data from that received. However, the design of such schemes is difficult when coding algorithms introduce dependencies in a compressed stream. In this case, loss concealments need to be applied not only to a lost frame but to all subsequent dependent frames that cannot be decoded even if they were received correctly.

Based on the end-to-end reconstruction approach discussed above, this paper develops new loss-concealment schemes for recovering lost data, when voice samples are compressed using low bit-rate coding standards, packetized, and sent in the Internet using the IPv4 UDP protocol. The schemes developed are general enough for reconstructing lost packets transmitted by UDP in IPv6.

The design of loss-concealment schemes for current low/very low bit-rate speech coding standards requires the understanding of dependencies introduced during coding. Most of these standards are based on the principle of linear prediction (LP) [10]. Using a linear predictor that models a vocal tract [22], an LP coder generates synthesized speech $\hat{s}_n$ from original speech s_n by exciting the vocal tract using multiplications of excitations and gains. Depending on whether the coder is close-looped (*LP analysis-by-synthesis coder* (LPAS) [26]) or open-looped (*vocoder*), excitations are chosen with or without minimizing the perceptual weighted mean squared errors between $\hat{s}_n$ and s_n. Because speech is quasistationary, it is coded frame-by-frame using linear prediction analysis, whose coefficients are commonly represented as *Line Spectral Pairs* (LSP) [25].

Manuscript received February 12, 2002; revised November 7, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hong Heather.

B. Wah is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801 USA (e-mail: wah@manip.crhc.uiuc.edu; http://manip.crhc.uiuc.edu).

D. Lin was with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801 USA. She is now with the Oracle Corporation, Nashua, NH 03002 USA (e-mail: dong.lin@oracle.com).

Digital Object Identifier 10.1109/TMM.2004.840593

TABLE I
MAJOR TECHNIQUES IN FOUR LOW BIT-RATE LP SPEECH CODERS AND THEIR BIT RATES (AC: ADAPTIVE CODE WORD; VQ: VECTOR QUANTIZATION)

Standard	bps	LSP Quant.	Excitation
FS 1016 CELP [17]	4800	scalar	stochastic code/AC
ITU G.723.1 (I) [30]	5300	pred.-split VQ	algebraic code/AC
ITU G.723.1 (II) [30]	6300	pred.-split VQ	multi pulse/AC
FS MELP [27]	2400	multi-stage VQ	mixed

Table I lists the bit rate, quantization method of LSPs, and excitation code-book structure for four representative LP speech coders. The top three are LPAS coders, and the last, a vocoder. Speech coding exploits temporal redundancies among speech signals and employs lossy quantization in order to compress the signals. Current compression algorithms are not robust to transmission errors because they assume an error-free channel when maximizing their coding gain. This is especially true in low bit-rate speech coding algorithms that remove as much redundancy as possible, leading to a great deal of dependencies in the coded sequence. As a result, the loss of a speech packet in packet-switched networks causes degradation in playback quality not only for the lost packet itself, but also for subsequent packets.

To deliver low bit-rate voice streams over packet networks with high quality, an active research area is to develop simple and robust loss-concealment and coding strategies. Existing techniques can be classified into those with redundancies and those without.

Adding packet-level redundancies is not the best for fault tolerance in the Internet because they require considerable increases in bandwidth over nonredundant schemes. Typical methods include adding copies of previous frames [18], using parity or forward error correction (FEC) codes [24] to protect every n packets by a redundant packet, using FEC to protect only sensitive information in LP-coders [3], and piggy-backing in a packet a redundant version of some previous packets obtained by a lower bit-rate coder [5]. In addition, redundant information can be sent to protect parts of each packet. Such information includes voiced/unvoiced indicators, pitch information [23], background noise, fricatives indicators [11], short-time energy, zero-crossings [12], pitch estimates with amplitude, and fundamental frequency information of previous packets [7]. Receivers then use waveform substitution to replace lost packets by finding a best match on the redundant segments received. The drawback of these approaches is that good quality can only be achieved by sending a considerable amount of redundant information.

In contrast, schemes with *zero-redundancy control* exploit implicit redundancies in voice streams and the property that voice transmissions can tolerate some loss without a lot of perceivable differences. Simple schemes typically perform loss-concealment actions at receivers alone. For instance, lost packets can be recreated by

a) padding silence or white noise [28];
b) repeating the last received packet [29];
c) pattern matching using small segments of samples immediately before or after a lost packet [34];
d) pitch-period replication by estimating pitch periods using speech segments immediately before a lost packet [34];
e) performing waveform substitution based on previously received frames on each subband of linear prediction residues [8];
f) copying coder parameters from the most recent error-free packet to both reconstruct the lost packet and update coder states [9];
g) repeating the parameters of the previous frame with simple modifications [2], [30];
h) periodically informing senders to restart encoding in order to reduce error propagation [21];
i) performing reconstruction differently for voice and unvoice frames [36], [37].

These strategies work well when losses are infrequent and when packet sizes are small [14], but fail in high-loss networks.

Dissimilar to the single description-coding (SDC) schemes above, multi-description coding (MDC) is a popular scheme that divides a data stream into equally important streams in such a way that the decoding quality with any subset is acceptable, and that better quality is obtained by more descriptions. It is assumed that losses to different descriptions are uncorrelated, and that the probability of losing all the descriptions is small. A straightforward way to implement MDC is *interleaving* (also called *sample-based MDC*) in which adjacent samples are distributed to different packets, thereby converting bursty losses to random losses that are much easier to recover. Receivers may reconstruct lost samples by odd-even sample interpolations [15], pattern-matching sample interpolations [35], Kalman-based sample interpolations [6], and adding redundant information to improve reconstruction quality [16]. We explain in Section II why such methods are not suitable for concealing errors in low bit rate-coded speech. Another redundant MDC [1] proposed for low bit-rate speech coders is to replicate important informations to all descriptions, such as linear predictors and pitch information, and to interleave excitation codewords. The drawback of this approach is that excitations cannot be reconstructed if loss happens.

In general, zero redundancy MDC schemes have not been developed specifically for LP speech coders. Since most current low bit-rate speech coding standards are based on linear prediction [10], our objective is to design LP coder-specific MDC loss-concealment methods.

This paper is organized as follows. Section II studies packet-loss patterns in the Internet and explains why MDC is attractive. Sections III and IV present our proposed LSP-based MDC algorithm and experimental results. Finally, Section V concludes the paper.

II. LOSS CHARACTERISTICS IN THE INTERNET

This section presents loss characteristics of real-time transmissions for domestic and international connections and concludes that MDC is suitable for concealing packet losses.

Our first set of experiments address the probability distribution of consecutive packet losses. During the experiments carried out in the first week of November 2001, a computer at UIUC periodically sent 2000 probe packets, at a rate of 30 packets/s and 500 bytes per packet, at the beginning of each hour over a 24-h period to the echo port of a remote computer, and monitored the packets bounced back. The inter-packet transmission

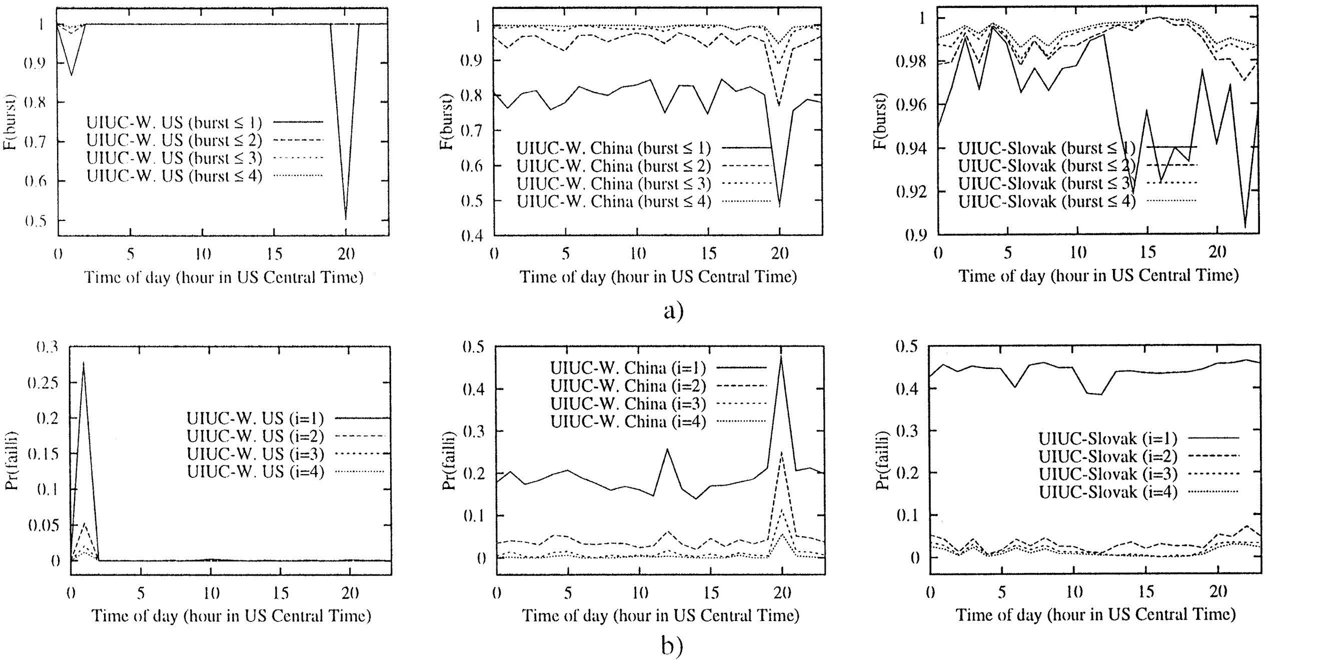


Fig. 1. Statistics on bursty losses in round-trip paths between UIUC and three remote locations: Berkeley (daedalus.cs.berkeley.edu), W. China (www.lzu.edu.cn), and Slovak (us.svf.stuba.sk). (a) $F(burst)$: probability distribution of raw bursty losses and (b) $Pr(fail \mid i)$: probability distributions of bursty losses that cannot be recovered under interleaving factor i.

time of around 30 ms represents the duration that packets are encoded in LSP-coded speech studied in this paper. In collecting the traces, we have used a packet size of 500 bytes instead of the actual packet size in LSP-coded speech because we have found in our previous experiments [19] that packet size does not affect the loss behavior as long as it is less than MTU. We have assumed uncompressed voice data in the form of 16-bit PCM and sampled at 8 kHz, leading to a packet size of 480 bytes. (Results on other packet transmission rates can be found elsewhere [20], [19].)

Statistics, such as the sending and arrival times of each packet, was collected. To account for "delayed losses," each packet received had a scheduled "playback" time calculated from the arrival time of the first received packet and the difference of their sequence numbers. A packet was considered lost if it had been delayed by more than 200 msec from its scheduled playback time. By using traces collected in trace-driven experiments, we were able to test the same network-loss condition under different algorithms that would not have been possible if streaming experiments were carried out in real time.

Fig. 1 demonstrates some typical results for a domestic and two international connections. It shows that burst lengths of one and two are predominant. Even for the two international connections with high losses, more than 80% of their losses were of burst length one or two.

Moreover, 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-Slovak connection, the unconditional probability for the current burst length to be four and the next burst length to be greater than or equal to four is only 0.001 [20], [32].

The fact that burst lengths are usually small (similar results have been shown in [4]) indicates that interleaving can be a good method to ease reconstruction. Define an *interleaving set* to be a collection of related information that is interleaved to different descriptions. When the burst length is less than the interleaving factor, or when a bursty loss involves information from different interleaving sets and some information in each interleaving set is received correctly, the information received can be used to recover the lost parts. For instance, with sample-based interleaving and an interleaving factor of two, a bursty loss of length one and a bursty loss of length two with samples belonging to different interleaving sets can be recovered by interpolations. With an interleaving factor of four, a bursty loss of length less than or equal to three and a bursty loss of length four, five, or 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 bursty losses of length in the range $[i,\ (2i-2)]$.

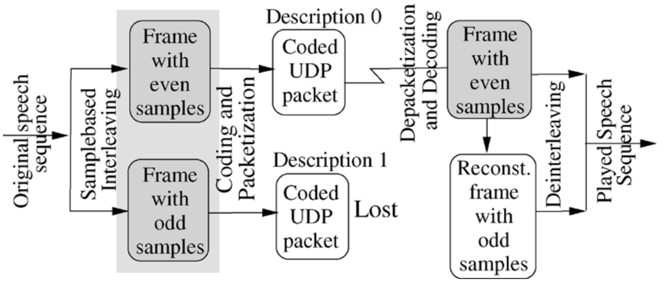
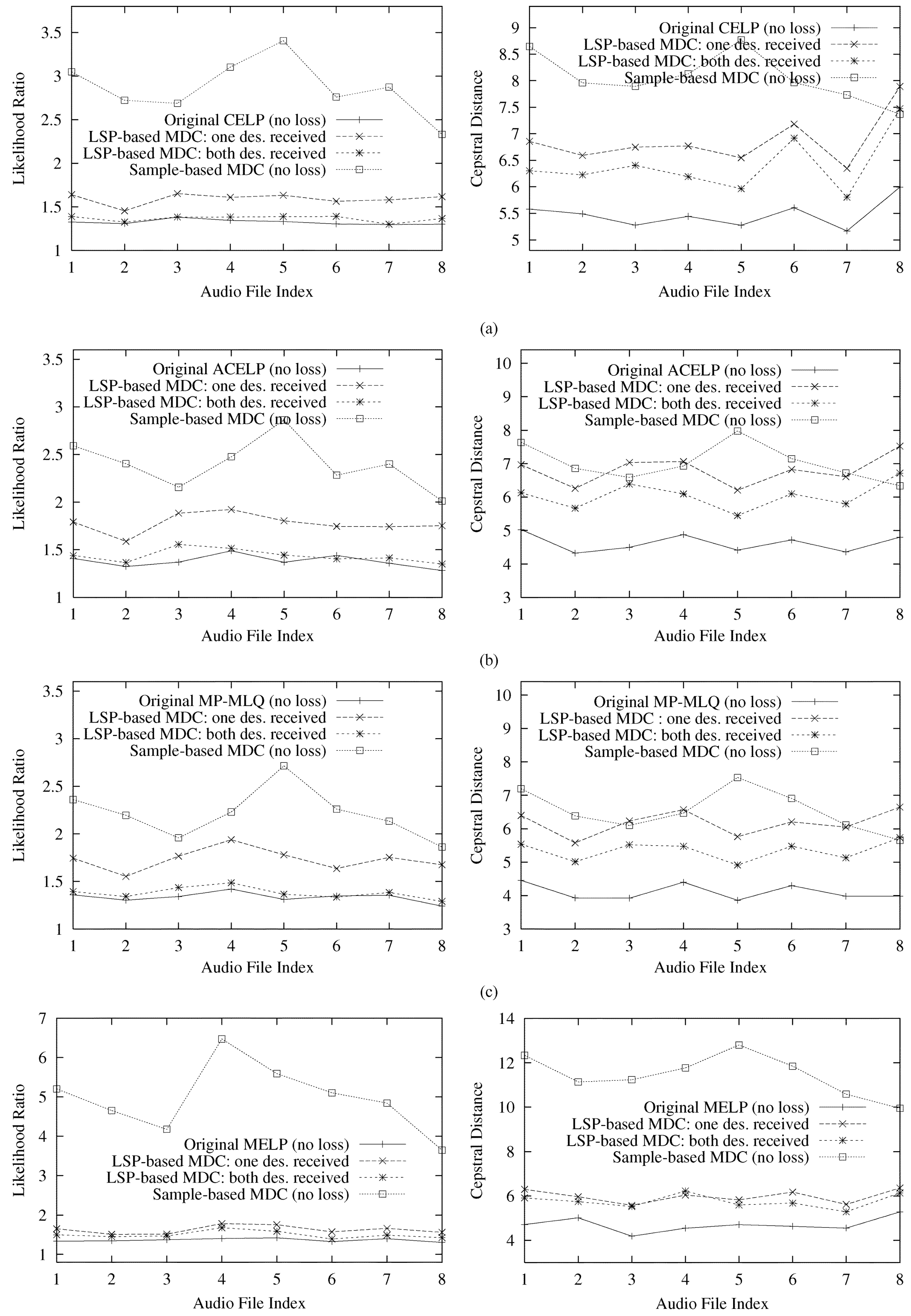


Fig. 2. Two-way sample-based MDC and reconstruction of a lost description at a receiver.

Let the total number of packets sent be n_p. Over all the interleaving sets, assuming that consecutive packet losses of length j, $j \le i$,[1] happen $m_{i,j}$ times, and that $n_s = \sum_{j=1}^{i} j \times m_{i,j}$ packets are lost (independent of i). We can derive $Pr(fail \mid loss, i)$, the

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



(a)

(b)

(c)

(d)

Fig. 3. Quality comparison in terms of LR and CD among SDC, sample-based MDC with both streams received, and two-way LSP-based MDC under two scenarios in synthetic experiments. (a) FS-1016 CELP. (b) ITU G.723.1 ACELP. (c) ITU G.723.1 MP-MLQ. (d) ITU G.723.1 FS MELP.

conditional probability that a packet cannot be recovered for interleaving factor i. This happens when all the packets in an interleaving set are lost.

$$Pr(fail \mid loss, i) = \frac{i \times m_{i,i}}{n_s}. \quad (1)$$

$Pr(fail \mid i)$, the unconditional probability that a packet cannot be recovered for interleaving factor i, is

$$Pr(fail \mid i) = \frac{i \times m_i^i}{n_s} \times \frac{n_s}{n_p} = \frac{i \times m_i^i}{n_p}. \quad (2)$$

TABLE II
VOICE STREAMS USED IN OUR EXPERIMENTS

Index	Length (ms)	Speakers
1	21432	2 male, 1 female
2	22560	2 male, 1 female
3	4424	1 female
4	5091	1 female
5	4160	1 male
6	4082	1 male
7	4867	1 male, 1 female
8	73615	1 male, 1 female

Streams 1 and 2 were obtained from John Hopkins University;
Streams 3 thru 7 were obtained from Cybernetics InfoTech. Inc.;
Streams 8 was obtained from VoiceAge Corp.

Fig. 1(b) shows that $Pr(fail \mid i)$ drops quickly when i increases. For all times and the three connections, $Pr(fail \mid i)$ is negligible when $i \geq 4$. Moreover, $i = 2$ works well for the connection to Berkeley, achieving $Pr(fail \mid i)$ well below 5%. For the connection to China, $i = 2$ is not always enough because about 20% of the total losses will not be recoverable.

The above results suggest that a small number of descriptions (between two to four) is adequate. In most cases, two-way MDC leads to good recovery.

III. LSP-BASED MDC

We present an LSP-based MDC scheme for concealing packet losses in low bit-rate LP coders.

First, we demonstrate that the traditional sample-based MDC is not suitable for linear predictive coders. Fig. 2 shows the 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 condition, both coded streams will be received, decoded separately, and de-interleaved to rebuild the original stream. Even in this case, the playback quality is very poor, as illustrated in Fig. 3 for the four coders in Table I and the eight sample voice streams in Table II. Here, performance is measured by the *Itakura-Saito Likelihood Ratio* (LR) and the *Cepstral Distance* (CD) [33]. LR for a speech frame is defined as

$$\text{LR} = \frac{\vec{a}_r R_o \vec{a}_r^T}{\vec{a}_o R_o \vec{a}_o^T} \tag{3}$$

where $\vec{a}_o$ and $\vec{a}_r$ are the vectors of linear prediction coefficients of the original and reconstructed speech sequences, respectively, and R_o is the correlation matrix derived from the original speech. CD for a speech frame is defined as

$$\text{CD} = 4.34 \left[(c_0 - c_0')^2 + 2 \sum_{i=1}^{\infty} (c_i - c_i')^2 \right]^{1/2} \text{[dB]} \tag{4}$$

where c_i and c_i' denote the cepstra of the original and reconstructed samples, respectively. The performance of each stream is the average over all the frames tested.

Fig. 3 shows that both LR and CD of sample-based MDC increase dramatically for the four coders and all the files tested under no loss, when compared to the decoding quality of SDC. 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. (Results on LSP-based MDC are described later.)

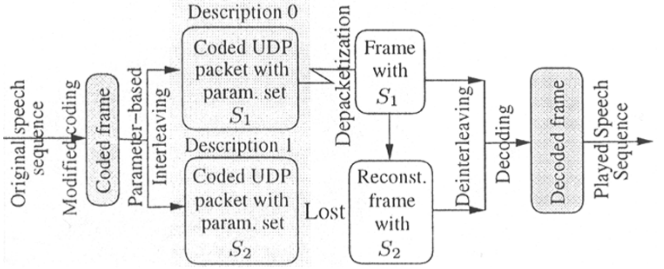


Fig. 4. Proposed two-way MDC for LP speech coders.

TABLE III
INTER-FRAME CORRELATIONS OF LSPS FOR THE EIGHT TEST STREAMS (8000-HZ SAMPLING RATE, 30-ms FRAMES, 45-ms HAMMING WINDOW, 10^{th} ANALYSIS ORDER, AND 8061 FRAMES)

Frame	LSP									
Distance	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	0.84	0.81	0.82	0.79	0.78	0.77	0.77	0.80	0.74	0.64
2	0.62	0.60	0.63	0.59	0.56	0.56	0.57	0.61	0.53	0.44
3	0.43	0.45	0.50	0.45	0.41	0.42	0.43	0.48	0.39	0.34

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 stream. To avoid these drawbacks, we investigate the interleaving of parameters after coding in LP coders, as shown in Fig. 4.

We first study the properties of coder parameters for loss concealments. As mentioned in Section I, the common part in modeling a vocal tract in most low bit-rate speech coders is the linear predictor, often represented by LSPs. From the physical point, since a vocal tract changes slowly and smoothly when one speaks, we expect that its LSPs will change slowly. Given the synthesis filter defined as

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + a_o z^{-1} + \cdots + a_{10} z^{-10}} \tag{5}$$

the LSPs are the zeros $x_1, x_2, \ldots, x_{10}$ on the upper unit circle of the following two functions [25]:

$$\begin{aligned} P(z) &= A(z) + z^{-11} A(z^{-1}) \\ &= (1 + z^{-1}) \prod_{k=1}^{5} (1 - 2\cos x_{2k-1} z^{-1} + z^{-2}) \end{aligned} \tag{6}$$

$$\begin{aligned} Q(z) &= A(z) - z^{-11} A(z^{-1}) \\ &= (1 - z^{-1}) \prod_{k=1}^{5} (1 - 2\cos x_{2k} z^{-1} + z^{-2}). \end{aligned} \tag{7}$$

There are three important properties of LSPs that make the linear interpolations of LSPs useful for loss concealments. First, the difference of adjacent LSPs is closely related to the formant bandwidths of speech [13], which suffice to specify the entire spectral envelope for vowels. This close relationship means that

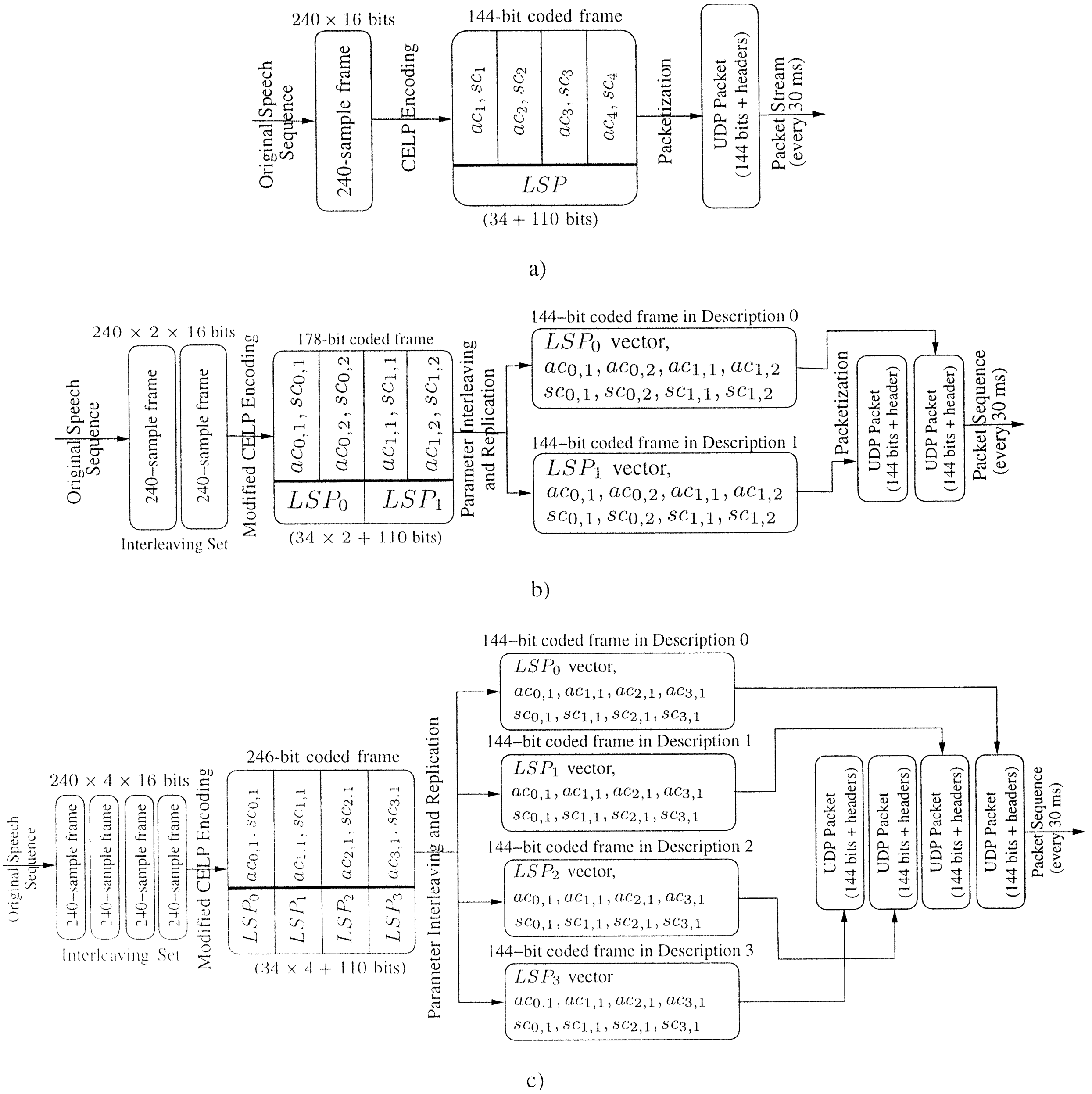

240 × 16 bits
Original Speech Sequence
240-sample frame
CELP Encoding
144-bit coded frame
ac_1, sc_1
ac_2, sc_2
ac_3, sc_3
ac_4, sc_4
LSP
(34 + 110 bits)
Packetization
UDP Packet (144 bits + headers)
Packet Stream (every 30 ms)
a)
240 × 2 × 16 bits
240-sample frame
Interleaving Set
Modified CELP Encoding
178-bit coded frame
$ac_{0,1}, sc_{0,1}$
$ac_{0,2}, sc_{0,2}$
$ac_{1,1}, sc_{1,1}$
$ac_{1,2}, sc_{1,2}$
LSP_0
LSP_1
(34 × 2 + 110 bits)
Parameter Interleaving and Replication
144-bit coded frame in Description 0
LSP_0 vector, $ac_{0,1}, ac_{0,2}, ac_{1,1}, ac_{1,2}$ $sc_{0,1}, sc_{0,2}, sc_{1,1}, sc_{1,2}$
144-bit coded frame in Description 1
LSP_1 vector, $ac_{0,1}, ac_{0,2}, ac_{1,1}, ac_{1,2}$ $sc_{0,1}, sc_{0,2}, sc_{1,1}, sc_{1,2}$
Packetization
UDP Packet (144 bits + header)
Packet Sequence (every 30 ms)
b)
240 × 4 × 16 bits
240-sample frame
Interleaving Set
Modified CELP Encoding
246-bit coded frame
$ac_{0,1}, sc_{0,1}$
$ac_{1,1}, sc_{1,1}$
$ac_{2,1}, sc_{2,1}$
$ac_{3,1}, sc_{3,1}$
LSP_0
LSP_1
LSP_2
LSP_3
(34 × 4 + 110 bits)
Parameter Interleaving and Replication
144-bit coded frame in Description 0
LSP_0 vector, $ac_{0,1}, ac_{1,1}, ac_{2,1}, ac_{3,1}$ $sc_{0,1}, sc_{1,1}, sc_{2,1}, sc_{3,1}$
144-bit coded frame in Description 1
LSP_1 vector, $ac_{0,1}, ac_{1,1}, ac_{2,1}, ac_{3,1}$ $sc_{0,1}, sc_{1,1}, sc_{2,1}, sc_{3,1}$
144-bit coded frame in Description 2
LSP_2 vector, $ac_{0,1}, ac_{1,1}, ac_{2,1}, ac_{3,1}$ $sc_{0,1}, sc_{1,1}, sc_{2,1}, sc_{3,1}$
144-bit coded frame in Description 3
LSP_3 vector $ac_{0,1}, ac_{1,1}, ac_{2,1}, ac_{3,1}$ $sc_{0,1}, sc_{1,1}, sc_{2,1}, sc_{3,1}$
UDP Packet (144 bits + headers)
Packet Sequence (every 30 ms)
c)

Fig. 5. FS-1016 CELP under single description and the decomposition of an interleaved set of frames into multiple descriptions in LSP-based MDC (*ac*: adaptive codeword; *sc*: stochastic codeword). (a) SDC at sender, (b) two-way LSP-based MDC at sender, and (c) four-way LSP-based MDC at sender.

linear interpolations of LSPs, which is equivalent to the interpolation of the difference of adjacent LSPs, is closely related to the generation of smooth formant information. Second, we have found experimentally that LSPs change slowly from one frame to the next and have high inter-frame correlations. This is illustrated in Table III that shows high correlations for all the ten indices under a typical frame size of 30 ms. The correlations found are general because the tested streams involve significant variations in speaker characteristics. These results mean that LSPs in a lost frame can be reconstructed from those received in adjacent frames. In contrast, we have found that excitation parameters across adjacent subframes have low correlations and cannot be reconstructed from those received in adjacent subframes [20]. Last, the vector of LSP indices in a frame are monotonically increasing ($0 < x_1 < x_2 < \cdots < x_{10}$). This means that they are stable linear predictors [25], and that a vector of interpolated LSPs will also be stable linear predictors when using linear interpolations to reconstruct lost LSPs.

Based on these observations, we propose an LSP-based MDC scheme for LP coders. The key idea [31] is to interleave the highly correlated LSP vectors in multiple descriptions and to reconstruct missing ones from those received in descriptions in the same interleaving set. For excitation parameters that have low correlations, we replicate them in multiple descriptions in an interleaving set and use any received in an interleaving set for decoding. Further, we reduce the number of excitation parameters generated in such a way that their replication leads to the same bandwidth as that in SDC.

We illustrate our proposed LSP-based MDC scheme on the FS-1016 CELP coder. Fig. 5(a) shows, in the original SDC coder, the generation of a 34-bit LSP vector for each 240-sample speech frame and four groups of adaptive and stochastic codewords (with 110 bits total), one for each 60-sample subframe. The 144-bit coded frame is then encapsulated in a UDP packet and sent to a receiver.

The single-description decoding process is the reverse process in Fig. 5(a). When one or more consecutive packets are lost, the receiver will not be able to recover the parameters in the corresponding coded frames and will play silence in the decoded speech frames. When valid packets are received again, the receiver will extract the parameters and feed them to the CELP decoder. Since decoding is state dependent, there will be several frame delays before the proper states of the decoder are restored and satisfactory playback quality is achieved. As a result, the performance of SDC is very sensitive to losses and burst lengths.

In our two-way LSP-based MDC design, the sender groups each pair of 240-sample frames in the original speech sequence into an interleaved set, performs linear prediction analysis, once for each frame, in order to generate a 34-bit LSP vector, and distributes the two LSP vectors to two frames in the two descriptions [see Fig. 5(b)]. However, instead of generating a 110-bit vector of excitation codewords for four 60-sample subframes, it extends the subframe size to 120 samples, generates a 110-bit vector of codewords for four 120-sample subframes, and replicates the 110-bit vector to the two frames in both descriptions. The vector of codewords is replicated because its elements are not correlated and cannot be reconstructed from those in adjacent frames. With replicated codewords, we need to extend the subframe size in coding in order to keep the coded frame size in each description to be 144 bits, the same size as a coded frame in SDC. After quantization of information and packetization, the LSP vector of frame $2n$ (*resp.* $2n+1$) is placed in UDP packet $2n$ (*resp.* $2n+1$), and the excitation codewords of frames $2n$ and $2n+1$ are replicated in packets $2n$ and $2n+1$. Finally, the sender sends the packet stream to the destination. Note that we have maintained the same packet size and packet rate as those in SDC 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 Fig. 5(b). It first deinterleaves the information received into a single coded stream by extracting the LSPs from the frames in both descriptions and the codewords from the 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. The decoded stream can be guaranteed to have better quality than sample-based MDC because we have preserved the precision of linear prediction analysis and have eliminated aliasing. However, it has worse quality than that of SDC because of its increased subframe size.

When some frames in one description are lost, the receiver only needs to reconstruct the lost LSPs using the LSPs in those frames received in the other description. It does not reconstruct the codewords because they are replicated in both descriptions. For example, if a frame in Description 1 of Fig. 5(b) is lost, then the receiver reconstructs the LSPs in the lost frame by averaging the LSPs of the immediately preceding and following frames in Description 0. It is easy to see that such reconstructions result in stable linear predictors. Moreover, since the receiver reconstructs the coding parameters of lost frames before decoding, it does not need to estimate the decoding states of lost frames as done 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 LSPs and codewords and will have to generate silence in the decoded speech sequence, similar to the recovery process in SDC. However, the chance for such bursty losses is much smaller than that in SDC because all losses of burst length one and some losses of burst lengths two can be recovered.

The above idea can be extended to four-way LSP-based MDC. Fig. 5(c) shows the way that the sender interleaves the LSPs of four 240-sample frames in an interleaved set into four frames in the four descriptions. Here, the frame size in linear prediction analysis is still 240, but the subframe size is increased to 240 in order to maintain the same 144-bit frame size after the codewords are replicated. After coding and parameter interleaving, the LSP vector of frame $4n+i$, $i = 0, 1, 2, 3$, is placed in UDP packet $4n+i$, and the excitation codewords of frames $4n$ through $4n+3$ are replicated in packets $4n$ through $4n+3$. Due to the longer subframe size, the quality of four-way MDC is expected to be worse than that of two-way MDC.

In four-way MDC, there are five loss patterns of frames in an interleaving set in which losses can be concealed at the 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, or 1–2–3, or 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 the other cases, the receiver can recover the lost LSPs by interpolating the LSPs received in immediately preceding and following frames. Similar to two-way MDC, we know that interpolations result in stable linear predictors.

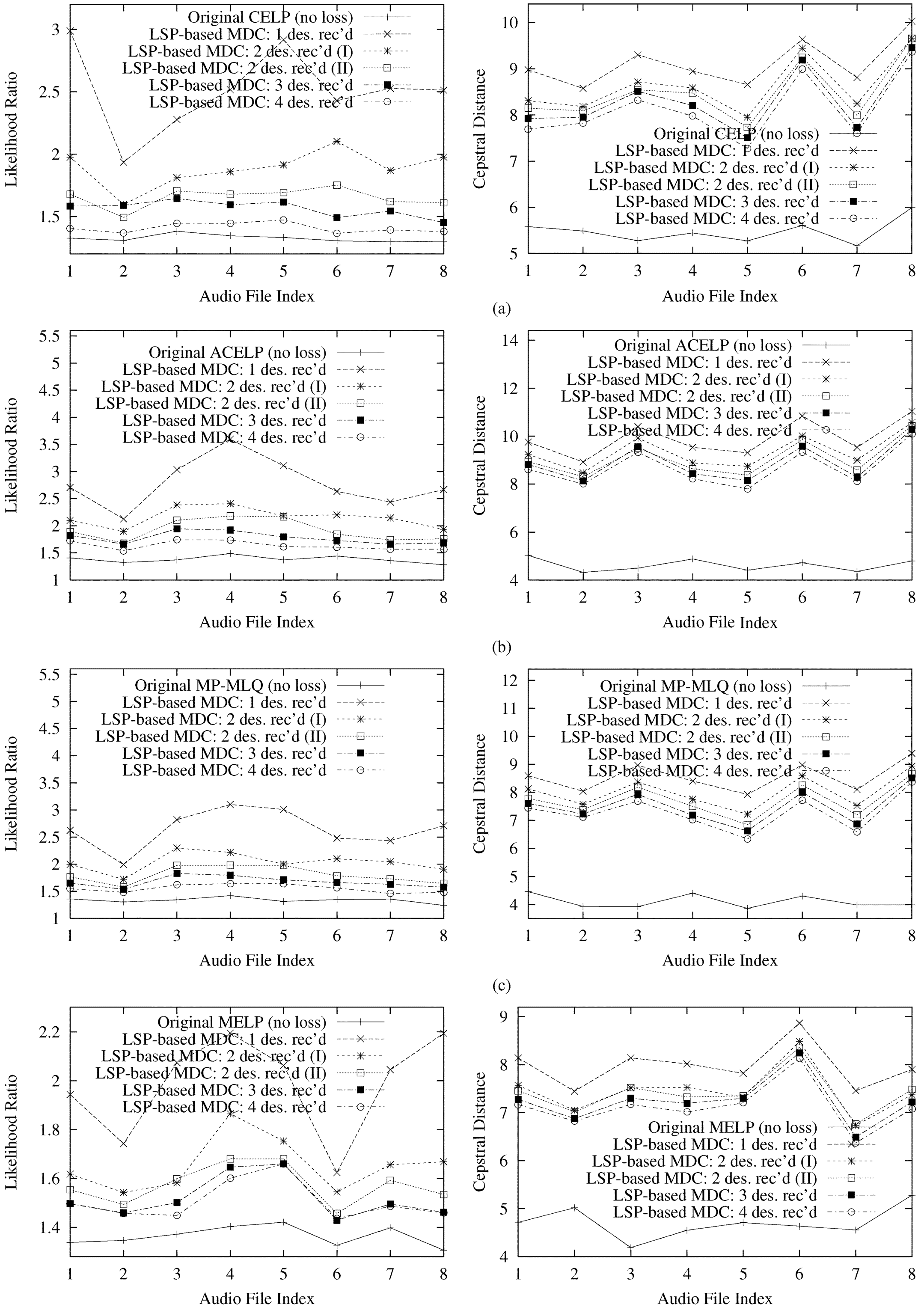
We do not study MDC beyond four ways because four-way interleaving will be enough to conceal errors in most, if not all, of the cases (see Fig. 1). Further, larger interleaving degrees will result in even larger subframe sizes that will degrade quality further at the receiver, even when there are no losses.

For the other three coders in Table I, we apply the same idea to construct two-way and four-way MDC, although there are some coder-specific variations [20]. In the next section, we present results to show the effectiveness of our proposed LSP-based MDC schemes.

IV. Experiment Results

In this section, we test our proposed two-way and four-way LSP-based MDC algorithms on the four coders in Table I using the eight test streams in Table II. We further present trace-driven simulation results using Internet packet traces collected under the conditions specified in Section II.

Fig. 3 compares the decoding quality of two-way LSP-based MDC in synthetic experiments between the cases when both descriptions are received and when only one description is received. When both descriptions are received, LSP-based MDC for the four coders in Table I has almost no degradation in terms of LR and about 10-20% degradation in terms of CD when



(a)

(b)

(c)

(d)

Fig. 6. Quality comparisons in terms of LR and CD under five scenarios for four-way LSP-based MDC in synthetic experiments (I: two consecutive descriptions were received, and II: two disjoint descriptions were received). (a) FS-1016 CELP, (b) ITU G.723.1 ACELP, (c) ITU G.723.1 MP-MLQ, and (d)ITU G.723.1 FS MELP.

compared to SDC, and significant improvements when compared to sample-based MDC. When only one description is received, LSP-based MDC still gives good quality and performs better than sample-based MDC with no loss. The performance is more pronounced for the very low bit-rate MELP coder. Our observations were further confirmed by subjective evaluations.

Note that the results plotted are averages and that some frames under the original SDC with no loss may actually have worse LR than that of frames under two-way LSP-based MDC with no loss. Hence, it is possible for the original SDC with no loss to perform a little worse on the average than LSP-based MDC with no loss. (For example, the original ACELP with no loss has

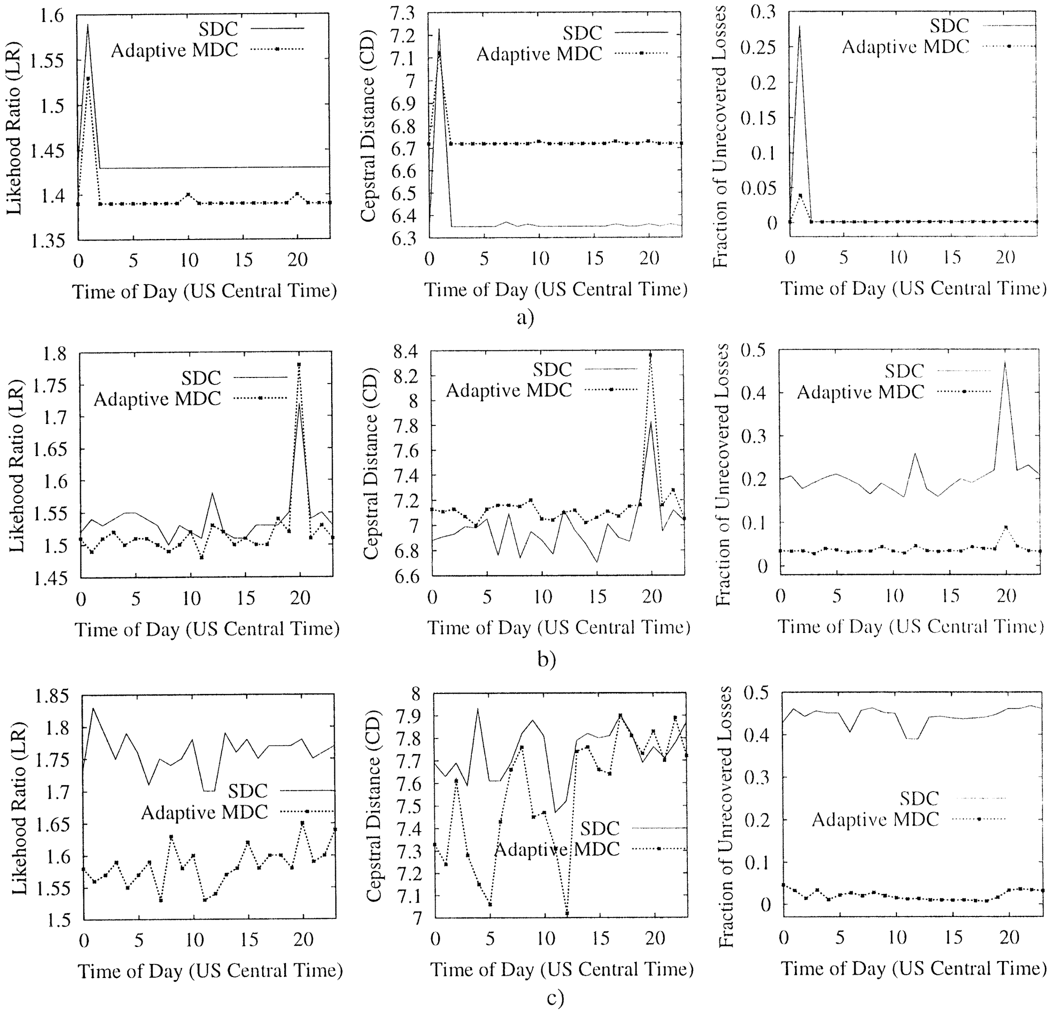


Fig. 7. Comparison of reconstruction quality between SDC and adaptive MDC for FS-1016 CELP on received and reconstructed frames over a 24-h period for the round-trip connections between UIUC and three destinations. The graphs on the right show the fraction of frames that were lost or cannot be reconstructed. (a)UIUC-Berkeley round-trip connection, (b)UIUC-Western China round-trip connection, and (c) UIUC-Slovak round-trip connection.

higher average LR than two-way LSP-based MDC with both descriptions received for Audio File 6 in Fig. 3(b).)

Next, Fig. 6 shows the reconstruction quality of four-way LSP-based MDC in synthetic experiments on five loss patterns where losses can be concealed, using the decoding quality of SDC with no loss as a baseline. In general, four-way MDC gives acceptable quality and performs worse than two-way MDC, especially in terms of CD. As is said earlier, such degradations are due to enlarged subframes in four-way MDC that lead to degraded precision in adaptive and stochastic codewords. The degree of degradation is also correlated to the degree of loss, which is illustrated by the consistently lower quality as more descriptions are lost. Our observations were confirmed by subjective evaluations.

To further verify our algorithms, we have built a prototype in Linux that codes input speech in real-time from a microphone using SDC or LSP-based MDC, transmits the coded packets to the echo port of a remote computer, reconstructs any lost information from the packets bounced back, and plays the reconstructed stream in another local computer.

In order 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 to the sender a one-bit message in UDP, indicating whether two-way or four-way MDC should be used. Since the feedbacks sent by the receiver are 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 coder 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. Our 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.

In the following, we used trace-driven simulations in order to compare various algorithms under the same operating condition.

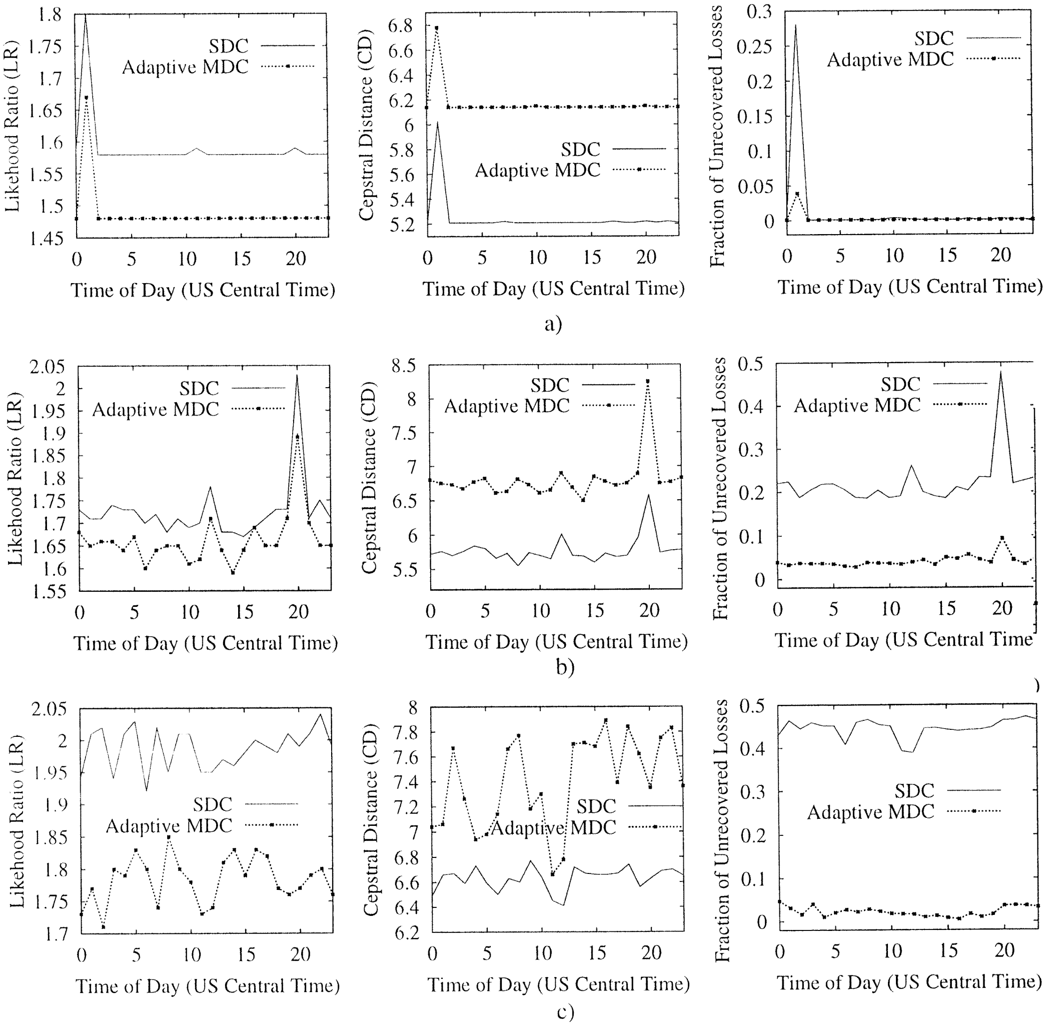


Fig. 8. Comparison of reconstruction quality between SDC and adaptive MDC for ITU G.723.1 ACELP on received and reconstructed frames over a 24-h period for the round-trip connections between UIUC and three destinations. The graphs on the right show the fraction of frames that were lost or cannot be reconstructed. (a) UIUC-Berkeley round-trip connection, (b) UIUC-Western China round-trip connection, and (c)UIUC-Slovak round-trip connection.

Using sites that represent typical low, medium, and high loss connections, we fed packet traces to our prototype and evaluated the statistics offline.

Figs. 7–9 show trace-driven results on SDC and adaptive MDC for FS-1016 CELP, ITU G.723.1 ACELP, and FS MELP, respectively. (Results for ITU G.723.1 MP-MLQ are similar to those of ACELP [20].) For each site tested, we show its playback quality measured in LR and CD averaged over all received and reconstructed frames. However, one must be careful in comparing the results 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 60-80% of all the frames transmitted (for the UIUC-China and UIUC-Slovak connections). This disparity in evaluations may lead to better LR but worse CD for adaptive MDC as compared to those of SDC in Figs. 7–9. To illustrate the reason leading to this disparity, we also plot the fraction of frames that were lost or unrecovered at the receiver for both schemes.

In general, adaptive MDC always have less distortions than SDC in terms of LR, but may sometimes have more distortions than SDC in terms of CD. To understand this difference, we need to consider two kinds of distortions. First, some distortions in adaptive MDC are introduced because excitations are extracted on larger subframes and are reflected more obviously in terms of CD. Other distortions in terms of both LR and CD are introduced when some frames are lost and unrecoverable, leading to incorrect decoding states for subsequent frames received. Such distortions happen to both SDC and adaptive MDC, but affect the quality of SDC more severely due to the large fraction of unrecoverable frames in SDC (see the rightmost graphs in Figs. 7–9). 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 far less 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.

As shown in Fig. 9, the advantage of applying adaptive MDC to FS MELP is more obvious. Except for slightly higher distortions in CD for the UIUC-Berkeley connection, adaptive MDC performs consistently better than SDC. The performance gain is particularly notable for the high-loss UIUC-Slovak connection.

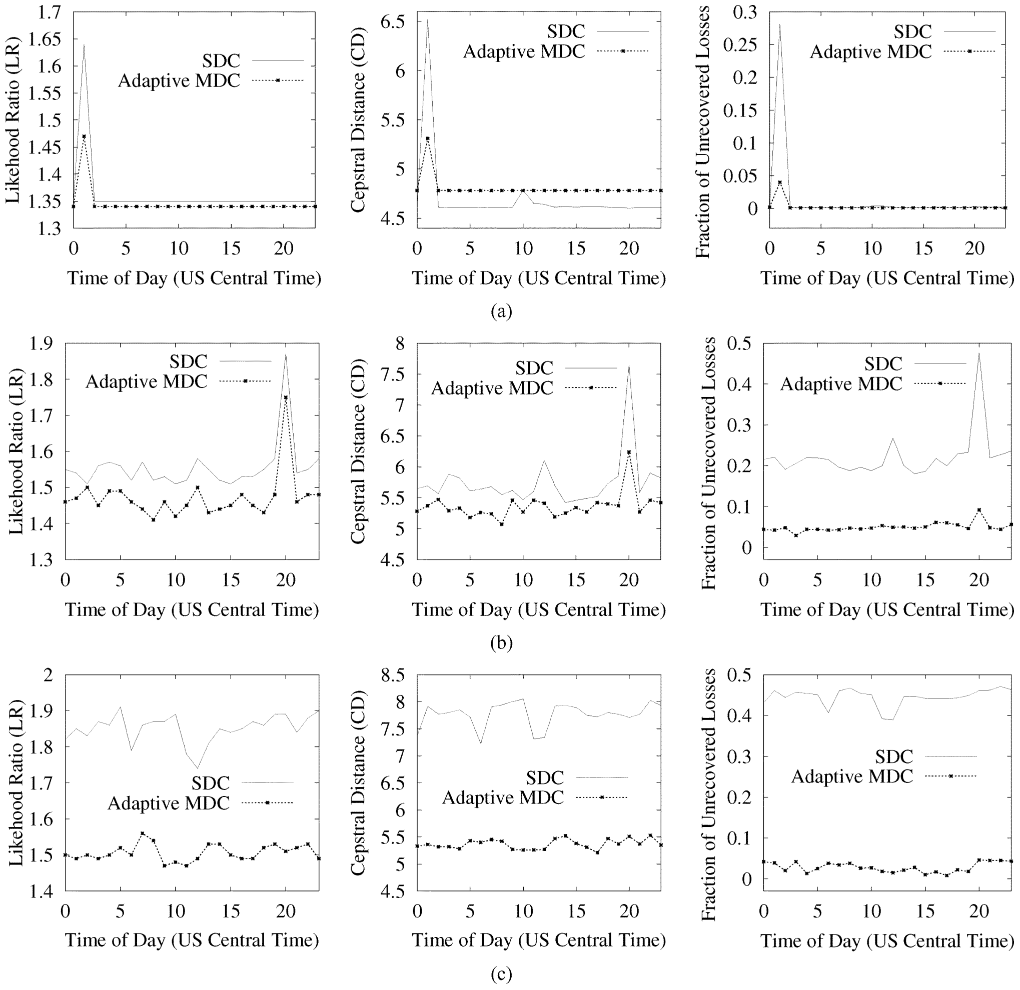


Fig. 9. Comparison of reconstruction quality between SDC and adaptive MDC for FS MELP on received and reconstructed frames over a 24-h period for the round-trip connections between UIUC and three destinations. The graphs on the right show the fraction of frames that were lost or cannot be reconstructed. (a) UIUC-Berkeley round-trip connection, (b) UIUC-Western China round-trip connection, and (c) UIUC-Slovak round-trip connection.

From the perspective of end users, SDC will give discontinuous playback for high-loss connections, such as UIUC-China and UIUC-Slovak with over 40% loss. 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 contract, 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. The delays experienced were also acceptable, given the round-trip delay of a packet traveling from UIUC to a remote echo port and back to UIUC is no more than 240 ms ($\approx$ 30 ms for Berkeley, between 180 and 240 ms for W. China, and $\approx$ 175 ms for Slovak). Subjective tests further verify that our proposed MDC scheme performs well under all conditions.

Although the quality in terms of CD in our Internet tests (Figs. 7–8) is still off from the quality of SDC under no loss (Fig. 6), significant improvements in CD can be achieved by proper adjustment of the perceptual weighting filter in closed-looped LP coders (such as FS-1016 CELP and ITU G.723.1 ACELP) that generate excitations by perceptually weighting their coding noises and in turn deciding code words for excitations [20].

V. Conclusions

In this paper, we have proposed and tested a novel LSP-based MDC scheme for low bit rate-coded speech transmissions over lossy IP networks. Without increasing the transmission bandwidth, our scheme represents a trade-off between the quality of the received packets and the ability to reconstruct lost packets. Our results show that LSP-based MDC can be applied to both close-looped and open-looped linear-prediction coders and greatly enhances their loss resilience.

References

[1] A. K. Anandakumar, A. V. McCree, and V. Viswanathan, “Efficient CELP-based diversity schemes for voip,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2000 (ICASSP’00)*, vol. 6, 2000, pp. 3682–3685.

[2] S. A. Atungsiri, A. M. Kondoz, and B. G. Evans, “Error control for low-bit-rate speech communication systems,” *Proc. Inst. Elect. Eng I: Commun., Speech Vis.*, vol. 140, no. 2, pp. 97–103, Apr. 1993.

[3] ——, “Multi-rate coding: A strategy for error control in mobile communication systems,” in *Proc. 1993 IEE Colloq. Low Bit-Rate Speech Coding for Future Applications*, 1993, pp. 3/1–3/4.

[4] J. C. Bolot, “Characterizing end-to-end packet delay and loss in the internet,” *High-Speed Networks*, vol. 2, no. 3, pp. 305–323, Dec. 1993.

[5] J. C. Bolot and A. V. Garcia, "Control mechanisms for packet audio in the internet," in *Proc. IEEE INFOCOM*, Apr. 1996, pp. 232–239.

[6] 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. Speech Audio Process.*, vol. 5, no. 3, pp. 220–231, May 1997.

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

[8] K. Cluver and P. Noll, "Reconstruction of missing speech frames using sub-band excitation," in *Proc. IEEE-SP Int. Symp. Time-Frequency and Time-Scale Analysis, 1996*, Jun. 1996, pp. 277–280.

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

[10] R. V. Cox and P. Kroon, "Low bit-rate speech coders for multimedia communication," *IEEE Commun. Mag.*, vol. 34, no. 12, pp. 34–41, Dec. 1996.

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

[12] N. Erdol, C. Castelluccia, and A. Zilouchian, "Recovery of missing speech packets using the short-time energy and zero-crossing measurements," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 3, pp. 295–303, Jul. 1993.

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

[14] V. Hardman, M. A. Sasse, M. Handley, and A. Watson, "Reliable audio for use over the internet," in *Proc. Int. Networking Conf.*, Jun. 1995, pp. 171–178.

[15] 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. Commun.*, vol. COM-29, no. 2, pp. 101–110, Feb. 1981.

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

[17] J. P. Campbell Jr., T. E. Tremain, and V. C. Welch, "The federal standard 1016 4800 bps CELP voice coder," *Digital Signal Process.*, vol. 1, no. 3, pp. 145–155, 1991.

[18] T. J. Kostas, M. S. Borella, I. Sidhu, G. M. Schuster, J. Grabiec, and J. Mahler, "Real-time voice over packet-switched networks," *IEEE Network*, vol. 12, no. 1, pp. 18–27, Jan.-Feb. 1998.

[19] D. Lin, "Real-Time Voice Transmissions over the Internet," M.S. thesis, Univ. Illinois at Urbana-Champaign, 1998.

[20] ——, "Loss Concealments for Low Bit-Rate Packet Voice," Ph.D. dissertation, Univ. Illinois, Dept. Elect. Comput. Eng., Urbana, IL, 2002.

[21] C. Montminy and T. Aboulnasr, "Improving the performance of ITU-T g.729a for VOIP," in *Proc. IEEE Int. Conf. Multimedia and Expo 2000*, vol. 1, 2000, pp. 433–436.

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

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

[24] N. Shacham and P. McKenney, "Packet recovery in high-speed networks using coding and buffer management," in *Proc. IEEE INFOCOM*, May 1990, pp. 124–131.

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

[26] A. S. Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, pp. 1441–1582, Oct. 1994.

[27] L. M. Supplee, R. P. Cohn, and J. S. Collura, "MELP: The new federal standard at 2400 bps," in *Proc. 1997 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, Apr. 1997, pp. 1591–1594.

[28] J. Suzuki and M. Taka, "Missing packet recovery techniques for low-bit-rate coded speech," *IEEE J. Sel. Areas. Commun.*, vol. 7, pp. 707–717, Jun. 1989.

[29] R. C. F. Tucker and J. E. Flood, "Optimizing the performance of packet-switch speech," in *Proc. IEEE Conf. Digital Processing of Signals in Communications*, Apr. 1985, pp. 227–234.

[30] *ITU-T G.723.1 — Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*, Int. Telecommun. Union, 1996.

[31] B. W. Wah and D. Lin, "Method and Program Product for Organizing Data Into Packets," U.S. Patent 6 754 203, Jun. 22, 2004.

[32] ——, "Transformation-based reconstruction for real-time voice transmissions over the internet," *IEEE Trans. Multimedia*, vol. 1, no. 4, pp. 342–351, Dec. 1999.

[33] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 819–829, Jun. 1992.

[34] 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. Acoust., Speech, Signal Process.*, vol. 36, no. 3, pp. 342–348, Mar. 1988.

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

[36] 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. Multimedia*, vol. 3, no. 1, pp. 98–107, Mar. 2001.

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

Benjamin W. Wah (SM'85–F'91) received the Ph.D. degree in computer science from the University of California, Berkeley, in 1979.

He is currently the Franklin W. Woeltge Endowed Professor of electrical and computer engineering and a Professor of the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign, Urbana. Previously, he had served on the faculty of Purdue University, West Lafayette, IN, from 1979 to 1985, as a Program Director at the National Science Foundation in 1988–1989, as Fujitsu Visiting Chair Professor of Intelligence Engineering, University of Tokyo, Japan, in 1992, and McKay Visiting Professor of electrical engineering and computer science, University of California, Berkeley, in 1994. His current research interests are in the areas of nonlinear optimization and multimedia signal processing. He is the Honorary Editor-in-Chief of *Knowledge and Information Systems*. He currently serves on the editorial boards of *Information Sciences*, *International Journal on Artificial Intelligence Tools*, *Journal of VLSI Signal Processing*, *World Wide Web*, and *Neural Processing Letters*.

Dr. Wah was the Editor-in-Chief of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING between 1993–1996. He had chaired a number of international conferences and was the International Program Committee Chair of the IFIP World Congress in 2000. He has served the IEEE Computer Society in various capacities; including Vice President for Publications (1998 and 1999) and President (2001). He is a Fellow of the Society for Design and Process Science, ACM, and AAA. In 1989, he was awarded a University Scholar of the University of Illinois; in 1998, he received the IEEE Computer Society Technical Achievement Award; and in 2000, the IEEE Millennium Medal.

Dong Lin received the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, in 1999 and 2002, respectively.

Following the completion of her Ph.D., she began working for Oracle Corporation, Nashua, NH, in its *inter*Media group. Her interests include speech, audio, and video coding, computer networking, signal processing, communication, cryptography, and security.