Model-Based Multirate Representation of Speech Signals and Its Application to Recovery of Missing Speech Packets

You-Li Chen, *Member, IEEE* and Bor-Sen Chen, *Senior Member, IEEE*

***Abstract*— When the samples of a critically sampled speech signal are lost, objectionable aliasing occurs and perfect recovery of the original speech becomes impossible. In this work, a multirate state-space representation of the autoregressive (AR) speech process is derived to describe the generation of regularly missing-sample speech sequences. Next, a new sample-interpolation algorithm based on the multirate Kalman reconstruction filter is proposed to reduce speech quality degradation caused by packet losses. This method is used together with packet interleaving configuration, thereby simplifying the recovery of missing packets to the interpolation of missing samples. Subjective tests indicate that the proposed Kalman-based sample-interpolation algorithm performs better than the conventional odd-even sample-interpolation procedure for mitigating the effects of random packet losses in 64 kb/s PCM codes. The tolerable packet loss rate P_L, which is strictly input-speech-dependent, can be as high as 10–20% with Kalman interpolation. These observations are based on computer simulations in terms of signal-to-noise ratio (SNR) values, waveform reconstruction plots, error spectral shapes, and summaries of informal listening tests.**

I. Introduction

MULTIRATE operations (decimation and interpolation) can occur concurrently in a digital signal processing system such as in the analysis-synthesis filter bank [1], [2]. In some situations, the decimation (or missing) of speech signals is due to some unexpected cause, e.g., a possible loss of speech samples in the digital transmission or recording systems. If the original speech signals are critically sampled (as in most cases), the decimation operation causes the aliasing effect. The missing samples must be interpolated somehow, such that the quality of the speech signal is not to be sacrificed. The deterministic lowpass interpolation filter is not suitable for critically sampled speech due to the aliasing effect caused by decimation.

In digital communication systems, a speech signal whose bandwidth extends from 0.2 to 3.4 kHz is typically sampled at 8 kHz to avoid aliasing effect while maintaining the bit rate lower. At the 8-kHz sampling rate, the speech signals can be well represented by an autoregressive (AR) generation model through frame-based linear predictive (LP) techniques [3]. The main applications of the AR generation model to speech signals include linear predictive coding (LPC), differential pulse code modulation (DPCM), and speech enhancement based on Kalman filtering. In these applications, the speech signals are first analyzed in the critical sampling rate and then synthesized or filtered in the same rate. For the missing speech cases, another multirate representation of the speech signals may be considered useful to eliminate the aliasing effect.

In this work, the conventional (single-rate) AR generation model of speech signal is transformed to a multirate state-space model to represent an incomplete sequence of speech signals that has at least one sample out of a block of L samples. First, the generation structure of the complete speech signal is described for block of L sampling periods by the dynamic equation in a multirate state-space model. Next, the available samples in each block of speech are described by the state vector through the observation equation. The state of the multirate state-space model is defined to contain both available and missing speech samples. Hence, an estimate of the state has the effect of interpolating the missing speech samples.

When temporary congestion in a packet switching network is considered [4]–[6], the interpolation technique of this work can be applied to reduce speech distortion caused by missing packets. If the Lth packet interleaving procedure is used in the transmitter of the packet switching network, the loss of one speech packet in the receiver will be transformed to the loss of speech samples; however, they are separated by $L-1$ samples. Hence, the recovery of the missing speech packets can be manipulated as a sample-interpolation problem by using the remaining samples in the arrived speech packets. Isolated sample interpolation is easier than missing packet recovery.

The multirate state space model of this work is an useful tool in describing such a missing speech packet system with packet interleaving configuration. The corresponding multirate Kalman reconstruction filter can be used to recover the missing speech packets. A comparison is made of its performance to other (sample-interpolation-based) packet recovery schemes. The performance is described in terms of signal-to-noise ratio (SNR) values, waveform reconstruction plots, error spectral shapes and summaries of informal listening tests.

The rest of this paper is organized as follows. In Section II, the speech signals are modeled as an AR process and reformulated into a block state-space representation. The multirate state-space representation is then derived to describe the gener-

Manuscript received January 18, 1995; revised August 25, 1996. This work was supported by the National Science Council under Contract NSC 83-0404-E-007-042. The associate edtior coordinating the review of this manuscript and approving it for publication was Prof. John H. L. Hansen.

Y.-L. Chen is with the Department of Electronic Engineering, Van-Nung Institute of Technology and Commerce, Chung-Li, Taur-Yuan, Taiwan, R.O.C.

B.-S. Chen is with the Department of Electrical Engineering of National Tsing Hua University, Hsinchu 300 , Taiwan, R.O.C.

Publisher Item Identifier S 1063-6676(97)03187-8.

ation of the regularly missing-speech-sample sequences. Based on packet interleaving techniques of Section III, the Kalman state estimation theory is applied to recover missing speech packets in a packet switching network. The proposed multirate Kalman reconstruction filters are simulated in Section IV to investigate the packet recovery performances. Further properties of the proposed Kalman-based packet recovery scheme are emphasized with a detailed discussion. Concluding remarks are finally made in Section V.

II. Multirate State-Space Representation of Speech Signals

The frame-based AR speech model is introduced in this section to describe the production of the critically sampled speech signals. A block state-space modeling of the packetized speech is then derived to describe the evolution of the speech signals in a block of sampling periods. Finally, a multirate state-space representation is proposed to account for possible losses of the speech packets.

A. AR Generation Model

The composite spectrum effects of radiation and vocal tract and glottal excitation of the speech signals can be accurately represented by a slowly time-varying AR generation model. Usually, frame-based LP techniques are employed to analyze the evolution of the speech production structure through time [3]. For the packet switching network with adequate packet length (4–64 ms), a packet of speech signals $x(k)$ can be regarded as a realization of a stationary AR process, i.e.,

$$x(k)=\sum_{i=1}^{p} a_i x(k-i)+v(k) \tag{1}$$

where $v(k)$ is a zero-mean, white driving noise with covariance $E[v^2(k)]=Q$. The practical value of the AR order p can range from 1 to 16 depending on the application.

The state of a system is generally defined to be a set of internal variables that can represent the effect of all past excitations, and is fundamental in determining the future evolution of the system [7]. For the purpose of multirate state-space representation of the AR process (1), the state must be defined to contain not only the p previous speech signals to describe the AR process model, but also the L previous speech signals to represent the possibly missing samples in a block of L sampling periods. Hence, a state-space description adequate for block representation of the AR speech process (1) would be

$$\begin{aligned}\mathbf{w}(k+1)&=\mathbf{A}\mathbf{w}(k)+\mathbf{b}v(k+1)\\ x(k)&=\mathbf{c}\mathbf{w}(k),\quad k=0,1,2,\cdots\end{aligned} \tag{2}$$

where the state vector $\mathbf{w}(k)=[x(k-N+1)\cdots x(k-1)\ x(k)]^T$, the state dimension $N=\max(p,L)$, and the parametric matrix/vectors $\mathbf{A},\mathbf{b}$, and $\mathbf{c}$ are, respectively, as follows:

$$\mathbf{A}=\begin{bmatrix}0 & & \uparrow & \\ \vdots & \leftarrow & \mathbf{I} & \rightarrow \\ 0 & & \downarrow & \\ a_N & \cdots & a_2 & a_1\end{bmatrix},\quad \mathbf{b}=\begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix},\quad \mathbf{c}=\begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix}^T \tag{3}$$

where $\mathbf{I}$ is a $(N-1)\times(N-1)$ identity matrix. Whenever $N=L$ occurs (i.e., $L>p$), the elements $a_i, i=p+1,\cdots,N$ are equal to zeros.

B. Block State-Space Model

The state-space equation (2) describes the evolution of the speech signals at every sampling time. In some cases, the observation of the speech samples is available in a block version. Hence, another state-space equation that describes the evolution of the speech signals in a block of sampling periods is considered here.

A block state-space description of the AR speech process (1) with block length $L=2$ can be derived from (2) as follows:

$$\begin{aligned}\mathbf{w}(k+2)&=\mathbf{A}^2\mathbf{w}(k)+[\mathbf{A}\mathbf{b}\quad \mathbf{b}]\begin{bmatrix}v(k+1)\\ v(k+2)\end{bmatrix}\\ \begin{bmatrix}x(k)\\ x(k+1)\end{bmatrix}&=\begin{bmatrix}\mathbf{c}\\ \mathbf{c}\mathbf{A}\end{bmatrix}\mathbf{w}(k)+\begin{bmatrix}0\\ \mathbf{c}\mathbf{b}\end{bmatrix}v(k+1)\end{aligned} \tag{4}$$

where $k=0,\ 2,\ 4,\cdots$. By a similar procedure, a block state-space representation of the AR speech process (1) with general block length L would be

$$\begin{aligned}\mathbf{w}(k+L)&=\mathbf{A}^L\mathbf{w}(k)+\mathbf{B}_L\mathbf{v}_L(k)\\ \mathbf{x}_L(k)&=\mathbf{C}_L\mathbf{w}(k)+\mathbf{D}_L\mathbf{v}_L(k)\end{aligned} \tag{5}$$

where $k=0,L,2L,\cdots$. The vectors $\mathbf{x}_L(k)$ and $\mathbf{v}_L(k)$ are block versions of $x(k)$ and $v(k)$, respectively, i.e.,

$$\begin{aligned}\mathbf{x}_L(k)&=[x(k)\quad x(k+1)\cdots x(k+L-1)]^T\\ \mathbf{v}_L(k)&=[v(k+1)\quad v(k+2)\cdots v(k+L)]^T\end{aligned} \tag{6}$$

and the parametric matrices $\mathbf{B}_L,\mathbf{C}_L$ and $\mathbf{D}_L$ are, respectively, as follows:

$$\mathbf{B}_L=[\mathbf{A}^{L-1}\mathbf{b}\cdots\mathbf{A}\mathbf{b}\quad \mathbf{b}],\quad \mathbf{C}_L=\begin{bmatrix}\mathbf{c}\\ \mathbf{c}\mathbf{A}\\ \vdots\\ \mathbf{c}\mathbf{A}^{L-1}\end{bmatrix}$$

$$\mathbf{D}_L=\begin{bmatrix}0 & & & & 0\\ \mathbf{c}\mathbf{b} & 0 & & & \\ \mathbf{c}\mathbf{A}\mathbf{b} & \mathbf{c}\mathbf{b} & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \\ \mathbf{c}\mathbf{A}^{L-2}\mathbf{b} & \cdots & \mathbf{c}\mathbf{A}\mathbf{b} & \mathbf{c}\mathbf{b} & 0\end{bmatrix}. \tag{7}$$

The observation $x(k)$ in (5) is considered as available in a full block version $\mathbf{x}_L(k)$ and there are now L samples by which the desired signal processing can be performed.

In the block state-space representation (5), both the state dynamic equation and the output observation equation contain the block driving noise $\mathbf{v}_L(k)$. This would result in a "correlating-noise" state-space description of the speech signals. This problem can be solved by the concept of the augmented state. Let the augmented state vector $\mathbf{z}(k)$ be defined by $\mathbf{z}(k)=[\mathbf{v}_L^T(k)\ \ \mathbf{w}^T(k)]^T$; then the standard block state-space model is

$$\begin{aligned}\mathbf{z}(k+L)&=\mathbf{F}\mathbf{z}(k)+\mathbf{G}\mathbf{v}_L(k+L)\\ \mathbf{x}_L(k)&=\bar{\mathbf{H}}\mathbf{z}(k),\quad k=0,L,2L,\cdots\end{aligned} \tag{8}$$

where the parametric matrices $\mathbf{F}$ (state transition matrix), $\mathbf{G}$ (block input matrix) and $\bar{\mathbf{H}}$ (block output matrix) are, respectively, as follows:

$$\mathbf{F} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{B}_L & \mathbf{A}^L \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}, \quad \bar{\mathbf{H}} = [\mathbf{D}_L \quad \mathbf{C}_L]. \tag{9}$$

In (8), the first equation describes the state transition of the AR speech process in a block of L sampling periods. Meanwhile, the second equation represents the block observation of the speech samples with complete measurements, i.e., no speech samples are missing to the observer.

C. Multirate State-Space Representation

The speech samples $x(k)$ of (1) are considered now to be available at the rate M out of L samples ($0 < M < L$), where M is constant over a segment of speech signals. Also, the missing samples are assumed to be lost at regular positions. The observation equation of the block state-space model (8) must be modified to adapt to this decimated measurement condition. If the actual measurement vector $\mathbf{y}(k)$ consists of a block of decimated speech samples $x(k)$, then it can be expressed as

$$\mathbf{y}(k) = \mathbf{E}\mathbf{x}_L(k) \tag{10}$$

where $\mathbf{E}$ is an $M \times L$ almost-zero matrix whose ijth element is set to one if jth element of the speech signal vector $\mathbf{x}_L(k)$ is the ith available observation of the actual measurement vector $\mathbf{y}(k)$. An illustrative example with $L = 4$ and $M = 3$ is provided as follows:

$$\begin{bmatrix} x(k) \\ x(k+2) \\ x(k+3) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x(k) \\ x(k+1) \\ x(k+2) \\ x(k+3) \end{bmatrix}$$

where the speech signals $x(k+1), k = 0, 4, 8, \cdots$ are regularly lost at the receiver.

With the combination of the block representation equation (8) and the missing measurement equation (10), the multirate state-space model of the AR speech process (1) becomes

$$\begin{aligned} \mathbf{z}(k+L) &= \mathbf{F}\mathbf{z}(k) + \mathbf{G}\mathbf{v}_L(k+L) \\ \mathbf{y}(k) &= \mathbf{H}\mathbf{z}(k), \quad k = 0, L, 2L, \cdots \end{aligned} \tag{11}$$

where the multirate output matrix $\mathbf{H} = \mathbf{E}\bar{\mathbf{H}}$ under the missing-speech-sample condition. Finally, by taking expectation on outer product of the block driving noise $\mathbf{v}_L(k)$ at different time points and using the zero-mean, white characteristics of the input driving noise $v(k)$ in (1), the block driving noise $\mathbf{v}_L(k)$ possesses the following white and diagonal covariance structure

$$E[\mathbf{v}_L(k)\mathbf{v}_L^T(l)] = Q\mathbf{I}_{L\times L}\delta(k-1) = \mathbf{Q}_L\delta(k-l) \tag{12}$$

where $\mathbf{I}$ is an $L \times L$ identity matrix and the time indices $k, l = 0, L, 2L, \cdots$. The multirate state-space model (11) is a special case of the well-known Kalman state-space form [8], [9], where the output observation $\mathbf{y}(k)$ is a perfect measurement vector, i.e., the measurement noise of the dynamic system (11) is zero.

III. Application to Recovery of Missing Speech Packets

Packetized speech has found application in telecommunication systems with combined voice and data services [10], [11]. Missing packets are a major cause of impairment in packet voice networks. Whenever a packet discarding occurs, the missing speech must be recovered somehow to mitigate the performance degradation of the reconstructed speech.

A. Packet Recovery Techniques

When one of two consecutive packets is lost, the discarded samples after depacketization distribute themselves differently according to an arrangement of the codewords in a packet. A variety of missing packet recovery techniques for PCM-coded speech have been investigated in [12]–[18] to reduce the degradation caused by the missing speech packets or to increase the maximum tolerable missing packet rate.

In the sample interpolation schemes [12]–[14], the samples in a certain segment are interleaved into several groups and packetized separately. If packet discarding occurs, the samples in the lost packet are reconstructed by sample interpolation using the remaining samples in the arrived packets. Jayant and Christensen's sample interpolation procedure [12] places odd and even numbered samples into consecutive packets. The odd-even sample-interpolation procedure mitigates the missing packet effects at the cost of an increased decoding delay. These authors report a tolerable packet loss rate of 5–10% based on informal listening tests.

In waveform substitution techniques [15]–[16], the output samples in a discarded packet are replaced by the previous waveform segment. This technique uses a template that is just before the missing packet. The waveform is selected by a pattern-matching procedure from the output speech segments already available at the receiver. The quality of waveform substitution is better than that of zero-amplitude stuffing because of the reduction in the waveform discontinuity. However, incorrect waveform substitution can occur for transient speech segments such as voiced-to-unvoiced transitions, and vice versa. The maximum tolerable packet loss rate reported in [16] is 10% for a target mean opinion score of 3.5.

The least significant bit (LSB) dropping scheme [17] discards bits from the LSB in the codewords, since the LSB's of the quantizer output have less significant effects on the reconstructed speech quality. It is reported in [17] that the mean opinion score (MOS) ratings of the waveform substitution technique [15] and the sample interpolation scheme [12] are approximately the same. Meanwhile, the subjective quality of the LSB-dropping scheme is somewhat better than the other two methods. However, priority discarding must be carefully performed in the LSB-dropping configuration to avoid the loss of the MSB packets.

These packet-recovery techniques produce different types of distortion. LSB-dropping produces amplitude-modulated quantization noise, the waveform substitution produces beep and chirplike distortion, and odd–even sample interpolation produces aliasing distortion. Because of these phenomena, most careful consideration should be given to the formulation of subjective tests between the different packet recovery schemes. In addition to different types of distortion, the reported maximum tolerable missing packet rates do not have an obvious difference with these techniques. Advanced efforts must be taken in each of these techniques to improve the worst-case tolerable ability. The following development is devoted to improve the packet recovery performance of the sample-interpolation procedure.

Remark A: Another waveform substitution technique based on the short-time energy and zero-crossing information was developed in [18] for missing packet recovery. The application is mainly geared toward packetized voice systems that employ digital speech interpolation (DSI). The reconstruction technique mitigates the missing packet effects at the cost of a considerable side information overhead (short-time energy and zero-crossing parameters). The zero-amplitude stuffing was used to establish a reference of mean opinion score in their subjective testing. The authors reported tolerable packet loss ratios up to 40% based on informal listening tests. However, the reported tolerable packet loss ratios may be questionable. The kept scores about the zero-amplitude stuffing were obviously higher than that of [16] and [17] for the same packet loss ratios. Specifically, the reconstruction quality of the zero-amplitude stuffing at $P_L = 20\%$ was claimed to be "fair" in [18] but "unsatisfactory" in [16] and [17]. Since the reported results were unclear and the techniques were developed particularly for DSI systems with extra overhead, the work of [18] is excluded from the above general discussion and differently considered here.

Remark B: Other than simple pulse code modulation (PCM) systems, more sophisticated speech coding techniques—such as adaptive differential PCM and code excited linear prediction (CELP)—may be applied in a packet-switching network. In these systems, the information contained in a packet is the "white" innovation after source compression. Hence, performance degradation caused by packet losses is more serious in these systems. A possible solution to withstand packet losses in a DPCM-based interleaving-packetized speech system has been proposed in [19]. The concepts of multiple-description source coding [20], [21] were used to organize the DPCM encoder and decoder. The encoder was designed to leave some correlation in the quantized prediction error sequence. The task of the decoder is to receive information over two channels in such a way that a good reproduction of the source sequence is obtained when both channels work, and that if either channel breaks down, a minimum degradation in performance is obtained. Although the packetization configuration of the work [19] is different from that of embedded DPCM [17], the concept of the multiple description coding was used in both speech compression systems. The main difference between them is that priority discarding must be performed in the embedded DPCM to avoid the loss of the most significant bit (MSB) packets of the quantized prediction errors. The performances of the coded speech without missing packets, the goodness of the reconstructed speech under missing packets, and other related issues must be further addressed for both configurations to make a detailed comparison between them.

B. Wiener-Based Sample Interpolation Procedure

Interleaving methods [22] are well-known digital transmission techniques for converting burst errors to separate errors by reordering a digital code sequence. Therefore, packet interleaving can be introduced to prevent phoneme or syllabic losses caused by packet losses, as well as to facilitate reconstruction of missing portions of speech. Restated, the packet recovery problem can be simplified to a more tractable sample interpolation problem. A proper sample interpolation method should be selected to provide sufficient speech quality.

In [12], an adaptive sample interpolation scheme was used together with the 2th (odd-even) packet interleaving configuration to reduce speech quality degradation caused by packet losses. The interpolation method used a second-order Wiener filter with forward parameter adaptation. The interpolation coefficients are based on the first- and second-order autocorrelation functions of the original speech packets, i.e.,

$$\begin{aligned} \hat{x}(k) &= \alpha x(k-1) + \beta x(k+1) \\ \alpha &= \beta = R_{xx}(1)/[1 + R_{xx}(2)] \end{aligned} \tag{13}$$

where $R_{xx}(m)$ is the normalized autocorrelation function of the original speech segment, i.e., $R_{xx}(m) = E[x(k)x(k+m)]/E[x^2(k)]$. The values of α (or equivalently β) were assumed in [12] to be included as part of side information in the headers of odd and even packets. In addition, the interpolation values are updated once for each $2B$ block.

Although the results in [17] have reported that the subjective quality of the LSB-dropping method is somewhat better than the sample interpolation scheme of [12], two factors cause this conclusion to be vague. First, subjective tests reported in [12] indicate that packet lengths most robust to losses are in the range 16–32 ms for the odd-even sample interpolation procedure. However, the packet length used in the listening experiment of [17] was fixed at 4 ms. Second, whether the interpolation coefficients used in [17] are the optimum values derived from (13) or the fixed value of $\alpha = \beta = 1/2$ remain unclear. These two factors may aptly influence the perceptual effect of the sample interpolation procedure [12].

Two additional causes merit proceeding with an investigation of the sample interpolation procedure. First, the packet-interleaving technique simplifies the recovery of missing packets to the interpolation of missing sample values. An interesting question is whether a higher packet interleaving factor L improves the final perceptual results. Second, the second-order Wiener interpolation method has been used in [12] to recover the missing speech packets. Is there another sophisticated sample-interpolation procedure better than the simple one of [12]? With these two problems in mind, the following Wiener- and Kalman-based sample interpolation schemes are derived

to recover the missing speech packets with the general Lth packet interleaving networks.

The design of a Wiener-based sample interpolation filter depends on the packet interleaving factor L, packet receiving factor M and filter order S. An illustrative example with $L = 4$, $M = 2$, and $S = 4$ is provided in Appendix A. By the same rules, similar results can be derived for other combinations of L, M and S. A noteworthy result is that a Sth-order Wiener interpolation filter would require $(L - M) \times S$ coefficients to perform interpolation, i.e., each lost packet needs S interpolation coefficients. The Wiener interpolation technique of Appendix A is a natural extensions of the Jayant's odd–even sample interpolation procedure [12]. They are introduced to make a comparison with the Kalman interpolation techniques employed in this work.

C. Kalman-Based Sample Interpolation Procedure

When the Lth packet interleaving configuration is used on the transmitter, the speech packets may be received by the rate M out of L at the receiver. Under this case, the multirate state-space representation (11) presents a suitable model to the missing-speech-packet network. The samples in the lost packets can be reconstructed by sample interpolation using the remaining samples in the arrived M packets.

Since the multirate state-space model (11) can be considered as a conventional state-space dynamic system with multi-input and multi-output (MIMO), the state $\mathbf{z}(k)$ can be optimally estimated in the minimum mean-square-error sense by using the Kalman state estimator [8], [9]. Thus, the optimal state estimate $\hat{\mathbf{z}}(k)$ based on the received measurement vector $\mathbf{y}(k)$ can be obtained by the following Kalman filter equations:

$$\hat{\mathbf{z}}(k) = [\mathbf{I} - \mathbf{K}(k)\mathbf{H}]\mathbf{F}\hat{\mathbf{z}}(k - L) + \mathbf{K}(k)\mathbf{y}(k) \tag{14}$$

where $\mathbf{I}$ is an identity matrix with adequate dimension and the time scale evolves by $k = 0, L, 2L, \cdots$. The Kalman gain $\mathbf{K}(k)$ in (14) must be recursively updated as follows:

$$\begin{aligned} \mathbf{K}(k+L) &= \mathbf{P}(k+L \mid k)\mathbf{H}^T[\mathbf{H}\mathbf{P}(k+L \mid k)\mathbf{H}^T]^{-1} \\ \mathbf{P}(k+L) &= \mathbf{F}\mathbf{P}(k \mid k)\mathbf{F}^T + \mathbf{G}\mathbf{Q}_L\mathbf{G}^T \\ \mathbf{P}(k+L \mid k+L) &= [\mathbf{I} - \mathbf{K}(k+L)\mathbf{H}]\mathbf{P}(k+L \mid k) \end{aligned} \tag{15}$$

where $\mathbf{P}(k+L \mid k)$ and $\mathbf{P}(k \mid k)$ are prediction and filtering state error covariance matrices, respectively. An adequate estimate of the initial state $\mathbf{z}(-L)$ would be

$$\hat{\mathbf{z}}(-L) = [0 \cdots 0 \quad x(-L-N+1) \cdots x(-L)]^T \tag{16}$$

where $x(-L-N+1), \cdots, x(-L)$ are the N speech samples before the present segment, which are available or have been estimated in the previous segment. Furthermore, a reasonable estimate of the initial state covariance matrix $\mathbf{P}(-L \mid -L)$ would be

$$\mathbf{P}(-L \mid -L) = \begin{bmatrix} Q\mathbf{I}_{L\times L} & 0 \\ 0 & \varepsilon\mathbf{I}_{N\times N} \end{bmatrix} \tag{17}$$

where Q is the covariance of the driving noise $v(k)$ and ε is an adequate value proportional to the average power of $[x(-L-N+1) \cdots x(-L)]$.

By the definition of (8), the state vector $\mathbf{z}(k)$ is composed of the noise vector $\mathbf{v}_L(k) = [v(k+1) \cdots v(k+L)]^T$ and the speech vector $\mathbf{w}(k) = [x(k-N+1) \cdots x(k)]^T$. Since the dimension N of the speech vector $\mathbf{w}(k)$ is $N = \max(p, L)$, the state set $\{\mathbf{z}(k), k = 0, L, 2L, \cdots\}$ in (11) contain all of the samples in the speech packets. Hence, after each of the state estimates $\hat{\mathbf{z}}(k), k = 0, L, 2L, \cdots$ is obtained, the optimal sample interpolation (or optimal packet recovery) of the speech segments can be obtained as follows:

$$\begin{bmatrix} \hat{x}(k-s-L+1) \\ \vdots \\ \hat{x}(k-s) \end{bmatrix} = [\mathbf{0}_{L\times(N-s)} \quad \mathbf{I}_{L\times L} \quad \mathbf{0}_{L\times s}]\hat{\mathbf{z}}(k) \tag{18}$$

where s is a lag factor $(N - L \geq s \geq 0)$. It is obvious from (18) that the Kalman state estimator (14) provides the optimal fixed-lag smoothed estimates of the missing speech samples, i.e., $\hat{x}(i) = E[x(i) \mid \mathbf{y}(0), \mathbf{y}(L), \cdots, \mathbf{y}(k)]; i = k-s-L+1, \cdots, k-s;\ k = 0, L, 2L, \cdots$. The performance of the fixed-lag estimates $\hat{x}(k)$ in (18) will be better than the filtering case, i.e., $s = 0$. As the lag increases, the estimation error variance decreases due to the information provided by the additional data.

The derivation of the above multirate state-space model (11) and the Kalman reconstruction filter (14)–(18) relies on the prior knowledge of the AR speech process (1). The problem associated with the above Kalman-based sample-interpolation procedure entails how to obtain the AR parameters a_i and the covariance Q of the driving noise $v(k)$. Theoretically, the optimum solution to this problem would involve computing these parameters from the original speech segments as well as including them into the packet headers as part of side information. However, such a *forward* adaptation of these parameters requires extra bits. Another practical solution is the *backward* computation of these parameters from the received incomplete speech packets. A simple but effective algorithm for the design of a Kalman interpolation filter with a backward parameter adaptation is described in the following subsection. Similar design rules can be applied to the construction of a Wiener interpolation filter with backward parameter adaptation.

D. Kalman-Based Packet Recovery Algorithm

The application of the multirate state-space model (11) toward the recovery of the missing speech packets is summarized in the following. For the Lth packet interleaving configuration, a speech segment with LB samples is interleaved into L packets each of which has B samples. Let $M(j)$ be the receiving factor which represents the value of the received packets in the jth speech segment. Clearly, the constraint $0 \leq M(j) \leq L$ holds for all j. If all of the L packets in the jth speech segment are received, i.e., $M(j) = L$, they are ready to be depacketized. Whenever all of the L companion packets are lost, i.e., $M(j) = 0$, zero-amplitude stuffing is assumed here for the entire segment of LB samples. If the missing-packet case occurs but the packets are not completely lost, i.e., $0 < M(j) < L$, the Kalman-based sample interpolation procedure is used to recover the missing speech packets through the following steps.

Step 1—Selection of Speech Segment: If the receiving factor $M(j-1)$ of the previous speech segment is greater than $M(j)$ of the present segment, then the previous segment of speech samples (which is available or has been estimated) is used to estimate the AR parameters a_i of the present speech segment. This is a reasonable assumption, since the speech signal is a short-time stationary process [3]. If $M(j-1)$ is smaller than or equal to $M(j)$, then the linearly interpolated speech samples of the present missing-packet segment are used to estimate the AR parameters a_i. This estimate is generally effective and especially suitable for the highly correlated signal such as the voiced speech segment.

Step 2—AR Parameter Estimation: By using the autocorrelation method of framed-based LP techniques [3], the AR parameters a_i of the AR speech process (1) can be estimated by solving the following normal equation ($\Phi \mathbf{a} = \varphi$):

$$\begin{bmatrix} \phi(0) & \cdots & \phi(p-1) \\ \vdots & & \vdots \\ \phi(p-1) & \cdots & \phi(0) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \phi(1) \\ \vdots \\ \phi(p) \end{bmatrix} \tag{19}$$

where $\phi(i)$ is the estimated autocorrelation function of the present speech segment, i.e.,

$$\phi(i) = \frac{1}{LB} \sum_{k=1}^{LB} \bar{x}(k-i)\bar{x}(k) \tag{20}$$

where the quantity $\bar{x}(k)$ is the speech samples selected in Step 1. Finally, the estimated covariance Q of the driving noise $v(k)$ is obtained by

$$Q = \phi(0) - \mathbf{a}^T \varphi \tag{21}$$

Step 3—Kalman Sample Interpolation: The parametric matrices of the multirate state-space model (11) are established according to the estimated AR parameters a_i of (19) and the estimated noise covariance Q of (21). Finally, the samples of the missing speech packets are interpolated using the Kalman state estimator (14)–(17) followed by sample reconstruction equation (18).

The autocorrelation parameter estimation method of Step 2 is used to generate the AR parameters a_i, thereby guaranteeing the stability of the AR speech process [23]. Hence, the Kalman state estimator (14) is asymptotically stable [9], and the stability of the interpolation estimate (18) is ensured. Furthermore, since the autocorrelation matrix Φ in (19) is a symmetric Toeplitz matrix, the Durbin's recursive procedure [23] can be used to efficiently solve the normal equation (19).

IV. Simulation Results

Subjective tests of the proposed Kalman packet recovery algorithms are of primary concern in this section. The five sentence-length Mandarin utterances of Appendix B were used as illustrative input speech throughout the following experiments. There were not obvious silent gaps in the illustrative utterances and, hence, the speech/silence discrimination was not put into the following simulations. The experiments were performed to study reconstructed speech quality as a function of interleaving factor L and probability of loss P_L with different sample-interpolation-based packet recovery schemes.

The 2th and 4th packet interleaving configurations were used in the simulated transmitter. Their associated packet lengths are 16 ms and 8 ms, respectively. Thus, the decoding delay is fixed at 32 ms (2×16 ms and 4×8 ms) in both interleaving configurations. An additional parameter in the proposed Kalman interpolation procedure is the order p of the AR speech process (1). It was set to be 4 as a compromise of the algorithm complexity and the reconstruction performance.

A. SNR's

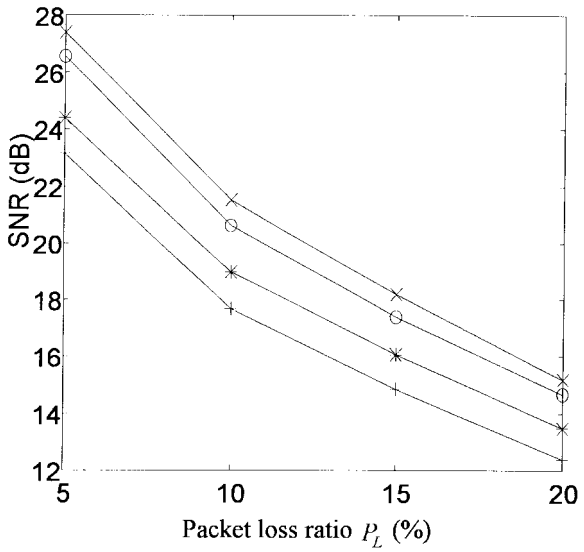
The reconstruction performance is evaluated in this subsection using SNR results. All SNR values, including the special versions SNRL and SNRSEG, to be defined subsequently, are obtained as ratios of appropriately averaged values of speech power x^2 and error power $(\hat{x} - x)^2$.

Four sample-interpolation schemes were simulated in the 2th packet interleaving configuration: linear interpolation, Jayant's interpolation [12], and Kalman interpolation with forward and backward parameter adaptation. The linear interpolation is a straightforward scheme that is used as a reference of the simulations. Jayant's interpolation (13) is a special case of the Wiener interpolation methods. The forward Kalman interpolation provides the optimum packet recovery performance based on AR speech modeling (1). The backward Kalman interpolation procedure was simulated to investigate the packet recovery performance with the proposed parameter estimation algorithm. The smoothing lag s in (18) was set to 2 in both of the Kalman interpolation techniques.

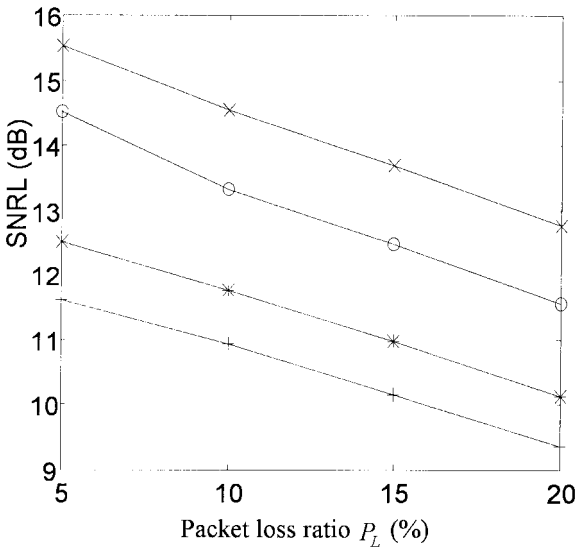
Fig. 1(a) indicates that the reconstruction performances of the linear interpolator are the worst of the four interpolation methods, since the linear interpolation is in no way optimum for the speech signals. The Jayant's interpolation improves the reconstruction performance up to 1 dB over linear interpolation due to a signal model (autocorrelation function model) that is somewhat related to the speech signals. The performance improvements of the backward Kalman interpolation over Jayant's interpolation are about 1–2 dB. This is owing to the exact interpolation of the Kalman filtering technique according to the AR structure of the speech signals.

An objective measure that more effectively isolates the effects of coding and interpolation noise is SNRL, the signal-to-noise ratio obtained by averaging SNR (dB) values over lost packets. Fig. 1(b) compares the four interpolation schemes of Fig. 1(a) on the basis of SNRL. Notably, unlike SNR, SNRL is not a significant function of P_L as expected. The backward Kalman interpolation provides a good reconstruction performance, the SNRL gains are about 1.5–2 dB over Jayant's interpolation.

Five sample-interpolation schemes were simulated in the fourth packet interleaving configuration: linear interpolation, the sixth-order backward and forward Wiener interpolation, the fourth-order backward and forward Kalman interpolation. For the Kalman-based sample interpolation procedures, the smoothing lag s in (18) was set to zero, since p (AR order) $= L$

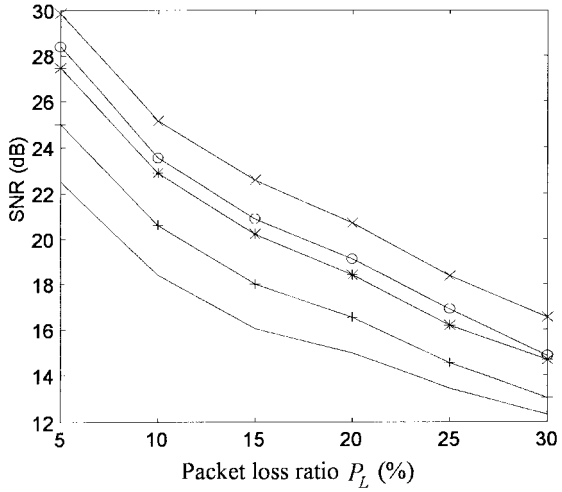


(a)

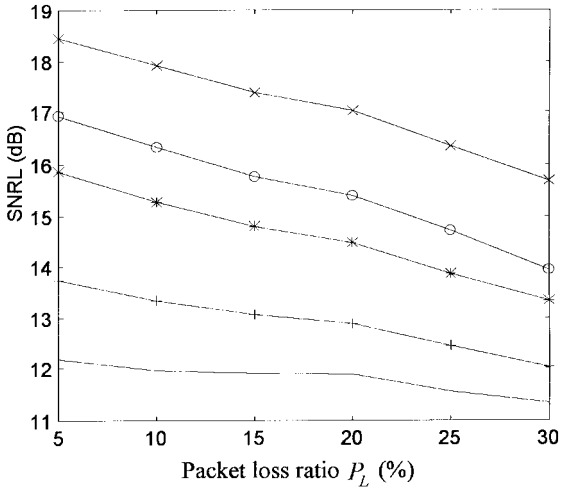


(b)

Fig. 1. SNR's versus probability of packet loss for four sample interpolation schemes. $\times$: Kalman interpolation with forward parameter adaptation. $\circ$: Kalman interpolation with backward parameter adaptation. $*$: Jayant's interpolation. $+$: Linear interpolation. Packet interleaving factor $L = 2$: (a) SNR; (b) SNRL.



(a)



(b)

Fig. 2. SNR versus probability of packet loss for five sample interpolation schemes. $\times$: Kalman interpolation with forward parameter adaptation. $\circ$: Kalman interpolation with backward parameter adaptation. $*$: Wiener interpolation with forward parameter adaptation. $+$: Wiener interpolation with backward parameter adaptation. —: Linear interpolation. Packet interleaving factor $L = 4$: (a) SNR; (b) SNRL.

(interleaving factor) $= 4$ and, hence, no additional state elements can be provided as smoothed estimates.

Fig. 2(a) and (b) reveal that the SNR or SNRL differences between the simplest (linear) and the most sophisticated (forward Kalman) interpolation schemes range from 4–7 dB. For the backward parameter adaptation configurations, the Kalman interpolation method possesses about 2–3 dB of SNR and SNRL improvements over Wiener interpolation technique. An interesting observation is that the backward Kalman interpolation performs better than the *forward* Wiener interpolation.

Fig. 3(a) and (b) shows SNR (SNRL) versus P_L plots for the 2th and 4th packet interleaving configurations in an illustration to observe the effect of the packet interleaving factor L. For the second interleaving configuration, the interesting interpolation schemes were the Jayant's and backward Kalman interpolation procedures. For the fourth interleaving configuration, the backward Wiener and Kalman interpolation schemes were used to make a comparison with the second interleaving configuration.

Two observations can be made from Fig. 3(b)–(d). First, for the same type of interpolation schemes (Kalman versus Kalman, Wiener versus Jayant), the fourth interleaving configurations provide significant improvements over the second interleaving configuration for the reconstruction performance. Second, the lost-packet SNRL in the second interleaving configuration is more sensitive to the packet loss ratio P_L than the fourth interleaving configuration. These phenomena are mainly contributed by the packet interleaving factor L and is accounted for in Remark 2 of the subsection D.

B. Waveform Reconstruction Plots

Fig. 4 demonstrates the benefits of Kalman interpolation by means of waveform plots and segment-specific SNR values

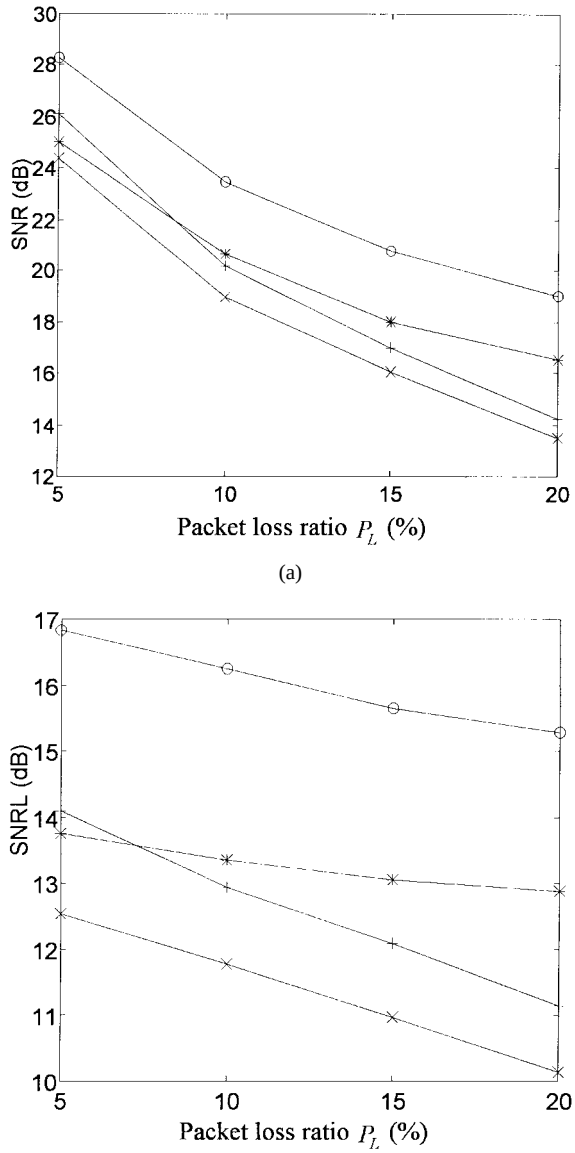


Fig. 3. Signal-to-noise ratio versus probability of packet loss for four sample interpolation schemes. ∘: Kalman interpolation with backward parameter adaptation ($L = 4$). ∗: Wiener interpolation with backward parameter adaptation ($L = 4$). +: Kalman interpolation with backward parameter adaptation. ($L = 2$). ×: Jayant's interpolation ($L = 2$). (a) SNR; (b) SNRL.

(SNRSEG, in dB). The packet receiving factor M in these illustrations is one, and the second packet interleaving configuration is used in the simulations. Each waveform in the illustrations is fixed at 15 ms long (120 samples, at 8 kHz). Fig. 4(a) is the waveform of the original speech segment. Fig. 4(b)–(d) refer, respectively, to the reconstructed versions of the waveform (a) by using linear, Jayant's, and backward Kalman interpolation procedures.

The Kalman interpolation performs better than the other two interpolation methods, as reflected by segment-specific SNR values and waveform details in the illustrations. The abrupt changes of the original speech signals are retained in the reconstruction of Kalman interpolation, but have been smoothed in reconstructions of linear and Jayant's interpolations. Subjective effect of the smoothed reconstruction is a

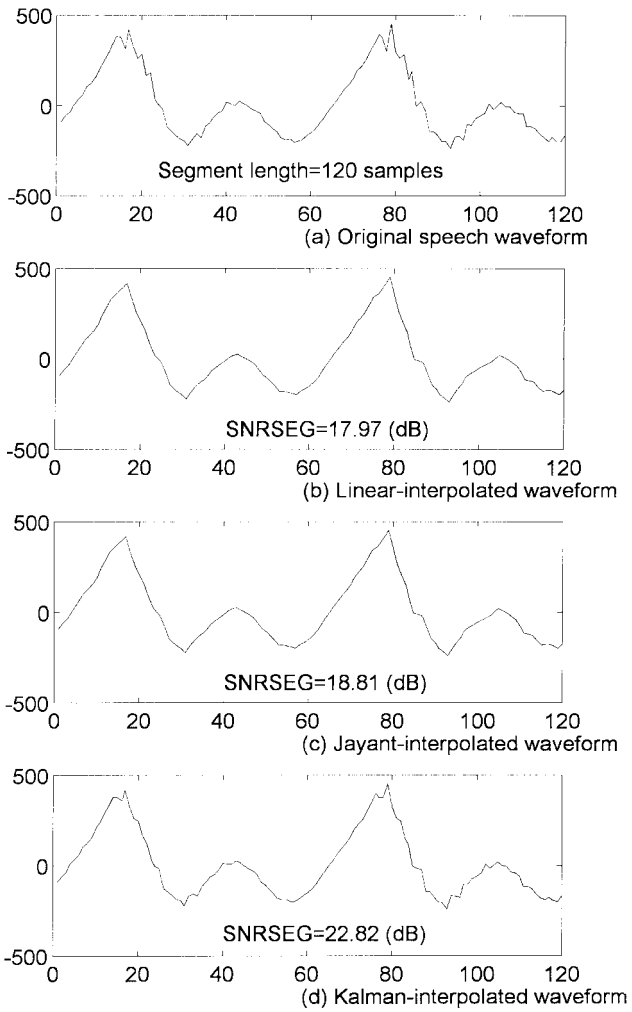


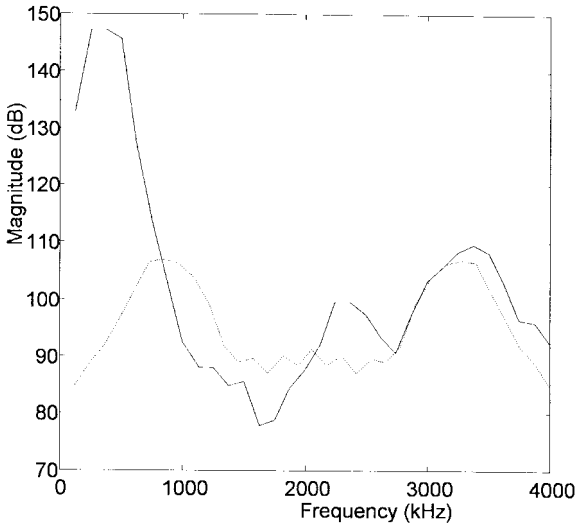
Fig. 4. Reconstruction of speech segment in odd–even packet interleaving-configuration. All odd samples are assumed lost in the packet transmission of (a).

deeper and distorted tone, which is more sensible for female speech.

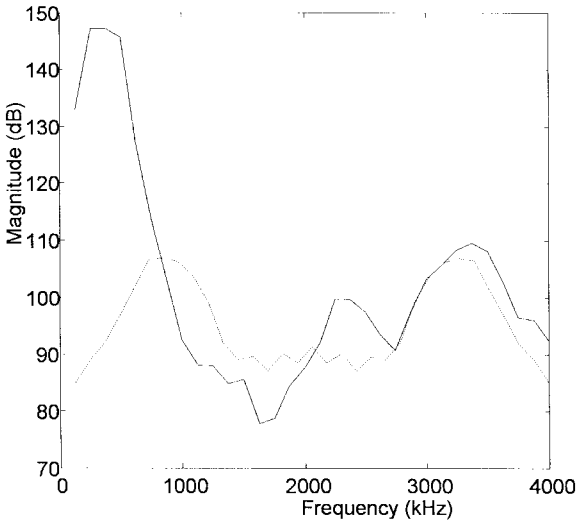
Fig. 5 shows the spectral shapes of the speech signal and reconstruction errors in Fig. 4. For the linear and Jayant's interpolation schemes, the error spectral magnitudes are somewhat comparable with the signal spectral magnitude, and that is particularly obvious in the speech spectral valleys and in the high-frequency band. The phenomenon of large high-frequency error is primarily due to the smoothing effect inherent in these two interpolation schemes. For the Kalman interpolation technique, the reconstruction error spectrum is well masked by the signal spectral envelope.

C. Results of Informal Listening Tests

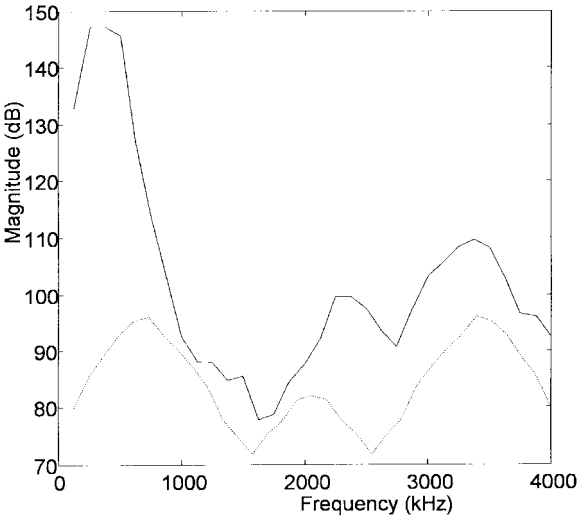
Based on subjective listening tests of the authors and their colleagues, the results obtained from Figs. 1–5 are well confirmed by differences in perceived quality of corresponding output samples. An informal listening test was carried out to assess the subjective quality of packetized PCM speech against packet losses. The missing packet ratios of 0, 5, 10, 15, and 20% were simulated via the following five methods:



(a)



(b)



(c)

Fig. 5. Spectral shapes of the speech signal (real lines) and reconstruction errors (dotted lines) in Fig. 4. (a) Linear interpolation error. (b) Jayant's interpolation error. (c) Kalman interpolation error.

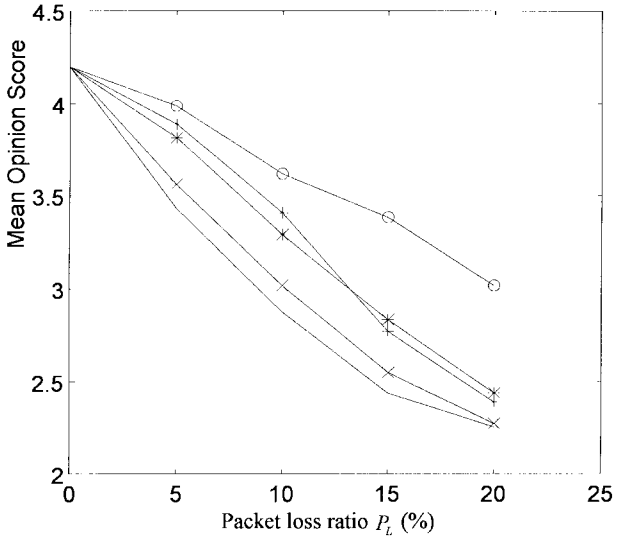


Fig. 6. MOS as a function of missing packet ratio P_L for five sample interpolation schemes. ∘: Kalman interpolation with backward parameter adaptation (L =4). +: Kalman interpolation with backward parameter adaptation (L = 2). ∗: Wiener interpolation with backward parameter adaptation (L = 4). ×: Jayant's interpolation (L = 2). —: Linear interpolation (L = 2).

backward Kalman and Wiener interpolation procedures with the fourth interleaving configuration; linear, Jayant's, and backward Kalman interpolation procedures with the second interleaving configuration. The listeners gave opinion scores for the processed speech arranged in a random order. The vote categories "excellent," "good," "fair," "poor," and "unsatisfactory" are represented by the category numbers 5, 4, 3, 2, and 1, respectively. The MOS ratings of the missing packet recovery methods are shown in Fig. 6. This figure displays a clear ranking of the effectiveness of the five packet recovery techniques.

Jayant's studies [12] showed that speech quality was sufficient up to a packet loss rate of P_L = 5–10% with the odd–even Wiener interpolation procedure. In the simulation results of Fig. 6, mean opinion scores with Jayant's interpolation are 3–3.6 for P_L = 5–10%. These values are considered to be the tolerable range of MOS to provide sufficient speech quality. Based on this assumption, the following observations can be drawn from Fig. 6. First, the (backward) second Kalman and fourth Wiener interpolation procedures provide nearly the same perceptual quality for the speech utterances used in this work. Both of them perform better than the Jayant's interpolation scheme and can raise the tolerable packet loss rate up to 8–13%. The SNR and SNRL gains of the fourth Wiener interpolation over the second Kalman interpolation (refer to Fig. 3) are not confirmed by the listening tests. Second, the fourth (backward) Kalman interpolation procedure is most robust to packet loss ratio, and provides a much higher speech quality than the other interpolation schemes. It can raise the tolerable level of P_L to significantly high values (10–20%) due to its sophisticated interpolation and the higher interleaving factor (L = 4).

D. Further Remarks

Objective SNR evaluating and subjective MOS tests of the packet loss effect have been discussed in the above

subsections. It is concluded that the Kalman-based sample interpolation procedure provides the optimum packet recovery performance among several methods. Further properties of the proposed Lth interleaving packetization and Kalman-based packet recovery scheme are emphasized in the following.

Remark 1: Missing packets were distributed randomly in time in our simulations. In anticipated applications, other patterns of packet loss occurred. If, in practice, the likelihood of consecutive missing packets can be reduced, a higher quality would result. Conversely, if the number of consecutive missing packets is higher than in a random distribution, speech quality would be lower than our tests indicate. Particularly, the damage of contiguous packet losses on backward-adaptive (Kalman and Wiener) interpolations is larger than that of forward-adaptive interpolations. The performance degradation is caused by using deficient statistical information of the incomplete speech samples to estimate the interpolation parameters.

Remark 2: The packet recovery performance of the Kalman-based sample interpolation procedure can be improved by raising the interleaving factor L. Higher interleaving factors will more uniformly distribute the missing samples to speech segments under the assumption of random and statistically independent missing packets. The effects of the uniformly distributed missing samples are

1) lower idle probability $(1-P_L)^L$ for the Kalman reconstruction filter;
2) less zero-amplitude stuffing (with probability P_L^L) for output speech packets;
3) generally more true samples in a speech segment to estimate the AR parameters;
4) generally more speech measurements ($M \times B$, in samples) to interpolate the missing samples ($(L-M) \times B$, in samples).

All of these effects demonstrate the benefits of a higher packet interleaving factor for the sample-interpolation methods.

Remark 3: When a packet transmission network sends packets fluently, a practical consideration involving the upper bound of the packetization configuration $L \times B$ is the tolerable decoding delay. If a packet network suffers from packet discarding due to temporary congestion, two perceptual considerations are important to the selection of the packetization configuration $L \times B$: i) the number of speech crackles per second, which is inversely proportional to $L \times B$; and ii) the probability of totally losing a phoneme, which increases rapidly for a large $L \times B$ value.

Remark 4: When a higher value of interleaving factor L is used in the packetization, a smaller value of packet length B should be adopted to avoid increased decoding delay and a higher probability of totally losing a phoneme. In a practical packet network, the packet length B is a significant factor of the side information overhead. Hence, the packet length B must be kept large enough to avoid system overload due to packet headers. In conclusion, the combination of a higher but bounded value of interleaving factor L and a smaller but large enough value of packet length B would be beneficial to the overall system performances. The speech packetization of $L \times B = 32$ ms (2×16 ms and 4×8 ms) were used in the simulations. It is a compromise that strikes a tradeoff between all of the practical considerations in Remarks 2–4.

Remark 5: At first glance, the Kalman filter equations (14)–(15) involve matrix/vector multiplications/additions and seem complex. However, this is just an illusion due to their expanded profiles. First, since the coefficient matrices/vectors $\mathbf{A}, \mathbf{b}, \mathbf{c}$ (see (3)) and $\mathbf{Q}_L$ (see (12)) possess sparse structure, many multiplication and addition operations are automatically eliminated in the computation of (14)–(15). Second, since the square matrices $\mathbf{P}(k+L \mid k)$ and $\mathbf{P}(k+L \mid k+L)$ of (15) are symmetric, a half of its elements would not be required to compute. Third, there are identity matrices that conceal themselves in the coefficient matrices $\mathbf{A}$ (see (3)) and $\mathbf{G}$ (see(9)). Hence, most elements of the square matrices $\mathbf{FP}(k \mid k)\mathbf{F}^T$ and $\mathbf{GQ}_L\mathbf{G}^T$ are merely a shift of the matrix $\mathbf{P}(k \mid k)$ and $\mathbf{Q}_L$. Only a few elements of the square matrix $\mathbf{P}(k+L \mid k)$ need to be calculated. All of the above observations will be beneficial to realistic implementation of the Kalman reconstruction filter.

V. Conclusion

The missing packet recovery problem with the Lth interleaving configuration has been formulated and treated in this work from a multirate state-space modeling perspective. The interleaving techniques simplify the recovery of missing packets to the interpolation of missing samples. Thus, the Kalman-based sample-interpolation procedure of this work can be used to recover the missing speech packets. The sample interpolation is accomplished through the Kalman state estimator and, hence, the output samples are the minimum mean-square-error estimates of the missing speech signals.

The new sophisticated interpolation method is based on the AR evolution of the speech signals to estimate the missing samples from their neighbors, thereby preserving the important structures of the speech formants and spectral valleys in the recovered packets as closely as possible. The resulting effects are that the final reconstructed speech not only possesses higher SNR's than the values of previous works, but also have a better subjective quality under the identical packet loss rates. Hence, the Kalman-based sample interpolation procedure will become a highly promising scheme for the missing packet recovery problem.

The packet recovery performance of the Kalman-based sample-interpolation procedure can be further improved by raising the interleaving factor L. Higher interleaving factors will more uniformly distribute the missing samples to speech segments. Hence, a higher packet interleaving configuration is more advantageous to the sample interpolation procedures than the second (odd-even) interleaving configuration. However, the packetization configuration $L \times B$ must be designed in consideration of four dominant factors:

1) the upper bound of the tolerable decoding delay;
2) the number of speech crackles per second;
3) the probability of totally losing a phoneme;
4) the overhead of the side information.

Two proposed packet interleaving configurations ($L \times B = 2\times16$ and 4×8 ms) have been simulated in this work. They are selected with regarding to the above practical considerations. The latter configuration ($L \times B = 4\times8$ ms) can raise the tolerable level of packet loss ratio P_L to 10–20% with the Kalman-based packet recovery algorithm.

APPENDIX A
WIENER INTERPOLATION PROCEDURE

The fourth-order Wiener interpolation filter is derived in what follows for packet interleaving factor $L = 4$. Assume only the first two packets of the present speech segment are received, i.e., the packet receiving factor $M = 2$, then the interpolation filters would be as follows:

$$\begin{aligned}\hat{x}(k) &= h_1^3 x(k-2) + h_2^3 x(k-1) \\ &\quad + h_3^3 x(k+2) + h_4^3 x(k+3); \quad k = 3,7,11,\cdots \\ \hat{x}(k) &= h_1^4 x(k-3) + h_2^4 x(k-2) \\ &\quad + h_3^4 x(k+1) + h_4^4 x(k+2); \quad k = 4,8,12,\cdots\end{aligned} \tag{22}$$

where the interpolation coefficients h_i^3 and h_i^4 must be decided to minimize the variance of the interpolation error

$$\begin{aligned} e_3(k) &= \hat{x}(k) - x(k); \quad k = 3,7,11,\cdots \\ e_4(k) &= \hat{x}(k) - x(k); \quad k = 4,8,12,\cdots . \end{aligned} \tag{23}$$

The error variances are minimized when the following equations hold:

$$\begin{aligned} \frac{\partial E\left[e_3^2(k)\right]}{\partial h_i^3} &= 0; \quad i = 1,2,3,4 \\ \frac{\partial E\left[e_4^2(k)\right]}{\partial h_i^4} &= 0; \quad i = 1,2,3,4 \end{aligned} \tag{24}$$

which subsequently leads to the following Wiener–Hopf equation ($\mathbf{Rh} = \mathbf{r}$)

$$\begin{bmatrix} R(0) & R(1) & R(4) & R(5) \\ R(1) & R(0) & R(3) & R(4) \\ R(4) & R(3) & R(0) & R(1) \\ R(5) & R(4) & R(1) & R(0) \end{bmatrix} \begin{bmatrix} h_1^3 & h_1^4 \\ h_2^3 & h_2^4 \\ h_3^3 & h_3^4 \\ h_4^3 & h_4^4 \end{bmatrix} = \begin{bmatrix} R(2) & R(3) \\ R(1) & R(2) \\ R(2) & R(1) \\ R(3) & R(2) \end{bmatrix} \tag{25}$$

where $R(i)$ is the ith-lag autocorrelation function of the present speech segment, i.e., $R(i) = E[x(k-i)x(k)]$. The resultant Wiener interpolation coefficients can be obtained by solving the Wiener–Hopf equation (25), i.e., $\mathbf{h} = \mathbf{R}^{-1}\mathbf{r}$. Unlike the normal equation (19), the Wiener–Hopf equation (25) does not have an efficient algorithm to solve it.

APPENDIX B
SPEECH DATA

The input speech was coded by 64 kb/s μ-law log PCM. Five sentence-length Mandarin utterances with no obvious silent gaps were used as inputs to the simulated interleaving packet network.

1) Yueh Liang De Lean Tou Tou Tzay Gae Biann. (Male speaker 1)
2) Sheau Ding Dang Shih Lieh Tsorng Shu. (Male speaker 1)
3) Ing Erl Yeu Muu Chin Shih Lieh Baw Dao. (Female speaker 1)
4) Ching Hwa Dah Shyue Diann Ji Yan Jiow Shoo. (Female speaker 1)
5) Sheue Hua Fei Shiang Shwei. (Female speaker 2)

ACKNOWLEDGMENT

The authors thank the reviewers for their constructive comments and suggestions, which have greatly improved the quality of this manuscript.

REFERENCES

[1] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing* Englewood Cliffs, NJ: Prentice-Hall, 1983.
[2] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[3] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.
[4] T. Bially, B. Gold, and S. Seneff, "A technique for adaptive flow control in integrated packet networks," *IEEE Trans. Commun.*, vol. COMM-28, no. 3, pp. 325–333.
[5] N. Yin, S. Q. Li, and T. E. Stern, "Congestion control for packet voice by selective packet discarding," in *Proc. IEEE GLOBECOM*, 1987, pp. 1782–1786.
[6] D. W. Petr, Luiz A. DaSilva, Jr., and V. S. Frost, "Priority discarding of speech in integrated packet networks," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 644–656, June 1989.
[7] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[8] T. Kailath, *Lectures on Wiener and Kalman Filtering*. New York: Springer-Verlag, 1981.
[9] J. M. Mendel, *Lessons in Digital Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
[10] M. Decina and D. Vlack, "Voice by the packet?," *IEEE J. Select. Areas Commun.*, vol. SAC-1, pp. 961–962, Dec. 1983.
[11] L. Turner, T. Aoyama, D. Pearson, D. Anastassiou, and T. Minami, "Packet speech and video," *IEEE J. Select. Areas Commun.*, vol. 7, p. 629, June 1989.
[12] N. S Jayant and S. W. Christensen, "Effects of packet losses in waveform coded speech and improvements due to an odd-even sample-interpolation procedure," *IEEE Trans. Commun.*, vol. COMM-29, pp. 101–109, Feb. 1981.
[13] N. Matsuo, M. Yuito, and Y. Tokunaga, "Packet interleaving for reducing speech quality degradation in packet voice communications," in *Proc. IEEE GLOBECOM*, 1987, pp. 1787–1791.
[14] M. Yuito and N. Matsuo, "A new sample-interpolation method for recovering missing speech samples in packet voice communications," in *Proc. of IEEE ICASSP*, 1989, pp. 381–384.
[15] D. J. Goodman, G. B. Lockhart, O. J. Wasem, and W. C. Wong, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1440–1448, Dec. 1986.
[16] O. J. Wasem, D. J. Goodman, C. A. Dvorak, and H. G. Page, "The effect of waveform substitution on the quality of PCM packet communications," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 342–348, Mar. 1988.
[17] J. Suzuki and M. Taka, "Missing packet recovery techniques for low-bit-rate coded speech," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 707–717, June 1989.
[18] N. Erdöl, C. Castelluccia, and A. Zilouchian, "Recovery of missing speech packets using the short-time energy and zero-crossing measurements," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 295–303, July 1993.
[19] A. Ingle and V. A. Vaishampayan, "DPCM system design for diversity systems with applications to packetized speech," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 48–58, Jan. 1995.
[20] L. Ozarow, "On a source coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, pp. 1909–1921, Dec. 1980.
[21] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 821–834, May 1993.

[22] J. L. Ramsey, "Realization of optimum interleavers," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 338–345, May 1970.
[23] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

You-Li Chen (S'93–M'95) was born in Taiwan, Republic of China, on December 16, 1965. He received the M.S. and Ph.D. degrees in electrical engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 1991 and 1994, respectively.

He is now an Associate Professor of electronic engineering at Van-Nung Institute of Technology and Commerce, Chung-Li, Taur-Yuan, Taiwan. His main research interests are in the areas of digital signal processing, multirate signal processing, and speech applications.

Bor-Sen Chen (M'82–SM'89) received the B.S. degree from Tatung Institute of Technology, Taipei, Taiwan, R.O.C., and the M.S. degree from National Central University, Chungli, Taiwan, in 1970 and 1973, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, in 1982.

He was a Lecturer, Associate Professor, and Professor at Tatung Institute of Technology from 1973 to 1987. He is now a Professor at National Tsing-Hua University, Hsinchu, Taiwan. His current research interests are in robust control, adaptive control, and signal processing.

Dr. Chen has received the Distinguished Research Award four times from National Science Council of Taiwan.