Efficient Source Decoding Over Memoryless Noisy Channels Using Higher Order Markov Models

Farshad Lahouti, *Member, IEEE,* and Amir K. Khandani, *Member, IEEE*

***Abstract*—Exploiting the residual redundancy in a source coder output stream during the decoding process has been proven to be a bandwidth-efficient way to combat noisy channel degradations. This redundancy can be employed to either assist the channel decoder for improved performance or design better source decoders. In this work, a family of solutions for the asymptotically optimum minimum mean-squared error (MMSE) reconstruction of a source over memoryless noisy channels is presented when the redundancy in the source encoder output stream is exploited in the form of a γ-order Markov model ($\gamma \geq 1$) and a delay of $\delta, \delta > 0$, is allowed in the decoding process. It is demonstrated that the proposed solutions provide a wealth of tradeoffs between computational complexity and the memory requirements. A simplified MMSE decoder which is optimized to minimize the computational complexity is also presented. Considering the same problem setup, several other maximum *a posteriori* probability (MAP) symbol and sequence decoders are presented as well. Numerical results are presented which demonstrate the efficiency of the proposed algorithms.**

***Index Terms*—Forward–backward recursion, joint source–channel coding, maximum *a posteriori* probability (MAP) detection, Markov sources, minimum mean-squared error (MMSE) estimation, residual redundancies, source decoding.**

I. Introduction

An important result of the Shannon's celebrated paper [1] is that the source and channel coding operations can be separated without any loss of optimality. This has been the basic idea of enormous research endeavors in separate treatment of source and channel coders. However, in practise, due to strict design constraints such as limited transmission bandwidth, high error protection requirements along with restricted delay and limitations on the complexity of the systems involved, the joint design of source and channel coders has found increasing interest.

Researchers have taken several paths toward the joint design of source and channel coders. A class of joint source channel coders are designed by attempting to optimally allocate a fixed bitrate between the source and channel coders to achieve the maximum overall system performance. Examples of the works in this class are present in [2]–[6]. The applications span across the areas of speech, image, and video coding such as that of Modestino *et al.* on image coding using the discrete cosine transform with convolutional channel coding [2], the work of Moore and Gibson on differential pulse-code modulation (DPCM) speech coding with self-orthogonal convolutional coding [3] and the work of Bystrom and Modestino on combined source channel coding for video transmission [4].

Other methods of joint source and channel coding include the systems designed based on unequal error protection (UEP), optimized index assignment, channel optimized quantization, and more recently exploiting the source residual redundancies. In some applications, a combination of these techniques is employed for a greater protection over noisy channels.

Systems designed with unequal error protection provide better error protection for the parts of the source coder output stream which have a greater contribution in the objective or subjective quality of the reconstructed source. One good example of this technique is the North American IS-641 [7] standard which accommodates three different classes of error protection for different output bits of a code excited linear predictive (CELP)-based speech coder. A related classical work is the work of Sundberg [8] in which he analyzed the effect of error in different bits on the reconstruction of a pulse-code modulation (PCM) coded signal. Examples of more recent applications of UEP is present in [9] and [10].

The index assignment technique provides more robustness to channel errors by assigning the quantizer outputs to encoder indices in a way that possible bit errors create a lower level of distortion in the reconstructed data. One usual advantage of the index assignment is that it does not degrade the performance during the clean channel conditions. For a review of different index assignment techniques refer to [11].

In channel optimized quantization, the quantization levels are designed to optimize the performance of the system in the presence of channel noise. Two classic works in this area are those of Kurtenbach and Wintz [12] on scalar quantization over noisy channels and Chang and Donaldson [13] on the design of a DPCM system for transmission over a discrete memoryless channel. Other works on channel optimized quantization include the works of Kumazawa *et al.* [14] and Farvardin and Vaishampayan [15] on vector quantization over noisy channels as well as the works of Dunham and Gray [16] and Ayanoglu and Gray [17] on joint source–channel trellis coding. Examples of more recent works in this class are present in [18]–[20]. For a more comprehensive review of the techniques for channel optimized quantization, the interested reader is referred to [11], [20], [21].

More recently in this venue, exploiting the residual redundancy [22] in the output of the source coders for improved

Manuscript received March 4, 2002; revised October 5, 2003. This work is supported in part by the Natural Sciences and Engineering Research Council of Canada. The material in this paper was presented in part at the International Symposium on Telecommunications, Tehran, Iran, 2001.

The authors are with the Coding and Signal Transmission Laboratory, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: farshad@cst.uwaterloo.ca; khandani@cst.uwaterloo.ca).

Communicated by R. Zamir, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2004.833337

reconstruction over noisy channels has found increasing attention [22]–[44]. This redundancy is due to the suboptimal source coding which is caused by, e.g., a constraint on complexity or delay. Researchers have used the residual redundancy for enhanced channel decoding, e.g., [24]–[28] or for effective source decoding, e.g., [29]–[33]. The problem is formulated in the form of a maximum *a posteriori* (MAP) detection or a minimum mean squared error (MMSE) estimation. In [34], instantaneous MAP and MMSE decoders as well as a MAP sequence decoder were proposed that exploit the residual redundancies using a first-order Markov model. Later, in [35], a sequence-based MMSE decoder was suggested that benefits from the redundancies of both the past and future samples. Source decoding over channels with memory using the residual redundancies has been considered in [37]–[39].

In [40], it was demonstrated that the use of residual redundancies both at the source and channel decoder could lead to improved performance. In the same direction, iterative source and channel decoding schemes were presented in [41]–[43]. The effectiveness of these techniques have lead the researchers to new horizons. In [44], it is suggested that intentional leaving of the redundancy of the source, through the use of simpler source coders, could result in higher performance when this redundancy is exploited effectively at the decoder. This higher performance is either attributed to lower overall system complexity or better tradeoffs of bandwidth between the source and channel coding.

A. Contributions of the Manuscript

The recent literature clearly demonstrates the benefit of exploiting the residual redundancies in reconstructing the data received over noisy channels. However, it has primarily limited itself to modeling the redundancy with a first-order Markov model, which does not necessarily encapsulate all the remaining redundancy.[1] In this work, we present a family of solutions for the asymptotically optimum MMSE reconstruction of a source over memoryless noisy channels when the redundancy in the source encoder output stream is exploited in the form of a γ-order Markov model ($\gamma \geq 1$) and a delay of δ, $\delta \geq 0$, is allowed in the decoding process. We demonstrate that the proposed solutions provide a wealth of tradeoffs between computational complexity and the memory requirements. We also present a simplified MMSE decoder which is optimized to minimize the computational complexity. Considering the same problem setup, we present several other MAP symbol and sequence decoders as well. Finally, we study the effect of different system parameters and characteristics on the performance of the proposed decoders. In a typical application, the parameters γ and δ will serve as design parameters. They are appropriately selected to match the proposed decoders to the specific settings of the problem under consideration.

This paper is particularly inspired by the work of Sayood and Borkenhagen in [22], the article by Phamdo and Farvardin [34], the article by Miller and Park [35], and the work of Skoglund in [38]. Throughout the text, we will compare the presented developments to these works and illuminate the connections. The organization of this paper is as follows. In Section II, an overview of the system and the channel model used is described. In Section III, the MMSE decoding problem statement and solutions are presented. In Section IV, the MAP decoding problem statement and solutions are presented. Section V includes the numerical results and various comparisons which demonstrate the effectiveness of the proposed schemes. We conclude this paper in Section VI.

[1]In the context of transmission of digital images over a noisy channel, [45] and [46] consider Markov models that exploit the dependencies across two dimensions. See Section IV-B for more on [45].

Fig. 1. Overview of the system.

II. Preliminaries

A. Notations

The notations used in this paper are as follows. Capital letters, e.g., I, represent random variables, while small letters, e.g., i, represent a realization. We replace the probability $P(I = i)$ by $P(I)$ in most instances when it does not lead to a confusion. The vectors are shown bold faced, e.g., $\boldsymbol{X}$. The lower index indicates the time instant, e.g., $\boldsymbol{X}_n$ is the vector $\boldsymbol{X}$ at time instant n. The upper index in parenthesis indicates components of a vector or bit positions representing an integer value, e.g., $\boldsymbol{X}_n = [X_n^{(1)}, \ldots, X_n^{(N)}]$ where N is the dimension of the vector $\boldsymbol{X}_n$. A sequence of variables over time, e.g., $(I_{n_1}, \ldots, I_{n_2})$, $n_1 \leq n_2$ is denoted by $\underline{I}_{n_2}^{n_1}$. For simplicity, we represent $\underline{I}_n^1$ by $\underline{I}_n$. The N-dimensional Cartesian product of a set $\mathcal{J}$ is represented by $\mathcal{J}^N$ that consists of N-dimensional vectors whose components are taken from $\mathcal{J}$.

B. System Overview

The block diagram of the system is shown in Fig. 1. The source coder is a mapping from an N-dimensional Euclidean space $\mathcal{R}^N$ into a finite index set $\mathcal{J}$ of M elements. It is composed of two components: the quantizer Q and the index generator $\mathcal{I}$. The quantizer maps the input sample $\boldsymbol{X} \in \mathcal{R}^N$ to one of the reconstruction points or *codewords* in the codebook $\mathcal{C}, \mathcal{C} \subset \mathcal{R}^N$. The index generator then maps this codeword to the an *index (symbol)* I in the index set $\mathcal{J}$. The bit rate of the quantizer r is given by $\lceil \log_2 M \rceil$ bits per symbol (or $\lceil \log_2 M \rceil / N$ bits per dimension). We assume the source encoder is memoryless, i.e., the mapping of $\boldsymbol{X}_n$ to I_n is independent from the past and future values of the encoder input and output.

At the receiver, for each transmitted r-bit index $I = i$, a vector J with r components is received which provides information about I. The reconstructor (source decoder) maps J to an output sample $\hat{\boldsymbol{X}}$. In this reconstruction, the source decoder may use the previously received signals or some of the future samples as well.

C. Channel Model

The channels considered in this work are described by a probability density function (pdf) $P(J_n|I_n)$. We assume that the channel is memoryless without intersymbol interference in the sense that, for a sequence of transmitted symbols

$\underline{I}_n = (I_1, I_2, \ldots, I_n)$ and the corresponding received signals $\underline{J}_n$, the following equality is valid:

$$P(J_n = j_n | \underline{I}_n = \underline{i}_n, \underline{J}_{n-1} = \underline{j}_{n-1}) = P(J_n = j_n | I_n = i_n). \tag{1}$$

This results in the followings (see Appendix for proof):

$$P(J_n = j_n | \underline{I}_n = \underline{i}_n) = P(J_n = j_n | I_n = i_n) \tag{2}$$

$$P(\underline{J}_n = \underline{j}_n | \underline{I}_n = \underline{i}_n) = \prod_{k=1}^{n} P(J_k = j_k | I_k = i_k). \tag{3}$$

An example is a binary phase-shift keying (BPSK) modulation over a channel with additive white Gaussian noise (AWGN) which produces soft outputs. In this work, we refer to such a channel as the soft-output channel model. Also in Section V, we present simulation results on the performance of the source decoders over the binary-symmetric channel. In this case, the channel output is discrete and $P(J_n|I_n)$ represents a conditional probability mass function. In the followings, for the development of the proposed source decoders, we assume that the probability distribution of $P(J_n|I_n)$ is given and the memoryless channel assumption of (1) is valid.

III. MMSE Decoding: Problem Statement and Solutions

Consider the case where there is a residual redundancy at the source coder output stream. This redundancy is in the form of a nonuniform distribution or a memory in the sequence of the transmitted symbols. Our objective is to design an effective reconstructor (source decoder), which exploits this redundancy and produces the MMSE estimate of the source sample $\boldsymbol{X}_n$, given the received sequence $\underline{J}_{n+\delta} = \underline{j}_{n+\delta} = [j_1, j_2, \ldots, j_{n+\delta}]$. To minimize the expected squared error of estimation

$$E[(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)'(\boldsymbol{X}_n - \hat{\boldsymbol{X}}_n)] \tag{4}$$

the reconstructed signal is given by

$$\hat{\boldsymbol{x}}_n = E[\boldsymbol{X}_n | \underline{J}_{n+\delta} = \underline{j}_{n+\delta}]. \tag{5}$$

In (5), we have

$$E[\boldsymbol{X}_n | \underline{J}_{n+\delta}] = \sum_{\underline{I}_{n+\delta} \in \mathcal{J}^{n+\delta}} E[\boldsymbol{X}_n | \underline{J}_{n+\delta}, \underline{I}_{n+\delta}] P(\underline{I}_{n+\delta} | \underline{J}_{n+\delta}) \tag{6}$$

and noting that condition on $\underline{I}_{n+\delta} = \underline{i}_{n+\delta}$, $\boldsymbol{X}_n$ is independent of $\underline{J}_{n+\delta}$, we have

$$E[\boldsymbol{X}_n | \underline{J}_{n+\delta}, \underline{I}_{n+\delta}] = E[\boldsymbol{X}_n | \underline{I}_{n+\delta}]$$

which forms the *decoder codebook*. Therefore, the optimal decoder at time n requires a sum over $M^{n+\delta}$ elements of the decoder codebook. It is seen that in this case both computational complexity and the memory requirement grow exponentially with time, leading to an impractical scheme. Similar observations have been also made in [34], [35], [38].

In the next subsection, we develop an asymptotically optimum MMSE decoder for the cases where the residual redundancy is modeled by a γ-order Markov model and show that it leads to a feasible decoder. Subsequently, in Section III-B, we present a simplified MMSE decoder.

A. An Asymptotically Optimum MMSE Decoder

Assuming that the source $\boldsymbol{X}$ has a memory that asymptotically decays with time, for sufficiently large values of τ, $\tau \in \mathcal{Z}$, we have

$$E[\boldsymbol{X}_n | \underline{I}_{n+\delta}] \approx E[\boldsymbol{X}_n | \underline{I}_{n+\delta}^{n-\tau}]. \tag{7}$$

This simplifies the optimum decoder to the following decoder:

$$\hat{\boldsymbol{x}}_n = \sum_{\underline{I}_{n+\delta}^{n-\tau}} E[\boldsymbol{X}_n | \underline{I}_{n+\delta}^{n-\tau}] P(\underline{I}_{n+\delta}^{n-\tau} | \underline{J}_{n+\delta}) \tag{8}$$

which is asymptotically optimal for $\tau \gg 0$. We refer to the decoder of (8) as the asymptotically optimum minimum mean-squared error (AOMMSE) decoder. It describes the decoded signal in the form of the weighted average of the codewords $E[\boldsymbol{X}_n | \underline{I}_{n+\delta}^{n-\tau}]$. Note that $E[\boldsymbol{X}_n | \underline{I}_{n+\delta}^{n-\tau}]$ provides a finer reconstruction of the source symbols than the codewords $E[\boldsymbol{X}_n | I_n]$ used at the transmitter side.

At each time instant, the decoder needs to calculate the instantaneous values of the weights in (8) or the probabilities $P(\underline{I}_{n+\delta}^{n-\tau} | \underline{J}_{n+\delta})$. To calculate these probabilities, we assume that due to the residual redundancy at the encoder output, the encoder output symbols form a γ-order, $\gamma > 0$, Markov model. In the following sections, we first present a solution to compute these probabilities and then consider alternative solutions to optimize the computational complexity. These solutions are valid for all values of $\tau \geq \gamma$ (the case $\tau < \gamma$ will be straightforward).

In [34] and [38], MMSE-based reconstruction schemes are proposed that exploit the residual redundancy using a first-order Markov model. Specifically, for $\gamma = 1, \delta = \tau = 0$, the decoding rule of (8) is simplified to the "instantaneous approximate MMSE decoder" of [34]. If $\gamma = \tau = 1, \delta \geq 0$, then the AOMMSE decoding rule is equivalent to the "Markov type 1" decoder of [38] used for a memoryless channel.

1) A Basic Solution: To compute the *a posteriori* probabilities required in the AOMMSE decoder we use a trellis structure. The trellis structure models the symbols I_n and their assumed γ-order Markov property due to the residual redundancies. In this trellis structure, the states at time n correspond to the ordered set

$$S_n = (I_{n-\gamma+1}, I_{n-\gamma+2}, \ldots, I_{n-1}, I_n), \quad I_{n-k} \in \mathcal{J},\ 0 \leq k < \gamma. \tag{9}$$

Hence, there are M^γ states in each time step (stage) $S_n \in \mathcal{J}^\gamma$. Each branch leaving the state at time step n corresponds to one particular symbol $I_{n+1} = i_{n+1}$. Therefore, there are M branches leaving each state. Each branch is identified by the pair $(S_n = s_n,\ S_{n+1} = s_{n+1})$ of the two states that the branch connects together. Having defined the trellis structure as such, there will be one *a priori* probability $P(I_{n+1} = i_{n+1} | S_n = s_n)$ corresponding to each branch which characterizes the γ-order Markov property of the source. The states now form a first-order Markov sequence. Using this property and the memoryless assumption of the channel (see (1)–(3)), in line with the

Bahl–Cocke–Jelinek–Raviv (BCJR) algorithm [47], the probability of a particular state $S_n = s_n$ given the observed sequence $\underline{J}_n$ is given by the following *forward* recursive equation:

$$P(S_n|\underline{J}_n) = C\,.\,P(J_n|I_n) \cdot \sum_{S_{n-1}\to S_n} P(I_n|S_{n-1})P(S_{n-1}|\underline{J}_{n-1}) \quad (10)$$

where the summation is over a subset of M states in time step $n-1$, that are connected to the state S_n. Throughout the paper, we use the notation C as a factor normalizing the sum of probabilities to one.

In the same direction, the probabilities of states given the observed sequence $\underline{J}_{n+\delta}$, $\delta \geq 0$ are calculated by the following equation:

$$P(S_n|\underline{J}_{n+\delta}) = C\,.\,P(S_n|\underline{J}_n)\,.\,P(\underline{J}_{n+\delta}^{n+1}|S_n) \quad (11)$$

where

$$\underline{J}_{n+\delta}^{n+1} = [J_{n+1}, J_{n+2}, \ldots, J_{n+\delta}].$$

Equation (11) is referred to as the *forward–backward* equation in which the first term is the forward equation given in (10) and the second term is referred to as the *backward* equation and can be calculated recursively as follows:

$$P(\underline{J}_{n+\delta}^{n+1}|S_n) = \sum_{I_{n+1}\in\mathcal{J}} P(J_{n+1}|I_{n+1})\,.\,P(I_{n+1}|S_n)\,.\,P(\underline{J}_{n+\delta}^{n+2}|S_{n+1}) \quad (12)$$

where the recursion starts from

$$P(J_{n+\delta}|S_{n+\delta-1}) = \sum_{I_{n+\delta}\in\mathcal{J}} P(J_{n+\delta}|I_{n+\delta}).P(I_{n+\delta}|S_{n+\delta-1}) \quad (13)$$

and continues backward in each time step. The details of the derivation of these equations are available in [48]. The presented trellis structure and either of the forward and backward equations are used in the following sections for calculation of different symbol or sequence probabilities. We note that in each time step, the forward recursion of (10) proceeds one step forward through the trellis while the backward term is recomputed over the entire backward window as indicated in (12) and (13).[2]

Now, using the presented trellis structure and the forward equation (10), the probabilities required for the asymptotically optimum MMSE decoding of (8) are calculated recursively by (14) at the bottom of the page ($\tau \geq \gamma > 0$, see Appendix for proof),

At each time instant the M^γ probabilities $P(S_{n-(\tau-\gamma)}|\underline{J}_{n-(\tau-\gamma)})$ corresponding to each state are stored to be used in (14) at the next time instant. In addition to the computations required to do this task, the complexity of the AOMMSE decoder is comprised of the cost to perform the multiplications and normalization in (14), as well as the weighted average of the reconstruction rule in (8). Clearly, these computations are not trivial and, therefore, efficient alternative solutions are of particular interest. In the followings, we present efficient alternative exact solutions based on construction of an extended trellis structure of the source.

2) Solutions Based on the Extended Trellis Structure: The general problem of calculating the weights required in (8) can be viewed as finding the probability of a sequence of states (symbols) within the structure of the source trellis given the entire history of the received signals or equivalently

$$P(S_{n-L}, \ldots, S_{n-1}, S_n|\underline{J}_n) = P(\underline{I}_n^{n-(\gamma+L)+1}|\underline{J}_n), \qquad L>0$$

where $L+1$ is the length of the sequence of states. The alternative solutions are provided based on different constructions of an extended trellis structure for the source coder output symbols. The states in this structure are referred to as the *super states* and are defined as

$$SS_n = (I_{n-(\gamma+L')+1}, \ldots, I_{n-1}, I_n), \quad I_{n-k} \in \mathcal{J},\ 0 \leq k < \gamma + L' \quad (15)$$

where L', $0 \leq L' < L$ is referred to as the (state set) extension factor and is the number of symbols by which the states (see (9)) have been extended to form the super states. Similar to the original trellis, there are M branches leaving a (super) state in the extended trellis, where each branch corresponds to one symbol $I_n = i_n, i_n \in \mathcal{J}$. Therefore, each stage of the extended trellis still corresponds to one time step.

Based on the proposed extended trellis, a family of solutions to calculate the required *a posteriori* probabilities are given as follows (see Appendix for proof):

$$P(\underline{I}_n^{n-(\gamma+L)+1}|\underline{J}_n) = C\,.\left[\prod_{k=0}^{L-L'-1} P(J_{n-k}|I_{n-k})P(I_{n-k}|SS_{n-k-1})\right] .P(SS_{n-(L-L')}|\underline{J}_{n-(L-L')}) \quad (16)$$

where

$$P(I_{n-k}|SS_{n-k-1}) = P(I_{n-k}|S_{n-k-1}) \quad (17)$$

[2]It is noteworthy that the forward–backward algorithm [47] has been used in different forms and applications such as channel decoding and the decoding in hidden Markov models. In another work [38], similar developments are related to the prediction and filtering within the context of Kalman filtering. There are recent efforts to develop general mathematical frameworks that contain these separately known algorithms [49]. Our focus here is to investigate the details of the calculation of the specific *a posteriori* probabilities required, so as to find more efficient solutions.

$$P(\underline{I}_{n+\delta}^{n-\tau}|\underline{J}_{n+\delta}) = C\,.\left[\prod_{k=-(\tau-\gamma)}^{\delta} P(J_{n+k}|I_{n+k})P(I_{n+k}|S_{n+k-1})\right] P(S_{n-(\tau-\gamma)-1}|\underline{J}_{n-(\tau-\gamma)-1}). \quad (14)$$

and $P(SS_{n-(L-L')}|\underline{J}_{n-(L-L')})$ are $M^{\gamma+L'}$ probabilities of super states which are stored in each time step. This term is updated by

$$P(SS_l|\underline{J}_l) = C.\sum_{SS_{l-1}\to SS_l} P(J_l|I_l)P(I_l|SS_{l-1})P(SS_{l-1}|\underline{J}_{l-1}) \quad (18)$$

in which $l = n-(L-L')+1$ and the terms within the summation are available during the process of calculating (16). The direct implementation of (16) leads to a computational complexity[3] CC of

$$CC = 2(L-L'+1)M^{L+\gamma} + (M+2)M^{L'+\gamma} \quad (19)$$

in which the first term is the cost due to the required multiplication of the terms and the required normalization and the second term includes the cost due to updating and normalizing the probability of the super states according to (18). The memory requirement includes the fixed amount of static memory (ROM) required to store $M^{\gamma+1}$ transition probabilities, as well as the dynamic memory (RAM) required for the operation of the algorithm, which is $O(M^{\gamma+L'})$ based on the number of super states. This indicates that the family of solutions of (16) provide a wealth of tradeoffs of computational complexity and the memory requirement. The increase of L' reduces the computational complexity at the cost of an increase in memory requirement. It is important to note that any increase of L', $0 \le L' < L$, beyond $L-1$ would lead to a solution which is suboptimal both in terms of computational complexity and the memory requirement.

More tradeoffs of computational complexity and memory are possible considering the fact that the structure of the extended trellis is still based on the redundancy of the source as indicated in (17). For example using (17), the multiplying terms of (16) can be calculated once for the M^{γ} states S_{n-k-1} and stored as an $M^{\gamma} \times M^{L-L'}$ matrix to be appropriately multiplied by the probability of super states. This reduces the corresponding computations in (19) from $2(L-L')M^{L+\gamma}$ to $2(L-L')M^{L-L'+\gamma}$.

B. A Simplified MMSE Decoder

An approximation of interest to the AOMMSE decoder is to consider a decoder that uses a codebook identical to that of the encoder. We refer to this decoder as the MMSE decoder. Simplifying (5), the MMSE decoder is given by

$$\hat{\boldsymbol{x}}_n = \sum_{I_n \in \mathcal{J}} E[\boldsymbol{X}_n|I_n]P(I_n|\underline{J}_{n+\delta}) \quad (20)$$

which describes the MMSE estimate in terms of the weighted average of the encoder (Linde, Buzo, and Gray (LBG) [50]) codewords. The weights are the probability of having transmitted the corresponding symbol given the received sequence $\underline{J}_{n+\delta}$. It is noteworthy that in the trivial case where there is no memory between the symbols I_n (corresponding to $\gamma = 0$), (20) collapses to the basic MMSE reconstruction rule

$$\hat{\boldsymbol{x}}_n = \sum_{I_n \in \mathcal{J}} E[\boldsymbol{X}_n|I_n]P(I_n|J_n) \quad (21)$$

in which the probability

$$P(I_n|J_n) = C.P(I_n).P(J_n|I_n), \qquad C = \frac{1}{P(J_n)}$$

includes the residual redundancy in the form of the nonuniform symbol *a priori* probabilities.

In the followings, we present efficient solutions to calculate the required probabilities in (20) based on a γ-order Markov redundancy model. It is noteworthy that, for $\gamma = 1$, the simplified MMSE decoder of (20) is reduced to the "sequence-based MMSE decoder" of [35], and if also $\delta = 0$, then the decoder is essentially the "instantaneous approximate MMSE decoder" of [34].

1) A Basic Solution: The *a posteriori* probability of a symbol I_n given the received sequence $\underline{J}_{n+\delta}$ is calculated as follows. Assuming that the encoded sequence contains a residual redundancy in the form of a γ-order, $\gamma \ge 1$, Markov model, we use the probabilities of states in the original trellis structure as described in Section III-A1. In particular, when no delay is allowed in the decoding process $\delta = 0$, we have

$$P(I_n|\underline{J}_n) = \sum_{I_{n-\gamma+1}} \cdots \sum_{I_{n-2}} \sum_{I_{n-1}} P(S_n|\underline{J}_n). \quad (22)$$

Expressions (10) and (22) together with the reconstruction rule of (20) provide the instaneous (no delay allowed, i.e., $\delta = 0$) MMSE decoding of the source samples given the history of the received channel outputs.

We observe that the required symbol *a posteriori* probabilities can be alternatively calculated using the *a posteriori* probabilities of any of the states S_{n+m} as long as S_{n+m} includes I_n, i.e., $0 \le m \le \gamma - 1$. As presented later, this is of particular interest when a delay of $\delta > 0$ is allowed in the decoding process. In such cases, this flexibility can be used to optimize the solution in terms of the complexity. We have

$$P(I_n|\underline{J}_{n+\delta}) = \cdots \sum_{I_{n+k}} \cdots P(S_{n+m}|\underline{J}_{n+\delta}), \quad \forall m \in \mathcal{Z}, 0 \le m \le \gamma - 1, \quad k = m - \gamma + 1, \ldots, m, \quad k \ne 0 \quad (23)$$

where the probabilities of states $P(S_{n+m}|\underline{J}_{n+\delta})$ as described in Section III-A1, are given by the following forward–backward equation:

$$P(S_{n+m}|\underline{J}_{n+\delta}) = C\,.P(S_{n+m}|\underline{J}_{n+m})\,.\,P(\underline{J}_{n+\delta}^{n+m+1}|S_{n+m}), \quad 0 \le m \le \delta \quad (24)$$

in which the forward and the backward terms are given in (10) and (12). In (24), if the number of computations required for the forward and backward recursions ((10) and (12)) per time step is denoted by CC_{fwd} and CC_{bwd}, respectively, we have

$$CC_{\text{fwd}} = (2M+3)\,M^{\gamma} \quad (25)$$

[3]In this work, the computational complexity is measured in terms of the number of floating-point operations. Each addition, multiplication, or comparison is considered as one floating-point operation (flop).

$$CC_{\text{bwd}} = 3(\delta - m)\, M^{\gamma+1} \tag{26}$$

where $\delta - m$ is the number of backward recursions required per time step. The overall complexity of computing (23) is then given by[4]

$$CC = CC_{\text{fwd}} + CC_{\text{bwd}} + 2M^{\gamma} + 2M. \tag{27}$$

Noting that only CC_{bwd} depends on m in (27), to minimize the overall computational burden, we solve the following for the optimum value of m:

$$\begin{aligned} &\text{Minimize} \quad CC_{\text{bwd}} = 3(\delta - m)\,.\,M^{\gamma+1} \\ &\text{subject to} \quad 0 \le m \le \gamma - 1; \quad 0 \le m \le \delta. \end{aligned} \tag{28}$$

Case 1. $\delta < \gamma$: In the cases where the delay is smaller than the assumed residual redundancy order, we are able to choose $m = \delta$ and eliminate the backward term. The probabilities in (20) are then given as follows:

$$P(I_n|\underline{J}_{n+\delta}) = \dots \sum_{I_{n+k}} \dots P(S_{n+\delta}|\underline{J}_{n+\delta}), \quad k = \delta - \gamma + 1, \dots, \delta,\ k \neq 0. \tag{29}$$

Case 2. $\delta \ge \gamma$: Alternatively, when the delay is larger than the assumed redundancy order, the CC_{bwd} is minimized when $m = \gamma - 1$, i.e., $\delta - \gamma + 1$ backward recursions are required. The probabilities in (20) are now given by

$$P(I_n|\underline{J}_{n+\delta}) = \sum_{I_{n+1}} \sum_{I_{n+2}} \dots \sum_{I_{n+\gamma-1}} P(S_{n+\gamma-1}|\underline{J}_{n+\delta}) \tag{30}$$

together with (10)–(13). The value $m = \gamma - 1$ sets up the solution of (30) based on the probabilities of states $S_{n+\gamma-1} = (I_n, \dots, I_{n+\gamma-1})$, in which I_n is located in the last position. Hence, it reduces the number of backward recursions while keeping the complexity due to the forward recursion unchanged. This motivates us that should we set up a solution based on the *a posteriori* probabilities of a sequence larger than a state, we could reduce the number of the backward recursions and its complexity even further. In the following subsection, we present such solutions and examine if this leads to a smaller *overall* complexity as compared to the solution of (30).

2) Alternative Solutions: In this subsection, we reconsider the problem of finding the *a posteriori* probability of a symbol I_n, given the observed sequence $\underline{J}_{n+\delta}$ for the cases where $\delta \ge \gamma$. Motivated by the results and discussions presented in Section III-B1, Case 2, we seek possibly more efficient solutions to calculate the required a posteriori symbol probability using the probability of a sequence larger than one state, i.e.,

$$(I_n, \dots, I_{n+\gamma+L-1}), \qquad 0 < L \le \delta - \gamma + 1$$

[4]Note that the computational complexity of the forward equation includes the cost of normalization as well ($2M^{\gamma}$). This is a common approach in this work in which the forward probabilities of *trellis states* which are stored to be used in the next time instant are always normalized. However, in practise, we perform the required normalization of (24) after we summed the multiplied forward and backward terms according to (23). This only costs $2M$ operations as opposed to the original $2M^{\gamma}$ according to (24).

we have

$$P(I_n|\underline{J}_{n+\delta}) = \sum_{I_{n+1}} \sum_{I_{n+2}} \dots \sum_{I_{n+\gamma+L-1}} P(I_n, \dots, I_{n+\gamma+L-1}|\underline{J}_{n+\delta}), \tag{31}$$

where $P(\underline{I}^{n}_{n+\gamma+L-1}|\underline{J}_{n+\delta})$ is given by the following forward–backward equation:

$$P(I_n, \dots, I_{n+\gamma+L-1}|\underline{J}_{n+\delta}) = C.P(I_n, \dots, I_{n+\gamma+L-1}|\underline{J}_{n+\gamma+L-1}).P(\underline{J}^{n+\gamma+L}_{n+\delta}|S_{n+\gamma+L-1}). \tag{32}$$

The first term in (32) is the *a posteriori* probability of a sequence of $L+\gamma$ symbols or L states ($\gamma, L > 0$), given the entire history of the received information from the channel. To calculate such a probability, in Section III-A2, we presented a set of solutions based on different constructions of an extended trellis structure for the source coder output with $M^{\gamma+L'}$, $0 \le L' < L$ states in each stage. The second term in (32) is a backward recursive term which is given by (12). The number of backward recursions is equal to $\delta - \gamma - L + 1$, which as expected reduces with the increase of L and is always smaller than that of the solution in Section III-B1, Case 2. However, an increase of L results in a more complex forward term.

Using the results of Sections III-A1 and III-A2, the overall complexity of the solution in (31) is given by

$$CC = 2(L - L' + 1)M^{L+\gamma} + 3(\delta - \gamma - L + 1)M^{\gamma+1} + (M+2)M^{\gamma+L'} + 2M \tag{33}$$

which includes the number of computations required to calculate the forward and backward terms, updating the forward probabilities of super states according to (18) and multiplications and normalization required in (32). The memory requirement includes the fixed amount of static memory required to store $M^{\gamma+1}$ transition probabilities and the dynamic memory required is $O(M^{\gamma+L'})$.

This set of solutions can be optimized over the choices of L (sequence length) and $L', 0 \le L' < L$ (the state set extension factor for the forward term). Interestingly, using the results of Section III-A2, it can be shown that $L = 1$ optimizes the solution in terms of the computational complexity disregarding the values of γ and δ for all $M \ge 2$. It is noteworthy that since $L = 1$ requires $L' = 0$ hence, $L = 1$ minimizes the memory requirement as well. Therefore, the optimum solution (for $\delta \ge \gamma$) in terms of the complexity, provided by the family of solutions of (31) is based on the original source trellis and is given by

$$P(I_n|\underline{J}_{n+\delta}) = \sum_{I_{n+1}} \sum_{I_{n+2}} \dots \sum_{I_{n+\gamma}} P(I_n, \dots, I_{n+\gamma}|\underline{J}_{n+\delta}) \tag{34}$$

where from (32) we have

$$P(I_n, \dots, I_{n+\gamma}|\underline{J}_{n+\delta}) = C\,.\,P(I_n, \dots, I_{n+\gamma}|\underline{J}_{n+\gamma})\,.P(\underline{J}^{n+\gamma+1}_{n+\delta}|S_{n+\gamma}) \tag{35}$$

in which, using (16), the forward term is given by

$$P(\underline{I}^{n}_{n+\gamma}|\underline{J}_{n+\gamma}) = C\,.\,P(J_{n+\gamma}|I_{n+\gamma})\,.\,P(I_{n+\gamma}|S_{n+\gamma-1})\,.\,P(S_{n+\gamma-1}|\underline{J}_{n+\gamma-1}) \tag{36}$$

and the backward term is calculated using (12) and (13). The probabilities of states in the forward term are then updated by the following:

$$P(S_{n+\gamma}|\underline{J}_{n+\gamma}) = \sum_{I_n \in \mathcal{J}} P(\underline{I}^n_{n+\gamma}|\underline{J}_{n+\gamma}). \tag{37}$$

From (33), the overall complexity of this solution is given by

$$CC = (3\delta - 3\gamma + 5)M^{\gamma+1} + 2M^{\gamma} + 2M. \tag{38}$$

Now, it remains to compare the above solution (based on $L = 1$, $L' = 0$) with that presented in Section III-B1, Case 2. Examining (38) and (27) indicates that the current solution maintains a lower complexity. Although both solutions are based on the same original source trellis, their distinction stems from using different forward recursive (and updating) equations ((36) and (37) versus (10)). This, in turn, leads to the reduction of the backward term by an additional step as seen comparing (35) and (24) for $m = \gamma - 1$.

IV. MAP Decoding: Problem Statement and Solutions

A. MAP Symbol Decoder

An instantaneous symbol MAP decoder exploiting the residual redundancies in the form of a first-order Markov model was presented in [34]. Later, in [35], a decoder that accommodates a certain delay in the decoding process was proposed for the same problem setup. Here, we present an optimal symbol MAP decoder when the residual redundancies are captured with a γ-order Markov model and a delay of δ is allowed in the decoding process.

The symbol MAP decoder receives the sequence $\underline{J}_{n+\delta}$ and determines the most probable transmitted symbol. Next, it outputs the corresponding codeword. We have

$$\begin{aligned} \hat{\boldsymbol{x}}_n &= E[\boldsymbol{X}_n | I_n = \hat{i}_n] \\ \hat{i}_n &= \arg \max_{I_n \in \mathcal{J}} P(I_n | \underline{J}_{n+\delta}). \end{aligned} \tag{39}$$

The required *a posteriori* probability of the symbol I_n in (39) can be efficiently calculated as described in Section III-B. The performance of this decoder is studied in Section V where it is referred to as the MAP decoder.

The presented MAP decoder uses a codebook identical to that of the encoder. Alternatively, we can use the decoder codebook corresponding to the asymptotically optimum MMSE decoding algorithm with the MAP decoder. In this case, a sequence is decoded such that

$$\hat{\underline{i}}^{n-\tau}_{n+\delta} = \arg \max_{\underline{I}^{n-\tau}_{n+\delta} \in \mathcal{J}^{\tau+\delta+1}} P(\underline{I}^{n-\tau}_{n+\delta} | \underline{J}_{n+\delta}). \tag{40}$$

using (14) or (16). Next, the source decoder reproduces

$$\hat{\boldsymbol{x}}_n = E[\boldsymbol{X}_n | \underline{I}^{n-\tau}_{n+\delta} = \hat{\underline{i}}^{n-\tau}_{n+\delta}]$$

at the output. We refer to this technique as the AOMAP decoder and present its performance in Section V.

B. MAP Sequence Decoder

A sequence MAP decoder exploiting the residual redundancies in the form of a first-order Markov model was presented in [22] for source decoding over noisy channels. Later, in [34], a similar but optimal decoder was proposed. Here we consider and analyze an optimal sequence MAP decoder when the residual redundancies are captured with a γ-order Markov model. A similar decoder was derived in [45].

The sequence MAP decoder receives the sequence $\underline{J}_T$ and determines the most probable transmitted sequence

$$\hat{\underline{i}}_T = \arg \max_{\underline{I}_T \in \mathcal{J}^T} P(\underline{I}_T | \underline{J}_T). \tag{41}$$

Using the same trellis structure as described in the previous section and considering the memoryless property of the channel as well as the Markov model for the source redundancy, it is straightforward to see that (41) is equivalent to (42) at the bottom of the page (see [48] for a proof), where $S_1 \triangleq (I_1, 0, \ldots, 0)$. The sequence MAP decoder in (42) can be implemented using the well-known Viterbi algorithm. We use the same trellis structure as defined in Section III-A1 and the metric corresponding to branch (S_{k-1}, S_k) is given by $\log[P(J_k|I_k)P(I_k|S_{k-1})]$. The optimum sequence MAP decoder, according to (42), requires the entire sequence $\underline{J}_T$ in order to decode the corresponding sequence $\hat{\underline{i}}_T$ and, hence, imposes a large delay. However, to limit the delay to a certain value, at each time instant, we identify the state with the maximum metric and decode the symbol at delay δ on the surviving path reaching that state accordingly. Subsequently, the corresponding codeword is reproduced at the source decoder output. We refer to this decoder as the sequence MAP (SMAP) decoder and will examine its performance in Section V.

Given that the values $\log P(J|I)$ received from the channel are available, the computational complexity of the SMAP algorithm per time step is given by

$$CC = 3M^{\gamma+1} + M^{\gamma} \tag{43}$$

which includes the computations required for updating the state metrics and selecting the one with the largest value.

The decoder codebook in the presented SMAP decoder is the same as the encoder codebook. Alternatively, we can use the decoder codebook corresponding to the asymptotically optimum MMSE decoding algorithm with the SMAP decoder. In this case, a sequence is decoded which in turn outputs one of the decoder codewords $E[\boldsymbol{X}_n|\underline{I}^{n-\tau}_{n+\delta} = \hat{\underline{i}}^{n-\tau}_{n+\delta}]$. We refer to this technique as the AOSMAP decoder and present its performance in Section V.

V. Numerical Results

To analyze the performance of the proposed MMSE decoders, we use a synthesized source similar to [22]. In addition to the fifth-order Gauss–Markov source from [22], two other tenth-

$$\hat{\underline{i}}_T = \arg \max_{\underline{I}_T \in \mathcal{J}^T} \left[\sum_{k=2}^{T} \log[P(J_k|I_k)P(I_k|S_{k-1})] + \log[P(J_1|I_1)P(S_1)] \right] \tag{42}$$

TABLE I
FILTER COEFFICIENTS OF THE SYNTHESIZED SOURCES

f_A	1.3822	-0.3399	-0.1772	-0.6760	0.6396	-0.0719	0.1386	-0.6024	0.4561	-0.1560
f_B	1.381	-0.599	0.367	-0.700	0.359					
f_C	1.7493	-2.4263	2.5733	-1.6585	0.7426	-0.1644	-0.3019	0.1157	-0.0350	-0.0018

TABLE II
REDUNDANCY OF THE SOURCE, $R(M,\gamma)$ (IN BITS), AT DIFFERENT REDUNDANCY MODEL ORDERS γ, $(M = 8, N = 1)$

Redundancy Order γ	0	1	2	3
R_A	0.254	1.162	1.482	1.667
R_B	0.319	1.182	1.327	1.388
R_C	0.439	0.540	0.867	1.248

TABLE III
NORMALIZED AUTOCORRELATION OF THE SOURCE SAMPLES AT DIFFERENT DELAYS

Delay δ	0	1	2	3
ρ_A	1.000	0.849	0.553	0.159
ρ_B	1.000	0.868	0.599	0.280
ρ_C	1.000	0.374	−0.417	−0.098

order Gauss–Markov sources are used whose coefficients have been picked from a speech linear prediction coding (LPC) database. Each 10 LPC coefficient set represents the short-time spectral information of speech within 20 ms. The source samples are given by

$$x_k = \sum_{i=1}^{\gamma_s} f(i).x_{k-i} + e_k \tag{44}$$

where E_k is a Gaussian independent and identically distributed (i.i.d.) random variable, γ_s is the order of the synthesizing filter, and the corresponding coefficients $f(i)$ are given in Table I. The source sample vector $\boldsymbol{X}_n = [X_{(n-1)N+1}, \ldots, X_{nN}]$ is quantized with an M point N-dimensional vector quantizer (VQ) producing the symbol I_n. At different redundancy model orders γ, the value $R(M,\gamma)$ in bits defined as

$$R(M,\gamma) \triangleq \log_2 M - H(I_n|S_{n-1}) \tag{45}$$

where $S_n = [I_n, \ldots, I_{n-\gamma+1}]$ provides an indication of the redundancy to be exploited and hence, the gains to be achieved. Table II presents the amount of $R(M,\gamma)$ for the selected synthesized sources at different values of γ when the source is quantized by a 3-bit LBG scalar quantizer. As given in Table II, for source A, the redundancy due to the nonuniform distribution $(\gamma = 0)$ is 0.25 bit. The redundancy exploited by means of a first-, second-, and third-order Markov model is 1.16, 1.48, and 1.67 bits, respectively. In Table III, the normalized autocorrelation of the source samples at different delays are also presented. In the followings, we investigate the performance of the decoders presented in the previous sections. Six source decoders are considered: i) AOMMSE decoder, ii) simplified MMSE (MMSE) decoder which uses identical encoder and decoder codebooks, iii) MAP symbol decoder (MAP), iv) AOMAP decoder which selects the codeword with the maximum *a posteriori* probability from the codebook corresponding to the AOMMSE decoder, v) the SMAP decoder, vi) and the AOSMAP decoder which is the SMAP decoder which uses the decoder codebook of the AOMMSE decoder. We begin with the performance comparison of the instantaneous decoders $(\delta = 0)$ over a binary-symmetric channel and we proceed to analyze the effect of delay, performance with a channel (decoder) with soft outputs, the effect of redundancy type, and the effect of quantizer bit rate.

A. Basic Comparison of the Decoders

In this subsection, we present a performance comparison of the instantaneous decoders $(\delta = 0)$. Fig. 2 demonstrates the performance of the instantaneous AOMMSE decoder for $\tau = \gamma$ (dotted lines) and the MMSE decoder (solid lines), for transmission of source A over a binary-symmetric channel when different levels of residual redundancy is exploited at the receiver $(\gamma = 1, 2, 3)$. As mentioned before, for $\gamma = 0$ (and $\tau = 0$) both schemes collapse to the basic MMSE decoder of (21). The performance of the basic MMSE decoder $(\gamma = 0)$ with equal symbol probability assumption (EPA) in which case no *a priori* information is used in the decoding process $(P(I = i) = \frac{1}{M}, \forall i \in \mathcal{J})$ is provided as a baseline for comparison.

For the AOMMSE decoder, Fig. 2 shows that using a redundancy model of order $\gamma = 2$ or $\gamma = 3$ provides a gain as high as 2.5 or 4 dB, respectively, compared to the case where the redundancy is modeled with a first-order Markov model. Using the simplified MMSE decoder, similar gains are achievable, however, at lower bit-error rates, the performance is upper-bounded by that provided at the encoder output. As mentioned before, in such cases the AOMMSE decoding provides a finer reconstruction of the source samples. This is due to using a larger decoder codebook which exploits the dependencies between the source coder output symbols, which was also observed in [35], [38]. The performance of the corresponding AOMAP and MAP decoders are presented in Fig. 3. In Fig. 4, the performance of a selected set of instantaneous AOMMSE, MMSE, AOMAP, and MAP decoders are redrawn for comparison. It is observed that the MMSE decoders constantly outperform the MAP decoders with gains as high as 1.4 dB.

Fig. 5 compares the performance of the MAP symbol decoder with that of the SMAP decoder for transmission of source A over a binary-symmetric channel. It is observed that the SMAP algorithm, although suboptimal in the sense of minimizing the symbol probability of error, performs very

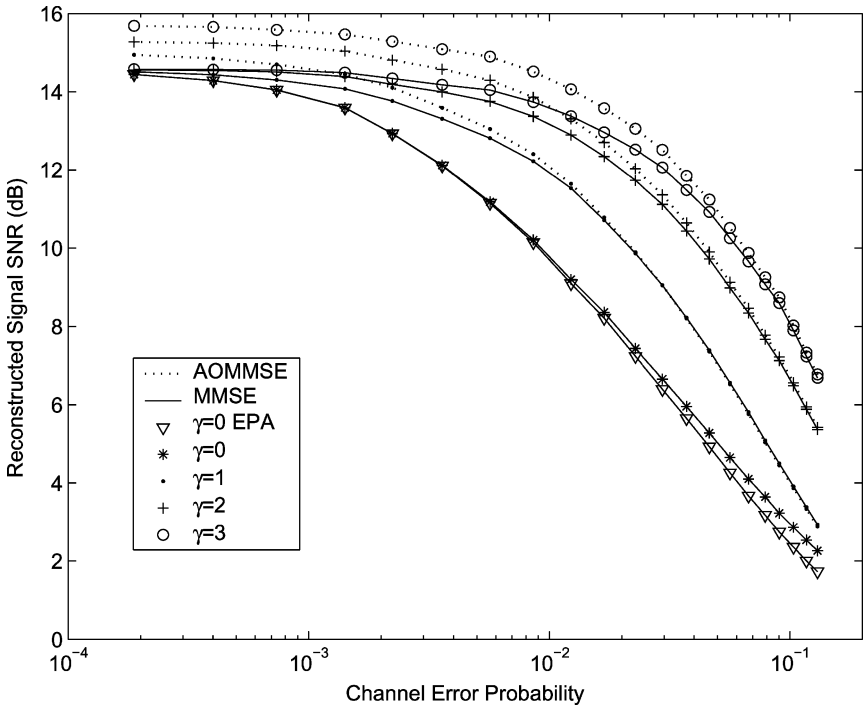

Fig. 2. Performance of the instantaneous ($\delta = 0$) AOMMSE (for $\tau = \gamma$) and MMSE decoders for transmission of the Gauss–Markov source A ($M = 8$, $N = 1$) over a binary-symmetric channel when different levels of residual redundancy are exploited at the decoder. Note that for $\gamma = 0$ and $\gamma = 0$ EPA the curves of AOMMSE and MMSE decoding have overlapped.

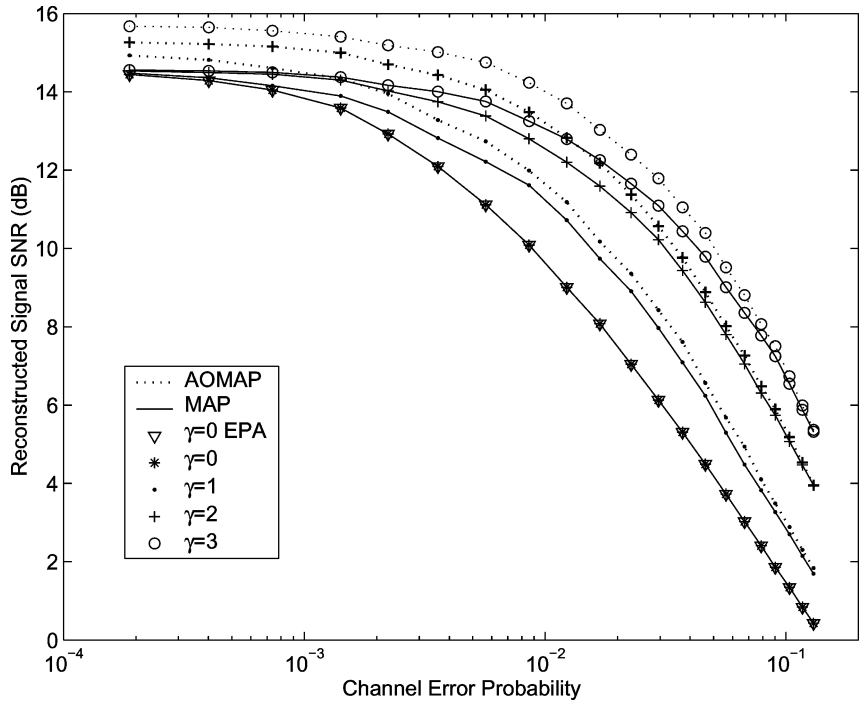

Fig. 3. Performance of the instantaneous ($\delta = 0$) AOMAP (for $\tau = \gamma$) and MAP decoders for transmission of the Gauss–Markov source A ($M = 8$, $N = 1$) over a binary-symmetric channel when different levels of residual redundancy is exploited at the decoder. Note that for $\gamma = 0$ and $\gamma = 0$ EPA the curves of AOMAP and MAP decoding have overlapped.

closely to the MAP algorithm in the mean-square error (MSE) sense. For any given delay of δ and redundancy order γ, similar observations are made in other cases, when comparing the AOMAP and the AOSMAP decoders (with the same τ) or the MAP and the SMAP decoders. Consequently, we will only discuss the performance of the MAP and AOMAP algorithms in the following sections. As seen in Section IV, the SMAP and the MAP algorithms can be implemented with comparable complexity. A more precise complexity comparison depends on the design parameters such as δ and γ and the actual decoder implementation. The AOSMAP decoder maintains a lower level of complexity as compared to the AOMAP decoder.

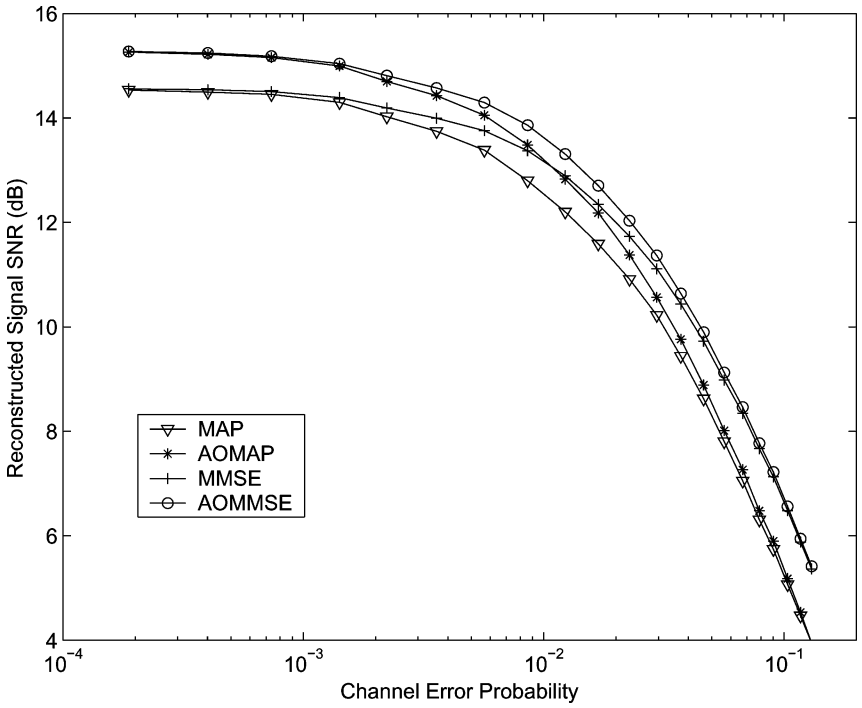

Fig. 4. Performance of the instantaneous ($\delta = 0$) MAP, AOMAP, MMSE, and AOMMSE decoders for transmission of the Gauss–Markov source A ($M = 8$, $N = 1$) over a binary-symmetric channel when the redundancy order $\gamma = 2$ and $\tau = \gamma$.

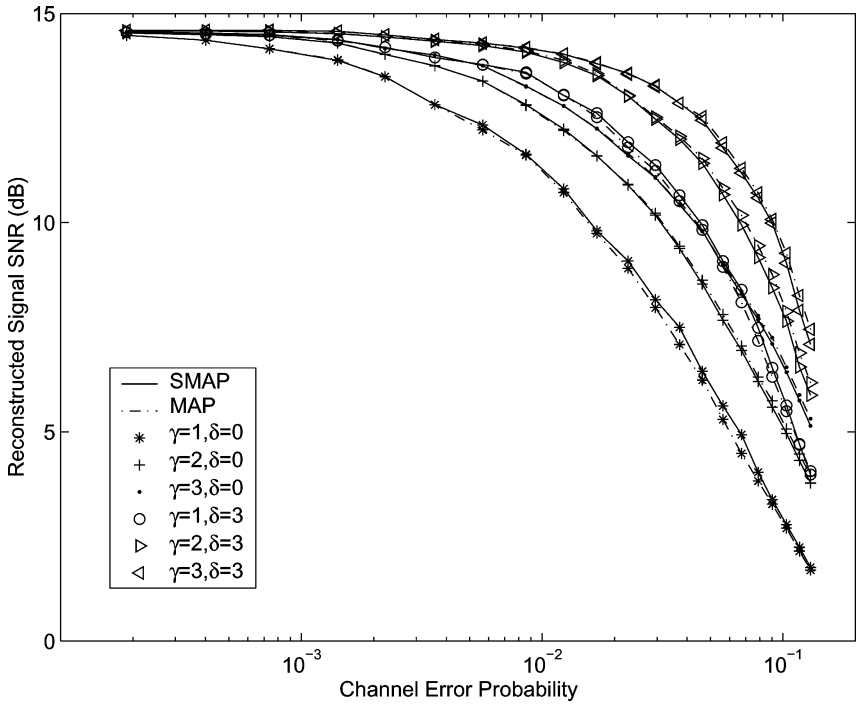

Fig. 5. Performance of the SMAP and MAP decoders for transmission of the Gauss–Markov source A ($M = 8$, $N = 1$) over a binary-symmetric channel when different levels of residual redundancy is exploited at the decoder, $\delta = 1, 3$.

B. Effect of Delay

To demonstrate the effect of delay, Fig. 6 depicts the performance of the MMSE decoder for reconstruction of the source A over a binary-symmetric channel at different delays ($\delta = 0, 1, 2, 3$) for the two scenarios of redundancy order $\gamma = 1$ and $\gamma = 3$. The curve corresponding to the basic MMSE decoder ($\gamma = 0$) of (21) provides a baseline for comparison. Also, the performance of the MAP decoder in similar scenarios are provided in Fig. 7. It is observed that for the case of transmission of source A over a noisy channel, a delay of $\delta = 3$ allows the decoder to capture almost all of the redundancy in the future samples. The gains achieved in this case are higher than 3.5, 3, and

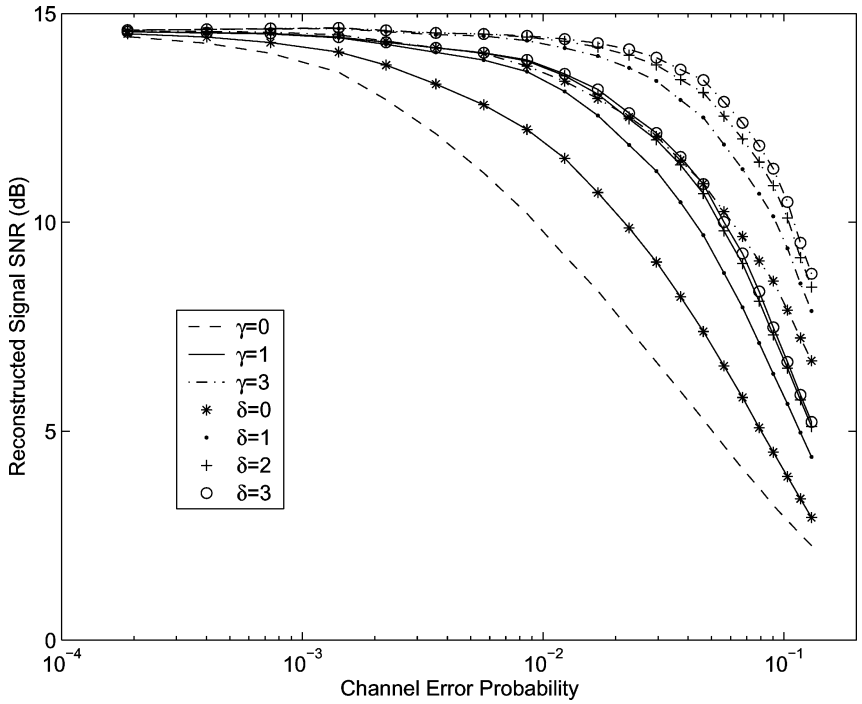

Fig. 6. Performance of the MMSE decoder for transmission of the Gauss–Markov source A ($M = 8$, $N = 1$) over a binary-symmetric channel when different delays are allowed ($\delta = 0, 1, 2, 3$) and the residual redundancy is exploited with a $\gamma = 1, 3$ order Markov model.

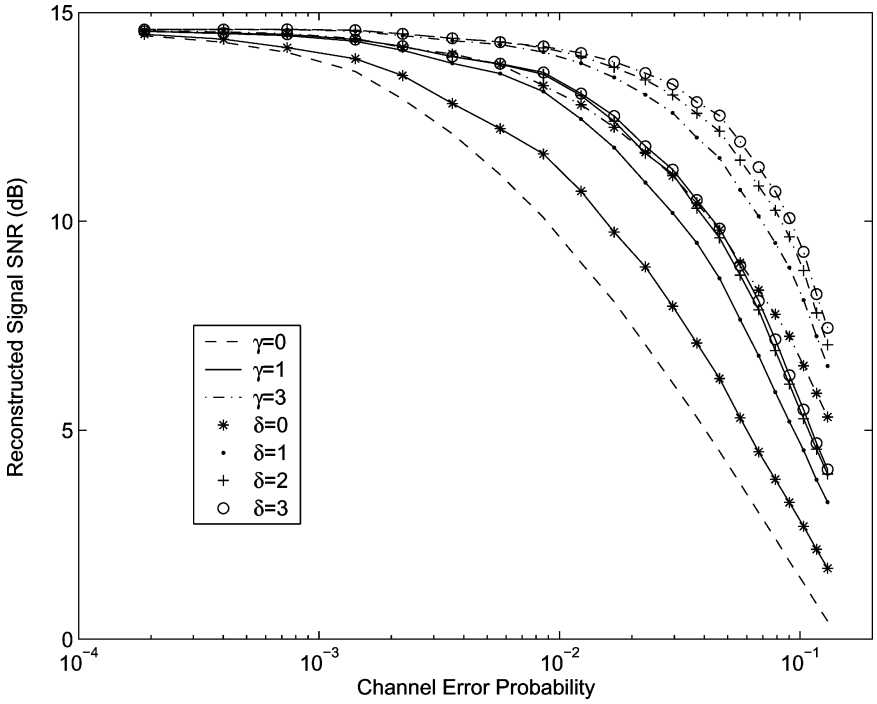

Fig. 7. Performance of the MAP decoder for transmission of the Gauss–Markov source A ($M = 8$, $N = 1$) over a binary-symmetric channel when different delays are allowed ($\delta = 0, 1, 2, 3$) and the residual redundancy is exploited with a $\gamma = 1, 3$ order Markov model.

2.5 dB for $\gamma = 1$, $\gamma = 2$, and $\gamma = 3$, respectively, when compared with the corresponding instantaneous decoding schemes.

C. Performance Using a Soft-Output Channel (Decoder)

Recently, channel decoding techniques using the soft channel information has found increasing attention in different applications for their improved performance. In techniques such as turbo decoding, iterative decoding, or soft-output Viterbi algorithm, soft outputs are readily available at the output of the channel decoder as well. Employing the soft outputs for improved source decoding have been shown to be fruitful, e.g., [51]. The source decoders proposed in this work are able to exploit the soft-output information and the source *a priori* information for effective source decoding. Alternatively, with appropriate considerations the proposed MAP (SMAP) decoders can be used for effective channel decoding using the

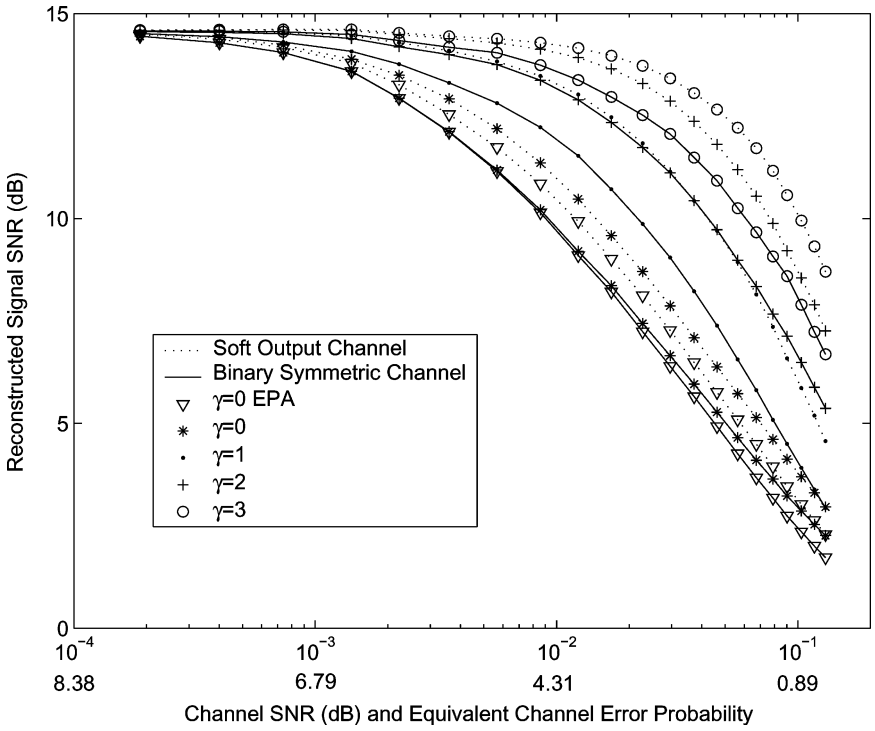

Fig. 8. Performance of the MMSE decoder for transmission of the Gauss–Markov source A ($M = 8$, $N = 1$) over the soft-output channel and the binary-symmetric channel when different levels of source redundancy are exploited at the decoder, $\delta = 0$.

soft channel information and assisted with the source *a priori* information.

To indicate the possible performance improvement due to using the soft output of the channel (decoder), Fig. 8 compares the performance of the instantaneous MMSE decoder for reconstruction of source A transmitted over the soft-output channel and the binary-symmetric channel. Alternatively, Fig. 9 depict the same performance results when MAP decoding is used. The performance results for the case, where a delay of $\delta = 3$ is allowed in the decoding process is available in [48]. It is observed that if channel (decoder) soft outputs are available, gains as high as 2–3 dB can be achieved. As the decoding schemes become stronger, i.e., δ and γ are increased, the maximum gains achieved move more toward the low channel signal-to-noise ratios (SNRs) or higher probabilities of error.

D. Effect of Redundancy

In this subsection, we study the effect of the type of source redundancy in the achievable gains using the proposed techniques. As well, we examine the effectiveness of the measures of redundancy as discussed before. We consider the instantaneous MMSE reconstruction of the sources A, B, and C over the soft-output channel as given in Figs. 8, 10, and 11, respectively. From these figures, it is observed that the amount of redundancy $R(M,\gamma)$, as defined in (45) and provided in Table II, correlates well with the achieved gains. On the other hand, the source autocorrelation as given in Table III does not seem to be a suitable indicator of the possible gains. This is in line with the observations in [22].

E. Effect of Quantizer Bitrate

Fig. 12 depicts the performance of the instantaneous MMSE decoder for transmission of source A quantized with an $M = 4, 8, 16$ point quantizer over the soft-output channel model. Since the higher rate quantizers are more sensitive to the channel errors, the effectiveness of the proposed decoder is more significant in such cases. Specifically, the gains at low error rates are noticeable.

VI. Conclusion

A family of solutions for the asymptotically optimum MMSE reconstruction of a source over a memoryless noisy channel is presented when the redundancy in the source encoder output stream is exploited in the form of a γ-order Markov model ($\gamma \geq 1$) and a delay of δ, $\delta > 0$, is allowed in the decoding process. Considering the same problem setup, we also present a simplified MMSE decoder as well as several other MAP symbol and sequence decoders. In each case, we investigate the alternative solutions and optimize them for the smallest computational complexity.

The numerical results and analysis demonstrate the effectiveness of the stronger models (higher Markov order γ) to capture the residual redundancy. The MMSE-based decoders outperform their equivalent MAP-based decoders. As expected, the asymptotically optimum MMSE (AOMMSE) decoder provides the best performance among the presented decoders. The simplified MMSE decoder has a smaller decoder codebook and a lower complexity, which is comparable to that of the SMAP decoder. The sequence MAP decoder and the symbol MAP decoder maintain the same level of performance. The AOSMAP

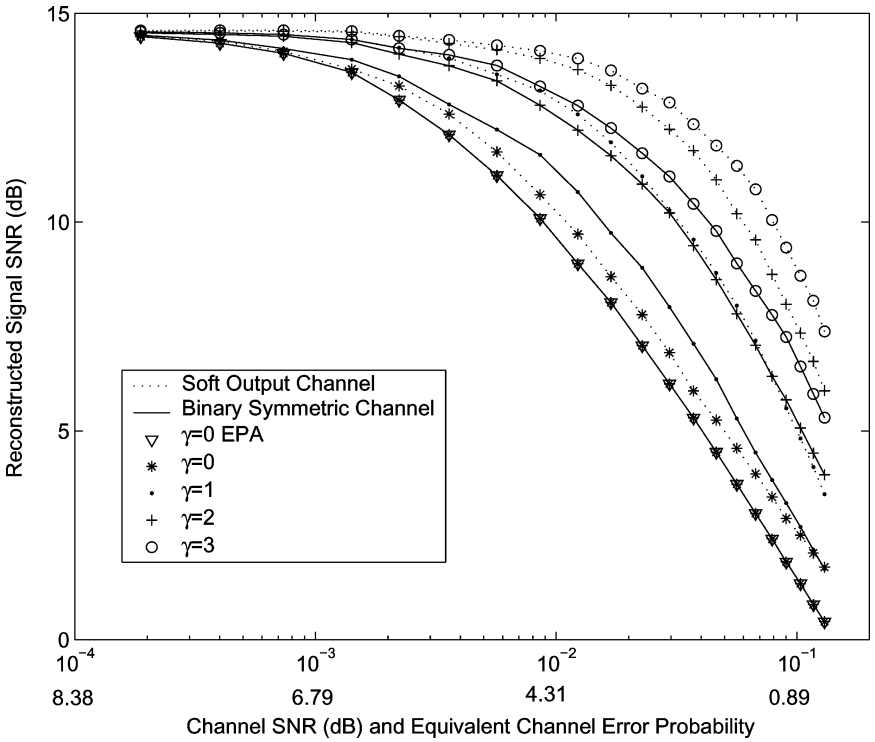

Fig. 9. Performance of the MAP decoder for transmission of the Gauss–Markov source A ($M = 8, N = 1$) over the soft-output channel and the binary-symmetric channel when different levels of source redundancy is exploited at the decoder, $\delta = 0$. Note that the curves corresponding to binary-symmetric channel with $\gamma = 0$, $\gamma = 0$ EPA, and soft-output channel with $\gamma = 0$ EPA have overlapped.

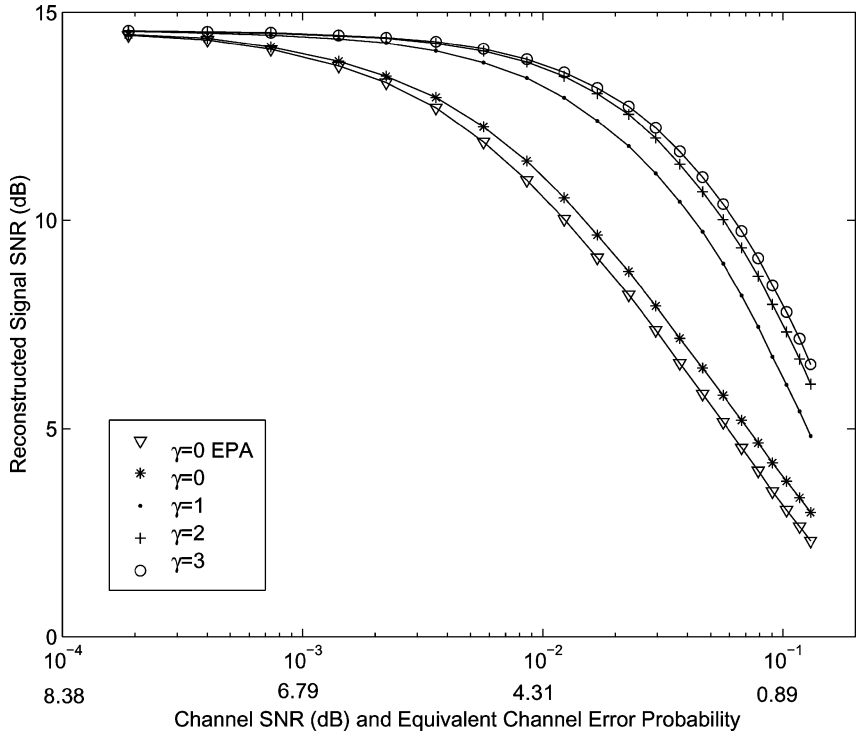

Fig. 10. Performance of the instantaneous MMSE decoder for transmission of the Gauss–Markov source B ($M = 8, N = 1$) over the soft-output channel when different levels of source redundancy are exploited.

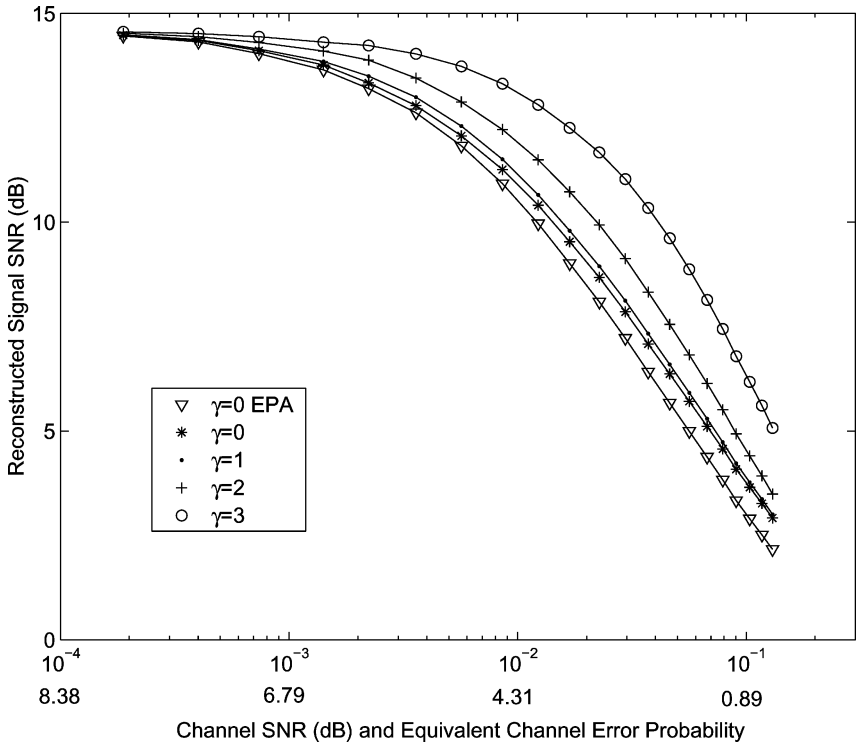

Fig. 11. Performance of the instantaneous MMSE decoder for transmission of the Gauss–Markov source C ($M = 8$, $N = 1$) over the soft-output channel when different levels of source redundancy are exploited.

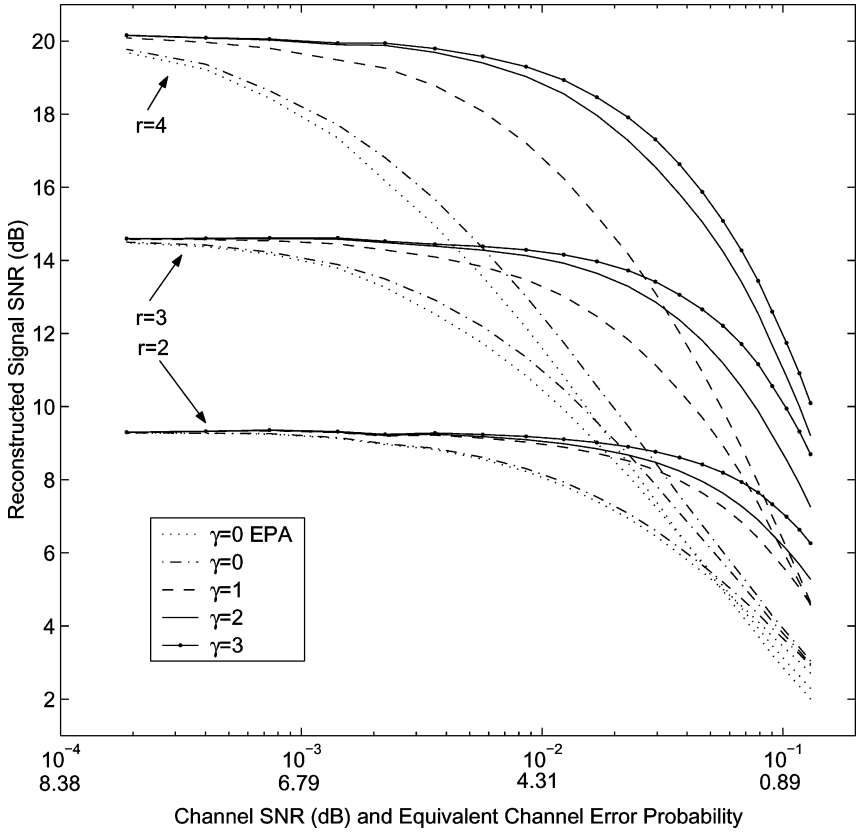

Fig. 12. Performance of the instantaneous MMSE decoder for transmission of the Gauss–Markov source A (quantized with rates $r = 2$, 3, and 4 bits, $N = 1$) over the soft-output channel when different levels of source redundancy are exploited, $M = 2^r$, $N = 1$.

$$\begin{aligned}P(\underline{I}_n^{n-(\gamma+L)+1}|\underline{J}_n) &= C_1.P(\underline{I}_n^{n-(\gamma+L)+1},\underline{J}_n)\\ &= C_1.P(SS_{n-(L-L')},\underline{I}_n^{n-(L-L')+1},\underline{J}_{n-(L-L')},\underline{J}_n^{n-(L-L')+1})\\ &= C.P(SS_{n-(L-L')}|\underline{J}_{n-(L-L')}).P(\underline{I}_n^{n-(L-L')+1}|SS_{n-(L-L')},\underline{J}_{n-(L-L')})\\ &\quad .P(\underline{J}_n^{n-(L-L')+1}|SS_{n-(L-L')},\underline{J}_{n-(L-L')},\underline{I}_n^{n-(L-L')+1}).\end{aligned}$$

decoder provides a lower complexity alternative to the AOMMSE decoder at the price of a certain loss in performance.

The possible future research in this direction includes the design of Channel Optimized Vector Quantizers based on the proposed decoders and more efficient approximate algorithms to the presented MMSE decoders.

APPENDIX

A. Proofs for Section II-C

Equation (2) is derived using the memoryless channel assumption of (1), as follows:

$$\begin{aligned}P(J_n|\underline{I}_n) &= \sum_{\underline{J}_{n-1}} P(J_n|\underline{I}_n,\underline{J}_{n-1})P(\underline{J}_{n-1}|\underline{I}_n)\\ &= \sum_{\underline{J}_{n-1}} P(J_n|I_n)P(\underline{J}_{n-1}|\underline{I}_n) \qquad (46)\\ &= P(J_n|I_n)\sum_{\underline{J}_{n-1}} P(\underline{J}_{n-1}|\underline{I}_n)\\ &= P(J_n|I_n). \qquad (47)\end{aligned}$$

Equation (1) is used in transition from (46) to (47).

Equation (3) is derived as follows:

$$\begin{aligned}&P(\underline{J}_n|\underline{I}_n)\\ &= \frac{1}{P(\underline{I}_n)}P(\underline{J}_n,\underline{I}_n)\\ &= \frac{1}{P(\underline{I}_n)}P(I_1)P(J_1|I_1)P(I_2|J_1,I_1)P(J_2|\underline{I}_2,J_1),\ldots,\\ &\quad .P(I_{n-1}|\underline{J}_{n-2},\underline{I}_{n-2})P(J_{n-1}|\underline{I}_{n-1},\underline{J}_{n-2})\\ &\quad .P(I_n|\underline{J}_{n-1},\underline{I}_{n-1})P(J_n|\underline{I}_n,\underline{J}_{n-1}). \qquad (48)\end{aligned}$$

Due to the causality of the communication channel and its independent with the input signal, we have

$$P(I_n|\underline{J}_{n-1},\underline{I}_{n-1}) = P(I_n|\underline{I}_{n-1}), \qquad n=1,2,\ldots$$

and using (1), (48) is now simplified to

$$\begin{aligned}P(\underline{J}_n|\underline{I}_n) &= \frac{1}{P(\underline{I}_n)}P(I_1)(I_2|I_1),\ldots,P(I_n|\underline{I}_{n-1})\\ &\quad .\prod_{k=1}^{n}P(J_k=j_k|I_k=i_k)\\ &= \prod_{k=1}^{n}P(J_k=j_k|I_k=i_k). \qquad (49)\end{aligned}$$

B. Proofs for Section III-A2)

Equation (16) provides the probability of a sequence of symbols within the structure of the trellis, given the entire history of the received signals. The derivation of this equation is presented in the following. Note that $L > L' \geq 0$ and for $L' = 0$ this collapses to the case with the original source trellis presented in (14) (see the equation at the top of the page). Using the memoryless property of the channel and the Markovian property of the source, this is simplified to

$$\begin{aligned}P(\underline{I}_n^{n-(\gamma+L)+1}|\underline{J}_n) &= C\,.\,P(\underline{J}_n^{n-(L-L')+1}|\underline{I}_n^{n-(L-L')+1})\\ &\quad .P(\underline{I}_n^{n-(L-L')+1}|SS_{n-(L-L')})\,.\,P(SS_{n-(L-L')}|\underline{J}_{n-(L-L')})\\ &= C.\left[\prod_{k=0}^{L-L'-1}P(J_{n-k}|I_{n-k})P(I_{n-k}|SS_{n-k-1})\right]\\ &\quad .P(SS_{n-(L-L')}|\underline{J}_{n-(L-L')}) \qquad (50)\end{aligned}$$

where in (50), $C = P(\underline{J}_{n-(L-L')})/P(\underline{J}_n)$. The overall computational complexity of the (50) is given by

$$CC = 2(L-L'+1)M^{\gamma+L} + (M+2)M^{\gamma+L'}. \qquad (51)$$

Now, we show that for a fixed value of L, $L > 0$, increasing the value of L', $0 \leq L' < L$, or the state set extension factor, reduces this complexity. The case of $L = 1$ requires $L' = 0$ and is trivial. For a given L, $L > 1$, $\forall L'$, $0 < L' < L$, we have

$$\begin{aligned}&CC(L,L') - CC(L,L'-1)\\ &\quad = -2M^{\gamma+L} + M^{\gamma+L'-1}(M^2+M-2)\\ &\quad = M^{\gamma+L'-1}(-2M^{L-L'+1} + M^2 + M - 2)\\ &\quad \leq M^{\gamma+L'-1}(-M^2+M-2) < 0\end{aligned}$$

which proves the point.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers, whose constructive comments substantially improved this work.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communications," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[2] J. W. Modestino, D. G. Daut, and A. L. Vickers, "Combined source channel coding of images using the block cosine transform," *IEEE Trans. Commun.*, vol. COM-29, pp. 1262–1274, Sept. 1981.

[3] C. C. Moore and J. D. Gibson, "Self-orthogonal convolutional coding for the DPCM-AQB speech encoder," *IEEE Trans. Commun.*, vol. COM-32, Aug. 1984.

[4] M. Bystrom and J. W. Modestino, "Combined source-channel coding schemes for video transmission over an additive white Gaussian noise channel," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 880–890, June 2000.

[5] B. Hochwald and K. Zeger, "Tradeoff between source and channel coding," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1412–1424, Sept. 1997.

[6] A. Méhes and K. Zeger, "Performance of quantizers on noisy channels using structured families of codes," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2468–2476, Nov. 2000.

[7] *TDMA Radio Interface, Enhanced Full-Rate Speech Codec*, Std. TIA/EIA PN-3467, 1996.

[8] C.-E. Sundberg, "The effect of single bit errors in standard nonlinear PCM systems," *IEEE Trans. Commun.*, vol. COM-24, pp. 1062–1064, June 1976.

[9] S. Gadkari and K. Rose, "Unequally protected multistage vector quantization for time-varying CDMA channels," *IEEE Trans. Commun.*, vol. 49, pp. 1045–1054, June 2001.

[10] I. Kozintsev and K. Ramchandran, "Robust image transmission over energy-constrained time-varying channels using multiresolution joint source-channel coding," *IEEE Trans. Signal Processing*, vol. 46, pp. 1012–1026, Apr. 1998.

[11] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2325–2383, Oct. 1998.

[12] A. J. Kurtenbach and P. A. Wintz, "Quantizing for noisy channels," *IEEE Trans. Commun.*, vol. COM-17, pp. 291–302, Apr. 1969.

[13] K. Y. Chang and R. W. Donaldson, "Analysis, optimization, and sensitivity study of differential PCM systems operating on noisy communication channels," *IEEE Trans. Commun.*, vol. COM-20, pp. 338–350, June 1972.

[14] H. Kumazawa, M. Kasahara, and T. Namekawa, "A construction of vector quantizers for noisy channels," *Electron. Eng. Japan*, vol. 67-B, no. 4, pp. 39–47, 1984.

[15] N. Farvardin and V. Vaishampayan, "On the performance and complexity of channel-optimized vector quantizers," *IEEE Trans. Inform. Theory*, vol. 37, pp. 155–160, Jan. 1991.

[16] J. G. Dunham and R. M. Gray, "Joint source and channel trellis encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 516–519, July 1981.

[17] E. Ayanoglu and R. M. Gray, "The design of joint source and channel trellis waveform coders," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 855–865, Nov. 1987.

[18] H. Jafarkhani and N. Farvardin, "Design of channel-optimized vector quantizers in the presence of channel mismatch," *IEEE Trans. Commun.*, vol. 48, pp. 118–124, Jan. 2000.

[19] Z. Guang-Chong and F. I. Alajaji, "Soft-decision COVQ for turbo-coded AWGN and Rayleigh fading channels," *IEEE Commun. Lett.*, vol. 5, pp. 257–259, June 2001.

[20] P. Hedelin, P. Knagenhjelm, and M. Skoglund, "Theory for transmission of vector quantization data," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. New York: Elsevier Science, 1995, pp. 347–396.

[21] ——, "Vector quantization for speech transmission," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. New York: Elsevier Science, 1995, pp. 311–345.

[22] K. Sayood and J. C. Borkenhagen, "Use of residual redundancy in the design of joint source/channel coders," *IEEE Trans. Commun.*, vol. 39, pp. 838–845, June 1991.

[23] N. Phamdo and N. Farvardin, "Scalar quantization of memoryless sources over memoryless channels using rate-one convolutional codes," in *Proc. IEEE Int. Symp. Information Theory (ISIT '94)*, Trondheim, Norway, p. 235.

[24] J. Hagenauer, "Source-controlled channel decoding," *IEEE Trans. Commun.*, vol. 43, pp. 2449–2457, Sept. 1995.

[25] K. Sayood, L. Fuling, and J. D. Gibson, "A constrained joint source/channel coder design," *IEEE Trans. Select. Areas Commun.*, vol. 12, pp. 1584–1593, Dec. 1994.

[26] F. I. Alajaji, N. Phamdo, and T. E. Fuja, "Channel codes that exploit the residual redundancy in CELP-encoded speech," *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 325–336, Sept. 1996.

[27] A. Ruscitto and E. M. Biglieri, "Joint source and channel coding using turbo codes over rings," *IEEE Trans. Commun.*, vol. 46, pp. 981–984, Aug. 1998.

[28] T. Fazel and T. E. Fuja, "Joint source-channel decoding of block-encoded compressed speech," in *Proc. Conf. Information Sciences and Systems*, Mar. 2000, pp. FA5-1–FA5-6.

[29] T. Fingscheidt and P. Vary, "Softbit speech decoding: A new approach to error concealment," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 240–251, Mar. 2001.

[30] F. Lahouti and A. K. Khandani, "Approximating and exploiting the residual redundancies- Applications to efficient reconstruction of speech over noisy channels," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, Salt Lake City, UT, May 2001, pp. 721–724.

[31] M. Adrat, J. Spittka, S. Heinen, and P. Vary, "Error concealment by near optimum MMSE-estimation of source codec parameters," in *Proc. IEEE Workshop on Speech Coding*, 2000, pp. 84–86.

[32] F. Lahouti and A. K. Khandani, "An efficient MMSE source decoder for noisy channels," in *Proc. Int. Symp. Telecommunications*, Tehran, Iran, Sept. 2001, pp. 784–787.

[33] ——, "Sequence MMSE source decoding over noisy channels using the residual redundancies," in *Proc. Ann. Allerton Conf. Communication, Control and Computing*, Urbana-Champaign, IL, Oct. 2001.

[34] N. Phamdo and N. Farvardin, "Optimal detection of discrete Markov sources over discrete memoryless channels—Applications to combined-source channel coding," *IEEE Trans. Inform. Theory*, vol. 40, pp. 186–193, Jan. 1994.

[35] D. J. Miller and M. Park, "A sequence-based approximate MMSE decoder for source coding over noisy channels using discrete hidden Markov models," *IEEE Trans. Commun.*, vol. 46, pp. 222–231, Feb. 1998.

[36] V. Kafedziski and D. Morrell, "Vector quantization over Gaussian channels with memory," in *Proc. IEEE Int. Communications Conf. (ICC '95)*, 1995, pp. 1433–1437.

[37] N. Phamdo, F. Alajaji, and N. Farvardin, "Quantization of memoryless and Gauss-Markov sources over binary Markov channels," *IEEE Trans. Commun.*, vol. 45, pp. 668–675, June 1997.

[38] M. Skoglund, "Soft decoding for vector quantization over noisy channels with memory," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1293–1307, May 1999.

[39] ——, "Bit-estimate based decoding for vector quantization over noisy channels with intersymbol interference," *IEEE Trans. Commun.*, pp. 1309–1317, Aug. 2000.

[40] T. Hindelang, T. Fingscheidt, N. Seshadri, and R. V. Cox, "Combined source/channel (de-)coding: Can a priori information be used twice?," in *Proc. IEEE Int. Commun. Conf.*, vol. 3, 2000, pp. 1208–1212.

[41] ——, "Combined source/channel (de-)coding: Can a priori information be used twice?," in *Proc. IEEE Int. Symp. Information Theory*, Sorrento, Italy, 2000, p. 266.

[42] N. Görtz, "On the iterative approximation of optimal joint source-channel decoding," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1662–1670, Sept. 2001.

[43] M. Adrat, P. Vary, and J. Spittka, "Iterative source-channel decoder using extrinsic information from softbit-source decoding," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 4, Salt Lake City, UT, May 2001, pp. 2653–2656.

[44] T. Fingscheidt, T. Hindelang, R. V. Cox, and N. Seshadri, "Joint source-channel (de-)coding for mobile communications," *IEEE Trans. Commun.*, vol. 50, pp. 200–212, Feb. 2002.

[45] S. Emami and S. L. Miller, "DPCM picture transmission over noisy channels with the aid of a Markov model," *IEEE Trans. Image Processing*, vol. 4, pp. 1473–1481, Nov. 1995.

[46] R. Link and S. Kallel, "Optimal use of Markov models for DPCM picture transmission over noisy channels," *IEEE Trans. Image Processing*, vol. 48, pp. 1702–1711, Oct. 2000.

[47] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, Mar. 1974.

[48] F. Lahouti, "Quantization and reconstruction of sources with memory," Ph.D., Dept. Elec. Comp. Eng., Univ. Waterloo, Waterloo, ON, Canada, 2002.

[49] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inform. Theory*, vol. 47, pp. 498–519, Feb 2001.

[50] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[51] V. A. Vaishampayan and N. Farvardin, "Joint design of block source codes and modulation signal sets," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1230–1248, July 1992.