ON SCALABLE CODING OF HIDDEN MARKOV SOURCES

*Mehdi Salehifar, Tejaswi Nanjundaswamy, and Kenneth Rose**

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA, 93106
E-mail:{salehifar, tejaswi , rose}@ece.ucsb.edu

ABSTRACT

While several real world signals, such as speech, image and sensor network data, are modeled as hidden Markov sources (HMS) for recognition and analysis applications, their typical compression exploits temporal correlations by modeling them as simple (non-hidden) Markov sources. However, the inherent hidden Markov nature of these sources implies that an observed sample depends, in fact, on all past observations, thus rendering simple Markov modeling suboptimal. Motivated by this realization, previous work from our lab derived a technique to optimally quantize and compress HMS. In this paper we build on, and considerably extend, these results to the problem of scalable coding of HMS. At the base layer, as proposed earlier, the approach tracks an estimate of the state probability distribution and adapts the encoder structure accordingly. At the enhancement layer, the state probability distribution is refined using available information from past enhancement layer reconstructed samples, and this refined estimate is further combined with quantization information from the base layer, to effectively characterize the probability density of the current sample conditioned on all available information. We update code parameters on the fly, at each observation, at both the encoder and the decoder and at both layers. Experimental results validate the superiority of the proposed approach with considerable gains over standard predictive coding employing a simple Markov model.

Index Terms— Scalable Coding, Hidden Markov Sources.

1. INTRODUCTION

The Hidden Markov model (HMM) is a discrete-time Markov chain observed through a memoryless channel. The random process consisting of the sequence of observations is referred to as a hidden Markov source (HMS). Markov chains are common models for information sources with memory, and the memoryless channel is among the simplest communication models. Thus HMMs are widely used in image understanding and speech recognition [1], source coding [2], communications, information theory, economics, robotics, computer vision and several other disciplines. Note that most signals modeled as Markov process are usually captured by imperfect sensors and are hence contaminated with noise, i.e., the resulting sequence is in fact an HMS. HMS is a special case of the broader family of multivariate Markov sources, which has been a focus of recent research, notably in the context of parameter estimation [3, 4]. Despite the importance of this model, there has been very limited work on quantizing HMS observations optimally and most practical applications simply assume a (non-hidden) Markov model for encoding these signals. The closest information theoretic work was on indirect rate distortion problems [5, 6], wherein, instead of the observations, the source hidden behind a layer of noise is encoded and reconstructed. In an alternative approach, a finite state quantizer was proposed in [7, 8, 9], but these techniques do not exploit all the available relevant information. Hence, in a recent paper from our lab [10], an optimal quantization scheme was proposed for HMS, which exploits all the available information on the source state. The probability distribution over the states of the underlying Markov chain captures all the available information and depends on the entire history of observations. Hence, we proposed to refine, with each observation, the estimate of the state probability distribution at both encoder and decoder, and correspondingly update the coding rule.

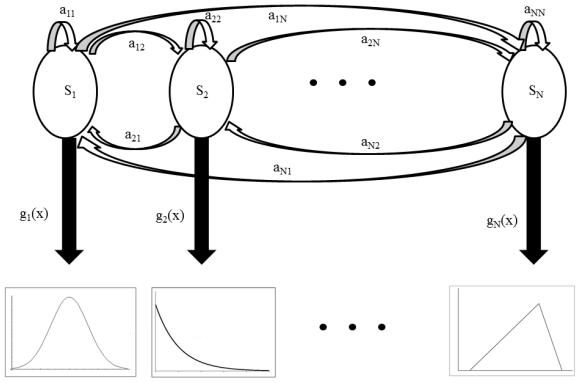

Fig. 1. A continuous emission hidden Markov source with N states.

This paper focuses on a significant extension of the optimal HMS quantization paradigm. Advances in internet and communication technologies, have created an extremely heterogeneous network scenario with data consumption devices of highly diverse decoding and display capabilities, all accessing the same content over networks of time varying bandwidth and latency. Thus it is necessary for coding and transmission systems to provide a scalable bitstream that allows decoding at a variety of bit rates (and corresponding levels of quality), where the lower rate information streams are embedded within the higher rate bitstreams in a manner that minimizes redundancy. That is, we need to generate layered bitstreams, wherein a base layer provides a coarse quality reconstruction and successive layers refine the quality, incrementally. Scalable coding with two quality levels, transmits at rate R_{12} to both decoders and at rate R_2 to only the second decoder.

*The work was supported in part by the National Science Foundation under grant CCF-1320599.

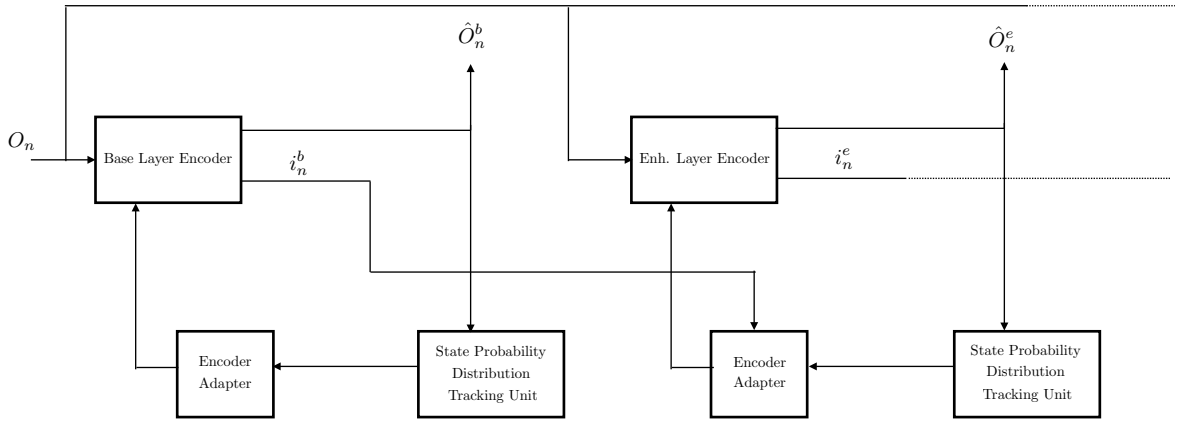

Fig. 2. Proposed encoder for the base and the enhancement layer.

The commonly employed technique for scalable coding of HMS, completely neglects the hidden states and employs a predictive coding approach such as DPCM (i.e., assumes a simple Markov model) at the base layer, and at the enhancement layer, the base layer reconstruction error is compressed and transmitted. That is, the base layer reconstruction is used as an estimate for the original signal, and the estimation error is compressed in the enhancement layer. A review of scalable coding techniques for audio and video signals can be found in [11, 12]. An estimation theoretically optimal scalable predictive coding technique was proposed in [13], which accounts for all the information available from the current base layer and past reconstructed samples to generate an optimal estimate at the enhancement layer. In this paper we propose a novel scalable coding technique for HMS, which accounts for all the available information while coding a given layer. At the base layer, we exploit all the available information by employing our previously proposed technique of tracking the state probability distribution at the encoder and the decoder, and using it to update the quantizers for encoding the current observation. At the enhancement layer, we again track the state probability distribution at the encoder and the decoder, but using the higher quality enhancement layer reconstruction for a better estimate, and then the enhancement layer quantizer is adapted to the interval determined by base layer quantization so as to enable full exploitation of all available information.

The rest of the paper is organized as follows. In Section 2, we define the problem. In Section 3, we present the proposed method. We substantiate the effectiveness of the proposed approach with comparative numerical results in Section 4, and conclude in Section 5.

2. PROBLEM DEFINITION

A hidden Markov source (HMS) is defined by the following set of parameters (please refer to [14] for more details):

1. Hidden states in the model: Assuming a finite number of the states, N, the set of all possible states is denoted $S = \{S_1, S_2, \ldots, S_N\}$, and the state at time t is denoted by q_t.
2. The state transition probability distribution: $A = \{a_{ij}\}$, where $a_{ij} = P[q_{t+1} = S_j | q_t = S_i], 1 \leq i, j \leq N$.
3. Observations: The set of observation symbols (or alphabet) is denoted by V, and in the discrete emission case $V = \{v_1, v_2, \cdots, v_M\}$, where M is the cardinality of the observation alphabet.
4. The observation (emission) probability density function (pdf) given state S_j, is denoted $g_j(\cdot)$, in the continuous emission case, and the observation (emission) probability mass function (pmf), $B = \{b_j(v_k)\}$, where $b_j(v_k) = p[O_t = v_k | q_t = S_j]$ in the discrete emission case, where O_t denotes the emitted observation at time t.
5. The initial state distributions $\pi = \{\pi_i\}$ where $\pi_i = P[q_1 = S_i], 1 \leq i \leq N$.

Fig. 1 depicts an example of a continuous hidden Markov source with N states. Clearly in hidden Markov sources, we have access only to the emitted observation, O_t, and not to the state of the source.

Our objective is to find the best scalable coding scheme for HMS.

3. PROPOSED METHOD

For optimal scalable coding of HMS, we need to exploit all the available information while encoding at each layer. To achieve this, we rely on the important HMS property that the state at time $t-1$, captures all the past information relevant to the emission of the next source symbol. Specifically, $P[O_t = v_k | q_{t-1} = S_j, O_{1\to(t-1)}] = P[O_t = v_k | q_{t-1} = S_j]$, which implies that all observations until time $t-1$ provide no additional information on the next source symbol, beyond what is captured by the state of the Markov chain at time $t-1$. Further note that the state of the HMS cannot be known with certainty. Thus the fundamental paradigm to optimally capture the correlations with the past is by tracking the state probability distribution of the HMS.

In this approach, each output of the encoder (quantized observation) at a given layer is sent into a unit called the *state probability distribution tracking function*. This unit estimates the state probability distribution, i.e., probabilities of the Markov chain being in each of the states, denoted by $\underline{\hat{p}}$. The base layer *encoder adapter unit* utilizes these probabilities to redesign the encoder optimally for the next input sample. At the enhancement layers, there is additional information of the quantization interval from the lower layers along with the state probability distribution. Thus the enhancement layer *encoder adapter unit* combines both types of available information to redesign the encoder optimally for the next input sample. Fig. 2 shows the proposed scalable encoder of HMS.

Our objective is to design the state probability distribution tracking function and the encoder adapter, given a training set of samples,

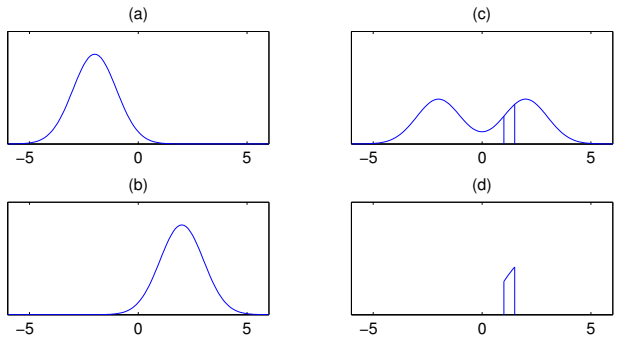

Fig. 3. Example of observation pdf. (a) and (b) are the observation pdf given states 1 and 2, $g_1(x)$ and $g_2(x)$ respectively. (c) is the overall observation pdf in the base layer given $\underline{\hat{p}}_{(t)} = \{0.5, 0.5\}$, $P(O_t|\underline{\hat{p}}_{(t)} = \{0.5, 0.5\})$. (d) is the overall observation pdf in the enhancement layer given $\underline{\hat{p}}_{(t)} = \{0.5, 0.5\}$ and the quantization interval information from the base layer (I_b), $P(O_t|\underline{\hat{p}}_{(t)} = \{0.5, 0.5\}, I_b)$.

so as to minimize the average reconstruction distortion at a given encoding rates for both base and enhancement layers. We first describe tracking of the state probability distribution, and then discuss the encoder adapter unit for the base layer and the enhancement layer.

3.1. State probability distribution tracking

We first estimate the HMS parameters using a training set, for which we follow standard procedure in HMM analysis. We then define forward variable at time $t-1$ as

$$\alpha_{t-1}(i) = P(O_{1\to(t-1)}, q_{t-1} = S_i), \tag{1}$$

i.e., the joint probability of emitting the observed source sequence up to time $t-1$, $O_{1\to(t-1)}$ and that the state at time $t-1$ is S_i. These forward variables can be computed recursively as given below:

1. $\alpha_1(i) = \pi_i g_i(O_1)$ for $1 \le i \le N$
2. $\alpha_{k+1}(j) = [\sum_{i=1}^{N} \alpha_k(i) a_{ij}] g_j(O_{k+1})$

We can then obtain probability of being in state i at time $t-1$, using Bayes' rule,

$$\begin{aligned} P[q_{t-1} = S_i|O_{1\to(t-1)}] &= \frac{P(q_{t-1} = S_i, O_{1\to(t-1)})}{P(O_{1\to(t-1)})} \\ &= \frac{P(q_{t-1} = S_i, O_{1\to(t-1)})}{\sum_{j=1}^{N} P(q_{t-1} = S_j, O_{1\to(t-1)})} \\ &= \frac{\alpha_{t-1}(i)}{\sum_{j=1}^{N} \alpha_{t-1}(j)} \end{aligned} \tag{2}$$

Using these we can compute the probability of being in state i at time t, $p_{(t,i)}$, given the observation sequence up to time $t-1$ as:

$$\begin{aligned} p_{(t,i)} &= P[q_t = S_i|O_{1\to(t-1)}] \\ &= \sum_{j=1}^{N} P[q_{t-1} = S_j|O_{1\to(t-1)}] a_{ij} \end{aligned} \tag{3}$$

The state probability distribution at time t, $\underline{p}_{(t)}$, is defined as,

$$\begin{aligned} \underline{p}_{(t)} &= \{p_{(t,1)}, \ldots, p_{(t,N)}\} \\ &= \{P[q_t = S_1|O_{1\to(t-1)}], \cdots, P[q_t = S_N|O_{1\to(t-1)}]\}. \end{aligned} \tag{4}$$

Given $\underline{p}_{(t)}$, the best estimate for the source distribution will be:

$$\sum_{j=1}^{N} p_{(t,j)} g_j(x) \tag{5}$$

3.2. Base Layer Design

At the base layer the state probability distribution tracking unit finds, $\underline{\hat{p}}^b$, using base layer reconstructed observation samples $\hat{O}^b$, as

$$\begin{aligned} \underline{\hat{p}}^b_{(t)} &= \{\hat{p}^b_{(t,1)}, \ldots, \hat{p}^b_{(t,N)}\} \\ &= \{P[q_t = S_1|\hat{O}^b_{1\to(t-1)}], \ldots, P[q_t = S_N|\hat{O}^b_{1\to(t-1)}]\}. \end{aligned} \tag{6}$$

Using the reconstructed observation, instead of the original samples, ensures the decoder can exactly mimic the operations of the encoder. Given $\underline{\hat{p}}^b_{(t)}$, the best estimate for the source distribution is:

$$\sum_{j=1}^{N} \hat{p}^b_{(t,j)} g_j(x) \tag{7}$$

One possible approach is to design a quantizer for this pdf for each source output from the ground up. But it clearly entails high complexity at the encoder and the decoder. We thus propose an alternative method with both low complexity and low memory requirement. For a set of T representative $\underline{p}$, we find the best codebook using Lloyd's (or other) algorithm offline. Then for a given $\underline{\hat{p}}^b_{(t)}$ in the process of encoding or decoding, we find the closest representative $\underline{p}$ from the set of T. Finally using the codebook of the closest representative as an initialization, we run one iteration of the Lloyd's algorithm, to find the codebook to be used for current sample.

Note that for the first symbol, we simply use a quantizer based on the assumption that the source symbol is from a fixed state (based on π) at both the encoder and decoder.

3.3. Enhancement Layer Design

The goal here is to design the best codebook for the enhancement layer based on:

- Quantization interval from the base layer, and
- State probability distribution based on the enhancement layer reconstructed observation samples $\hat{O}^e$.

The enhancement layer state probability distribution tracking unit finds $\underline{\hat{p}}^e_{(t)}$, similar to the base layer, but using enhancement layer reconstructed observation samples $\hat{O}^e$, as

$$\begin{aligned} \underline{\hat{p}}^e_{(t)} &= \{\hat{p}^e_{(t,1)}, \ldots, \hat{p}^e_{(t,N)}\} \\ &= \{P[q_t = S_1|\hat{O}^e_{1\to(t-1)}], \ldots, P[q_t = S_N|\hat{O}^e_{1\to(t-1)}]\}. \end{aligned}$$

Given just $\underline{\hat{p}}^e_{(t)}$, the best estimate for the source distribution is:

$$\sum_{j=1}^{N} \hat{p}^e_{(t,j)} g_j(x) \tag{8}$$

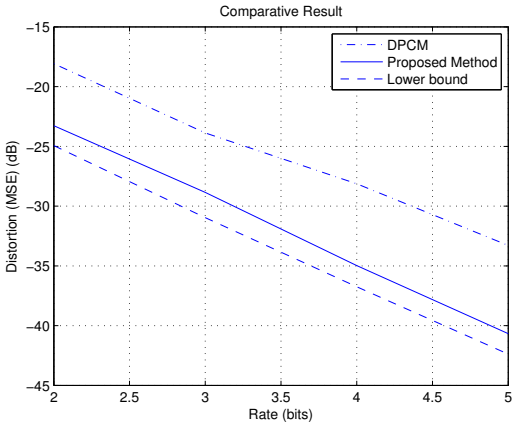

Fig. 4. Rate-Distortion plots for the proposed method, DPCM and the lower bound, at base layer rate of 3 and enhancement layer rate ranging from 2 to 5 bits per sample.

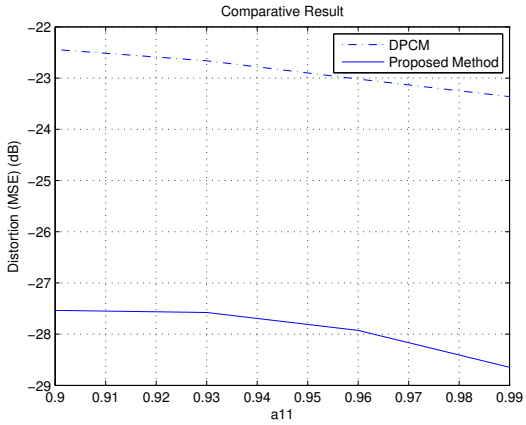

Fig. 5. Distortion plots for the proposed method and DPCM, at base and enhancement layer rates of 3 bits/sample, and transition probability, $a_{11} = a_{22}$, varying from 0.9 to 0.99.

However, in combination with the quantization interval information from the base layer, the best estimate for the source distribution is:

$$\sum_{j=1}^{N} \hat{p}^{e}_{(t,j)} \hat{g}_j(x) \tag{9}$$

where, $\hat{g}_j(x)$ is the observation pdf in state j truncated and normalized to the interval determined by quantization at the base layer.

We design the quantizer for this pdf of the current sample, by using uniform codebook as an initialization and running one iteration of Lloyd's algorithm. Note that we use uniform codebook as an initialization for the enhancement layer quantizer design as observation pdf within the quantization interval of the base layer has lesser variations and is closer to uniform distribution. However, this assumption is not true for the base layer, as illustrated by an example source distribution for a specific $\underline{\hat{p}}_{(t)}$ in Fig. 3. Using the uniform codebook as the initialization also reduces the memory requirements of the encoder and the decoder. For the first symbol, we assume that the source symbol is from a fixed state (based on π) at both the encoder and decoder, and use this in combination with the quantization interval information from the base layer.

Note that we can generalize our proposed approach to any number of enhancement layers, by combining the refined estimate of state probability distribution based on observation reconstruction of the given layer, with the quantization interval information from its lower layers.

4. EXPERIMENTAL RESULTS

For the the first experiment, we use a HMS which has two Gaussian subsources, one of them with mean $\mu_1 = -1.5$ and variance $\sigma_1^2 = 1$ the other one with mean $\mu_2 = +1.5$ and variance $\sigma_2^2 = 1$, and the transition probabilities are, $a_{11} = a_{22} = 0.99$. For the scalable coder, we set the base layer rate to be $R_{12} = 3$ bits and the enhancement layer rate varies as $R_2 = 2, 3, 4, 5$ bits. We compare the proposed method with the prior approach which assumes a simple Markov model and uses DPCM in the base layer and employs a quantizer designed via Lloyd's algorithm for encoding the base layer reconstruction error in the enhancement layer. Also we compare our results with the theoretical lower bound for using a switched scalar quantizer [9] operating at the total rate of $R_{12} + R_2$ bits. The rate distortion plots for the different approaches and the lower bound are shown in Fig. 4. The superiority of the proposed method is evident from the plots with substantial gains of around 5 dB over the prior approach. Note that there is a gap of around 2 dB from the lower bound due to the scalable coding penalty of the hierarchical structure employed, as this source distortion pair is not successively refinable.

In the second experiment, the same source is used, but with varying transition probability, $a_{11} = a_{22}$, from 0.9 to 0.99, and fixed coding rate of 3 bits in the base and the enhancement layer. Results for this experiment, as shown in Fig. 5, demonstrates the gains marginally increasing with values of a_{11}.

Note that calculating $\underline{\hat{p}}$ at base and enhancement layers of encoder and decoder does not impose any significant computational burden, as forward variables are easily updated recursively for each sample (as given in Section 3.1) and then $\underline{\hat{p}}$ is obtained with a few more manipulations.

5. CONCLUSION

We have proposed a novel technique for scalable coding of hidden Markov sources which utilizes all the available information while coding a given layer. Contrary to the existing approaches, which assume a simple Markov model, we use the hidden Markov model and exploit the dependency efficiently in the hidden states.

In the base layer, the state probability distribution is estimated for each sample using past base layer observation reconstruction and the quantizer for the current sample is updated accordingly. In the enhancement layer, the state probability distribution is refined for each sample using past enhancement layer observation reconstruction. Then this information is combined with the quantization interval available from the base layer, and the quantizer for the current sample is updated accordingly. The decoder mimics this quantizer updates of the encoder in both base and enhancement layers.

Experimental results show substantial performance improvements of the proposed approach over the prior approach of assuming a simple Markov model.

6. REFERENCES

[1] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden markov models to speaker-independent, isolated word recognition," *Bell System Technical Journal*, vol. 62, no. 4, pp. 1075–1105, 1983.

[2] A. Gersho and R. Gray, *Vector quantization and signal compression*, Springer, 1992.

[3] Y. Ephraim and B.L. Mark, "On forward recursive estimation for bivariate markov chains," *46th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, March 2012.

[4] Y. Ephraim and B.L. Mark, "Causal recursive parameter estimation for discrete-time hidden bivariate markov chains," *IEEE Transactions on Signal Processing*, vol. 63, no. 8, pp. 2108–2117, April 2015.

[5] R. L. Dobrushin and B. S. Tsybakov, "Information transmission with additional noise," *IRE Trans. Inform. Theory*, vol. IT-18, pp. S293–S304, 1962.

[6] H. S. Witsenhausen, "Indirect rate distortion problems," *IEEE Transactions on Information Theory*, vol. IT-26, no. 5, pp. 518–521, September 1980.

[7] J. Foster, R. Gray, and M. Dunham, "Finite-state vector quantization for waveform coding," *IEEE Transactions on Information Theory*, vol. 31, no. 3, pp. 348–359, 1985.

[8] M. Dunham and R. Gray, "An algorithm for the design of labeled-transition finite-state vector quantizers," *IEEE Transactions on Communications*, vol. 33, no. 1, pp. 83–89, 1985.

[9] D.M. Goblirsch and N. Farvardin, "Switched scalar quantizers for hidden markov sources," *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1455–1473, September 1992.

[10] M. Salehifar, E. Akyol, K. Viswanatha, and K. Rose, "On optimal coding of hidden markov sources," *IEEE Data Compression Conference*, March 2014.

[11] J. Ohm, "Advances in scalable video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 42–56, 2005.

[12] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[13] K. Rose and S. Regunathan, "Toward optimality in scalable predictive coding," *IEEE Transactions on Image processing*, vol. 10, no. 7, pp. 965–976, July 2001.

[14] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.