ON RELAXING THE STRICT HIERARCHICAL CONSTRAINTS IN LAYERED CODING OF AUDIO SIGNALS

Tejaswi Nanjundaswamy, Kumar Viswanatha and Kenneth Rose

Department of Electrical and Computer Engineering,
University of California, Santa Barbara, CA 93106-9560, USA
{tejaswi, kumar, rose}@ece.ucsb.edu

ABSTRACT

Scalable coders generate hierarchically layered bitstreams to serve content at different quality levels, wherein the base layer provides a coarse quality reconstruction and successive layers incrementally refine the quality. However, it is widely recognized that there is an inherent performance penalty due to the scalable coding structure, when compared to independently encoded copies. To mitigate this loss we propose a layered compression framework, having roots in information theoretic concepts, which relaxes the strict hierarchical constraints, wherein only a "subset" of the information of a lower quality level is shared with higher quality levels. In other words, there is flexibility to also have "private" information at each quality level, beside information that is common to multiple levels. We employ this framework within the MPEG Scalable AAC and propose an optimization scheme to jointly select parameters of all the layers. Experimental evaluation results demonstrate the utility of the flexibility provided by the proposed framework.

Index Terms— Audio compression, audio streaming, layered coding, scalable audio coding, joint optimization

1. INTRODUCTION

In today's networks, diverse data consumption devices (e.g., smartphones, tablets, smart-TVs, PC) may access the same multimedia content (e.g., live event broadcast or precompressed and stored content) over networks of time varying bandwidth and latency. This implies that the same content is served to various users at different data rates and quality levels. Obviously, different quality levels of a signal share some common information between them. Thus, it is beneficial to exploit this shared information to ensure efficient use of resources for storage and transmission across the network. Note that allocation of resources in this scenario has to consider the cost of storing content at all the required quality levels at a data center (total storage rate), or cost of transmitting content at all the required quality levels to intermediary nodes in the network (total transmit rate, same as total storage rate) and cost of transmitting content to various users at their corresponding quality requirement over the final links (total receive rate). In the simplest approach, the signal is independently coded at different quality levels. This approach has the lowest total receive rate, as independent encoding is unconstrained and can ensure best performance at the given quality levels. However, it has the highest total transmit rate, as no common information is exploited across quality levels. An alternative approach, known as scalable coding [1, 2], encodes the signal in hierarchical layers, wherein the base layer provides a coarse quality reconstruction and successive layers refine the quality incrementally. This approach clearly has a low total transmit rate as the bitstream of a lower quality level is completely subsumed in the bitstream of higher quality level. However, it incurs a high total receive rate, as the rigid hierarchical structure is well known to compromise the performance at higher layers. These two approaches clearly operate close to the two extremes in term of the trade-off between total transmit rate and total receive rate. Most audio content providers resort to independently encoding different quality levels, to avoid the scalable coding penalty.

Drawing inspiration from information theoretic principles related to the concept of common information [3], we propose a layered compression framework with relaxed hierarchical constraints, wherein only a (properly selected) subset of the information at a lower quality level is shared with the higher quality level, thus providing the flexibility of having private information at each quality level. This flexibility provides the opportunity to optimally encode the common information between the quality levels separately and also importantly provides the opportunity to control the layered coding penalty and achieve intermediate operating points in terms of the trade-off between total transmit rate and total receive rate. In our recent work [4] we have derived theoretical (asymptotic) characterization of this framework and demonstrated via example its performance gains over conventional scalable coding in terms of asymptotic sum rate, when the source-distortion pair is not "successively refinable" (i.e., scalability penalty is unavoidable even at asymptotic delay).

This work was supported in part by the Broadcom Foundation, and the NSF under grants CCF-1016861 and CCF-1320599

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

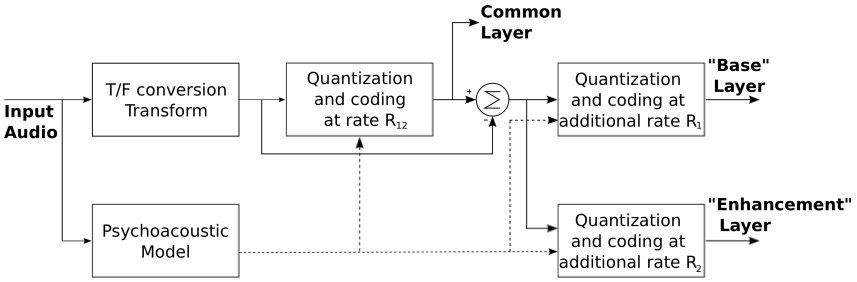

Fig. 1. The proposed audio encoder structure for two quality levels

In this paper we employ this framework within MPEG scalable advanced audio coder (AAC) [5, 6] for two quality levels with three layers: common, base and enhancement layers. Note that the common layer information is sent to both the base and enhancement decoders, while the base and enhancement layers contain information that is "private" for their respective layers. We then propose a technique for joint optimization of all the layers to minimize an overall cost which achieves a specific trade-off between total transmit rate and total receive rate. The optimization technique is based on our previously proposed scalable coder optimization [7], wherein parameters are optimized iteratively at each layer, while accounting for the impact of encoding decisions in each layer on the overall cost, until convergence. Experimental evaluation results demonstrate the ability of the proposed approach to achieve intermediate points in the trade-off between total transmit rate and total receive rate at a better overall cost than existing techniques. We note in passing some connections to the problem of multiple description coding [8] where descriptions are designed to individually achieve prescribed levels of distortion and improve performance for any subset of more descriptions. However, note that in our setting, the common layer only serves the purpose of aiding reconstruction in conjunction with other layers.

2. PROPOSED FRAMEWORK IN MPEG SCALABLE AAC

MPEG AAC is a transform domain perceptual audio coder that efficiently exploits statistical redundancy and psychoacoustic irrelevancy in audio signals. The AAC encoder segments audio signal into 50% overlapped frames, transforms each frame via the modified discrete cosine transform (MDCT), transform domain coefficients are then grouped into bands, all coefficients of a band are quantized using scaled version of a generic quantizer, and finally the quantized coefficients are entropy coded using one of the available Huffman codebooks (HCB). The scaling factor (SF) of the quantizer controls the distortion, and the HCB index controls the rate in each band. Psychoacoustic irrelevancy is exploited by controlling the noise in each band with respect to the masking threshold estimated by a psychoacoustic model. In this paper, we specifically minimize the distortion of maximum quantization noise to masking threshold ratio (MNMR) over all frequency bands of all frames. The AAC encoder optimization problem is of selecting the parameters of SFs and HCBs to minimize the perceptual distortion for a given rate, or minimize the rate given a distortion constraint. As the AAC standard only dictates the bitstream syntax and decoder, numerous encoder optimization techniques have been proposed, however, in this paper, we optimize the encoder via the trellis based search [9, 10].

Scalable AAC layers multiple AAC codecs, each usually encoding the reconstruction error of the preceding layer. Given a distortion constraint at each layer, the scalable AAC encoder optimizes SFs and HCBs to minimize the rate at each layer successively. For the proposed framework with two quality levels, we modify the scalable coder to include one common layer, providing coarse information that is then refined by the "base" and "enhancement" layers, as illustrated in Fig. 1. The overall paradigm is shown in Fig. 2, with total transmit rate, $R_t = R_{12} + R_1 + R_2$, and total receive rate, $R_r = 2R_{12} + R_1 + R_2$. Note that setting $R_{12} = 0$ in the framework results in independent encoding, and setting $R_1 = 0$ results in conventional scalable coding. Although in this paper we specifically address the case of two quality levels, the proposed framework can be extended to arbitrary number of levels by segregating common information corresponding to any subset of quality levels, similar to the structure proposed in our recent work on multiple descriptions coding [11]. However, note that this extended framework will entail combinatorial growth in the number of layers. Thus as an alternative, the proposed framework can be extended via a linearly growing rate-splitting approach, wherein a bitstream for each quality level includes an individual layer for itself, a common layer it shares with higher quality levels and other common layers from lower quality levels. Further analysis and evaluation of these extensions to arbitrary number of quality levels will be covered in future work.

3. JOINT OPTIMIZATION OF LAYERS

For two quality levels, let D_1, D_2 denote the distortion at the base and enhancement layers, respectively ($D_2 < D_1$) and

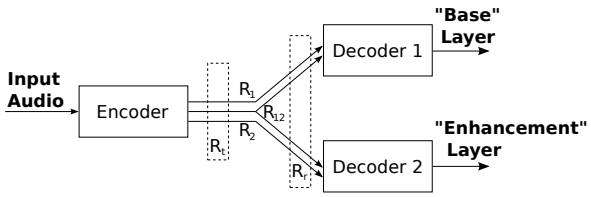

Fig. 2. The proposed paradigm for two quality levels

D_1^*, D_2^* be the corresponding distortion constraints ($D_2^* < D_1^*$). Given these constraints, we define the overall cost to optimize the trade-off between total transmit rate and total receive rate as,

$$\begin{aligned} J &= (1-\alpha)R_t + \alpha R_r \\ &= (1+\alpha)R_{12} + R_1 + R_2, \end{aligned} \quad (1)$$

where, α controls the trade-off. Note that while the base and enhancement layers have to satisfy distortion constraints (D_1^*, D_2^*), this is not the case for the common layer. It is hence viewed as a free (unconstrained) parameter and adjusted to minimize the overall cost. Let $\mathcal{P}$ denote the set of all parameters, which includes SFs and HCBs at all the layers and the distortion constraint at common layer. The optimization problem at hand is given as,

$$\begin{aligned} \mathcal{P}^* &= \arg\min_{\mathcal{P}} J \\ &\quad \text{s.t. } D_1 < D_1^*, D_2 < D_2^*. \end{aligned} \quad (2)$$

Let $\mathcal{P}_{12}$ denote the distortion constraint, SFs and HCBs at the common layer, and $\mathcal{P}_1, \mathcal{P}_2$ denote the SFs and HCBs at the base and enhancement layers, respectively. Joint optimization of all the layers is performed by successively and iteratively selecting parameters for each layer while accounting for impact on the overall cost in (1), similar to [7]. Optimization of a single layer to find SFs and HCBs given a distortion constraint is done via the trellis based search [9, 10]. To find the optimal distortion constraint of the common layer, the trellis based search is repeated for many distortion constraints and the one which minimizes the overall cost is retained. In every iteration of the overall algorithm, the common layer optimization minimizes (1) over all $\mathcal{P}_{12}$, while maintaining $\mathcal{P}_1, \mathcal{P}_2$ as unchanged from previous iteration. Note that, although $\mathcal{P}_1, \mathcal{P}_2$ are fixed at this step, changing $\mathcal{P}_{12}$ modifies the input to base and enhancement layer, thus not only R_{12}, but also R_1, R_2 have to be re-estimated for a new combination of $\mathcal{P}_{12}$. Next, the base layer parameters $\mathcal{P}_1$ are optimized, given fixed $\mathcal{P}_{12}$. In this step only R_1 is affected, so minimizing R_1 is equivalent to minimizing the overall cost (1). Similarly, the enhancement layer parameters $\mathcal{P}_2$ are optimized, given fixed $\mathcal{P}_{12}$, and only R_2 is affected, so minimizing R_2 is equivalent to minimizing the overall cost (1).

The first iteration of common layer needs an initialization of base and enhancement layers. Similar to [7], we adopt an "informed" initialization where every node in the common layer trellis for a given distortion constraint is associated with the best available "guess" of the base and enhancement layer parameters, and using these parameters the overall cost of (1) at every node of the trellis is calculated. Optimizing such trellis for multiple common layer distortion constraints gives us the initial values of $\mathcal{P}_{12}$. The overall algorithm can be summarized as:

1. Produce an "informed" initialization of common layer parameters, $\mathcal{P}_{12}$.
2. Repeat the following three steps until convergence, or a prespecified exit condition is met:
   (a) Optimize base layer to find parameters, $\mathcal{P}_1$, that minimize R_1, given the current choice of $\mathcal{P}_{12}$.
   (b) Optimize enhancement layer to find parameters, $\mathcal{P}_2$, that minimize R_2, given the current choice of $\mathcal{P}_{12}$.
   (c) Optimize common layer to find parameters, $\mathcal{P}_{12}$, that minimize overall cost in (1), given the current choice of $\mathcal{P}_1$ and $\mathcal{P}_2$.

Note that convergence is guaranteed as the overall cost is monotonically non-increasing in every step of the iteration. Also note that this iterative algorithm can be easily extended to optimize arbitrary number of layers.

4. EXPERIMENTAL RESULTS

For experimental evaluation we compared the following three coders:

- Single layer non-scalable AAC coder (NS-AAC)
- Conventional two layered scalable AAC coder (CS-AAC)
- Proposed framework based two quality level AAC coder (Prop-AAC)

All coders employ a simple psychoacoustic model with a fixed signal-to-mask ratio similar to the MPEG reference software. We observed that the proposed coder converges quickly due to the "informed" initialization, and thus we limit the algorithm described in Section 3 to just two iterations to minimize the impact on complexity. The audio test samples, which are single channel at 48 kHz, are obtained from the standard MPEG dataset. For time efficient evaluation only the first 5 seconds of each audio file is tested. The test samples include:

- Speech signal: vocal (vega)
- Single instrument: harpsichord (harp)
- Simple sound mixture: plucked strings (stri)

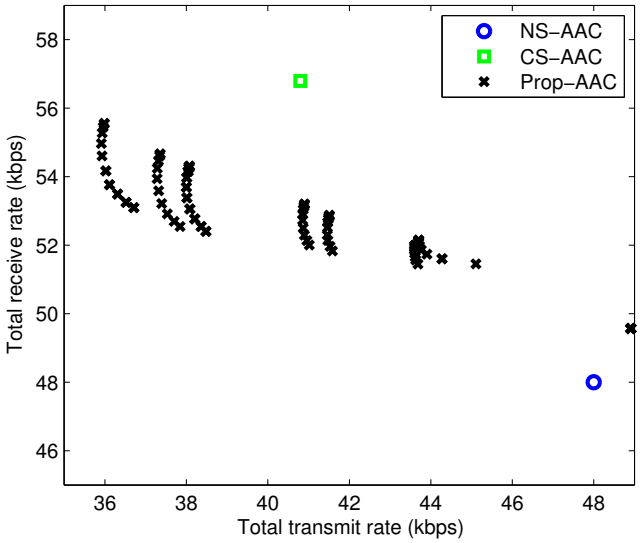

Fig. 3. Scatter plot of total transmit rate versus total receive rate achieved by the various coders

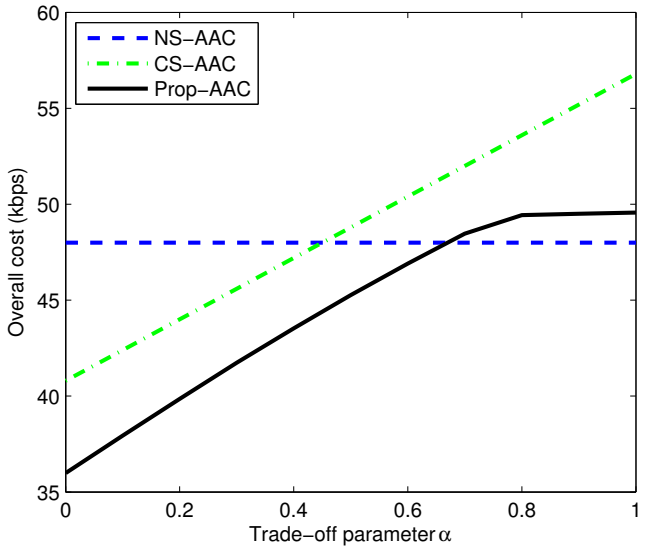

Fig. 4. Overall cost as given in equation (1) for the various coders

- Complex sound mixture: orchestra (orch)

The distortion constraints are chosen as $D_1^* = 7.1$ dB and $D_2^* = 2.13$ dB for the two levels such that the non-scalable AAC coders achieve bitrate of $R_1 = 16$ kbps and $R_2 = 32$ kbps. Obviously $R_{12} = 0$ for independently encoding at the two quality levels using the non-scalable coder. For this coder the total transmit rate and total receive rate are, $R_t = R_r = 48$ kbps. For a conventional scalable coder, these distortion constraints resulted in $R_{12} = 16$ kbps and $R_2 = 24.8$ kbps. Obviously $R_1 = 0$ for this coder, as there is no private information sent to the base layer. Consequently, the total transmit rate and the total receive rate are, $R_t = R_{12} + R_2 = 40.8$ kbps and $R_r = 2R_{12} + R_2 = 56.8$ kbps, for the conventional scalable coder. The proposed coder can achieve various combinations of R_t and R_r based on the trade-off parameter, α.

A scatter plot of total transmit rate versus total receive rate achieved by the various coders is shown in Fig. 3. The many scatter points of the proposed coder clearly demonstrates that it achieves various intermediate operating points based on the trade-off parameter, α. A good operating point would be that of having total transmit rate less than the conventional scalable coder and total receive rate close to that of the non-scalable coder. Our proposed framework allows achieving operating points close to this good operating point, and as part of our future work we plan to further close this gap with the good operating point.

The overall cost function of (1) achieved by the various coders is shown in Fig. 4. This result clearly demonstrates the proposed coder's ability to achieve an overall cost lower than that of both the competing coders for most of the trade-off points. Note that the proposed coder under performs the non-scalable coder at the highest alpha values due to the preliminary, proof-of-concept implementation's inefficiency in signaling zero-quantized bands at the common layer. We plan to mitigate this inefficiency in a fully optimized implementation.

5. CONCLUSIONS

Conventional scalable coding and independent coding are both unsatisfactory solutions for storing and transmitting content at various quality levels. Thus we propose a novel framework for layered coding of information at various quality levels, wherein there is flexibility to share only a subset of information in lower quality level with a higher quality level. This flexibility is critical for achieving intermediate operating points in the trade-off between total transmit rate and total receive rate, and also for optimally encoding common information between various quality levels separately. We employed the proposed framework within the MPEG scalable AAC and proposed an optimization technique for joint selection of parameters of all layers. Evaluation results demonstrate the utility of the flexibility provided by the proposed framework for storage and transmission system designers.

6. REFERENCES

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

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

[3] A. Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163–179, Mar. 1975.

[4] K. Viswanatha, E. Akyol, T. Nanjundaswamy, and K. Rose, "On common information and the encoding of sources that are not successively refinable," in *Proc. IEEE Information Theory Workshop (ITW)*, Sep. 2012.

[5] ISO/IEC JTC1/SC29 14496-3:2005, "Information technology - Coding of audio-visual objects - Part 3: Audio - Subpart 4: General audio coding (GA)," 2005.

[6] B. Grill, "A bit rate scalable perceptual coder for MPEG-4 audio," in *Proc. 103rd AES Conv.*, Sep. 1997, Preprint 4620.

[7] E. Ravelli, V. Melkote, T. Nanjundaswamy, and K. Rose, "Joint optimization of base and enhancement layers in scalable audio coding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 711–724, 2013.

[8] A.E. Gamal and T. Cover, "Achievable rates for multiple descriptions," *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 851–857, Nov. 1982.

[9] A. Aggarwal, S.L. Regunathan, and K. Rose, "A trellis-based optimal parameter value selection for audio coding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 623–633, 2006.

[10] V. Melkote and K. Rose, "Trellis-based approaches to rate-distortion optimized audio encoding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 330–341, 2010.

[11] K. Viswanatha, E. Akyol, and K. Rose, "A strictly improved achievable region for multiple descriptions using combinatorial message sharing," in *Proc. IEEE Information Theory Workshop (ITW)*, Oct. 2011.