SNR Scalability, Multiple Descriptions, and Perceptual Distortion Measures

Jerry D. Gibson
Department of Electrical & Computer Engineering
University of California, Santa Barbara
gibson@mat.ucsb.edu

***Abstract*—SNR Scalability and Multiple Descriptions (MD) coding are two important functionalities for speech coding applications. Previous analyses of these structures have not included frequency weighted error criteria. We present rate distortion theoretic results showing that the weighting functions in the core and enhancement layers for SNR scalable coding and in the side descriptions and the joint description for MD coding are necessarily different. We present simulation results for SNR scalable speech coding and MD speech coding to illustrate the theory.**

***Index Terms*—speech coding, SNR scalability, multiple descriptions coding, frequency weighted error criteria**

I. Introduction

Advances in speech and audio coding in the last decade have been driven by the incorporation of perceptual distortion measures into the source compression process. In particular, code-excited linear predictive (CELP) coding is the dominant coding paradigm in speech coding standards, and the use of a perceptually based distortion measure is key to its success [1]. As speech and audio coders are integrated into packet-switched network applications, new functionalities, such as SNR scalable coding and multiple descriptions coding, have taken on new importance. SNR scalable coding, also called bit rate scalable coding or layered coding, consists of sending a minimum rate bit (core) stream that produces acceptable coded speech quality, with the possibility of sending additional incremental rate enhancement (refinement) bit streams, which when combined with the core bit stream, yield successively improved output speech quality [2, 3]. Multiple descriptions (MD) coding at a given rate R bits/sec consists of providing two or more bit streams coded at fractional rates of the total bit rate, which if decoded individually, any one bit stream will provide acceptable performance, but if all bit streams are available and jointly decoded, much-improved performance is obtained [4]. SNR scalable coding and MD coding are contrasted by noting that the enhancement layers in SNR scalable coding cannot generate acceptable reconstructed output speech if decoded alone, while any one of the MD bit streams is designed to do so.

SNR scalability allows efficient network utilization for users with different bandwidth capabilities and performance requirements, while MD coding provides diversity transmission to compensate for possibly degraded network conditions. Numerous SNR scalable speech coders have been proposed and studied, with the most familiar being the MPEG-4 scalable coders described within the MPEG-4 Bit Rate Scalable toolbox [5]. Fewer MD speech coders have been developed, and no MD speech coder has yet appeared in a standard, but some recent efforts are promising [6].

The basic theory underlying SNR scalable coding and multiple descriptions coding has been developed primarily under the assumptions of memoryless sources and an unweighted mean squared error (MSE) fidelity criterion. Since the most important speech coders in use today rely heavily on a perceptually weighted distortion measure, it is of interest to investigate the interaction of perceptually based, frequency weighted squared error distortion measures with the desirable functionalities of SNR scalability and multiple descriptions coding. We present expressions for the rate distortion performance of weighted and unweighted squared error distortion measures for SNR scalable and multiple descriptions coders, and investigate particular applications involving code-excited speech coding. These results reveal that the different layers in SNR scalable coding and the different descriptions in multiple descriptions coding with perceptually weighted error criteria can have conflicting requirements on the distortion measures, and hence, that optimal performance may be compromised

In Section II, we outline the rate distortion theory essentials needed for the development, and in Secs. III and IV, we present some key (new) rate distortion theory results for SNR scalable codes and MD codes, respectively. Section V demonstrates the effects of using perceptual weighting in an SNR scalable coder, and Section VI provides similar results for multiple descriptions coding with a frequency-weighted squared error distortion measure. Section VII contains an analysis and conclusions.

II. Rate Distortion Theory Basics

The rate distortion function is the minimum rate required to send a source subject to a constraint on the average distortion. It is defined as [7]

$$R(D) = \min_{E[d(X,\hat{X})]\le D} I(X;\hat{X}) \qquad (1)$$

where $I(X;\hat{X})$ is the mutual information between the input source X and the reconstructed output $\hat{X}$, $d(X,\hat{X})$ is the distortion measure, and the average distortion constraint determines the set of admissible transition probabilities between the input and the reconstructed output. One of the most quoted results from rate distortion theory is the rate distortion function of a memoryless Gaussian source with arbitrary mean and variance σ_X^2 subject to the mean squared error (MSE) distortion measure, which is given by

This research was supported, in part, by the National Science Foundation under Grant Nos. CCF-0429884 and CNS-0435527, and by the University of California Micro Program, Dolby Laboratories, Inc., Lucent Technologies, Inc., Mircosoft Corp., and Qualcomm, Inc.

$$R_G(D) = \frac{1}{2}\log\frac{\sigma_X^2}{D} \text{ for } \sigma_X^2 > D \qquad (2)$$

and zero for $\sigma_X^2 \le D$. This result, in its distortion rate form, has served as the basis for optimal quantizer design and for bit allocation in transform coding of speech, audio, and still images [8].

Since it is difficult to find a closed form expression for *R(D)* in most cases, one often resorts to investigating bounds on the rate distortion function. The rate distortion function for a non-Gaussian, memoryless source with respect to the MSE distortion measure is upper bounded by $R_G(D)$ in Eq. (2) and lower bounded by

$$R_L(D) = \frac{1}{2}\log\frac{Q_1}{D} \qquad (3)$$

where Q_1 is the entropy power (or entropy rate power) of the source, given by

$$Q_1 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi}\log S(\omega)d\omega\right\}. \qquad (4)$$

An important observation for speech coding is that Q_1 is the one-step mean squared prediction error for Gaussian sequences, and can be calculated from the autocorrelation matrix of the source [7].

The parametric form of the rate distortion function for a time discrete Gaussian source with power spectral density $S(\omega)$ subject to the MSE fidelity criterion is given by

$$R(D_\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\max\left[0, \log\frac{S(\omega)}{\theta}\right]d\omega \qquad (5)$$

and

$$D_\theta = \int_{-\pi}^{\pi}\min\left[\theta, S(\omega)\right]d\omega \qquad (6)$$

where $R(D_\theta)$ traces out the rate distortion function as the parameter θ is varied.

For a frequency-weighted squared error fidelity criterion with weighting function $W(\omega)$, the rate distortion function is [9]

$$R(D) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\log\left[S(\omega)/\min\left\{S(\omega), \theta/W(\omega)\right\}\right]d\omega \qquad (7)$$

where

$$D = \frac{1}{2\pi}\int_{-\pi}^{\pi}W(\omega)\min\left[S(\omega), \theta/W(\omega)\right]d\omega \qquad (8)$$

For small distortions

$$R(D) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\log S(\omega)d\omega + \frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{W(\omega)}{D}d\omega \qquad (9)$$

This is the form that is useful to us in our investigations of frequency-weighted distortion measures for speech coding.

III. Successive Refinement of Information

SNR scalability has been investigated from the rate distortion theory viewpoint as successive refinement of information [2]. A sequence of random variables $X_1, \ldots, X_n$ is successively refined by a two-stage description that is rate distortion optimal at each stage. The X sequence is encoded as $\hat{X}$ at rate R_1 bits/symbol with average distortion D_1. Then, information is added to the first message at the rate $R_e = R_2 - R_1$ bits/symbol so that the resulting two-stage reconstruction $\hat{X}_r$ now has average distortion $D_2 \le D_1$ at rate $R_2 \ge R_1$.

Most rate distortion theory research for SNR scalability has been concerned with finding the conditions under which successive refinement is achievable. The successive refinement problem was first introduced by Koshelev as the problem of divisibility, and he proved the sufficiency of a Markov chain relationship between the source and the refined reconstructions [10]. Equitz and Cover proved necessity and showed, using the Shannon backward channel formulation, that the Markov chain condition holds for Gaussian sources and squared error distortion, Laplacian sources and the absolute error criterion, and all discrete sources and Hamming distortion measures [2]. The Markov chain condition to be satisfied for successive refinement of X is that $X \to \hat{X}_r \to \hat{X}$, or equivalently, $\hat{X} \to \hat{X}_r \to X$. This condition was extended by Rimoldi to the case where a different distortion measure is used at each layer [12].

Recently, the nomenclature, *successive refinement with no excess rate* has been coined to allow a distinction between rate distortion optimal successive refinement and SNR scalable coding in general that may not be rate distortion optimal.

IV. Multiple Descriptions

The simplest form of the Multiple Descriptions (MD) problem is shown in Fig. 1, and consists of representing a source with two descriptions at rates R_1 and R_2 such that if both descriptions are received, a central decoder achieves average distortion D_0, while if either description is lost, the side decoder can achieve average distortion D_1 or D_2 for rates R_1 or R_2, respectively. Since the rate of the central decoder is $R = R_1 + R_2$, then clearly, $D_0 \le D_1$ and $D_0 \le D_2$. On the theoretical side, much of the interest in the MD problem has been on characterizing the achievable rate distortion region. Trivially, one can write the achievable region as [4]

$$\begin{aligned} R_1 &\ge R(D_1) \\ R_2 &\ge R(D_2) \\ R_1 + R_2 &\ge R(D_0) \end{aligned} \qquad (10)$$

where the $R(D_i), i = 0, 1, 2,$ represent values of the rate distortion function at those distortions. Much of the challenge of the MD problem is captured in these simple expressions, and there are two primary cases of interest. In one case, denoted as the *no excess marginal rate* case, the individual descriptions are rate distortion optimal, and so the joint reconstruction that is decoded when both descriptions are

received is necessarily suboptimal, since the two individual descriptions must be very similar; hence, the average distortion D_0 that is obtained is larger than would be obtained by a single rate distortion optimal description at rate $R = R_1 + R_2$. The second case, called the *no excess joint rate* case, is when the joint description is rate distortion optimal, and hence the individual descriptions are independent and therefore individually far away from the rate distortion bound.

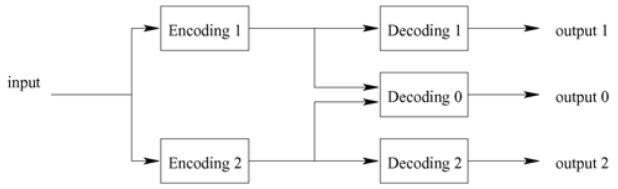

Figure 1. The Multiple Descriptions Problem

For a memoryless Gaussian source with variance σ_X^2, Ozarow [11] characterized the MD distortion rate region, which can be rewritten in terms of rate distortion functions as

$$R_1 \geq \frac{1}{2}\log\frac{\sigma_X^2}{D_1}$$
$$R_2 \geq \frac{1}{2}\log\frac{\sigma_X^2}{D_2} \qquad (11)$$
$$R_1 + R_2 \geq \frac{1}{2}\log\frac{\sigma_X^2}{D_1} + \frac{1}{2}\log\frac{\sigma_X^2}{D_2} + \Delta R$$

where ΔR can be interpreted as the rate used to minimize the distortion when both descriptions are received. If $\Delta R = 0$, the individual descriptions are rate distortion optimal.

V. Perceptual Distortion Measures and SNR Scalability

We consider two stage SNR scalable coding wherein a frequency weighted squared error distortion measure is used at each stage. From Eq. (9), we can write the rate distortion function for the core layer as

$$R_1(D_1) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\log S(\omega)d\omega + \frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{W_1(\omega)}{D_1}d\omega$$

or using Eq. (4) as

$$R_1(D_1) = \frac{1}{2}\log\frac{Q_x}{D_1} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log W_1(\omega)d\omega \qquad (12)$$

for $0 \leq D_1 \leq \delta_1$, where Q_x is the entropy power of the source, D_1 is the average distortion in the core layer, $W_1(\omega)$ is the weighting factor for the core layer, and δ_1 is the minimum of the frequency weighted source spectrum.

The rate distortion function for the enhancement layer can be written as

$$R_e(D_e) = \frac{1}{2}\log\frac{Q_e}{D_e} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log W_e(\omega)d\omega \qquad (13)$$

for $0 \leq D_e \leq \delta_e$, where Q_e is the entropy power of the enhancement layer coding error, D_e is the average distortion in the enhancement layer, $W_e(\omega)$ is the weighting factor for the enhancement layer, and δ_e is the minimum of the frequency weighted core layer error spectrum. The total rate for the core and enhancement layers is thus

$$R(D) = R_1(D_1) + R_e(D_e)$$
$$= \frac{1}{2}\log\frac{Q_x}{D_1} + \frac{1}{2}\log\frac{Q_e}{D_e} \qquad (14)$$
$$+ \frac{1}{4\pi}\int_{-\pi}^{\pi}\log W_1(\omega)W_e(\omega)d\omega$$

We can check this result against the memoryless source, unweighted MSE case by letting $W_1(\omega) = W_e(\omega) = 1$, so with $Q_e = D_1$ and $D_e = D_2$, the result agrees with Equitz and Cover.

If we contrast the two stage successive refinement result in Eq. (14) when $Q_e = D_1$ and $D_e = D$, with a one stage rate distortion optimal rate distortion function as in Eq. (9), we see that for the two stage refinable result to equal the one stage rate distortion optimal encoding, we need

$$W_1(\omega)W_e(\omega) = W(\omega) \qquad (15)$$

This result implies that if $W(\omega)$ is optimal for single stage encoding, then the core layer and enhancement layer frequency weighting should not be the same $W(\omega)$! In the following example, we investigate the SNR scalable coders standardized as part of the MPEG-4 Natural Audio Coding Suite with respect to this result.

Example: MPEG-4 Bit Rate Scalable Tool

A CELP SNR scalable coder was standardized as a part of the MPEG-4 natural audio coding toolbox in 1998 [5]. The MPEG-4 CELP operates at more than fifty bit rates by changing its frame size and coding parameters for both wideband and narrowband speech. SNR scalability in the MPEG-4 CELP coder is achieved by encoding the speech signal using a combination of the core coder and the bit rate scalable tool. The core coder is based on a CELP algorithm, and for wideband speech, encodes the input speech signal at predetermined bit rates between 10.9 and 23.8 kbps. In the bit rate scalable tool, a residual signal that is produced at the core coder is encoded utilizing multi-pulse vector quantization to enhance the coding quality by an analysis-by-synthesis structure. The bit rate of each enhancement layer is 4 kbps for wideband speech, and up to 3 enhancement layers may be combined for better quality.

In each enhancement layer, the linear prediction filter and the perceptual weighting filter are the same as those in the core layer. The algebraic-structure codebook at the enhancement layer is obtained by minimizing the perceptually weighted distortion between the reconstruction error signal from the core and the output signal from the enhancement layer.

As a result, the weighting function for a single layer coder and for each enhancement layer is the same, $W(\omega)$. In Fig. 2, we show the input speech and error spectra for a speech frame encoded at the core rate of 10.9 kbps by the MPEG-4 coder, along with one 4 kbps enhancement layer. We also show the input speech and error spectra for a single layer coder at 14.8 kbps for comparison to the SNR scalable coder.

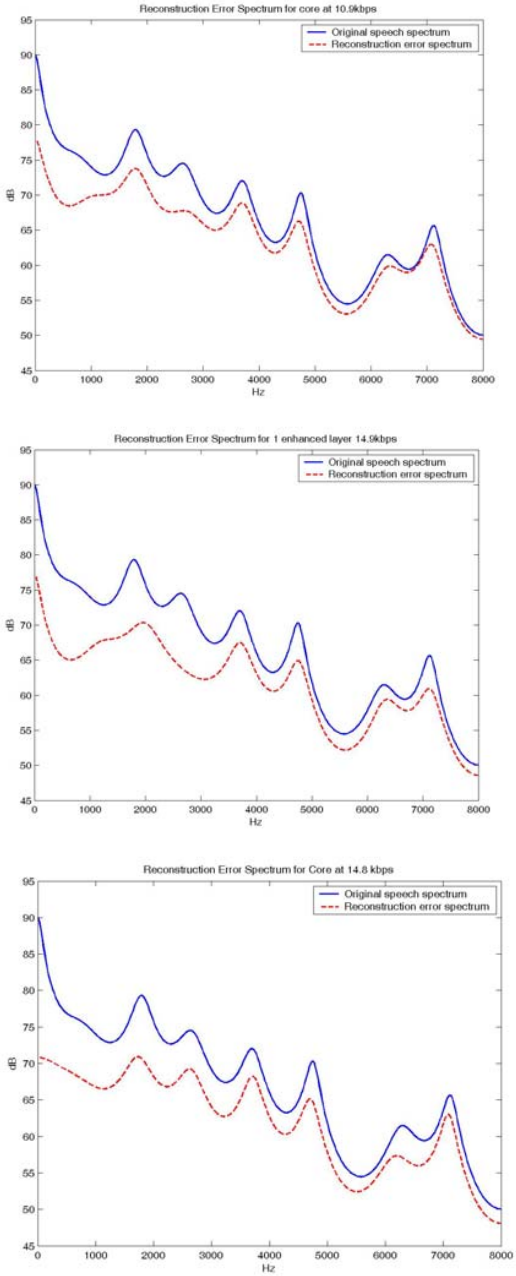

Figure 2. Reconstruction Error Spectra for SNR Scalable Coding. (a) Core at 10.9 kbps, (b) Two Layer Reconstruction at 14.9 kbps, (c) Single Layer at 14.8 kbps

The different shaping of the error spectrum between single stage encoding and scalable encoding at the same rate is evident (no postfiltering is being used). The error spectra for the single stage encoding seems to follow the input spectrum better, while the enhancement layer encoding has an error spectrum that appears less related to the input spectrum. Results not shown indicates this continues with each subsequent layer. These results support the concept that SNR scalable coding using the same perceptual weighting filter at each layer does not provide the same shaping as single stage encoding at the same rate with the same weighting filter. However, it is important to note that the theoretical rate distortion results are for optimal encoding for small distortions, and since these qualities cannot be verified at these rates for the MPEG-4 coder, the specific quantitative relationships may not hold.

VI. Perceptual Distortion Measures and Multiple Descriptions Coding

To investigate frequency weighted distortion measures for multiple descriptions coding, we must consider the separate cases of no excess marginal rate and no excess joint rate. For the no excess marginal rate case and two side channels, we have that the rate distortion functions for the side channels are

$$R_1(D_1) = \frac{1}{2}\log\frac{Q_x}{D_1} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log W_1(\omega)d\omega$$
$$R_2(D_2) = \frac{1}{2}\log\frac{Q_x}{D_2} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log W_2(\omega)d\omega \quad (16)$$

so that the expression for the joint description becomes

$$R_1(D_1) + R_2(D_2) = \frac{1}{2}\log\frac{Q_x^2}{D_1 D_2} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log\left[W_1(\omega)W_2(\omega)\right]d\omega \quad (17)$$

Single stage optimal encoding at the total rate of $R = R_1 + R_2$ has the rate distortion function

$$R(D) = \frac{1}{2}\log\frac{Q_x}{D_0} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log W_0(\omega)d\omega \quad (18)$$

so for equality between Eqs. (17) and (18) we need $\frac{Q_x D_0}{D_1 D_2} = 1$ and $W_0(\omega) = W_1(\omega)W_2(\omega)$.

For the no excess joint rate case, Eq. (18) represents the joint description rate distortion function, and if we factor this into equal rate side channels, we have

$$R_1(D_1) = \frac{1}{2}\log\frac{\sqrt{Q_x}}{\sqrt{D_0}} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log\sqrt{W_0(\omega)}d\omega$$
$$R_2(D_2) = \frac{1}{2}\log\frac{\sqrt{Q_x}}{\sqrt{D_0}} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log\sqrt{W_0(\omega)}d\omega \quad (19)$$

which clearly implies that $W_1(\omega) = W_2(\omega) = \sqrt{W_0(\omega)}$ and $D_1 = D_2 = \sqrt{D_0}$.

The results in this section suggest that it may be difficult to optimize MD coders that employ frequency weighted distortion measures.

Example: AMR-WB Multiple Descriptions Coder

We consider an MD coder based upon the AMR-WB codec [13]. We obtain an MD coder by finding the best joint description (equivalent to the no excess joint rate case) and then splitting the bits into two bit streams, with sufficient redundancy between the two streams to get good performance should only one side channel be received. The resulting bit allocations are shown in Table I, and differ from the MD coder bit allocations in [6] by the inclusion of the first 6 bits of the second stage vector quantizer for the immittance spectrum pairs (ISPs) in both descriptions here.

Table I. Splitting of 12.65 kbps Joint Description into Two Side Descriptions. Boldface in Both Descriptions, ( ) denotes Description 1 and { } denotes Description 2

ISF	Stage 1	**8**		**8**	(34), {34}
	Stage 2	**6**{7}	(75)	{5}	
	1st Subfame	2nd Subframe	3rd Subframe	4th Subframe	
P-Delay	(9)	(6)	{9}	{6}	(15), {15}
A-Code	(36)	{36}	(36)	{36}	(72), {72}
Gains	(7)	{7}	(7)	{7}	(14), {14}

The bit rates for each single description is 6.9 kbps and for the joint description is 13.8 kbps. The quality of the joint description at 13.8 kbps is equivalent to the single stage coder at 12.65 kbps in the WB-AMR coder. The side descriptions achieve different quality output speech and both can be compared to the WB-AMR codec performance at 6.6 kbps.

Although space precludes including the plots here, a comparison of the input speech and reconstruction error spectra clearly show that the error spectra in the side descriptions differ from the joint rate description as well as from each other. Since the MD coder here was designed to have good joint rate performance, the joint description shaping is good but the shaping in the side descriptions is inadequate.

VII. Analysis and Conclusions

The results in Fig. 2 for SNR scalability and in Sec. VI for the MD coder can be interpreted by considering the shaping provided by perceptual weighting with $W(\omega), W^2(\omega)$, and $\sqrt{W(\omega)}$ shown in Fig. 3 for a specific speech frame. So, if in MD coding, the joint rate weighting is $W(\omega)$, the side channels are weighted much differently and much less. Thus, frequency weighted error criteria add another constraint to the design of SNR scalable and multiple descriptions speech coders.

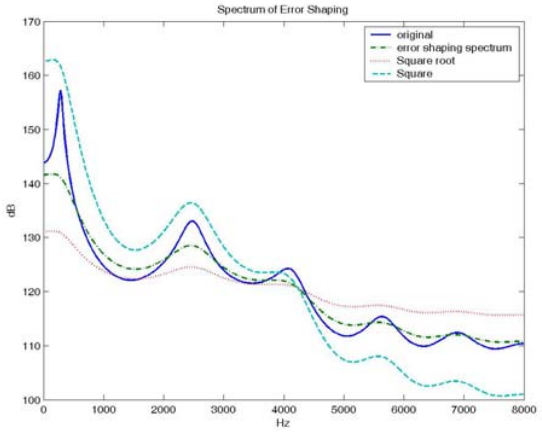

Figure 3. Perceptual Weighting Spectra for SNR Scalable and MD Coding

References

[1] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*, Amsterdam, Elsevier, 1995.

[2] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. on Information Theory*, vol. 37, pp. 269-274, March 1991.

[3] H. Dong and J. D. Gibson, "Structures for SNR scalable speech coding," *IEEE Trans. on Speech and Audio Processing*, accepted for publication, 2004.

[4] A. A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. on Information Theory*, vol. IT-28, pp. 851-857, Nov. 1982.

[5] K. Brandenburg, O. Kunz, and A. Sugiyama, "MPEG-4 natural audio coding," *Signal Processing: Image Communication*, vol. 15, pp. 423-444, 2000.

[6] H. Dong, A. Gersho, J. D. Gibson, and V. Cuperman, "A multiple description speech coder based on the AMR-WB for mobile ad hoc networks," *2004 IEEE ICASSP,* May 17-24, Montreal, Canada.

[7] T. Berger, *Rate Distortion Theory*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[8] J. D. Gibson, T. Berger, T. Lookabaugh, D. Lindbergh, and R. L. Baker, *Digital Compression for Multimedia: Principles & Standards,* Morgan-Kaufmann, 1998.

[9] L. D. Davisson, "Rate-distortion theory and application," *Proc. IEEE*, vol. 60, pp. 800-808, July 1972.

[10] V. Koshelev, "Hierarchical coding of discrete sources," *Probl. Pered. Inform*., vol. 16, pp. 31-49, 1980.

[11] L. Ozarow, "On a source-coding problem with two channels and three receivers," *BSTJ*, Vol. 59-10, pp. 1909-1921, 1980.

[12] B. Rimoldi, "Successive refinement of information: Characterization of achievable rates," *IEEE Trans. on Information Theory*, vol. 40, pp. 253-259, Jan. 1994.

[13] B. Bessette, et al, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 620-636, Nov. 2002.