Rate Distortion Bounds for Voice and Video

Jerry D. Gibson
University of California, Santa Barbara
gibson@ece.ucsb.edu

Jing Hu
University of California, Santa Barbara
jinghu@ece.ucsb.edu

Full text available at: http://dx.doi.org/10.1561/0100000061
Editorial Scope

Topics

Foundations and Trends® in Communications and Information Theory publishes survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2013, Volume 10, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328. Also available as a combined paper and online subscription.
Rate Distortion Bounds for Voice and Video

Jerry D. Gibson
University of California, Santa Barbara
gibson@ece.ucsb.edu

Jing Hu
University of California, Santa Barbara
jinghu@ece.ucsb.edu
Contents

1 Introduction 2
   1.1 Rate Distortion Functions for Speech Sources ............... 5
   1.2 Rate Distortion Functions for Video Sources ............... 6
   1.3 Conclusion ............................................. 7

2 Overview of Voice and Video Coding Techniques and Standards 9
   2.1 Voice Codecs .............................................. 9
       2.1.1 Characteristics of Voice Signals ...................... 11
       2.1.2 Performance Measures ................................. 12
       2.1.3 Speech Coding Methods ................................ 13
       2.1.4 Current and Developing Standards .................... 20
   2.2 Video Codecs .............................................. 24
       2.2.1 Characteristics of Video Signals ...................... 24
       2.2.2 Performance Measures ................................. 26
       2.2.3 Motion-Compensated Transform Coding ................. 29
       2.2.4 Current and Developing Standards .................... 31

3 The Rate Distortion Problem 37
   3.1 Rate Distortion Theory Basics ............................... 37
3.2 Rate Distortion Results for Gaussian Sources and Squared Error Distortion ................................................................. 40
  3.2.1 Scalar Gaussian Source with Mean Squared Error .......... 40
  3.2.2 Reverse Water-filling ............................................. 40
  3.2.3 Stationary Gaussian Sources with Memory ................. 41
  3.2.4 Rate Distortion Function for a Gaussian Autoregressive Source ................................................................. 43

3.3 Composite Source Models ............................................... 45

3.4 Conditional Rate Distortion Functions ............................ 46

3.5 Estimating Composite Source Model Parameters ............... 48

4 Rate Distortion Bounds for Voice ........................................ 50
  4.1 Related Prior Work .................................................. 51
  4.2 Composite Source Models for Speech ................................ 53
  4.3 Marginal and Conditional Rate Distortion Bounds based on MSE Distortion Measure .................................................. 59
  4.4 Mapping MSE to PESQ-MOS/WPESQ .............................. 62
    4.4.1 PESQ-MOS/WPESQ .............................................. 64
    4.4.2 ADPCM Speech Coders ......................................... 65
    4.4.3 Mapping Function ................................................. 67
  4.5 New Theoretical Rate Distortion Bounds for Speech ................. 69
    4.5.1 Rate Distortion Bounds and Operational Rate Distortion Performance for Narrowband Speech .................. 73
    4.5.2 Rate Distortion Bounds and Operational Rate Distortion Performance for Wideband Speech .................. 79
    4.5.3 Modifications to the MSE Mapping Function and Other Distortion Measures ........................................... 83
  4.6 Conclusions .......................................................... 85

5 Rate Distortion Bounds for Video ........................................ 86
  5.1 Related Prior Work .................................................. 86
    5.1.1 Statistical Models of Images and Videos .................. 86
    5.1.2 Statistical Models of Practical Video Compression Systems ................................................................. 88
  5.2 A New Block-Based Conditional Correlation Model for Video .................................................................................. 91
5.2.1 The Conditional Correlation Model in the Spatial Domain 91
5.2.2 Correlation Among Pixels Located in Nearby Frames 101

5.3 New Theoretical Rate Distortion Bounds of Natural Videos 105
5.3.1 Formulation of Rate Distortion Bound without Local Texture as Side Information 107
5.3.2 Formulation of Rate Distortion Bound with Local Texture as Side Information 108
5.3.3 Rate Distortion Bounds for One Video Frame 109
5.3.4 Rate Distortion Bounds for a Sequence of Video Frames 113

5.4 Constrained Rate Distortion Bounds for Blocking and Intra-frame Prediction 119
5.4.1 Constrained Rate Distortion Bound for Blocking 121
5.4.2 Constrained Rate Distortion Bound for Blocking and Optimal Intra-frame Prediction 121

5.5 Conclusion 129

References 130
Abstract

Numerous voice, still image, audio, and video compression standards have been developed over the last 25 years, and significant advances in the state of the art have been achieved. However, in the more than 50 years since Shannon’s seminal 1959 paper, no rate distortion bounds for voice and video have been forthcoming. In this volume, we present the first rate distortion bounds for voice and video that actually lower bound the operational rate distortion performance of the best-performing voice and video codecs. The bounds indicate that improvements in rate distortion performance of approximately 50% over the best-performing voice and video codecs are possible. Research directions to improve the new bounds are discussed.


DOI: 10.1561/0100000061.
Numerous voice, still image, audio, and video compression standards have been developed over the last 25 years, and significant advances in the state of the art have been achieved. There are several reasons for researchers and standards bodies to consider developing new voice or video codecs. One motivation might be a new application that has different constraints than those imposed on prior codecs. For example, a new application might require better quality, lower complexity, a different transmitted bit rate, or improved robustness to channel impairments. A second motivation might be that the input source changes, namely a different resolution for video, a requirement for 3D video, or a different bandwidth and sampling rate for audio. A third motivation might be that a particular codec is relatively old and that there is the possibility of improving performance, perhaps by increasing complexity because of advances due to Moore’s Law.

In each of these cases, it would seem natural to ask what is the best possible performance theoretically achievable by a new codec? Or, alternatively, given the operational rate distortion performance of a particular codec, how close is the operational rate distortion performance to the optimal performance theoretically achievable?
To answer this question, one natural place to look in order to characterize the best possible performance of any lossy source codec would appear to be rate distortion theory. In particular, it would be of great utility if the host of existing rate distortion theory results could be applied to bounding the performance of practical codecs or if new rate distortion bounds for such practical sources and their attendant perceptual distortion measures could be obtained. However, no such applications of existing rate distortion theory results, nor any appropriate new results, have been forthcoming. While there are many reasons for this lack of progress, one main reason is that such an effort is not easy—in fact, it is particularly difficult.

The particular challenges involved were anticipated by experts in Information Theory very early. Specifically, Robert Gallager, in his classic text on Information Theory [18], summarizes the challenges at the end of his rate distortion theory chapter where he notes that information theory has been more useful for channel coding than for source coding and that the reason, “...appears to lie in the difficulty of obtaining reasonable probabilistic models and meaningful distortion measures for sources of practical interest.” He goes on to say, “…it is not clear at all whether the theoretical approach here will ever be a useful tool in problems such as speech digitization...” [18].

Finding suitable statistical models for video has been considered a very difficult topic as well. In 1998, almost 40 years after Shannon’s landmark paper developing rate distortion theory [76], Ortega and Ramchandran wrote, “Unfortunately, to derive bounds one needs to first characterize the sources and this can be problematic for complex sources such as video. Indeed, bounds are likely to be found only for the simpler statistical models” [67].

Thus, like all rate distortion problems, the two primary challenges are (1) finding good source models for speech and video, and (2) identifying a distortion measure that is perceptually meaningful, yet computationally tractable. There have been only a few prior research efforts in the last 25 years that have attempted to address various aspects of this problem for either speech or video, and broad-based bounds of significance have not been obtained. It is clear, however, that the utility
of such bounds would be substantial.

In this volume, we present our recent results on obtaining rate distortion functions for both voice and video sources. For both sources, we overcome past limitations on source modeling by employing composite source models to achieve more accurate modeling of the different voice and video source modes. Although we use composite source models for both voice and video, the treatments of the distortion measure for the two sources are distinctly different. For speech, we devise a mapping technique to extend existing MSE $R(D)$ results to the perceptually meaningful PESQ-MOS distortion measure. For video, no such mappings are developed and the MSE distortion measure, or equivalently peak SNR (PSNR), is used directly to develop our video $R(D)$ bounds. This is because although MSE and PSNR are widely criticized as not having a direct interpretation in terms of reconstructed video quality, PSNR is known to order the performance of codecs in the same class correctly. In fact, since optimizing MSE/PSNR often produces competitive performance in terms of perceptual measures, and its limitations are well known, it is still a dominant performance measure in video codec standardization efforts.

For future progress, as well as for the development of future practical rate distortion results, it is critical to note from the above outline of the approaches used here that there are two key elements in play in order to obtain the rate distortion bounds presented in this volume. These are (1) a grasp and fundamental understanding of key rate distortion theory results, and (2) a deep understanding of the real-world sources and their codec performance evaluation methods. Either one alone is not sufficient. Indeed, the first author has emphasized to his students repeatedly over the past 30 years that in order to utilize significant theoretical results for practical problems, one must also have an understanding of the physical problem being addressed. This combination is not often present, perhaps because, as noted by Berger and Gibson [7], rate distortion theorists and voice and video codec designers are mostly non-intersecting sets of researchers.

We summarize the contents of this volume for each source in the following subsections.
1.1 Rate Distortion Functions for Speech Sources

We develop new rate distortion bounds for narrowband and wideband speech coding based on composite source models for speech and perceptual PESQ-MOS/WPESQ distortion measures. It is shown that these new rate distortion bounds do in fact lower bound the performance of important standardized speech codecs, including, G.726, G.727, AMR-NB, G.729, G.718, G.722, G.722.1, and AMR-WB.

Our approach is to calculate rate distortion bounds for mean squared error (MSE) distortion measures using the classic eigenvalue decomposition and reverse water-filling method for each of the subsource modes of the composite source model, and then use conditional rate distortion theory to calculate the overall rate distortion function for the composite source. While composite source models for speech have been considered previously for obtaining $R(D)$ functions for speech, our method of choosing the subsources based on a knowledge of speech signals and on successful multi-mode voice codecs, as well as the inclusion of diverse subsources in the composite source models, are new.

In order to develop $R(D)$ bounds for speech in terms of a meaningful distortion measure that still allows a tractable mathematical calculation of the bounds required a new innovation as well. Mapping functions are developed to map rate distortion curves based on MSE to rate distortion curves subject to the perceptually meaningful distortion measures PESQ-MOS and WPESQ. These final rate distortion curves are then compared to the performance of the best known standardized speech codecs based on the code-excited linear prediction paradigm.

In addition to the striking result that these new bounds do in fact lower bound the best known narrowband and wideband standardized speech codecs, the bounds are revealing in that performance comparisons show that current linear predictive codecs do a relatively good job of coding voiced speech, but are much less effective for other subsources, such as unvoiced speech, Onset, and Hangover modes. Equally important is that the procedure used in developing our bounds can easily be reproduced by other researchers, and thus other, perhaps more refined, rate distortion curves can be generated. For example, one could
utilize a different composite source model with the known MSE rate distortion theory results outlined here, and then employ our mapping functions to determine new bounds for the utterances considered in this paper.

1.2 Rate Distortion Functions for Video Sources

For the video source we address the difficult task of modeling the correlation in pixel values by first proposing a new spatial correlation model for two close pixels in one frame of digitized natural video sequences that is conditional on the local texture. This new spatial correlation model is dependent upon five parameters whose optimal values are calculated for a specific image or specific video frames. The new spatial correlation model is simple, but it performs very well, as strong agreement is discovered between the approximate correlation coefficients and the correlation coefficients calculated by the new correlation model, with a mean absolute error (MAE) usually smaller than 5%.

Further, we extend the correlation coefficient modeling from pixels within one video frame to pixels that are located in nearby video frames. We show that for two pixels located in nearby video frames, their spatial correlation and their temporal correlation are approximately independent. Therefore the correlation coefficient of two pixels in two nearby video frames, denoted by $\rho$, can be modeled as the product of $\rho_s$, the texture dependent spatial correlation coefficient of these two pixels, as if they were in the same frame, and $\rho_t$, a variable to quantify the temporal correlation between these two video frames. $\rho_t$ does not depend on the textures of the blocks the two pixels are located in and is a function of the indices of the two frames.

With the new block-based local-texture-dependent correlation model, we first study the marginal rate distortion functions of the different local textures. These marginal rate distortion functions are shown to be quite distinct from each other. Classical results in information theory are utilized to derive the conditional rate distortion function when the universal side information of local textures is available at both the encoder and the decoder. We demonstrate that by involving
1.3. Conclusion

this side information, the lowest rate that is theoretically achievable in *intra-frame* video compression can be as much as 1 bit per pixel lower than that without the side information; and the lowest rate that is theoretically achievable in *inter-frame* video compression can be as much as 0.7 bit per pixel lower than that without the side information. The rate distortion bounds with local texture information taken into account while making no assumptions on coding, are shown indeed to be valid lower bounds with respect to the operational rate distortion curves of both *intra-frame* and *inter-frame* coding in Advanced Video Coding (AVC/H.264) and in the newly standardized High Efficiency Video Coding (HEVC/H.265).

The incorporation of the new correlation model into existing operational models of practical image and video compression systems is also promising. We demonstrate this by studying the common “blocking” scheme used in most video compression standards [32, 33, 34, 35], which divides a video frame into $16 \times 16$ macroblocks (MB) or smaller blocks before processing. With the block based nature of the new correlation model, we study the penalty paid in average rate when the correlation among the neighboring MBs or blocks is disregarded completely or is incorporated partially through predictive coding. A constrained rate distortion bound is calculated for the scenario when the texture information is coded losslessly and optimal predictive coding is employed. This lower bound is shown to be reasonably tight with respect to the operational rate distortion curves of intra-frame coding in AVC/H.264. Furthermore, it is near linear in terms of average bit rate per pixel versus PSNR of a video frame and can easily be utilized in future video codec designs.

1.3 Conclusion

In this volume, we present the first rate distortion bounds for voice and video that actually lower bound the operational rate distortion performance of the best-performing voice and video codecs. Members of the Panel on “New Perspectives on Information Theory” held at the IEEE Information Theory Workshop at Paraty, Brazil, on October
20, 2011, repeatedly expressed their concern about the gap between lossy compression theory and practice [52]. The new rate distortion bounds presented here, for the first time, make the gap specific for voice and video, and as discussed later, aid in pointing the way forward to improving the performance of practical voice and video codecs.
References


References


References


References


[83] Koen Vos, Karsten Vandborg Sørensen, Søren Skak Jensen, and Jean-Marc Valin. The opus codec.


