Foundations and Trends® in Communications and Information Theory
Vol. 10, No. 4 (2013) 379–514

DOI: 10.1561/0100000061

now
the essence of knowledge

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

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.

J. D. Gibson and J. Hu. *Rate Distortion Bounds for Voice and Video*. Foundations and Trends® in Communications and Information Theory, vol. 10, no. 4, pp. 379–514, 2013.

DOI: 10.1561/0100000061.

1

Introduction

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 ρ, can be modeled as the product of ρ_s, the texture dependent spatial correlation coefficient of these two pixels, as if they were in the same frame, and ρ_t, a variable to quantify the temporal correlation between these two video frames. ρ_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

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 16×16 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 [82]. 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.

2

Overview of Voice and Video Coding Techniques and Standards

The purpose of this chapter is to provide an overview of voice and video coding techniques, especially those techniques that lay the foundation for voice and video coding standards. The chapter also presents a summary of relevant voice and video coding standards, primarily emphasizing the current highest performing codecs and the codecs that motivated the model building and the selection of the fidelity criteria for the rate distortion bounds presented in later chapters. There is no intention of providing the history of voice and video coding, nor a complete discussion of voice and video coding methods and standards.

2.1 Voice Codecs

The goal of speech coding is to represent speech in digital form with as few bits as possible while maintaining the intelligibility and quality required for the particular application. Speech coding is fundamental to the operation of the public switched telephone network (PSTN), videoconferencing systems, digital cellular communications, and voice over Internet protocol (VoIP) applications. There was almost an exponential growth of speech coding standards in the 1990's for a wide range of net-

works and applications, including the wired Public Switched Telephone Network (PSTN), digital cellular systems, and multimedia streaming over the Internet [20]. Standards activities are still being pushed forward aggressively today, with efforts on combined voice/audio coding, called fullband coding as later elaborated, recently being established or currently being pursued. Included in these standards are the USAC codec [71], the Opus codec [83], and the Enhanced Voice Standard (EVS) codec for Long Term Evolution (LTE) digital cellular [54].

Interestingly, new standards for speech coding have never been driven by rate distortion bounds that show the existence of better rate/distortion operating points. Instead, the standards activities have been motivated by new applications with new requirements and expected demand for a service. Engineers often have a good idea that a new, better performing voice codec can be developed because of work performed in pursuing recently established standards, but no basic rate distortion bounds have been obtained or relied upon that actually indicate that there is a rate distortion performance gap, or how large that gap might be.

Like many fields, improvements in codec performance have been greatly facilitated by Moore's Law; in fact, it is abundantly clear that improvements in the operational rate distortion performance of voice codecs standardized in the last decade or more have been acquired with ever-increasing codec complexity.

With respect to Shannon theory, this increase in complexity is only of slight concern, since the primary goal of generating useful bounds on the rate distortion performance of voice codecs is not tied to complexity at all. The question posed by Shannon is simply, what is the best rate/distortion performance obtainable by any codec of any complexity? This separation from penalties due to increasing complexity should be valuable in searching for rate distortion performance bounds, since the bounds are obtained off line and can use any complex model of the speech generation process that makes sense physically and that can be handled analytically. These new degrees of freedom have not been fully exploited, even in the work presented herein.

We discuss voice codecs in this chapter only to the extent that

we use the codecs to develop meaningful distortion measures or if the codec performance is compared to the rate distortion bounds developed in this book. We compare the rate distortion performance of the best known voice codecs to our rate distortion bounds in Chapter 4.

As is typical in the literature, we use the terms speech coding and voice coding interchangeably in this book. Generally, it is desired to reproduce the voice signal, since we are interested in not only knowing what was said, but also in having a voice quality reproduction sufficient to identify the speaker. This goal is reflected in the choice of the fidelity criterion or distortion measure.

2.1.1 Characteristics of Voice Signals

Speech and audio coding can be classified according to the bandwidth occupied by the input and the reproduced source [12]. Indeed, the particular bandwidth occupied by the original speech source is closely tied to the particular applications of interest and can also sufficiently narrow the design requirements enough to admit simpler or higher performing codecs than if all bandwidth are of interest. Narrowband or telephone bandwidth speech occupies the band from 200 to 3400 Hz, or sometimes 100 to 3700 Hz, or slight variations, refers to the band associated with classical wired telephone quality speech as well the basic digital cellular and Voice over Internet Protocol (VoIP) services. The sampling rate for narrowband speech is 8,000 samples/sec.

In the mid to late 1980's, a new bandwidth of 50 Hz to 7 kHz, called wideband speech, with a sampling rate of 16,000 samples per second became of interest for videoconferencing applications. The importance of this band deriving from the fact that videoconferences or audio conference calls of a half hour to an hour require a long attention span by the participants and it was discovered that listening fatigue was greatly reduced with the wider band in comparison to narrowband speech. This band is still of great importance today in videoconferencing, digital cellular, and VoIP applications.

High quality audio is generally taken to cover the range of 20 Hz to 20 kHz, and this bandwidth is designated today as fullband. This band is associated with high quality audio and the codecs designed

for high quality music coding and playback. In recent years, quite a few other bandwidths have attracted attention, primarily for audio over the Internet applications, and the bandwidth of 50 Hz to 14 kHz, designated as superwideband, has gotten considerable recent attention in standardization activities.

The discussions in this book emphasize narrowband and wideband speech, which encompass the vast majority of voice codecs in use today. Rate distortion bounds for superwideband and fullband voice/audio require extensions beyond the approaches developed in this book, although now there is a roadmap in place. These directions for future work are covered in later chapters.

Beyond occupied bandwidth, voice signals can be further classified into the types of sounds produced or into one of several speech modes, such as voiced, unvoiced, onset, silence, and so on [87, 86]. Such classifications can be useful for developing voice codecs and also for producing accurate models of the voice source. We defer discussions of these classifications methods to the later speech source modeling discussions, but they will prove crucial in what follows.

2.1.2 Performance Measures

Given a particular source, the classic tradeoff in lossy source compression is rate versus distortion–the higher the rate, the smaller the average distortion in the reproduced signal. Of course, since a higher bit rate implies a greater channel or network bandwidth requirement, the goal is always to minimize the rate required to satisfy the distortion constraint, or alternatively, to minimize the distortion for the specified rate constraint. For speech coding, we are interested in achieving a quality as close to the original speech as possible within the rate, complexity, latency, and any other constraints that might be imposed by the application of interest. Encompassed in the term quality are intelligibility, speaker identity, and naturalness.

There has been considerable research into objective and subjective methods to evaluate the perceived quality of the speech produced by voice codecs. Absolute category rating (ACR) tests are subjective tests of speech quality and involve listeners assigning a category and rat-

ing for each speech utterance according to the classifications, such as, Excellent (5), Good (4), Fair (3), Poor (2), and Bad (1). The average for each utterance over all listeners is the Mean Opinion Score (MOS) [84]. ACR tests have been the dominant approach to evaluating codec reconstructed voice quality for narrowband speech for half a century.

Because of the widespread use of subjective ACR tests that produce MOS values between 1 and 5, the perceptual evaluation of speech quality (PESQ) method was developed as an objective method to provide an assessment of speech codec performance, particularly in conversational voice communications. Objective methods remove the need for human listeners and costly and time-consuming testing with human subjects. The PESQ has been standardized by the ITU-T as P.862 and can be used to generate MOS values for both narrowband and wideband speech [48]. It is a full reference method in that the original and the coded voice signals are used as inputs, and the output is a PESQ-MOS value on the same scale as MOS. PESQ-MOS has been accepted as a reasonable substitute for actual listening tests since it was standardized, although subjective tests are always preferred. The narrowband PESQ performs well for the situations for which it has been qualified, and the wideband PESQ (WPESQ) MOS, while initially not very accurate, has become more reliable, and relied upon, in recent years.

These techniques are the primary focus of the current work, and descriptions of other methods for determining the quality of coded speech and audio are left to the references.

2.1.3 Speech Coding Methods

The most common approaches to speech coding today center around three paradigms, namely, waveform-following coders, analysis-by-synthesis methods, and subband/transform domain methods [20]. Waveform-following coders attempt to reproduce the time domain speech waveform as accurately as possible, and for good performance, they are among the highest rate, but by far, the least complex codecs. Although waveform-following codecs produce bit streams with a relatively high bit rate, and are certainly not state-of-the-art in terms of

operational rate distortion performance, waveform-following coders are relevant to standards because they appear in many VoIP and other such packet switched backbone applications. Waveform-following coders are also of interest to us in this book because for these codecs, mean squared error is meaningful as a performance indicator in a way to be explained shortly, and thus we use them to map the mean squared error distortion values into PESQ-MOS values that are important for the analysis-by-synthesis codecs. The basic approach is elaborated in Chapter 4.

Analysis-by-synthesis methods utilize the linear prediction model, excitation codebooks, and a distortion measure based on a perceptually motivated spectral shaping to reproduce only those characteristics of the input speech determined to be most important [60]. As a consequence, a much lower bit rate is obtained for equivalent perceptual performance, but at the price of much increased algorithmic and implementation complexity. The analysis-by-synthesis structure is the basis for all high-performing narrowband and wideband voice codecs today, including codecs in the Code-Excited Linear Prediction (CELP) category. We will be comparing the operational rate distortion performance of CELP-based standardized codecs to our rate distortion bounds in what follows, after we describe the most important analysis-by-synthesis CELP codecs.

Subband/transform based codecs are utilized primarily for wideband, super wideband, and fullband speech/audio, and these approaches serve as the basis for MP3 players, audio on movies, and many audio streaming applications. Codecs for the fullband and super wideband regime have as their goal what might be called transparent reproduction of the source in the sense that no audible perceptual distortion is evident to most listeners, and these codecs use perceptual distortion measures beyond those considered in the rate distortion performance bounds obtained thus far and presented in this book. The basic approach used is still viable but a different distortion mapping must be developed to extend the bounds down to the very small distortion region. Codecs using subband/transform coding used for wideband speech are included in our current comparisons. Standards based on subband/transform coding are discussed briefly in the following to

allow the reader to have a clear context for the field and the work that remains to be performed.

Waveform Coding

Familiar waveform-following methods are logarithmic pulse code modulation (log-PCM) and adaptive differential pulse code modulation (ADPCM), and both have found widespread applications. Log PCM at 64 kilobits/sec (kbps) was developed in the 1960's and officially standardized in the United States as G.711 in 1972 [23, 20]. Historically, it is the speech codec long-used in the long distance public switched telephone network at a rate of 64 kbps since the 1960's, and it is the most widely employed codec for VoIP applications in the backbone network. It is an extremely simple, sample-by-sample nonuniform memoryless quantizer and it achieves what is called toll quality, which is the standard level of performance against which all other narrowband speech coders are judged. As such, a G.711 codec is almost always included in ACR subjective listening tests as a benchmark [14].

There are two closely related types of log-PCM quantizer used in the World– -law, which is used in North America and Japan, and A-law, which is used in Europe, Africa, Australia, and South America [23]. Both achieve toll quality speech, and in terms of the MOS value, it is usually between 4.0 and 4.5 for log-PCM, with the exact value depending on the particular set of listeners, language, and other test conditions [14]. It is not a predictive coder, though, and therefore, we do not utilize these codecs for MSE to PESQ-MOS mapping, for detailed reasons explained later.

Adaptive Differential Pulse Code Modulation (ADPCM) usually operates at 32 kbps or lower, and at 32 kbps, it achieves performance comparable to log-PCM by using an adaptive linear predictor to remove short-term redundancy in the speech signal before sample-by-sample adaptive quantization [20, 43]. A block diagram of an ADPCM speech encoder and decoder is shown in Figure 2.1. The most common form of ADPCM uses what is called backward adaptation of the predictors and quantizers to follow the waveform closely. Backward adaptation means that the predictor and quantizer are adapted based upon past

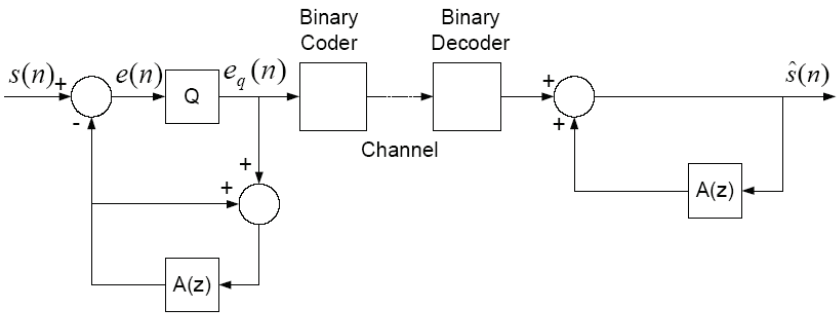


Figure 2.1: ADPCM Speech Encoder and Decoder

reproduced values of the signal that are available at the decoder as well as the encoder. No predictor or quantizer parameters are sent as side information along with the quantized waveform values. As will be seen, this codec appears in several standards.

Another type of ADPCM codec is embedded ADPCM, which uses embedded quantization, with the coarse quantization being utilized in the feedback prediction loop [20, 44]. An embedded quantizer characteristic has the property that coarse quantizers have step points that are a subset of the step points of a finer grained quantizer and the output points that satisfy the coarse quantizer are midpoints of the output levels of the next finer grain quantizer. The end result is that some number of least significant bits can be discarded by the network and the decoder can still reconstruct a good version of the signal without re-encoding and without losing coding synchronization with the encoder. For embedded ADPCM, the finer quantized error signal is summed with a predicted value calculated using the coarser quantized prediction error to obtain the reconstructed value.

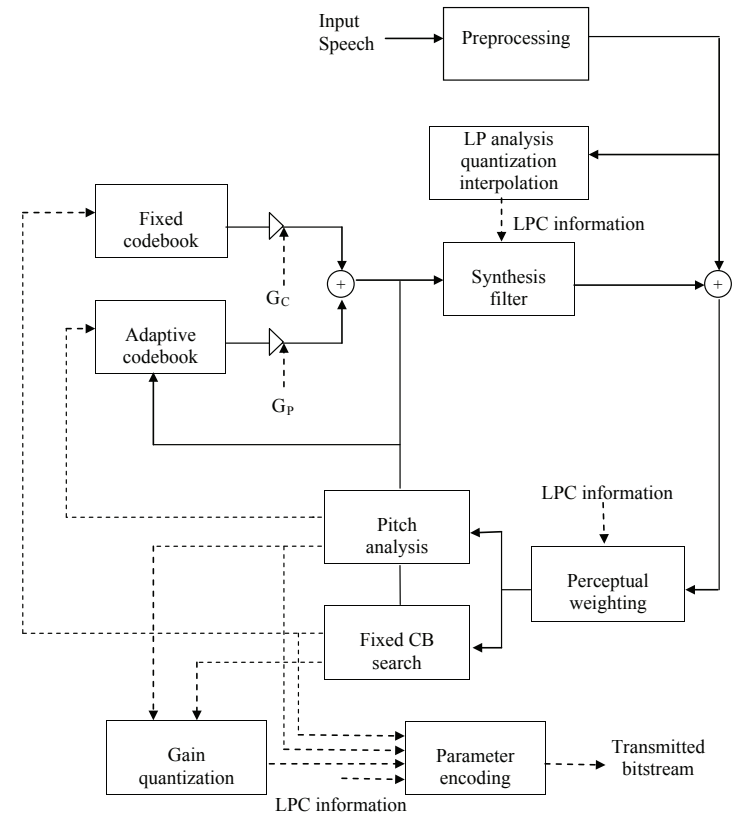
The performance of ADPCM can be measured using mean squared reconstruction error (MSE) or by using PESQ-MOS. MSE is not a perceptual measure, but the MSE does place different ADPCM systems in the correct order of performance. MSE or SNR for ADPCM should not be used for direct comparison to quantization methods that do not use prediction, since results indicate that the reconstruction noise

in ADPCM is less objectionable audibly than for log-PCM [68]. More directly, the MSE for ADPCM may be larger than for log-PCM, yet ADPCM may be preferred perceptually over log-PCM with a lower MSE. Apparently, the effect is that the reconstruction error in ADPCM is correlated with the input but the error in log-PCM is not. The correlated error is less objectionable perceptually.

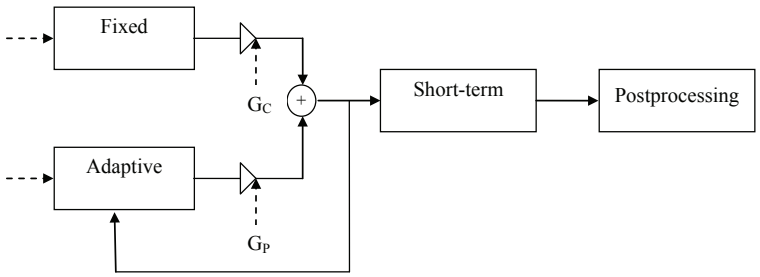
Analysis-by-Synthesis Methods

Analysis-by-synthesis (AbS) methods are a considerable departure from waveform-following techniques. The most common and most successful analysis-by-synthesis method in terms of widespread applications is code-excited linear prediction (CELP). In CELP speech coders, a segment of speech (say, 5 to 10 ms) is synthesized by passing entries from what is called a codebook into a long-term redundancy predictor, and the resulting combined excitation is used as input to a linear prediction model. The term analysis-by-synthesis derives from the fact that this process is repeated for all possible excitations in the codebook. For each excitation, an error signal, corresponding to the difference between the input speech and the synthesized speech, is calculated and passed through a perceptual weighting filter. The excitation that produces the minimum perceptually weighted coding error is selected for use at the decoder. Therefore, the best excitation out of all possible excitations for a given segment of speech is selected by synthesizing all possible representations at the encoder, and hence the name analysis-by-synthesis. The predictor parameters and the excitation codeword are sent to the receiver to decode the speech [57].

In recent years, it has become common to use an adaptive codebook structure to model the long term memory rather than a cascaded long term predictor. An encoder block diagram with the adaptive codebook structure is shown on Fig. 2.2(a) and a corresponding decoder using the adaptive codebook approach is shown in Fig. 2.2(b). The analysis-by-synthesis procedure is computationally intensive, and it is fortunate that algebraic codebooks, which have mostly zero values and only a few nonzero pulses, have been discovered and work well for the fixed codebook [59, 20].



(a) Encoder for Code-Excited Linear Predictive (CELP) Coding with an Adaptive Codebook



(b) CELP Decoder with an Adaptive Codebook and Postfiltering

Figure 2.2: CELP Encoder and Decoder

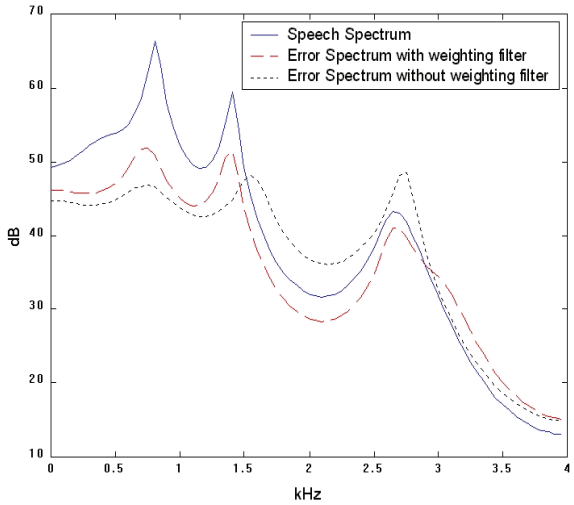


Figure 2.3: Perceptual Weighting of the Coding Error as a Function of Frequency

The perceptual weighting is key to obtaining good speech coding performance at the targeted lower rates, and the basic idea is that the coding error is spectrally shaped to fall below the envelope of the input speech across the frequency band of interest. Figure 2.3 illustrates the concept wherein the spectral envelope of a speech segment is shown, along with the coding error spectrum without perceptual weighting (unweighted denoted by short dashes) and the coding error spectrum with perceptual weighting (denoted by long dashes). The perceptually weighted coding error falls below the spectral envelope of the speech across most of the frequency band of interest, just crossing over around 3100 Hz. The coding error is thus masked by the speech signal itself. In contrast, the unweighted error spectrum is above the speech spectral envelope starting at around 1.6 kHz, which produces audible coding distortion for the same bit rate. CELP coding performs very well in terms of PESQ-MOS at bit rates as low as 8 kbps. However, CELP codecs do not attempt to match the time domain waveform, and as a result, MSE in the time domain is not a reasonable indicator of CELP codec performance. This is one of the challenges of developing rate distortion bounds for speech that are actual lower bounds to the best known voice codecs, which are based on CELP.

2.1.4 Current and Developing Standards

In this section, we describe the relevant details of current and some past standardized speech codecs for digital cellular and packet switched VoIP for wireless access points. We begin the discussion with ITU-T standardized codecs since some of those codecs have served as the basis for cellular codecs, and since some of these codecs also are used for VoIP applications. The performance of these codecs are compared to our rate distortion bounds in later chapters of the book.

ITU-T Standards

Tables 2.1 and 2.2 list some of the narrowband and wideband/fullband voice codecs that have been standardized by the ITU-T over the years, including details concerning the codec technology, transmitted bit rate, performance, complexity, and algorithmic delay. Those shown include G.711, G.726, G.728, and G.729 for narrowband (telephone bandwidth) speech (200 to 3400 Hz), G.722, G.722.1 [41], G.722.2 [42], and G.718 for wideband speech (50 Hz to 7 kHz) [38], and G.719 for fullband audio [39]. G.711 at 64 kilobits/sec. (kbps) is the voice codec most often used in VoIP backbone applications today. This codec is based on a nonlinear quantization method called logarithmic pulse code modulation (log-PCM), which allows low amplitude speech signal samples to be quantized finely while the larger amplitude samples are subjected to larger step sizes. The basic goal of this quantization approach is to preserve the quality of low amplitude samples, which are important to perceptual quality, and to maintain a relatively constant SNR performance over the full range of input signal power. This codec is the benchmark for narrowband toll quality voice transmission.

G.726 is a fully backward adaptive ADPCM codec that operates at selectable rates of 16, 24, 32, and 40 kbps [43]. G.727 is an embedded ADPCM codec with fine/coarse bits/sample quantization combinations of 5/4,3,2, 4/4,3,2, 3/3,2, and 2/2 [44]. Each combination of fine/coarse quantization yields a different bit rate and different performance. To obtain the flexibility afforded by embedding coding, the embedded codec suffers a loss in performance compared to non-embedded ADPCM at the same rate.

Table 2.1: ITU Narrowband Speech Codecs

Standards Body	ITU	ITU	ITU	ITU
Recommendation	G.711	G.726	G.728	G.729
Coder type	Companded PCM	ADPCM	LD-CELP	CS-ACELP
Dates	1972	1990	1992/4	1995
Bite rate (kbps)	64	16-40	16	8
Peak quality	Toll	≤ Toll	Toll	Toll
Complexity (MIPS)	<<1	~1	~30	≤20
Frame size (ms)	0.125	0.125	0.625	10
Lookahead (ms)	0	0	0	5
Codec delay (ms)	0.25	0.25	1.25	25

The G.729 codec is an analysis-by-synthesis codec based on algebraic code excited linear prediction (ACELP), and it uses an adaptive codebook to incorporate the long term pitch periodicity [45]. The G.729 codec structure has been very influential on subsequent voice coding standards for VoIP and digital cellular networks.

Even though we are quite comfortable communicating using telephone bandwidth speech (200 to 3400 Hz) for regular, relatively short telephone conversations, there is considerable interest in compression methods for wideband speech covering the range of 50 Hz to 7 kHz. The primary reasons for the interest in this band are that wideband speech improves intelligibility, naturalness, and speaker identifiability. The first application of wideband speech coding was to videoconferencing, and the first standard, G.722, separated the speech into two subbands and used ADPCM to code each band. The G.722 codec is relatively simple and produces good quality speech at 64 kbps, and lower quality speech at the two other possible codec rates of 56 and 48 kbps [59]. G.722 at 64 kbps is often employed as a benchmark for the performance of other wideband codecs.

G.722.2 is actually an ITU-T designation for the adaptive multirate wideband (AMR-WB) speech coder standardized by the 3GPP [5]. This coder operates at rates of 6.6, 8.85, 12.65, 14.25, 15,85, 18.25, 19.85, 23.05, and 23.85 kbps and is based upon an algebraic CELP (ACELP)

Table 2.2: ITU-T Wideband Speech Codecs

	ITU-T G.722	ITU-T G.722.1	ITU-T G. 722.2 3GPP AMR-WB	ITU-T G.718	ITU-T G.719
Coder Type	Subband ADPCM	MLT	ACELP	ACELP, MDCT	Adaptive resolution MDCT, FLVQ
Audio Bandwidth (Hz)	50-7000	50-7000	50-7000	50-7000	20-20000
Bitrate(s) (kbps)	48, 56, 64	24, 32	6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85	8, 12, 16, 24, 32 & 12.65 (G.722.2, AMR-WB, VMR-WB Interop Mode)	32…128 steps of 4 kbps up to 96kbps, steps of 8 kbps up to 128 kbps
Frame Length (mssec)	0.125	20	20	20	20
Algorithmic Delay (msec)	1.625	40	25	32.875 to 43.875	40
Comp. Complexity	10 MIPS	<5.5 WMOPS	27.2-39.0 WMOPS	57 WMOPS	15.39-21 WMOPS

analysis-by-synthesis codec. Since ACELP utilizes linear prediction, the coder works well for speech but less well for music, which does not fit the linear prediction model. G.722.2 achieves good speech quality at rates greater than 12.65 kbps and performance equivalent to G.722 at 64 kbps with a rate of 23.05 kbps and higher.

G.718 is a newer wideband speech codec that has an embedded codec structure and that operates at 8, 12, 16, 24, and 32 kbps, plus a special alternate coding mode that is bit stream compatible with AMR-WB [38]. G.719 is a fullband audio codec that has relatively low complexity and low delay for a fullband audio codec. This codec is targeted toward real-time communications such as in videoconferencing systems and the high definition telepresence applications, even though the algorithmic delay is getting somewhat high for real-time interactions [39].

The operational rate distortion performance of many of these codecs is compared to our newly obtained rate distortion bounds for narrowband and wideband speech in Chapter 4.

Digital Cellular Standards

The rapid deployment of digital cellular communications was facilitated by the development of efficient, high quality voice codecs. An important and somewhat dominant voice codec today is the Adaptive Multirate (AMR) Codec, both narrowband (NB) and wideband (WB) versions [1, 2]. The AMR codecs are based on the CELP method, which utilizes the

analysis-by-synthesis approach wherein fixed and adaptive codebooks excite a linear prediction model and the best excitation is chosen by minimizing a perceptual weighting criterion [64]. The rates for AMR-NB are 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2, or 12.2 kbits/s, and the rates for AMR-WB are 6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, and 23.85 kbits/s. The operational rate distortion performance of the AMR-NB and AMR-WB codecs are compared to our rate distortion bounds in Chapter 4.

Initial studies on 4G LTE cellular utilized the AMR codecs, but a new codec, designated as EVS, is being developed for later deployment. That standardization process is on-going at the time of this writing. Our rate distortion bounds are expected to lower bound the performance of the new EVS codec for narrowband and wideband speech sources.

VoIP Standards

For voice codecs normally implemented in VoIP solutions, we find that at this point in time, almost all codecs are borrowed from other standards bodies. The codecs that are often available in VoIP systems are G.711 (NB@64 kbits/s), G.722 (WB@48, 56, and 64 kbits/s), G.729 (NB@8 kbits/s), and AMR-NB/WB [64]. These codecs are all included in the later performance comparisons in Chapter 4.

There is another codec standardization effort that has the goal of coding narrowband voice all the way up to fullband audio and with the constraint of low delay. The Opus Audio Codec is being designed for interactive voice and audio and has three modes: (a) a linear prediction-based mode for low bit rate coding up to 8 kHz bandwidth, (b) a hybrid linear prediction and MDCT (Modified Discrete Cosine Transform) mode for fullband speech/audio at medium bit rates, and (c) an MDCT-only mode for very low latency coding of speech and audio. Details of this codec can be found in [83]. Although not yet tested, it is expected that our rate distortion bounds will apply to the Opus codecs based on linear prediction in (a) and any narrowband and wideband speech codecs in (c).

2.2 Video Codecs

2.2.1 Characteristics of Video Signals

Digital videos are sequences of digital images. Each image is composed of many PICture ELements, commonly referred to as pixels. The pixels in a digital image are arranged in rows and columns on the video display. They are so close spatially that they appear to be connected when viewed by human eyes. The number of pixels in an image frame, denoted as the "resolution" of a digital video or a digital image, varies from thousands to millions. For digital videos, the resolution is often described by commonly known code names and their acronyms, such as Video Graphics Array (VGA), or as the dimension (height and/or width) of the image in pixels, such as 720p and 1080p; for still digital images, the resolution is often described by the total number of pixels (in Megapixels) of the image. In Table 2.3 we list a few common resolutions of digital videos. Another parameter related to video resolution is aspect ratio, which is defined as the ratio of the width over height of a video frame. As seen in Table III, the aspect ratio of the digital videos is 4 : 3 for the older and lower resolutions; and it is 16 : 9 for the newer and higher resolutions.

Table 2.3: Resolution of typical digital videos

Video Resolution	Width × Height in Pixels	Megapixels per Frame	Aspect Ratio
QSIF	176 × 120	0.021	4:3
QCIF	176 × 144	0.025	
QVGA	320 × 240	0.077	
SIF	352 × 240	0.084	
CIF	352 × 288	0.10	
VGA	640 × 480	0.31	
4SIF	704 ×480	0.34	
4CIF	704 × 576	0.41	
w360p	640 × 360	0.23	16:9
w448p	768 × 448	0.34	
720p	1280 × 720	0.92	
1080p	1920 × 1080	2.07	

Each pixel of a video frame is represented by three values, normally in the red, green and blue (RGB) color palette. Each of the three red, green and blue pixel values is traditionally quantized using 8 bits, and

hence the color space is composed of 256 levels each of the red, green and blue components. Recently, higher fidelity in the video has been sought after and video products with 10 bits and 12 bits per color plane have emerged in consumer markets.

The content of a video frame is often identified by the objects that can be seen dominantly on the video frame. The objects are considered to be on the "foreground" of the video frame while naturally the rest of the video frame is considered to be the "background". The content of a sequence of video frames is then identified by the movement, or "motion", of the foreground objects on the relatively static background. The composition of the foreground objects and the background is often referred to as a video "scene". The transition in a video from one scene to another is often referred to as a "scene change" or a "scene cut". Although there is not an exact definition of the "scene" since the details on the foreground objects and on the background often change in natural videos (unlike in some animated videos), the video "scene" is nevertheless a convenient concept to denote the rather similar content exhibited in a segment of video sequence.

Although video content is generated at a high data rate, there is also a huge amount of redundancy in a video sequence, both spatially, among nearby pixels, and temporally, among successive frames. When it comes to compress the videos either for storage on a physical medium such as a DVD or for communication over a network, the spatial frequency transformation takes a central role, while video content modeling in the spatial and temporal domains take a back seat. Either entire video frames are transformed into the frequency domain, such as in the wavelet based compression methods including motion JPEG, or the video frames are divided into blocks of pixels, such as 16 by 16 pixel macroblocks (MB), and each block (after motion estimation and compensation) is transformed into the frequency domain separately. The most popular video coding standards, including MPEG-2, H.264/AVC, and the newest standard HEVC, utilize this motion-compensated spatial frequency domain transformation paradigm. This line of video coding standards is fairly successful at reducing the video data rate to less than one percent of the video raw data rate while providing satisfac-

tory visual quality. However, its macroblock-based, spatial frequency domain centric nature has not facilitated the development of a video source model suitable for characterizing videos, especially for the development of rate distortion bounds. And the video content modeling is left with the original challenge of dealing with the sheer volume of information contained in a video, for example, a data rate of 180 megabytes per second for an uncompressed high definition 1080p video.

2.2.2 Performance Measures

Before getting into the details of video coding and video coding standards, it is important to examine the criteria for measuring the effectiveness of a video coding algorithm used in a video coding standard. Unlike voice where codecs are usually designed for a specified quality at a given compression rate, the video coding standards do not specify a compressed video bit rate and corresponding quality. Instead, the international video coding standards [32, 33, 34, 35] provide vast flexibility for each application to design its own encoder according to its specific compression requirement. A video coding technique or a video coding standard is considered more effective if for a specific original video, at a specified compression rate, the compressed video yields a higher quality; or for a specified compressed video quality, it can achieve a higher compression rate.

In addition to the operational rate distortion performance, the computational complexity involved in compressing the video is a critical criterion in selecting a video coding technique for a standard. However, the purpose of reviewing practical video coding techniques and standards in this chapter is to highlight the basic video coding approaches and the distortion measures used to characterize their operational rate distortion performance, and thus lay the groundwork for later comparisons to our new rate distortion bounds. As a consequence, video codec implementation complexity is not discussed here.

Video quality measurement is a very difficult problem. The most commonly used objective video quality measure is the mean squared error (MSE) of the distorted video with respect to the original or reference video, or the peak signal-to-noise ratios (PSNR). The relationship

between these two quantities, assuming an 8-bit representation of pixel value and hence a peak signal value of 255, is

$$PSNR(dB) = 10 \log_{10} \frac{255^2}{MSE}. \tag{2.2.1}$$

The MSE/PSNR based measures compute only the pixel-to-pixel difference of the processed (reconstructed) video and the referenced (original) video. They rely on the availability of the original video as the reference video. MSE or PSNR do not explore the perceptual effects of any distortion, and therefore, they are criticized for correlating poorly with human perception in some scenarios.

On the other hand, the objective perceptual video quality measures based on the lower order processing of human vision systems (HVS) may correlate better with human perception but they are computationally very intensive [69, 88]. These objective video quality measures can be divided into three different categories according to the availability of the reference video. These three categories are:

- Full Reference (FR) - the reference video is available;
- Reduced Reference (RR) - the reference video is not available but a description of the reference video is available;
- No Reference (NR) - no information about the reference video is available.

Table 2.4 summarizes standardization status of the objective video quality measures by these three categories.

Despite the advances in objective perceptual video quality measures and the well known criticisms and weaknesses of MSE/PSNR measures, MSE/PSNR remains the most popular video quality measure for evaluating the performance of a video coding standard or a specific codec implementation of the standard. In fact, during the standardization process of the newest video coding standard HEVC in the past three years, although some subjective tests were conducted, "[b]it rate savings is more frequently measured by using simpler 'objective metrics,' especially with the simple peak signal-to-noise ratio (PSNR) metric"

Table 2.4: Standardization status of objective video quality measures

Category	Standards	Dates	Remarks
Pixel level - Full Reference	ITU-R BT.1683/ ITU-T J.144 for SDTV	Jun./Mar. 2004	Includes four models: British Telecom (UK) Yonsei University/Radio Research, Laboratory/SK Telecom (South Korea) Center for Telecommunications Research and Development (Brazil) NTIA/ITS (USA)
	ITU-T J.247 for multimedia (VGA, CIF, QCIF)	Aug. 2008	Includes four models: NTT (Japan) OPTICOM (Germany) Psytechnics (UK) Yonsei University (South Korea)
	ITU-T J.341 for HDTV	Jan. 2011	Includes one model: SwissQual (Switzerland)
Pixel level - Reduced Reference	ITU-T J.246	Aug. 2008	Includes one model: Yonsei University (South Korea)
Pixel level - No Reference	N/A	Attempted but failed in 2008	

[66]. Fortunately for our work, MSE is also deeply rooted in the rate distortion theory of Gaussian random variables, as reviewed in Chap. 3. Although as shown in Table 2.4, there are standardized full reference perceptual objective video quality measures, these measures are inherently very complicated. In order for these perceptual objective video quality measures to be sufficiently accurate to approximate human perception, they typically employ transformation of video sequences into perceptually meaningful domains and engage multiple paths and/or layers of computation in those domains. They are highly unlikely, if not impossible, to allow for a tractable mathematical calculation of the rate distortion bounds. For these three reasons we use MSE as the performance measure in both deriving the theoretical rate distortion bounds for videos and in collecting the operational rate distortion curves of the video coding standards.

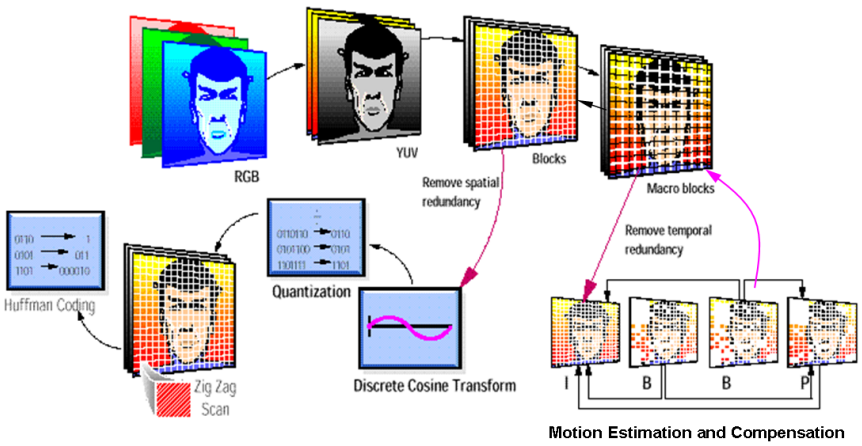


Figure 2.4: Block-based hybrid video coding with motion estimation and compensation

2.2.3 Motion-Compensated Transform Coding

In this section, we highlight some of the important details for a group of video coding techniques called "block-based hybrid video coding with motion estimation and motion compensation."' This basic structure is used by a majority of video codecs in the dominant standards. Figure 2.4 illustrates key steps in this video coding approach.

As shown in Fig. 2.4, a first step in video compression is to convert the red, green and blue (RGB) values of each pixel into one luminance value and two chrominance values, often referred to as the YUV color planes. The conversion from RGB to YUV separates the luminance signal, which is perceptually much more important, from the chrominance signal, which is less important perceptually and can be represented at a lower resolution and a lower data rate to achieve more efficient data compression. Different codecs use different chrominance downsampling ratios as appropriate for their applications.

After color conversion, each video frame is partitioned into arrays of pixel blocks, for example, 16×16 pixel blocks that are commonly referred to as macroblocks (MB). Each block then undergoes motion estimation to locate a similar block in previously coded frame(s) and

motion compensation to account for the difference in pixel values of the current block and its reference block. The basic premise of motion estimation and compensation is that for many typical video sequences of interest, consecutive video frames are quite similar, except for changes largely induced by objects moving within the frames. Therefore, the frames are usually processed in groups, called a group of pictures or GOP.

One frame in the GOP (usually the first) is encoded simply as a still image. This frame is called an intra-coded frame, or I-frame. Other frames in the GOP are called predicted frames, or P-frames, and bi-directionally predicted frames called B-frames. The P-frames are predicted from the I-frame or other P-frames that come before them in time. B-frames can be predicted from future frames, and usually, B-frames are predicted from two directions, from an I- or P-frame that precedes them and from an I- or P-frame that follows them in time sequence. In this case the future frames need to be encoded before the predicted frames and thus, the encoding order does not necessarily match the frame order. A GOP is composed of I-, P-, and B- frames in unlimited number of combinations and orders. For instance, one GOP can be described as IBBPBBPBBPBB, which consists of all three types of frames: the first frame is an I-frame; the second and third frames are B-frames; the fourth frame is a P-frame, and so on.

When a frame is predictively coded, each block in this frame is predicted from a block of equal size in a previously coded, called reference frame, and the difference of the positions of the two blocks is called a motion vector. The block in the reference frame is subtracted from the original block in the frame being predicted to form a residue (or residual or prediction error) block of pixels. The two steps in this process are motion estimation and motion compensation, respectively. The reference block for a block in a B-frame is formed as a weighted average of the two blocks from the two reference frames. It is possible to reference a block that is shifted from the current block by a non-integer vector, such as a half pixel or a quarter pixel. This is called sub-pixel precision motion estimation and compensation. For sub-pixel motion estimation and compensation, the reference block is formed by interpolating the

nearby full pixel values.

Motion estimation and compensation reduces the temporal redundancy across the video frames. To reduce the spatial redundancy within a frame, the intra-coded MBs and the residue block of the inter-coded MBs are transformed to the frequency domain through 8x8 discrete cosine transform (DCT) or similar integer transforms. The coefficients produced by the DCT or other type of frequency domain transform are quantized, ordered in the block according to horizontal and vertical harmonics, and then zig-zag scanned, after which entropy (lossless) coding is applied. Uniform scalar quantization is often chosen to quantize the coefficients because of its simplicity. The entropy coding normally uses variable-length coding tables, among which the run-length coding combines a number of consecutive zero-valued quantized coefficients and the value of the next non-zero quantized coefficient into a single symbol.

The decoding process consists of performing, to the extent possible, an inversion of each stage of the encoding process. For the stages that cannot be exactly inverted such as quantization and DCT, a best-effort approximation of inversion is performed.

2.2.4 Current and Developing Standards

Video codec designs are often standardized, i.e., specified precisely in published documents. Since the 1980's, video compression standardization has been dominated by two international organizations: the International Telecommunications Union - Telecommunications standardization sector (ITU-T) 's Video Coding Experts Group (VCEG) and the International Standardization Organization (ISO) and International Electro-technical Commission (IEC) - Joint Technical Committee (JTC) 's Moving Picture Experts Group (MPEG). Figure 2.5 is a time table of the major image and video coding standards published by VCEG and MPEG. The older video coding standards in this table - H.261/2/3, MPEG-1/2/4 had been powerful engines behind the commercial success of digital video compression. They had played pivotal roles in establishing the technology by providing interoperability among products developed by different manufacturers.

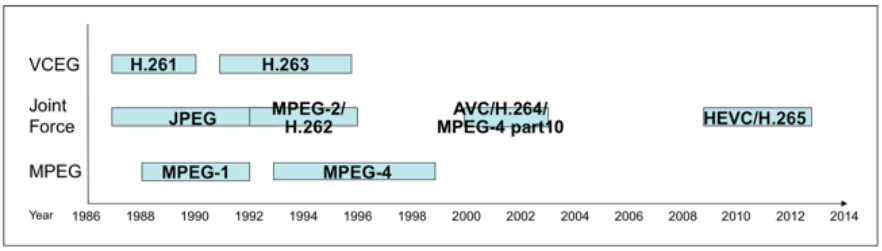


Figure 2.5: A time table of the major image and video coding standards published by VCEG and MPEG.

The more recently established Advanced Video Coding (AVC) standard, also named ITU-T Recommendation H.264 and MPEG-4 Part 10, offers a coding efficiency improvement by a factor of two over previous standards and its network abstraction layer (NAL) transports the coded video data over networks in a more "network-friendly" way [89]. Because of these two features, the AVC/H.264 standard emerged as the method of choice for the next generation video networks.

All the video coding standards listed in Fig. 2.5 follow the general framework of block-based motion-compensated transform coding that is discussed briefly in the previous section of this chapter. The newer standards normally include new schemes or refinements of the schemes that already exist in the older standards. In the following we briefly explain the concepts of three such new schemes in AVC/H.264: intra-frame prediction, the integer transform, and quantization with scaling. These features, or the underlying ideas, are relevant to later chapters of this book.

Intra-frame prediction is a new feature in AVC/H.264 which removes, to a certain extent, the spatial redundancy in neighboring MBs (16×16 blocks) or smaller 4×4 blocks. If a MB or a 4×4 block is to be encoded in intra-mode, a prediction MB/block is first formed based on previously encoded and reconstructed surrounding pixels. The prediction block is then subtracted from the current block prior to encoding. For the luminance samples, there are a total of nine prediction modes for each 4×4 block and a total of four prediction modes for each 16×16 MB (modes 0 to 3 of the nine modes for 4×4 blocks).

Intra-frame prediction can be better explained by reference to Fig. 2.6. In this figure, the small blocks in each of the nine big 5×5 blocks represent individual pixels and the different shades of gray in the small blocks represent the luminance values of the individual pixels. The pixels in the 4×4 block on the bottom right corner of the 5×5 blocks are to be encoded with intra-frame prediction. The nine pixels surrounding the 4×4 blocks (four on top, four on the left and one on the top left corner) are previously encoded and reconstructed, here assigned different luminance values as an example. For each of the nine intra-frame prediction modes, a predefined formula is applied to form the prediction block P. The 4×4 blocks on the bottom right corner of the nine 5×5 blocks in Fig. 2.6 show the prediction blocks for all nine intra-frame prediction modes given the example surrounding pixel values. Mode 2, DC mode is designed for a prediction block to have the same luminance value across all pixels. The other eight modes, and their corresponding formulas for calculating the prediction block, are designed to capture a different direction of the gradience of the local texture, with a 22.5 degree difference from one mode to the next.

The idea of intra-frame prediction is that one of the nine prediction blocks is sufficiently similar to the actual 4×4 block to be encoded. And hence only the difference needs to be encoded besides the intra-frame prediction mode itself. For modes 3 and 7, the four pixels on top right of the current 4×4 block (not shown in Fig. 2.6) are also used for calculating the prediction block. Only the surrounding pixels on the left and on the top of the current blocks are used in forming the prediction block because the blocks are encoded from left to right in each row of the blocks and from top row to bottom row in an image frame.

To transform the residual MB after intra-/inter- prediction to the frequency domain for further processing, an integer transform, instead of an 8×8 DCT, is used in AVC/H.264. The integer transform can be formulated in Eq. (2.2.2) where $\mathbf{E_f}$ contains the post-scaling factors. The integer transform is applied to each 4×4 block $\mathbf{X}$. An integer transform can avoid the mismatch between encoder and decoder inherent to the implementation of the forward DCT at the encoder and the

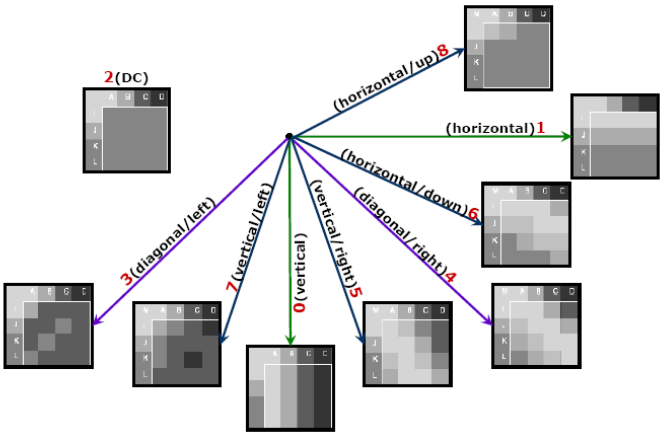


Figure 2.6: The intra-prediction modes for 4×4 blocks in AVC/H.264

inverse DCT at the decoder.

$$\mathbf{Y} = \left\{ \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \mathbf{X} \begin{bmatrix} 1 & 2 & 1 & 1 \\ 1 & 1 & -1 & -2 \\ 1 & -1 & -1 & 2 \\ 1 & -2 & 1 & -1 \end{bmatrix} \right\} \otimes \mathbf{E_f}. \quad (2.2.2)$$

The post-scaling factors contained in $\mathbf{E_f}$ are combined with uniform quantization following the integer transform. A total of 52 values of quantization step sizes (Qstep) are supported in AVC/H.264 and they are indexed by quantization parameters (QP), as shown in Table 2.5. Note that Qsteps and QPs are arranged in a way that an increase of 1 in QP yields an increase of Qstep by approximately 12%, which results in an approximate bit rate reduction of 12%.

At the decoder side, the quantized 4×4 block $\hat{\mathbf{Y}}$ is post-scaled by the components in $\mathbf{E_i}$ and then inverse transformed into the reconstructed

Table 2.5: Quantization stepsizes supported in AVC/H.264

QP	0	1	2	3	4	5	6	7	8
QStep	0.625	0.6875	0.8125	0.875	1	1.125	1.25	1.375	1.625
QP	9	10	11	12	...	18	...	24	...
QStep	1.75	2	2.25	2.5	...	5	...	10	...
QP	30	...	36	...	42	...	48	...	51
QStep	20	...	40	...	80	...	160	...	224

residual block $\hat{\mathbf{X}}$, as

$$\hat{\mathbf{X}} = \begin{bmatrix} 1 & 1 & 1 & 1/2 \\ 1 & 1/2 & -1 & -1 \\ 1 & -1/2 & -1 & 1 \\ 1 & -1 & 1 & -1/2 \end{bmatrix} \left\{ \hat{\mathbf{Y}} \otimes \mathbf{E_i} \right\} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1/2 & -1/2 & -1 \\ 1 & -1 & -1 & 1 \\ 1/2 & -1 & 1 & -1/2 \end{bmatrix}. \tag{2.2.3}$$

It is perhaps surprising but critical to note that only the decoding process is standardized. More specifically, it is the syntax of the bit stream that serves as input to the decoder that is standardized. The encoding process is typically not specified at all in a standard, and hence developers are free to design their encoder however they want, as long as the video can be decoded by a standard compatible decoder, which means that the encoder need only produce a bit stream that fits the standardized syntax. This allows flexibility and ingenuity in encoder optimization and allows the technology to be molded to fit a given application, thus adjusting to the cost-performance trade-offs suited to particular requirements [63].

The newest big thing in video coding standardization is the High Efficiency Video Coding (HEVC) standard. HEVC is the most recent joint video coding standardization project of ITU-T Video Coding Experts Group (ITU-T Q.6/SG 16) and ISO/IEC Moving Picture Experts Group (ISO/IEC JTC 1/SC 29/WG 11). Also referred to as H.265, HEVC has recently been finalized and published. The goal of HEVC is to achieve video quality comparable to H.264 High Profile at about half of the bit rate. Current estimates suggest the computation cost of improved coding efficiency is about two times at the decoder and four to ten times at the encoder, compared to H.264 High Profile. HEVC

is targeted toward higher resolutions such as 720p, 1080p, and next-generation HDTV displays of Ultra HDTV (7680×4320). Many new coding techniques are included in HEVC. Please refer to its standard document [36] for details. One particular technique, extended coding tree block sizes of up to 64×64 and a larger set of prediction modes, will affect the comparison of the operational rate distortion bounds of HEVC with theoretical rate distortion bounds. This will be discussed in detail in Chapter 5.

3

The Rate Distortion Problem

In this book we devise new source models and distortion measures for natural voice and video signals and utilize these models and distortion measures to obtain meaningful rate distortion bounds for real sources that bound the performance of the very best performing voice and video codecs. In this chapter we review the theoretical underpinnings of information theory and rate distortion theory for common source models and distortion measures and present relevant rate distortion theory results. We cover only those models, distortion measures, and results needed in later chapters of the book. More general treatments and almost all proofs are left to the references.

3.1 Rate Distortion Theory Basics

The "Rate for a Source Subject to a Fidelity Evaluation" was introduced by Shannon in his original paper [75] in 1948. He returned to the concept and dealt with it exhaustively in his 1959 paper [77], where among many significant contributions, he defined the rate distortion function, proved positive and negative coding theorems, derived $R(D)$ for several important cases, and derived the extremely useful Shannon lower

bound to $R(D)$. Meanwhile, Kolmogorov [58] and Pinsker [70] in the Soviet Union began to develop rate distortion theory in the mid-1950's.

In information theory, the mutual information of two random variables or processes is a quantity that measures the mutual dependence of the two random variables or processes. For two continuous amplitude random variables X and $\hat{X}$, their mutual information is defined as

$$I(X;\hat{X}) := \int_x \int_{\hat{x}} p(x,\hat{x}) \log\left(\frac{p(x,\hat{x})}{p(x)p(\hat{x})}\right) dxd\hat{x}. \tag{3.1.1}$$

For lossy source coding problems and rate distortion theory, $\hat{X}$ refers to the reconstruction of the source X, and $I(X;\hat{X})$ is derived as

$$\begin{gathered} I(X;\hat{X}) = \textstyle\int_x \int_{\hat{x}} p(\hat{x}|x)p(x) \log\left(\frac{p(x|\hat{x})p(\hat{x})}{p(x)p(\hat{x})}\right) dxd\hat{x} = \\ -\textstyle\int_x p(x)\log\left(p(x)\right)dx + \int_{\hat{x}}\int_x p(x|\hat{x})p(\hat{x})\log\left(p(x|\hat{x})\right)dxd\hat{x} = \\ h(X) - h(X|\hat{X}) \end{gathered} \tag{3.1.2}$$

The last form of the equation is often useful in writing closed form expressions for mutual information.

A mathematical characterization of the rate distortion function is given by the following fundamental theorem of rate distortion theory [6, 10]. (Note that the mathematical characterization given in the theorem is given physical meaning by the proof of a coding theorem, which shows that there exists an encoder/decoder pair generated according to what is called a test channel $p(\hat{x}|x)$ that achieves the rate distortion pair $(R(D), D)$ specified in the theorem. More specifically, the proof involves specifying encoder and decoder functions f_n and g_n, respectively, to generate a length-n code with a codebook of 2^{nR} sequences. Then, using a random coding argument and distortion typicality for encoding and decoding, it can be shown that the average distortion given by

$$d(\underline{x},\underline{\hat{x}}) = \sum_{i=1}^{n} \frac{1}{n} d(x_i,\hat{x}_i), \tag{3.1.3}$$

asymptotically approaches D as the code block length n gets large.

The mathematical characterizations of rate distorion bounds in later theorems are obtained in a similar fashion.)

Theorem 3.1. *(Shannon's third theorem)* The minimum achievable rate to represent an i.i.d. source X with a probability density function $p(x)$, by $\hat{X}$, with a bounded distortion function $d(X, \hat{X})$ such that $\int_x \int_{\hat{x}} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D$, is equal to

$$R(D) = \min_{p(\hat{x}|x): \int_x \int_{\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}), \tag{3.1.4}$$

where $I(X; \hat{X})$ is the mutual information between X and $\hat{X}$.

Given the source density $p(x)$, the minimization is over all admissible test channels, that is, all $p(\hat{x}|x)$ satisfying the average distortion constraint. The critical roles of the probabilistic source model $p(x)$ and the distortion measure $d(x, \hat{x})$ are thus evident in the definition of $R(D)$ above.

The challenges in developing rate distortion functions for real sources should now be evident since somehow the probability density function $p(x)$ needs to capture all of the intricacies of a real source, and the distortion measure must indicate the subjective quality of the reproduced source output $\hat{X}$. Additionally, both the source model and the distortion measure must yield an analytically tractable optimization problem.

The voice and video signals in real life can be represented as continuous amplitude random processes on continuous time. In the case of video signals, the time includes the spatial dimension as well as the temporal dimension. In this book, we deal with digitized natural voice and video signals, and consequently we consider voice and video signals to be discrete in time, i.e., voice samples and video pixels, but to be continuous amplitude. As a result, the expression for $R(D)$ given above fits our physical problems well.

To make further progress, however, we need to specialize both the source probability density and the form of the distortion measure.

3.2 Rate Distortion Results for Gaussian Sources and Squared Error Distortion

Whether or not there exists a closed-form solution for the rate distortion function $R(D)$ depends on the distribution of the source and the criterion selected to measure the fidelity of reproduction between the source and its reconstruction. One of the most tractable formulations has been for Gaussian sources and the squared error difference distortion measure. To lay the groundwork for developing our bounds, we briefly review rate distortion functions for time-discrete Gaussian sources subject to the squared error distortion measure.

3.2.1 Scalar Gaussian Source with Mean Squared Error

The rate distortion function of a scalar Gaussian source with squared error distortion is as follows [6].

Theorem 3.2. The rate distortion function for a scalar Gaussian random variable $X \sim N(0, \sigma^2)$ with squared error distortion measure $d(x, \hat{x}) = (x - \hat{x})^2$ is

$$\begin{aligned} R(D) &= \min_{p(\hat{x}|x):\int_x \int_{\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x})dxd\hat{x} \leq D} I(X;\hat{X}) \\ &= \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2 \\ 0, & D > \sigma^2 \end{cases} \end{aligned} \tag{3.2.5}$$

The Shannon lower bound is particularly useful in proving this result and in displaying the probability densities that achieve the minimum. To move toward real source models, we now consider a vector of independent but not identically distributed Gaussian random variables.

3.2.2 Reverse Water-filling

The rate distortion function of a vector of independent (but not identically distributed) Gaussian sources is calculated by the reverse water-filling theorem [10]. This theorem says that one should encode the independent sources with equal distortion level λ, as long as λ does not exceed the variance of the transmitted sources, and that one should not

transmit at all those sources whose variance is less than the distortion λ.

Theorem 3.3. *(Reverse water-filling theorem)* For a vector of independent random variables $X_1, X_2, ..., X_n$ such that $X_i \sim N(0, \sigma_i^2)$ and the distortion measure $d(\underline{x}, \underline{\hat{x}}) = \sum_{i=1}^{n}(x_i - \hat{x}_i)^2$, the rate distortion function is

$$R(D) = \min_{p(\underline{\hat{x}}|\underline{x}):\int_{\underline{x}}\int_{\underline{\hat{x}}} p(\underline{\hat{x}}|\underline{x})p(\underline{x})d(\underline{x},\underline{\hat{x}})d\underline{\hat{x}}d\underline{x} \leq D} I(X;\hat{X}) = \sum_{i=1}^{n} \frac{1}{2}\log\frac{\sigma_i^2}{D_i}, \tag{3.2.6}$$

where

$$D_i = \begin{cases} \lambda & 0 \leq \lambda \leq \sigma_i^2 \\ \sigma_i^2 & \lambda > \sigma_i^2 \end{cases}, \tag{3.2.7}$$

and $\sum_{i=1}^{n} D_i = D$.

Reverse water-filling is a classical result in rate distortion theory and it plays a major role in future chapters. The following section provides the road map for applying the reverse water-filling result for voice and video sources.

3.2.3 Stationary Gaussian Sources with Memory

In this section, we show how to connect the usual statistics that we have about speech or video sources, namely the autocorrelation or covariance function, with a decomposition that allows us to use the reverse water-filling theorem on parallel Gaussian sources to calculate the rate distortion function. The following derivation of rate distortion theory for stationary Gaussian sources apparently first appeared in [15].

Let A be an unitary matrix denoting an orthonormal linear transformation from a vector of random variables $\underline{X}$ to another vector of random variables $\underline{\Theta}$ as

$$\underline{\Theta} = A\underline{X}, \ \underline{\hat{X}} = A^{-1}\underline{\hat{\Theta}} = A^T\underline{\hat{\Theta}}. \tag{3.2.8}$$

The following relations between $\underline{X}$ and $\underline{\Theta}$ can be derived:

Mean squared error:

$$\begin{array}{rcl} D(\underline{X},\underline{\hat{X}}) & = & E[(\underline{X}-\underline{\hat{X}})^T(\underline{X}-\underline{\hat{X}})] \\ & = & E[(\underline{\Theta}-\underline{\hat{\Theta}})^T A^T A(\underline{\Theta}-\underline{\hat{\Theta}})] \\ & = & E[(\underline{\Theta}-\underline{\hat{\Theta}})^T(\underline{\Theta}-\underline{\hat{\Theta}})] \\ & \text{orthonormal} & \\ & = & D(\underline{\Theta},\underline{\hat{\Theta}}); \end{array} \tag{3.2.9}$$

Mutual information:

$$I(\underline{X};\underline{\hat{X}}) \underset{|A|\neq 0}{=} I(\underline{\Theta};\underline{\hat{\Theta}}) \geq \sum_{i=1}^{n} I(\Theta_i;\hat{\Theta}_i), \tag{3.2.10}$$

with equality if and only if (iff) Θ_i's are independent.

As shown above, both the distortion (chosen here to be the summation of squared errors) and the mutual information of a random process $\underline{X}$ are equal to those of the unitary transform of the random process $\underline{\Theta}$, and therefore the rate distortion function of $\underline{X}$ equals the rate distortion function of $\underline{\Theta}$.

Now think of $\underline{X}$ as pixel values of a digital image with correlated Gaussian elements. To utilize the Reverse-water-filling theorem discussed in the previous section, the goal is to find $\underline{X}$'s unitary transform $\underline{\Theta}$ with independent elements. To achieve this goal, we utilize the well known Karhunen Lòeve Transform (KLT), which is also called principal component analysis, to decorrelate the source $\underline{X}$ as follows.

Letting the covariance function of a stationary zero-mean Gaussian source be denoted by

$$\phi(n) = E[x_i x_{i+n}], \tag{3.2.11}$$

and letting Φ_n be the $n \times n$ covariance matrix of the source, with its entries defined in Eq. (3.2.11), then, $\Phi_n = \{\phi(|i-j|), i,j = 1,...,n\}$. Denoting $\{\underline{\psi}_i, i = 1,...,n\}$ as the normalized eigenvectors of Φ_n with corresponding eigenvalues $\{\lambda_i, i = 1,...,n\}$, so that

$$\Phi_n \underline{\psi}_i = \lambda_i \underline{\psi}_i, \tag{3.2.12}$$

then we write

$$\Phi_n = \Psi_n \Lambda \Psi_n^T, \tag{3.2.13}$$

where $\Psi_n = [\underline{\psi}_1, \underline{\psi}_2, ..., \underline{\psi}_n]$.

Since covariance matrices are symmetric, there always exists an eigenvalue decomposition of the covariance matrix with real eigenvalues, and furthermore, covariance matrices are positive semi-definite, therefore all their eigenvalues are non-negative, yielding

$$\underline{\Theta} = \Psi_n^T \underline{X}. \tag{3.2.14}$$

Thus, the rate distortion function of a stationary Gaussian source $\underline{X}$ with covariance matrix Φ_n can be computed as the rate distortion function of a stationary Gaussian source $\underline{\Theta}$, where $\underline{\Theta}$ has independent Gaussian elements, each of variance λ_i, which are eigenvalues of the covariance matrix Φ_n. The rate distortion function of $\underline{\Theta}$ is in turn solved by the reverse-water filling theorem.

3.2.4 Rate Distortion Function for a Gaussian Autoregressive Source

Since Shannon's rate distortion theory requires an accurate source model and a meaningful distortion measure, and both of these are difficult to express mathematically for real physical sources such as speech, these requirements have limited the impact of rate distortion theory on the lossy compression of speech. There have been some notable advances and milestones, however. Berger [6] and Gray [26], in separate contributions in the late 60's and early 70's, derived the rate distortion function for Gaussian autoregressive (AR) sources for the squared error distortion measure. Since the linear prediction model, which is an AR model, has played a major role in voice codec design for decades and continues to do so, their results are highly relevant to our work. The basic result is summarized in the following theorem [6]:

Theorem 3.4. Let $\{X_t\}$ be an mth-order autoregressive source generated by an i.i.d. $N(0, \sigma^2)$ sequence $\{Z_t\}$ and the autoregression constants $a_1, ..., a_m$. Then the MSE rate distortion function of $\{X_t\}$ is given parametrically by

$$D_\vartheta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \min \left[\vartheta, \frac{1}{g(\omega)} \right] d\omega, \tag{3.2.15}$$

and

$$R(D_\vartheta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \max \left[0, \frac{1}{2} \log \frac{1}{\vartheta g(\omega)} \right] d\omega, \tag{3.2.16}$$

where

$$g(\omega) = \frac{1}{\sigma^2}\left|1 + \sum_{k=1}^{m} a_k e^{-jk\omega}\right|^2 . \tag{3.2.17}$$

The points on the rate distortion function are obtained as the parameter ϑ is varied from the minimum to the maximum of the power spectral density of the source. ϑ can be associated with a value of the average distortion, and as illustrated in Fig. 3.1, only the shape of the power spectral density, $\Phi(\omega)$, above the value of ϑ is reproduced at the corresponding distortion level. The reverse water-filling interpretation is clearly evident from the shaded region in the figure.

ϑ is related to the average distortion through the slope of the rate distortion function at the point where the particular average distortion is achieved. This idea will prove important later when we work with composite source models.

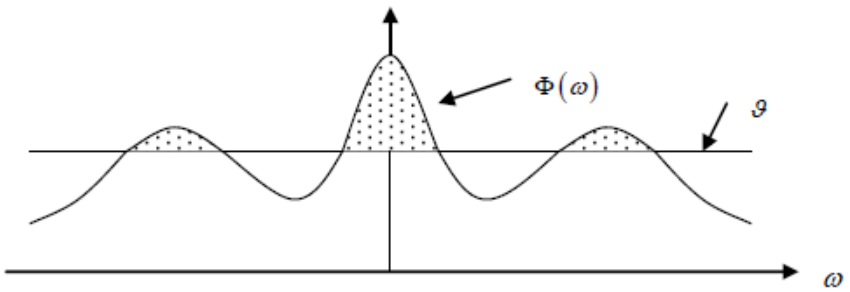


Figure 3.1: Example source, error, and reconstruction spectral densities

The importance of this theorem is that autoregressive sources have played a principal role in the design of leading narrowband and wideband voice codecs for decades. The rate distortion function in this theorem offers a direct connection to these codecs, although as we shall demonstrate in the following, one single source model will not do.

3.3 Composite Source Models

It was recognized early on that sources may have multiple modes and could switch between modes probabilistically, and such sources were called composite sources in the rate distortion theory literature [6]. In particular, a composite source is defined as source with probability depending on the side information Y [6, Sec. 6.1], as shown in Fig. 3.2. The choice of subsources is according to a probabilistic switch process, which is the side information. The power of composite sources derives from the individual subsources being able to capture local or finer dependence, while the switch process can represent changes that happen more globally and also model discontinuities. Given an appropriate number of carefully selected subsources and accurate switch modeling, time-varying or spatially-varying complex real world sources can be represented accurately.

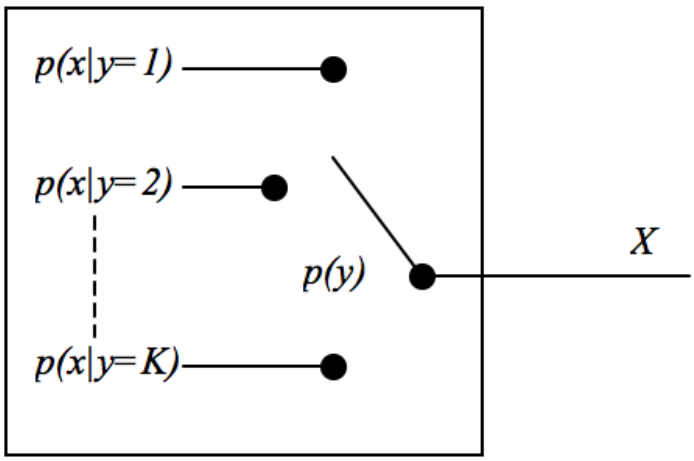


Figure 3.2: A Composite Source Model with K Subsources

Motivated by the work of Berger and others, research efforts explored, from the theoretical side, the properties of composite sources and also attempted to obtain expressions or bounds for the rate distortion functions of composite sources. In particular, Fontana [17] studied the stationarity, ergodicity, and mixing properties of composite sources

with stationary transitions and then went on to examine nonstationary composite sources. He also proved theorems characterizing the distortion rate function and the entropy rate of composite sources with slowly varying switch processes. Garde [19] turned to developing rate distortion bounds based on Fontana's work, wherein he considered symmetric Markov switch processes. Wallace [85] considered separable composite sources with memoryless subsources and Markov switch processes and found techniques for calculating the entropy rate. Among other things, Carter [8] studied the rate distortion theory of classes of composite sources, and he used conditional rate distortion theory as his primary tool to develop these bounds. For the class of regenerative composite sources, he derived upper and lower bounds to the rate distortion function of the composite source, and for the notion of an interrupted source, obtained an exact expression for the conditional rate distortion function.

Composite sources and conditional rate distortion functions place a key role in developing our voice and video rate distortion bounds in later chapters. We examine conditional rate distortion theory in more detail in the following section.

3.4 Conditional Rate Distortion Functions

The conditional rate distortion function is the rate of a source subject to a fidelity criterion when the encoder and decoder both have access to side information [27]. Thus, the conditional rate distortion function describes the rate required for a composite source subject to a fidelity criterion, where the side information is the switch process that selects the appropriate subsource at any time. We assume that all subsources have the same source alphabet and that all subsources are subject to the same distortion measure. The following definition of the conditional rate distortion function is from Gray [27].

Definition 3.1. The conditional rate distortion function of a source $\underline{X}$ with side information Y, which serves as the subsource information, is defined as

$$R_{\underline{X}|Y}(D) = \min_{p(\underline{\hat{x}}|\underline{x},y):D(\underline{X},\underline{\hat{X}}|Y)\leq D} I(\underline{X};\underline{\hat{X}}|Y), \tag{3.4.18}$$

where

$$D(\underline{X},\underline{\hat{X}}|Y) = \sum_{\underline{x},\underline{\hat{x}},y} p(\underline{x},\underline{\hat{x}},y) D(\underline{x},\underline{\hat{x}}|y),$$

$$I(\underline{X};\underline{\hat{X}}|Y) = \sum_{\underline{x},\underline{\hat{x}},y} p(\underline{x},\underline{\hat{x}},y) \log \frac{p(\underline{x},\underline{\hat{x}}|y)}{p(\underline{x}|y)p(\underline{\hat{x}}|y)}. \tag{3.4.19}$$

$R_{\underline{X}|Y}(D)$ is the lowest rate given that both encoder and decoder are allowed to observe perfectly the sequence Y. It can be proved [27] that the conditional rate distortion function in Eq. (3.4.18) can also be expressed as

$$R_{\underline{X}|Y}(D) = \min_{D_y's:D(\underline{X},\underline{\hat{X}}|Y)=\sum_y D_y p(y)\leq D} \sum_y R_{\underline{X}|y}(D_y)p(y), \tag{3.4.20}$$

and the minimum is achieved by adding up the individual, also called marginal, rate-distortion functions at points of equal slopes of the marginal rate distortion functions.

Utilizing the classical results for conditional rate distortion functions in Eq. (3.4.20), the minimum is achieved at D_y's where the slopes $\frac{\partial R_{\underline{X}|Y=y}(D_y)}{\partial D_y}$ are equal for all y and $\sum_y D_y P[Y=y] = D$.

This conditional rate distortion function $R_{\underline{X}|Y}(D)$ can be used to write the following inequality involving the overall source rate distortion function $R_{\underline{X}}(D)$ [27]

$$R_{\underline{X}|Y}(D) \leq R_{\underline{X}}(D) \leq R_{\underline{X}|Y}(D) + I(\underline{X};Y), \tag{3.4.21}$$

where $I(\underline{X};Y)$ is the average mutual information between $\underline{X}$ and Y and the equality in the leftmost inequality is achieved if and only if $\underline{X}$ and Y are independent. We can bound $I(\underline{X};Y)$ by $H(Y)$, entropy of the side information Y, which is further bounded by $\frac{1}{M}\log K$, as

$$I(\underline{X};Y) \leq H(Y) \leq \frac{1}{M}\log K, \tag{3.4.22}$$

where K is the number of subsources and M is the number of samples representing how often the subsources change in the speech utterance.

Since for voice, $K = 5$ here and M is on the order of 100 or more, the second term on the right in Eq. (3.4.21) is negligible, and the rate distortion for the source is very close to the conditional rate distortion function in Eq. (3.4.20). Therefore, we use the conditional rate distortion function $R_{\underline{X}|Y}(D)$ to develop our performance bounds for voice[56, 55].

3.5 Estimating Composite Source Model Parameters

With the prior rate distortion results and the general composite source model in our tool set, we see that to use these tools, we need good models for the subsources and a model for the switch process. While we develop these ideas in detail for voice and video sources in later chapters, we provide a high level overview of some prior approaches to obtain the subsource models and the switch process here.

Most prior work has assumed either Gaussian memoryless subsources or Gaussian autoregressive subsources and almost always a MSE distortion measure. The switch process is often assumed to be independent and identically distributed or deterministic but unknown. While these are idealized assumptions made primarily for analytical tractability reasons, it is often argued that the mixture process has a distribution that accurately models the actual source, even when it is not Gaussian. Plus, it is well known that the rate distortion function of Gaussian memoryless source upper bounds the rate distortion function of any other memoryless source with the same variance. Therefore, an upper bound to $R(D)$ is being obtained in the worst case.

Hence, it is the parameters of the Gaussian subsources and the switch process probabilities, or the points where the subsource switching takes place, that must be selected. Prior work has started with a joint maximum likelihood approach to estimate the needed parameters and switch process information, and then various assumptions are invoked to simplify the calculations. In particular, the switch process is chosen to be i.i.d., the sequence of subsources or frames are assumed independent, and the initial conditions are assumed to have minimal impact on the results. Some of the work has also assumed that the

structure of the subsources is identical and only the subsource parameters are changing. For example, all subsources may be assumed to be 20th order autoregressive processes [56, 55].

In these instances, the maximum likelihood estimates can be simplified substantially, and the source parameter estimation process often becomes that of estimating the parameters of an Nth order autoregressive process, which is identical to the methods used for the estimation of the parameters of the linear prediction model for speech coding introduced in the late 1960's.

When the rate distortion bounds are developed in Chapters 4 and 5, the prior work is examined in more detail.

4

Rate Distortion Bounds for Voice

We now turn our attention to developing rate distortion bounds specifically for voice sources. Of course, the two principal things that we have to do are to define adequate models of voice sources and to determine distortion measures that provide meaningful comparisons to existing voice codecs, and both need to be analytically tractable. We model the voice source using a composite source model, where we choose the subsources based upon phonetic speech models that have been useful in voice codec designs, and use conditional rate distortion theory for the MSE distortion measure, as discussed in Section 4.2. To obtain a distortion measure that is analytically tractable, we could develop bounds based upon weighted MSE distortion measures, but this approach has its own challenges, as we elaborate later. Our chosen approach is to devise a mapping function to map the MSE based rate distortion results into rate versus PESQ-MOS distortion results. To accomplish this, consider waveform-following coders, for which MSE produces a reasonable ordering of rate versus distortion results, and for which PESQ-MOS values can be obtained. Considerations in selecting the codecs and justification of this approach are presented in Section 4.4. Our rate distortion bounds for both narrowband and wideband voice and comparisons to known standardized voice codecs are given in Section 4.5.

There has been some limited prior work on developing rate distortion bounds for speech, and we begin the chapter by surveying this work in Section 4.1.

4.1 Related Prior Work

There have been only a small number of prior research efforts in the last 25 years that have attempted to develop rate distortion bounds for speech. Most have used the MSE distortion measure, and a few have been based on using subsources or on explicitly using composite source models. We discuss these contributions in this section.

Brehm and Trottler [29] utilize the mean squared error (MSE) distortion measure and focus their efforts on modeling the speech source for narrowband speech. They utilize spherically invariant random process (SIRP) models that allow the inclusion of correlation in the source probability density function (pdf) and then note that with the autocorrelation function and the first order pdf known, then the first order pdf and all higher order pdfs can be expressed in terms of G-functions, which are a class of higher-transcendental functions. The first order pdf based on the G-function is then fit to speech data from one male speaker and shown to be a good fit to the first order histogram of the data. They then characterize the SIRP speech model as a decomposition of Gaussian subsources with a variable standard deviation, where the pdf of the standard deviation is expressed in terms of a G-function. The overall rate distortion bound is obtained by averaging the rate distortion functions of the subsources at points of equal slope. The rate distortion bounds actually calculated and presented in the paper, however, only use the initial first order pdf fit to the experimental data. Their rate distortion bounds are only for the MSE distortion measure, thus limiting their applicability to CELP codec performance comparisons.

A composite source model is a collection of subsources accessed by a probabilistic switching process, as illustrated in Fig. 3.2. In [56, 55], composite source models for speech are obtained by segmentation of the speech into equal order (20th order) Gaussian autoregressive sub-

sources. Each subsource is parametrized by the predictor coefficients and the residual variance which are estimated by maximum likelihood estimation, assuming i.i.d. subsources and an i.i.d. switch process. The rate distortion functions of composite sources are calculated using conditional rate distortion functions for the MSE distortion measure. In their experiments, they calculated lower bounds to the rate distortion function for different numbers of subsources, and showed that a relatively small number of subsources (6 in the cited paper) is needed to have a good composite source model for speech. Only one male speaker is considered in this work, and it is what is categorized as wideband speech since the input band is 30 Hz to 7 kHz and the sampling rate is 16,000 samples per second. The work also utilizes a Hamming window of length 30 msec., and since the analysis windows are only 10 msec., there is considerable overlap between adjacent windows. This windowing is done to improve the autoregressive subsource parameter estimation. No comparisons to standardized speech codecs are provided since MSE is not a meaningful distortion measure for these codecs.

A cochlear model serves as the basis for a perceptual distortion measure for speech in [16], and the speech source model is merged with the cochlear models and used to characterize the rate distortion function for speech. With the cochlear variational distance as the distortion measure, a lower bound to the rate distortion function is calculated , and Blahut's algorithm is applied for the direct evaluation of the rate distortion function with the cochlear directed divergence and variational distance. Four speech coders were compared with the rate distortion bound generated by Blahut's algorithm. Among the interesting results are that the Shannon lower bound for the cochlear variational distance distortion measure is only tight at very small distortions and that the voice codecs evaluated required more than twice the minimum rate to achieve the same distortion. This work emphasizes the distortion measure, and de-emphasizes the source modeling aspect. The drawback of the approach is that the method is not easily employed by non-skilled users to compare to their codec performance.

Gibson, Hu, and Ramadas [21] obtained rate distortion bounds for speech coding based on composite source models and unweighted and

weighted MSE distortion measures. The composite source models are constructed by classifying each sentence as Voiced (V), Unvoiced (UV), Onset (ON), Hangover (H), and Silence (S) by hand. The V, ON, and H modes are modeled as autoregressive with different orders, and the UV mode is modeled as uncorrelated. The marginal and conditional rate distortion bounds for two English sequences were shown, and the operational rate distortion performance of the waveform following codec, G.727, was compared with the rate distortion bounds based on unweighted MSE. Since the G.727 codec does not use voice activity detection to efficiently model and compress silence and since it is a relatively high rate codec, it has performance well above the calculated rate distortion bound. The particular weighted MSE considered in this paper was based on an average weighting across the entire utterance and hence was not able to meaningfully capture perceptual distortion. The performance of important CELP based codecs were far above the calculated $R(D)$ bound for the weighted MSE measure, and therefore, the rate distortion curves based on this weighted MSE criterion are not useful. Further research is needed to determine if varying the weighting for different subsource modes or much more often, say every 10 msec or so might yield better results. These $R(D)$ curves are presented in a later section after composite source models are more fully elaborated and as a step toward developing the final $R(D)$ bounds that are valid for all speech codecs.

4.2 Composite Source Models for Speech

It was recognized early on in rate distortion theory that sources may have multiple modes and can switch between modes probabilistically, and as we have seen, such sources were called composite sources in the rate distortion theory literature [6]. Prior work on rate distortion bounds for speech coding, as discussed in Section 4.1, have utilized different types of composite source models to provide good models for speech signals. We also rely on composite source models for our work, but we construct these models in a more straightforward way than prior authors by drawing on prior research on speech codec design. We also

allow a greater diversity of subsource models.

Multimodal models have played a major role in speech coding, including the voiced/unvoiced decision for the excitation in linear predictive coding (LPC) [3] and the long-term adaptive predictor in adaptive predictive coding (APC) [4]. Further, phonetic classification of the input speech into multiple modes and coding each mode differently has lead to some outstanding voice codec designs [87, 86]. We build on the phonetic classification methods in these successful codec designs to surmise useful composite source models.

In particular, the work of Ramadas and Gibson [73] on speech coding has been guided by these prior contributions, and we have developed a mode classification method that breaks the input speech into Voiced (V), Onset (ON), Hangover (H), Unvoiced (UV), and Silence (S) modes, each of which may be coded at a different rate. We use these modes to develop a composite source model for speech here.

For narrowband speech, we model Voiced (V) speech as a 10^{th} order AR Gaussian source since most narrowband speech codecs, such as AMR-NB, use 10^{th} order linear prediction in the codec. Onset (ON) is modeled as a 4^{th} order AR Gaussian source, Hangover (H) is modeled as a 4^{th} order AR Gaussian source, Unvoiced (UV) speech is modeled as a memoryless Gaussian source, and Silence (S) is treated by sending a code for comfort noise generation. In addition to the five-mode (V, UV, H, ON, and S) composite source models, we also try two modes (V and S) and one mode (V) as a source model for comparison.

Table 4.1 presents the autocorrelation values and mean-squared prediction error for five narrowband English sentences. There is no mode classification on each sequence in this table, so whole sequences are treated as Voiced. This represents the case where the source is purely autoregressive and is a good fit to the classical rate distortion approach for Gaussian autoregressive sources in the rate distortion theory literature [6]. Table 4.2presents the autocorrelation values, mean-squared prediction error for the two modes, and the probability of non-silence and silence for five narrowband English sentences. The two mode classifications fit the common classifications used in early linear prediction voice codecs.

Table 4.1: Autocorrelation coefficients and Mean Square Prediction Error for Narrowband Speech Sentences

Sequence	Autocorrelation coefficients	Mean Square Prediction Error
T07 (Female) (active speech level: −25.0 dBov) (sampling rate: 8 kHz)	[1 0.8833 0.6985 0.5154 0.3357 0.2025 0.1015 0.0195 −0.0381 −0.0833 −0.1290]	0.0326
T08 (Female) (active speech level: −15.6 dBov) (sampling rate: 8 kHz)	[1 0.7975 0.4655 0.2195 0.0618 −0.0137 −0.0250 0.0109 0.0579 0.0472 −0.0289]	0.0679
T13 (Female) (active speech level: −24.8 dBov) (sampling rate: 8 kHz)	[1 0.8018 0.5018 0.2526 0.0540 −0.0462 −0.0942 −0.1480 −0.1946 −0.1748 −0.1293]	0.0780
"lathe" (Female) (active speech level: −18.1 dBov) (sampling rate: 8 kHz)	[1 0.8076 0.5507 0.3444 0.1470 0.0221 −0.0521 −0.0745 −0.0878 −0.1441 −0.2321]	0.0813
"we were away" (Male) (active speech level: −16.5 dBov) (sampling rate: 8 kHz)	[1 0.8014 0.5176 0.2647 0.0432 −0.1313 −0.2203 −0.3193 −0.3934 −0.4026 −0.3628]	0.0780

The five-mode composite source models for five narrowband English sentences are shown in Table 4.3. The five-mode composite source model has seen some applications in voice codec design but it is not a dominant paradigm. For accurate source modeling, however, it is known that more than two subsources are necessary to capture accurately the source behavior [55]. The selection of the subsource models is based on experience with both speech models and on the availability of existing rate distortion results for those source models.

There are a few things to note about the data in Tables 4.2 and 4.3. First, the average frame energy for the UV mode and the mean-squared prediction errors for the other modes are normalized to the average energy over the entire sentence since the MSE of the mapping function is normalized by the average energy. Second, the sentence, "We were away" is only 1.05% classified as Silence. T07 is 35.81%, T08 is 33.09%, T13 is 35.66%, and "lathe" is 36.85% classified as Silence. These Silence segments are assumed to be transmitted using a fixed length code to represent the length of the Silence intervals and to represent comfort

Table 4.2: Silence or Non-silence Source Models for Narrowband Speech Sentences

Sequence	Mode	Autocorrelation coefficients for V	Mean Square Prediction Error	Probability
T07 (Female) (active speech level: −25.0 dBov) (sampling rate: 8 kHz)	V	[1 0.8833 0.6985 0.5154 0.3357 0.2025 0.1014 0.0194 −0.0382 −0.0833 −0.1291]	0.0325	0.6419
	S			0.3581
T08 (Female) (active speech level: −15.6 dBov) (sampling rate: 8 kHz)	V	[1 0.7975 0.4655 0.2195 0.0618 −0.0137 −0.0250 0.0109 0.0579 0.0472 −0.0289]	0.0679	0.6691
	S			0.3
T13 (Female) (active speech level: −24.8 dBov) (sampling rate: 8 kHz)	V	[1 0.8018 0.5018 0.2525 0.0540 −0.0463 −0.0942 −0.1480 −0.1946 −0.1748 −0.1293]	0.0780	0.6435
	S			0.3565
"lathe" (Female) (active speech level: −18.1 dBov) (sampling rate: 8 kHz)	V	[1 0.8076 0.5507 0.3444 0.1470 0.0221 −0.0521 −0.0745 −0.0878 −0.1441 −0.2321]	0.0813	0.6315
	S			0.3685
"we were away" (Male) (active speech level: −16.5 dBov) (sampling rate: 8 kHz)	V	[1 0.8014 0.5176 0.2647 0.0432 −0.1313 −0.2203 −0.3193 −0.3934 −0.4026 −0.3628]	0.0780	0.9895
	S			0.0105

noise inserted in the decoded stream.

AMR-WB [42] uses 16^{th} order linear prediction at 12.8 kHz sampling rates for wideband speech. Hence, we also model Voiced speech as a 16^{th} order AR Gaussian source at a 12.8 kHz. In order to model the wideband speech source at 12.8 kHz sampling rate, we down-sample the wideband speech from 16 kHz to 12.8 kHz using the decimation filter in AMR-WB for wideband speech. While Voiced speech is modeled as 16^{th} order AR Gaussian sources, Onset and Hangover are modeled as 4^{th} order AR Gaussian sources. Unvoiced speech is modeled as a memoryless Gaussian source, and Silence is treated by sending a code for comfort noise generation at 12.8 kHz sampling rates.

In addition to the five-mode (V, UV, H, ON, and S) composite source models, we also try two modes (V or S) and one mode (V) as source model. In particular, Table 4.4 presents the autocorrelation values and mean-squared prediction error for two wideband English sentences and one wideband Japanese sentence. There is no mode classification on each sequence. Whole sequences are treated as Voiced mode. Table 4.5 presents the autocorrelation values, mean-squared prediction

Table 4.3: Composite Source Models for Narrowband Speech Sentences

Sequence	Mode	Autocorrelation coefficients for V, ON, H Average frame energy for UV	Mean Square Prediction Error	Probability
T07 (Female) (active speech level: −25.0 dBov) (sampling rate: 8 kHz)	V	[1 0.8850 0.6999 0.5151 0.3351 0.2014 0.1003 0.0181 −0.0394 −0.0844 −0.1298]	0.0310	0.4876
	ON	[1 0.8233 0.6751 0.5848 0.4221]	0.0883	0.0193
	H	[1 0.9282 0.8579 0.8067 0.7503]	0.0186	0.0138
	UV	0.0159	0.0159	0.1212
	S			0.3581
T08 (Female) (active speech level: −15.6 dBov) (sampling rate: 8 kHz)	V	[1 0.7975 0.4647 0.2184 0.0608 −0.0145 −0.0253 0.0109 0.0580 0.0471 −0.0293]	0.0675	0.4654
	ON	[1 0.9209 0.8119 0.7002 0.5630]	0.0181	0.0132
	H	[1 0.9211 0.8622 0.8105 0.7556]	0.0225	0.0074
	UV	0.0142	0.0142	0.1831
	S			0.3309
T13 (Male) (active speech level: −24.8 dBov) (sampling rate: 8 kHz)	V	[1 0.8024 0.5015 0.2517 0.0528 −0.0474 −0.0952 −0.1489 −0.1953 −0.1757 −0.1295]	0.0767	0.4393
	ON	[1 0.8588 0.7319 0.6532 0.5077]	0.0561	0.0118
	H	[1 0.9099 0.7933 0.6992 0.6416]	0.0260	0.0074
	UV	0.0060	0.0060	0.1849
	S			0.3566
"lathe" (Female) (active speech level: −18.1 dBov) (sampling rate: 8 kHz)	V	[1 0.8217 0.5592 0.3435 0.1498 0.0200 −0.0517 −0.0732 −0.0912 −0.1471 −0.2340]	0.0656	0.5265
	ON	[1 0.8495 0.5962 0.3979 0.2518]	0.0432	0.0093
	H	[1 0.2709 0.2808 0.1576 0.1182]	0.7714	0.0186
	UV	0.1439	0.1439	0.0771
	S			0.3685
"we were away" (Male) (active speech level: −16.5 dBov) (sampling rate: 8 kHz)	V	[1 0.8014 0.5176 0.2647 0.0432 −0.1313 −0.2203 −0.3193 −0.3934 −0.4026 −0.3628]	0.0780	0.9842
	ON	[1 0.8591 0.7215 0.6128 0.5183]	0.0680	0.0053
	H			0
	UV			0
	S			0.0105

error for the two modes (V or S), and the probability of non-silence (V) and silence for three wideband sentences. The five-mode composite source models for three wideband sentences are shown in Table 4.6. The average frame energy for the UV mode and the mean-squared prediction errors for the other modes are normalized to the average energy over the entire sentence since the MSE of the mapping function is normalized by the average energy. F1 has 55.23%, M3 has 24.67%, and F2 has 30.53% classified as Silence. These Silence segments are assumed to be transmitted using a fixed length code to represent the length of the Silence intervals and to represent comfort noise to be inserted into the decoded stream.

Examining all of the tables of composite source models, it is evident that each of these sentences has quite different characteristics, and this,

Table 4.4: Autocorrelation coefficients and Mean Square Prediction Error for Wideband Speech Sentences

Sequence	Autocorrelation coefficients	Mean Square Prediction Error
F1 (Female) (active speech level: −25.968 dBov) (sampling rate: 12.8 kHz)	[1 0.8460 0.5931 0.4187 0.3214 0.2722 0.2169 0.1507 0.0645 −0.0929 −0.2951 −0.4027 −0.3753 −0.3034 −0.2629 −0.2833 −0.3239]	0.0257
M3 (Male) (active speech level: −29.654 dBov) (sampling rate: 12.8 kHz)	[1 0.8005 0.6683 0.4887 0.3019 0.2538 0.2127 0.2252 0.2262 0.2261 0.2011 0.1586 0.1282 0.1031 0.1281 0.1688 0.1595]	0.0836
F2 (Female) (active speech level: −26.009 dBov) (sampling rate: 12.8 kHz) (Japanese)	[1 0.9274 0.8251 0.6843 0.5071 0.3616 0.2176 0.1010 0.0119 −0.0655 −0.1265 −0.1897 −0.2423 −0.2746 −0.3007 −0.3078 −0.2957]	0.0086

Table 4.5: Silence or Non-silence Source Models for Wideband Speech Sentences

Sequence	Mode	Autocorrelation coefficients for V	Mean Square Prediction Error	Probability
F1 (Female) (active speech level: −25.968 dBov) (sampling rate: 12.8 kHz)	V	[1 0.8448 0.5891 0.4132 0.3156 0.2670 0.2122 0.1462 0.0599 −0.0987 −0.3028 −0.4109 −0.3816 −0.3084 −0.2673 −0.2879 −0.3293]	0.0253	0.4761
	S			0.5523
M3 (Male) (active speech level: −29.654 dBov) (sampling rate: 12.8 kHz)	V	[1 0.7954 0.6612 0.4775 0.2864 0.2398 0.2005 0.2169 0.2215 0.2248 0.2023 0.1614 0.1333 0.1076 0.1334 0.1759 0.1662]	0.0861	0.7533
	S			0.2467
F2 (Female) (active speech level: −26.009 dBov) (sampling rate: 12.8 kHz) (Japanese)	V	[1 0.9275 0.8249 0.6835 0.5054 0.3592 0.2146 0.0976 0.0085 −0.0688 −0.1296 −0.1926 −0.2449 −0.2775 −0.3039 −0.3110 −0.2988]	0.0084	0.6947
	S			0.3053

in turn, hints at what we will see later – namely that the these sentences have distinctively different rate distortion functions. Indeed, the subsources in each sentence have different $R(D)$ curves and their probabilities of occurrence are different across sentences as well. These facts highlight one of the weaknesses of much earlier work on $R(D)$ bounds for speech (and video) that used average source models, averaged over all samples in each source and over multiple utterances, to obtain $R(D)$ curves. In the end, the resulting $R(D)$ curves did not lower bound the performance of codecs on many individual utterances. The key lesson is an old one, and one that we have been emphasizing – to obtain valid $R(D)$ functions, the models must be accurate.

Table 4.6: Composite Source Models for Wideband Speech Sentences

Sequence	Mode	Autocorrelation coefficients for V, ON, H Average frame energy for UV	Mean Square Prediction Error	Probability
F1 (Female) (active speech level: −25.968 dBov) (sampling rate: 12.8 kHz)	V	[1 0.8448 0.5891 0.4132 0.3156 0.2670 0.2122 0.1462 0.0599 −0.0987 −0.3028 −0.4109 −0.3816 −0.3084 −0.2673 −0.2879 −0.3293]	0.0253	0.4406
	ON	[1 0.1226 −0.2917 0.2239 −0.0034]	0.5241	0.0043
	H			0
	UV	0.0009	0.0009	0.0028
	S			0.5523
M3 (Male) (active speech level: −29.654 dBov) (sampling rate: 12.8 kHz)	V	[1 0.7954 0.6612 0.4775 0.2864 0.2398 0.2004 0.2169 0.2214 0.2248 0.2022 0.1613 0.1333 0.1075 0.1334 0.1759 0.1662]	0.0861	0.6939
	ON	[1 0.9564 0.9334 0.9104 0.8862]	0.0066	0.0069
	H	[1 0.9387 0.9028 0.8696 0.8257]	0.0129	0.0461
	UV	0.0015	0.0015	0.0064
	S			0.2467
F2 (Female) (active speech level: −26.009 dBov) (sampling rate: 12.8 kHz) (Japanese)	V	[1 0.9285 0.8251 0.6841 0.5056 0.3593 0.2148 0.0975 0.0087 −0.0690 −0.1296 −0.1928 −0.2450 −0.2777 −0.3040 −0.3112 −0.2990]	0.0079	0.6539
	ON	[1 −0.8659 0.6094 −0.2720 −0.0584]	0.0212	0.0056
	H	[1 0.9606 0.9140 0.8444 0.7646]	0.0045	0.0281
	UV	0.0610	0.0610	0.0070
	S			0.3053

Using the composite models just presented, the next section begins the development of our $R(D)$ bounds using conditional rate distortion theory.

4.3 Marginal and Conditional Rate Distortion Bounds based on MSE Distortion Measure

Given the source models in the tables, conditional rate distortion theory and reverse water-filling as outlined in Chapter 3, can be used to calculate rate distortion functions for each of the sentences for the MSE distortion measure. The specific steps are to use the Karhunen-Loeve decomposition outlined there for each subsource and then use reverse water-filling on each of the resulting subsource models. Then, the subsource rate distortion bounds are combined at points of equal slope applying the weighting indicated by the subsource probabilities.

The resulting marginal and conditional MSE rate distortion bounds of the composite source models for two narrowband English sequences are shown in Figs. 4.1 and 4.2, and results for two English wideband sentences are shown in Figs. 4.3 and 4.4. It is interesting, but perhaps not surprising, to see that for each sentence, the several subsources

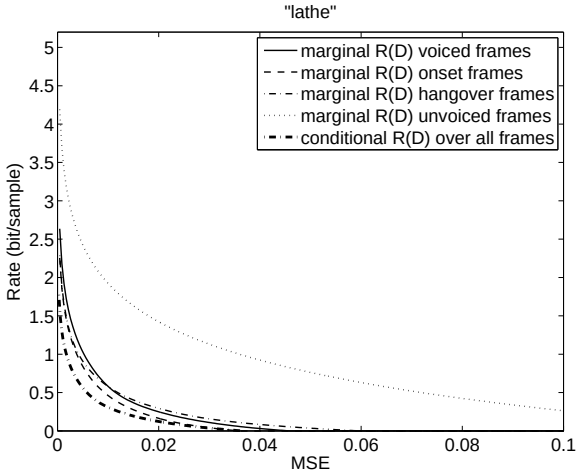


Figure 4.1: The MSE rate distortion bounds of narrowband sequence "A lathe is a big tool. Grab every dish of sugar."

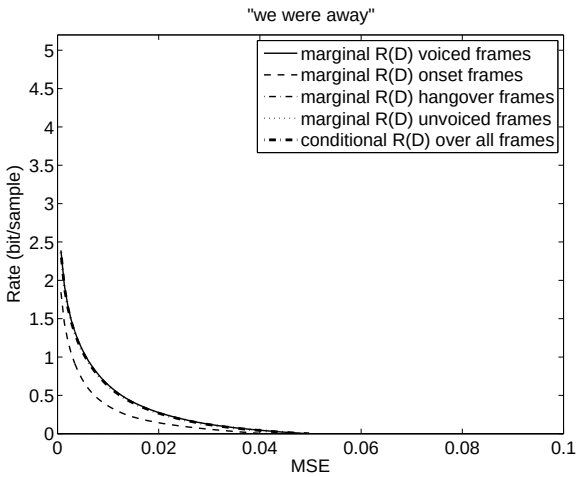


Figure 4.2: The MSE rate distortion bounds of narrowband sequence "We were away a year ago."

(modes) have different rate distortion functions; furthermore, the rate distortion functions for the subsources differ across the four sentences, since the model of each subsource is different for each sentence.

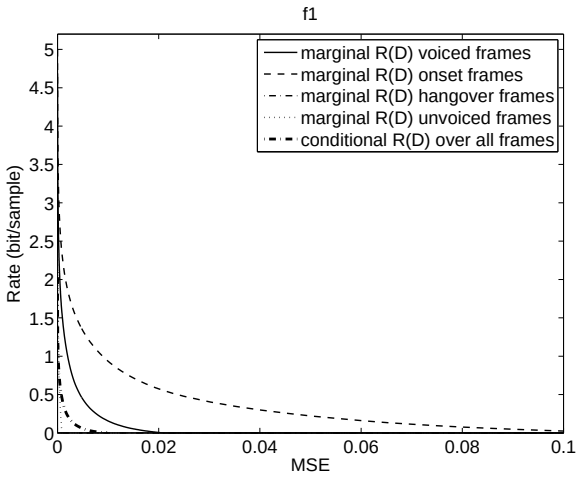


Figure 4.3: The MSE rate distortion bounds of wideband sequence F1, "You must go and do it at once. There were several small outhouses."

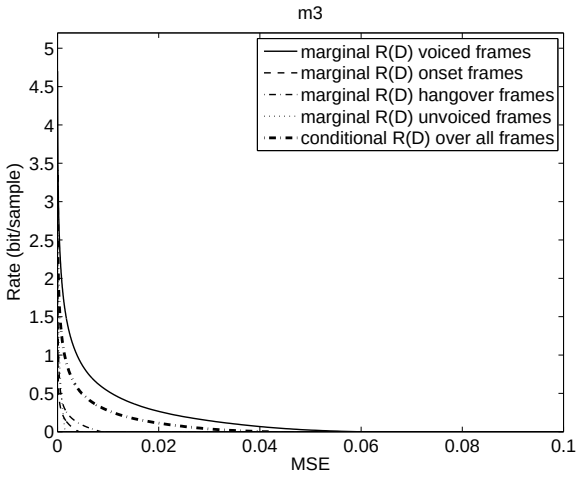


Figure 4.4: The MSE rate distortion bounds of wideband sequence M3, "I don't know, the vampire said, and he smiled."

Another important point is that the probabilities of the different subsources have a very profound effect. A speech sequence with considerably more voiced or unvoiced segments would weight the marginal

rate distortion functions differently and thus produce a quite different conditional rate distortion bound even for exactly the same subsources. This implies that the rate distortion bounds based on speech models obtained by using average autocorrelation functions over many sequences will not be very useful if the average results are interpreted as bounds for a more restrictive subset of the source models.

In Fig. 4.2, since the sequence is 98.42% Voiced, the conditional rate distortion function is dominated by the marginal rate distortion function of the voiced mode. In Figs. 4.1, 4.3, and 4.4, since each sequence has at least 24% Silence, the final conditional rate distortion functions are lower than the marginal rate distortion functions of Voiced frames.

While the observations above are significant in terms of the source models and indicate that the composite source model approach advocated here is a viable way forward, the MSE distortion measure prohibits any comparisons to voice codecs for which MSE is not a reasonable performance indicator. This precludes useful comparisons to CELP based codecs, for example, which are the dominant narrowband and wideband voice codec structures at the time of this writing. We therefore consider a mapping approach for the average distortion as discussed in the following subsection.

4.4 Mapping MSE to PESQ-MOS/WPESQ

The rate distortion theory results reported in Chapter 3 are built on the assumption of the mean squared error distortion measure, which unfortunately, is not a reliable or widely used indicator of speech codec performance. One approach would be to use an analytically tractable weighted MSE criterion, and such an approach with an average weighting over the entire utterance was reported in Gibson, Hu, and Ramadas [21]. Furthermore, it may be possible to change the weighting throughout the utterance adaptively to obtain rate distortion bounds. However, weighting functions for the squared error distortion measure that have a strong connection to the perceptual quality achieved by state-of-the-art voice codecs are not known at this time.

Alternatively, it is well known that PESQ-MOS and WPESQ are

standardized objective methods for narrowband and wideband speech quality assessment, and both are widely used in categorizing the perceptual performance of standardized speech codecs. Hence, if PESQ-MOS and WPESQ could be used as the distortion measure in the $R(D)$ calculations, more valid curves should be obtained. Therefore, in order to extend the utility of the prior theoretical rate distortion theory results using the MSE distortion measure, we developed a procedure for mapping MSE into PESQ-MOS or WPESQ, as is respectively appropriate for the bandwidth of interest.

The basic approach to generating the mapping functions is to first note that MSE is a reasonable performance indicator for waveform coders in that MSE correctly orders the perceptual performance of these waveform codecs, although the difference in MSE might not be a good indicator of the exact perceptual quality difference. Since PESQ-MOS can be obtained for these same codecs as well, we have codec performance for both distortion measures, and calculating an appropriate mapping function would appear possible. However, developing such a mapping is not straightforward and a number of constraints need to be imposed to make the mapping meaningful.

In particular, the mapping of MSE to PESQ-MOS must be performed with several key points in mind. First, the mapping must be done for a codec for which MSE is a valid performance indicator and to which PESQ-MOS can be applied. Second, the codec must be a predictive coder since it is well known that MSE for predictive coders and for non-predictive coding have different correspondences with subjective performance. Third, the existing theoretical $R(D)$ results assume that the source model parameters are known exactly at the decoder, and therefore do not include bit rate, and the associated distortion, for representation of the source model parameters. Therefore the codec used for the mapping should not transmit side information for the parameters either; however, the effect of the unknown source parameters on bit rate and average distortion must be incorporated in some fashion. In order to meet this constraint, we chose backward adaptive waveform coders to generate the MSE/PESQ-MOS pairs for the mapping. Fourth, the codecs used for the mapping must have a range of bit rates suffi-

cient to generate the mapping over the bit rates of interest. Fifth, the mapping function must be convex $\cup$ in order to maintain the relative order of the MSE values and PESQ-MOS values. Sixth, the mapping must be matched to each individual utterance to be evaluated.

Another critical consideration is the active speech level of the test sequence. For the PESQ, we need to avoid peak clipping (mentioned in P.862.3), and therefore, the active speech level should not be too high. Further, if the energy of the speech utterance is too low, the MSE will blow up. The active speech level of test sequences we use is between -15 dBov and -30 dBov and thus satisfy both requirements.

In light of these key constraints, for narrowband speech, we focus on the particular class of predictive waveform coders represented by the G.726 and G.727 standards. It is known that MSE orders the performance of these codecs accurately, while PESQ-MOS values can be obtained for these codecs in order to get a meaningful perceptual performance indicator. For wideband MSE to WPESQ mapping, ADPCM speech coders for wideband speech are needed. The standard, G.722, is a wideband speech coder based on ADPCM and for which MSE (SNR) has been used as a performance indicator in the past, however, there are only three coding rates. As a result, there are not enough rate/distortion points to develop a good mapping. Therefore, we developed our own ADPCM coder for wideband speech based on G.722 and G.727 to generate the mapping function for wideband speech. The details of the wideband ADPCM coder are described in Section 4.4.2. We first describe the PESQ-MOS and WPESQ standards since they are widely employed for speech codec performance evaluation. This information also plays a role in the development of the mapping function.

4.4.1 PESQ-MOS/WPESQ

Perceptual evaluation of speech quality (PESQ) [48] is an objective, full reference method for end-to-end speech quality assessment of narrowband speech codecs. Full reference means that the original utterance to be coded is available, and the distance between the original and degraded speech signal, called the PESQ score, is calculated based on the PESQ perceptual model. The PESQ score is then mapped to a

MOS-like scale by a monotonic function. The MOS-like PESQ (PESQ-MOS) is a single number in the range of -0.5 and 4.5, although for most cases the output range will be between 1.0 and 4.5, which is the normal range of MOS values found in an Absolute Category Rating (ACR) experiment with human listeners.

Even though PESQ-MOS is not the same as MOS, and it has known limitations, it is a standardized objective measure for evaluating the perceptual performance of speech codecs that is widely used and quoted. WPESQ is an extension to PESQ for wideband telephone networks and speech codecs.

The wideband extension is mapped from the raw scores provided by the P.862 model. The details of WPESQ are described in the ITU-T P.862.2 Recommendation [50].

4.4.2 ADPCM Speech Coders

ADPCM coders are waveform coders, that is, they attempt to follow the time-domain waveform. As a result, MSE is an indicator of how well the codec is reproducing the input speech signal. MSE (SNR) is also useful in establishing the relative ordering of the performance of ADPCM speech coders [53]. In addition, the PESQ-MOS/WPESQ of ADPCM coders can be generated, thus providing a perceptual distortion value that corresponds with the MSE achieved by the codec at the given rate for the selected input utterance.

Even though both MSE and PESQ-MOS of other waveform coders, such as linear PCM and log-PCM, can be computed as well, we focus on backward adaptive ADPCM coders which use backward adaptive prediction, since we are interested in applying the resulting mapping functions to a broader class of predictive coders, such as CELP codecs. Another reason it is important to use the ADPCM codecs is that prior work has shown that the SNR or MSE of this class of predictive coders corresponds to a better perceptual preference than the equivalent SNR for a nonpredictive codec such as PCM. This phenomenon appears to be due to the quantization error being correlated with the speech signal, and thus the higher energy error is less objectionable perceptually.

G.726/G.727 Narrowband ADPCM Speech Coders

G.726 [43, 37] and G.727 [37, 44] are standardized narrowband ADPCM speech coders. Both use backward adaptive prediction and backward adaptive quantization, so that the coded residual error signal is all that is needed to reconstruct the speech. Both of these codecs have four selectable transmitted bit rates of 40, 32, 24, and 16 kbps.

Since G.727 is an embedded coder, it has enhancement and core bits, and the transmitted bit rate can be reduced up to the number of bits per sample indicated by the core bits. It is important to know the full transmitted bit rate as well as the minimum rate, so G.727 is often referred to by using (x, y) pairs, where x refers to the total of both enhancement and core bits, which sets the transmitted bit rate, and y refers to the number of core bits used in the predictor coefficient adaptation loop.

The full rate can be pruned to y bits/sample, so ITU-T G.727 Recommendation [44] provides coding rates of 40 kbps for the 3 combinations $(5, 4)$,$(5, 3)$, and $(5, 2)$, 32 kbps for 3 combinations $(4, 4)$, $(4, 3)$, and $(4, 2)$, 24 kbps for 2 combinations $(3, 3)$ and $(3, 2)$, and 16 kbps for one combination $(2, 2)$, resulting in 9 pairs of coding rates. Therefore, with the 4 coding rates for non-embedded G.726 and the 9 coding rates for G.727, we have 13 MSE and PESQ pairs to generate a mapping function for each narrowband sentence.

The development of the mapping function is presented after the wideband codecs used for generating the wideband mapping are discussed.

Wideband ADPCM Speech Coders

G.722 [40] is a well-established, standardized wideband ADPCM speech coder for which both MSE values and WPESQ scores can be obtained. However, G.722 has only three bit rates and so there are only three MSE/WPESQ pairs that can be generated by G.722. Moreover, the average WPESQ of the lowest bit-rate, 48 kbps, of G.722 is greater than 3.0, and the average WPESQ of the highest bit-rate of G.722, 64 kbps, is about 4.0. Thus, based on such a small amount of data

whose range is much smaller than the mapping range, we cannot get a reasonable range of MOS values and therefore getting a good curve fitting result is not possible.

To obtain additional rate/distortion pairs, we created a new wideband ADPCM speech coder based on G.722 and G.727. The frequency band of the wideband signal is split into two sub-bands (higher and lower) by using the quadrature mirror filters from G.722. The upper sub-band still uses the coding method used in the upper sub-band of G.722. For the lower sub-band, we use G.726 and G.727 as the lower sub-band ADPCM coders. Since there are 9 coding rates for G.727 as discussed in 4.4.2 and 4 for G.726, we have 13 MSE and WPESQ pairs to generate a mapping function for each wideband sentence.

In this way, we generate a mapping function for each wideband sentence. In addition, the range of WPESQ generated is from 1.8–3.9, which is much wider than using G.722 only.

4.4.3 Mapping Function

In this section, we outline the specific process used to generate the mapping functions for narrowband and wideband speech sources. This mapping function is then applied to map the theoretical rate distortion curves for the MSE distortion measure to rate distortion performance curves versus a PESQ/WPESQ-MOS distortion measure. For each speech sentence (sequence), we calculate the MSE of each coded sequence and normalize the MSE by the average energy of the original sequence. The PESQ-MOS/WPESQ of each coded sequence is evaluated by the software provided by ITU-T Recommendation P.862/P.862.2 [48, 50].

As mentioned in Section 4.4.2, there are 13 pairs of MSE and PESQ that we use for curve fitting for each narrowband sequence, and 13 MSE/WPESQ pairs for each wideband sequence. Since MSE is increasing and PESQ/WPESQ is decreasing as the bit rate is reduced, two candidate mapping functions are considered; namely, the inverse function $z = \frac{a}{w} + b$, and the exponential function $z = ae^{-bw} + c$, where w is MSE and z is PESQ-MOS/WPESQ.

We chose the exponential function to perform the curve fitting since

it provides a better fit across all rates and distortion pairs. The range of PESQ-MOS/WPESQ is between -0.5 and 4.5 [48], so we set the PESQ-MOS/WPESQ to 4.5 when MSE is 0, and we forced $f(0) = 4.5$. Therefore, the explicit mapping function is modeled as

$$z = f(w) = ae^{-bw} + 4.5 - a, \tag{4.4.1}$$

where a and b are estimated by the least squares fit of the MSE and PESQ/WPESQ pairs of the ADPCM waveform codecs.

Several clean English sequences are used to illustrate the results of designing the mapping functions for both narrowband and wideband sequences. There is a different mapping function for each sentence, since it is well known that speech codec performance in terms of both MSE and particularly PESQ-MOS/WPESQ are highly source dependent. The active speech level of each sequence is computed based on ITU-T P.56 [37, 46]. ITU-T Recommendation P.830 [47] mentions that the nominal value for mean active speech level is -26 dBov, and that the active speech level should be observed during recording. In addition, ITU-T Recommendation P.862.3 [51] recommends that the active speech level of reference speech files and degraded signals should be stored around -30 dBov to avoid clipping. Therefore, we only used sequences with active speech level greater than -30 dBov, and we recommend that our approach to developing mapping functions not be used on low energy sequences. The active speech level of each sequence is also listed in Table 4.3 for narrowband speech and in Table 4.6 for wideband speech.

The mapping functions of the five narrowband sequences are shown in Figures 4.5 through 4.9, while the mapping functions of the three wideband sequences are shown in Figures 4.10–4.12. The results show that the exponential function provides a good fit to the MSE-and-WPESQ pairs.

Later, after rate distortion functions for both narrowband and wideband speech have been presented, we discuss how varying the fit of the mapping to the points can be used to study the tightness of the bounds.

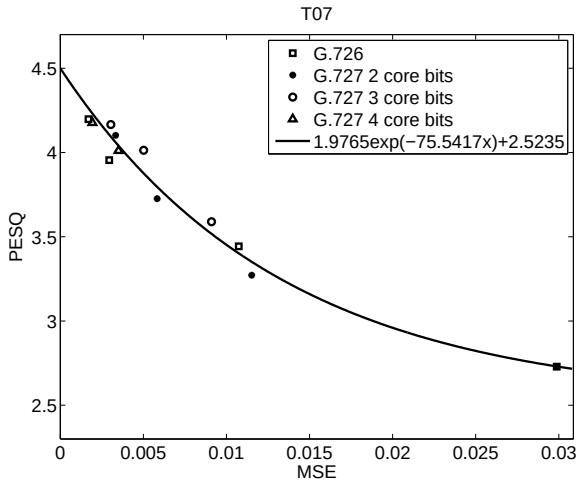


Figure 4.5: The mapping function of narrowband speech T07

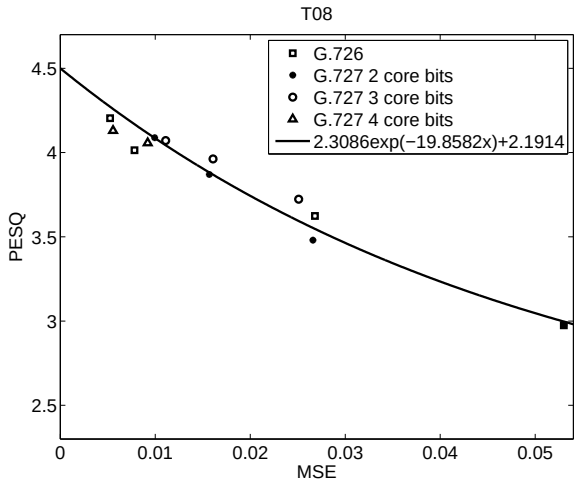


Figure 4.6: The mapping function of narrowband speech T08

4.5 New Theoretical Rate Distortion Bounds for Speech

The rate distortion bounds using MSE as distortion measures are calculated by the classical eigenvalue decomposition [10] and reverse water-filling approach described in Section 3.2.2 on each subsource of the

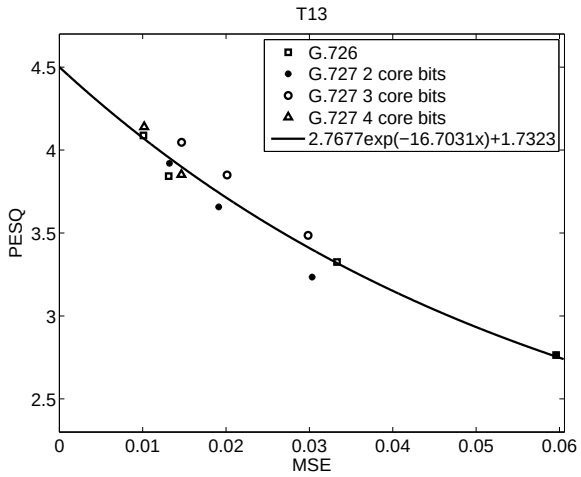


Figure 4.7: The mapping function of narrowband speech T13

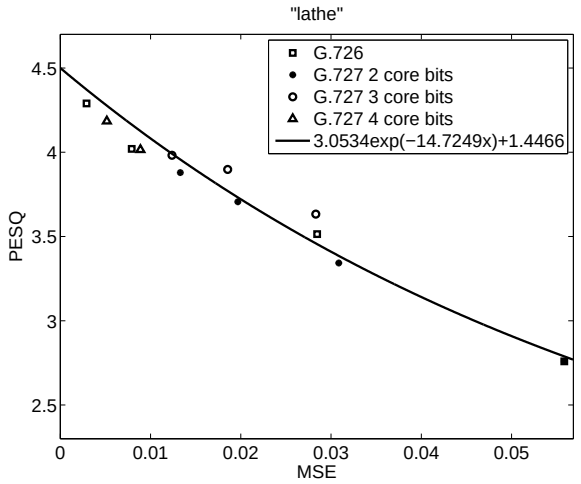


Figure 4.8: The mapping function of narrowband speech "A lathe is a big tool"

composite source models presented in Section 4.2. Then the rate distortion bounds based on MSE are mapped to PESQ-MOS/WPESQ values by the mapping function generated by the ADPCM waveform codecs as described in Section 4.4. Rate distortion bounds are generated for the three different source models for each narrowband and

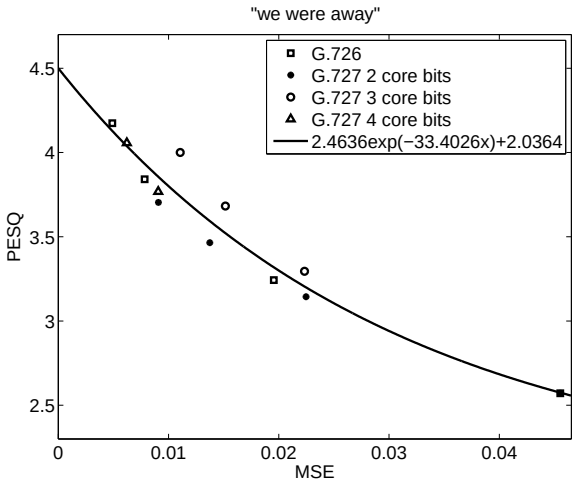


Figure 4.9: The mapping function of narrowband speech "We were away a year ago"

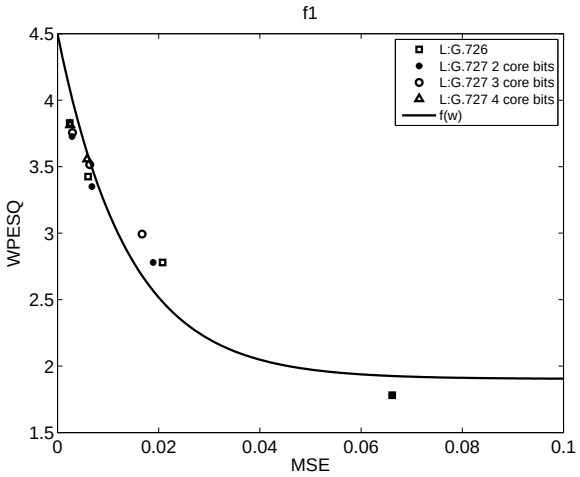


Figure 4.10: The mapping function of wideband speech F1

wideband voice utterance presented in the composite source model section. These are indicated on the plots as: (1) A single voiced model for all frames, labeled as "$R(D)$ over all frames (1 mode)"; (2) A two subsource model with one for speech and one for silence, labeled as "$R(D)$

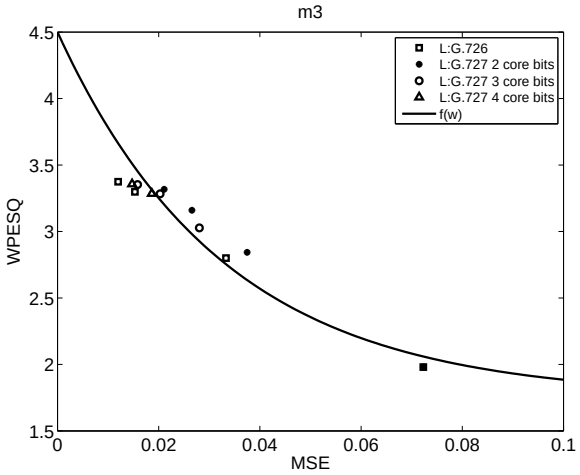


Figure 4.11: The mapping function of wideband speech M3

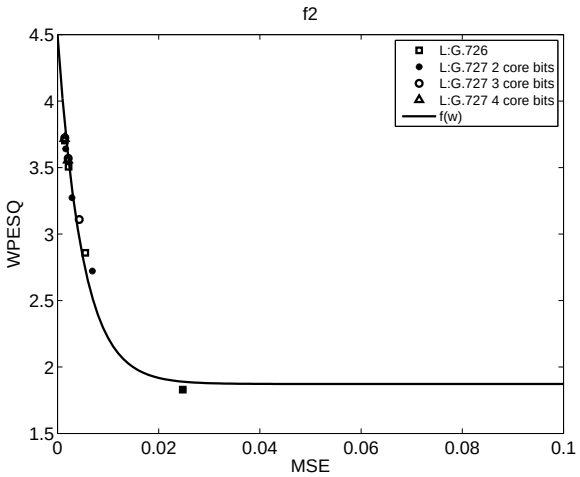


Figure 4.12: The mapping function of wideband speech F2

over all frames (V or S)"; and (3) A five subsource model with subsources corresponding to voiced, unvoiced, onset, hangover, and silence subsources, labeled "$R(D)$ over all frames (5 modes)". The mapping is performed to obtain theoretical rate distortion curves for rate versus PESQ and WPESQ MOS distortion, which can be compared to the

operational rate distortion curves for the speech codecs.

The operational rate distortion performance of six different narrowband speech codecs are compared with the conditional rate distortion bounds based on PESQ in Section 4.5.1, while four wideband speech codecs are compared with the conditional rate distortion bounds based on WPESQ in Section 4.5.2. The results show that our new rate distortion bounds based on perceptual PESQ-MOS/WPESQ distortion measure are indeed lower bounds to the performance of the standardized speech codecs, G.726 (with and without CNG), G.727 (with and without CNG), AMR-NB, G.728 (with and without CNG), G.729, G.718, G.722, G.722.1, and AMR-WB. Detailed discussions of the results follow in those sections.

4.5.1 Rate Distortion Bounds and Operational Rate Distortion Performance for Narrowband Speech

The rate distortion bounds based on PESQ-MOS are compared with CELP codecs such as AMR-NB [1], G.729 [45], and G.718 [38], and ADPCM coders, G.726 and G.727, in Figures 4.13–4.17. For AMR-NB, 8 different bit-rates, 12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15, and 4.75 kbps, are used, and source controlled rate operation is enabled. For G.729, 3 different bit-rates, 6.4, 8, and 11.8 kbps, are used, and DTX/CNG is enabled. For G.718, 2 different bit-rates, 8 and 12 kbps, are used, and DTX/CNG is enabled as well. For G.726 and G.727, 4 bit-rates, 16, 24, 32, and 40 kbps are compared, along with these same rates for speech but with DTX/CNG from AMR at 12.2 kbps implemented. Since G.727 is an embedded speech codec, codecs with 2 core bits are used in our experiments. G.728 operates at 16 kbps and that rate along with G.728 combined with DTX/CNG from AMR at 12.2 kbps are also used for comparisons. The PESQs of all speech codecs are computed by ITU-T P.862 [48].

Examining the three theoretical $R(D)$ curves corresponding to the three source models, we see that the $R(D)$ bound for the single mode, all-voiced model is clearly higher than the other two for all utterances, except for the nearly all voiced utterance, "We were away a year ago" where it is only slightly so. As the models include more subsources, their

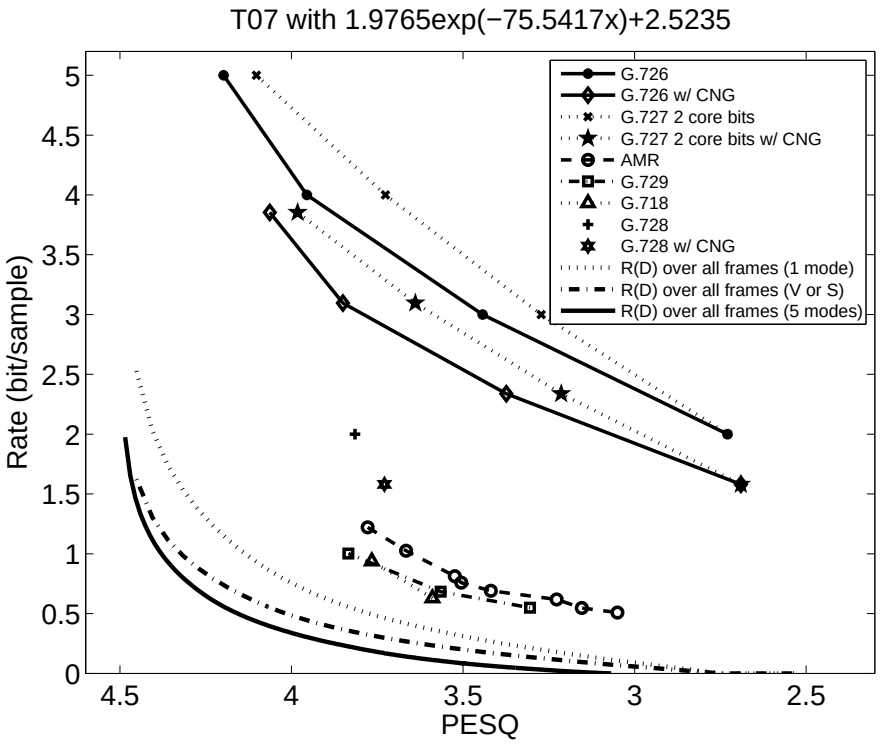


Figure 4.13: The rate distortion bounds and the operational rate distortion performance of narrowband speech T07 using PESQ as the distortion measure. The MSE rate distortion bound is mapped to PESQ as the distortion measure by using the mapping function.

corresponding $R(D)$ bounds move lower. So, a better model yields a more precise lower bound. While this is not surprising, it is instructive in that it illustrates the futility in attempting to lower bound the performance for a given source with a bound computed on a single source model averaged over several subsources. If the averaging were over all five utterances in the figures, the performance curves could not reasonably be expected to lower bound the performance of codecs for any one of the utterances. It is clearly evident that composite source models are important to get reasonable rate distortion bounds.

From Figs. 4.13–4.17, we see that the performance of all narrowband

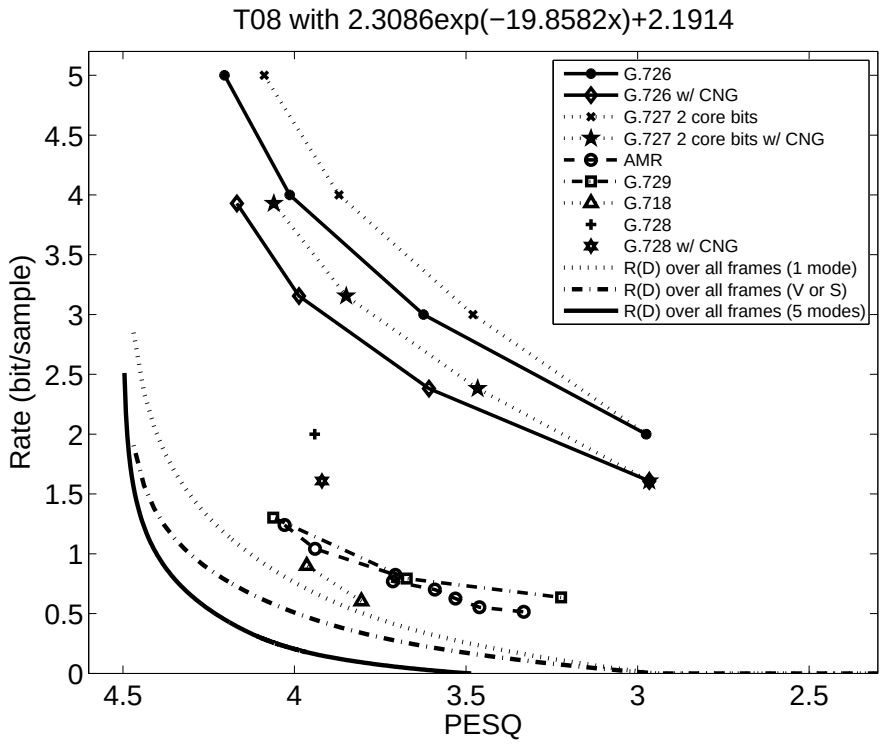


Figure 4.14: The rate distortion bounds and the operational rate distortion performance of narrowband speech T08 using PESQ as the distortion measure. The MSE rate distortion bound is mapped to PESQ as the distortion measure by using the mapping function.

codecs is lower bounded by the rate distortion curves for the source models with 5 and 2 subsources, but the curve with the single source model is actually beaten by the G.718 codec at 8 kbps. Additionally, the 5 source composite source model yields the lowest bound of all three.

As expected, CELP codecs such as AMR-NB, G.729, and G.718 are much closer to the rate distortion bounds than the ADPCM coders. Since G.727 is an embedded ADPCM coder, the performance of G.727 with 2 core bits is worse than that of G.726. In addition, the operational rate distortion performance of G.726 and G.727 is far above the rate

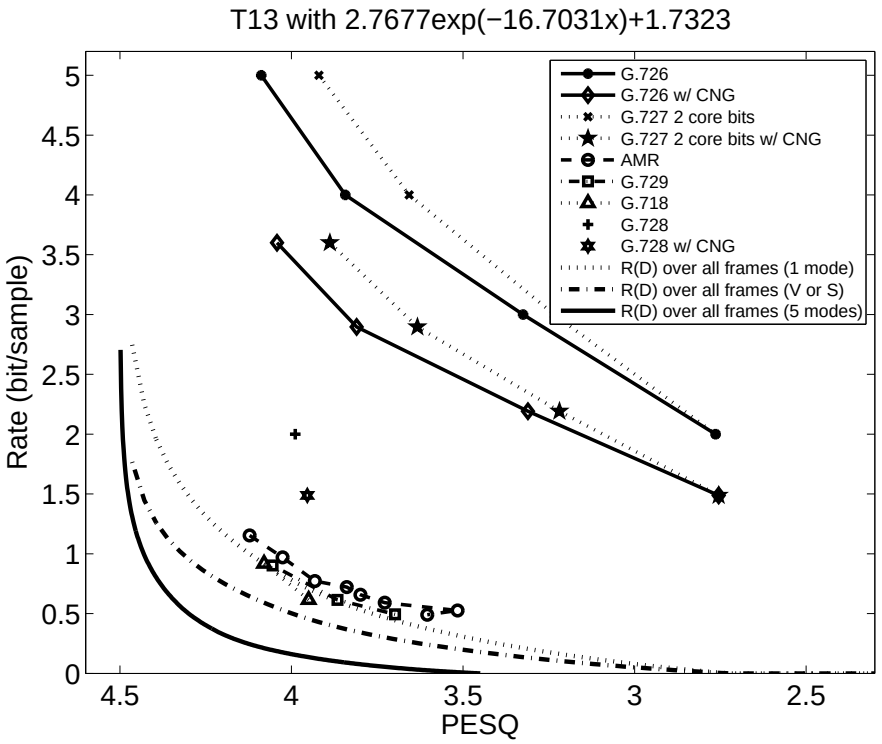


Figure 4.15: The rate distortion bounds and the operational rate distortion performance of narrowband speech T13 using PESQ as the distortion measure. The MSE rate distortion bound is mapped to PESQ as the distortion measure by using the mapping function.

distortion bound since they do not detect silence and code it separately, and they are fully waveform following codecs as opposed to source modeling codecs, the latter of which have the potential for lower rates at the risk of lower quality. The operational rate distortion performance curves of AMR-NB, G.729, and G.718 are quite close. Since they have Voice Activity Detection (VAD) and encode silence by comfort noise generation, the average bit-rate of these codecs is between 1 bit/sample and 1.5 bit/sample for a PESQ-MOS near 4.0 or better.

It is revealing to compare the performance of the standardized codecs to the rate distortion bounds across the five utterances shown.

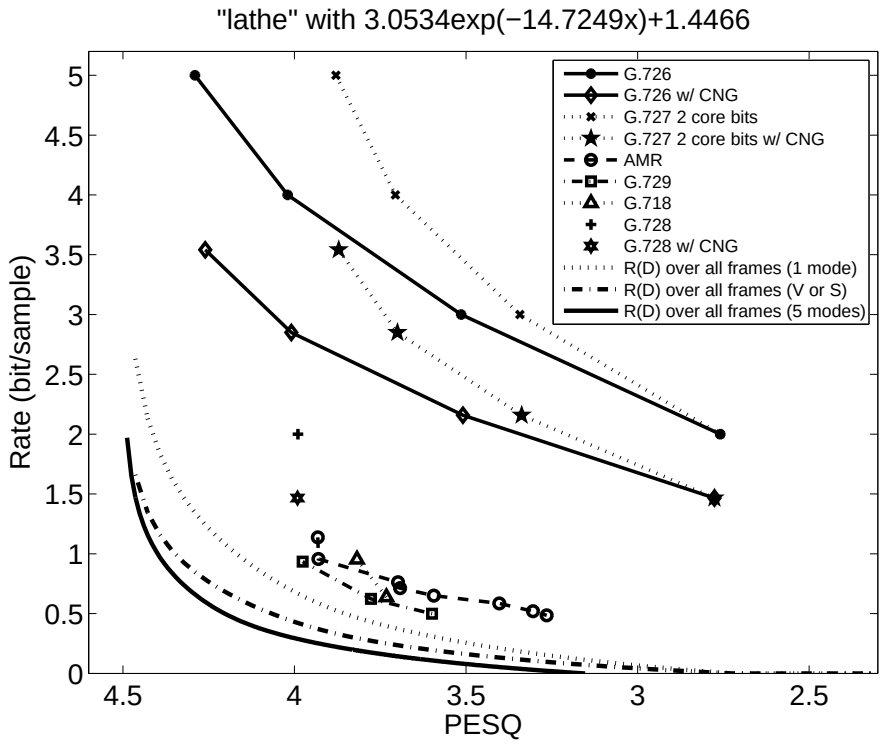


Figure 4.16: The rate distortion bounds and the operational rate distortion performance of narrowband speech "A lathe is a big tool" using PESQ as the distortion measure. The MSE rate distortion bound is mapped to PESQ as the distortion measure by using the mapping function.

The performance of the codecs, AMR-NB, G.729, G.718, and G.728, for the utterance "We were away a year ago" are all significantly closer to the rate distortion bound than for the other sequences. This is because "We were away a year ago" is a fully voiced sequence, and the composite source model is dominated by the voiced mode, which is modeled as a 10^{th} order AR Gaussian source. Therefore, it is evident that the AMR-NB, G.729, and G.718 voice codecs, all based on the CELP predictive coding paradigm are quite efficient at coding voiced speech. However, other speech modes are perhaps less well-modeled by these codecs, and hence, less efficiently coded, as implied by the

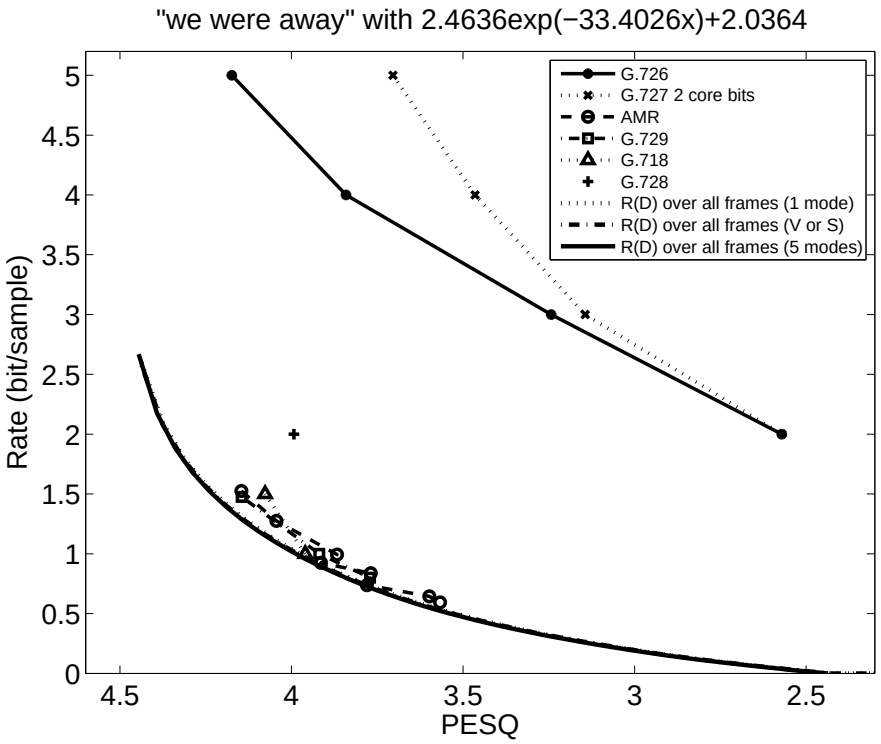


Figure 4.17: The rate distortion bounds and the operational rate distortion performance of narrowband speech "We were away a year ago" using PESQ as the distortion measure. The MSE rate distortion bound is mapped to PESQ as the distortion measure by using the mapping function.

gap between the operational performance of these codecs and the rate distortion bounds in the figures.

For the utterances other than "We were away a year ago", comparisons of the best voice codec operational performance to the lower $R(D)$ bounds reveal that there is no less than approximately 0.5 dB improvement possible for PESQ values between 3.5 and 4.0 with improved codec designs. Since the best of these codecs operate at or near 0.5 bit/sample already, this observation implies that there is a relatively large percentage increase in performance available. It is often said about information theoretic rate distortion bounds that, while the

best possible performance may be indicated, the bounds provide no guidance to how to achieve these bounds. While such comments may be true at first blush, hints at good approaches may be available in the proofs of the bounds, a prominent example of which is that random coding arguments rather naturally imply training mode vector quantizer designs.

However, in the current situation, the way forward to better codec designs may be available more explicitly. At the outset, it is clear that the CELP based codecs are very effective at modeling voiced speech with their linear prediction model as indicated by the $R(D)$ bounds in Fig. 4.17. However, the standardized codecs do not perform as well for the other utterances considered and this is where the opportunity lies. To begin, for these other utterances, one can compare the relative frequency of the subsources for the several utterances and the accuracy of each of the subsource models, and then consider the speech sounds present in each of the utterances. It is possible that better modeling of particular subsources, such as Onset and Hangover, is necessary. It may also be necessary to add more subsources to the codec designs. A good place to look to pursue both of these latter efforts is in some prior low rate multimode codec designs to see how they code the various modes and how effectively they use other speech modes.

These latter steps, better modeling of subsources and adding more subsources, can also be used to refine the $R(D)$ bounds themselves. There are some limitations on codec designs and designers, however, that are not present when one is developing models to calculate rate distortion bounds. Limits on complexity may deter codec designers from pursuing more exotic source models, plus the addition of more codec modes may add to the transmitted bit rate. As a result, separate studies of model building, that is, designing composite source models for speech are needed.

4.5.2 Rate Distortion Bounds and Operational Rate Distortion Performance for Wideband Speech

The mapped rate distortion bounds with the WPESQ distortion measure using different numbers of subsources are compared with CELP

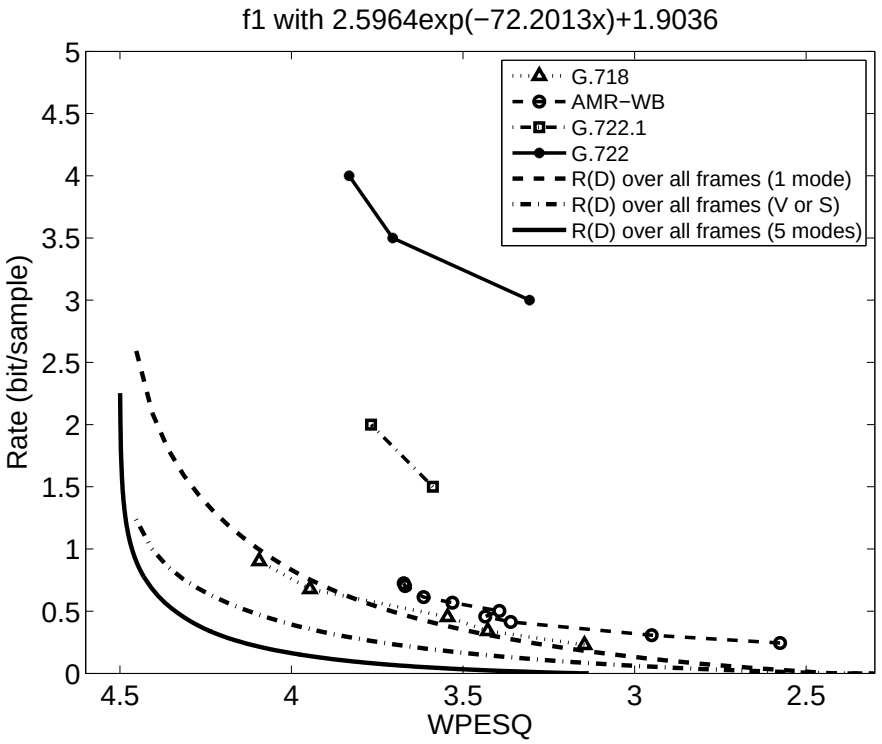


Figure 4.18: The rate distortion bounds and the operational rate distortion performance of wideband speech F1 using WPESQ as the distortion measure. The MSE rate distortion bound is mapped to WPESQ as the distortion measure by using the mapping function (13 pairs).

codecs such as AMR-WB [42], and G.718 [38], G.722.1 [41], and ADPCM coder, G.722 in Figures 4.18–4.20. For AMR-WB, 9 different bit rates, 23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85, and 6.60 kbps, are used, and source controlled rate operation is enabled. For G.718, 5 different bit rates, 8, 12, 16, 24, and 32 kbps, are used, and DTX/CNG is enabled. For G.722, 3 different bit rates, 64, 56, and 48 kbps, are used. There is no DTX/CNG for G.722. The WPESQs of all speech codecs are computed by the ITU-T P.862 [49] wideband version.

In Figures 4.18–4.20, the mapped rate distortion bounds with the WPESQ distortion measure using different numbers of subsources are

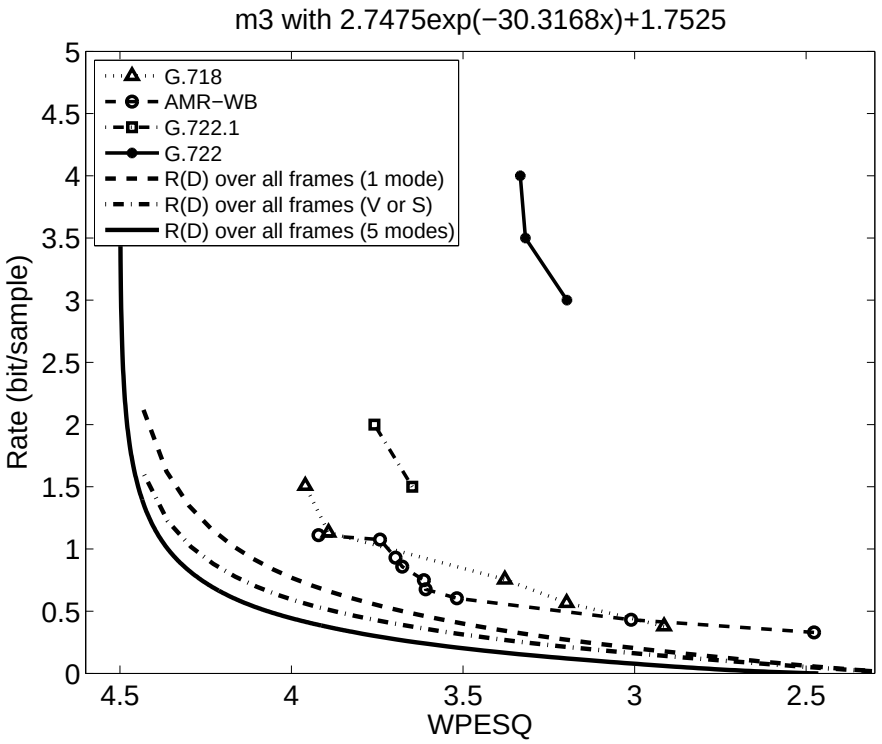


Figure 4.19: The rate distortion bounds and the operational rate distortion performance of wideband speech M3 using WPESQ as the distortion measure. The MSE rate distortion bound is mapped to WPESQ as the distortion measure by using the mapping function (13 pairs).

calculated based on the composite source models shown in Tables 4.4–4.6. These figures show that when the number of subsources increases, the rate distortion bounds get lower, which is similar to the narrowband results. In addition, the performance of all the codecs are bounded by the mapped rate distortion curves which are calculated using five-subsource model. In Figure 4.18, the performance of G.718 is not lower bounded by the curve calculated by one subsource model, but it is lower bounded by five-subsource model curve. This is because the sequence is coded with DTX/CNG, and the source model based on a single subsource does not capture the full complexity of the speech source.

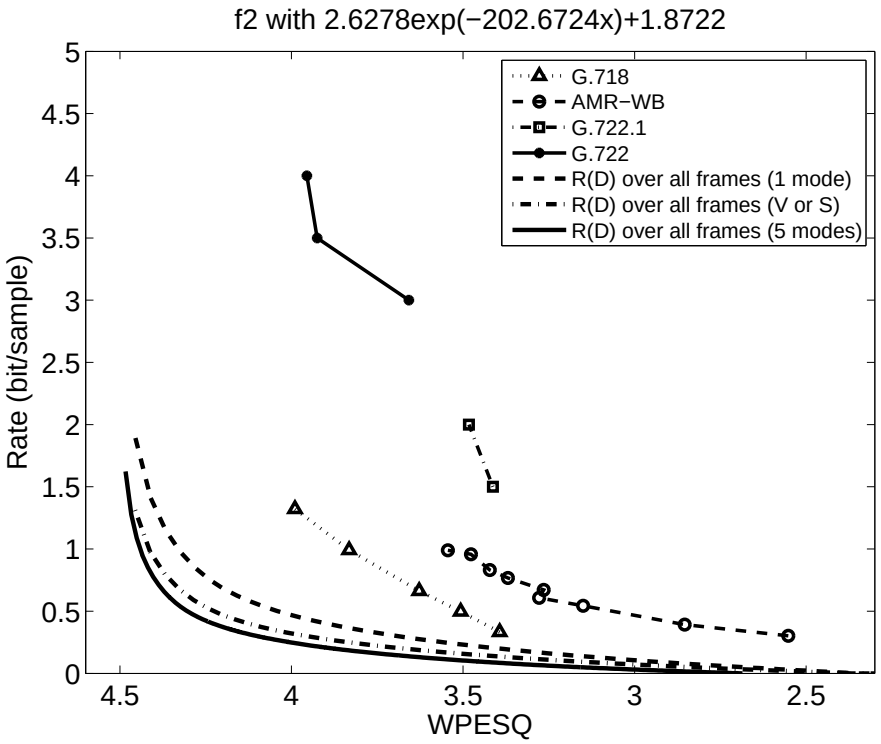


Figure 4.20: The rate distortion bounds and the operational rate distortion performance of wideband speech F2 using WPESQ as the distortion measure. The MSE rate distortion bound is mapped to WPESQ as the distortion measure by using the mapping function (13 pairs).

From Figures 4.18–4.19, as expected, the CELP codecs, AMR-WB and G.718, are much closer to the rate distortion bounds than G.722 and G.722.1. Since AMR-WB and G.718 have Voice Activity Detection (VAD) and encode silence by comfort noise generation, the operational rate distortion curves for AMR-WB and G.718 are quite close to the $R(D)$ bound and also quite close to each other. The operational rate distortion performances of G.722 and G.722.1 are far above the rate distortion bound since they do not detect silence and code it separately. For 4.20, which is Japanese, the AMR-WB and G.718 codecs perform quite differently, with the G.718 codec substantially outper-

forming AMR-WB. Interestingly, the G.718 codec performs relatively close to the rate distortion curve at a PESQ of about 3.4, but tracks away from the $R(D)$ bound rather dramatically at small distortions. The difference in performance for the G.718 and AMR-WB codecs and fact that G.718 performs less well for small distortions are both opportunities to study these two codec designs for this particular Japanese utterance to determine if the codec designs can be improved. The existing $R(D)$ bounds imply that an improvement of at least 0.5 bit/sample in codec performance is possible at a WPESQ-MOS of 4.0 or smaller distortions.

4.5.3 Modifications to the MSE Mapping Function and Other Distortion Measures

The MSE to PESQ/WPESQ/MOS mappings are far from arbitrary and are carefully designed based on several key principles as outlined in Section 4.4. Experiments have been performed that explore the sensitivity of the final rate distortion bounds to variations in the mapping function. Some explicit results for wideband speech sources are given in Gibson and Li [22], which we discuss here. In the cited paper, all rates available in embedded codecs are used to generate MSE versus WPESQ/MOS values upon which to base the mapping function. There is some scatter in the points caused primarily by embedded coding of the higher subband. Two mapping functions are designed based on all original data points and on a subset of the data points after outliers are removed. The points labeled as outliers were determined by listening to the reconstructed speech and determining that the WPESQ/MOS values being obtained were overly optimistic and not representative of the quality as judged by human listeners. (The reader should note that for wideband utterances, it is often difficult to obtain a good objective measure of reconstructed speech quality in the higher frequencies since there is less energy in the higher band and thus conditions for the validity of the WPESQ/MOS tests may not be satistfied fully.)

The mapping with the outliers included mapped higher MSE distortion values into higher values of WPESQ/MOS, which pushes the resulting $R(D)$ bound lower, and therefore, when compared to the op-

erational rate distortion performance of actual voice codecs, the bound is less tight; that is, the bound is more optimistic about possible performance gains achievable by new codec designs. With the outliers removed, did not weight the points that mapped the larger MSE values to better WPESQ/MOS scores, and as a result, the $R(D)$ bound produced is tighter.

Generally, in x-y mappings of MSE to WPESQ/MOS, shifting the mapping upward, lowers, or shifts to the left, the final $R(D)$ function and shifting the mapping downward, raises the $R(D)$ curve, or effectively shifts it to the right. Of course, this conclusion holds for narrowband as well as wideband speech sources. Based on extensive studies, in no cases that we have seen do justifiable variations in the mapping function cause the resulting $R(D)$ bounds to no longer lower bound the best performing codecs. For the tightest bounds we have found, such as in Fig. 4.17, which has the corresponding mapping function in Fig. 4.9, moving the mapping down to overlay the lower data points, will increase the tightness of the, already tight, bound, while moving the mapping up to overlay the upper data points will open up a slight gap between the $R(D)$ function and the operational rate distortion performance of the codecs. Just as in theoretical results, it is always important to investigate how tight the bounds are and take the tightness into account when evaluating codecs.

Perhaps a more natural approach to obtaining a distortion measure for our rate distortion bounds would be to use a weighted MSE fidelity criterion. Indeed, from the prior discussion in Chap. 2, the reader may recall that the CELP codecs do in fact use a weighting based on the spectrum calculated from the linear prediction coefficients every frame, which suggests that the same weighting function should be a strong candidate to serve as the basis for a good fidelity criterion. There are several pitfalls to this approach. First, while the weighting functions are absolutely essential to the success of these analysis-by-synthesis codecs, the weighting often does not achieve the full promise of shaping the coding error such that it always lies below the input source spectrum across the input band of frequencies. Second, this weighting is accepted in codec design but it is not accepted as a valid measure of perceptual

performance for any voice codecs, as MOS and PESQ/MOS are.

As an initial step, however, prior work as in [21] has explored this research direction. Unfortunately, the bounds obtained in Gibson, et al [21], were not very encouraging in terms of the hope of producing tight $R(D)$ bounds. In that work, the weighting was calculated as an average over the entire utterance, and as a consequence, the weighting in any particular segment or frame was not particularly accurate. One alternative would be to recalculate the weighting for each frame of speech, thus producing a more local distortion measure and then averaging the distortion values. Many details of this approach have yet to be investigated. However, as pointed out in the prior paragraph, even if this approach generates what appear to be more reasonable $R(D)$ curves, voice codecs and the voice coding community have not accepted this weighting approach as being reflective of perceptual performance.

4.6 Conclusions

The results show that our new rate distortion bounds do lower bound the PESQ-MOS and WPESQ performance of the best known standardized narrowband and wideband speech codecs. While there is room to improve the bounds by better mode selection and better modeling of the modes, these are the first true bounds on the rate distortion performance of standardized speech codecs to date, and they offer deep insights into how the existing codecs can be improved. Exploration of $R(D)$ bounds based on the approach presented in this chapter should yield valuable insights into research directions to improve voice codecs in the future, and these bounds are a valuable tool in determining whether additional codec design effort might be rewarded. The current $R(D)$ bounds imply that a reduction in rate of 0.5 bit/sample, or approximately 50% is possible.

5

Rate Distortion Bounds for Video

In this chapter we first propose a new correlation model for a digitized natural video that has a local texture dependent spatial component and a temporal component. We then derive theoretical rate distortion bounds that are solely based on the statistical model of the video source with distortion measured in MSE. In the last section of this chapter, we study a constrained rate distortion bound where the constraint is imposed on the channel transition probability by the incorporation of blocking and prediction across neighboring blocks, two common coding steps performed in current video coding standards AVC/H.264 and HEVC.

5.1 Related Prior Work

5.1.1 Statistical Models of Images and Videos

The research on statistically modeling the pixel values within one image goes back to the 1970s when two correlation functions were studied. Both assume a Gaussian distribution of zero mean and a constant variance for the pixel values.

The first correlation model is

$$\rho(\Delta i, \Delta j) = e^{(-\alpha|\Delta i| - \beta|\Delta j|)}, \tag{5.1.1}$$

with Δi and Δj denoting offsets in horizontal and vertical coordinates of any two pixels in a digital image. The parameters α and β control the correlation in the horizontal and vertical directions, respectively, and

their values can be chosen for different images [30]. The separability in spatial coordinates of this correlation model facilitates the analysis of the two-dimensional rate distortion behavior of images using the one-dimensional Karhunen Lòeve transform (KLT).

The second correlation model is an isotropic function

$$\rho(\Delta i, \Delta j) = e^{-\alpha\sqrt{\Delta i^2+\Delta j^2}}, \tag{5.1.2}$$

again with Δi and Δj denoting offsets in horizontal and vertical coordinates of any two pixels in a digital image. This model implies that the correlation between two pixels within an image depends only on the Euclidean distance between them [52]. The major advantage of this model is that it has a closed-form two-dimensional Fourier transform and therefore leads to a closed-form rate function and a closed-form distortion function on a common parameter.

These two correlation models for natural images are simple yet effective in providing insights into image coding and analysis. However image and video coding schemes have advanced significantly and statistical image and video models that are relevant to these more sophisticated methods are needed. Let us start with a close look at the approximate correlation coefficients among the pixel values of some real videos.

Let $X(i, j)$ denote the pixel value at the i^{th} row and the j^{th} column of a digitized image, and let M and N denote the numbers of rows and columns in the image. The approximate correlation coefficient $\hat{\rho}(\Delta i, \Delta j)$ of this image can be expressed as

$$\hat{\rho}(\Delta i, \Delta j) = \frac{\frac{\sum[X(i,j)X(i+\Delta i,j+\Delta j)]}{\sqrt{\sum[X^2(i,j)]\sum[X^2(i+\Delta i,j+\Delta j)]}}}{(M-\Delta i)(N-\Delta j)}, \tag{5.1.3}$$

for $0 \le \Delta i \le M-1$, $0 \le \Delta j \le N-1$. The summations in Eq. (5.1.3) are taken over all pixels whose coordinates satisfy $0 \le i \le M-1-\Delta i$, $0 \le j \le N-1-\Delta j$. Fig. 5.1 plots the approximate correlation coefficients $\hat{\rho}(\Delta i, \Delta j)$ of two digitized natural images, selected from two digitized natural video sequences, paris.cif and football.cif, respectively. We can see in Fig. 5.1 that when Δi and Δj are larger than 50, which is still much smaller than the image size we encounter in present applica-

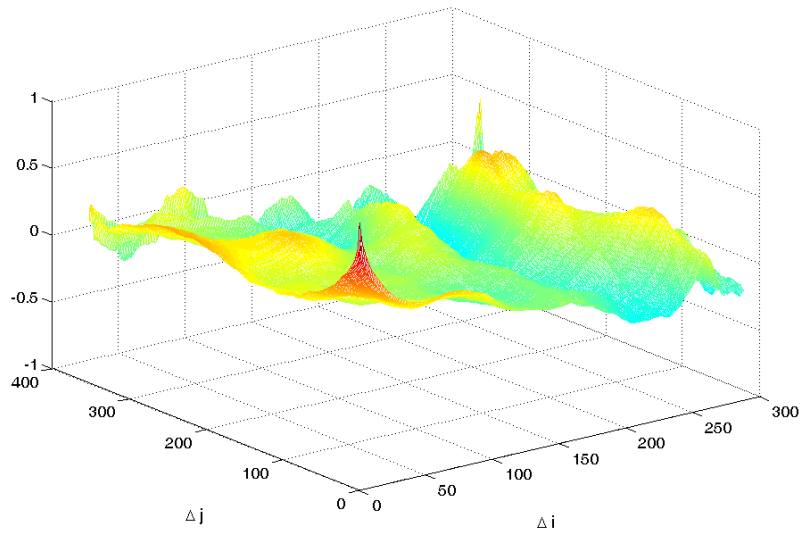
tions, for example 352×288 in this figure, the approximate correlation coefficients $\hat{\rho}(\Delta i, \Delta j)$ are rather random and neither of the two correlation functions can model this behavior. Correspondingly, the rate distortion analysis of natural images based on these two correlation functions will be inaccurate. This is confirmed later in this chapter as the rate distortion bounds calculated based on these two correlation functions are shown actually to be much higher than the operational rate distortion curves of the current video coding schemes.

For the same reason, more recent rate distortion theory work for videos, such as [24, 25, 81] that adopt these two spatial correlation models, is limited in scope. For example, in [24, 25], distortion-rate performance is analyzed by deriving the power spectral density of the prediction error with respect to the probability density function of the displacement error. This is shown, however, to be incapable of describing, with sufficient accuracy, the measured distortion-rate performance of a typical video encoder [79].

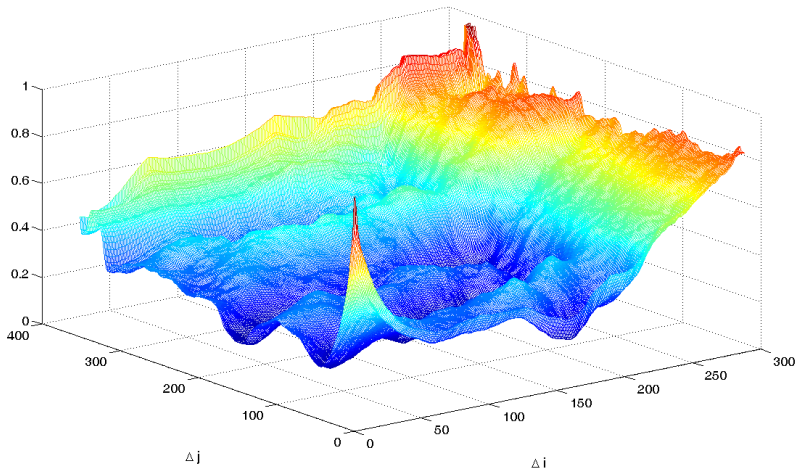
5.1.2 Statistical Models of Practical Video Compression Systems

Researchers working on video compression have developed statistical models of images in the transformed domain. The most popular among them treats the discrete cosine transform (DCT) coefficients in the predicted frames of a video sequence as uncorrelated Laplacian random variables [72, 78]. If the absolute magnitude distortion measure $d(x, \hat{x}) = |x - \hat{x}|$ is used, there is a closed form rate distortion function for the memoryless Laplacian source that can be expanded into a Taylor series and approximated by $R(D) \sim= aD^{-1} + bD^{-2}$.

This quadratic operational rate distortion function is the foundation of the rate control schemes [9, 61, 74] that are adopted by the international video coding standards, such as ISO MPEG-2/4 [32, 33] and ITU-T H.263 [34]. In these rate control schemes, the distortion D in the quadratic operational rate distortion function is approximated by q, the average of the quantization scales used in the video frame. The quantization scales, which are indexed by the quantization parameters (QPs), are hence chosen optimally based on the quadratic rate distortion function $R(q) \sim= aq^{-1} + bq^{-2}$, number of bits left to con-



(a) paris.cif



(b) football.cif

Figure 5.1: The approximate correlation coefficient $\hat{\rho}(\Delta i, \Delta j)$ of two digitized natural images

sume and the approximate coding complexity. The bits spent coding the other syntax elements, considered to be mainly the motion vectors, are monitored and predicted through simple linear or nonlinear functions.

The memoryless Laplacian model for DCT coefficients becomes less appropriate, even for practical video compression system design purposes, since the emergence of new coding standards such as AVC/H.264. The new schemes and refinements in AVC/H.264 [89] reduce the applicability of the memoryless Laplacian model of the DCT coefficients for at least two reasons. First, with all the options offered in the codecs and the very small processed block sizes, the majority of the bandwidth is likely to be allocated to transmit the coding parameters and the motion vectors of each block rather than the DCT coefficients, especially in the low to medium bit rate applications. Since the Laplacian model only treats the DCT coefficients, it becomes insufficient to represent the information in the video source. Second and more importantly, these coding options and parameters are to be chosen, in an optimal way if possible, before the DCT or DCT-like transforms can be applied to the residue block. This is considered as a rate distortion optimization problem and the most popular solution to this problem is to conduct the optimization with a fixed quantization parameter. However, from the perspective of rate control, the quantization parameter is to be optimally chosen based on the residue data after the rate distortion optimization is performed. Therefore there is a "chicken and egg" dilemma artificially caused by modeling the statistics in the transformed domain that has prevented a global optimum from being obtained, even for a specific codec [65, 62, 13].

Two other schemes following in the same vein [65, 62] try to tackle this dilemma by either engaging a "two pass scheme" or defining a "basic unit". This is an ongoing research direction and for more recent activities please refer to [13]. Another recent work on rate distortion modeling for H.264 [90] treats the residue blocks after intra/inter prediction in the spatial domain as Laplacian random vectors with separable correlation coefficients that depend only on one *a priori* parameter. The statistics in the spatial domain are then used to calculate rate

distortion models in the transformed domain. Even though this work also studies the statistics in the spatial domain of videos, it relies on a very simple model of the residue block, and therefore does not address the interdependence between the rate control and rate distortion optimization.

In summary, a new statistical correlation model for digitized natural videos is much needed in both theory and application. This correlation model should be independent of any coding schemes, rather than modeling the processed values, such as the DCT coefficients, in a coding scheme, so that the theoretical rate distortion bounds can be derived to predict the fundamental limit on the number of bits (per pixel) needed to represent a video at a given distortion level. This correlation model should also be more sophisticated than the old correlation models in Eqs. (5.1.1) and (5.1.2) so that the derived theoretical rate distortion bounds are valid. It will be a plus if this correlation model has a simple form with parameters that can be calculated for a specific video, which makes the incorporation of the correlation model into a practical video codec design and evaluation possible. In the next section we propose such a correlation model.

5.2 A New Block-Based Conditional Correlation Model for Video

In this section we propose a new correlation model for a digitized natural video. We assume that all pixel values within one natural video form a three dimensional Gaussian random vector with memory, and each pixel value is of zero mean and the same variance σ^2. We first propose a new correlation model for a digitized natural image or an image frame in a digitized natural video, and then extend the spatial correlation model to the temporal dimension to pixels located in nearby frames of a video sequence.

5.2.1 The Conditional Correlation Model in the Spatial Domain

From the discussion in Section 5.1.1, we know that to study the correlation between two pixel values within one natural image, these two

pixels should be located close to each other compared to the size of the image. Also for a sophisticated correlation model, the correlation between two pixel values should not only depend on the spatial offsets between these two pixels but also on the other pixels surrounding them. When developing the composite source models for speech, we took some cues from successful multimode voice codec designs. Similarly, this can be done for video sources as well. In particular, a coding technique, called "intra-frame prediction", in the video coding standard AVC/H.264, gave us hints on how to deal with the two aforementioned requirements. Intra-frame prediction is explained briefly in Section 2.2.4.

To quantify the effect of the surrounding pixels on the correlation between pixels of interest, we utilize the concept of local texture, which is simplified as local orientation, i.e., the axis along which the luminance values of all pixels in a local neighborhood have the minimum variance. The local texture is similar to the intra-prediction modes in AVC/H.264, but with a generalized block size and an arbitrary number of total textures. To calculate the local texture of a block, we also employ the pixels on the top and to the left of this block as surrounding pixels. However, since we are deriving a source model, we use the original values of these surrounding pixels rather than the previously encoded and reconstructed values used in intra-frame prediction of AVC/H.264.

The block can have any rectangular shape as long as its size is small compared to the size of the image. The local textures need not to be restricted to those defined in AVC/H.264. For example, in Fig. 5.2, the numbered arrows represent a few local textures that are defined as intra-prediction modes in AVC/H.264 and the unnumbered arrows represent a few local textures that are not defined as intra-prediction modes in AVC/H.264. Once the block size and the available local textures are fixed, the local texture of the current block is chosen as the one that minimizes the mean absolute error (MAE) between the original block and the prediction block constructed based on the surrounding pixels and the available local textures. It is important to point out that even though we choose a very simple and computationally inexpensive

way to calculate the local texture, there are other, more sophisticated schemes of doing so, as described in [80], for example, which should produce even better results in rate distortion modeling.

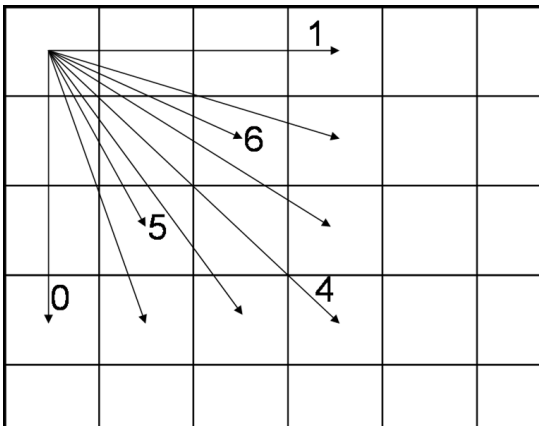


Figure 5.2: The numbered arrows represent a few local textures that are defined as intra-prediction modes in AVC/H.264 and the unnumbered arrows represent a few local textures that are not defined as intra-prediction modes in AVC/H.264

The local texture reveals which one, out of the different available local textures, is the most similar to the texture of the current block. It is reasonable to conjecture that the difference in local texture also affects the correlation between two close pixels within one video frame. To confirm this we first calculate the approximate correlation coefficient between one block of size $M \times N$ whose left top pixel is on i^{th} row and j^{th} column of a video frame, and another nearby block of the same size, shifted by Δi vertically and Δj horizontally, according to the following formula

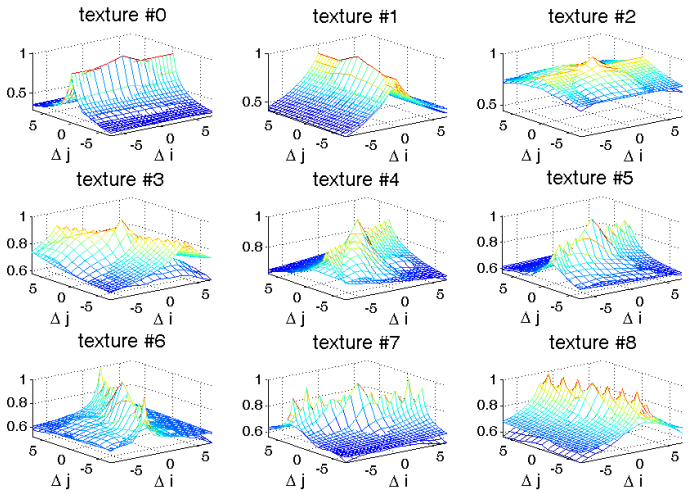
$$\hat{\rho}_s(i,j,\Delta i,\Delta j) = \frac{1}{MN} \frac{\sum[X(i,j)X(i+\Delta i, j+\Delta j)]}{\sqrt{\sum[X^2(i,j)]\sum[X^2(i+\Delta i, j+\Delta j)]}}, \quad (5.2.4)$$

for $-I \leq \Delta i \leq I$, $-J \leq \Delta j \leq J$. This formula is similar to Eq. (5.1.3), except that 1) $M \times N$ is not the size of a whole image, but the size of a block, usually much smaller than the image size; 2) the ranges

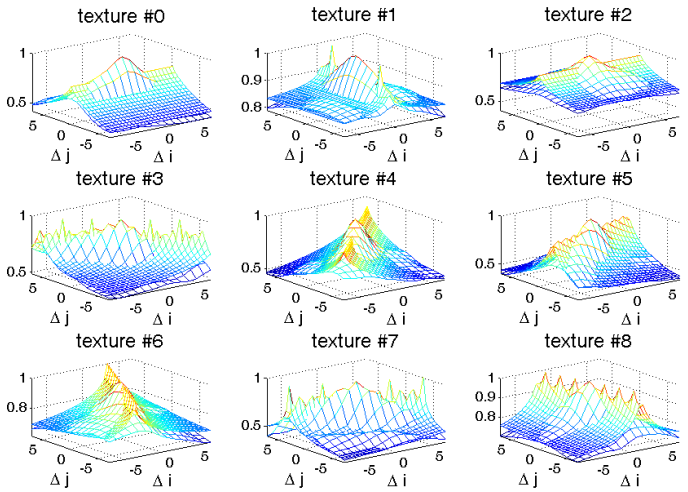
for Δi and Δj are different and need not be smaller than M and N. $\hat{\rho}_s(i, j, \Delta i, \Delta j)$ is first calculated for each $M \times N$ block in an image frame, then they are averaged among the blocks that have the same local texture. We denote this average approximate correlation coefficient for each local texture as $\hat{\rho}_s(\Delta i, \Delta j|y)$ where y denotes the local texture. In other words, $\hat{\rho}_s(\Delta i, \Delta j|y)$ is the average of $\hat{\rho}_s(i, j, \Delta i, \Delta j)$ over all values of i and j in the video frame for which the block that starts at i^{th} row and j^{th} column is of local texture y.

In Figs. 5.3(a) and 5.3(b), we plot $\hat{\rho}_s(\Delta i, \Delta j|y)$ (shown in the figures as the loose surfaces, i.e., the mesh surfaces that look lighter with fewer data points) for the first frames from paris.cif and football.cif, respectively. The dense surfaces, i.e., the mesh surfaces that look darker with more data points, are the correlation coefficients calculated using the proposed conditional correlation model, which is discussed later in this section. The block size is $M = N = 4$. The available nine local textures are chosen to be those plotted in Fig. 2.6. We set Δi and Δj to be very small, ranging from -7 to 7, to concentrate on the dependence of the statistics on local texture in an image frame. Figure 5.3 shows that the average approximate correlation coefficient $\hat{\rho}_s(\Delta i, \Delta j|y)$ is very different for blocks with different local textures. If we average $\hat{\rho}_s(\Delta i, \Delta j|y)$ across all the blocks in the picture, we should get what is shown in Fig. 5.1 in the corresponding region of Δi and Δj, but the important information about the local texture is lost. Not surprisingly $\hat{\rho}_s(\Delta i, \Delta j|y)$ demonstrates certain shapes that agree with the orientation of the local textures. It is also interesting that although the average approximate correlation coefficients of the same local texture in both images demonstrate similar shapes, their actual values are quite different.

Motivated by these observations, in the following we present the formal definition of the new correlation coefficient model for a digitized natural image or an image frame in a digitized natural video that is dependent on the local texture. To distinguish from the approximate average correlation coefficients $\hat{\rho}_s(\Delta i, \Delta j|y)$ calculated from pixel values of video frames, that is what we have discussed so far, we use $\rho_s(\Delta i, \Delta j|y)$ to denote the proposed correlation coefficient model.



(a) paris.cif



(b) football.cif

Figure 5.3: The loose surfaces (the mesh surfaces that look lighter with less data points) are $\hat{\rho}_s(\Delta i, \Delta j|y)$, the approximate correlation coefficients of two pixel values in the first frame from paris.cif and football.cif respectively, averaged among the blocks that have the same local texture; the dense surfaces (the mesh surfaces that look darker with more data points) are $\rho_s(\Delta i, \Delta j|y)$, the correlation coefficients calculated using the proposed conditional correlation model, along with the optimal set of parameters

Definition 5.1. The correlation coefficient of two pixel values with spatial offsets Δi and Δj within a digitized natural image or an image frame in a digitized natural video is defined as

$$\rho_s(\Delta i, \Delta j | Y_1 = y_1, Y_2 = y_2) = \frac{\rho_s(\Delta i, \Delta j | y_1) + \rho_s(\Delta i, \Delta j | y_2)}{2}, \quad (5.2.5)$$

where

$$\rho_s(\Delta i, \Delta j | y) = a(y) + b(y) e^{-|\alpha(y)\Delta i + \beta(y)\Delta j|^{\gamma(y)}}. \quad (5.2.6)$$

Y_1 and Y_2 are the local textures of the blocks the two pixels are located in. They are random variables of integer values between 0 and $|Y| - 1$, where $|Y|$ denotes the total number of local textures. The parameters a, b, α, β and γ are functions of the local texture Y. Furthermore we restrict that $b(y) \geq 0$ and $a(y) + b(y) \leq 1$.

This definition satisfies $\rho_s(\Delta i, \Delta j | Y_1 = y_1, Y_2 = y_2) = \rho_s(-\Delta i, -\Delta j | Y_1 = y_1, Y_2 = y_2)$. To satisfy the other restrictions for a function to be a correlation function: $\rho_s(\Delta i, \Delta j | Y_1 = y_1, Y_2 = y_2) \in [-1, 1]$ and $\rho_s(0, 0 | Y_1 = y_1, Y_2 = y_2) = 1$, we need $a(y) + b(y) = 1$ and $a(y) \geq -1$. In order for the correlation model to approximate as closely as possible the average correlation coefficients in a video, we loosen the requirement $a(y) + b(y) = 1$ to $b(y) \geq 0$ and $a(y) + b(y) \leq 1$. The blocks the two pixels are located in are of the same rectangular shape. The size of the rectangular blocks can potentially affect the accuracy of the correlation coefficient model, which will be discussed later in this section.

This new correlation model discriminates different local textures. As the spatial offsets between the two pixels, Δi and Δj, increase, $\rho_s(\Delta i, \Delta j | Y_1 = y_1, Y_2 = y_2)$ decreases at a different speed depending on the five parameters a, b, α, β and γ, which will be shown to be quite different for different local textures. For each local texture, we choose the combination of the five parameters that jointly minimizes the MAE between the approximate correlation coefficients, averaged among all the blocks in a video frame that have the same local texture, i.e., $\hat{\rho}_s(\Delta i, \Delta j | y)$, and the correlation coefficients calculated using the new model, $\rho_s(\Delta i, \Delta j | y)$.

These optimal parameters for one frame from paris.cif and football.cif respectively and their corresponding MAEs are presented in Table 5.1. The local textures are calculated for each one of the 4 by 4 blocks; the available nine local textures are chosen to be those plotted in Fig. 2.6; Δi and Δj range from -7 to 7. We can see from this table that the parameters associated with the new model are quite distinct for different local textures while the MAE is always less than 0.05. The values of all five parameters are also different for the two videos. In Fig. 5.3 we plot $\rho_s(\Delta i, \Delta j|y)$ of all the local textures for the same images from paris.cif and football.cif using these optimal parameters as the dense surfaces, i.e., the mesh surface with more data points. We can see that the new spatial correlation model does, in fact, capture the dependence of the correlation on the local texture and fits the average approximate correlation coefficients $\hat{\rho}_s(\Delta i, \Delta j|y)$ very well.

The parameters a, b, α, β and γ should have different optimal values when the block size used to calculate the local texture is different. Generally speaking, when the available local textures are fixed, the larger the block size, the less the actual average correlation coefficients should agree with the shape designated by the local texture. What also matters are the ranges of spatial offsets Δi and Δj over which the MAE between $\hat{\rho}_s(\Delta i, \Delta j|y)$ and $\rho_s(\Delta i, \Delta j|y)$ is calculated. The larger the range of spatial offsets, the more average correlation coefficients the model needs to approximate which will normally yield a larger MAE. These two aspects are shown in Fig. 5.4 for four different videos. As we can see in Fig. 5.4 the average MAE over all local textures increases, when the block size and/or the ranges of Δi and Δj increase. Therefore, when we employ the proposed correlation model and its corresponding optimal parameters in applications such as rate distortion analysis, we need to choose the block size and spatial offsets that yield a small MAE, chosen here to be 0.05.

The new spatial correlation model with its optimal parameters a, b, α, β and γ is expected to capture the characteristics of the content of the frames of a video scene. Therefore, the change of the optimal parameters a, b, α, β and γ from one frame to another in a video clip with the same scene needs to be investigated. To study this de-

Table 5.1: The optimal parameters for one frame in paris.cif and football.cif and their corresponding mean absolute errors (MAE's)

paris.cif						
	a	b	γ	α	β	MAE
texture #0	0.3	0.6	0.7	0.0	0.6	0.022
texture #1	0.3	0.6	0.9	-0.2	0.0	0.024
texture #2	0.6	0.3	0.9	0.0	-0.1	0.035
texture #3	0.6	0.3	0.9	-0.2	-0.1	0.043
texture #4	0.6	0.3	0.7	0.1	-0.2	0.034
texture #5	0.6	0.3	0.7	0.2	-0.6	0.028
texture #6	0.6	0.4	0.5	-1.3	0.4	0.026
texture #7	0.6	0.4	0.5	0.4	1.1	0.030
texture #8	0.6	0.4	0.6	0.4	0.1	0.046

football.cif						
	a	b	γ	α	β	MAE
texture #0	0.2	0.6	0.8	0.0	-0.1	0.045
texture #1	0.8	0.2	0.3	-1.0	0.1	0.017
texture #2	0.6	0.3	0.8	0.0	-0.2	0.043
texture #3	0.5	0.5	0.5	0.4	0.5	0.048
texture #4	0.3	0.6	0.7	-0.1	0.1	0.040
texture #5	0.4	0.5	0.9	0.1	-0.3	0.034
texture #6	0.6	0.4	0.5	-0.2	0.1	0.031
texture #7	0.4	0.6	0.5	-0.3	-0.7	0.044
texture #8	0.7	0.3	0.6	0.4	0.1	0.029

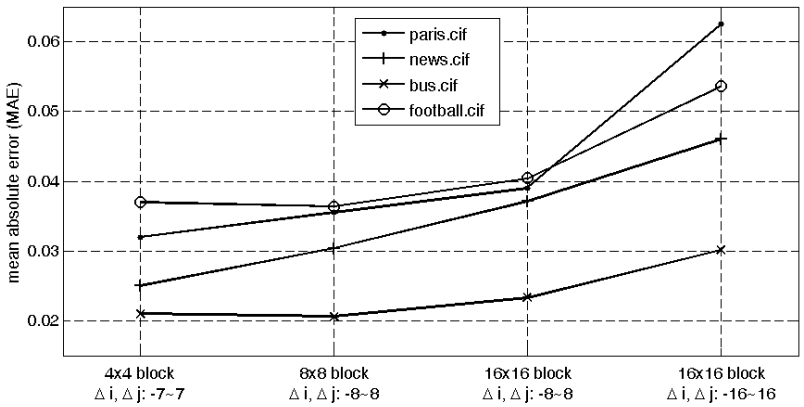


Figure 5.4: The average MAE's over all local textures, for different block sizes and spatial offsets of four videos

pendence, instead of calculating the optimal parameters of each local texture for each frame in a video clip and studying their variations, we use the optimal parameters calculated based on the average correlation coefficients of the first frame, and then study the average MAE over all local textures between the model-calculated correlation coefficients using these parameters and the average correlation coefficients of the following frames in the video clip.

In Fig. 5.5 we plot such MAE's for 90 frames of four CIF videos. We can see that for paris and news, both of which have low motion, the MAE's throughout the whole video sequences are almost the same as that of the first frame. This is not true for football, however, whose MAE's quickly reach beyond 0.1 at frame # 21 and jump to 0.3 at frame # 35. This behavior becomes less surprising when we look at a few of the video frames in this clip as presented in Fig. 5.6. With the high motion in the football video, the frames in this video do not have the same scene any more. For example, frame # 35 looks completely different than the first frame. Therefore, the optimal parameters generated based on one frame can be used in the other frames of the same scene. Different optimal parameters need to be calculated for different scenes, however, even though the frames might reside in the same video.

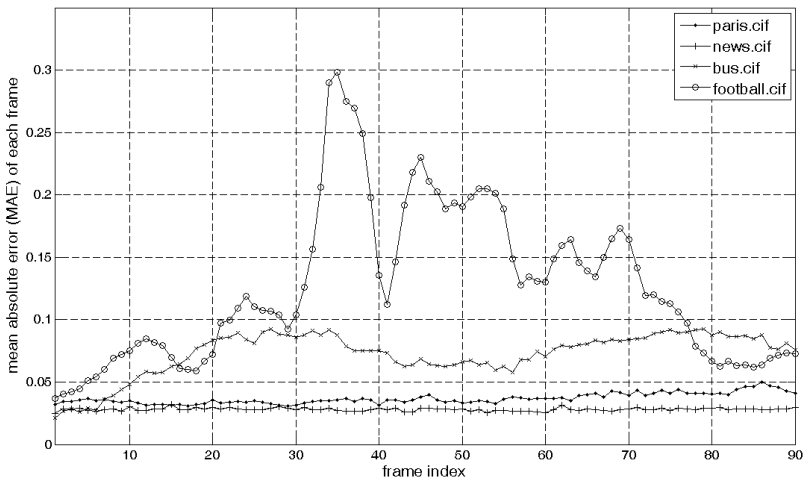


Figure 5.5: The average MAE over all local textures, between the model-calculated correlation coefficients using the optimal parameters of the first frame in a video clip, and the average correlation coefficients of the following frames in the video clip

(a) frame #1

(b) frame #21

(c) frame #35

(d) frame #89

Figure 5.6: Four frames in video clip football.cif

5.2.2 Correlation Among Pixels Located in Nearby Frames

In this section we extend the correlation coefficient modeling from pixels within one video frame to pixels that are located in nearby video frames. Similar to the approach we take in deriving the spatial correlation model, we first study the approximate correlation coefficient between one block of size $M \times N$ in frame k_1 of a video, and another block of the same size, shifted by Δi vertically and Δj horizontally, in frame k_2 of the same video. Equation (5.2.4) is used to calculate the approximate correlation coefficient of each pair of blocks, which is then averaged over all blocks with the same local texture. We denote this extended average approximate correlation coefficient as $\hat{\rho}_s(\Delta i, \Delta j, k_1, k_2|y)$. In Fig. 5.7 we plot $\hat{\rho}_s(\Delta i, \Delta j, k_1 = 1, k_2 = 16|y)$, with y being one of 9 local textures for video silent.cif. As shown in this figure, even though silent.cif is a video of a medium level of motion, the pixels in the first frame and the pixels in the sixteenth frame have quite high correlation; and furthermore, the approximate correlation coefficients between these pixels show certain shapes that are similar to those modeled by the spatial correlation coefficient model we proposed in our previous work.

To isolate the temporal correlation between two frames from the overall correlation, and to apply the spatial correlation coefficient model we already investigated, we first divide, element by element, the overall approximate correlation coefficients $\hat{\rho}(\Delta i, \Delta j, k_1 = 1, k_2 = 16|y)$, by the spatial approximate correlation coefficients $\hat{\rho}_s(\Delta i, \Delta j|y)$ of the first frame, i.e., $\hat{\rho}(\Delta i, \Delta j, k_1 = k_2 = 1|y)$. The results for paris.cif are plotted in Fig. 5.8. As shown in this figure (note that the scales in this figure are different than those in Figs. 5.3 and 5.7), although the fractions are not exactly constant across all the values of Δi and Δj, their variations are much smaller than the variations of the overall approximate correlation coefficients and the spatial approximate correlation coefficients. As a result, we calculate the temporal approximate correlation coefficients, denoted by $\hat{\rho}_t(k_1, k_2|y)$, as the fractions of $\hat{\rho}(\Delta i, \Delta j, k_1, k_2|y)$ over $\hat{\rho}(\Delta i, \Delta j, k_1 = k_2|y)$, averaged over all values of Δi and Δj.

Now let us take a closer look at the temporal approximate correlation coefficients $\hat{\rho}_t(k_1, k_2|y)$ for all frames k_1's, k_2's and local textures

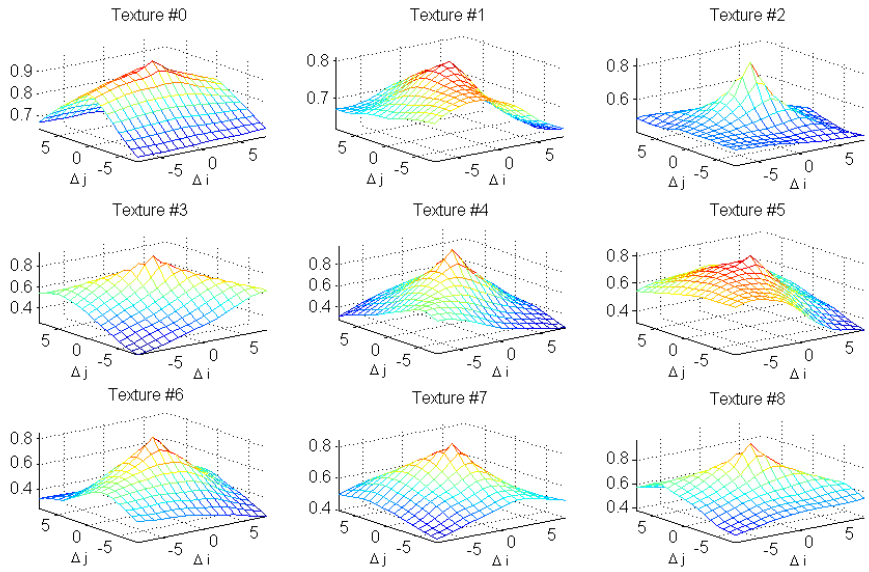


Figure 5.7: $\hat{\rho}(\Delta i, \Delta j, k_1 = 1, k_2 = 16|y)$, the overall approximate correlation coefficients of two blocks, each in the 1^{st} and 16^{th} frames of silent.cif, respectively, averaged among the blocks that have the same local texture

y's of interest. If we investigate the correlation among 16 frames of a video and there are 9 different local textures, for example, we need to calculate and store a $16 \times 16 \times 9$ matrix in order to specify the temporal correlation among all pixels within these 16 video frames. One attempt to reduce the dimension of this matrix is to take the averages of $\hat{\rho}_t(k_1, k_2|y)$ over all local textures y, the result of which is plotted in Fig. 5.9 for paris.cif. Looking at this plot, we notice that when $k_2 > k_1$, $\hat{\rho}_t(k_1, k_2)$ is almost a constant for all values of k_1 and k_2 with the same shift $\Delta k := k_2 - k_1$. We therefore further take the average of $\hat{\rho}_t(k_1, k_2)$ over all values of k_1 and k_2 with the same temporal shift Δk, which results in the curve plotted in Fig. 5.10. As seen from this plot, $\hat{\rho}_t(\Delta k)$ descends as Δk increases from $\Delta k \geq 0$ and it is not quite symmetric with respect to $\Delta k = 0$. The asymmetry in this plot is the result of dividing the overall correlation by the spatial correlation of different frames when isolating temporal correlation from the overall correlation.

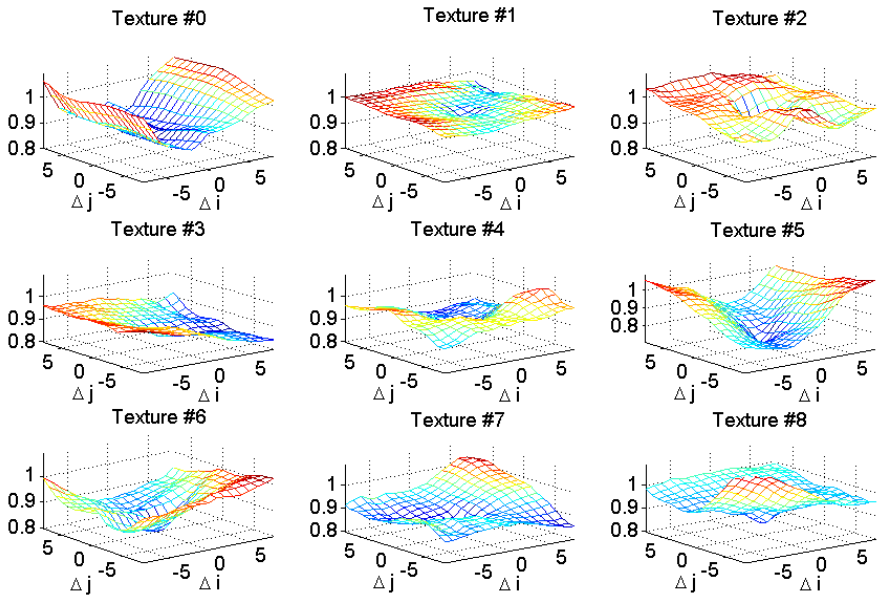


Figure 5.8: $\frac{\hat{\rho}(\Delta i,\Delta j,k_1=1,k_2=16|y)}{\hat{\rho}(\Delta i,\Delta j,k_1=k_2=1|y)}$, the element by element fraction of the overall approximate correlation coefficient over the spatial approximate correlation coefficient of the first frame, of the video paris.cif

Another attempt to reduce the dimension of $\hat{\rho}_t(k_1, k_2|y)$ is to take its average over all values of k_1 and k_2 with the same shift Δk for each local texture y. These results are left to the references [31], but in the next section, we show that for paris.cif, the rate distortion bounds when either $\hat{\rho}_t(\Delta k|y)$ or $\hat{\rho}_t(\Delta k)$ is used are nearly identical except for small distortions. Therefore, for simplicity, we use $\hat{\rho}_t(\Delta k)$, the average of $\hat{\rho}_t(k_1, k_2|y)$ over all k_1 and k_2 with the same shift $\Delta k = k_2 - k_1$ and over all local texture y's, to specify approximately the temporal correlation coefficient between two video frames with index difference Δk.

We conclude this section with the following definition of the overall correlation coefficient model of natural videos that is dependent on the local texture.

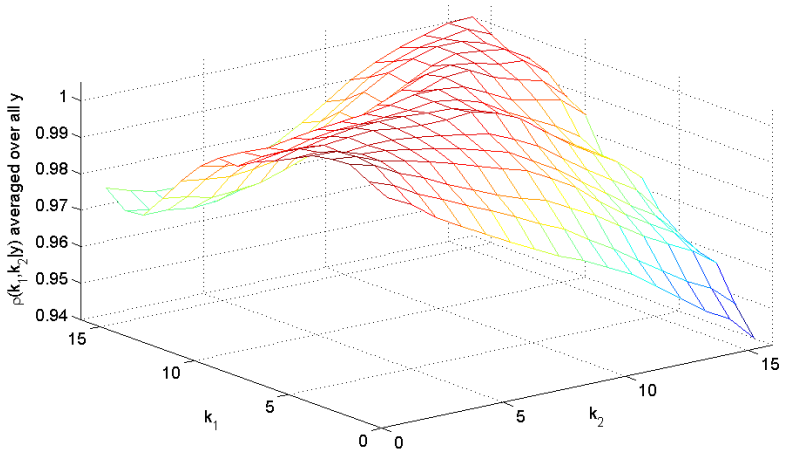


Figure 5.9: $\hat{\rho}_t(k_1, k_2)$, the average of $\hat{\rho}_t(k_1, k_2|y)$ over all texture y's

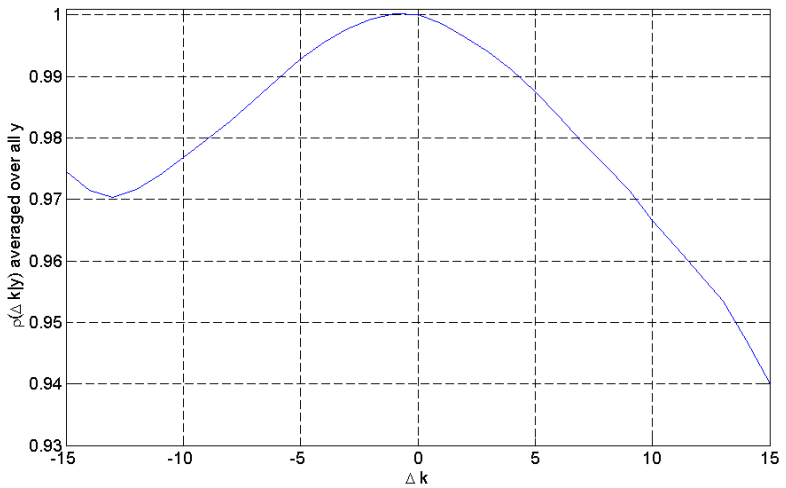


Figure 5.10: $\hat{\rho}_t(\Delta k)$, the average of $\hat{\rho}_t(k_1, k_2|y)$ over all k_1 and k_2 with the same shift $\Delta k = k_2 - k_1$ and all local texture y's, for paris.cif. This average is used to specify approximately the temporal correlation coefficient between two video frames with index difference Δk

Definition 5.2. The correlation coefficient of two pixel values within a digitized video, with spatial offsets Δi and Δj, and temporal offset Δk, is defined as

$$\begin{aligned} &\rho(\Delta i, \Delta j, \Delta k | Y_1 = y_1, Y_2 = y_2) \\ &= \rho_s(\Delta i, \Delta j | Y_1 = y_1, Y_2 = y_2)\rho_t(\Delta_k) \end{aligned} \tag{5.2.7}$$

where $\rho_s(\Delta i, \Delta j | Y_1 = y_1, Y_2 = y_2)$ is the spatial correlation coefficient as defined in Definition 5.1 and $\rho_t(\Delta_k)$ can be calculated by averaging the approximate temporal correlation coefficients $\hat{\rho}_t(\Delta_k|y)$, over all local texture y's.

In the following section, we study the rate distortion bounds of digitized natural videos which depend not only on the correlation model, but also on the pixel variance. Therefore we discuss briefly here the change in pixel variance from one frame to another in a video clip as plotted in Fig. 5.11. The results in Fig. 5.11 agree with those in Fig. 5.5 very well: for videos paris.cif and news.cif which have low motion and therefore can be considered as having only one scene in the entire clips, the change in pixel variance throughout the video clip is almost negligible; for videos with higher motion and frequent scene changes, such as bus.cif and football.cif, a new pixel value variance should be calculated based on the frames in each scene of the video.

5.3 New Theoretical Rate Distortion Bounds of Natural Videos

In this section, we study the theoretical rate distortion bounds of videos based on the correlation coefficient model as defined in Definition 5.2. We compare these bounds to the *intra-frame* and *inter-frame* coding of AVC/H.264 and the High Efficiency Video Coding (HEVC) video coding standards.

To facilitate the comparison with the operational rate distortion functions, we construct the video source in frame k by two parts: $\underline{X}_k$ as an M by N block (row scanned to form a vector of length $M \times N$) and $\underline{S}_k$ as the surrounding $2M+N+1$ pixels ($2M$ on the top, N to the left and the one on the left top corner as shown in Fig. 5.12, forming a

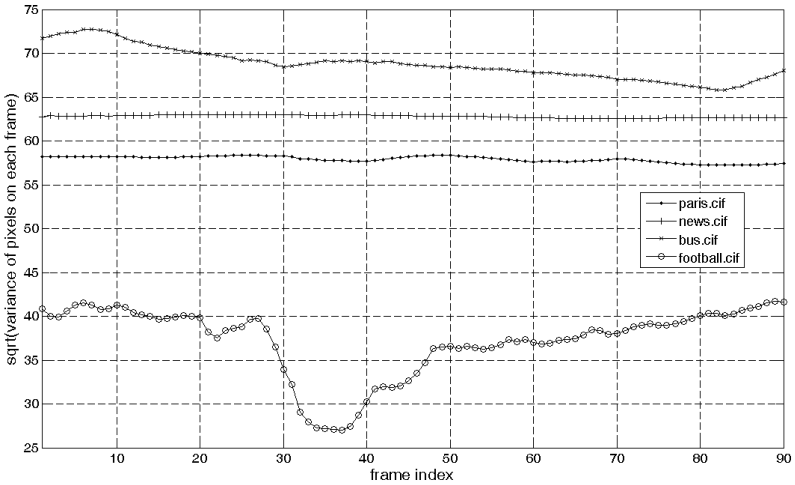


Figure 5.11: Pixel value variance of 90 frames in four video clips

vector of length $2M + N + 1$). When we investigate the rate distortion bounds of a few frames k_1, k_2, ..., k_l, the video source across all these frames is defined as a long vector $\underline{V}$, where

$$\underline{V} = [\underline{X}_{k_1}^T, \underline{S}_{k_1}^T, \underline{X}_{k_2}^T, \underline{S}_{k_2}^T, \ldots, \underline{X}_{k_l}^T, \underline{S}_{k_l}^T]^T. \tag{5.3.8}$$

We assume that $\underline{V}$ is a Gaussian random vector with memory, and all entries of $\underline{V}$ have zero mean and the same variance σ^2. The value of σ is different for different video scenes however, as we discussed at the end of the previous section. The conditional correlation coefficient between each two entries of $\underline{V}$ can be calculated using Definition 5.2 and the spatial offsets Δi and Δj, and temporal offset Δk between these two entries.

We use Y to denote the information of local textures formulated from a collection of natural videos and Y is considered as universal side information available to both the encoder and the decoder. We only employ the first order statistics of Y, $P[Y = y]$, i.e., the frequency of occurrence of each local texture in the natural videos. In simulations, when available, $P[Y = y]$ is calculated as the average over a number of natural video sequences commonly used as examples in video coding

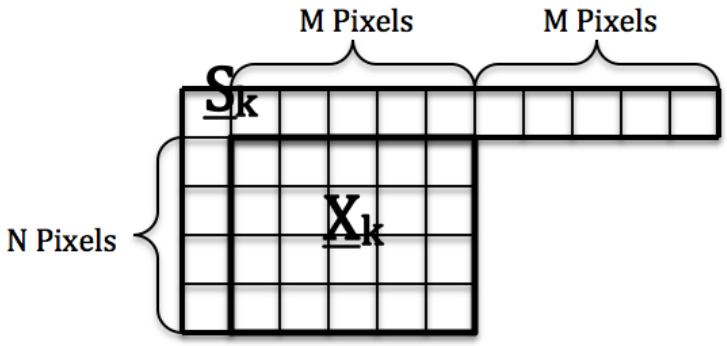


Figure 5.12: The construction of an $M \times N$ block and its surrounding $2M + N + 1$ pixels

studies.

In the following we first investigate briefly the rate distortion bound of $\underline{V}$ without the universal side information Y, the case normally studied in prior rate distortion work for video; we then focus on the case when Y is taken into account in the rate distortion analysis, where the interesting new bounds lie.

5.3.1 Formulation of Rate Distortion Bound without Local Texture as Side Information

The rate distortion bound for the MSE distortion measure without taking into account the texture as side information is a straightforward rate distortion problem of a (single mode) source with memory (no conditioning on textures), which has been studied extensively. It can be expressed as

$$R_{\text{no texture}}(D) = \min_{p(\underline{\hat{v}}|\underline{v}):d(\underline{\hat{V}},\underline{V})\leq D} I(\underline{V};\underline{\hat{V}}), \tag{5.3.9}$$

which is the minimum mutual information between the source $\underline{V}$ and the reconstruction $\underline{\hat{V}}$, subject to a mean square distortion measure $d(\underline{\hat{v}},\underline{v}) = \frac{1}{|\underline{v}|}|\underline{\hat{v}} - \underline{v}|^T|\underline{\hat{v}} - \underline{v}|$. To facilitate the comparison with the case when side information Y is taken into account, we calculate the corre-

lation matrix as

$$E\left[\underline{V}\underline{V}^T\right] = \sum_{y=0}^{|Y|-1} \sigma^2 \rho\left(\underline{V}|y\right) P[Y=y], \tag{5.3.10}$$

i.e., by taking the average of the texture dependent correlation coefficients $\rho\left(\underline{V}|y\right)$ over all local textures. With $\underline{V}$ being a random vector, $\rho\left(\underline{V}|y\right)$ is a correlation coefficient matrix. Each entry of this matrix represents the conditional correlation coefficient between two corresponding entries of $\underline{V}$, which can be calculated using Definition 5.2 and the spatial offsets Δi and Δj, and temporal offset Δk between the two entries of $\underline{V}$.

To calculate $R_{\text{no texture}}(D)$, we first de-correlate the entries of the video source $\underline{V}$ by taking an eigenvalue decomposition of the correlation matrix $E\left[\underline{V}\underline{V}^T\right]$. Reverse water-filling [11] is then utilized to calculate the rate distortion bound of $\underline{V}$, whose entries are independent Gaussian random variables after de-correlation. The details of this calculation for a generic Gaussian source model are in Section 3.2.

5.3.2 Formulation of Rate Distortion Bound with Local Texture as Side Information

The rate distortion bound with the local texture as side information is a conditional rate distortion problem of a source with memory. It is defined in Sec. 3.4 as [6, Sec. 6.1]

$$R_{\underline{V}|Y}(D) = \min_{p(\underline{\hat{v}}|\underline{v},y):E[d(\underline{V},\underline{\hat{V}}|Y)]\leq D} I(\underline{V};\underline{\hat{V}}|Y), \tag{5.3.11}$$

where

$$d(\underline{V},\underline{\hat{V}}|Y) = \textstyle\sum_{\underline{v},\underline{\hat{v}},y} p(\underline{v},\underline{\hat{v}},y) d(\underline{v},\underline{\hat{v}}|y), \tag{5.3.12}$$

and

$$I(\underline{V};\underline{\hat{V}}|Y) = \textstyle\sum_{\underline{v},\underline{\hat{v}},y} p(\underline{v},\underline{\hat{v}},y) \log \dfrac{p(\underline{v},\underline{\hat{v}}|y)}{p(\underline{v}|y)p(\underline{\hat{v}}|y)}. \tag{5.3.13}$$

It can be shown [28] that the conditional rate distortion function in Eq. (5.3.11) can also be expressed as

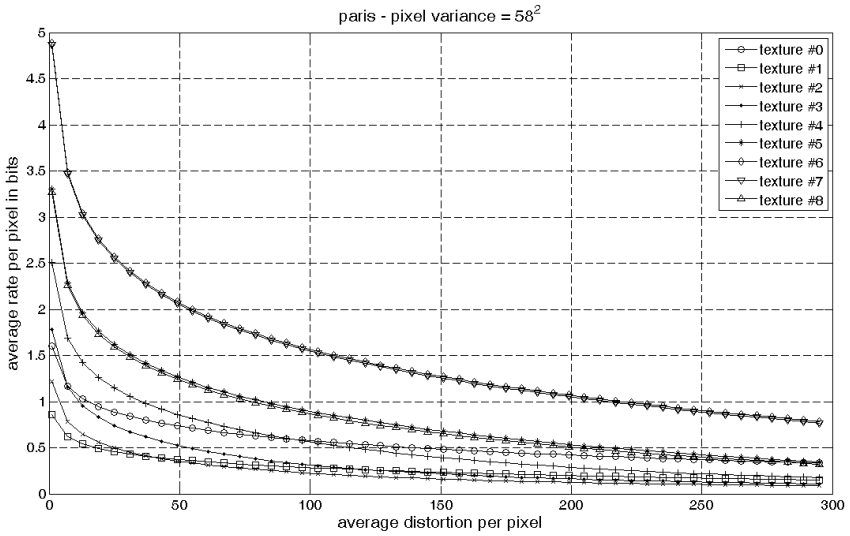
$$R_{\underline{V}|Y}(D) = \min_{D_y:\sum_y D_y p(y)\leq D} \sum_y R_{\underline{V}|y}(D_y) p(y), \tag{5.3.14}$$

and the minimum is achieved by adding up $R_{\underline{V}|y}(D_y)$, the individual, also called marginal, rate-distortion functions, at points of equal slopes of the marginal rate distortion functions, i.e., when $\frac{\partial R_{\underline{V}|y}(D_y)}{\partial D_y}$ are equal for all y and $\sum_y D_y p(y) = D$. These marginal rate distortion bounds can also be calculated using the classic results on the rate distortion bound of a Gaussian vector source with memory and a mean square error criterion as reviewed in Section 3.2, but now the correlation matrix of the source is dependent on local texture y for each subsource.

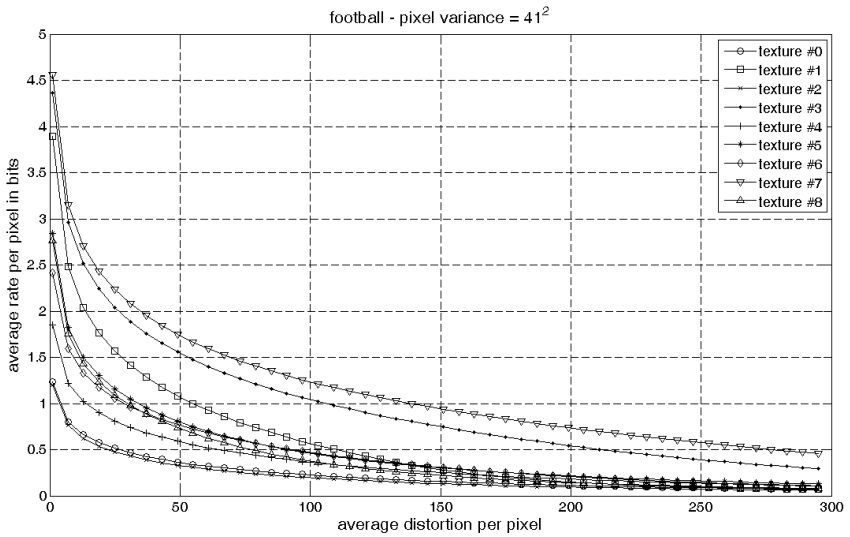
5.3.3 Rate Distortion Bounds for One Video Frame

Because the proposed correlation model discriminates all the different local textures, we can calculate the marginal rate distortion functions for each local texture, $R_{\underline{V}|Y=y}(D_y)$, as plotted in Fig. 5.13 for one frame in paris.cif and football.cif, respectively. The local textures are calculated for each one of the 4 by 4 blocks, the available nine local textures are chosen to be those plotted in Fig. 2.6, and the spatial offsets Δi and Δj are set to range from -7 to 7. The two plots in Figs. 5.13(a) and 5.13(b) show that the rate distortion curves of the blocks with different local textures are very different. Without the conditional correlation coefficient model proposed in this book, this difference could not be calculated explicitly.

The relative order of the nine local textures in terms of the average rate per pixel depends not only on the texture but also on the parameters associated with the correlation coefficient model for each local texture. For example, texture # 1, which is horizontal prediction, by intuition should consume less rate compared to other more complicated textures (# 3 through #8), which is the case for paris.cif. However for football.cif, texture # 1 consumes higher rate for some of the more complicated textures. This can be explained by looking at Fig. 5.3. In Fig. 5.3(b) both the approximate correlation coefficients and the model-calculated correlation coefficients of texture #1 are above 0.8, which is very high compared to those of the other textures. This means that the marginal rate distortion functions depend not only on the local texture, but also on the characteristics of a specific video. The latter dependence is captured by the five parameters a,b,α,β,γ in the new



(a) paris.cif



(b) football.cif

Figure 5.13: Marginal rate distortion functions for different local textures, $R_{\underline{V}|Y=y}(D_y)$, for a frame in paris.cif and football.cif, respectively

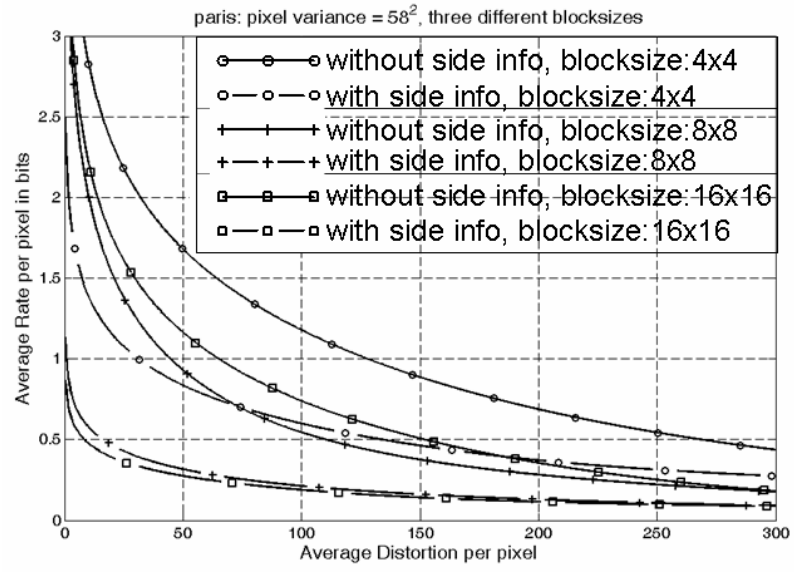
correlation model.

In Fig. 5.14 we plot $R_{\underline{V}|Y}(D)$ and $R_{\text{no texture}}(D)$ as dashed and solid lines, respectively, for two videos and three different blocksizes. The rate distortion curves of paris.cif are generally higher than those of football.cif due to the higher pixel variance in paris.cif. For both videos, the larger the blocksizes, the lower the rate distortion curves. This is reasonable because when correlation among a larger set of pixels is explored, the average rate per pixel should be lower. The difference between each pair of curves (solid line - without side information; dashed line - with side information, the same markers for the same blocksize) in Figs. 5.14(a) and 5.14(b), however, does not have a monotonic relationship with the block size at any distortion level. For example, at distortion 50, for paris.cif, this difference for blocksize 8×8 is lower than those of the other two blocksizes; but for football.cif, this difference for blocksize 8×8 is higher than those of the other two blocksizes.

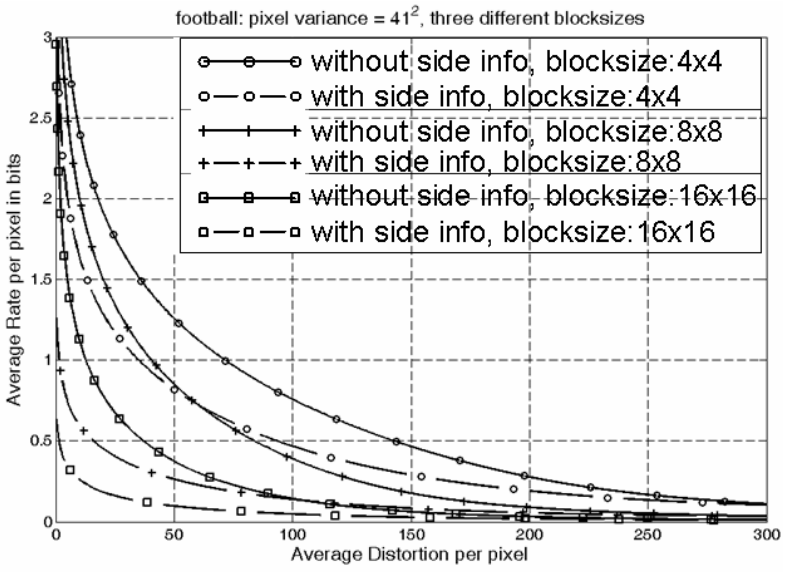
In Fig. 5.16 we plot the two rate distortion bounds $R_{\underline{V}|Y}(D)$ and $R_{\text{no texture}}(D)$ as dashed and solid lines, respectively, as well as the operational rate distortion functions of *intra-frame* coding in AVC/H.264 and in HEVC, for the first frame of paris.cif. The dash dotted line in this figure plots a rate distortion bound calculated based on the new texture dependent correlation model for the scenario where optimal predictive coding is engaged. It will be discussed in detail in the next section.

In AVC/H.264, for both intra-frame and inter-frame coding, we choose the main profile with context-adaptive binary arithmetic coding (CABAC), which is designed to generate the lowest bit rate among all profiles. Rate distortion optimized mode decision and a full hierarchy of flexible block sizes from MBs to 4×4 blocks are used to maximize the compression gain. In HEVC, for both intra-frame and inter-frame coding, we choose CABAC and allow prediction unit sizes from 64×64 to 8×8 and transform block sizes from 32×32 to 4×4 . We also allow the encoder to use two-level hierarchical B frames. For the rate distortion bounds, we choose the block size 16x16 and the spatial offsets as from -16 to 16.

The rate distortion bound without local texture information, plot-



(a) paris.cif



(b) football.cif

Figure 5.14: Comparison of the theoretical rate distortion bounds for two videos and three different blocksizes: solid lines – $R_{\text{no texture}}(D)$ (Eq. (5.3.9)); dashed lines – $R_{\underline{V}|Y}(D)$ (Eq. ((5.3.11))

ted as a black solid line, is higher than the actual operational rate distortion curves of H.264/AVC and HEVC. However, the rate distortion bounds with local texture information taken into account while making no assumptions in coding, plotted as a red dashed line, is indeed a lower bound with respect to the operational rate distortion curves of AVC/H.264 and HEVC.

In order to have a better idea of the region of interest for the average distortion levels, we plot in Fig. 5.15 the correspondence between peak signal to noise ratio (PSNR) and the average distortion when the maximum pixel value is 255. Comparing the two rate distortion bounds $R_{\underline{V}|Y}(D)$ and $R_{\text{no texture}}(D)$ in Fig. 5.16 also shows that engaging the first-order statistics of the universal side information Y saves at least 1 bit per pixel at low distortion levels (distortion less than 25, PSNR higher than 35 dB), which corresponds to a reduction of about 100 Kbits per frame for the CIF videos and 1.5 Mbps if the videos only have intra-coded frames and are played at a medium frame rate of 15 frames per second. This difference decreases as the average distortion increases but remains between a quarter of a bit and half a bit per pixel at high distortion level (distortion at 150, PSNR at about 26 dB), corresponding to about 375 Kbps to 700 Kbps in bit rate difference. A 50% or higher bit rate reduction from the operational rate distortion curve of HEVC intra-frame coding to the new theoretical rate distortion bound can be achieved across the wide range of average distortion shown in this figure.

5.3.4 Rate Distortion Bounds for a Sequence of Video Frames

Now for multiple video frames, we calculate $R_{\text{no texture}}(D)$ and the conditional rate distortion bounds $R_{\underline{V}|Y}(D)$ with the temporal correlation coefficient ρ_t as defined in Eq. (5.3.11) and with correlation coefficients exactly those specified in Definition 5.2.

As before, our approach is to first decorrelate the entries of the video source $\underline{V}$ by taking an eigenvalue decomposition of the correlation matrix, and then reverse water-filling [11] is utilized to calculate the rate distortion bound of $\underline{V}$, whose entries are independent Gaussian random variables after decorrelation. The details of this calculation for

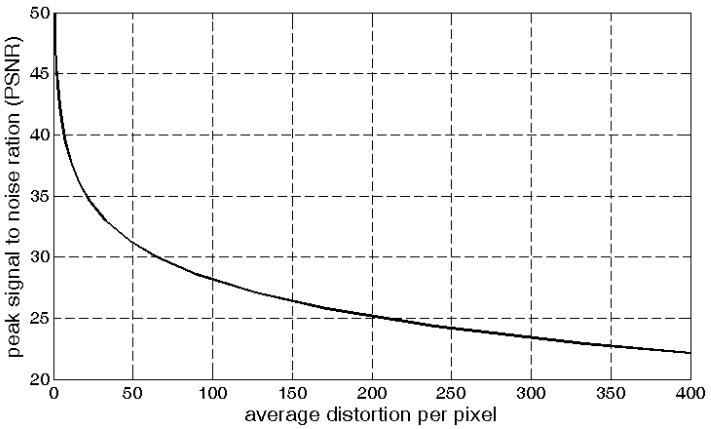


Figure 5.15: The correspondence between peak signal to noise ration (PSNR) in dB and the average distortion when the maximum pixel value is 255

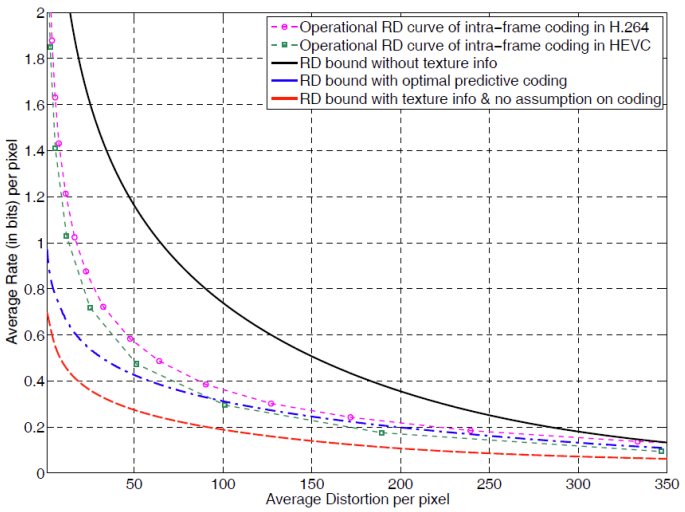


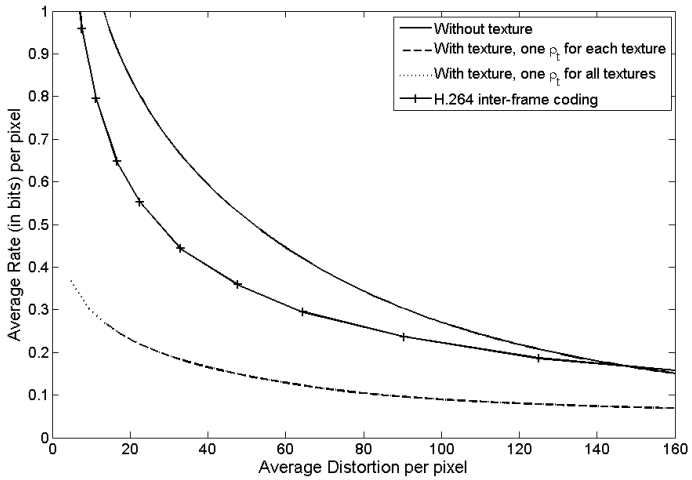
Figure 5.16: Comparison of the rate distortion bounds and the operational rate distortion curves of paris.cif intra-coded in AVC/H.264 and in HEVC

a generic Gaussian source model are in Section 3.2.

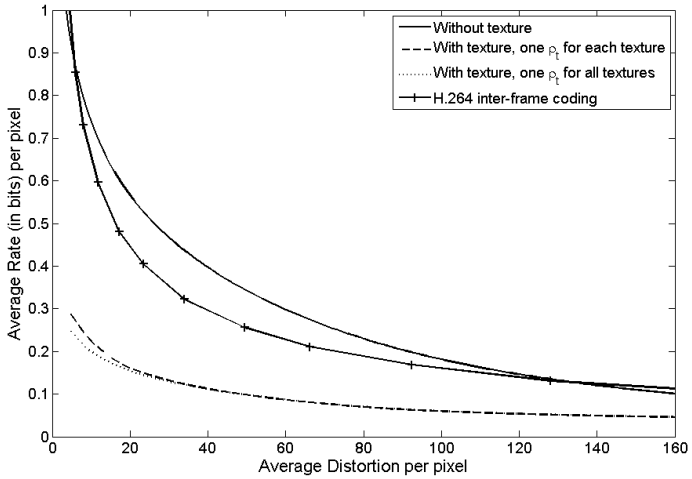
In Figs. 5.17 and 5.18 we plot the conditional rate distortion bound as well as $R_{\text{no texture}}(D)$ in Eq. (5.3.9) for paris.cif and the operational rate distortion curves for paris.cif, inter-frame coded in AVC/H.264. As shown in Figs. 5.17 and 5.18 , the rate distortion bound without local texture information, plotted as solid lines, are higher than, or intersect with, the actual operational rate distortion curve of AVC/H.264. The rate distortion bounds with local texture information taken into account while making no assumptions in coding, using one ρ_t for all textures, plotted as dotted lines, is indeed a lower bound with respect to the operational rate distortion curves of AVC/H.264. In Section 5.2.2 we propose to use $\hat{\rho}_t(\Delta k)$, the average of $\hat{\rho}_t(k_1, k_2|y)$ over all k_1 and k_2 with the same shift $\Delta k = k_2 - k_1$ and over all local texture y's, to specify approximately the temporal correlation coefficient between two video frames with index difference Δk. In Figs. 5.17 and 5.18 it is shown that the rate distortion bounds when either $\hat{\rho}(\Delta k|y)$ or $\hat{\rho}(\Delta k)$ is used are indeed close in values.

Figure 5.19 is similar to Fig. 5.18(b) with the operational rate distortion function of HEVC also included. As can be seen from Fig. 5.19, the theoretical rate distortion bound without the texture information is not a valid lower bound to the operational rate distortion function of HEVC inter-frame coding with a group of pictures size of 5.

Comparing $R_{\text{no texture}}(D)$ (solid lines) and the conditional rate distortion bound (dotted lines) in Fig. 5.17(a) shows that by engaging the first-order statistics of the universal side information Y saves 0.5 bit per pixel at low distortion levels (distortion less than 25, PSNR higher than 35 dB), which corresponds to a reduction of about 50 Kbits per frame for the CIF videos and 750 Kbps if the videos have a group of picture size equal to 2 and are played at a medium frame rate of 15 frames per second. This difference decreases as the average distortion increases but remains 0.1 bit per pixel at high distortion level (distortion at 150, PSNR at about 26 dB), corresponding to about 150 Kbps in bit rate difference. Similar to intra-frame coding, a 50% or higher bit rate reduction from the operational rate distortion curve of HEVC inter-frame coding to the new theoretical rate distortion bound

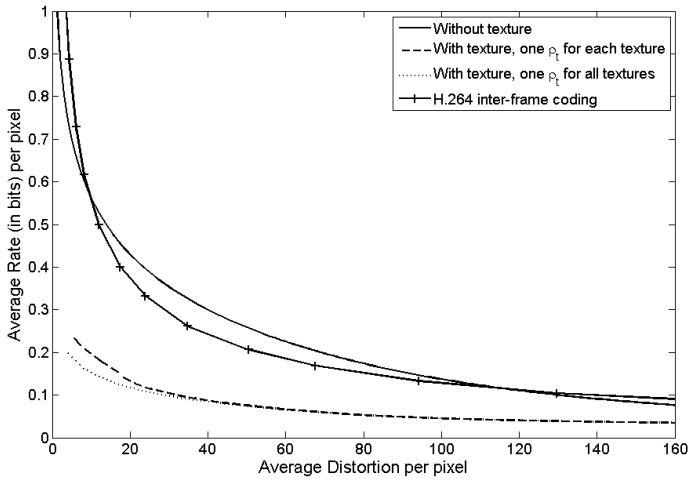


(a) Frames 1 and 2

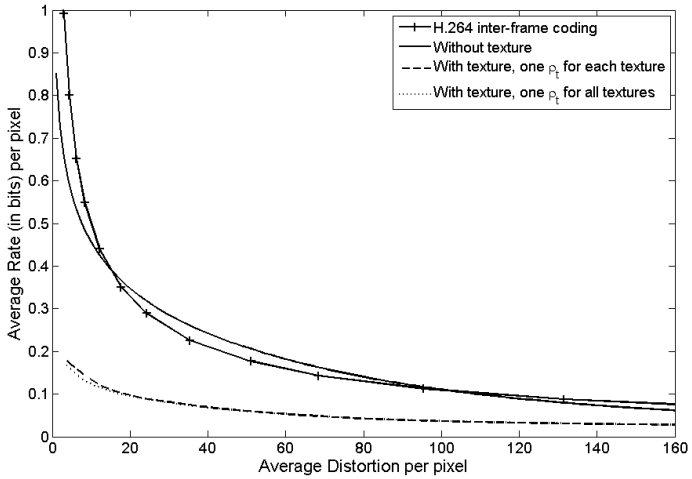


(b) Frames 1, 2 and 3

Figure 5.17: Theoretical rate distortion bounds and the rate distortion curves of inter-frame coding in AVC/H.264 - part 1 of 2



(a) Frames 1, 2, 3 and 4



(b) Frames 1, 2, 3, 4 and 5

Figure 5.18: Theoretical rate distortion bounds and the rate distortion curves of inter-frame coding in AVC/H.264 - part 2 of 2

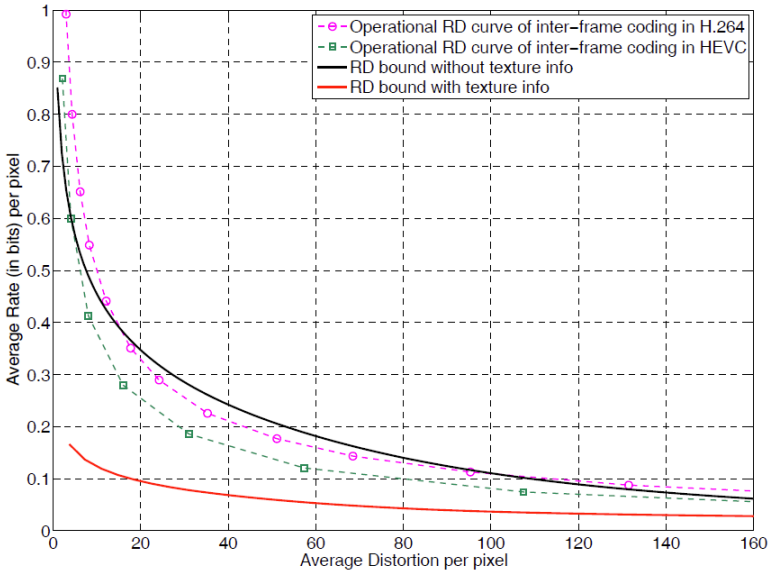


Figure 5.19: Comparison of the rate distortion bounds and the operational rate distortion curves of paris.cif inter-coded in AVC/H.264 and in HEVC

can be achieved across the wide range of average distortion shown in Fig. 5.19.

Another interesting observation of Figs. 5.17 and 5.18 is that as more video frames are coded, the actual operational rate distortion curves of inter-frame coding in AVC/H.264 become closer and closer to the theoretical rate distortion bound when no texture information is considered. This is because in AVC/H.264, only the intra-coded frames (i.e., only the 1^{st} frame in our simulation) take advantage of the local texture information through intra-frame prediction, while the inter-coded frames are blind to the local texture information. Therefore, when more frames are inter-coded, the bit rate saving achieved by intra-frame prediction in the 1^{st} frame is averaged over a larger number of coded frames. This suggests a possible coding efficiency improvement in video codec design by involving texture information even for inter-coded frames.

5.4 Constrained Rate Distortion Bounds for Blocking and Intra-frame Prediction

Breaking an image frame into 16×16 pixel MBs and processing one MB at a time, commonly known as the "blocking" scheme, has been employed in the most popular image coding standards such as JPEG and almost all video coding standards such as MPEG-2/4 and the H.26x series [32, 33, 34, 35]. In AVC/H.264, intra-frame prediction is utilized to reduce the spatial redundancy in the intra-coded frames, as discussed in Section 2.2.4. With the new block-based local-texture-dependent correlation model, an explicit study of the rate distortion behavior of these key schemes, such as blocking and intra-prediction, is feasible. In this last section of this chapter, we derive a constrained rate distortion bound where the constraint is imposed on the test channel transition probability by the incorporation of blocking and prediction across neighboring blocks, two common coding steps performed in current video coding standards AVC/H.264 and HEVC.

The basic setup can be summarized in the block diagram in Fig. 5.20. $\underline{X}$ denotes the M by N block currently being processed. The surrounding $2M + N + 1$ pixels ($2M$ on the top, N to the left and the one on the left top corner), denoted by $\underline{S}$, are used to form a prediction block for each one of the available local textures, as

$$\underline{Z} = \underline{X} - P_d^{(A)} \underline{S}, \tag{5.4.15}$$

where $P_d^{(A)}$ is a $M \times N$ by $2M + N + 1$ matrix, different for each local texture. A is the local texture chosen for the current block which yields the smallest prediction error. $\underline{Z}$ and A are further coded and transmitted to the decoder, where the predicted value is added in to obtain

$$\underline{\hat{X}} = \underline{\hat{Z}} + P_d^{(\hat{A})} \underline{\hat{S}}. \tag{5.4.16}$$

In the block diagram in Fig. 5.20, Y denotes the information of local textures formulated from a collection of natural images and is considered as universal side information available to both the encoder and the decoder. The number of available local textures is denoted by $|Y|$.

With the block based nature of the new correlation model, we study the penalty paid in average rate when the correlation among the

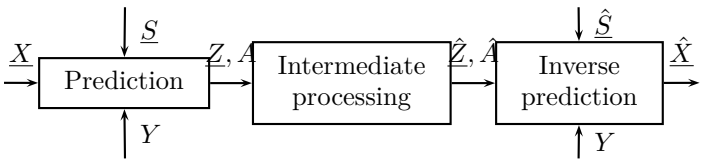


Figure 5.20: Coding of one M by N block $\underline{X}$ and the surrounding $2M + N + 1$ pixels $\underline{S}$

neighboring MBs or blocks is disregarded completely (blocking, Section 5.4.1) or is incorporated partially through the predictive coding (blocking and intra-frame prediction, Section 5.4.2).

Two different distortion constraints are considered in this section, denoted by "avgD" and "sepD" respectively ($|\underline{S}|$ denotes the length of $\underline{S}$ and $|\underline{X}|$ denotes the length of $\underline{X}$):

Average distortion constraint (avgD):

$$\frac{1}{|\underline{S}| + |\underline{X}|} \left\{ E[||\underline{S} - \hat{\underline{S}}||^2] + E[||\underline{X} - \hat{\underline{X}}||^2] \right\} \leq D. \tag{5.4.17}$$

Separate distortion constraint (sepD):

$$\frac{1}{|\underline{S}|} E[||\underline{S} - \hat{\underline{S}}||^2] \leq D \text{ and } \frac{1}{|\underline{X}|} E[||\underline{X} - \hat{\underline{X}}||^2] \leq D. \tag{5.4.18}$$

The average distortion constraint is used dominantly in image and video compression, while recent research in perceptual quality measurement of videos has suggested the importance of the separate distortion constraint on maintaining perceptual video quality, because the variation in video quality from frame to frame or from one region to another in the same frame induces an unpleasant viewing experience of the human users.

In the previous two sections the lowest rate that can be achieved by coding $\underline{X}$ and $\underline{S}$ together is studied; therefore, we only use the average distortion constraint "avgD". In this section we use the separate distortion measure, "sepD" since in video coding each MB is processed sequentially and only local distortion is considered. The rate distortion bounds calculated using "sepD" should be slightly higher than those when "avgD" is used.

5.4.1 Constrained Rate Distortion Bound for Blocking

Since in this subsection we are interested in the penalty paid in average rate when the correlation among the neighboring MBs or blocks are disregarded completely, $\underline{S}$ and $\underline{X}$ are coded separately with the separate distortion constraint "sepD" in Eq. (5.4.18). The total rate can be calculated as

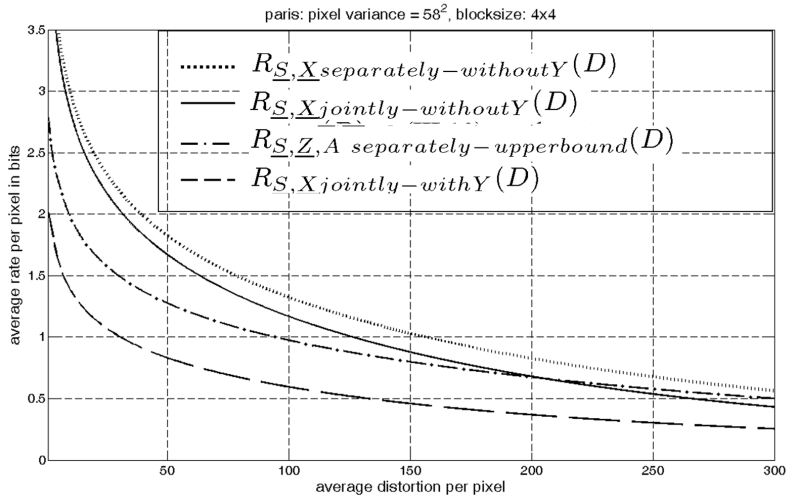
$$R_{\underline{S},\underline{X} separately-without Y}(D) = \frac{R_{\underline{X}}(D)|\underline{X}|+R_{\underline{S}}(D)|\underline{S}|}{|\underline{S}|+|\underline{X}|}, \tag{5.4.19}$$

which is the average of the rate distortion functions of $\underline{X}$ and $\underline{S}$. We plot $R_{\underline{S},\underline{X}\ separately-without Y}(D)$ as dotted lines in Figs. 5.21, 5.22, and 5.23 for two videos and three different block sizes. Not surprisingly for both videos and all three block sizes, coding $\underline{S}$ and $\underline{X}$ separately costs more bits than coding them jointly. The difference in bit rate decreases as the block size increases, since for smaller block sizes information on stronger correlation across the blocks is disregarded. With the new correlation coefficient model and the corresponding rate distortion curves, we can calculate explicitly the bit rate increase caused by blocking. For example, this penalty is one sixth bit per pixel in this plot at all distortion levels in Fig. 5.21(a), which is quite significant.

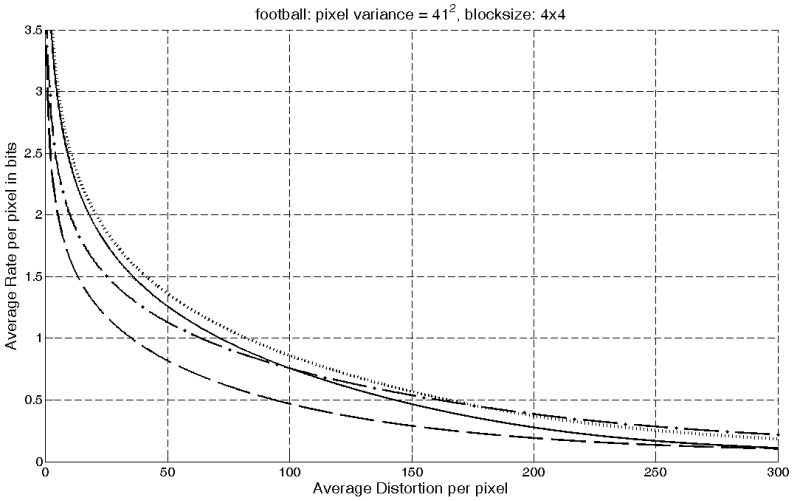
5.4.2 Constrained Rate Distortion Bound for Blocking and Optimal Intra-frame Prediction

In the following we focus on the scenario when the video frames are processed block by block sequentially but the correlation among the blocks is utilized through predictive coding. We restrict ourselves to the separate distortion measure "sepD" in Eq. (5.4.18) and therefore $\underline{S}$ is coded with no consideration of $\underline{X}$, after which $\underline{Z}$ and A are calculated by using intra-prediction in Eq. (5.4.15). The rate distortion function for this scenario is

$$\begin{aligned} R_{\underline{S},\underline{Z},A\ separately-without Y}(D) = \Bigg(& \min_{p(\underline{\hat{s}}|\underline{s}):\ \frac{E[||\underline{S}-\underline{\hat{S}}||^2]}{|\underline{S}|} \le D} I(\underline{S};\underline{\hat{S}}) \\ & + \min_{p(\underline{\hat{z}},\hat{a}|\underline{z},a,\underline{s},\underline{\hat{s}}):\ \frac{E[||\underline{X}-\underline{\hat{X}}||^2]}{|\underline{X}|} \le D} I(\underline{Z},A;\underline{\hat{Z}},\hat{A}) \Bigg) / (|\underline{S}|+|\underline{X}|) \end{aligned} \tag{5.4.20}$$

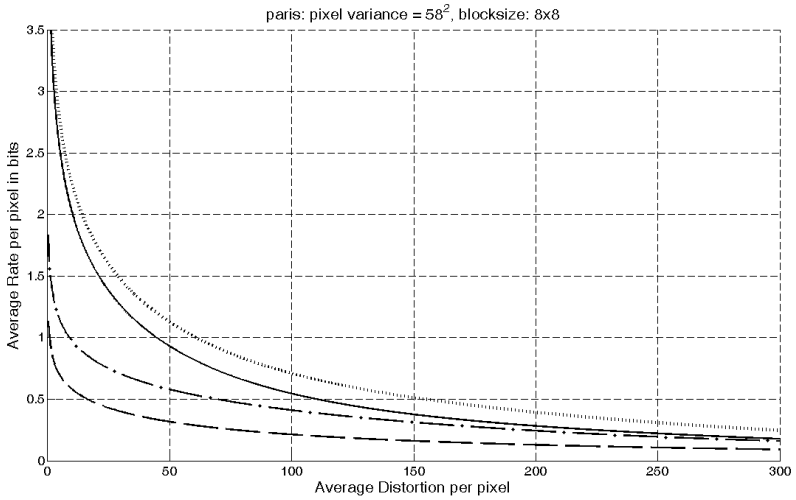


(a) paris, block size 4×4

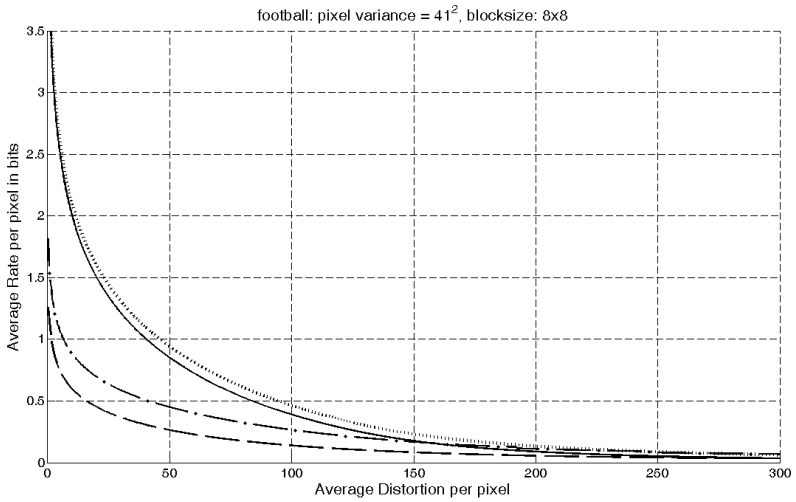


(b) football, block size 4×4

Figure 5.21: Comparison of rate distortion bounds of Section 5.4 for two videos and three blocksizes (part 1 of 3): solid lines – $R_{\underline{S},\underline{X}\ jointly-withoutY}(D)$ in Eq. (5.3.9); dashed lines – $R_{\underline{S},\underline{X}\ jointly-withY}(D)$ in Eq. (5.3.11); dotted lines – $R_{\underline{S},\underline{X}\ separately-withoutY}(D)$ in Eq. (5.4.19) ; dash dot lines – $R_{\underline{S},\underline{Z},A\ separately-sep-upperbound}(D)$ in Eq. (5.4.26)

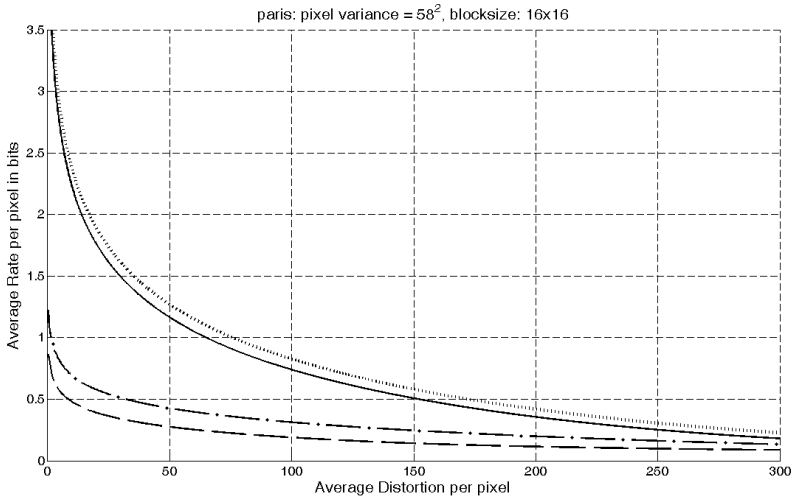


(a) paris, block size 8×8

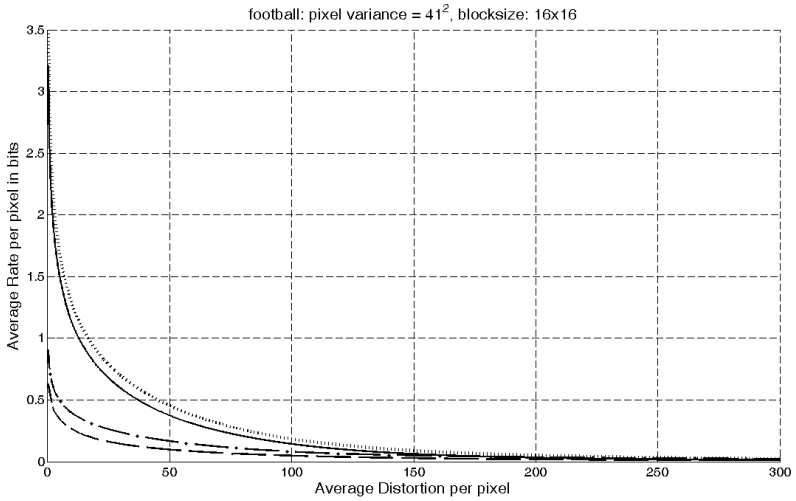


(b) football, block size 8×8

Figure 5.22: Comparison of rate distortion bounds of Section 5.4 for two videos and three blocksizes (part 2 of 3): solid lines – $R_{\underline{S},\underline{X}\ jointly-without_Y}(D)$ in Eq. (5.3.9); dashed lines – $R_{\underline{S},\underline{X}\ jointly-with_Y}(D)$ in Eq. (5.3.11); dotted lines – $R_{\underline{S},\underline{X}\ separately-without_Y}(D)$ in Eq. (5.4.19) ; dash dot lines – $R_{\underline{S},\underline{Z},A\ separately-sep-upperbound}(D)$ in Eq. (5.4.26)



(a) paris, block size 16x16



(b) football, block size 16x16

Figure 5.23: Comparison of rate distortion bounds of Section 5.4 for two videos and three blocksizes (part 3 of 3): solid lines – $R_{\underline{S},\underline{X}\ jointly-without_Y}(D)$ in Eq. (5.3.9); dashed lines – $R_{\underline{S},\underline{X}\ jointly-with_Y}(D)$ in Eq. (5.3.11); dotted lines – $R_{\underline{S},\underline{X}\ separately-without_Y}(D)$ in Eq. (5.4.19) ; dash dot lines – $R_{\underline{S},\underline{Z},A\ separately-sep-upperbound}(D)$ in Eq. (5.4.26)

If we restrict that $A = \hat{A}$, i.e., we code the local texture A losslessly, the second part in Eq. (5.4.20) becomes

$$\begin{aligned} &\min_{p(\underline{\hat{z}},\hat{a}|\underline{z},a,\underline{s},\underline{\hat{s}}):\frac{1}{|\underline{X}|}E[||\underline{X}-\underline{\hat{X}}||^2]\le D} I(\underline{Z},A;\underline{\hat{Z}},\hat{A}) = \\ &\min_{p(\underline{\hat{z}}|\underline{z},a,\underline{s},\underline{\hat{s}}):\frac{1}{|\underline{X}|}E[||\underline{X}-\underline{\hat{X}}||^2]\le D} I(\underline{Z};\underline{\hat{Z}}|A) + H(A), \end{aligned} \tag{5.4.21}$$

which forms an upper bound for all the scenarios when A is coded either losslessly or subject to a fidelity criterion. Also when $A = \hat{A}$, we have

$$\begin{aligned} &E[||\underline{X}-\underline{\hat{X}}||^2] = \textstyle\sum_a Pr(a)E[||(\underline{Z}+P_d^{(a)}\underline{S}) - (\underline{\hat{Z}}+P_d^{(a)}\underline{\hat{S}})||^2|a] \\ &\quad = \textstyle\sum_a Pr(a)\int_{\underline{s}}\int_{\underline{\hat{s}}}\int_{\underline{z}}\int_{\underline{\hat{z}}} p(\underline{z},\underline{\hat{z}},\underline{s},\underline{\hat{s}}|a)(\underline{\hat{z}}-\underline{z})^T(\underline{\hat{z}}-\underline{z}) + \\ &(\underline{\hat{s}}-\underline{s})^T P_d^{(a)T} P_d^{(a)}(\underline{\hat{s}}-\underline{s}) + 2(\underline{\hat{s}}-\underline{s})^T P_d^{(a)T}(\underline{\hat{z}}-\underline{z}) d\underline{s}d\underline{\hat{s}}d\underline{z}d\underline{\hat{z}}. \end{aligned} \tag{5.4.22}$$

In order to investigate the lowest rate when predictive coding is employed, we use the optimal linear predictor $P_{opt}^{(a)} = E[\underline{X}\underline{S}^T|a](E[\underline{S}\underline{S}^T|a])^{-1}$ assuming that $E(\underline{S}\underline{S}^T|a)$ is non-singular. Since the source is assumed to be zero-mean Gaussian, the optimal linear predictor is also the optimal conditional mean predictor. The optimality is in the sense of minimizing MSE of $\underline{X}$. When the optimal linear predictor $P_{opt}^{(A)}$ is used, the cross-product term in Eq. (5.4.22) disappears. Let

$$D'_{\underline{S}} = \sum_a Pr(a)\int_{\underline{s}}\int_{\underline{\hat{s}}} p(\underline{s},\underline{\hat{s}}|a)(\underline{\hat{s}}-\underline{s})^T P_{opt}^{(a)T} P_{opt}^{(a)}(\underline{\hat{s}}-\underline{s}) d\underline{s}d\underline{\hat{s}}. \tag{5.4.23}$$

Eq. (5.4.22) becomes

$$E[||\underline{X}-\underline{\hat{X}}||^2] = |\underline{Z}|D_{\underline{Z}} + D'_{\underline{S}}. \tag{5.4.24}$$

Since $\underline{S}$ is optimally coded without consideration of $\underline{X}$ as in the first part of Eq. (5.4.20), $D'_{\underline{S}}$ is fixed as well. The constraint on the distortion of $\underline{Z}$ becomes

$$D_{\underline{Z}} \le (|\underline{X}|D - D'_{\underline{S}})/|\underline{Z}|. \tag{5.4.25}$$

An upper bound for Eq. (5.4.20) is thus

$$\begin{aligned} &R_{\underline{S},\underline{Z},A\ separately-upperbound}(D) = \frac{1}{|\underline{S}|+|\underline{X}|} \\ &\left(|\underline{S}|R_{\underline{S}}(D) + |\underline{Z}|R_{\underline{Z}|A}(\frac{|\underline{X}|D-D'_{\underline{S}}}{|\underline{Z}|}) + H(A)\right) \end{aligned} \tag{5.4.26}$$

The conditional rate distortion function $R_{\underline{Z}|A}(D_{\underline{Z}})$ in Eq. (5.4.26) is again calculated based on the "equal slope" theorem of the marginal rate distortion functions $R_{\underline{Z}|A=a}(D_a)$ [28]. In this case since the actual local texture A is coded without any loss, the exact statistics of A are available at both the encoder and the decoder; therefore, whether the universal side information Y is available or not becomes insignificant. The only complexity in computation is caused because $E(\underline{S}\underline{S}^T|a)$ is usually singular when the direction of the local texture is DC, horizontal, vertical, or too close to horizontal/vertical. In these cases we use the pseudo-inverse matrix of $E(\underline{S}\underline{S}^T|a)$ in the calculation.

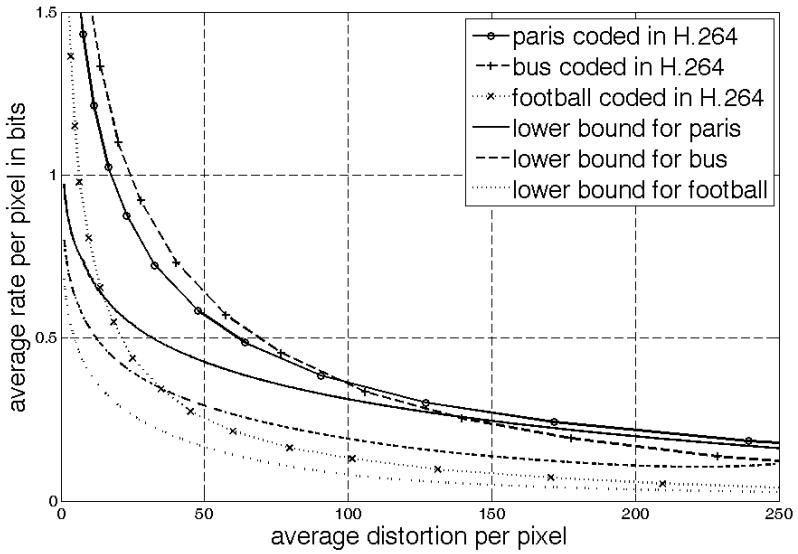
The bit rate decrease from the dotted lines (coding $\underline{S}$ and $\underline{X}$ separately, Eq. (5.4.19)) to the dash-dotted lines (the upper bound of coding $\underline{S}$, $\underline{Z}$ and A separately with optimal prediction, Eq. (5.4.26)) is truly phenomenal in all the plots in Figs. 5.21, 5.22, and 5.23 at low distortion levels, corresponding to about 1 bit per pixel for paris and between half a bit to 1 bit per pixel for for football at distortion 25 (corresponding to PSNR 35 dB). This bit rate saving decreases as the distortion increases, and interestingly, it vanishes for football at certain distortions. This is because spending bits coding the local texture A losslessly becomes unjustifiable at high distortion levels. This is especially true when the bit rate is low and the processing block size is small. We can see that in Fig. 5.21(b) the dash-dotted line and the dotted line intersect at a distortion of about 180, corresponding to an average rate of 0.4 bits per pixel. The average bit rate spent on coding the local texture A losslessly is simply the entropy of A, divided by the number of pixels per block, which is 16 in Fig. 5.21(b) since 4×4 blocks are investigated. This average rate is about 0.2 bits per pixel, or 50% of the total average rate. This is to say that for this particular video football.cif, processed in 4×4 blocks, 0.4 bits per pixel is the threshold in average rate that depicts when incorporating the correlation among the neighboring blocks through optimal intra-frame predictive coding and coding the local texture A losslessly, becomes worse than discarding the correlation among the neighboring blocks. This crossover average rate is different for different videos and different processing blocksizes, as can be seen in Figs. 5.21, 5.22, and 5.23 . It

can be calculated along with the rate distortion bounds we derive in this chapter and be utilized in real video codecs.

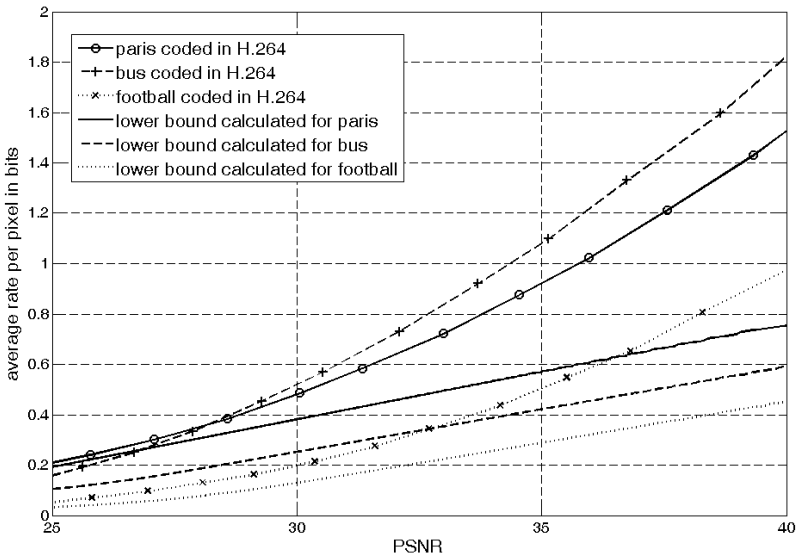
Among all the rate distortion functions we investigate in this chapter, engaging prediction and coding $\underline{S}$, $\underline{Z}$ and A separately with the separate distortion constraint, as in Section 5.4.2, is the most similar to intra-fame coding in state-of-the-art codecs such as AVC/H.264. The upper bound $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ in Eq. (5.4.26) is achieved when the local texture A is losslessly coded and optimal prediction is employed. Since in AVC/H.264, for intra-coded frames, the intra-modes are always coded losslessly, $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ should be a lower bound for the operational rate distortion function of intra-frame coding in AVC/H.264. If we remove all the assumptions on coding, the rate distortion bound of a video frame is $R_{\underline{V}|Y}(D)$ in Eq. (5.3.11). It is the theoretical rate distortion bound that is solely based on the proposed correlation model of the video source and takes advantage of the universal side information on the local texture. $R_{\underline{V}|Y}(D)$ should always be lower than $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ according to the data processing theorem [11].

In Fig. 5.16 we plot $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$, the rate distortion bound calculated based on the new texture dependent correlation model for the scenario where optimal predictive coding is engaged to code $\underline{S}$, $\underline{Z}$ and A separately with separate distortion constraint, as a dash dotted line. As shown in this figure, $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ is a reasonably tight lower bound to the operationally rate distortion curve of AVC/H.264, especially at medium to high distortion levels.

In Fig. 5.24(a) we also plot the lower bound $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ (Eq. (5.4.26)) and the operational rate distortion function using AVC/H.264 for two other videos. We can see that although the lower bounds are calculated based on only five parameters generated from each video, they do agree with the operational rate distortion curves of the corresponding video reasonably well. If we further plot these lower bounds as average rate per pixel versus PSNR of a video frame as in Fig. 5.24(b), the lower bounds appear to be nearly, linear which shows promises in codec design.



(a) average rate vs. average distortion



(b) average rate vs. PSNR

Figure 5.24: The lower bounds calculated based on the new correlation coefficient model and its corresponding optimal parameters for three videos, compared to the operational rate distortion curves of these videos coded in AVC/H.264

5.5 Conclusion

In this chapter we revisit the classic problem of developing a correlation model for natural videos and studying their theoretical rate distortion bounds. We propose the correlation coefficient of two pixels in two nearby video frames as the product of the spatial correlation coefficient of these two pixels, as if they were in the same frame, and a variable to quantify the temporal correlation between these two video frames. The spatial correlation model for pixels within one video frame is a conditional correlation model. The conditioning is on local texture and the optimal parameters can be calculated for a specific video with a mean absolute error (MAE) usually smaller than 5%. We use this conditional correlation model to calculate the conditional rate distortion function when universal side information on local texture is available at both the encoder and the decoder. This rate distortion bound with local texture information taken into account while making no assumptions on coding, is shown indeed to be a valid lower bound, and the only valid theoretical rate distortion bound to our best knowledge, with respect to the operational rate distortion curves of both intra-frame and inter-frame coding in AVC/H.264 and in HEVC/H.265. A 50% or higher bit rate reduction from the operational rate distortion curve of HEVC to the new theoretical rate distortion bound can potentially be achieved across the whole range of average distortion typically encountered in video coding.

References

[1] 3GPP. Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions. TS 26.090, 3rd Generation Partnership Project (3GPP), Mar. 2011.

[2] 3GPP. Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions. TS 26.190, 3rd Generation Partnership Project (3GPP), Mar. 2011.

[3] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustic Society of America*, 50:637–655, 1971.

[4] B. S. Atal and M. R. Schroeder. Adaptive predictive coding of speech signals. *The Bell System technical journal*, pages 1973–1986, 1970.

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

[6] T. Berger. *Rate Distortion Theory*. Prentice-Hall, 1971.

[7] T. Berger and J. D. Gibson. Lossy Source Coding. *IEEE Trans. on Information Theory*, 44(6):2693 –2723, Oct. 1998.

[8] M.J. Carter. *Soource coding of composite sources*. PhD thesis, The University of Michigan, 1984.

[9] Tihao Chiang and Ya-Qin Zhang. A new rate control scheme using quadratic rate distortion model. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(1):246–251, Feb. 1997.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, Aug. 1991.

[11] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, 1991.

[12] Richard V Cox, Simao Ferraz De Campos Neto, Claude Lamblin, and Mostafa Hashem Sherif. Itu-t coders for wideband, superwideband, and fullband speech communication [series editorial]. *Communications Magazine, IEEE*, 47(10):106–109, 2009.

[13] D.-K. Kwon, M.-Y. Shen and C.-C. J. Kuo. Rate control for H.264 video with enhanced rate and distortion models. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(5):517–529, May 2007.

[14] W. Daumer. Subjective Evaluation of Several Efficient Speech Coders. *IEEE Trans. on Communications*, 30(4):655 – 662, Apr. 1982.

[15] L. D. Davisson. Rate-distortion theory and application. *Proceedings of the IEEE*, 60(7):800 – 808, July 1972.

[16] A. De and P. Kabal. Rate-distortion function for speech coding based on perceptual distortion measure. *IEEE Global Telecommunications Conference*, pages 452–456, Orlando, Dec. 1992.

[17] R.J. Fontana. *A class of composite sources and their erfodic and information theoretic properties*. PhD thesis, Standford University, 1978.

[18] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, 1968.

[19] S.A. Garde. *Communication of composite sources*. PhD thesis, Univeristy of California, Berkeley, 1980.

[20] J. D. Gibson. Speech Coding Methods, Standards, and Applications. *IEEE Circuits and Systems Magazine*, 5(4):30 – 49, Fouth Quarter 2005.

[21] J. D. Gibson, J. Hu, and P. Ramadas. New Rate Distortion Bounds for Speech Coding Based on Composite Source Models. *Information Theory and Applications Workshop (ITA)*, UCSD, Jan. 31 - Feb. 5, 2010.

[22] J. D. Gibson and Y.-Y. Li. Rate Distortion Performance Bounds for Wideband Speech. *Information Theory and Applications Workshop*, San Diego, CA, Feb. 5-10, 2012.

[23] Jerry D. Gibson, Toby Berger, Tom Lookabaugh, Dave Lindbergh, and Richard L. Baker. *Digital compression for multimedia: principles and standards*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1998.

[24] Bernd Girod. The efficiency of motion-compensating prediction for hybrid coding of video sequences. *IEEE Journal on selected areas in communications*, SAC-5(7):1140–1154, Aug. 1987.

[25] Bernd Girod. Motion-compensating prediction with fractional-pel accuracy. *IEEE Transactions on Communications*, 41:604–612, Apr. 1993.

[26] R. M. Gray. Information rates of autoregressive processes. *IEEE Trans. on Information Theory*, 16(4):412 – 421, Jul. 1970.

[27] R. M. Gray. A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions. *IEEE Trans. on Information Theory*, IT-19(4):480–489, July 1973.

[28] R. M. Gray. A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions. *IEEE Tran. Inform. Theory*, IT-19(4):480–489, Jul. 1973.

[29] H. Brehm and K. Trottler. Rate distortion functions for speech-model signals. *Signal Processing III: Theories and Applications*, pages 353–356, EURASIP, 1986.

[30] Ali Habibi and Paul A. Wintz. Image coding by linear transformation and block quantization. *IEEE Transactions on Communication Technology*, Com-19(1):50–62, Feb. 1971.

[31] J. Hu and J. D. Gibson. New rate distortion bounds for natural videos based on a texture dependent correlation model in the spatial-temporal domain. *the 46th Annual Allerton Conference on Communication, Controls, and Computing*, Sept. 2008.

[32] ISO/IEC 13818-1:2000. Information technology – generic coding of moving pictures and associated audio information: Systems. 2000.

[33] ISO/IEC 14496-1:2001. Information technology – coding of audio-visual objects – part 1: Systems. 2001.

[34] ITU Recommendations. Video coding for low bit rate communication. *ITU-T rec. H.263*, Jan. 2005.

[35] ITU-T and ISO/IEC JTC 1. Advanced video coding for generic audiovisual services. 2003.

[36] ITU-T and ISO/IEC JTC 1. H.265 : High efficiency video coding. *http://www.itu.int/rec/T-REC-H.265-201304-I*, Apr. 2013.

[37] ITU-T Recommendation G.191. Software tools for speech and audio coding standardization. Mar. 2010.

[38] ITU-T Recommendation G.718. Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s. June 2008.

[39] ITU-T Recommendation G.719, Series G. Transmission systems and media, digital systems and networks, Digital terminal equipments-Coding of analogue signals by pulse code modulation. June 2008.

[40] ITU-T Recommendation G.722. 7 kHz Audio-Coding within 64 kbits/s . Nov. 1988.

[41] ITU-T Recommendation G.722.1. Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss . May 2005.

[42] ITU-T Recommendation G.722.2. Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB). July 2003.

[43] ITU-T Recommendation G.726. 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM) . Dec. 1990.

[44] ITU-T Recommendation G.727. 5-, 4-, 3- and 2-bit/sample embedded Adaptive Differential Pulse Code Modulation (ADPCM). Dec. 1990.

[45] ITU-T Recommendation G.729. Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP) . Jan. 2007.

[46] ITU-T Recommendation P.56. Objective measurement of active speech level. Mar. 1993.

[47] ITU-T Recommendation P.830. Subjective performance assessment of telephone-band and wideband digital codecs. Feb. 1996.

[48] ITU-T Recommendation P.862. Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end Speech Quality Assessment of Narrow-band telephone networks and Speech Codecs. Feb. 2001.

[49] ITU-T Recommendation P.862. Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end Speech Quality Assessment of Narrow-band telephone networks and Speech Codecs. Feb. 2001.

[50] ITU-T Recommendation P.862.2. Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs. Nov. 2007.

[51] ITU-T Recommendation P.862.3. Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2. Nov. 2007.

[52] J. B. O'neal Jr. and T. Raj Natarajan. Coding isotropic images. *IEEE Transactions on Information Theory*, IT-23(6):697–707, Nov. 1977.

[53] N. S. Jayant and P. Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, Mar. 1984.

[54] K. Järvinen, I. Bouazizi, L. Laaksonen, P. Ojala, and A. Rämö. Media coding for the next generation mobile system lte. *Computer Communications*, 33(16):1916 – 1927, 2010.

[55] H. Kalveram and P. Meissner. Itakura-saito clustering and rate distortion functions for a composite source model of speech. *Signal Processing*, 18(2):195 – 216, 1989.

[56] H. Kalveram and P. Meissner. Rate Distortion Bounds for Speech Waveforms based on Itakura-Saito-Segmentation. *Signal Processing IV: Theories and Applications*, EURASIP, 1988.

[57] W Bastiaan Kleijn and Kuldip K Paliwal. *Speech coding and synthesis*. Elsevier Science Inc., 1995.

[58] A Kolmogorov. On the shannon theory of information transmission in the case of continuous signals. *IEEE Transactions on Information Theory*, 2(4):102–108, 1956.

[59] A. M. Kondoz. *Digital Speech: Coding for Low Bit Rate Communication Systems*. John Wiley & Sons, West Sussex, England, 2004.

[60] Ahmet M Kondoz. *Digital speech: coding for low bit rate communication systems*. Wiley. com, 2005.

[61] Hung-Ju Lee, Tihao Chiang, and Ya-Qin Zhang. Scalable rate control for MPEG-4 video. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(6):878–894, Sep. 2000.

[62] Q. Li and M. van der Schaar. Providing adaptive QoS to layered video over wireless local area networks through real-time retry limit adaptation. *IEEE Transactions on Multimedia*, 6(2):278–290, Apr. 2004.

[63] A. K. Luthra, G. J. Sullivan, and T. Wiegand. Introduction to the special issue on the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7), Jul. 2003.

[64] et al M. Jelinek. On the architecture of the cdma2000ő variable-rate multimode wideband (VMR-WB) speech coding standard. *Proc. ICASSP*, pages I–281–I–284, 2004.

[65] Siwei Ma, Wen Gao, and Yan Lu. Rate control on JVT standard. *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, JVT-D030*, Jul. 2002.

[66] Jens-Rainer Ohm and Gary J. Sullivan. High efficiency video coding: The next frontier in video compression. *IEEE Signal Processing Magazine*, 30(1):152–158, Jan. 2013.

[67] A. Ortega and K. Ramchandran. Rate-distortion methods for image and video compression. *IEEE Signal Processing Magazine*, 15(6):23Ű50, Nov. 1998.

[68] P. Cummiskey, N. S. Jayant, and J. L. Flanagan. Adaptive Quantization in Differential PCM Coding of Speech. *The Bell System technical journal*, 52:1105–1118, Sept. 1973.

[69] T. N. Pappas and R. J. Safranek. Perceptual criteria for image quality evaluation. *Handbook of Image & Video Processing (A. Bivok eds.), Academic Press*, 2000.

[70] Mark S Pinsker. Mutual information between a pair of stationary gaussian random processes. In *Dokl. Akad. Nauk. USSR*, volume 99, pages 213–216, 1954.

[71] Quackenbush. MPEG unified speech and audio coding. *IEEE Multimedia*, 20(2):72–78, April-June 2013.

[72] R. C. Reininger and J. D. Gibson. Distributions of the two-dimensional DCT coefficients for images. *IEEE Transactions on Communications*, 31:835–839, Jun. 1983.

[73] P. Ramadas and J. D. Gibson. Phonetically Switched Tree coding of speech with a G.727 Code Generator. *the 43rd Annual Asilomar Conference on Signals, Systems, and Computers*, Nov. 1-4, 2009.

[74] Jordi Ribas-Corbera and Shawmin Lei. Rate control in DCT video coding for low-delay communications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(1):172–185, Feb. 1999.

[75] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27:379–423, 1948.

[76] C. E. Shannon. Coding Theorems for a Discrete Source with a Fidelity Criterion. *IRE Conv. Rec.*, 7:142–163, 1959.

[77] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, 4:142–163, 1959.

[78] S. R. Smoot and L. A. Rowe. Study of DCT coefficient distributions. *SPIE Symposium on Electronic Imaging, San Jose, CA*, 2657, Jan. 1996.

[79] Klaus Stuhlmuller, Niko Farber, Michael Link, and Bernd Girod. Analysis of video transmission over lossy channels. *IEEE Journal on Selected Areas in Communications*, 18(6), Jun. 2000.

[80] T. Aach, C. Mota, I. Stuke, M. Mhlich, and E. Barth. Analysis of superimposed oriented patterns. *IEEE Transactions on Image Processing*, 15(12):3690–3700, Dec. 2006.

[81] G. Tziritas. Rate distortion theory for image and video coding. *International Conference on Digital Signal Processing, Cyprus*, 1995.

[82] Sergio Verdu, Venkat Anantharam, Giuseppe Caire, Max Costa, Gerhard Kramer, and Raymond Yeung. Panel on new perspectives for information theory. *Information Theory Workshop, Paraty, Brazil*, Oct. 2011.

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

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

[85] M.S. Wallace. Some techniques in universal source coding and coding for composite sources. Master's thesis, University of Illinois at Urbana-Champaign, 1982.

[86] S. Wang and A. Gersho. Improved Phonetically- Segmented Vector Excitation Coding at 3.4 Kb/s. In *Proceedings, IEEE ICASSP*, San Francisco, Mar. 1992.

[87] S. Wang and A. Gersho. Phonetically-based vector excitation coding of speech at 3.6 kbit/s. *Proceedings, IEEE ICASSP*, pages 49–52, Glasgow, May 1989.

[88] Z. Wang, H. R. Sheikh, and A. C. Bovik. Objective video quality assessment. *The Handbook of Video Databases: Design and Applications (B. Furht and O. Marqure, eds.), CRC Press*, pages 1041–1078, Sep. 2003.

[89] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and Ajay Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13:560–576, Jul. 2003.

[90] Y. K. Tu, J.-F. Yang and M.-T. Sun. Rate-distortion modeling for efficient H.264/AVC encoding. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(5):530–543, May 2007.