RATE DISTORTION BOUNDS FOR SPEECH CODING BASED ON A PERCEPTUAL DISTORTION MEASURE (PESQ-MOS)

Ying-Yi Li and Jerry D. Gibson

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106, USA
Email: yingyi_li@umail.ucsb.edu, gibson@ece.ucsb.edu

ABSTRACT

We develop practical rate distortion bounds for speech coding based on composite source models and the PESQ-MOS distortion measure. Specifically, the bounds are formulated using composite source models for speech, the rate distortion function for Gaussian autoregressive sources, the classical reverse water-filling result, and conditional rate distortion theory, along with a recently devised MSE-to-PESQ-MOS mapping. The resulting rate distortion bounds are shown to lower bound the performance of the AMR, G.729, and G.718 standardized codecs, and based on the tightness of these bounds, to indicate how the performance of voice codecs might be improved.

Index Terms— Speech coding, Rate distortion bounds, Speech codec performance

1. INTRODUCTION

Speech coding plays a significant role in digital cellular, Voice over IP (VoIP), and Voice over Wireless LAN (VoWLAN) applications, and extraordinary progress has been made in developing standardized speech codecs for these applications. In order to evaluate these codecs, it would be very useful if meaningful rate distortion bounds on the performance of speech codecs were available, thus helping to guide future directions in speech coding research.

In particular, it would be of great utility if the past 50 years of rate distortion theory results could be applied to bounding the performance of practical codecs. Gallager, in his classic text on Information Theory [1], summarizes the challenges in doing so when 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 ..." [1].

This research has been supported by NSF Grant Nos. CCF-0728646 and CCF-0917230.

In this paper, we utilize recently formulated composite source models for speech [2], the classical water-filling result for each composite subsource [3, 4], conditional rate distortion theory [5], and results on mapping MSE to PESQ-MOS, to obtain true rate distortion bounds for voice codecs subject to a perceptually meaningful distortion measure. These rate distortion bounds are shown to lower bound the best known standardized voice codecs, including AMR-NB, G.718, and G.729, thus revealing the limitations of the voice codecs for different speech sources and indicating where voice codec performance can be improved.

The paper is organized as follows. Some relevant prior work is described in Section 2. Section 3 provides some needed background on rate distortion functions for autoregressive sources, and a brief description of reverse water-filling is given in Section 4. Specifics concerning the composite source models are described in Section 5, and conditional rate distortion functions based on MSE are briefly explained in Section 6. The standardized PESQ-MOS, and the steps performed to generate the MSE-to-PESQ-MOS mapping function are given in Section 7. Rate distortion bounds based on the PESQ-MOS distortion measures are contained in Section 8, and comparisons of the performance of standardized codecs to these bounds are given. Conclusions are presented in Section 9, wherein the contributions of the current work are summarized.

2. RELEVANT PRIOR WORK

In [6], composite source models for speech are obtained by Itakura-Saito segmentation of the speech into equal order autoregressive subsources, and by calculating lower bounds to the rate distortion function for different numbers of subsources, it is shown that a relatively small number of subsources (6 in the cited paper) is needed to have a good composite source model for speech. No comparisons to standardized speech codecs are given.

A cochlear model serves as the basis for a perceptual distortion measure for speech in [7], and the cochlear models are used to characterize the rate distortion function for speech and to compare to the operational rate distortion performance of common voice codecs. Among the interesting results are that

the Shannon lower bound for this 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.

Gibson, Hu, and Ramadas [2] obtained rate distortion bounds for speech coding based on composite source models and unweighted and weighted mean squared error (MSE) distortion measures. The composite source models are constructed by classifying each sentence as Voice (V), Unvoiced (UV), Onset (ON), Hangover (H), and Silence (S). The V, ON, and H modes are modeled as autoregressive with different orders, and the UV mode is modeled as uncorrelated. Unfortunately, the performance of code excited linear prediction (CELP) codecs, such as G.729 and AMR-NB, is not accurately represented by unweighted MSE and useful weighted MSE distortion measures were not found. In the current work, we are able to develop composite source models following [2] and combine them with a perceptual PESQ-MOS distortion measure to obtain valid, meaningful bounds on speech codec performance.

3. RATE DISTORTION BACKGROUND

A natural starting point for the development of fundamental limits for speech coding is Shannon's rate distortion theory [8], a historical discussion of which is presented by Berger and Gibson [9]. 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 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 [3] and Gray [10], 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, as summarized in the following theorem:

Theorem 3.1 *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*

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

and

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

where

$$g(\omega) = \frac{1}{\sigma^2}\left|1 + \sum_{k=1}^{m} a_k e^{-jk\omega}\right|^2. \tag{3}$$

A limitation of these results is that the distortion measure is MSE and the source model is assumed to be known, even the predictor coefficients, which are actually changing frame-to-frame, so rate distortion bounds calculated using coefficients averaged over an entire sentence do not provide useful bounds. As a result, we model the input speech to be compressed as a composite source consisting of subsources with different characteristics and that occur with some probability. Further we incorporate a perceptual distortion measure that allows easy evaluation of standardized speech codecs.

Our work here is motivated by the prior work on rate distortion bounds for video by Hu and Gibson [11, 12, 13, 14].

4. REVERSE WATER-FILLING

To calculate rate distortion functions for the subsources in the composite source model, we use the squared error fidelity criterion and the classic eigenvalue decomposition [4] and reverse water-filling approach [15]. This standard result is given in the theorem below [16].

Theorem 4.1 *Rate Distortion Function for Parallel Gaussian Sources*

Let $X_i \sim N(0,\sigma_i^2), i = 1, 2, ..., N$, be independent Gaussian random variables and let the distortion measure be

$$D(x^N, \hat{x}^N) = \sum_{i=1}^{N} (x_i - \hat{x_i})^2. \tag{4}$$

Then the rate distortion function is

$$R(D) = \sum_{i=1}^{N} \frac{1}{2}\log\frac{\sigma_i^2}{D_i}, \tag{5}$$

where

$$D_i = \begin{cases} \lambda & \text{if} \quad \lambda < \sigma_i^2 \\ \sigma_i^2 & \text{if} \quad \lambda \geq \sigma_i^2 \end{cases}. \tag{6}$$

In the following, we view each of the parallel Gaussian sources as decompositions in the frequency domain. The main departure of the current work from prior efforts to calculate rate distortion functions for AR sources and speech is that we utilize a perceptual distortion measure, and we employ reverse water-filling for each of the identified modes of the composite source model and then combine the resulting rate distortion functions using conditional rate distortion theory.

5. 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 [3]. Multimodal models have played a major role in speech coding, including the voiced/unvoiced decision for the excitation in linear predictive coding (LPC) [17] and the long-term adaptive predictor in adaptive predictive coding (APC) [18]. Further, phonetic classification of the input speech into multiple modes and coding each mode differently has lead to some outstanding voice codec designs [19, 20].

Ramadas and Gibson's [21] work on speech coding has built on these prior contributions and they have developed a mode classification method that breaks the input speech into Voiced (V), Onset (ON), Hangover (H), Unvoiced (UV), and

Table 1. Composite Source Models for Speech Sentences

Sequence (Gender) (Active speech level)	Mode	Autocorrelation coefficients for V, ON, H Average frame energy for UV	Mean Square Prediction Error	Probability
"lathe" (Female) (-18.1 dBov)	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) (-16.5 dBov)	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

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. We model Voiced speech as a 10^{th} order AR Gaussian source, Onset as a 4^{th} order AR Gaussian source, Hangover as a 4^{th} order AR Gaussian source, Unvoiced speech as a memoryless Gaussian source, and silence is treated by sending a code for comfort noise generation. In particular, Table 1 presents the autocorrelation values and mean squared prediction error for the several modes for two sentences. The probability of each mode is also shown in Table 1. For example, the sentence, "We were away" has 98.42% classified as Voiced. We have calculated similar data for many speech utterances, but only these two are presented here due to space limitations.

There are a few things to note about the data in Table 1. 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" has only 1.05% classified as Silence, while the sentence, "Lathe" has 36.85% classified as Silence. These Silence sections 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 in the decoded stream.

Further work on developing appropriate composite models for speech is underway to optimize the phonetic classification of the modes, the AR model order for the Voiced, Onset, and Hangover modes, and to investigate alternative models for the Onset and Hangover modes. Since these operations are done off-line and only once per utterance, complexity is not a major issue.

6. CONDITIONAL RATE DISTORTION FUNCTIONS BASED ON MSE

Given the composite source models from Section 5, a rate distortion bound based on MSE [2] is derived using the conditional rate distortion results from Gray [5]. The conditional rate distortion function of a source $\underline{X}$ with side information Y is defined as

$$R_{\underline{X}|Y}(D) = \min_{p(\underline{\hat{x}}|\underline{x},y):D(\underline{X},\underline{\hat{X}}|Y)\leq D} I(\underline{X};\underline{\hat{X}}|Y), \qquad (7)$$

where

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

$$I(\underline{X};\underline{\hat{X}}|Y) = \sum_{\underline{x},\underline{\hat{x}},y} p(\underline{x},\underline{\hat{x}},y) \log \frac{p(\underline{x},\underline{\hat{x}}|y)}{p(\underline{x}|y)p(\underline{\hat{x}}|y)}. \qquad (8)$$

It can be proved [5] that the conditional rate distortion function in Eq. (7) can also be expressed as

$$R_{\underline{X}|Y}(D) = \min_{D'_y s:D(\underline{X},\underline{\hat{X}}|Y)=\sum_y D_y p(y)\leq D} \sum_y R_{\underline{X}|y}(D_y)p(y), \qquad (9)$$

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. (9), 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 $\sum_y D_y P[Y=y] = D$.

7. MAPPING MSE TO PESQ-MOS

Perceptual evaluation of speech quality (PESQ) [22] is a standardized objective method for end-to-end speech quality assessment of narrow-band speech codecs. Therefore, the PESQ-MOS of each speech codec is easy to measure, and a rate distortion bound based on PESQ-MOS is particularly useful.

The distance between the original and degraded speech signal, PESQ score, is calculated based on the PESQ perceptual model. The PESQ score is 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. 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.

ADPCM coders are waveform coders. Thus, MSE is an

indicator of how well these codecs reproduce the input speech waveform, and it is also useful in establishing the relative ordering of the performance of ADPCM speech coders. In addition, the PESQ-MOS of ADPCM coders can be evaluated easily, thus providing a perceptual distortion that can be aligned with the MSE achieved by each codec at the given rate for the selected input utterance.

G.726 [23, 24] and G.727 [25, 24] are standardized ADPCM speech coders. These codecs have four selectable transmitted bit rates of 40, 32, 24, and 16 kbps. Since G.727 is an embedded coder, ITU-T G.727 Recommendation [25] provides coding rates of 40 kbps for the 3 combinations, 32 kbps for 3 combinations, 24 kbps for 2 combinations, and 16 kbps for one combination, resulting in 9 pairs of coding rates. Therefore with the 4 coding rates for G.726 and the 9 coding rates for G.727, there are 13 MSE and PESQ pairs to generate a mapping function for each sentence.

For each speech sequence, the MSE of each coded sequence is calculated and normalized by the average energy of the original sequence. The PESQ-MOS of each coded sequence is evaluated by the software provided by ITU-T Recommendation P.862 [22], and 13 pairs of MSE and PESQ are used for curve fitting for each sequence. Since MSE is increasing and PESQ is decreasing as the bit rate is reduced, an exponential function is chosen as the mapping function since it provides a good fit across all rates and distortion pairs. The range of PESQ-MOS is between -0.5 and 4.5 [22], so the PESQ-MOS is 4.5 when MSE is 0. Therefore, the mapping function is modeled as

$$z = f(w) = ae^{-bw} + 4.5 - a, \tag{10}$$

where w is MSE, z is PESQ-MOS, and a and b are estimated by the least squares fit of the 13 MSE and PESQ pairs of G.726 and G.727.

Fig. 1 is an example of the mapping function generated by ADPCM waveform codecs. The MSE/PESQ-MOS pairs are fitted with the exponential mapping function, and Fig. 1 shows that the exponential function provides a good fit to the MSE/PESQ-MOS pairs.

8. RATE DISTORTION BOUNDS FOR SPEECH BASED ON PESQ-MOS

The rate distortion bound using MSE as distortion measure is calculated by the classical eigenvalue decomposition [4] and reverse water-filling approach described in Section 4 with the composite source models presented in Section 5. Then the rate distortion bound based on MSE is mapped to PESQ-MOS measure by the mapping function generated by ADPCM waveform codecs as described in Section 7.

The details of each test sequence are also shown in Table 1. The active speech level of each sequence is computed based on ITU-T P.56 [26, 24]. ITU-T Recommendation P.830 [27] mentions that the nominal value for mean active speech level is -26 dBov, and that the active speech level should be observed during recording. Therefore, the active speech level of the test sequences we used is greater than -26 dBov.

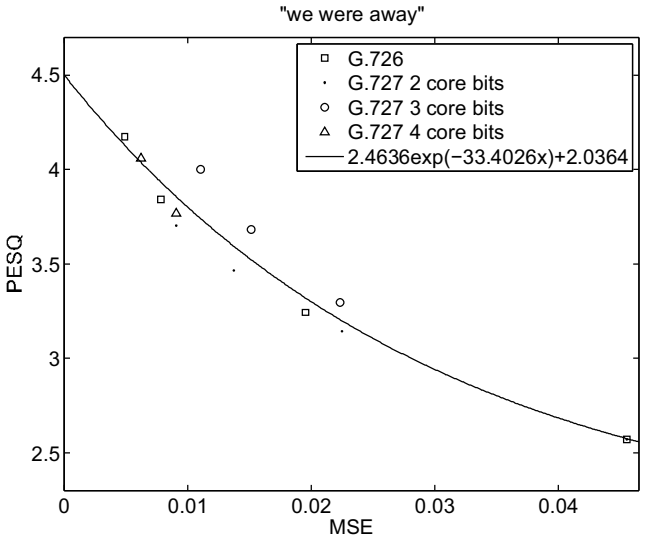

Fig. 1. The mapping function of sequence "We were away a year ago."

The rate distortion bounds for each of the composite source modes of the two sequences are shown in Figs. 2 and 3. It is interesting to see that the rate distortion functions for the modes differ across the two sentences. It is also interesting to note the very profound effect of the probabilities of the different modes. 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. In Fig. 3, since the sequence is 98.42% Voiced, the conditional rate distortion function is dominated by the marginal rate distortion function of the voiced mode. Since there is 36.85% Silence in the sequence "A lathe is a big tool," the final conditional rate distortion function is lower than all of the marginal rate distortion functions.

The rate distortion bounds based on PESQ-MOS are compared with CELP codecs such as AMR-NB, G.729, and G.718 [28], and ADPCM coders, G.726 and G.727 in Figs. 4 and 5. 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. Since G.727 is an embedded speech codec, codecs with 2 core bits are used in our experiments. The PESQs of all speech codecs are computed by ITU-T P.862 [22, 24].

From Figs. 4 and 5, we see that the performance of all codecs is lower bounded by the rate distortion bounds. In addition, CELP codecs such as AMR-NB, G.729, and G.718 are much closer to the rate distortion bounds than ADPCM coders, which fits our intuition. Since G.727 is an embedded

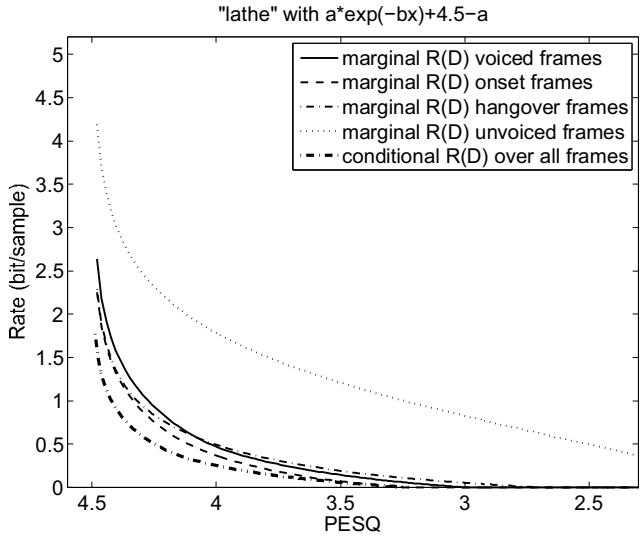

Fig. 2. The marginal and conditional rate distortion bounds based on PESQ of the sequence "A lathe is a big tool."

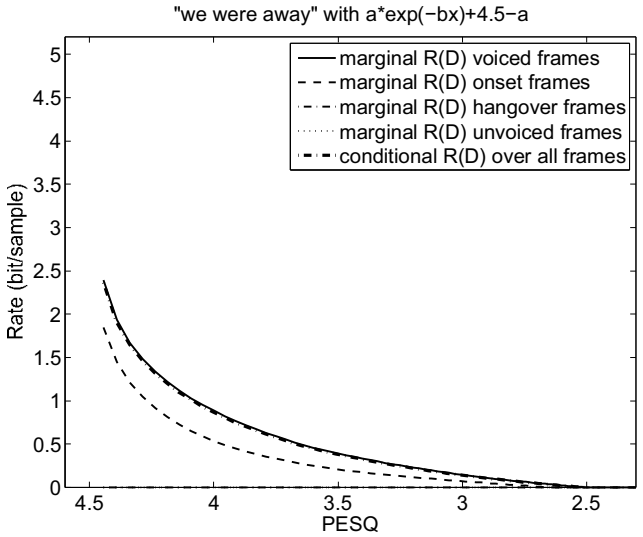

Fig. 3. The marginal and conditional rate distortion bounds based on PESQ of the sequence "We were away a year ago."

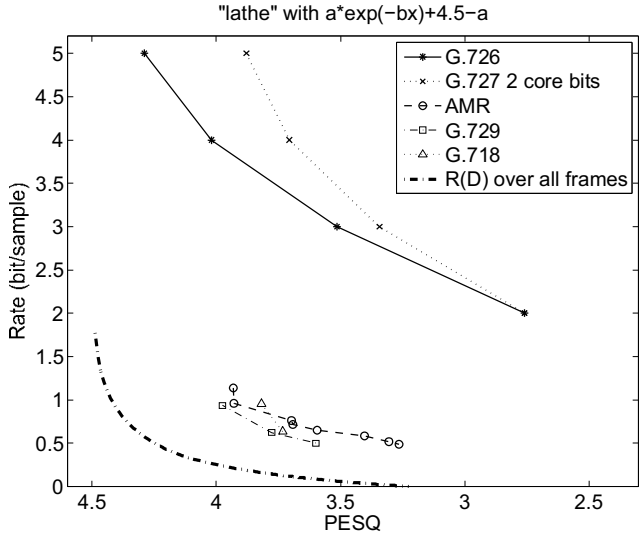

Fig. 4. The rate distortion bounds and operational rate distortion performance of speech codecs of the sequence "A lathe is a big tool."

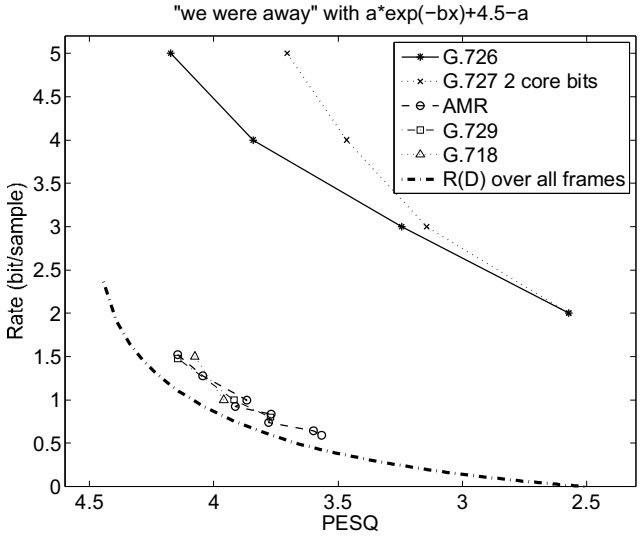

Fig. 5. The rate distortion bounds and operational rate distortion performance of speech codecs of the sequence "We were away a year ago."

ADPCM coder, the performance of G.727 with 2 core bits is worse than that of G.726. The performance 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 of 4.0 or better.

The performance of the codecs for the utterance "We were away a year ago" is closer to the rate distortion bound than 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 10th 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.

These results show that our new rate distortion bounds do lower bound the PESQ-MOS performance of the best known standardized speech codecs. However, there is room to improve the bounds by better mode selection and better modeling of the modes. This is the subject of on-going work. However, 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.

9. CONCLUSIONS

We present practical rate distortion bounds for speech coding based on composite source models and the PESQ-MOS distortion measure. Comparisons of the rate versus PESQ-MOS performance of standardized CELP codecs such as AMR-NB, G.729, G.718, and ADPCM coders, G.726 and G.727, show that the new rate distortion bound developed here does in fact lower bound the PESQ-MOS performance of the best known standardized speech codecs. Further, because of the decomposition of the speech into various source modes, it is suggested by the tightness of the rate distortion bounds how the performance of existing codecs might be improved.

10. REFERENCES

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

[2] 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.

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

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

[5] 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*, vol. IT-19, no. 4, pp. 480–489, July 1973.

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

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

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

[9] T. Berger and J. D. Gibson, "Lossy Source Coding," *IEEE Trans. on Information Theory*, vol. 44, no. 6, pp. 2693 –2723, Oct. 1998.

[10] R. M. Gray, "Information rates of autoregressive processes," *IEEE Trans. on Information Theory*, vol. 16, no. 4, pp. 412 – 421, Jul. 1970.

[11] J. Hu and J. D. Gibson, "New Block-Based Local-Texture-Dependent Correlation Model of Digitized Natural Video," *the 40th Annual Asilomar Conference on Signals, Systems and Computers*, Oct. 29-Nov. 1 2006.

[12] J. Hu and J. D. Gibson, "New Rate Distortion Bounds for Natural Videos Based on a Texture Dependent Correlation Model," *IEEE International Symposium on Information Theory*, June 2007.

[13] 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.

[14] J. Hu and J. D. Gibson, "New Rate Distortion Bounds for Natural Videos Based on a Texture-Dependent Correlation Model," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 19, no. 8, pp. 1081–1094, Aug. 2009.

[15] R. A. McDonald and P. M. Schultheiss, "Information rates of gaussian signals under criteria constraining the error spectrum," *Proceedings of the IEEE*, vol. 52, pp. 415–416, Apr. 1964.

[16] L. D. Davisson, "Rate-distortion theory and application," *Proceedings of the IEEE*, vol. 60, no. 7, pp. 800 – 808, July 1972.

[17] 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*, vol. 50, pp. 637–655, 1971.

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

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

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

[21] 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.

[22] 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.

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

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

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

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

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

[28] 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.