NEW RATE DISTORTION BOUNDS FOR SPEECH CODING BASED ON COMPOSITE SOURCE MODELS

Jerry D. Gibson*, Jing Hu†, and Pravin Ramadas*
*Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA
†Digital Signal processing Group, Cisco systems
Email: gibson@ece.ucsb.edu, jinghu@cisco.com, pravin_ramadas@umail.ucsb.edu

***Abstract*—We present new rate distortion bounds for speech coding based upon new composite source models for speech and conditional rate distortion theory. The composite source models are constructed by classifying each sentence as Voiced (V), Unvoiced (UV), Onset (ON), Hangover (H), and Silence (S). A 10^{th} order AR model is used for the V mode, 4^{th} order AR models are used for the ON and H modes, and the UV mode is modeled as uncorrelated. Marginal rate distortion functions are computed for each mode and combined to produce conditional rate distortion bounds based on unweighted and weighted mean squared error distortion measures. For unweighted distortion measures, the new bounds imply that good performance is attainable at rates as low as 0.25 bits/sample for narrowband speech.**

***Index Terms*—rate distortion bounds for speech, composite source models for speech, speech codec performance**

I. INTRODUCTION

Speech coding is an essential part of efficient human communications today, playing a dominant role in voice over IP (VoIP), digital cellular systems, and videoconferencing. As standards bodies continue to improve performance and extend the functionalities of voice codecs for a myriad of applications [1], it is important to examine fundamental limits on the performance of speech coding. In this paper, we present new rate distortion bounds for speech coding based upon new composite source models and conditional rate distortion theory. We describe the development of the new composite models, briefly outline the relevant conditional rate distortion theory, and compare the performance of existing high performance voice codecs to the new rate distortion bounds based on the unweighted squared error distortion measure. We also develop bounds based on the weighted mean squared error distortion measure, and indicate additional work needed to perform meaningful comparisons to CELP speech codes. We begin with some background on rate distortion theory, speech coding, and prior work.

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

II. BACKGROUND AND PRIOR WORK

A natural starting point for the development of fundamental limits for speech coding is Shannon's rate distortion theory [2], a historical discussion of which is presented by Berger and Gibson [3]. 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 [4] and Gray [5], in separate contributions in the late 60's and early 70's, derived the rate distortion function for Gaussian autoregressive (AR) sources for both the squared error and weighted squared error distortion measures. Since the linear prediction (LP) model was just starting to have a significant impact on the design of efficient voice coders, these were promising results indeed [6–8]. However, a limitation of these results is that the source model is assumed to be known, even the predictor coefficients, which are actually changing frame-to-frame, so that voice codecs based on the linear prediction model must calculate these coefficients in either a backward adaptive or forward adaptive fashion, the latter of which requires that the predictor coefficients (or related parameters) be quantized and coded for transmission to the decoder. Rate distortion bounds that incorporate the AR coefficients by modeling them as correlated Gaussian sources was presented by Jones and Gupta [9] and asymptotic rate distortion bounds including a bound on coding the coefficients and the estimation error are derived by Gibson [10].

In [11], composite source models for speech are obtained by Itakura-Saito segmentation of the speech into 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. A cochlear model is used as the basis for a perceptual distortion measure for speech in [12], 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

TABLE I
COMPOSITE SOURCE MODELS FOR SPEECH SENTENCES

Sequence	Mode	Autocorrelation coefficients for V, ON, H and Average Frame Energy for UV	Mean Square Prediction Error for V, ON, H and Average Frame Energy for UV	Probability of each mode
"We were away" (male)	V	[1 0.8057 0.5197 0.2569 0.02730 -0.1497 -0.2392 -0.3537 -0.4467 -0.4686 -0.4330]	0.0935	0.3415
	ON	[1 0.8479 0.5463 0.2383 -0.0339]	0.0851	0.0123
	H	[1 0.9394 0.7793 0.5559 0.3131]	0.0008775	0.0154
	UV	0.4177	0.4177	0.16
"Lathe" (female)	V	[1 0.8271 0.5711 0.3543 0.1489 0.01977 -0.0569 -0.0797 -0.09609 -0.1637 -0.2701]	0.0699	0.4211
	ON	[1 0.7601 0.4639 0.3537 0.2739]	0.010647	0.0246
	H	[1 0.087 0.3478 0.087 0.0435]	0.000000004769	0.0188
	UV	0.0693	0.0693	0.1135

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.

Our work here is motivated by and draws heavily on the approach to calculating rate distortion bounds for video as presented by Hu and Gibson [13–16].

III. 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 [17] and reverse water-filling approach [18]. The standard result with a minor modification to accommodate weighted distortion measures is given in the theorem below [19].

Theorem 1. *Rate distortion Function for a Parallel Gaussian Source*

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

$$D(x^N, \hat{x}^N) = \sum_{i=1}^{N} W_i (x_i - \hat{x}_i)^2$$

Then the rate distortion function is

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

where

$$D_i = \begin{cases} \lambda & \text{if } \lambda < W_i \sigma_i^2 \\ W_i \sigma_i^2 & \text{if } \lambda \geq W_i \sigma_i^2 \end{cases}$$

In the following, we will view each of the parallel Gaussian sources as decompositions in the frequency domain, and the weights are adjusted to represent the perceptual weighting as appropriate.

The main departure of the current work from prior efforts to calculate rate distortion functions for autoregressive sources and speech is that we employ reverse water-filling for each of the identified modes of the composite source model for speech and combine the resulting rate distortion functions using results from conditional rate distortion theory.

IV. 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 [4]. Multimodal models have played a major role in speech coding, beginning with the voiced/unvoiced decision for the excitation in linear predictive coding (LPC) [7] and the long-term adaptive predictor in adaptive predictive coding (APC) [6]. Other modes and classification methods have been investigated, with phonetic classification of the input speech into multiple modes and coding each mode differently, leading to some outstanding voice codec designs [20, 21].

Our recent work on speech coding has built on these prior contributions and we have developed a mode classification method, which breaks the input speech into Voiced, Unvoiced, Silence, Onset, and Hangover modes, each of which is coded at a possibly different rate [22]. 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 Gaussian source, Hangover as a 4^{th} order Gaussian source, Unvoiced speech as a memoryless Gaussian source, and silence is treated by sending a code for comfort noise generation. In particular, Table I presents the autocorrelation values and mean squared prediction error for the several modes for two sentences. We have calculated similar data for many speech, audio, and combined speech and audio sequences, but only these two are presented here due to space limitations.

There are a few things to note about the data in Table I. First, the average frame energy for the UV mode and the mean squared prediction errors for the other modes are normalized to the average frame energy over the entire sentence, excluding the segments classified as Silence. Second, the sentence, "We were away ..." has slightly over 47% classified as Silence, while the sentence, "Lathe ..." has slightly over 42% 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.

Another important point is that the mode classification method used here is a very simplified one based on frame prediction error energy [22], and as a result, some low energy voiced segments may be classified as Unvoiced. This appears to be happening for the sentence, "We were away a year ago," since it is widely known that this is a completely voiced sentence.

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.

V. CONDITIONAL RATE DISTORTION FUNCTIONS

Given the composite source models from the prior section, we propose to use the conditional rate distortion results from Gray to derive a rate distortion bound [23]. In particular, the conditional rate distortion function of a source X with side information Y is defined as $R_{X|Y}(D) = \min I(X;\hat{X}|Y)$, where the minimum is taken over $p(\hat{x}|x,y)$: $D(X,\hat{X}|Y) \leq D$ with $I(X;\hat{X}|Y)$ and $D(X,\hat{X}|Y)$ given by the usual expressions. Gray shows that this conditional rate distortion function can be expressed as [23] $R_{X|Y}(D) = \min \sum_y R_{X|y}(D_y)p(y)$ with the minimum taken over D_y's: $D(X,\hat{X}|Y) = \sum_y D_y p(y) \leq D$. The minimum is achieved at D_y's where the slopes $\partial R_{X|Y=y}(D_y)/\partial D_y$ are equal for all y and $\sum_y D_y P[Y=y] = D$.

Since we are assuming that the Voiced, Onset, and Hangover modes can be represented by Gaussian AR models and since we are using the squared error fidelity criterion, we use the classical eigenvalue decomposition [17] and reverse water-filling approach [18] to calculate these conditional rate distortion functions as outlined in a prior section.

VI. RATE DISTORTION BOUNDS FOR SPEECH

Figures 1 and 2 show rate distortion bounds for the sentence, "Lathe is a big tool" and "We were away a year ago," respectively, for the unweighted squared error distortion measure. Each figure contains plots of the marginal rate distortion functions for each speech mode, as well as the conditional rate distortion function over all modes. It is interesting to see that for each sentence the modes have dramatically different rate distortion functions and 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

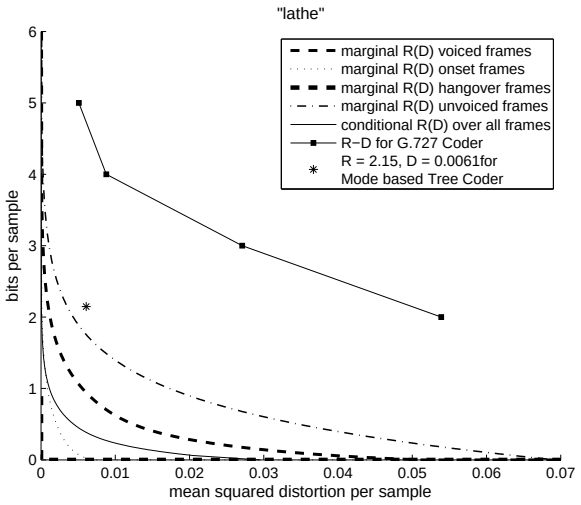

Fig. 1. Marginal and Conditional Rate Distortion Bounds for the Sentence, "The lathe is a big tool" and Operational Rate Distortion Performance of Speech Codecs

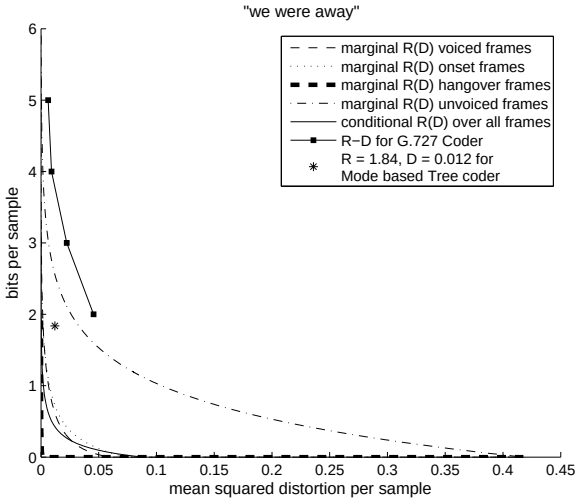

Fig. 2. Marginal and Conditional Rate Distortion Bounds for the Sentence, "We were away a year ago" and Operational Rate Distortion Performance of Speech Codecs

rate distortion functions differently and thus produce a quite different conditional rate distortion bound. This implies that $R(D)$ 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.

The operational rate distortion curves of the G.727 embedded DPCM codec for each sentence are also plotted in Figs. 1 and 2. The G.727 operational rate distortion performance is far above the $R(D)$ bound in both cases, which is not surprising since G.727 does not detect silence and code it separately. A more meaningful comparison is to a new multimode tree coder recently developed by Ramadas and Gibson [22] and the unweighted mean squared error performance achieved by this

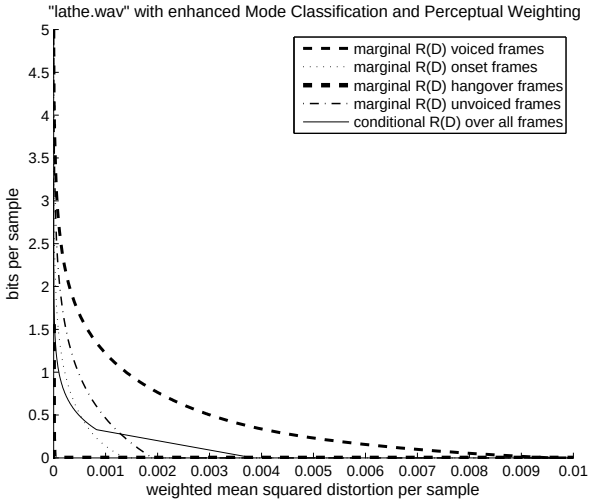

Fig. 3. Marginal and Conditional Rate Distortion Bounds for the Sentence, "The lathe is a big tool" and a Weighted Distortion Measure

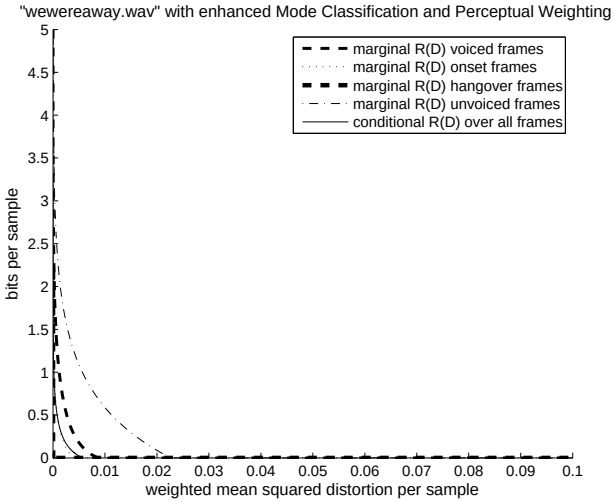

Fig. 4. Marginal and Conditional Rate Distortion Bounds for the Sentence, "We were away a year ago" and a Weighted Distortion Measure

codec is shown by the asterisk-like symbols in the two figures. Here we see that for an average distortion of 0.01, this codec requires a rate of 2.15 bits/sample for the "Lathe" sentence and a rate of 1.84 bits/sample for the "We were away" sentence. These values are much closer to the theoretical rate distortion bounds but are still far from the most efficient coding of the sentences as represented by their $R(D)$ bounds.

The comparison to the new multimode tree coder is not entirely fair since that codec uses perceptual weighting in the search through the tree for the best excitation sequence. However, the multimode tree coder is more of a waveform-following codec than are the code excited linear prediction (CELP) codecs, such as G.729 and AMR. We have not compared the more efficient CELP codecs to the bounds in Figures 1 and 2, since the rate distortion bounds computed here are based on the unweighted mean squared error distortion measure, and CELP codecs employ a perceptually weighted distortion measure. As a result, the performance of CELP codecs, such as G.729 and AMR, is not accurately represented by unweighted mean squared error and cannot be compared to the current bounds. However, an important implication of the rate distortion bounds in Figs. 1 and 2 is that good performance with average distortion is possible with rates as low as 0.25 bits/sample for the current composite source models.

The basic approach can be extended to cover perceptually weighted distortion measures and we present some preliminary results here that incorporate weighting as indicated in the water-filling expression in the earlier section. Figures 3 and 4 show rated distortion bounds for the sentences, "Lathe is a big tool" and "We were away a year ago," respectively, for weighted squared error distortion measures. As before, each figure contains plots of the marginal rate distortion functions for each speech mode, as well as the conditional rate distortion function over all modes.

It is striking to compare these weighted results with the prior unweighted rate distortion bounds. We see that the $R(D)$ curves for the weighted distortion measure are substantially lower than those for the unweighted distortion. These results indicate that the choice of the distortion measure can have a dominating impact on the bounds obtained. Taking this statement in isolation seems far from surprising, but the implication for obtaining meaningful rate distortion bounds for real sources is profound. Generally, the theoretical rate distortion expressions that we employ here are all more than 35 years old, but it is clear that applying them to real sources deserves much more attention and research.

We have begun comparisons of the performance of the standardized CELP speech codecs that utilize perceptual weighting, but we have yet to come up with a meaningful characterization of the codecs performance that we can compare to the bounds we have obtained. The challenge is that the performance of these codecs cannot be captured easily with squared error type calculations. Of course, for those who work in speech coding, this is not surprising since the evaluation of the speech quality produced by state of the art codecs is determined by Mean Opinion Scores (MOS), and the standardized PESQ-MOS algorithm is often used to develop such results. Unfortunately, if one examines the documentation of the PESQ-MOS [24], it is not transparent how to map those results into a weighted squared error fidelity criterion. Thus, another implication of the current work is that more research on tractable distortion measures for speech coding is needed.

We should also add that in calculating the weighted squared error for the latter $R(D)$ curves, we summed the weighted squared error for each mode over the entire sentence, and this may not be representative of real codecs, including our multimode codec, since the perceptual weighting is changing every 10 to 20 milliseconds. Furthermore, this reveals a possible short coming of the classic water-filling bound as we have applied it, since the weights there are fixed. We are

continuing to investigate these issues.

VII. CONCLUSIONS

We present rate distortion bounds for speech coding based upon composite source models for speech and conditional rate distortion theory. Marginal rate distortion functions are computed for each mode and combined to produce a conditional rate distortion bound based on the unweighted and weighted mean squared error distortion measures. For the unweighted case, it is observed that the modes have different rate distortion functions for each sentence and that the rate distortion functions for the modes differ across the two sentences. It is also noted that the probabilities of the different modes dramatically effect the conditional rate distortion bound averaged over all modes. These new bounds imply that good performance is attainable at rates as low as 0.25 bits/sample. The $R(D)$ curves for the weighted squared error distortion measure are much lower than those for the unweighted case, thus indicating the important role of the distortion measure in real applications. Future work is needed to optimize the composite source models and to understand more clearly the implications of the weighted distortion measure results, particularly as to how they may indicate the performance of standardized CELP based speech codecs that rely on perceptually weighting during encoding.

APPENDIX

The speech sequences used in this paper are narrowband and sampled at 8000 samples/second. They are designated as:
(1) "Lathe" (female speaker) – A lathe is a big tool, grab every dish of sugar
(2) "We were away" (male speaker) – We were away a year ago

REFERENCES

[1] J. D. Gibson, "Speech Coding Methods, Standards, and Applications," *IEEE Circuits and Systems Magazine*, Vol. 5, No. 4, Fourth Quarter 2005, pp. 3049 ff.

[2] C.E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Conv. Rec.*, vol. 7, pp. 142-163, 1959.

[3] T. Berger and J. D. Gibson,"Lossy Source Coding," *IEEE Trans. on Information Theory*, Vol. 44, pp. 2693-2723, Oct. 1998.

[4] T. Berger, *Rate distortion theory*. New York: Wiley, 1971.

[5] R.M. Gray, "Information rates of autoregressive processes," *IEEE Transactions on Information Theory*, pp. 412-421, 1970.

[6] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell Syst Tech. J.*, pp. 1973-1986, 1970.

[7] 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.

[8] M. R. Schroeder and B. S. Atal, "Rate distortion theory and predictive coding," *IEEE ICASSP*, pp. 201 - 204, 1981.

[9] R.A. Jones and S.C. Gupta, "A rate-distortion bound on channel requirements for transmission of parameters of adaptive predictive coders, with speech applications," *IEEE Transactions on Information Theory*, Vol. IT-19, pp. 458-471, 1973.

[10] J. D. Gibson, "A bound on the rate of a system for encoding an unknown Gaussian autoregressive source," *IEEE Transactions on Information Theory*, vol. 40, pp. 230-236, 1994.

[11] H. Kalveram and P. Meissner, "Rate distortion bounds for speech waveforms based on Itakura-Saito segmentation," *Signal Processing IV: Theories and Applications*, pp. 137-140, EURASIP, 1988.

[12] De and P. Kabal, "Rate-distortion function for speech coding based on perceptual distortion measure," *Proc. IEEE Globecom Conference*, Orlando FL, pp. 452-456, 1992.

[13] J. Hu and J. D. Gibson, "New Block-Based Local-Texture-Dependent Correlation Model of Digitized Natural Video," *Proceedings of the Fortieth Annual Asilomar Conference on Signals, Systems, and Computers*, October 29 - November 1, 2006.

[14] 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*, Jun 2007.

[15] 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," *Forty-Sixth Annual Allerton Conference on Communication, Control, and Computing*, Sep. 2008.

[16] J. Hu and J. D. Gibson, "New Rate Distortion Bounds for Natural Videos Based on a Texture Dependent Correlation Model," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol 19, Issue 8, pp. 1081-1094, Aug. 2009

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

[18] R. A. McDonald and P. M. Schultheiss, "Information rates of Gaussian signals under criteria constraining the error spectrum," *Proc. IEEE*, vol. 52, pp. 415-416, Apr. 1964.

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

[20] S. Wang and A. Gersho, "Improved Phonetically- Segmented Vector Excitation Coding at 3.4 Kb/s," *in Proceedings, IEEE ICASSP*, San Francisco, pp. 349-352, March 1992.

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

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

[23] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Information Theory*, pp. 480-489, 1973.

[24] *PESQ: An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,* ITU-T P.862 Recommendation, Feb. 2001.