Rate Distortion Performance Bounds for Wideband Speech

Jerry D. Gibson and Ying-Yi Li

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

***Abstract*—We develop new rate distortion bounds for wideband speech sources based on phonetically-motivated composite source models, conditional rate distortion theory, and perceptual wideband PESQ (WPESQ) distortion measures. The approach is to calculate rate distortion bounds for MSE distortion for each subsource of the composite source model and use conditional rate distortion theory to calculate the MSE R(D) for the composite source. Since MSE is not a useful distortion measure for today's best-performing voice codecs, we generate a mapping of MSE-to-WPESQ using fully backward adaptive waveform coders, which have MSE distortion values that correctly order their performance, and for which WPESQ values can be generated. We generate the final rate distortion functions with the mapping and show that our new rate distortion curves lower bound the performance of the best known standardized wideband speech codecs.**

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

I. Introduction

Speech codecs based on linear prediction play a significant role in digital cellular, Voice over IP (VoIP), and Voice over Wireless LAN (VoWLAN) applications; however, while speech researchers have been extremely innovative in optimizing speech codecs, meaningful rate distortion bounds for speech sources have not been.

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.

Like all rate distortion problems, the two primary challenges are (1) finding good source models for speech, 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.

We develop new rate distortion bounds for 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 wideband speech codecs, including, G.718, 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. Mapping functions are developed to map the rate distortion curves based on MSE to rate distortion curves subject to the perceptually meaningful distortion measures 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.

The paper is organized as follows. The relevant prior work is briefly described in Section II. Section III collects the essential results from rate distortion theory. The composite source models developed for speech and used in our rate distortion analyses are discussed in Section IV, and the resulting rate distortion bounds for the MSE distortion measure are presented in Section V. The mapping of the distortion measure from MSE to WPESQ is developed and discussed in Section VI. Rate distortion bounds based on the WPESQ distortion measures and our composite speech model are given in Section VII, where the new rate distortion curves are compared to common, standardized high-performance speech codecs and shown to lower bound the performance of all of the codecs. Conclusions are presented in Section VIII.

II. Related Prior Work

There have only been a handful of efforts to calculate rate distortion bounds for speech, see for example, [1], [2], [3], [4], [5]. However, only Kalveram and Meissner [2], [3], have considered wideband speech. They focus on the MSE distortion measure and obtain composite source models for wideband speech by segmentation of the speech into equal order autoregressive subsources. Each subsource is parametrized by the predictor coefficients and the residual variance, which are estimated by maximum likelihood estimation. 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

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

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. No comparisons to standardized speech codecs are provided since MSE is not a meaningful distortion measure for these codecs.

The work in the present paper extends our work in [5], [6] by providing new composite source models and new rate distortion bounds for wideband speech, and by comparing the performance of standardized wideband speech codecs to the new rate distortion bounds.

III. Rate Distortion Theory

Even though the target here is to generate rate distortion bounds based on a perceptual distortion measure, we formulate the rate distortion bounds based on MSE distortion measures first, and then use a mapping function of MSE and WPESQ to generate the rate distortion bounds based on WPESQ distortion measure. We start by stating the classical result for 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 [7]. 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 III.1. (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 $D(\underline{X}, \hat{\underline{X}}) = E\left[\sum_{i=1}^{n}(X_i - \hat{X}_i)^2\right] \leq D$, the rate distortion function is*

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

where

$$D_i = \begin{cases} \lambda & 0 \leq \lambda \leq \sigma_i^2 \\ \sigma_i^2 & \lambda > \sigma_i^2 \end{cases} . \quad (2)$$

A. Conditional Rate Distortion Functions based on MSE

Given a general composite source model, a rate distortion bound based on the MSE distortion measure can be derived [5] using the conditional rate distortion results from Gray [8]. The conditional rate distortion function of a source $\underline{X}$ with side information Y, which serves as the subsource information, is defined as

$$R_{\underline{X}|Y}(D) = \min_{p(\hat{\underline{x}}|\underline{x}, y): D(\underline{X}, \hat{\underline{X}}|Y) \leq D} I(\underline{X}; \hat{\underline{X}}|Y), \quad (3)$$

where

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

$$I(\underline{X}; \hat{\underline{X}}|Y) = \sum_{\underline{x}, \hat{\underline{x}}, y} p(\underline{x}, \hat{\underline{x}}, y) \log \frac{p(\underline{x}, \hat{\underline{x}}|y)}{p(\underline{x}|y) p(\hat{\underline{x}}|y)}. \quad (4)$$

It can be proved [8] that the conditional rate distortion function in Eq. (3) can also be expressed as

$$R_{\underline{X}|Y}(D) = \min_{D_y' s: D(\underline{X}, \hat{\underline{X}}|Y) = \sum_y D_y p(y) \leq D} \sum_y R_{\underline{X}|y}(D_y) p(y), \quad (5)$$

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. The equal slope requirement means that the marginal rate distortion functions are combined at points of equal average distortion.

Utilizing the results for conditional rate distortion functions in Eq. (5), 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 $R_{\underline{X}}(D)$ [8]

$$R_{\underline{X}|Y}(D) \leq R_{\underline{X}}(D) \leq R_{\underline{X}|Y}(D) + I(\underline{X}; Y), \quad (6)$$

where $I(\underline{X}; Y)$ is the average mutual information between $\underline{X}$ and Y. We can bound $I(\underline{X}; Y)$ by

$$I(\underline{X}; Y) \leq H(Y) \leq \frac{1}{M} \log K, \quad (7)$$

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 $K = 5$ here and M is on the order of 100 or more, the second term on the right in Eq. (6) is negligible, and the rate distortion function for the source is very close to the conditional rate distortion function in Eq. (5). Therefore, we use the conditional rate distortion function $R_{\underline{X}|Y}(D)$ to develop our performance bounds [2], [3].

IV. Composite Source Models

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 [9].

In our earlier work [10] on narrowband speech coding, we developed a mode classification method that breaks the speech into Voiced (V), Onset (ON), Hangover (H), Unvoiced (UV), and Silence (S) modes, each of which may be coded at a different rate. For wideband speech, we down-sample the wideband speech from 16 kHz to 12.8 kHz using a decimation filter, as is done for AMR-WB and for the lower rates of G.718, and model Voiced speech as a 16^{th} order AR Gaussian source at 12.8 kHz. 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 rate. Table I presents the autocorrelation values, mean squared prediction error, and the probability for the several modes for two wideband English sentences.

In Table I, 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.

TABLE I
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

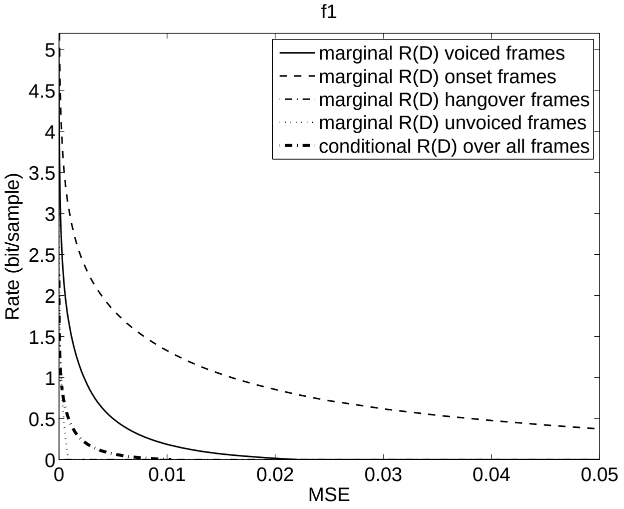

Fig. 1. The MSE rate distortion bounds of wideband sequence F1, “You must go and do it at once. There were several small outhouses.”

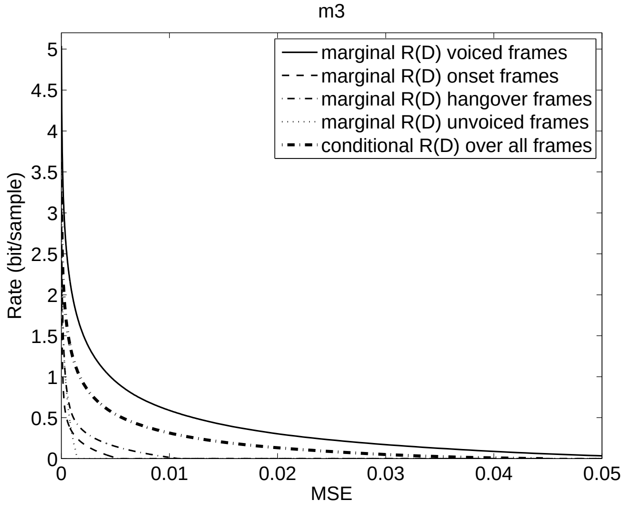

Fig. 2. The MSE rate distortion bounds of wideband sequence M3, “I don’t know, the vampire said, and he smiled.”

The sentence F1 has 55.23% and M3 has 24.67% 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.

V. MARGINAL AND CONDITIONAL RATE DISTORTION BOUNDS BASED ON MSE DISTORTION MEASURE

The resulting marginal and conditional MSE rate distortion bounds of the composite source models for two English wideband sentences are shown in Figs. 1 and 2. It is interesting, but perhaps not surprising, to see that for each sentence, the several modes have different rate distortion functions; furthermore, the rate distortion functions for the modes differ across the two sentences, since the model of each mode is different for each sentence. Another important point is that the probabilities of the different modes 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. 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.

VI. MAPPING MSE TO WPESQ

The rate distortion theory results are built on the assumption of the MSE distortion measure, which unfortunately, is not a reliable or widely used indicator of speech codec performance. Alternatively, WPESQ is a standardized objective methods for wideband speech quality assessment, and is widely used in categorizing the perceptual performance of standardized speech codecs. Therefore, in order to extend the utility of the prior theoretical rate distortion theory results, we have developed a procedure for mapping MSE into WPESQ. The way we do this is to use waveform coders for which MSE is a reasonable performance indicator, in that MSE correctly orders the perceptual performance of these waveform codecs, although the difference in MSE might not be an exact indicator of the perceptual quality difference. However, developing such a mapping is not straightforward and several constraints need to be imposed to make the mapping meaningful.

In particular, the mapping of MSE to WPESQ must be performed with several key points in mind. First, the mapping must be done with a codec for which MSE is a valid performance indicator and to which WPESQ 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 do not include bit rate for the separate encoding of the prediction coefficients, and therefore the codec used for the mapping should not do so as well; however, the effect of the coefficients on bit rate must be incorporated in some fashion. In order to meet this constraint, we chose backward adaptive waveform coders to perform the mapping. Fourth, the codecs used for the mapping must have a range of bit rates sufficient 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 WPESQ values. Sixth, the mapping must be matched to each individual utterance to be evaluated. Another critical consideration is the active speech level of test sequence. For the WPESQ, we need to avoid peak clipping (mentioned in P.862.3), and therefore, the active speech level should not be too low or too high. The active speech level of test sequences we use is between -15 dBov and -30 dBov. Further, if the energy of the speech utterance is too low, the MSE will blow up.

To satisfy these constraints for wideband MSE to WPESQ mapping, we use several ADPCM speech coders for wideband speech, including the G.722 standard and several extensions. We have 81 MSE and WPESQ pairs to generate a mapping function for each wideband sentence.

The MOS-like WPESQ 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, the normal range of MOS values found in an ACR listening quality experiment. The details of WPESQ are described in the ITU-T P.862.2 Recommendation [11].

A. Mapping Function

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 WPESQ of each coded sequence is evaluated by the software provided by ITU-T Recommendation P.862.2 [11]. Since MSE is increasing and 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 WPESQ. We chose the exponential function to perform the curve fitting since it provides a good fit across all rates and distortion pairs. The range of WPESQ is between -0.5 and 4.5 [12], so the WPESQ is 4.5 when MSE is 0, and we forced $f(0) = 4.5$. Therefore, the mapping function is modeled as

$$z = f(w) = ae^{-bw} + 4.5 - a, \tag{8}$$

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

Several clean English sequences are used to illustrate the results of designing the mapping functions for 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 WPESQ are highly source dependent.

The mapping functions of wideband sequences F1 and M3 are shown in Figs. 3 and 4. We found that there are points that should be considered as outliers for curve fitting in the two figures. In particular, the WPESQs of the points marked as outliers do not match our separate subjective listening tests, which reveal poorer audible performance than the WPESQ values obtained. Therefore, we fit two mapping functions for each sentence, one that includes all points and one that removes the outliers. For sentence F1, two mapping functions are shown in Fig. 3; the line with all points including outliers (dotted line) uses 81 points, while the solid line is for the mapping when the outliers are removed, which uses 65 MSE/WPESQ pairs. For sentence M3, in Figure 4, the mapping function including the outliers is the dotted line and uses 81 points, while the mapping without outliers uses 33 points and is the solid line. We plot rate distortion curves for both mappings shown for F1 and M3 in the following.

VII. Rate Distortion Bounds for Speech

The rate distortion bounds based on MSE from Sec. V are mapped to WPESQ values by the mapping function generated by the ADPCM waveform codecs as described in Section VI.

A. Rate Distortion Bounds and Operational Rate Distortion Performance for Wideband Speech

The rate distortion bounds based on WPESQ are compared with wideband speech codecs, such as AMR-WB [13] and G.718, and G.722.1 [14] in Figs. 5 and 6. For AMR-WB, 9 different bit-rates, 6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, and 23.85 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.722.1, 2 different bit-rates, 24 and 32 kbps, are used for comparison. The WPESQs of all speech codecs are computed by ITU-T P.862.2 [11].

In Fig. 5, the rate distortion curves generated by the two mappings from Fig. 3 are almost coincident since the mapping functions are very close for WPESQ values greater than about 2.8. For Fig. 6, the rate distortion curve generated without using the outliers is higher and hence the performance of the speech codecs is closer to the bound. From both Figs. 5 and 6, we see that the performance of all wideband codecs is lower bounded by the rate distortion curves obtained here. In addition, CELP codecs, AMR-WB and G.718, are much closer to the rate distortion bounds than G.722.1. This is as expected because the two CELP codecs have VAD and encode silence by comfort noise generation. By comparing these two CELP codecs with our rate distortion bounds, we observe that for sentence F1, AMR-WB is about 0.4 bit/sample above our bound for a WPESQ of 3.5 and G.718 is about 0.3 bit/sample above the rate distortion curves. The performance of the codecs is further away from the bounds obtained for M3,

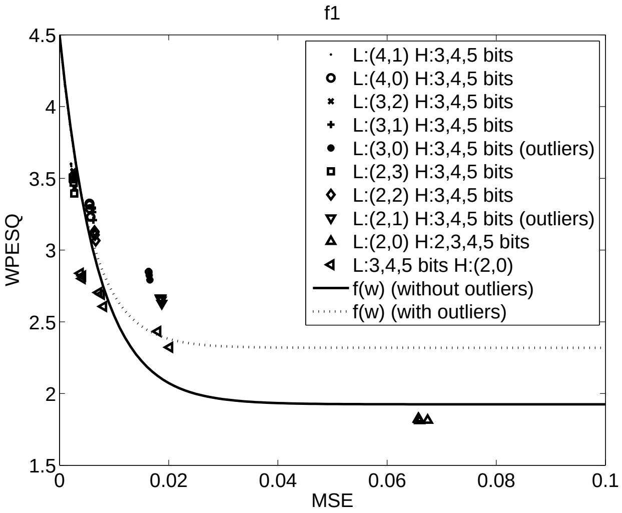

Fig. 3. The mapping function of wideband sequence F1.

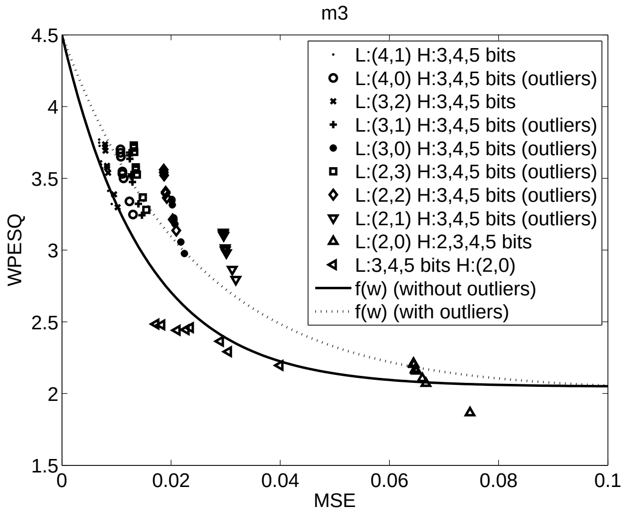

Fig. 4. The mapping function of wideband sequence M3.

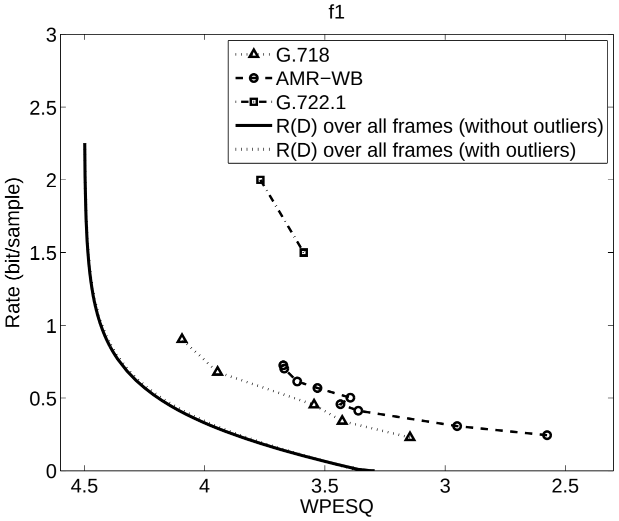

Fig. 5. The rate distortion bounds and operational rate distortion performance of speech codecs of the sequence of F1.

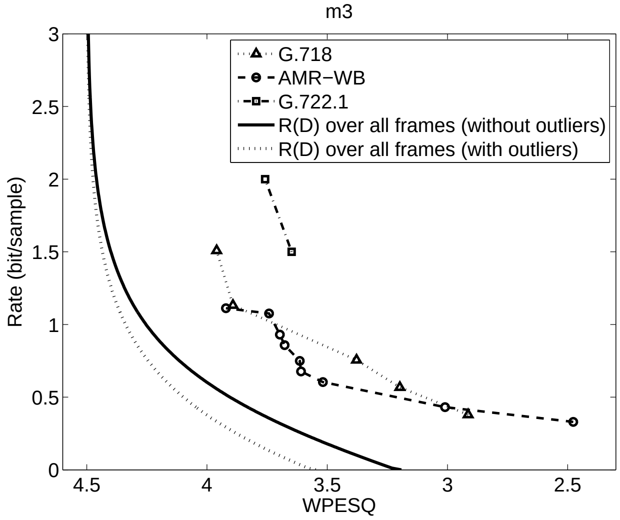

Fig. 6. The rate distortion bounds and operational rate distortion performance of speech codecs of the sequence of M3.

with AMR-WB 0.4 to 0.5 bit/sample above the rate distortion curves and G.718 over 0.5 bit/sample above for a WPESQ of 3.5. Since these codecs utilize a sampling rate of at least 12.8 kHz, these results imply that the codec bit rates can be reduced by 4 to 6 kilobits/sec.

These are the first meaningful bounds on the rate distortion performance of standardized speech codecs to date, and by studying the several subsources of the composite source models and the performance of each codec for each sentence, it is possible to obtain deep insights into how the existing codecs can be improved.

VIII. Conclusions

We develop new rate distortion bounds for wideband speech coding based on composite source models for speech and perceptual 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.718, G.722.1, and AMR-WB.

In addition, 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 unvoiced speech, Onset, and Hangover modes.

References

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

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

[3] ——, "Itakura-saito clustering and rate distortion functions for a composite source model of speech," *Signal Processing*, vol. 18, no. 2, pp. 195 – 216, 1989.

[4] 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.

[5] 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.

[6] Y.-Y. Li and J. D. Gibson, "Rate Distortion Bounds for Speech Coding based on a Perceptual Distortion Measure (PESQ-MOS)," *IEEE International Conference on Multimedia and Expo (ICME'11)*, Barcelona, July 2011.

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

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 3GPP, "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions," 3rd Generation Partnership Project (3GPP), TS 26.190, Mar. 2011.

[14] 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.