Rate Distortion Analyses and Bounds on Speech Codec Performance

Jerry D. Gibson, Fellow, IEEE and Ying-Yi Li, Student Member, IEEE

EDICS Name: 1-CPRS

Abstract—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. The approach is to calculate rate distortion bounds for MSE distortion measures using the classic eigenvalue decomposition and reverse water-filling approach 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 speech source. Mapping functions are developed to map the rate distortion functions based on MSE to rate distortion curves subject to the perceptually meaningful distortion measures PESQ-MOS and WPESQ. These final rate distortion functions are then compared to the performance of the best known standardized speech codecs based on the code-excited linear prediction paradigm. An analysis of the tightness of these bounds indicates how the performance of existing voice codecs might be improved.

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, and extraordinary progress has been made in developing standardized speech codecs for these applications. Indeed, code-excited linear prediction (CELP) is the basis for most speech coding standards for narrowband (200 Hz to 3400 Hz) and wideband (50 Hz to 7000 Hz) speech today. Additionally, recent and on-going standardization efforts toward the development of fullband codecs (20 Hz to 20 kHz) utilize a combination of the CELP approach and the transform/filter bank approaches with well-designed switching between the coding methods, as is evident in the recently standardized USAC codec [1] and as is expected in the EVS codec for LTE Mobile Systems [2]. While speech researchers have been extremely innovative in optimizing the form of speech codecs, it would be very useful at this point if meaningful rate distortion analyses relating to the performance of speech codecs were available.

One natural place to look in order to characterize the performance of lossy source codecs would 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. However, such an effort is not easy, and experts in Information Theory have stated the difficulties succinctly. Specially, Robert Gallager, in his classic text on Information Theory [3], 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 …” [3].

Thus, 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 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. Mapping functions are developed to map the 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 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 unvoiced speech, Onset, and Hangover modes. Finally, 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 the mapping functions given in this paper to
determine new bounds for the utterances considered in this
paper.

The paper is organized as follows. The relevant prior work
is briefly described in Section II. Section III collects and
organizes the essential results from rate distortion theory in
a manner that shows the theoretical underpinnings of our
results and also provides a road map for those who want
to extend or reproduce our results. More specifically, this
section presents the rate distortion bounds based on MSE
distortion measures, including the rate distortion functions for
Gaussian autoregressive sources, and a brief description of
reverse water filling, composite source models, and conditional
rate distortion functions based on MSE. 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 PESQ-MOS/WPESQ is developed and discussed in Section
VI. Rate distortion bounds based on the PESQ-MOS/WPESQ
distortion measures and our composite speech model are given
in Section VII, where the new rate distortion curves are
discussed in Section V, and the resulting rate
distortion bounds for the MSE distortion measure presented in
Section VII. Where the new rate distortion curves are
compared to common, standardized high-performance speech
codes and shown to lower bound the performance of all of
the codecs. Conclusions are presented in Section VIII, wherein
the contributions of the current work are summarized and
suggestions for improving or extending the bounds and for
future work are given.

II. RELATED PRIOR WORK

Brehm and Trottler [4] 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 inclu-
sion 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. 1. In [5], [6], composite source models for speech
are obtained 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 distor-
tion functions of composite sources are calculated using condi-
tional 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
citation 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.

A cochlear model serves as the basis for a perceptual distor-
tion measure for speech in [7], and the speech source model
is merged with the cochlear models and used to characterize
the rate distortion function for speech. The rate distortion
bound is calculated with the cochlear variational distance, 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 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 [8] obtained rate distortion
bounds for speech coding based on composite source mod-
els 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). 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. The
particular weighted MSE considered in this paper was not able
to meaningfully capture the performance of important CELP
based codecs and thus the rate distortion curves based on this
weighted MSE criterion are not useful.

The work in the present paper augments and extends our
conference paper [9] by providing new composite source
models and new rate distortion bounds for wideband speech,
and by comparing the performance of standardized wideband

![Fig. 1. A Composite Source Model with K Subsources](image-url)
speech codecs to the new rate distortion bounds. Additionally, an extensive new development of the relevant rate distortion results is presented, which clarifies the theoretical foundations of our approach and clearly delineates the steps in calculating the bounds. Further, a detailed development of the MSE to PESQ-MOS/WPESQ mapping procedure is given, which specifies the critical constraints in performing such a mapping.

III. RATE DISTORTION THEORY

For a specific source model and distortion measure, rate distortion theory answers the following fundamental question: What is the minimum number of bits (per data sample) needed to represent the source at a chosen average distortion? A mathematical characterization of the rate distortion function is given by the following fundamental theorem of rate distortion theory.

**Theorem III.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 $d(x, \hat{x}) \leq D$, is equal to

$$R(D) = \min_{p(\hat{x}|x): \sum_{x, \hat{x}} p(x)p(\hat{x}|x)I(x; \hat{x}) \leq D} I(X; \hat{X}),$$

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

Whether or not there exists a closed-form 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. 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.

A. Scalar Gaussian Source with Mean Squared Error

The rate distortion function of a scalar Gaussian source with squared error distortion is as follows [10].

**Theorem III.2.** The rate distortion function for a scalar Gaussian random variable $X \sim N(0, \sigma^2)$ with squared error distortion measure $\sum_{x, \hat{x}} p(x)p(\hat{x}|x)(x-\hat{x})^2 \leq D$ is

$$R(D) = \min_{p(\hat{x}|x): \sum_{x, \hat{x}} p(x)p(\hat{x}|x)(x-\hat{x})^2 \leq D} I(X; \hat{X}) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2 \\ 0, & D > \sigma^2 \end{cases}.$$  

B. Rate Distortion Background

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 [10] and Gray [11], 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 [10]:

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

where

\[ D_i = \begin{cases} \lambda \sigma_i^2, & 0 \leq \lambda \leq \sigma_i^2 \\ \lambda, & \lambda > \sigma_i^2 \end{cases} \]
In the following section, we show how to connect the usual information that we have about speech sources, namely the autocorrelation/covariance function, with a decomposition that allows us to use the prior result on parallel Gaussian sources to calculate the rate distortion function.

D. Stationary Gaussian Sources with Memory

The following derivation of rate distortion theory for stationary Gaussian sources first appeared in [13].

Let \( A \) be an unitary matrix denoting an orthonormal linear transformation
\[
\Theta = AX, \quad \hat{X} = A^{-1}\Theta = A^{T}\Theta.
\]
The following relation between \( X \) and \( \Theta \) can be derived:
Mean squared error:
\[
D(X, \hat{X}) = E[(X - \hat{X})^T(X - \hat{X})] = E[(\Theta - \hat{\Theta})^T A^T A (\Theta - \hat{\Theta})] = D(\Theta, \hat{\Theta});
\]
Mutual information:
\[
I(X; \hat{X}) = I(\Theta; \hat{\Theta}) \geq \sum_{i=1}^{n} I(\Theta_i; \hat{\Theta}_i),
\]
with equality if and only if (iff) \( \Theta_i \)'s are independent.

Therefore the rate distortion function of \( X \) equals to the rate distortion function of \( \Theta \).

So now the question becomes, “How do we get the vector \( \Theta \)?” Of course, the answer is that we use the well known Karhunen L`oeve Transform (KLT), which is also called principal component analysis, to decorrelate the source as in what follows. Supposing the covariance function of a stationary Gaussian source is
\[
\phi(n) = E[x_i x_{i+n}],
\]
let \( \Phi_n \) be the \( n \times n \) covariance matrix of the source, with its entries defined in Eq. (11), i.e., \( \Phi_n = \{\phi(i - j), i, j = 1, ..., n\} \). Let \( \{\psi_i, i = 1, ..., n\} \) be the normalized eigenvectors of \( \Phi_n \) with corresponding eigenvalues \( \{\lambda_i, i = 1, ..., n\} \), so that
\[
\Phi_n \psi_i = \lambda_i \psi_i
\]
i.e.,
\[
\Phi_n = \Psi_n \Lambda \Psi_n^T,
\]
where \( \Psi_n = [\psi_1, \psi_2, ..., \psi_n] \).

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.

Let
\[
\Theta = \Psi_n^T X.
\]
The entries of the vector \( \Theta_i \) for \( i = 1, ..., n \) are uncorrelated, and if they are Gaussian, they are independent and the variance of \( \Theta_i \) is equal to the corresponding eigenvalue \( \lambda_i \) of \( \Phi_n \).

After the KLT, the rate distortion function of the decorrelated sources can then be solved using the reverse water filling theorem as discussed in the previous section.

E. Conditional Rate Distortion Functions based on MSE

We now have a procedure for calculating the rate distortion function of an autoregressive source subject to a MSE distortion measure. However, a single source model is not effective in modeling real speech sources as indicated by prior work outlined in Sec. II. More specifically, composite source models, which combine multiple subsources according to a switch process, can serve as a good model for speech when characterizing achievable rate distortion performance. Therefore, given a general composite source model, a rate distortion bound based on the MSE distortion measure can be derived [8] using the conditional rate distortion results from Gray [14]. The conditional rate distortion function of a source \( X \) with side information \( Y \), which serves as the subsource information, is defined as
\[
R_{X|Y}(D) = \min_{P(\hat{x}|x, y): D(X, \hat{X}|Y) \leq D} I(X; \hat{X}|Y),
\]
where
\[
D(X, \hat{X}|Y) = \sum_{x, \hat{x}} p(x, \hat{x}|y) D(x, \hat{x}|y),
\]
\[
I(X; \hat{X}|Y) = \sum_{x, \hat{x}} p(x, \hat{x}|y) \log \frac{p(x, \hat{x}|y)}{p(x|y)p(\hat{x}|y)}.
\]

It can be proved [14] that the conditional rate distortion function in Eq. (15) can also be expressed as
\[
R_{X|Y}(D) = \min_{D_y: D(X, \hat{X}|Y) = \sum_y D_y p(y) \leq D} \sum_y R_{X|y}(D_y) p(y),
\]
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 classical results for conditional rate distortion functions in Eq. (17), the minimum is achieved at \( D_y \)'s where the slopes \( \frac{\partial R_{X|y}(D_y)}{\partial D_y} \) are equal for all \( y \) and \( \sum_y D_y p(Y = y) = D \).

This conditional rate distortion function \( R_{X|Y}(D) \) can be used to write the following inequality involving the overall source rate distortion function \( R_X(D) \) [14]
\[
R_{X|Y}(D) \leq R_X(D) \leq R_{X|Y}(D) + I(X; Y),
\]
where \( I(X; Y) \) is the average mutual information between \( X \) and \( Y \). We can bound \( I(X; Y) \) by
\[
I(X; Y) \leq H(Y) \leq \frac{1}{M} \log K,
\]
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. (18) is negligible, and the rate distortion for the source is very close to the conditional rate distortion function in Eq. (17). Therefore, we use the conditional rate distortion function \( R_{X|Y}(D) \) to develop our performance bounds [5], [6].
TABLE I

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Mode</th>
<th>Autocorrelation coefficients for V, ON, H</th>
<th>Mean Square Prediction Error</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;lathe&quot; (Female)</td>
<td>V</td>
<td>0.8217 0.5592 0.3435 0.1498 0.0200</td>
<td>0.0650</td>
<td>0.5265</td>
</tr>
<tr>
<td></td>
<td>ON</td>
<td>-0.0517 -0.0732 -0.0912 -0.1471 -0.2340</td>
<td>0.0432</td>
<td>0.0093</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>0.1209 0.2708 0.1576 0.1182</td>
<td>0.7714</td>
<td>0.0186</td>
</tr>
<tr>
<td></td>
<td>UV</td>
<td>0.1439</td>
<td>0.1439</td>
<td>0.0771</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.3685</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;we were away&quot; (Male)</td>
<td>V</td>
<td>0.8014 0.5176 0.2647 0.0432 -0.1313</td>
<td>0.0780</td>
<td>0.9842</td>
</tr>
<tr>
<td></td>
<td>ON</td>
<td>-0.2203 -0.3193 -0.3934 -0.4026 -0.3628</td>
<td>0.0680</td>
<td>0.0053</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UV</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.6125</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE II

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Mode</th>
<th>Autocorrelation coefficients for V, ON, H</th>
<th>Mean Square Prediction Error</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 (Female)</td>
<td>V</td>
<td>0.8438 0.5891 0.4132 0.3156 0.2670 0.2122</td>
<td>0.0233</td>
<td>0.4406</td>
</tr>
<tr>
<td></td>
<td>ON</td>
<td>0.1402 0.0599 -0.0987 -0.3028 -0.4109</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>-0.3816 -0.3084 -0.2673 -0.2879 -0.3293</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UV</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0028</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.5523</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3 (Male)</td>
<td>V</td>
<td>0.7954 0.6612 0.4775 0.2864 0.2398 0.2004</td>
<td>0.0801</td>
<td>0.6039</td>
</tr>
<tr>
<td></td>
<td>ON</td>
<td>0.2169 0.2214 0.2284 0.2022 0.1613</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>0.1333 0.1075 0.1334 0.1759 0.1662</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UV</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0064</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>0.2467</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 [10]. Prior work on rate distortion bounds for speech coding, as discussed in Sec. II, 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.

Multimodal models have played a major role in speech coding, including the voiced/unvoiced decision for the excitation in linear predictive coding (LPC) [15] and the long-term adaptive predictor in adaptive predictive coding (APC) [16]. Further, phonetic classification of the input speech into multiple modes and coding each mode differently has lead to some outstanding voice codec designs [17], [18]. We build on the phonetic classification methods in these successful codec designs to surmise useful composite source models.

In particular, our earlier work [19] 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 speech as a 10th order AR Gaussian source since most narrowband speech codecs, such as AMR-NB, use 10th order linear prediction in the codec. Onset is modeled as a 4th order AR Gaussian source, Hangover is modeled as a 4th order AR Gaussian source, Unvoiced speech is modeled 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 narrowband English sentences. The probability of each mode is also shown in Table I.

For wideband speech, AMR-WB uses 16th order linear prediction at 12.8 kHz sampling rate. Therefore, we downsample the wideband speech from 16 kHz to 12.8 kHz using the decimation filter in AMR-WB, and model Voiced speech as a 16th order AR Gaussian source at 12.8 kHz. Onset and Hangover are modeled as 4th 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 II presents the autocorrelation values, mean squared prediction error, and the probability for the several modes for two wideband English sentences.

There are a few things to note about the data in Tables I and II. 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 other sentences, “lathe” has 36.85%, 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.

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. 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 narrowband English sequences are shown in Figs. 3 and 4, and results for two English wideband sentences are shown in Figs. 5 and 6. 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 four 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. In Fig. 4, 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. 3, 5, and 6, 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.

VI. MAPPING MSE TO PESQ-MOS/ WPESQ

The rate distortion theory results 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. Alternatively, 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. Therefore, in order to extend the utility of the prior theoretical rate distortion theory results, we have developed a procedure for mapping MSE into PESQ-MOS or WPESQ, as is appropriate for the bandwidth of interest. 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 PESQ-MOS 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 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 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 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 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 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 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. For wideband MSE to WPESQ mapping, ADPCM speech codecs 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 VI-B2. We first describe the PESQ-MOS and WPESQ standards as widely employed for speech codec performance evaluation. This information also plays a role in the development of the mapping function.

A. PESQ-MOS/ WPESQ

Perceptual evaluation of speech quality (PESQ) [20] is an objective method for end-to-end speech quality assessment of narrowband speech codecs. 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 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 $4.5$, the normal range of MOS values found in an ACR listening quality experiment. 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 P.862 model. The details of WPESQ are described in the ITU-T P.862.2 Recommendation [21].

B. ADPCM Speech Coders

ADPCM coders are waveform coders, and as a result, MSE is an indicator of how well the codec is reproducing the input speech waveform. MSE (SNR) is also useful in establishing the relative ordering of the performance of ADPCM speech coders [22]. In addition, the PESQ-MOS,WPESQ of ADPCM coders can be evaluated, thus providing a perceptual distortion function for each narrowband sentence.

1) G.726/G.727 Narrowband ADPCM Speech Coders: G.726 [23], [24] and G.727 [24], [25] are standardized narrowband 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, it is referred to by (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. ITU-T G.727 Recommendation [25] provides coding rates of 40 kbps for the 2 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 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.

2) Wideband ADPCM Speech Coders: G.722 [26] is a standardized wideband ADPCM speech coder. However, 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. We cannot get a good curve fitting result based on such a small amount of data whose range is much smaller than the mapping range. Therefore, we create a 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) like G.722 by using the quadrature mirror filters from G.722, and the signals in each sub-band are encoded using G.727 with different bit-rates. Since there are 9 coding rates for G.727 as discussed in VI-B1, we have 81 MSE and WPESQ pairs to generate a mapping function for each wideband sentence.

**C. 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 PESQ-MOS WPESQ of each coded sequence is evaluated by the software provided by ITU-T Recommendation P.862/P.862.2 [20], [21]. As mentioned in Section VI-B, there are 13 pairs of MSE and PESQ that we use for curve fitting for each narrowband sequence, and 81 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 \( z = \frac{w}{a} + b \), and the exponential function \( z = a e^{-bw} + c \), where \( w \) is MSE and \( z \) is PESQ-MOS/WPESQ. We chose the exponential function to perform the curve fitting since it provides a good fit across all rates and distortion pairs. The range of PESQ-MOS/WPESQ is between \(-0.5\) and \(4.5\) [20], so the PESQ-MOS/WPESQ is \(4.5\) when MSE is \(0\), and we forced \( f(0) = 4.5 \). Therefore, the mapping function is modeled as

\[
z = f(w) = a e^{-bw} + 4.5 - a,
\]

where \( a \) and \( b \) are estimated by the least squares fit of the MSE and PESQ/WPESQ pairs of 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 [24], [27]. ITU-T Recommendation P.830 [28] 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 [29] 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 I and Table II.

The mapping functions of the narrowband sequences “lathe” and “we were away” are shown in Figs. 7 and 8. The results show that the exponential function provides a good fit to the MSE/PESQ-MOS pairs. The mapping functions of wideband sequences F1 and M3 are shown in Figs. 9 and 10. We found that there are some outliers for curve fitting in Fig. 9. 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 remove those outliers, and use 63 MSE/ WPESQ pairs to generate the mapping function for sequence F1. Figure 10 shows that the exponential function fits most MSE/ WPESQ pairs well.

**VII. RATE DISTORTION BOUNDS FOR SPEECH**

The rate distortion bounds using MSE as distortion measures are calculated by the classical eigenvalue decomposition [12] and reverse water-filling approach described in Section III-C with the composite source models presented in Section IV. 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 VI. The operational rate distortion performance of five different narrowband speech codecs are compared with the conditional rate distortion bounds based on PESQ in Section VII-A, while four wideband speech codecs are compared with the conditional rate distortion bounds based on WPESQ in Section VII-B. 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, G.727, AMR-NB, G.729, G.718, G.722, G.722.1, and AMR-WB. Detailed discussions of the results follow.

A. 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 [30], G.729 [31], and G.718 [32], and ADPCM coders, G.726 and G.727, in Figs. 11 and 12. 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 [20].

From Figs. 11 and 12, we see that the performance of all narrowband codecs is lower bounded by the rate distortion curves calculated as described in this work. 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 distortion bound since they do not detect silence and code it separately. The performance of AMR-NB, G.729, and G.718 is 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 two sentences shown. The performance of the codecs, AMR-NB, G.729, and G.728, for the utterance “We were away a year ago” is significantly closer to the rate distortion bound than the other sequence. 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.

B. 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 [33] and G.718, G.722.1 [34], and the standardized subband/ADPCM coder, G.722, in Figs. 13 and 14. 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. For G.722, 3 different bit-rates, 48, 56, and 64 kbps are compared. The WPESQs of all speech codecs are computed by ITU-T P.862.2 [21].

From Figs. 13 and 14, we see that the performance of all wideband codecs is lower bounded by the rate distortion curves obtained here, just as for the narrowband codecs. In addition, CELP codecs, AMR-WB and G.718, are much closer to the rate distortion bounds than the subband/ADPCM coder, G.722,
which follows the narrowband results. 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 can conclude that AMR-WB and G.718 can be improved by about 0.4 bit/sample and achieve the same WPESQ. The operational rate distortion performance of G.722.1 and G.722 is far above the rate distortion bound because both codecs operate at a fixed bitrate.

These 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. 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.

VIII. CONCLUSIONS

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. The approach is to calculate rate distortion bounds for MSE distortion measures using the classic eigenvalue decomposition and reverse water-filling approach 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 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.
Like all rate distortion bounds, these new bounds are valid only for the source models and distortion measures used in their calculation. For example, the theoretical calculations for the subsources are based on the assumption that the subsources are Gaussian. Gaussian sources are known to be the most pessimistic compared to other source distributions; that is, Gaussian source models yield rate distortion performance that upper bounds the rate distortion performance of sources with other distributions, such as Laplacian or Gamma. Additionally, the number of subsources and the models for certain subsources are somewhat arbitrary. Namely, using a 4\textsuperscript{th} order AR model for Onset speech is not fully justified. However, Onset mode has a very low probability in most utterances, and hence, does not greatly impact the final rate distortion performance. Therefore, it is clear that these new bounds can be improved by more accurate composite source models or a more accurate method for determining the average distortion.

In addition to the striking result that these new bounds do in fact lower bound the best known 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 unvoiced speech, Onset, and Hangover modes. Finally, 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, and then employ the mapping functions given here to determine new bounds for the utterances considered in this paper.

REFERENCES


