NTIA Technical Report TR-10-462

Low Rate Speech Coding and Random Bit Errors: A Subjective Speech Quality Matching Experiment

Andrew A. Catellier
Stephen D. Voran

report series

U.S. DEPARTMENT OF COMMERCE • National Telecommunications and Information Administration

NTIA Technical Report TR-10-462

Low Rate Speech Coding and Random Bit Errors: A Subjective Speech Quality Matching Experiment

Andrew A. Catellier
Stephen D. Voran

U.S. DEPARTMENT OF COMMERCE
Gary Locke, Secretary

Lawrence E. Strickling, Assistant Secretary
for Communications and Information

October 2009

DISCLAIMER

Certain commercial equipment, instruments, or materials are identified in this report to adequately describe the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is necessarily the best available for the purpose.

CONTENTS

Page

FIGURES

TABLES

Page

LOW RATE SPEECH CODING AND RANDOM BIT ERRORS: A SUBJECTIVE SPEECH QUALITY MATCHING EXPERIMENT

Andrew A. Catellier and Stephen D. Voran*

When bit errors are introduced between a speech encoder and a speech decoder, the quality of the received speech is reduced. The specific relationship between speech quality and bit error rate (BER) can be different for each speech coding and channel coding scheme.

This report describes a subjective experiment concerning the relationships between BER and perceived speech quality for the TIA Project 25 Full Rate (FR), Enhanced Full Rate (EFR), and Enhanced Half Rate (EHR) speech codecs. Using the FR codec with 2% random bit errors as a reference, we sought to characterize the BER values for which the EFR (or EHR) codec produces speech quality that is equivalent to the reference. We used an adaptive paired-comparison subjective testing algorithm to efficiently adapt BER values for the EFR and EHR codecs to quickly locate the BER values where listeners found the speech quality to be the same as the reference.

The results from sixteen listeners reveal ranges of BER values that were judged to produce speech quality equivalent to the reference. When these ranges are reduced to central values, those values indicate that on average, the EFR and EHR codecs are more robust to bit errors than the FR codec. We provide a set of additional results from a popular objective speech quality estimator for comparison purposes.

Key words: bit errors, listening tests, speech coding, speech quality, subjective testing

1 INTRODUCTION

An important parameter in many digital communication links, and especially in digital radio links, is bit error rate (BER). Error detecting codes and error correcting codes can increase robustness to bit errors. When the desired signal becomes sufficiently weak, or interference becomes sufficiently strong, uncorrected bit errors will inevitably result. When uncorrected bit errors are introduced between a speech encoder and a speech decoder, the quality of the received speech is reduced.

The TIA Project 25 Full Rate (FR) speech codec (encoder/decoder pair) is the cornerstone of Public Safety digital voice communications in the United States [1]. The encoder produces 4400 bits/second of compressed speech data. This data is protected by several different error correcting codes resulting in a 7200 bits/second data stream that can then be transmitted by radio. Some bit error patterns will be corrected by this scheme. Patterns associated with poor radio channels will not be corrected and the quality of the speech produced by the decoder will suffer. For this codec,

*The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, Colorado 80305.

a BER of 2% has been used as a speech quality benchmark. In this report, all BER values refer to randomly distributed or independent bit errors.

Since the introduction of the FR speech codec, the TIA Project 25 Enhanced Full Rate (EFR) and Enhanced Half Rate (EHR) have been introduced [2]. Like the FR codec, the EFR codec uses a net data rate of 4400 bits/second and a gross data rate of 7200 bits/second. True to its name, the EHR codec has a net rate of 2450 bits/second and a gross rate of 3600 bits/second.

Our first goal was to characterize the BER values for which the EFR codec produces speech quality that is equivalent to the benchmark associated with the FR codec when it is subjected to 2% BER. Similarly, our second goal was to characterize the BER values for which the EHR codec produces speech quality that is equivalent to that same benchmark. This benchmark is key to all the work described in this report and we use the terms "reference speech" and "reference condition" to point to this benchmark. That is, "reference condition" refers to FR encoding, followed by 2% random bit errors, then FR decoding. "Reference speech" refers to speech that has been processed by the reference condition.

In order to complete our goals, we designed a subjective listening experiment. An obvious choice when comparing speech codecs in subjective experiments is a paired-comparison test. Paired-comparisons are more sensitive than single-stimulus evaluations, and the listener's job is very concrete and objective. In such a test, two recordings are played and listeners are then asked to indicate which stimulus they prefer. In this case, one recording is reference speech, and the other is from the EFR or EHR codec using the current BER value under test.

In order to use paired-comparison testing to find the BERs in question, one would have to select a range of candidate BER values and sample this range with some BER step size determined in advance. This could result in a very large experiment, or rather coarse BER steps.

Instead, we significantly reduced the size of the experiment while maintaining high BER precision by adapting a new subjective testing technique. This technique, described in [3], uses a listener's votes to intelligently adapt the test contents in order to quickly zero in on the desired point in a parameter space. In this case, that means adapting the current BER value under test for the EFR or EHR codec in order to quickly zero in on the BER value that gives a speech quality match between the EFR or EHR codec and the reference speech.

Section 2 describes the listening experiment and a set of additional experiments that use a popular objective estimator of speech quality. We give special attention to the speech recordings used, the processing conditions and methods, the experiment procedure, and how the listeners interacted with the experiment. Section 3 presents the results of these experiments. Finally, we present and discuss the conclusions that can be drawn from this work.

2 EXPERIMENT DESCRIPTION

The goal of the listening experiment was to characterize the BER values for which the EFR or EHR codec produces speech quality that is equivalent to the benchmark associated with the FR codec when it is subjected to 2% BER. For simplicity, we refer to the EFR or EHR codec as the codec under test (CUT), and the FR codec at 2% BER as the reference condition.

The TIA Project 25 FR, EFR, and EHR encoders decompose input speech according to a multiband excitation (MBE) model. This model allows for a mixture of voiced and unvoiced components in different spectral bands. This model is efficient and can provide good speech quality at low bit rates. It is also robust in the sense that speech quality can be maintained even when significant acoustic background noise is combined with the input speech. This makes the family of codecs well-suited for use in the two-way land-mobile radio systems used by public safety officials. Further details are available in [1, 2, 4].

The listening experiment uses an adaptive paired-comparison technique to efficiently arrive at BER values that produce speech quality matches. From the listener's perspective, the experiment procedure is quite simple. This apparent simplicity is made possible by fairly complex control and bookkeeping mechanisms operating behind the scenes. In order to understand the system, we must first look at its parts. At the most basic level, there are source sentences. These are the raw materials that will be processed and presented to listeners. There are also three codecs: FR, EHR, and EFR.

Each source sentence must be processed with the EHR and EFR codecs at varying BERs to facilitate comparison with the reference condition. We can represent each source sentence/codec combination with the term SSxCUT. The goal is to find a BER where each SSxCUT is judged to have the same perceived speech quality as the reference condition. In order to do this, we present a listener with the reference speech paired with the SSxCUT processed at varying BERs. Each time a SSxCUT is processed with a different BER, the listener compares it with the reference speech and indicates which he prefers, or indicates that he has no preference. The result of this trial provides information that allows the software to intelligently select the next BER to evaluate for that SSxCUT. The iterative process of presenting each SSxCUT and its corresponding reference condition to a listener, recording a listener's vote, adjusting the BER accordingly, and (in the best case) eventually arriving at the BER that produces a speech quality match is called a "task."

Each listener was given 32 tasks to complete. The individual trials that comprise each task were presented in a random order. Thus, on each trial a listener contributes to the completion of one of the 32 tasks.

2.1 Speech Recordings

The speech recordings used in this experiment come from a database that was recorded at the Institute for Telecommunication Sciences. The database contains four men and four women speaking Harvard phonetically-balanced sentences [5]. All speakers were recorded using high quality recording equipment in a sound isolated chamber with an average background acoustic noise level

of less than 20 dBA. Each person spoke a total of sixteen sentences for a total of 128 recordings. The recordings were processed with two different software tools defined in [6]. The first process was sample rate conversion from 48,000 samples/second to 8,000 samples/second using a 160 to 3640-Hz bandpass filter (the G.712 or "PCM" filter). In the second process all recordings were normalized to have an active speech level matched to the software speech codecs described below. That level is 25 dB below the level of the maximum unclipped sine wave in a 16 bit representation. The average length of the recordings was 2.2 seconds. We refer to each of these processed recordings of sentences as a "source sentence."

In order to distribute these recordings evenly among 16 listeners, the 128 source sentences were divided into eight groups. Two sentences from each speaker were placed in a group, resulting in a total of sixteen sentences per group. Each group served as the basis for 32 tasks (16 for EHR, 16 for EFR) for two listeners. This use of groups allowed for the inclusion of a wider selection of speech and speakers in a balanced fashion while maintaining a manageable experiment length.

2.2 Experiment Procedure

Listeners took part in the experiment one at a time. Upon arrival, each listener read and signed an informed-consent form and was then seated in a sound isolation chamber with an average background noise level of less than 20 dBA. The experiment administrator then read scripted instructions to each listener. A PDA was used to interact with the experiment software and was connected via wireless ethernet and a VNC program to a computer that displayed the human interface.

The computer that ran the experiment software also produced the signals heard by listeners via a high quality sound card (the Mia card from Echo Digital Audio Corporation). The listening instrument was a pair of Model RS1 headphones, powered by a Model RA1 amplifier, both produced by Grado Labs. Each listener was instructed to place the headphones on his head such that the earpieces covered his ears. Each listener was instructed to adjust the volume control on the RA1 amplifier to his or her preferred listening level.

After the door to the sound isolation chamber was closed, listeners were instructed to use the PDA's stylus to push a button displayed on the PDA labeled "Begin." The interface shown in Figure 1 then appeared on the PDA's display. A source sentence was processed by two different codecs and the results were then played through the headphones one right after another. There was a 0.5 second delay between the playing of the two recordings.

After both processed recordings finished playing, the listener was instructed to answer "Which version do you prefer?" The listener's choices were "first," "no preference," or "second." Listeners were allowed to replay the pair of processed recordings by using the stylus to push a button labeled "Play Again." Doing so replayed both processed recordings just as before. Listeners were allowed to replay the processed recordings until they were ready to enter their vote. After a vote had been submitted, software processed the next source speech sample for presentation to the listener. When processing was complete, the next pair of processed recordings played through the headphones. This procedure continued until the software indicated all necessary data had been collected.

Figure 1. The human interface used during the experiment.

This experiment differs from many in that the speech signals to be presented on a given trial were generated on-the-fly immediately before that trial. In the worst case, the time required to do this was 750 ms. Thus, the on-the-fly generation was not perceived as an unnatural delay in the experiment by the listeners.

2.3 Listeners

Sixteen listeners participated in the experiment. These listeners were a subset of a large pool of employees that were randomly selected from the U.S. Department of Commerce Boulder Laboratories telephone directory and invited to participate. A total of twelve males and four females participated. Three of the listeners were estimated to be in their 20's, five in their 30's, six in their 40's, and two in their 50's. Two of the females were non-native speakers of English. Listeners were not familiar with the goals of the experiment nor the speech codecs under test. Depending on the listener, the duration of the experiment ranged from 10 to 30 minutes.

2.4 Speech Quality Matching Algorithm

In addition to the functions already identified, the experiment control software also implements a speech quality matching algorithm (QMA). This algorithm seeks to find a BER value where the CUT has the same perceived quality as the reference condition. That is, when the CUT and the reference condition are played sequentially as a pair with randomized order, the listener votes "no preference." Thus the algorithm searches a line segment in BER space extending from 0% to 8% BER. This upper limit of 8% was chosen after a preliminary listening to CUT recordings. It was chosen to be well above the range where "no preference" votes would be likely to occur.

In order to avoid conducting an exhaustive search of the line segment, we adapt the gradient ascent paired-comparison method described in [3]. Adaptation is necessary because the original method locates an area of maximum quality in parameter space, but the present problem requires quality matching.

Since we are comparing either CUT to the reference condition, one recording in the paired comparison pair will always be the reference condition. The other recording in the pair is the CUT at a given BER. The matching algorithm iterates to find a point of equivalence for each SSxCUT. As mentioned in Section 2, the iteration to find a point of equivalence is also called a task. The result of each task is a BER value for the SSxCUT.

For each task the algorithm starts with random BER, b, drawn from the uniform distribution between 0% and 8%. A source sentence is processed by the CUT encoder, and the resulting channel file is modified to reflect the BER b. The modified channel file is then decoded by the CUT decoder. The software then encodes the same source sentence with the FR encoder, modifies the resulting channel file to reflect a 2% BER, and finally uses the FR decoder to decode said channel file to create the reference condition. The channel file modification process is described in Section 2.5. Both recordings are then presented to the listener as described in Section 2.2.

Instead of collecting subjective scores as described in [3], we ask listeners "Which version do you prefer?" Listeners may respond with "first," "no preference," or "second." Selecting "first" suggests that the first recording played has a higher quality than the second. Likewise, selecting "second" suggests that the second recording played has a higher quality than the first, and "no preference" indicates that the listener could not distinguish the quality between the two.

After the listener has entered a vote the software stores it and performs a calculation. If the vote was "no preference," it stores the BER b as a point of equivalency for the source sentence processed by the CUT. Otherwise b is adapted, following the premise that BER and speech quality are inversely related.

The adaptation of b is described in (1). L_{lim} and U_{lim} are the current lower and upper limits of the line segment to search and are initialized to 0 and 8% respectively. If the listener preferred the CUT over the reference, the quality of the CUT (Q_{CUT}) must be greater than the quality of the reference condition (Q_{REF}), and therefore the point of equivalence must be a BER greater than the BER b. Hence the upper branch of (1) is used to update b from $b_{current}$ to b_{new} and L_{lim} is set to $b_{current}$ as well.

$$b_{new} = \begin{cases} \dfrac{b_{current} + U_{lim}}{2}, & Q_{CUT} > Q_{REF} \\ \dfrac{L_{lim} + b_{current}}{2}, & Q_{CUT} < Q_{REF} \end{cases} \tag{1}$$

If the listener preferred the reference condition over the CUT, Q_{CUT} must be less than Q_{REF}, and therefore the point of equivalence must be a BER less than the BER b. Hence, the lower branch of (1) is used to update b from $b_{current}$ to b_{new} and U_{lim} is set to $b_{current}$ as well.

In both of these cases ($Q_{CUT} > Q_{REF}$ and $Q_{CUT} < Q_{REF}$), the result of (1) gives the BER b_{new}. BER b_{new} is then stored and used to modify the channel file during the next iteration.

The software uses (1) to bisect a search segment until a listener reports both processed recordings have the same quality, or the listener has voted ten times on a given task. This limit was determined heuristically and represents a good tradeoff; obtaining as much information as is practical without overburdening the listener. After ten iterations, the length of the line segment in BER space that remains to be searched will be between $2^{-10} \times 8\%$ and $2^{-9} \times 8\%$ (i.e., 0.008% to 0.016%) depending on the initial random starting point b.

2.5 Speech Coding and Bit Errors

When the experiment control software asked for a speech signal from a particular codec at a particular BER, the following steps, shown graphically in Figure 2, were executed. Licensed reference software provided by Digital Voice Systems Inc. (DVSI) was used to encode the 16 bit/sample speech file using either the FR, EFR, or EHR encoder. The software encoders performed speech coding and forward error correction (FEC), resulting in channel files that contained representations of 7200 or 3600 bits/second data streams.

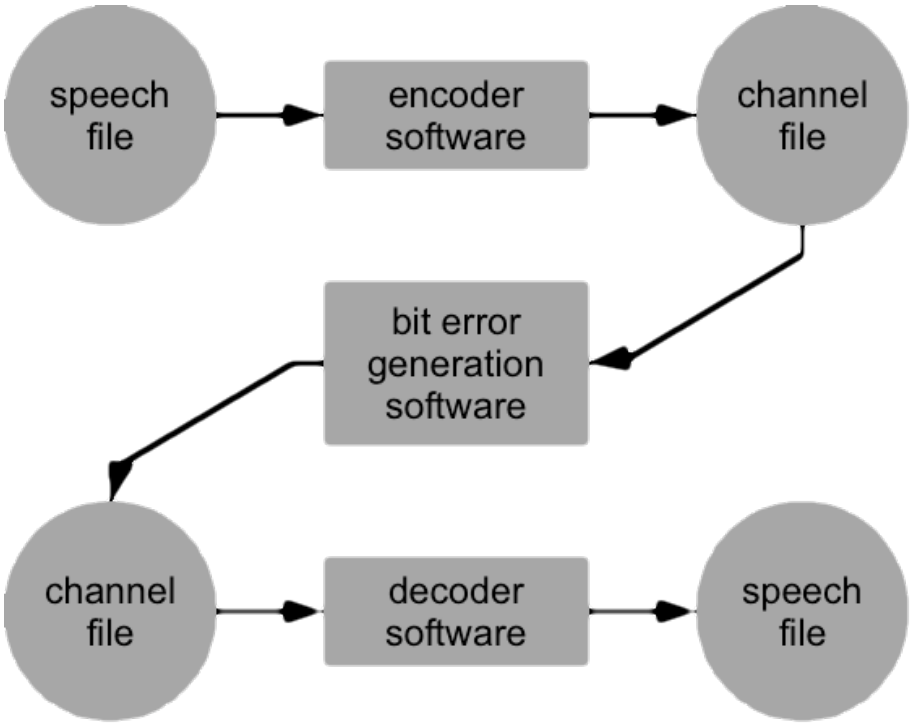

Figure 2. Signal flow through the encoding, bit error, and decoding processes.

These channel files were then read and processed to induce bit errors at the requested BER. First, the total number of bits read from the channel file were counted, multiplied by the BER, and rounded to the nearest integer in order to find the exact number of bits *NE* to invert. Next, the first *NE* locations were extracted from the appropriate master bit error location pattern (master bit error

location patterns are described below). Then the bits at those *NE* locations were inverted. The other bits remained untouched, and the resulting bitstream was written to a new channel file. Note that this procedure achieves a BER that matches the target BER exactly (within the constraint that the total number of bits in error must be an integer).

Finally, licensed reference software provided by DVSI was used to decode the new channel file to produce the requested 16 bit/sample speech file. The decoder was instructed to use hard decision decoding when reading processed channel files.

The bit error generation process described above includes the use of multiple master bit error location patterns. These patterns are simply lists of locations. Each list includes enough locations to allow for the highest BER needed in the experiment. We use these patterns to appropriately control variance in speech quality due to locations of bit errors. More specifically, in order to obtain the most accurate and efficient speech quality matches, we need to minimize this source of variance within the trials that comprise a given task. In order to accurately reflect actual operating conditions, we need to allow this source of variance to appear between tasks. Thus, when a bit error pattern is needed for a given task, either to generate the reference (FR) speech, or the CUT (EFR or EHR) speech, the same master bit error location pattern is always used for that task. There is one master pattern for each task. If a given task is done by more than one listener, a different master pattern is used for each listener.

The use of a master bit error location pattern for each task results in a bit error location subset property: If $B(NE)$ is the set of bit error locations used for a given task and listener when *NE* bit errors are needed, then

$$NE_1 < NE_2 \Rightarrow B(NE_1) \subset B(NE_2). \tag{2}$$

This subset property follows directly from the fact that the master bit error location pattern is a list of locations and we always extract the needed number of locations from the front of the list. The bit error location repeatability property follows:

$$NE_1 = NE_2 \Rightarrow B(NE_1) = B(NE_2). \tag{3}$$

Figure 3 provides an example of a set of bit error patterns generated by this technique. The patterns cover a group of 100 bits, and the value 1 is used to indicate a bit error. The top, middle and bottom panels show bit error patterns for BERs of 15, 10, and 5% respectively ($NE = 15$, 10, and 5).

Due to the bit error location subset property, when the BER of the CUT is increased, new bit error locations are added to the existing bit error locations. When the BER of the CUT is decreased, bit error locations are removed from the existing bit error locations. The master bit error location patterns provide a simple way to achieve these properties while maintaining randomly distributed bit error locations.

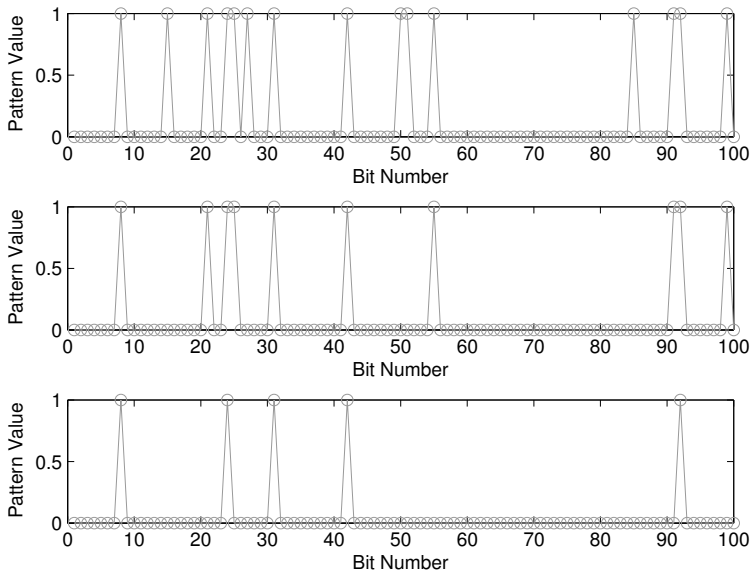

Figure 3. Example bit error patterns that conform with the subset property. BER is 15, 10, and 5% in the top, middle and bottom panels, respectively.

For example, if the CUT BER is above 2%, the CUT bit error locations are all of the reference bit error locations plus additional bit error locations. If the CUT BER is exactly 2%, the CUT bit error locations match the reference bit error locations. And if the CUT BER is below 2%, the CUT bit error locations are a subset of the reference bit error locations.

There is one additional complexity to address. For tasks where the CUT is EHR, we need a bit error pattern of length *NB* bits with BER $= 2\%$ for the FR codec and a pattern of length $NB/2$ bits with BER $= b$ for the EHR codec. We use a single master bit error location pattern to create a pattern of length *NB* bits and BER $= 2\%$ for the FR codec and a pattern of length *NB* bits with BER $= b$ for the EHR codec. We then subsample the second pattern by a factor of two to arrive at a pattern with $NB/2$ bits with BER $\approx b$. Next we add or remove bit errors at random locations as necessary to force BER $= b_2$. This time-domain subsampling is a desirable technique because it preserves the BER and the temporal relationships among the error patterns, the speech signals, and the speech coding frames to the extent possible.

Finally, we describe how the locations of errored bits can influence the resulting decoded speech quality. There are three distinct factors to consider. First, the location of errored bits relative to each other will influence how successful FEC can be at correcting those errors. Even randomly distributed bit errors will show different amounts of clustering from trial to trial. Second, once FEC has failed, the location of the uncorrected bit errors within each frame of encoded speech data determines the speech coding parameter(s) that are corrupted and the way in which they are corrupted. Finally, the location of a frame containing one or more corrupted speech coding parameters relative to the speech signal will determine how audible and annoying that corruption is once it has passed through the decoder.

2.6 Bookkeeping

Our software treats each task as a separate entity. Upon initialization, memory space is allocated to store all necessary data and metadata for each task. All memory required for every task is organized into a one-dimensional array where each task has its own index in the array. Memory for a given source sentence and both CUTs is allocated at the same time. Random starting values of BER are drawn from the uniform distribution at this time as well. The starting place, $b_{initial}$, is used as the starting place for the given source sentence with both EHR and EFR codecs. When the equivalency algorithm is initiated for a given source sentence and either CUT, $b_{initial}$ is the same.

Each listener was associated with one group of source sentences. A group of source sentences consists of 16 sentences, and for each of these a speech quality match must be made for the EHR codec and the EFR codec. Thus each listener performed 32 tasks.

Tasks are presented to each listener in a random order. That is, with each trial the listener makes one step of progress on one task, and that task is taken at random from the list of all unfinished tasks. When a task encounters a terminating condition in the matching algorithm, that task is removed from the list of unfinished tasks. When the list of unfinished tasks is empty, the listener has finished the experiment.

Each listener used a unique set of starting BER values that were generated at the beginning of an experiment session. The seed used to generate these unique starting points was an input string specified by the experiment administrator that was unique for each listener. This also caused each listener to hear a unique set of bit error patterns.

2.7 Objective Speech Quality Estimation

Listening experiments directly access human hearing and human judgement. This means they are well suited to answering questions like "How does this system sound?" or "Which system sounds better?" However, listening experiments require significant resources and can take weeks or months to design, implement, and analyze. Objective estimators of speech quality seek to provide similar information through digital signal processing algorithms. These algorithms can replace weeks or months of listening experiment work with minutes or hours of computer processing time, but they can only provide estimates of perceived speech quality.

The most effective objective speech quality estimator presently available (and also the most popular) is the "Perceptual Evaluation of Speech Quality" (PESQ) algorithm [7, 8, 9]. Measuring speech quality produced by MBE codecs in the presence of bit errors is not within the application scope of the PESQ algorithm (see Tables 1 and 3 in [9]) so our use of PESQ here is purely experimental. Having completed a formal listening experiment, it is natural to ask how the results would compare with those produced by the best available objective estimator. The PESQ algorithm compares an original and distorted speech signal. It produces a raw quality score that ranges from -0.5 to 4.5 and a second value called MOS-LQON, an acronym for Mean Opinion Score, Listening Quality, Objective, Narrowband [10] that ranges from 1.0 to 4.5. MOS-LQON

was developed to allow direct comparison with mean opinion scores from subjective tests. We use the MOS-LQON output from PESQ to compare with our subjective results.

We also use a previously explored variant of PESQ that applies PESQ creatively in an attempt to more accurately account for quality loss due to channel interference. This method, known as DPESQ (where "D" means disturbance)[1], uses multiple iterations of the PESQ algorithm and some simple math to determine a score. In order to calculate DPESQ, several versions of a signal are needed. The original source sentence, *SS*, and the source sentence processed by the CUT using a BER of 0%, SS_{CUT}, are compared by the PESQ algorithm, and a score δ_{clean} is stored according to the following equation:

$$\delta_{clean} = 4.5 - \text{PESQ}(SS, SS_{CUT}). \tag{4}$$

Then SS_{CUT} is compared with the source sentence processed by the CUT using a given BER, SS_{BER}, and a score δ_{BER} is stored according to (5).

$$\delta_{BER} = 4.5 - \text{PESQ}(SS_{CUT}, SS_{BER}) \tag{5}$$

A measure of total disturbance, δ_{total}, is calculated:

$$\delta_{total} = \sqrt{{\delta_{clean}}^2 + {\delta_{BER}}^2}. \tag{6}$$

Finally, a DPESQ score is calculated:

$$\text{DPESQ} = 4.5 - \delta_{total}. \tag{7}$$

One approach we used to compare the subjective and objective results was to keep the listening experiment design and QMA intact, and simply replace each of the 16 listeners with the PESQ and DPESQ algorithms. (Since different listeners heard different source speech and different random starting BER values, applying PESQ or DPESQ in place of each listener does not yield 16 identical sets of results as it would in some experiments.)

Using PESQ, each pair of recordings was used to generate a vote: either "first," "second," or "no preference." PESQ produced a MOS-LQON speech quality estimate S_1 for the first recording and a MOS-LQON estimate S_2 for the second recording. The PESQ vote was calculated as follows:

[1] Alan Wilson, "DPESQ and MOS for Phase 2," Presentation to the TIA APCO Project 25 Interface Committee Vocoder Task Group, Document Number 08-013-VTG, Apr. 22, 2008.

$$\begin{array}{ccccc} \Delta & < & S_1 - S_2 & \text{First,} \\ |S_1 - S_2| & <= & \Delta & \text{No Preference,} \\ S_1 - S_2 & < & -\Delta & \text{Second,} \end{array} \tag{8}$$

where Δ is a "sensitivity" parameter discussed in Section 3.2 below. Once the decision had been calculated, the result was given to the experiment system which would then calculate and produce the next pair of recordings to evaluate. This process continued until the same termination conditions imposed on the listening experiment were met.

In order to use DPESQ in this construct, we used MOS-LQON values to perform the calculations needed to create a DPESQ score (not raw PESQ scores). This allows us to compare results more directly with the PESQ and subjective results. DPESQ scores, DS_1 and DS_2, were calculated for the first and second recordings in each pair, respectively. The DPESQ vote was then calculated as shown in (8), substituting DS_1 for S_1 and DS_2 for S_2.

While the experiment design and QMA used were identical to those used during the listening experiment, each objective test had a unique set of starting places as well as unique bit error patterns.

Another method, a more common exhaustive search (ES) of objective scores in the BER space, was also conducted. We used PESQ and DPESQ to directly estimate MOS-LQON scores for all of the source sentences used in this experiment at BERs ranging from 0% to 8% with an interval of 0.1%. We then looked at the mean MOS-LQON value and the confidence interval that the FR codec achieved at 2% BER over all of the source sentences, and determined at which BERs the EFR and EHR codecs perform equivalently.

3 RESULTS

In this section we describe in detail the results of the listening experiment. Next we present the results of the parallel efforts with an objective estimator of speech quality. Finally we describe our work to verify the randomness of the bit error patterns used in the experiments.

3.1 Listening Experiment

As previously described, each listener performed 16 speech quality matching tasks for the EFR codec and 16 tasks for the EHR codec. Sixteen listeners participated in the experiment, so 512 tasks were performed: 256 for EHR and 256 for EFR.

The sixteen listeners completed a total of 1292 paired-comparison trials in the listening experiment. Of these trials, 596 were trials in EFR tasks and 696 were trials in EHR tasks. The comparisons between EFR and FR resulted in a "no preference" vote 247 times, thus terminating 247 of the 256 EFR tasks. The comparisons between EHR and FR resulted in a "no preference" vote 244 times, thus terminating 244 of the 256 EHR tasks. Thus 491 of the 512 tasks were terminated due to a vote of "no preference" and the other 21 tasks were terminated because the limit of 10 trials per task had been reached.

Thus 41% of the EFR trials produced a speech quality match and 35% of the EHR trials did the same. In other words, relative to an exhaustive search of the BER line segment, a small amount of effort was expended to find points of equivalency. The average task length for EFR was 2.3 trials, and the average task length for EHR was 2.7 trials. This indicates that the listeners were not burdened with an unreasonably large number of trials.

The key results from this experiment are the BERs for which a pair formed by a CUT and the reference is voted "no preference." Figure 4 shows a histogram of BERs where EFR-FR pairs were voted "no preference." The central 95% of the results are shown; 2.5% of the results have been removed from each tail of the histogram. The votes in the central 95% of results range in BER from 0.29% to 7.75%, the mean BER value is 3.70%, and the median value is 3.69%. If all results are considered, calculating a mean results in a BER of 3.71%, and this mean has a 95% confidence interval that covers the interval from 3.47% to 3.96%. More detailed statistics can be found in Table 1.

Table 1. Results of the Listening Experiment

	% BER					
	min	max	mean	median	% complete	avg. trials
EHR	0.22	6.63	2.92	2.86	95	2.72
EFR	0.29	7.75	3.70	3.69	96	2.33

Likewise Figure 5 shows a histogram of BERs where EHR-FR pairs were voted "no preference." The votes in the central 95% of results range in BER from 0.22% to 6.63%, the mean BER value

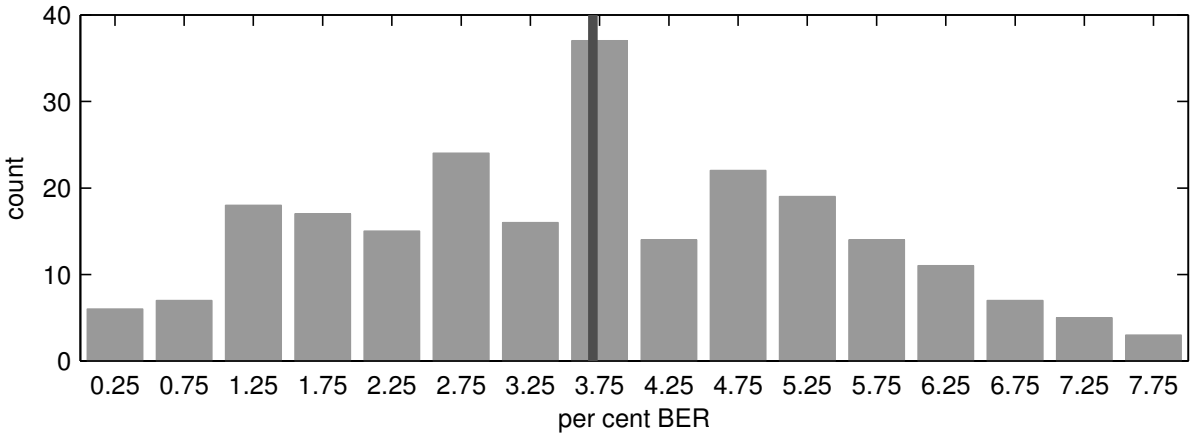

Figure 4. Histogram of BER equivalence values for the EFR codec, mean of all values indicated by the dark gray line.

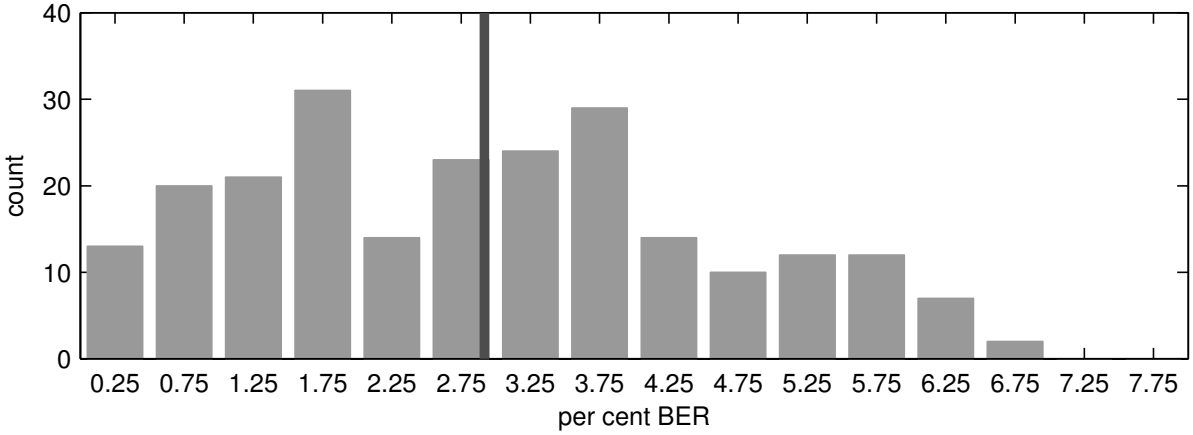

Figure 5. Histogram of BER equivalence values for the EHR codec, mean of all values indicated by the dark gray line.

is 2.92%, and the median is 2.86%. If all results are considered, calculating a mean results in a BER of 2.96%, and this mean has a 95% confidence interval that covers the interval from 2.73% to 3.18%.

3.2 Objective Estimation

Section 2.7 describes parallel experiments using objective estimators of speech quality called PESQ and DPESQ. (8) shows how PESQ and DPESQ results are translated to automated votes using the parameter Δ. We investigated Δ values of 0.05, 0.1, 0.2, 0.25, and 0.5. Given the decision criteria listed in (8), using $\Delta = 0.05$ would model a highly sensitive listener while using $\Delta = 0.5$ would model a less sensitive listener. For each value of Δ, the objective estimator performed the experiment using the same QMA as the listening experiment. Table 2 shows the results obtained for each value of Δ, along with the listening experiment results for comparison. Values

in the column marked "% BER" describe point of equivalence statistics.

Table 2. Results of the Listening Experiment and Objective Experiment

			% BER					
		Δ	min	max	mean	median	% complete	avg. trials
listening	EHR	n/a	0.22	6.63	2.92	2.86	95	2.72
	EFR	n/a	0.29	7.75	3.70	3.69	96	2.33
QMA PESQ	EHR	0.5	0.20	5.45	2.58	2.69	100	1.54
		0.25	0.21	4.08	1.93	1.82	97	2.24
		0.2	0.19	4.41	1.87	1.84	94	2.73
		0.1	0.09	4.42	1.70	1.61	83	3.88
		0.05	0.10	4.07	1.68	1.57	71	5.14
	EFR	0.5	0.20	6.42	3.15	3.21	100	1.32
		0.25	0.39	5.97	2.98	3.06	98	1.71
		0.2	0.38	5.71	2.96	3.05	96	2.07
		0.1	0.47	5.84	2.94	2.79	87	3.43
		0.05	0.48	5.71	3.15	3.20	73	4.94
QMA DPESQ	EHR	0.5	0.28	5.09	2.61	2.75	100	1.58
		0.25	0.42	4.80	2.42	2.34	97	2.12
		0.2	0.25	4.61	2.21	2.03	94	2.55
		0.1	0.26	4.80	2.26	2.10	89	3.55
		0.05	0.23	5.12	2.48	2.43	75	4.84
	EFR	0.5	0.28	6.67	3.26	3.37	99	1.38
		0.25	0.42	6.18	3.35	3.36	98	1.80
		0.2	0.50	6.23	3.27	3.32	97	2.13
		0.1	0.61	6.40	3.51	3.59	91	3.31
		0.05	0.94	6.11	3.76	3.82	72	5.15

The results from the objective estimators do not reproduce the listening experiment results, but the DPESQ results come closer than the PESQ results. For both PESQ and DPESQ the range of BER values is narrower than those produced by the listening experiment. For the PESQ results, the mean and median BERs of those ranges are markedly lower than those produced by the listening experiment. However, when $\Delta = 0.2$ the completion rate (94% for EHR, 96% for EFR in the PESQ results, 94% for EHR, 97% for EFR in the DPESQ results) and average number of trials per task (2.7 trials per task for EHR, 2.1 trials per task for EFR in the PESQ results, 2.6 trials per task for EHR, 2.1 trials per task for EFR in the DPESQ results) resemble those of the listening experiment. Perhaps this Δ value models the average discrimination abilities of the listeners in the listening experiment.

In the second part of our objective investigation, we used both PESQ and DPESQ to survey the BER space for each codec. PESQ gave an estimate for the reference condition of 3.28 with a 95% confidence interval of ± 0.05 points, while DPESQ gave an estimate of 3.22 with a 95% confidence interval of ± 0.05 points.

The PESQ estimate indicated that the EHR codec would have the same mean quality at 0.6% BER. In order to state a range of BERs where we are 95% certain that the equivalent BER lies we need to take the 95% confidence intervals on the means into account. This results in the rather wide range from 0 to 2% BER.

Similarly, the PESQ estimate indicated that the EFR codec provided mean equivalent quality at 2.5% BER, with an equivalence range of 0 to 3.6% BER. The mean BERs calculated by this exhaustive search approach are significantly lower than those of our subjective results, and the equivalence ranges exclude what the subjective results show are points of equivalence.

The DPESQ exhaustive search estimate indicated that the EHR codec provided equivalent quality at 1.9% BER, and had an equivalence range of 0 to 3.1% BER. Similarly, the DPESQ estimate indicated that the EFR codec provided equivalent quality at 3.5% BER and had an equivalence range of 2.2 to 4.3% BER. These results are closer to the subjective results than the PESQ results are, yet differences of up to 1% BER remain.

As described, DPESQ[1] uses raw PESQ scores. However, the MOS-LQON output from the PESQ algorithm is specifically designed for comparison with subjective results. We performed all objective measurements using both raw PESQ scores and PESQ MOS-LQON values. We found that results derived from MOS-LQON values do give better agreement with our subjective test results than those derived from raw PESQ scores and we have included only results derived from MOS-LQON in this report.

However, based on the lack of agreement between the listening experiment results and the PESQ and DPESQ results, we conclude that objective measures are not a suitable substitute for subjective listening experiments in this application area. This is consistent with the stated scope of the PESQ algorithm. Despite the increased performance that using DPESQ in either objective method provides, it is not safe to assume that DPESQ is a suitable replacement for subjective tests in this context.

3.3 Bit Error Statistics

The constraint when generating bit errors at various BERs was that individual bit errors must be equally likely to occur at any given position in a channel file and that probability is the BER. In other words, the bit errors are to be random. If bit errors are not random they are said to be bursty. With bursty errors the probability of an error in bit position n is increased when there is an error in bit position $n-1$. Because we used known seeds to generate all necessary bit error patterns (via master bit error location patterns), it is possible for us to recreate all bit error patterns and test them for randomness.

Let P_{UE} be the unconditional probability of bit error. Our empirical estimate of P_{UE} is (total bit errors) / (total bits). Let P_{CE} be the conditional probability of bit error. Our empirical estimate of P_{CE} is found by analyzing only the bit positions that immediately follow a bit error. If there are NE such positions then P_{CE} is estimated as (number of the NE bit positions that have a bit error) / NE.

A bursty bit error pattern will have $P_{CE} > P_{UE}$, meaning it would be more likely that an error occurred directly after another error. A random bit pattern will have $P_{CE} \approx P_{UE}$, where an error is equally likely to occur on its own as it is to occur after another error. Thus $P_{CE} - P_{UE}$ is a measure of burstiness.

We are interested in examining the bit error patterns associated with "no preference" votes. These bit error patterns are interesting because it is important to ensure that a less-than-random pattern did not cause adverse quality effects during testing. Figure 6 shows the burstiness measurement for bit error patterns associated with the "no preference" vote for the EFR codec. The figure shows that P_{CE} is not reliably greater or less than P_{UE} at any BER and we conclude that the error patterns are random, not bursty. The variation in $P_{CE} - P_{UE}$ reflects the trial-to-trial variation inherent when small samples of a random process are observed.

Note that good estimates of P_{CE} become increasingly difficult as BER goes to zero because *NE* becomes small so the estimates are "data poor." In Figure 6 P_{CE} has been estimated as zero for a number of BER values between 0 and .005. When this happens $P_{CE} - P_{UE}$ becomes $-P_{UE}$ and this generates the line with slope -1 in the lower left.

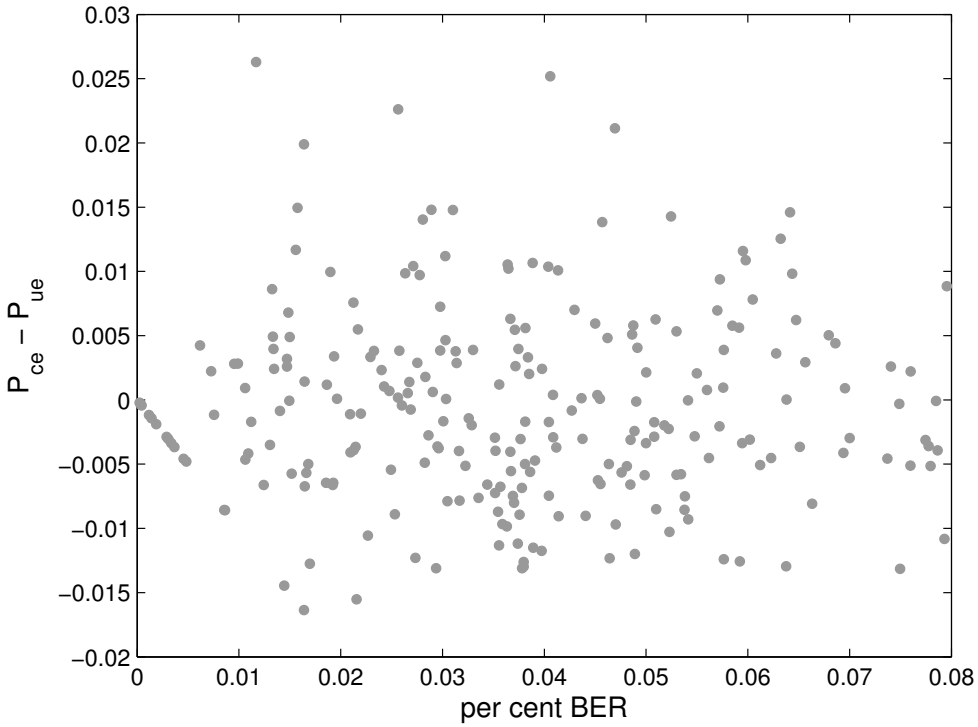

Figure 6. Burstiness measurements for the EFR BER points of equivalence.

4 CONCLUSIONS

We adapted an efficient new paired-comparison subjective testing method to find BER values where the EFR and EHR codecs produce speech quality equivalent to that produced by the FR codec at 2% BER. Sixteen listeners made 1292 comparisons that resulted in 491 votes for "no preference." The BER values associated with these "no preference" votes are the output of the experiment. Meticulous attention was paid to the generation of bit error patterns, and post experiment analysis confirms their randomness.

As both Figure 4 and Figure 5 show, there is a range of BER values that listeners perceive as sounding the same as the reference. This, however, is not unexpected and the variance can logically be attributed to sources from two main classes. The first class is the signal production class and includes the temporal relationships among the bit error patterns, the bits in each transmitted frame, and the speech content in each transmitted frame. The second class is the signal perception class and includes the hearing, personal auditory preferences, and other personal characteristics (diligence, patience, etc.) of the listeners who participated in the experiment.

It is critical that the distribution of BER values resulting from this work is not ignored, but is instead taken into consideration when setting radio channel performance goals, and when defining coverage areas. The distribution of BER values indicates that a distribution of user experiences can be expected depending on the factors cited above, and other factors (e.g., acoustic background noise at transmitting and receiving locations) that occur outside the laboratory environment.

Nonetheless, simplification is sometimes desirable in order to draw the most basic conclusions. Toward that end, we note that when one considers the measures of central tendency presented in this report (mean and median) to characterize the BER distributions in the most simplistic terms, those measures for EHR and EFR are higher than the 2% reference BER value used with FR. Thus on average, under the conditions of this experiment, both the EFR and EHR codecs appear to be more robust (in terms of decoded speech quality) to random bit errors than the FR codec.

We also observe that the PESQ algorithm, widely considered to be the most effective objective speech quality estimator presently available, produces results that are pessimistic (lower BER) when compared with this listening experiment. This lack of agreement is not unexpected as this experiment concerns an application area that is outside the stated scope of PESQ. DPESQ has been developed to address additional disturbances that were not considered during the development of the PESQ algorithm. DPESQ results agree with subjective test results better than PESQ results do. Even so, DPESQ does not serve as a reliable replacement for the subjective tests described here.

Additional insights might follow from knowledge of the *range* of BERs that would generate a "no preference" vote for a given task and a given listener. A modified version of this experiment could be designed to investigate this topic and is a potential topic for future work. In fact, there are multiple experimental approaches that could be applied to investigate the more general topic of listener resolution in paired-comparison subjective testing.

5 ACKNOWLEDGEMENTS

This work was funded by the Department of Homeland Security Office of Interoperability and Compatibility, through the National Institute of Standards and Technology Office of Law Enforcement Standards. The work was conducted at the Institute for Telecommunication Sciences under the supervision of J.R. Bratcher.

We also wish to thank the anonymous listeners from the US Department of Commerce Boulder Laboratories who participated in these experiments.

6 REFERENCES

[1] Telecommunications Industry Association, TIA-102.BABA-1998, "Project 25 Vocoder Description," May 1998. Reaffirmed Dec. 2008.

[2] D. W. Griffin and J. S. Lim., "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, Aug. 1988.

[3] S.D. Voran and A. Catellier, "Gradient ascent paired-comparison subjective quality testing," *Proc. First International Workshop on Quality of Multimedia Experience*, San Diego, CA, Jul. 2009.

[4] Telecommunications Industry Association, TIA-102.BABA, Annex A, "APCO Project 25 Half-Rate Vocoder Addendum," Apr. 2009.

[5] "IEEE recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, Sep. 1969.

[6] ITU-T Recommendation P.191, "Software tools for speech and audio coding standardization," Geneva, 2005.

[7] A. Rix, M. Hollier, A. Hekstra, and J. Beerends, "Perceptual evaluation of speech quality (PESQ)—The new ITU standard for end-to-end speech quality assessment, Part I—Time-delay compensation," *J. Audio Engineering Society*, vol. 50, pp. 755–764, Oct. 2002.

[8] J. Beerends, A. Hekstra, A. Rix, and M. Hollier, "Perceptual evaluation of speech quality (PESQ)—The new ITU standard for end-to-end speech quality assessment, Part II—Psychoacoustic model," *J. Audio Engineering Society*, vol. 50, pp. 765–778, Oct. 2002.

[9] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Geneva, 2001.

[10] ITU-T Recommendation P.862.1 "Mapping function for transforming P.862 raw result scores to MOS-LQO," Geneva, 2003.

FORM **NTIA-29** (4-80)

U.S. DEPARTMENT OF COMMERCE
NAT'L. TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO.	2. Government Accession No.	3. Recipient's Accession No.
TR-10-462		

4. TITLE AND SUBTITLE	5. Publication Date
Low Rate Speech Coding and Random Bit Errors: A Subjective Speech Quality Matching Experiment	Oct. 2009
	6. Performing Organization NTIA/ITS

7. AUTHOR(S)	9. Project/Task/Work Unit No.
Andrew A. Catellier and Stephen D. Voran	
8. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Telecommunication Sciences National Telecommunications & Information Administration U.S. Department of Commerce 325 Broadway Boulder, CO 80305	6513000-320
	10. Contract/Grant No.

11. Sponsoring Organization Name and Address	12. Type of Report and Period Covered
Office of Law Enforcement Standards National Institute of Standards and Technology 100 Bureau Drive, M/S 8102 Gaithersburg, MD 20899-8102	

14. SUPPLEMENTARY NOTES

15. ABSTRACT

When bit errors are introduced between a speech encoder and a speech decoder, the quality of the received speech is reduced. The specific relationship between speech quality and bit error rate (BER) can be different for each speech coding and channel coding scheme.

This report describes a subjective experiment concerning the relationships between BER and perceived speech quality for the TIA Project 25 Full Rate (FR), Enhanced Full Rate (EFR), and Enhanced Half Rate (EHR) speech codecs. Using the FR codec with 2% random bit errors as a reference, we sought to characterize the BER values for which the EFR (or EHR) codec produces speech quality that is equivalent to the reference. We used an adaptive paired-comparison subjective testing algorithm to efficiently adapt BER values for the EFR and EHR codecs to quickly locate the BER values where listeners found the speech quality to be the same as the reference.

The results from sixteen listeners reveal ranges of BER values that were judged to produce speech quality equivalent to the reference. When these ranges are reduced to central values, those values indicate that on average, the EFR and EHR codecs are more robust to bit errors than the FR codec. We provide a set of additional results from a popular objective speech quality estimator for comparison purposes.

16. Key Words
bit errors; listening tests; speech coding; speech quality; subjective testing

17. AVAILABILITY STATEMENT	18. Security Class. (This report)	20. Number of pages
☐ UNLIMITED.	Unclassified	20
	19. Security Class. (This page) Unclassified	21. Price:

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities. Subsets of this series include:

NTIA RESTRICTED REPORT (RR)

Contributions that are limited in distribution because of national security classification or Departmental constraints.

NTIA CONTRACTOR REPORT (CR)

Information generated under an NTIA contract or grant, written by the contractor, and considered an important contribution to existing knowledge.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail info@its.bldrdoc.gov.

This report is for sale by the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161,Tel. (800) 553-6847.