NTIA Report 17-522

Intelligibility of Selected Speech Codecs in Frame-Erasure Conditions

Andrew A. Catellier
Stephen D. Voran

report series

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

NTIA Report 17-522

Intelligibility of Selected Speech Codecs in Frame-Erasure Conditions

Andrew A. Catellier
Stephen D. Voran

U.S. DEPARTMENT OF COMMERCE

November 2016

DISCLAIMER

Certain commercial equipment and materials are identified in this report to specify adequately the technical aspects of the reported results. 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 the best available for this purpose.

PREFACE

The work described in this report was performed by the Public Safety Communications Research Program (PSCR) on behalf of the Department of Homeland Security (DHS) Science and Technology Directorate. The objective was to quantify the speech intelligibility associated with selected digital speech coding algorithms subjected to erased data frames. This report constitutes the final deliverable product for this project. The PSCR is a joint effort of the National Institute of Standards and Technology and the National Telecommunications and Information Administration.

CONTENTS

FIGURES

TABLES

ABBREVIATIONS/ACRONYMS

3GPP	Third Generation Partnership Project
AES/EBU	Audio Engineering Society/European Broadcasting Union
AMR	Adaptive Multi-Rate
AMR-WB	Adaptive Multi-Rate Wideband
ANSI	American National Standards Institute
b/smp	bits/sample
CA	Channel Aware
DHS	Department of Homeland Security
EVS	Enhanced Voice Services
FEC	Forward Error Correction
FER	Frame Erasure Rate
ITU-T	International Telecommunication Union, Telecommunication Standardization Sector
kb/s	kilobits per second
LTE	Long-Term Evolution
MCV	Mission-Critical Voice
MRT	Modified Rhyme Test
ms	millisecond
NPSTC	National Public Safety Telecommunications Council
PSCR	Public Safety Communications Research Program
smp/s	samples per second
SNR	Signal-to-Noise Ratio
SWB	Super Wideband
TS	Technical Specification

USB	Universal Serial Bus
WB	Wideband

EXECUTIVE SUMMARY

The importance of speech intelligibility in mission critical voice (MCV) services is undisputed. When MCV services are supplied over the LTE wireless infrastructure, speech intelligibility can be reduced by degradations in the Radio Access Network (RAN). The relationship between RAN degradation and speech intelligibility reductions is mediated by the robustness of the speech codec selected for the MCV service. More robust codecs can tolerate more degradation while producing a given level of speech intelligibility. Characterizing the per-codec relationships between RAN degradation and speech intelligibility is particularly important for MCV because the events that stress the RAN may very well be events that also have critical intelligibility requirements.

This report describes the design, implementation, and analysis of a speech intelligibility test designed to characterize those relationships for five important speech codec modes. The five codec modes were selected for their potential compatibility with LTE. The test includes four levels of RAN degradation, characterized by four frame-erasure rates. The frame erasure statistics were selected to match those used in LTE standardization work. The test includes two background noise environments: coffee shop noise and siren noise.

The test protocol required twenty listeners to repeat all words that they heard in short messages with median length of seven words. This was done using professional audio equipment in sound-isolated chambers. Each of the 40 test conditions was tested using approximately 1100 words. Listeners' responses were scored against the original message transcripts to produce a count of words correctly repeated and thus a measure of speech intelligibility. The results are presented in tabular and graphical form. They show exactly how this measure of speech intelligibility drops as frame-erasure rate increases for three of the five codec modes. These results can support the goal of properly provisioning radio resources based on the most critical user experience factor—speech intelligibility.

The remaining two codec modes did not produce valid results due to defects in the reference software provided to us. This unexpected result is also valuable in that it reinforces the importance of thorough testing before a codec is selected for MCV services. These two codec modes were specifically designed for enhanced robustness, so additional testing with stable updated software is an important goal.

INTELLIGIBILITY OF SELECTED SPEECH CODECS IN FRAME-ERASURE CONDITIONS

Andrew A. Catellier and Stephen D. Voran[1]

We describe the design, implementation, and analysis of a speech intelligibility test. The test included five codec modes, four frame-erasure rates, and two background noise environments, for a total of 40 conditions. The test protocol required twenty listeners to repeat all words that they heard in short messages with median length of seven words. Each condition was tested using approximately 1100 words total. Listeners' responses were scored against the original message transcripts to produce a count of words correctly repeated and thus a measure of speech intelligibility. We present results that show exactly how this measure of speech intelligibility drops as frame-erasure rate increases for three of the five codec modes. The remaining two codec modes did not produce valid results due to defects in the reference software provided to us.

Keywords: acoustic noise, audio coding, background noise, frame erasure, packet loss, speech coding, speech intelligibility

1. BACKGROUND AND MOTIVATION

The National Public Safety Telecommunications Council (NPSTC) has defined seven high-level requirements for public safety Mission-Critical Voice (MCV) networks [1]. Six of these requirements relate to operating modes or system capabilities and one requirement lists four related yet separate dimensions of audio performance:

> "Audio Quality: This is a vital ingredient for mission critical voice. The listener MUST be able to understand without repetition, and can identify the speaker, can detect stress in a speaker's voice, and be able to hear background sounds as well without interfering with the prime voice communications."

The report [1] also prioritizes the four audio quality issues:

> "Audio Quality
> The transmitter and receiver audio quality must be such that, in order of importance:
> 1. The listener can understand what is being said without repetition.
> 2. The listener can identify the speaker (assuming familiarity with the speaker's voice).
> 3. The listener can detect stress in the speaker's voice, if present.
> 4. The background environment audio shall be sufficiently clear to the listener that sounds such as sirens and babies crying can be determined."

[1] The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

In early work the Public Safety Communications Research Program (PSCR) investigated items two and three on the list: speaker identification [2] and detection of speaker stress [3]. Our results suggest that if a system preserves speech intelligibility, then it will also preserve the ability to identify speakers and detect urgency in speakers' voices [4]. These results reaffirm that speech intelligibility (item one on the NPSTC list) is indeed the critical issue.

The primary focus of subsequent PSCR work has been speech intelligibility. More specifically, we have emphasized speech intelligibility in background noise [5]–[7], and this emphasis was motivated by field reports from our public safety partners. Most recently PSCR performed a detailed study of speech intelligibility for some of the digital speech and audio codecs that can be used to provide voice over Long-Term Evolution (LTE) based radio networks [8]. Acoustic noise environment and audio coding bandwidth were key variables in that work. The effort considered 83 codec modes at the outset, and ultimately provided comprehensive intelligibility results for 28 of those codec modes.

Speech intelligibility can also be influenced by degradations in a Radio Access Network (RAN). When RANs and the individual underlying radio links are stressed, the frames or packets of data that carry encoded voice can be corrupted and with sufficient corruption they will become useless. This is often described as "frame erasure" or "packet loss." The speech decoder consumes data frames and produces a short segment (typically 20 ms) of speech based on each frame. In the event of frame erasure, the speech decoder has no new data to process but must still produce speech output. This task is described as "loss concealment" and can include elements of extrapolation, time-scale modification, and the use of redundant information. Loss concealment algorithms range from rather simple to truly elaborate.

The success of these concealment algorithms depends not only on the algorithm, but also on the frequency and duration of the data losses or erasures. In general, it is easier to conceal shorter erasures and less frequent erasures. In addition, it is easier to conceal erasures that happen when the speech signal is changing slowly (e.g. a sustained vowel sound) and more difficult to conceal erasures that happen coincident with abrupt changes in speech.

In some cases concealment is completely successful, but in general it is only partially successful. As radio channel degradations become more severe, frame erasures become longer and more frequent. It becomes impossible to completely hide this and speech quality will drop. Speech intelligibility may be reduced as well. The relationship between erased frames and speech intelligibility reductions can be called the robustness of the speech codec. More robust codecs can tolerate higher levels of frame erasure while producing a given level of speech intelligibility. Viewed alternately, a more robust codec produces higher speech intelligibility at a given level of frame erasure.

Characterizing the relationship between the condition of the RAN and the intelligibility is particularly important for MCV because the events that stress the RAN may very well be events that also have critical intelligibility requirements. One public safety related example would be first responders moving deeper into a building to perform critical functions. Unless mitigation measures are in place, the radio link to outside parties will likely suffer additional attenuation and there may then be negative consequences for the speech intelligibility even as it is becoming more and more critical.

A second example would be an event that is escalating and requiring additional personnel to report to the scene. As more and more personnel share radio resources on the scene, those radio resources will (barring any mitigating measures) inevitably be spread thinner and thinner. Here again there may be negative consequences for speech intelligibility even as it becomes particularly important to coordinate the new personnel.

Both of these examples include the concept of mitigating actions. These actions could include the use of deployable infrastructure or reallocation of RAN resources. The goal is to provide the required radio resources when and where they are critically needed. These decisions should not be based on arbitrary engineering thresholds, but rather on the critical user experience factor: speech intelligibility. Our goal is to provide stakeholders with actionable numerical results to guide codec selection and to inform the design, provisioning, and adaptation of MCV services and the underlying RAN.

This report describes recent PSCR work in support of that goal. It addresses the robustness of several speech codecs. More specifically, we performed a speech intelligibility measurement experiment using five different speech codec modes in two different acoustic noise environments while forcing varying levels of frame erasure between the encoder and the decoder. The five codec modes were selected for their potential compatibility with LTE and the frame erasure statistics were selected to match those used in LTE standardization work. The report continues with the experiment design, experiment implementation, and finally the results obtained. Additional implementation details are provided in the appendices.

2. EXPERIMENT DESIGN

PSCR has previously used the Modified Rhyme Test (MRT) to provide a measure of speech intelligibility in loud noise conditions [9]. A trial in an MRT consists of a carrier sentence followed by a one-syllable test word. For example, “Please select the word ban.” That test word must be correctly identified in order for that trial to be considered a success. If one uses the MRT protocol to test imperfect radio channels, the effect of the radio channel would only have an impact on the success of a trial if there was a channel impairment at the same instant as the critical portion of the test word. To be most efficient with our testing resources and the limited time available, we pursued a different protocol. Put simply, listeners would listen to recorded public safety messages played back through a loudspeaker and then repeat the words that they understood into a microphone attached to a recording device. The resulting recordings were then scored for accuracy resulting in a number of words correctly identified. This very intuitive and direct definition of intelligibility has been used to measure the effectiveness of various speech enhancement (or noise reduction) algorithms; see [10] for one example.

The original recorded messages contained between one and 15 words. It was necessary to prevent learning in order to hold the amount of linguistic context constant. Test design did not present any multiword message more than one time. However, the single word messages were presented more than one time (e.g.: “thanks” or “affirmative”). Listeners were required to report what they heard after a single play—no repeats were allowed.

Twenty listeners participated and each listener heard all conditions but different combinations of source and condition. The result of this design is that around 1100 words are used to evaluate each condition.

2.1 Source Material

A corpus of public safety messages was transcribed from public safety scanner traffic. We made studio quality recordings of four males and two females reading these transcripts aloud in a sound isolated chamber (noise measured below 26.5 dBA) with absorbing materials on all surfaces. These recordings were stored in the `.wav` format at 16 bits per sample (b/smp) and 48,000 samples per second (smp/s). Approximately one-sixth of the 360 messages used came from each of the six talkers. Message length ranged from one word to 15 words with a median length of 7 and a mean length of 6.1 words. The 360 messages used in this experiment contain a total of 2204 words. Examples of messages used include: “Negative.”, “Receiving loud and clear.”, and “Five assorted vehicles in a no parking zone.”

2.2 Codec Modes

One key attribute of a speech or audio codec is the audio bandwidth that it encodes and reproduces. The present study includes audio codecs with two different nominal bandwidths. Wideband (WB) codecs support the range from approximately 50 Hz to 7 kHz and super-wideband (SWB) codecs have a nominal range from 50 Hz to 16 kHz. Compared to the original bandwidth used for telephony (300-3500 Hz), the WB and SWB options can improve the “realism” or “presence” of communications.

Five codec modes were selected for this experiment and are listed below. In light of the LTE context for MCV services, these codec modes include Adaptive Multi-Rate Wideband (AMR-WB) and Enhanced Voice Services (EVS) WB and SWB. The Adaptive Multi-Rate Wideband (AMR-WB) codec is specified in Third Generation Partnership Project (3GPP) Technical Specification (TS) 26.204. The software implementation used in this study is distributed as part of that technical specification and is available from 3gpp.org. The Enhanced Voice Services (EVS) codec was standardized by the 3GPP in September 2014. We used the latest available software (version 12.5.0) which was provided as part of TS 26.442 via 3gpp.org.

The five codec modes considered in this experiment are:

- EVS WB CA @ 13.2 kilobits per second (kb/s) [11]
- EVS SWB CA @ 13.2 kb/s [11]
- EVS WB @ 13.2 kb/s [11]
- AMR WB @ 12.65 kb/s [12]
- AMR WB @ 12.65 kb/s decoded by G.718 I/O [12], [13]

The selected bit rates of 12.65 and 13.20 kb/s have identical requirements in terms of LTE resource block allocation. Of particular interest are the channel aware (CA) modes of EVS. In the CA modes selectively applied redundant coding is used to enable significantly higher robustness to erased frames. This enhanced robustness has been already demonstrated in the speech quality domain for some frame-erasure environments [14].

In order to enable EVS CA mode, parameters for forward error correction (FEC) must be specified. In this test we used the `HI` indicator for parameter `FEC` and we used FEC offset 3 (the FEC offset is further discussed in Section 4.2.3). We are also interested in EVS WB at 13.2 b/s as a reference. Inclusion of this mode also allows for the most direct comparison to AMR WB at 12.65 kb/s (since CA mode will produce slightly higher end-to-end delay, compared to EVS when not utilizing CA mode). Additionally, we tested AMR WB at 12.65 kb/s decoded by G.718 I/O as this would give an indication of how AMR WB could perform with a more recently developed concealment algorithm.

Unfortunately after the experiment was completed we were notified of defects in the EVS reference software implementation that we used (Version 12.5.0). These defects prevented proper activation of the CA modes. Thus this report contains no valid results for those two modes.

2.3 Frame Erasure Rates

Four levels of frame erasure rate (FER) were selected for this experiment. To serve as a baseline, the first level was 0%. Consistent with previous work in 3GPP bodies, a two-state Markov model was used to generate packet loss patterns, and we selected non-zero FER conditions using Table A.1.2-3 of [14] “Markov parameters for 3 km/h.” Note that 3 km/h is “walking around” speed.

This means that channel conditions that depend on location will change more slowly than if the user were in a vehicle. Slowly changing channel conditions can produce multiple consecutive frame erasures, and this is an especially challenging environment for maintaining speech intelligibility. We chose FERs of 5%, 10%, and 20% for this experiment representing, mild, intermediate and extreme channel impairment scenarios.

The Markov models referenced above define each state to persist for 10 ms so we use every fourth state to simulate a downlink with 40 ms frames. Each of these 40 ms frames would contain two 20 ms EVS or AMR-WB frames, so the value of each 40 ms frame is repeated once. The end result is a loss pattern where each entry represents a 20 ms EVS or AMR-WB frame. The loss patterns were further processed such that any frame with no loss indicated was modified to indicate a minimal amount of jitter. The loss plus delay patterns were then used as input to programs that formatted the data in a way that was appropriate for each codec. Exact command line calls for all cases processing one file can be found in Appendix A. Further documentation for this process can be found in [15] which has block diagrams describing encoding and decoding processes in Clause 4 as well as command line syntax listings in Clause 5.

2.4 Background Noise

Speech intelligibility in noisy environments is very important to the public safety community. For this experiment we chose two background noise types relevant to the public safety community: coffee shop and U.S. police car siren. In order to maintain a practical experiment size, we selected one signal-to-noise ratio (SNR) for each type. After weighing numerous factors, we selected an SNR of 10 dB for the coffee shop noise and an SNR of 5 dB for the siren noise.

2.5 Processed Speech Assignments

With five codec modes, four FER levels, and two noise types, a total of 40 conditions are included in this speech intelligibility test. Each condition was tested using 180 messages. For maximum comparability, one group of 180 messages (containing a total of 1085 words) was used to test all coffee shop noise conditions and a second group of 180 messages (containing a total of 1119 words) was used to test all siren noise conditions. The 20 coffee shop noise conditions were tested in a single session and the 20 siren noise conditions were tested in a single session. Half of the listeners did the coffee shop session first, and half of the listeners did the siren session first.

Each session covered 20 conditions, contained 180 messages, and was heard by 20 listeners. In each session, each listener heard each of the 180 messages once and only once. Each condition was paired with each of the 180 messages once and only once. Once both sessions were completed, each listener had heard all of the 360 messages once and only once.

This balance was accomplished by breaking the list of 180 messages into 20 groups of 9 messages. These 20 groups were assigned to the 20 conditions for listener one (Group 1 → Condition 1, Group 2 → Condition 2, …, Group 20 → Condition 20). Then these assignments were shifted for listener two (Group 2 → Condition 1, Group 3 → Condition 2, …, Group 1 → Condition 20), and shifted again for listener three (Group 3 → Condition 1, Group 4 →

Condition 2, …, Group 2 → Condition 20). Continuing this process results in these assignments for listener 20: Group 20 → Condition 1, Group 1 → Condition 2, …, Group 19 → Condition 20.

In any session, the 20 conditions were presented in a random order under the following constraints. All WB conditions were grouped and all SWB conditions were grouped. Half the sessions had WB then SWB, the other half had SWB then WB. In addition each constrained random presentation order was balanced by using the reverse order in a different session. Thus the position of the conditions within the sessions was balanced.

3. EXPERIMENT IMPLEMENTATION

This section documents various diverse processes required as part of experiment implementation.

3.1 Speech Processing

3.1.1 Addition of Noise

The steps for combining speech with noise at the chosen SNR are as follows:

1) Calculate relative A-weighted SPL for both speech and noise signals.

2) Calculate initial SNR by subtracting noise SPL from speech SPL.

3) Calculate scale factor to multiply noise by in order to achieve desired SNR.

4) Add speech signal and scaled noise signal.

3.1.2 Speech Coding and Decoding

Specific processing steps are required to prepare raw speech source recordings for software encoding and decoding. In order to properly downsample, normalize, mix, code, and decode all of the source speech recordings we reused software previously developed for the work described in [8]. That software assumed a specified directory structure and then automatically processed all available source material. The steps are as follows:

1) The directory structure was parsed and all source files and their properties (sample rate, file type, source type—speech or noise) were calculated and stored.

2) All source files in `.wav` format were converted to standard raw PCM format.

3) All source files were downsampled to 16,000 smp/s using high quality G.191 filters and stored for use with WB codecs [16].

    a) All downsampled files were normalized to -28 dB using ITU Rec. P.56 [17].
    b) All downsampled files were modified to account for filter delay.

4) All source files were downsampled to 32,000 smp/s using high quality G.191 filters and stored for use with SWB codecs.

    a) All downsampled files were normalized to -28 dB using ITU Rec. P.56.

5) Speech and noise were mixed at the specified levels using the noise insertion process specified in 3.1.1. Speech and noise were mixed separately at each sample rate.

6) The proper resulting source files were then coded and decoded by each selected codec mode.

a) The process for coding and decoding for each codec mode included looping through each channel impairment level.

7) The resulting processed audio was then upsampled to 48,000 smp/s.

8) The resulting upsampled audio was then converted from standard raw PCM format to `.wav` format.

Appendix A provides the specific calls for the various codec modes as well as bitstream processing functions.

3.2 Experiment Logistics

We conducted the test sessions in December 2015. Twenty listeners (twelve male and eight female) participated in this test. The listeners were recruited from PSCR personnel and the average estimated listener age was in the mid-thirties. Listeners had no prior knowledge of the contents or nature of the experiment.

Listeners participated one at a time in a quiet sound isolation chamber. The coded speech plus noise signals were played back through a high-quality loudspeaker (described below) and listeners were allowed to adjust the volume of the loudspeaker to a comfortable level. A soft beep sound marked the start and end of each recording. Listeners were asked to repeat the message as they heard it. The displayed prompt read: “After the second beep, please exactly repeat each word that you understood between the two beeps.” A studio quality microphone captured the repetition which was then digitized and stored for later analysis. The experiment was double-blind. The only information provide to listeners was their progress through the session. This was done via a prompt that read, for example, “This is trial 175 of 180 total trials.”

Each listener started with a practice session that included six trials. This session allowed listeners to familiarize themselves with the experiment protocol and also allowed for verification of the operation of all equipment involved. After the practice session any procedural questions or issues were resolved but any questions regarding the contents of the experiment were deferred until after the completion of the experiment. After the practice session, each listener performed one session (180 trials) with siren noise and one session (180 trials) with coffee shop noise. Ten of the listeners did the siren noise session first, and the other ten did the coffee shop noise session first. On average, each session of 180 trials required approximately 20 minutes to complete. The total time spent on the experiment by each listener was typically just under one hour. Once both sessions were completed, each listener had heard all of the 360 messages once and only once (and had also heard six additional messages in the practice session.)

3.3 Listening Lab

We conducted the experiment in two matched sound-isolated rooms with inside dimensions 305 cm long, 274 cm wide and 213 cm high (approximately 10 by 9 by 7 feet). In each room the floor is carpeted and all of the walls and the ceiling are covered with sound absorbing materials. Under normal conditions as would be experienced in the experiment, the noise level inside either

room is below 26.5 dBA measured with a Brüel and Kjær Type 2250 sound level meter. When the air conditioning for a room is turned off, that level drops below 19.5 dBA for each room. These are extremely low noise levels and these measurements demonstrate that background noise is well-controlled in these labs.

Both rooms were configured so that the listener sits on a chair in the center of the room behind a 76 cm by 152 cm (2.5 by 5 foot) work table. This table supports a loudspeaker, an LCD monitor screen, a mouse, and a keyboard.

The recording format is digital files with 48,000 smp/s and 16 b/smp. The playback path includes a digital audio interface (USB to AES/EBU) so that the AES/EBU digital audio format is provided to the digital input of a Fostex Model 6301D Digital Personal Monitor loudspeaker. Listeners were encouraged to adjust the volume knob on the front of this loudspeaker to achieve preferred listening level.

We used pink noise playback to characterize the combined frequency response of the playback electronics, the loudspeaker, and the room. Our spectral analysis was performed at the listener's head location using octave-wide analysis bands (see ANSI S1.11). The composite response in the octaves centered at 125, 250, 500, 1000, 2000, 4000, 8000, and 16,000 Hz deviate no more than ±5 dB relative to the response in the octave centered at 1000 Hz.

3.4 Scoring

After listeners had completed their tasks an engineer listened to every recording produced by every listener and marked each word of each corresponding transcript as either correct or incorrect. This task was facilitated by a software tool that we wrote specifically for this task. This tool automated blind playing of the recordings, replaying of recordings as required, presentation of the transcripts, and collection of the results. The replaying option was frequently used so that the engineer could discriminate the smaller differences between listeners' recordings and original transcripts. A set of rules that concretely define the word scoring process are available in Appendix B.

In this step the engineer was effectively measuring the number of words successfully received in each condition, using a total of about 1100 words and 20 listeners. The corresponding success rate forms a relative measure of speech intelligibility for the various conditions. Table 1 provides the results. We acknowledge that this success rate may also include some additional human factors on the part of the listeners and the scoring engineer. But we emphasize again the balance achieved in the test design—all 20 listeners and all 180 messages contribute equally to the results for each condition. The goal of this design is that factors other than condition will average out and condition will dominate the results. Indeed the results presented here show a strong and clear dependence on condition.

Table 1. Number of words correctly received and success rates for all 40 conditions. Final 16 conditions are not valid (per discussion in Section 4.2.1) but are included for completeness.

		Coffee Shop Noise (1085 words total)		**Siren Noise (1119 Words Total)**	
Codec Mode	FER (%)	Words Correct	Success Rate	Words Correct	Success Rate
AMR-WB @ 12.65 kb/s	0	1008	0.929	1064	0.951
	5	976	0.900	1033	0.923
	10	917	0.845	947	0.846
	20	774	0.713	833	0.744
AMR-WB @ 12.65 kb/s decoded by G.718 I/O	0	1010	0.931	1068	0.954
	5	994	0.916	1052	0.940
	10	935	0.862	978	0.874
	20	785	0.723	873	0.780
EVS WB @ 13.2 kb/s	0	1045	0.963	1057	0.945
	5	986	0.909	1023	0.914
	10	937	0.864	939	0.839
	20	813	0.749	804	0.719
Intended to be EVS WB CA @ 13.2 kb/s but results were corrupted by defect in reference codec software.	0	1017	0.937	1052	0.940
	5	968	0.892	1010	0.903
	10	932	0.859	950	0.849
	20	779	0.718	792	0.708
Intended to be EVS SWB CA @ 13.2 kb/s but results were corrupted by defect in reference codec software.	0	1021	0.941	1056	0.944
	5	975	0.899	1008	0.901
	10	936	0.863	951	0.850
	20	787	0.725	869	0.777

In order to characterize variation due to the scoring engineer, we asked a second engineer to score all recordings for the coffee noise. The results agreed closely with those of the first engineer. Across the twenty coffee noise conditions, word success rates calculated for the two independent scorings correlate at the level of 0.998. The differences in success rates for the two scorings are also of interest. Across the twenty conditions with coffee shop noise, the mean difference is 0.001, indicating nearly zero consistent bias between the two scoring engineers. Indeed, seven of these differences are negative and thirteen are positive. The average magnitude of these differences is 0.005, and we conclude that variation due to the scoring engineer is minimal.

4. EXPERIMENT RESULTS

4.1 Speech Intelligibility Results

The speech intelligibility results presented here are limited to three of the five codec modes tested. This is because the operation of both the EVS WB CA codec mode and the EVS SWB CA codec mode was corrupted by defects in the reference software provided to us, and this corruption was only revealed after the experiment was completed. More information is provided in Section 4.2.

The valid speech intelligibility results obtained from the experiment are shown in Figure 1. Each figure shows the speech intelligibility computed as a word success rate (words correctly repeated divided by total words) for three valid codec modes at the four FER values included in the experiment. Having defined speech intelligibility in this manner for this experiment, we use the terms “speech intelligibility” and “success rate” as synonyms in this report. Figure 1 (a) shows results for coffee shop noise and Figure 1 (b) addresses siren noise.

Each figure clearly shows the expected trend of decreasing intelligibility as FER increases. Each noise type sets an upper limit on intelligibility as FER goes to zero. In the case of coffee shop noise, this upper limit depends on codec mode as well. In broadest terms (across both noises and all codec modes), we observe that intelligibility does not drop below 0.7 at the worst FER (20%) used in this experiment. This result, as well as those at 5 and 10% FER, may be advantageously used to inform the provisioning of radio resources and other network resources from the very relevant perspective of speech intelligibility.

Each condition was evaluated using 180 messages for a total of 1085 (coffee shop noise) or 1119 (siren noise) words. In either case this is around 1100 Bernoulli Trials per condition. A chi-squared test is an appropriate way to test for significant differences among Bernoulli Trials. Due to the nature of Bernoulli Trials, the interval that defines a statistically significant difference varies with success rate. With this number of trials, and using a significance level of 95%, a success rate of 0.93 is significantly greater than a success rate of 0.90 but a success rate of 0.92 is not significantly greater than a success rate of 0.90. The difference required for significance increases slightly towards the middle of the scale: a success rate of 0.74 is significantly greater than a success rate of 0.70 but a success rate of 0.73 is not significantly greater than a success rate of 0.70. These values are intended to provide a general sense of the magnitude of difference in success rates required for significance in this experiment. In the following, we calculate and report significance using the exact number of trials for each case following the chi-squared test for independence of categorical data, [18]–[20]. We apply the test in the same manner as detailed in [8].

In the case of coffee shop noise at 10 dB SNR the only significant difference (95% level) between codec modes at any given FER occurs at FER = 0%. Here EVS WB shows a significantly higher success rate than the other two codec modes. At all other FER values, we found no statistically significant differences between the three codec modes.

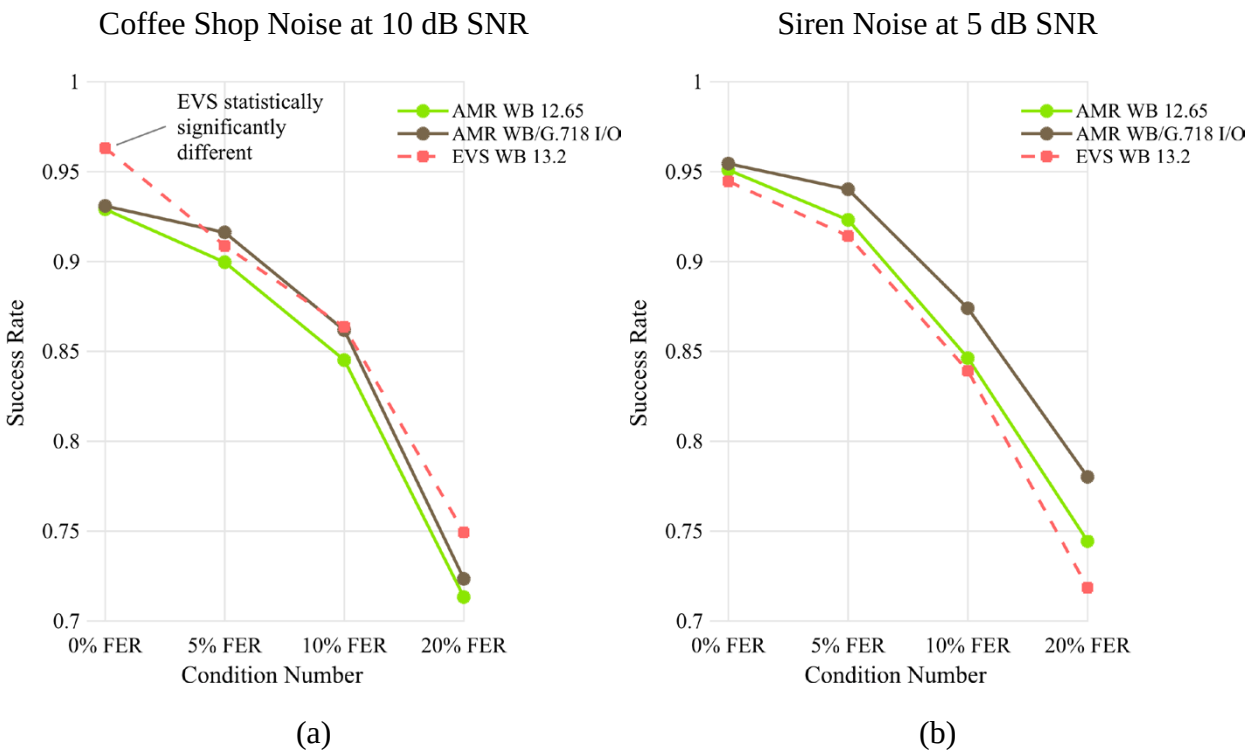

Figure 1. Word success rates for three codec modes in (a) coffee shop noise at 10 dB SNR and (b) siren noise at 5 dB SNR.

For the case of siren noise at 5 dB SNR the opposite is true. Here there are no significant differences between codec modes at FER = 0%, but consistent significant differences for the non-zero FER values. Specifically, AMR WB decoded with G.718 I/O has a higher success rate than EVS WB at FER = 5, 10, and 20%. In addition, AMR WB decoded with G.718 I/O has a higher success rate than conventional AMR WB in the 20% FER case.

4.2 Other Results and Observations

4.2.1 EVS CA Software Problem

After our experiment was completed we were notified of defects in the EVS reference software implementation that we used during our experiment (Version 12.5.0). The software error caused the jitter buffer manager to lock up in the case of low jitter and high FER. Thus the CA mode of EVS did not properly activate. We completed speech intelligibility test sessions in December 2015 and we were notified of this defect on January 11, 2016. Newer versions of the EVS reference software have been released since that date. It is an important goal for our future work to assess the extent to which the EVS CA modes can preserve speech intelligibility as FER increases.

4.2.2 Loss Patterns not Identical

When conducting experiments of this sort, a common goal is to minimize all extraneous sources of variation. After the experiment we discovered a minor short-coming in this regard. While it is of no practical consequence, we detail the shortcoming here for the sake of completeness.

We sought to use identical pseudo-random frame erasure patterns for each of the codec modes under test, yet also allow different random patterns for each message in the test. This approach was to be implemented by strategically setting the seed in the pseudo random pattern generator. The seed was to be unique based on the combination of source file, the noise type/SNR and channel impairment level. To achieve this, a hash function was used to generate a numerical input to seed the random number generator. The input to this hash function was the path (which contained information about noise type/SNR) and filename of the source file combined with a number representing the channel impairment level.

A post-experiment audit revealed that the method used to compensate for delay when using the AMR WB codec caused an error in this process. The delay compensation method created a temporary version of the source file with a slightly different name, and this file was used as the input to the encoder and as the input to the hashing function. This is evident in Appendix A.

Therefore, the loss patterns generated for the AMR WB and AMR WB/G.718 codecs differ from those generated for the EVS codecs. Additionally, the delay compensation for AMR WB and AMR WB/G.718 differed by 16 samples (1 ms). These factors could conceivably add a minimal amount of additional variation between the conditions under consideration, but it is impossible for such factors to systematically favor one codec over another.

We performed a post-hoc analysis on the actual loss patterns seen by the codecs. The results listed in Tables 2 through 5 show that the median loss lengths for both loss pattern types were identical and that the mean loss rate, the mean loss length, and the standard deviation of loss length for both loss pattern types had only insignificantly small differences.

Table 2. Measured mean loss rates for AMR and EVS loss patterns.

Average Loss Rate	**5% FER**	**10% FER**	**20% FER**
AMR	4.71%	10.09%	20.34%
EVS	4.71%	10.18%	20.40%

Table 3. Measured mean loss length in frames (20 ms) for AMR and EVS loss patterns.

Average Loss Rate	**5% FER**	**10% FER**	**20% FER**
AMR	2.6883	2.9718	3.6160
EVS	2.7052	2.9736	3.5938

Table 4. Measured standard deviation of loss length in frames (20 ms) for AMR and EVS loss patterns.

Average Loss Rate	**5% FER**	**10% FER**	**20% FER**
AMR	1.38	1.75	2.45
EVS	1.46	1.71	2.46

Table 5. Measured median loss lengths in frames (20 ms) for AMR and EVS loss patterns.

Average Loss Rate	**5% FER**	**10% FER**	**20% FER**
AMR	2	2	2
EVS	2	2	2

4.2.3 End-to-End Delay Considerations

An additional consideration in MCV applications is end-to-end delay. Several delay contributions must be added together to arrive at end-to-end delay. One of these contributions is the delay induced by buffering frames of encoded speech data at the decoder, often called jitter buffering. Motivations for this buffering include mitigation of delay jitter and robustness against lost data. The only delay jitter introduced in the present experiment was the 20 ms scheduling delay variation mandated when two 20 ms speech data frames must be carried in a single 40 ms downlink frame.

The AMR WB and G.718 decoders do not use multiple frames of encoded speech data to achieve robustness. Thus the buffering required in front of these decoders in this test is one frame (20 ms). More precisely, when a downlink frame containing speech frames N and N+1 is received, speech frame N can be decoded immediately and speech frame N+1 must be buffered while frame N is played. When frame N has finished playing, frame N+1 can be decoded and played. If any frame of speech data is declared lost, the AMR WB or G.718 decoder will activate a concealment algorithm that produces a playable output signal without waiting for any additional frames of speech data.

The EVS decoder can operate in exactly the same fashion unless CA mode is used. In order for the CA mode to be effective, additional frames of speech data must be available at the decoder and this increased buffering requirement will increase the end-to-end delay.

The EVS CA mode gains robustness through the use of multiple encodings spread across multiple speech frames. In order to take advantage of these multiple encodings, they must be available at the decoder. In this experiment we set the EVS CA FEC offset option to 3 (invoked on the command line by specifying `-rf 3`) as instructed by EVS experts. This means that when the decoder needs to conceal losses using CA mode, the decoder must have access to 3 frames in addition to the one that it is currently decoding for playout. When this requirement is combined with the requirement that two 20 ms speech data frames be carried in a single 40 ms downlink frame, the equivalent buffering requirement for EVS CA operation in this test is four frames of speech data (80 ms). For example, consider three consecutive downlink frames that contain speech frames (N, N+1), (N+2, N+3), and (N+4, N+5). If speech frame N+1 is lost, then CA

decoding requires speech frame *N*+4 in order to produce the speech output associated with frame *N*+1. But speech frame *N*+4 and *N*+5 arrive at the same time, so a total of four frames (80 ms) must be available in order to play out frame *N*+1. This is an increase of three frames (60 ms) over the requirement for non-CA mode.

Note that the value of `rf` defines a trade-off between decoding delay and robustness. Non-CA mode defines a baseline decoding delay and a baseline level of robustness. The example `-rf 3` used above gives an increased level of robustness and also decoding delay. This is the value of `rf` used in most or all of the SA4 EVS CA testing to date. Intermediate values of `rf` may give intermediate levels of robustness and decoding delay.

5. CONCLUSIONS

Our first conclusion is that the test protocol we adopted, customized, and implemented does provide a sensitive measurement of speech intelligibility as a function of FER. As expected, speech intelligibility drops as FER is increased. Most generally (across two noise environments and three codec modes) our intelligibility measure drops from the range 0.93–0.96 at 0% FER to the range 0.71–0.78 at 20% FER.

The CA modes of EVS use selectively applied redundant coding to enable higher speech quality in frame erasure environments. Our present test was unable to produce parallel results in the speech intelligibility domain because the reference software implementation provided to us was defective. Thus our conclusions with respect to the EVS CA modes are necessarily limited to the observation that additional testing with stable updated software is an important future goal.

While we cannot compare the expected enhanced robustness of EVS CA with the other three codec modes, we can report several significant differences measured among those three codec modes. For coffee shop noise (10 dB SNR) the only significant difference between codec modes is at 0% FER, where EVS WB shows higher intelligibility than the two AMR modes. With siren noise (5 dB SNR), AMR WB decoded with G.718 I/O has higher intelligibility than EVS WB at FER = 5, 10, and 20%. And for 20% FER, AMR WB decoded with G.718 I/O also has higher intelligibility than conventional AMR WB.

We expect that these test results can inform MCV codec selection as well as the design, provisioning, and adaptation of MCV services and the underlying RAN. Most importantly, these results can allow those engineering activities to be driven by the critical user experience factor—speech intelligibility.

6. REFERENCES

[1] National Public Safety Telecommunications Council (NPSTC), "Mission critical voice communications requirements for public safety," Littleton, CO, 2011.

[2] A. Catellier and S. Voran, "Speaker identification in low-rate coded speech," in *Proc. 7th International Measurement of Audio and Video Quality in Networks Conference*, Prague, 2008. Available http://www.its.bldrdoc.gov/publications/2626.aspx.

[3] S. Voran, "Listener detection of talker stress in low-rate coded speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, 2008. Available http://www.its.bldrdoc.gov/publications/2627.aspx.

[4] A. Catellier and S. Voran, "Relationships between intelligibility, speaker identification, and the detection of dramatized urgency," NTIA Report 09-459, Boulder CO, 2008. Available http://www.its.bldrdoc.gov/publications/2496.aspx.

[5] D. Atkinson and A. Catellier, "Intelligibiliity of selected radio systems in the presence of fireground noise: Test plan and results," NTIA Report 08-453, Boulder CO, 2008. Available http://www.its.bldrdoc.gov/publications/2490.aspx.

[6] D. Atkinson and A. Catellier, "Intelligibility of analog FM and updated P25 radio systems in the presence of fireground noise: Test plan and results," NTIA Report 13-495, Boulder CO, 2013. Available http://www.its.bldrdoc.gov/publications/2720.aspx.

[7] D. Atkinson, S. Voran and A. Catellier, "Intelligibility of the adaptive multi-rate speech coder in emergency-response environments," NTIA Report 13-493, Boulder CO, 2013. Available http://www.its.bldrdoc.gov/publications/2693.aspx.

[8] S. Voran and A. Catellier, "Speech Codec Intelligibility Testing in Support of Mission-Critical Voice Applications for LTE," NTIA Report 15-520, Boulder CO, 2015. Available http://www.its.bldrdoc.gov/publications/2811.aspx.

[9] ANSI, "ANSI S3.2 American national standard method for measuring the intelligibility of speech over communication systems," New York, 1989.

[10] P. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47-56, 2011.

[11] 3GPP, *TS 26.442: Codec for Enhanced Voice Services (EVS); ANSI C code (fixed-point)*, ETSI.

[12] 3GPP, *TS 26.204: Speech codec speech processing functions; Adaptive Multi-Rate – Wideband (AMR-WB) speech codec; ANSI-C code*, ETSI.

[13] ITU-T, *G.718: Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s*, Geneva: ITU-T, 2009.

[14] 3GPP, *TR 26.989: Technical Specification Group Services and System Aspects; Mission Critical Push To Talk (MCPTT); Media, codecs and Multimedia Broadcast/Multicast Service (MBMS) enhancements for MCPTT over LTE (Release 13),* 3GPP.

[15] 3GPP, *Tdoc S4-141392 EVS Permanent Document EVS-7c: Processing functions for characterization phase,* Santa Cruz de Tenerife, Spain: ETSI.

[16] ITU-T, "Rec. P.191: Software tools for speech and audio coding standardization," Geneva, 2012.

[17] ITU-T, "Rec. P.56: Objective measurement of active speech level," Geneva, 2011.

[18] E. L. Crow, F. A. Davis and M. W. Maxfield, Statistics Manual, New York: Dover, 1960.

[19] A. M. Mood, F. A. Graybill and D. C. Boes, Introduction to the Theory of Statistics, New York: McGraw-Hill, 1974.

[20] R. V. Hogg and E. A. Tanis, Probability and Statistical Inference, New York: Macmillian, 1977.

ACKNOWLEDGEMENTS

Funding for this work was provided by the DHS Science and Technology Directorate, Cuong Luu, Program Manager. The work was conducted by the PSCR, Andrew Thiessen and Dereck Orr, Program Managers. This work builds on the work of the late DJ Atkinson. DJ established the speech intelligibility work in the PSCR and the present work would not have been possible without DJ's vision, leadership, and hard work. We are deeply indebted to DJ and we seek to honor his memory through this present work.

We offer our since thanks to Ellen Ryan for scheduling and recruiting listeners from across the PSCR, to those many PSCR employees who took time to participate as listeners in this experiment, and to Adam Danielson for tirelessly scoring thousands of recordings. This experiment would not have been possible without this generous support. Finally, we recognize the technical reviewers who provided input to this report and we offer our thanks to the ITS Publications Officer Lilli Segre for her overall support and her thorough editorial revisions that have produced this final product.

APPENDIX A. EXAMPLE PROCESSING COMMANDS

1. EVSCA-WB, no losses

Encode string:

['../../exes/win_EVSencoder.exe', '-max_band', 'WB', '-dtx', '-rf', 'HI', '3', '13200', '16',
'../../audioOneFile/uncoded/raw16k/speech/CF_medium_150.raw',
'../../bitstreamFilesNoNorm/EVSCA_132_WB_None/raw16k/speech/CF_medium_150.bit']

Decode string:

['../../exes/win_EVSdecoder.exe', '16',
'../../bitstreamFilesNoNorm/EVSCA_132_WB_None/raw16k/speech/CF_medium_150.bit',
'../../audioOneFile/EVSCA_132_WB/raw16k/speech/CF_medium_150.raw']

2. EVSCA-WB, with losses

Encode String:

['../../exes/win_EVSencoder.exe', '-max_band', 'WB', '-dtx', '-rf', 'HI', '3', '13200', '16',
'../../audioOneFile/uncoded/raw16k/speech/CF_medium_150.raw',
'../../bitstreamFilesNoNorm/EVSCA_132_WB_2/raw16k/speech/CF_medium_150.bit']

Decode process:

count number of frames in bitstream files

multiply by two to get correct number of 10ms frames

generate loss pattern using markov chain

get every fourth value in markov chain output

repeat each frame once to get back to 20ms

if loss indicated, do nothing, else, change to either 120 or 100

write resulting loss pattern to disk

Network Simulator String:

['../../exes/networkSimulator_g192.exe',
'../../bitstreamFilesNoNorm/EVSCA_132_WB_2/raw16k/speech/CF_medium_150.bit.dat',
'../../bitstreamFilesNoNorm/EVSCA_132_WB_2/raw16k/speech/CF_medium_150.bit',
'../../bitstreamFilesNoNorm/EVSCA_132_WB_2/raw16k/speech/CF_medium_150.bit.rtp',
'tracefile', '1', '0']

Decode string:

['../../exes/win_EVSdecoder.exe', '-VOIP', '-Tracefile', 'evsTrace.tmp', '16',
'../../bitstreamFilesNoNorm/EVSCA_132_WB_2/raw16k/speech/CF_medium_150.bit.rtp',
'../../audioOneFile/EVSCA_132_WB_2/raw16k/speech/CF_medium_150.raw']

3. EVSCA-SWB, no losses

Encode String:

['../../exes/win_EVSencoder.exe', '-max_band', 'SWB', '-dtx', '-rf', 'HI', '3', '13200', '32',
'../../audioOneFile/uncoded/raw32k/speech/CF_medium_150.raw',
'../../bitstreamFilesNoNorm/EVSCA_132_SWB_None/raw32k/speech/CF_medium_150.bit']

Decode String:

['../../exes/win_EVSdecoder.exe', '32',
'../../bitstreamFilesNoNorm/EVSCA_132_SWB_None/raw32k/speech/CF_medium_150.bit',
'../../audioOneFile/EVSCA_132_SWB/raw32k/speech/CF_medium_150.raw']

4. EVSCA-SWB, with losses

Encode String:

['../../exes/win_EVSencoder.exe', '-max_band', 'SWB', '-dtx', '-rf', 'HI', '3', '13200', '32',
'../../audioOneFile/uncoded/raw32k/speech/CF_medium_150.raw',
'../../bitstreamFilesNoNorm/EVSCA_132_SWB_2/raw32k/speech/CF_medium_150.bit']

Decode process:

count number of frames in bitstream files

multiply by two to get correct number of 10ms frames

generate loss pattern using markov chain

get every fourth value in markov chain output

repeat each frame once to get back to 20ms

if loss indicated, do nothing, else, change to either 120 or 100

write resulting loss pattern to disk

Network Simulator String:

['../../exes/networkSimulator_g192.exe',
'../../bitstreamFilesNoNorm/EVSCA_132_SWB_2/raw32k/speech/CF_medium_150.bit.dat',
'../../bitstreamFilesNoNorm/EVSCA_132_SWB_2/raw32k/speech/CF_medium_150.bit',

'../../bitstreamFilesNoNorm/EVSCA_132_SWB_2/raw32k/speech/CF_medium_150.bit.rtp', 'tracefile', '1', '0']

Decode String:

['../../exes/win_EVSdecoder.exe', '-VOIP', '-Tracefile', 'evsTrace.tmp', '32', '../../bitstreamFilesNoNorm/EVSCA_132_SWB_2/raw32k/speech/CF_medium_150.bit.rtp', '../../audioOneFile/EVSCA_132_SWB_2/raw32k/speech/CF_medium_150.raw']

5. EVS-WB, no losses:

Encode String:

['../../exes/win_EVSencoder.exe', '-max_band', 'WB', '-dtx', '13200', '16', '../../audioOneFile/uncoded/raw16k/speech/CF_medium_150.raw', '../../bitstreamFilesNoNorm/EVS_132_WB_None/raw16k/speech/CF_medium_150.bit']

Decode String:

['../../exes/win_EVSdecoder.exe', '16', '../../bitstreamFilesNoNorm/EVS_132_WB_None/raw16k/speech/CF_medium_150.bit', '../../audioOneFile/EVS_132_WB/raw16k/speech/CF_medium_150.raw']

6. EVS-WB, with losses:

Encode String:

['../../exes/win_EVSencoder.exe', '-max_band', 'WB', '-dtx', '13200', '16', '../../audioOneFile/uncoded/raw16k/speech/CF_medium_150.raw', '../../bitstreamFilesNoNorm/EVS_132_WB_2/raw16k/speech/CF_medium_150.bit']

[decode process]

Network Simulator String

['../../exes/networkSimulator_g192.exe', '../../bitstreamFilesNoNorm/EVS_132_WB_2/raw16k/speech/CF_medium_150.bit.dat', '../../bitstreamFilesNoNorm/EVS_132_WB_2/raw16k/speech/CF_medium_150.bit', '../../bitstreamFilesNoNorm/EVS_132_WB_2/raw16k/speech/CF_medium_150.bit.rtp', 'tracefile', '1', '0']

Decode String:

['../../exes/win_EVSdecoder.exe', '-VOIP', '-Tracefile', 'evsTrace.tmp', '16',
'../../bitstreamFilesNoNorm/EVS_132_WB_2/raw16k/speech/CF_medium_150.bit.rtp',
'../../audioOneFile/EVS_132_WB_2/raw16k/speech/CF_medium_150.raw']

7. AMR WB, no losses:

correct for AMR delay

Encode String:

['../../exes/win_AMRWBencoder.exe', '-dtx', '-itu', '2',
'../../audioOneFile/uncoded/raw16k/speech/CF_medium_150.raw.temp',
'../../bitstreamFilesNoNorm/amr_126_WB_None/raw16k/speech/CF_medium_150.bit']

Decode String:

['../../exes/win_AMRWBdecoder.exe', '-itu',
'../../bitstreamFilesNoNorm/amr_126_WB_None/raw16k/speech/CF_medium_150.bit',
'../../audioOneFile/amr_126_WB/raw16k/speech/CF_medium_150.raw']

8. AMR WB, with losses:

correct for AMR delay

Encode String:

['../../exes/win_AMRWBencoder.exe', '-dtx', '-itu', '2',
'../../audioOneFile/uncoded/raw16k/speech/CF_medium_150.raw.temp',
'../../bitstreamFilesNoNorm/amr_126_WB_2/raw16k/speech/CF_medium_150.bit']

[decode process]

Delay to Error Pattern string:

['../../exes/dlyerr_2_errpat.exe', '-L', '22000', '-d', '200', '-f', '1', '-w', '-s', '0', '-i',
'../../bitstreamFilesNoNorm/amr_126_WB_2/raw16k/speech/CF_medium_150.bit.dat', '-o',
'../../bitstreamFilesNoNorm/amr_126_WB_2/raw16k/speech/CF_medium_150.bit.dat.errorpat']

Error Insertion Device String:

['../../exes/eid-xor.exe', '-vbr', '-fer',
'../../bitstreamFilesNoNorm/amr_126_WB_2/raw16k/speech/CF_medium_150.bit',
'../../bitstreamFilesNoNorm/amr_126_WB_2/raw16k/speech/CF_medium_150.bit.dat.errorpat',
'../../bitstreamFilesNoNorm/amr_126_WB_2/raw16k/speech/CF_medium_150.bit.rtp']

Decode String:

['../../exes/win_AMRWBdecoder.exe', '-itu',
'../../bitstreamFilesNoNorm/amr_126_WB_2/raw16k/speech/CF_medium_150.bit.rtp',
'../../audioOneFile/amr_126_WB_2/raw16k/speech/CF_medium_150.raw']

9. AMR WB/G.718 IO, no losses:

[correct for AMR delay]

Encode String:

['../../exes/win_AMRWBencoder.exe', '-dtx', '-itu', '2',
'../../audioOneFile/uncoded/raw16k/speech/CF_medium_150.raw.temp',
'../../bitstreamFilesNoNorm/amrG718_126_WB_None/raw16k/speech/CF_medium_150.bit']

Decode String:

['../../exes/win_G718decoder.exe', '-IO_G722_2', '16',
'../../bitstreamFilesNoNorm/amrG718_126_WB_None/raw16k/speech/CF_medium_150.bit',
'../../audioOneFile/amrG718_126_WB/raw16k/speech/CF_medium_150.raw']

10. AMR WB/G.718 IO, with losses:

[correct for AMR delay]

Encode String:

['../../exes/win_AMRWBencoder.exe', '-dtx', '-itu', '2',
'../../audioOneFile/uncoded/raw16k/speech/CF_medium_150.raw.temp',
'../../bitstreamFilesNoNorm/amrG718_126_WB_2/raw16k/speech/CF_medium_150.bit']

[decode process]

Delay to Error Pattern string:

['../../exes/dlyerr_2_errpat.exe', '-L', '22000', '-d', '200', '-f', '1', '-w', '-s', '0', '-i',
'../../bitstreamFilesNoNorm/amrG718_126_WB_2/raw16k/speech/CF_medium_150.bit.dat', '-o',
'../../bitstreamFilesNoNorm/amrG718_126_WB_2/raw16k/speech/CF_medium_150.bit.dat.error
pat']

Error Insertion Device String:

['../../exes/eid-xor.exe', '-vbr', '-fer',
'../../bitstreamFilesNoNorm/amrG718_126_WB_2/raw16k/speech/CF_medium_150.bit',
'../../bitstreamFilesNoNorm/amrG718_126_WB_2/raw16k/speech/CF_medium_150.bit.dat.error
pat', '../../bitstreamFilesNoNorm/amrG718_126_WB_2/raw16k/speech/CF_medium_150.bit.rtp']

Decode String:

['../../exes/win_G718decoder.exe', '-IO_G722_2', '16',
'../../bitstreamFilesNoNorm/amrG718_126_WB_2/raw16k/speech/CF_medium_150.bit.rtp',
'../../audioOneFile/amrG718_126_WB_2/raw16k/speech/CF_medium_150.raw']

APPENDIX B. INSTRUCTIONS AND EXAMPLES FOR SCORING LISTENER RECORDINGS

Recall that the instruction given to the listener is: “After the second beep, please exactly repeat each word that you understood between the two beeps.”

- Score is based on exact repetition of words, not meanings.
- Order of words spoken by listener does not matter.
- Extra words spoken by listener do not hurt.

Due to similarity and lack of consistent pronunciation by readers:

- “en route” and “in route” are considered to be equivalent in this experiment.
- “complainant” and “complaintant” are considered to be equivalent in this experiment.
- “til” and “until” are considered to be equivalent in this experiment.

The transcript sometimes breaks hyphenated or compound words into single words if they can reasonably stand alone in the context. For example, twenty-seven becomes twenty seven because either word could be independently errored (e.g. to either thirty-seven, or twenty-two.) By this reasoning then, the hyphenated word twenty-seven, counts as two words for the purpose of this experiment.

Contractions are equivalent to full words. Here are some hypothetical examples:

Transcript: We will get cats.	Listener: We will get cats.	Score is 4 out of 4.
Transcript: We will get cats.	Listener: We’ll get cats.	Score is 4 out of 4.
Transcript: We will get cats.	Listener: We get cats.	Score is 3 out of 4.
Transcript: We will get cats.	Listener: Will get cats.	Score is 3 out of 4.
Transcript: We’ll get cats.	Listener: We’ll get cats.	Score is 3 out of 3.
Transcript: We’ll get cats.	Listener: We will get cats.	Score is 3 out of 3.
Transcript: We’ll get cats.	Listener: We get cats.	Score is 2 out of 3.
Transcript: We’ll get cats.	Listener: Will get cats.	Score is 2 out of 3.

NTIA FORM 29
(4-80)

U.S. DEPARTMENT OF COMMERCE
NATIONAL TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO.	2. Government Accession No.	3. Recipient's Accession No.
TR-17-522		

4. TITLE AND SUBTITLE	5. Publication Date
Intelligibility of Selected Speech Codecs in Frame-Erasure Conditions	November 2016
	6. Performing Organization Code NTIA/ITS.P

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	10. Contract/Grant Number. 6776000-300
11. Sponsoring Organization Name and Address Office of Interoperability and Compatibility Science and Technology Directorate U.S. Department of Homeland Security 245 Murray Lane SW Washington, DC 20528	12. Type of Report and Period Covered

14. SUPPLEMENTARY NOTES

15. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)

We describe the design, implementation, and analysis of a speech intelligibility test. The test included five codec modes, four frame-erasure rates, and two background noise environments, for a total of 40 conditions. The test protocol required twenty listeners to repeat all words that they heard in short messages with median length of seven words. Each condition was tested using approximately 1100 words total. Listeners' responses were scored against the original message transcripts to produce a count of words correctly repeated and thus a measure of speech intelligibility. We present results that show exactly how this measure of speech intelligibility drops as frame-erasure rate increases for three of the five codec modes. The remaining two codec modes did not produce valid results due to defects in the reference software provided to us.

16. Key Words (Alphabetical order, separated by semicolons)

acoustic noise, audio coding, background noise, frame erasure, packet loss, speech coding, speech intelligibility

17. AVAILABILITY STATEMENT	18. Security Class. (This report)	20. Number of pages
☒ UNLIMITED.	Unclassified	47
☐ FOR OFFICIAL DISTRIBUTION.	19. Security Class. (This page)	21. Price:
	Unclassified	N/A

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.

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 itsinfo@ntia.doc.gov.