NTIA Report 18-529

Intelligibility Robustness of Five Speech Codec Modes in Frame-Erasure and Background-Noise Environments

Stephen D. Voran
Andrew A. Catellier

report series

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

NTIA Report 18-529

Intelligibility Robustness of Five Speech Codec Modes in Frame-Erasure and Background Noise-Environments

Stephen D. Voran
Andrew A. Catellier

U.S. DEPARTMENT OF COMMERCE

December 2017

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 Institute for Telecommunications Sciences in collaboration with 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 five different speech codec modes in operating environments that include frame erasures or background noise. This report constitutes the final deliverable product for this project.

CONTENTS

Figures viii

Tables ix

Acronyms and Abbreviations x

Executive Summary xi

1. Background 1

2. Intelligibility Test 1 — Targeted Consecutive Frame Erasures 4
   2.1 Speech Codec Modes 4
   2.2 Modified Rhyme Test Considerations 5
   2.3 Frame Erasures 6
   2.4 Background Noise 8
   2.5 Operating Environments 8
   2.6 Modified Rhyme Test Execution 9
   2.7 Results 9

3. Intelligibility Test 2 — Random Frame Erasures and Background Noise 19
   3.1 Frame Erasures and Operating Environments 19
   3.2 Modified Rhyme Test Execution 22
   3.3 Results 23
      3.3.1 Frame-Erasure Results 25
      3.3.2 Noise Results 30
      3.3.3 General Results 32

4. Summary 35

5. References 36

Acknowledgements 38

FIGURES

Figure 1. Example patterns of consecutive erased frames used in Test 1. Black indicates erased frame. 7

Figure 2. Intelligibility results for five codec modes and twelve frame-erasure conditions. 12

Figure 3. Thresholds for upward and downward intelligibly deviations for 95% significance when 480 MRT trials are used. 15

Figure 4. Intelligibility results averaged over three non-CA codec modes and two CA codec modes for twelve fame erasure conditions. Statistically significant differences are marked with yellow. 17

Figure 5. Two-state Gauss-Markov model for generating frame-erasure patterns. 20

Figure 6. Intelligibility versus FER for five codec modes. 25

Figure 7. Thresholds for upward and downward intelligibly deviations for 95% significance when 480 and 2160 MRT trials are used. 26

Figure 8. Intelligibility versus FER for AMR-WB and two CA modes averaged. Red arrow shows that AMR-WB 2% FER intelligibility matches CA 5.1% intelligibility. Black arrow shows that AMR-WB 10% FER intelligibility matches CA 12.9% intelligibility. 28

Figure 9. Intelligibility versus SNR for five codec modes. 31

TABLES

Table 1. Five codec modes tested. 4

Table 2. Twelve operating environments used in Test 1. Each environment is applied to all 5 codec modes to produce a total of 60 conditions. 8

Table 3. Results of Test 1 (480 trials per condition)....10

Table 4. Test 1 intelligibility results for five codec modes....12

Table 5. Example table comparing results for codec mode X with codec mode Y. 13

Table 6. Increases in intelligibility that are significant at the 95% level. 15

Table 7. Increases in intelligibility that are significant at the 95% level, presented with symbols to aid in visual grouping. 16

Table 8. Properties of twelve operating environments. Each environment is applied to each of the five codec modes to produce a total of 60 conditions. 21

Table 9. Measured FER statistics across 1080 MRT recordings. 22

Table 10. Measured MEL statistics across 1080 MRT recordings. 22

Table 11. Results of Test 2 (2160 trials per condition)....23

Table 12. Test 2 intelligibility results for five codec modes....25

Table 13. Increases in intelligibility that are significant at the 95% level, frame-erasure environments only. 27

Table 14. Increases in intelligibility that are significant at the 95% level, frame-erasure environments only, presented with symbols to aid in visual grouping. 27

Table 15. Average FER increase tolerated by CA relative to non-CA. 29

Table 16. Increases in intelligibility that are significant at the 95% level, noise environments only. 31

Table 17. Increases in intelligibility that are significant at the 95% level, noise environments only, presented with symbols to aid in visual grouping. 32

Table 18. Equivalences between FER and SNR based on intelligibility. 33

ACRONYMS AND ABBREVIATIONS

3GPP	Third Generation Partnership Project
AMR-WB	Adaptive Multi-Rate Wideband
ANSI	American National Standards Institute
APCO	Association of Public-Safety Communications Officials
CA	Channel Aware
CMRT	Crowdsourced Modified Rhyme Test
EVS	Enhanced Voice Services
FER	Frame-Erasure Rate
FM	Frequency Modulation
Hz	Hertz
ITU-T	International Telecommunication Union, Telecommunication Standardization Sector
kb/s	Kilobits/second
kHz	Kilohertz
LMR	Land Mobile Radio
LTE	Long-Term Evolution
MCV	Mission-Critical Voice
MEL	Mean Erasure Length
MRT	Modified Rhyme Test
NB	Narrowband
NPSTC	National Public Safety Telecommunications Council
P25	Project 25
POLQA	Perceptual Objective Listening Quality Assessment
PSCR	Public Safety Communications Research Program
RAN	Radio Access Network
smp/s	Samples per second
SNR	Signal-to-Noise Ratio
SWB	Super-Wideband
WB	Wideband

EXECUTIVE SUMMARY

This report describes speech intelligibility measurements for five speech codec operating modes from the adaptive multi-rate (AMR) and enhanced voice services (EVS) speech codec families. All five codec modes use bit rates near 13 kb/s. Four wideband (WB) and one super wideband (SWB) codec mode are included. These codec modes are of interest because of their potential to carry public safety mission-critical voice (MCV) communications over LTE radio access networks (RANs). Two EVS Channel Aware (CA) modes are included. These modes selectively apply redundant coding to enable higher robustness to erased frames. In addition, a more robust version of AMR is considered by pairing the AMR-WB encoder with the decoder specified in ITU-T Recommendation G.718. Thus the five codec modes measured in this report are denoted as: AMR-WB, AMR-WB/G.718, EVS-WB, EVS-WB CA, and EVS-SWB CA.

Measurements are applied to these codec modes under ideal conditions and also under a range of frame-erasure conditions and background-noise conditions. Both frame erasures and background noise will be encountered in actual operational scenarios. Erased frames occur when RANs are stretched to their limits by high loading, marginal signal strengths, or interference. The added robustness to erased frames offered by the CA modes, and how that is manifested in speech intelligibility, is of particular interest. The relationships between intelligibility and erased frames are particularly important for MCV because scenarios that stress the RAN may well be scenarios that also require high intelligibility.

Two separate tests are described, and both use the Modified Rhyme Test (MRT) protocol. This protocol requires listeners to identify which of six different words were presented and the success rate for this task forms a measure of intelligibility. Both tests were implemented using the Crowdsourced MRT (CMRT) protocol. This protocol relinquishes tight laboratory controls, allows remote participation of large numbers of self-selecting listeners, and efficiently generates the required number of trials. A previous study has demonstrated that CMRT produces results that are equivalent to laboratory MRT results.

Test 1 focused on consecutive frame erasures targeted at MRT keywords to achieve maximum sensitivity. The number of consecutive frames erased ranged from one to eleven. The test used 480 MRT trials for each of 60 conditions — a total of 28,800 trials. Statistical analysis of the trials revealed numerous differences in speech intelligibility that were significant at the 95% level. At the highest level there is a clear common thread — all of these differences were cases where a CA codec mode produced higher intelligibly than a non-CA codec mode. But few consistent trends are present within that broader result. Each of the significant differences falls into the range of three to nine consecutive frames erased. Yet for each CA codec mode, there are multiple frame-erasure cases in that range where no significant intelligibility advantage is found. In addition, there is no consistency with respect to which of the three non-CA modes is exceeded by a CA mode, and neither CA mode shows an advantage over the other. Together these results informed the design of Test 2, including the frame-erasure environments and the number of trials per condition.

Test 2 used randomly occurring frame erasures produced by a two-state Gauss-Markov Model. Through proper selection of the model parameters, the model produced frame-erasure rates (FERs) ranging from 5 to 30% and mean erasure lengths ranging from 4 to 9 frames. Test 2

included the simulation of background noise at the transmitting location at signal-to-noise ratios ranging from -5 to +20 dB. The test used coffee shop noise recordings that include a rich mix of sources including multiple moving talkers, music, and coffee making sounds. Test 2 used 2160 trials (instead of 480) for each condition thus increasing the resolving power of the test over that of Test 1. Test 2 produced a grand total of 129,600 trials.

The results of Test 2 show numerous small but statistically significant intelligibility improvements for the CA codec modes across the frame-erasure environments. The CA codec modes show fairly consistent intelligibility improvements over the AMR-WB codec mode — these improvements occur in nine out of twelve cases. The CA codec modes show fewer intelligibility improvements over the EVS-WB codec mode with improvements occurring in just six of twelve cases. And when the AMR-WB encoder is coupled with the decoder specified in ITU-T Recommendation G.718, the CA codec modes offer improvement in only four of the twelve cases.

In addition to comparing intelligibility at fixed FER values, the report also provides comparisons of FER values at fixed intelligibility levels (by means of interpolation.) For a fixed reference intelligibility set by a non-CA mode, the CA modes can tolerate absolute FER values that are higher by 2.9 to 3.8%. The measurements show no differences between EVS-WB CA and EVS-SWB CA in the frame-erasure environments. In the noise environments EVS-SWB CA has intelligibly advantages in some cases while EVS-WB CA is exceeded by other codec modes in several cases. Additional analyses show how noise and frame-erasure environments compare in terms of the intelligibility results they produce. The report also compares five conditions that are similar between Test 1 and Test 2 to draw conclusions about test repeatability and sources of variation.

Overall, it is clear that when using large numbers of trials the MRT results show that the CA codec modes (EVS-WB CA and EVS-SWB CA) offer small but statistically significant speech intelligibility improvements in numerous frame-erasure environments. The detailed results reported here are available to inform some of the design and provisioning choices required in the development, deployment, and tuning of LTE based mission-critical voice equipment, applications, and services.

INTELLIGIBILITY ROBUSTNESS OF FIVE SPEECH CODEC MODES IN FRAME-ERASURE AND BACKGROUND NOISE-ENVIRONMENTS

Stephen D. Voran[1] and Andrew A. Catellier[2]

Frame erasures and background noise are two factors that can interact with speech coding to reduce speech intelligibility and thus impair public safety mission-critical voice communications. We conducted two tests of intelligibility in the face of these factors. The tests covered five adaptive multi-rate (AMR) and enhanced voice services (EVS) speech coding modes, each using a bit rate near 13 kb/s. Two EVS Channel Aware (CA) modes were included. Both tests use the Modified Rhyme Test (MRT) protocol and together they comprise over 150,000 trials. The first test used frame erasures targeted at critical consonants for maximum sensitivity and the second used frame erasures generated at random by a two-state Gauss-Markov model. By using these large numbers of MRT trials we found that the CA codec modes offer small but statistically significant speech intelligibility improvements in numerous frame-erasure environments.

Keywords: AMR, EVS, channel aware, frame erasure, frame loss, MRT, noise, packet loss, speech coding, speech intelligibility, speech quality

1. BACKGROUND

A fundamental requirement for many telecommunications services is the delivery of intelligible speech. In public safety telecommunications, high intelligibility supports efficient execution of time-critical tasks. Lower intelligibility can lead to requests for repetitions that slow operations, or can even create misunderstandings that jeopardize operations, safety, or lives. In its published requirements for Mission-Critical Voice (MCV) networks for public safety, the National Public Safety Telecommunications Council (NPSTC) addresses this requirement under the heading of "audio quality" in [1]. The requirement contains a set of four quality-of-service thresholds and prioritizes them:

> "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 author is with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.
[2] The author was formerly with the Institute for Telecommunication Sciences.

We have investigated items 2 and 3 on the list: speaker identification and detection of speaker stress [2]–[4]. We tested speech intelligibility in parallel with the ability to identify a speaker from a set of speakers, and the ability to detect dramatized urgency in a speaker's voice. These tests were repeated across multiple telecommunication systems and we found that speech intelligibility degraded more rapidly than the ability to identify speakers or detect dramatized urgency. 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. These results reaffirm that item 1 on the NPSTC list above is the critical issue.

Accordingly the Public Safety Communications Research Program (PSCR) has performed numerous tests of speech intelligibility for various telecommunication systems and environments that are relevant to public safety operations. Our earliest work was motivated by field reports from our public safety partners and was primarily focused on analog FM and APCO Project 25 (P25) land mobile radio (LMR) systems, with an emphasis on difficult noise environments at the transmitting location [5]–[7]. We then shifted to consider the effects of digital speech and audio coding options that might be used to provide MCV over Long-Term Evolution (LTE) based radio networks [7], [8]. This work initially focused on effects of background noise at the transmitting location, and then extended to include effects of erased data frames in a radio access network (RAN) [9].

The motivation for this extension is as follows. When a RAN and the individual underlying radio links are stressed, the frames of data that carry encoded voice can be corrupted, excessively delayed, lost, or deleted. (In this report we used the term "erased" to cover all of these situations.) The erasure of speech data frames will inevitably cause at least some reduction in speech intelligibility and ultimately can completely obliterate any possibility of communications. Characterizing the relationship between the radio link and intelligibility is particularly important for MCV because the very activities that stress the RAN are likely to be activities that also require high intelligibility.

For example, when an emergency event is escalating additional personnel will typically report to the scene. When personnel share radio resources on the scene, those radio resources will (barring any mitigating measures) inevitably be spread thinner and thinner. This can result in negative consequences for speech intelligibility even as it becomes particularly important to coordinate the new personnel. Consider also when personnel must move deeper into a building to perform critical functions. Unless mitigation measures are in place, the radio link will suffer additional attenuation and there may then be negative consequences for speech intelligibility even as it is becoming more and more critical. Finally, every radio system has a finite coverage area and as users approach the limits of that area, weaker radio signals result in data losses that at some point will lead to reductions in speech intelligibility.

Viable strategies to mitigate these issues exist. These could include the use of deployable radio resources or the adjustment of RAN priorities. The goal is to provide the required radio resources when and where they are most critically needed. These decisions should not be based on untested theoretical thresholds, but rather on the critical user experience factor: speech intelligibility. Thus it is crucial that the relationship between frame erasures and speech intelligibility be well understood.

An additional key factor is the ability of a speech codec to preserve speech intelligibility even when data frames are erased. This property is called frame-erasure robustness and if all other things are equal, then a codec with greater frame-erasure robustness is more desirable than one with lesser robustness. One motivation for our work in [9] was to characterize the potential additional robustness provided by the Channel Aware (CA) modes offered by the Enhanced Voice Services (EVS) codec [10]. Unfortunately, after our testing was completed we were notified of defects in the EVS reference software implementation (Version 12.5.0) provided to us. These defects prevented CA from operating properly and thus CA modes showed no additional robustness in those tests.

Since the completion of our work described in [9], corrected versions of the EVS reference software implementation have become available. Our preliminary testing showed that the problem was resolved and some minimal additional robustness was measurable but only with rather sensitive testing. Those tests also showed strong interactions between background noise at the transmitting location and frame erasures. These preliminary test results allowed us to design, implement, and analyze the two tests described in this report. Both tests use the Modified Rhyme Test (MRT) paradigm for quantifying speech intelligibility.

Section 2 describes Test 1 in detail. This test was designed for maximum sensitivity and modest size (480 trials per condition). Test 1 used twelve levels of deterministically controlled frame erasures targeted at the critical consonants of the MRT keywords to maximally and precisely affect intelligibility. It used a single, minimal level of background noise to avoid interactions with the frame-erasure effects.

The results of Test 1 test informed the design of Test 2, described in Section 3. This test used a more conventional stochastic model for frame erasures. It included a range of background-noise levels held independent from the levels of frame erasure. Test 2 used 2180 trials per condition and identified numerous situations where the CA modes offer modest improvements in MRT speech intelligibility. Results in Section 3.3 detail every case where there is a statistically significant difference between the speech intelligibilities delivered by two different codec modes. Intelligibility increases are also viewed as robustness increases, thus leading to values of frame-erasure rate increases that can be tolerated by the more robust codec modes. These frame-erasure rate increases are based on constant measured speech intelligibility and can be compared with previously published analogous results that are based on constant estimated speech quality.

2. INTELLIGIBILITY TEST 1 — TARGETED CONSECUTIVE FRAME ERASURES

The first test described in this report used deterministically placed frame erasures to achieve maximum effect with modest test size. In this section we step through each test factor or consideration, culminating with exposition and discussion of the results of Test 1.

2.1 Speech Codec Modes

This work is driven by the need to provide MCV services to public safety users over an LTE based RAN. The adaptive multi-rate (AMR) and EVS speech codec families are well-suited to this application and they are the focus of our present tests, consistent with our earlier work described in [9].

A defining characteristic of a speech codec is its nominal audio bandwidth. 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 narrowband (NB) bandwidth used for telephony (300-3500 Hz), the WB and SWB options can improve the "realism" or "presence" of communications. While the vast majority of speech information is passed by NB systems, WB and SWB systems have the potential to offer minor improvements in speech intelligibility, depending on the relative spectral content of the speech and the background noise. As a point of reference, when there is no background noise, the articulation index predicts that word intelligibly in sentence context will increase from 99.3% to 99.9% when NB is replaced with WB [11]. Test 1 includes four WB codec modes and one SWB codec mode.

The five codec modes used in this test are specified in Table 1. The selected bit rates are very close in the absolute sense and they are actually equivalent in terms of their consumption of resource blocks in an LTE-based radio access network. Note that all five codec modes process speech in 20 ms frames. For each such speech frame the speech encoder produces a "speech codec data frame" or "frame" for short.

Table 1. Five codec modes tested.

Codec Mode	**Description**	**Bit Rate**	**Nominal Audio Bandwidth**
AMR-WB	Adaptive Multi-Rate wideband encoding and decoding	12.65 kb/s	50 Hz to 7 kHz
AMR-WB/G.718	Adaptive Multi-Rate wideband encoding and G.718 wideband decoding.	12.65 kb/s	50 Hz to 7 kHz
EVS-WB	Enhanced Voice Services wideband encoding and decoding.	13.20 kb/s	50 Hz to 7 kHz
EVS-WB CA	Enhanced Voice Services wideband encoding and decoding with channel aware mode activated for robustness to frame erasures.	13.20 kb/s	50 Hz to 7 kHz
EVS-SWB CA	Enhanced Voice Services super-wideband encoding and decoding with channel aware mode activated for robustness to frame erasures.	13.20 kb/s	50 Hz to 16 kHz

The AMR-WB [12] codec is specified in Third Generation Partnership Project (3GPP) Technical Specification (TS) 26.204. The software implementation used in this study is version 7.0.0[3], which is distributed as part of that technical specification and available from 3gpp.org. The G.718 decoding option [13] for AMR-WB offers updated and enhanced decoding with potential additional robustness to erased frames. We used software version 1.7 available from itu.int.

The EVS codec [10] was standardized by the 3GPP in September 2014. We used the latest available software which was provided as part of TS 26.442 via 3gpp.org. Thus in Test 1 we used version 13.3.0 and in Test 2 we used 13.4.0. The differences between the two versions were intended to resolve some identified issues in specific cases but are not expected to impact baseline speech intelligibility nor robustness to frame erasures.

Given the motivation for our testing, the channel aware (CA) modes of EVS are of particular interest. These modes selectively apply redundant coding to enable higher robustness to erased frames [14]. This redundant coding exploits time diversity — if a frame is erased then some of the lost information may be available in a subsequent frame, and if that frame arrives in time for playout then the erasure may be mitigated. The CA modes do not increase the bit rate. Instead they selectively enforce a minor reduction in the bits available for the primary coding so that some bits can be available for the redundant coding.

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 an FEC offset value of three frames (invoked on the command line by specifying '-rf 3'). This parameter controls the level of time diversity — greater values increase the time diversity, the robustness to isolated frame erasures, and the end-to-end delay that is required in order to take advantage of that robustness. When FEC offset is set to three frames, then the CA decoder must have access to the three frames that follow the frame that it is currently decoding for playout.

2.2 Modified Rhyme Test Considerations

The PSCR measures speech intelligibility using the Modified Rhyme Test (MRT). This selection was made in collaboration with public safety stakeholders when the PSCR undertook the original work in this area [5] and more details are provided in [8]. The MRT is fully defined in [15]. In an MRT trial, a subject must identify the word presented from a set of six words and the success at this task forms the measure of speech intelligibly. In a high-intelligibility system, this task is easy and success rates are high. In a low intelligibly system the task is difficult and success rates are lower. Other tests of speech intelligibility are available. Thus the intelligibility results presented in this report would be most precisely described as "MRT intelligibility" results. In the interest of conciseness we often use the simplified term "intelligibility" in this report.

The MRT protocol specifies 50 groups containing 6 words each. Twenty-five of the word groups contain words that differ only in the initial consonant sound, for example "bed," "led," "fed," "red," "wed," and "shed." The other twenty-five word groups contain words that differ only in the final consonant sound, for example "dug," "dung," "duck," "dud," "dub," and "dun." In the

[3] AMR-WB software is relatively stable and version 7.0.0 is equivalent to version 13.0.0.

MRT, each word is presented in a carrier sentence: “Please select the word —.” So when the keyword is “bed,” the entire presentation is “Please select the word bed.”

Our MRT source material was recorded by two female and two male talkers. Each is a native speaker of North American English. Each talker recorded 300 sentences, consisting of the 50 groups of 6 words, each in the standard carrier sentence. This is a total of 1200 recorded sentences. We used professional audio equipment and a 48,000 smp/s sample rate to obtain full bandwidth, low noise, low distortion recordings. Additional details are provided in [8].

Since MRT intelligibility hinges on a leading or trailing consonant sound, we targeted this “critical consonant” with frame erasures for maximum impact and thus maximum sensitivity testing. Thus we deterministically erased 0, 1, 2, …, 11 frames located at the critical consonant.

To further maximize impact and sensitivity, we also selected the keywords with the shortest duration of delivery from each talker. A fixed number of frames corresponds to a fixed time duration which in turn becomes a larger fraction of a shorter keyword. For example an 80 ms (4 frame) erasure is 25% of a 320 ms keyword but it is only 16% of a 500 ms keyword.

We determined that we could include ten word groups with initial critical consonants and ten word groups with final critical consonants from each talker. This selection was done independently for each talker and for each category (initial or final critical consonant). To select these 10 groups from the 25 possible groups with initial critical consonants we first found the longest keyword in each of the 25 groups of keywords (“the maximal keyword”) and then used the lengths of these 25 maximal keywords to sort the 25 word groups. We then selected the ten word groups with the shortest maximal keywords. We repeated the process for the final critical consonant word groups. That is, we selected 10 word groups for each talker from the 25 word groups with final critical consonants.

This process produced 20 word groups for each of the 4 talkers, or a total of 80 word groups. Since each word group contains 6 keywords this produced a total of 480 keywords, and each condition was thus tested with 480 MRT trials.

2.3 Frame Erasures

Our goal in frame erasure is to impair the critical consonant sound in the MRT keyword. We studied the lengths of these consonant sounds in order to come up with a general rule for their typical locations within the keyword. On average the critical consonant sounds occupy the first 15 frames (300 ms) or last 18 frames (360 ms) of the keyword. Thus we developed the following frame-erasure procedure. The procedure involves selecting a constrained random location for the center of the frame erasure and then allowing the frame-erasure pattern to grow symmetrically about that location. This means that when n frames (n = 1 to 10) are erased, those frames are a subset of the frames that will be erased when n+1 frames are erased. The motivation behind this approach is to minimize the variation in erased frame *locations* as we adjust the *number* of erased frames.

The procedure for erasing frames at the initial consonant is as follows. Our code picks at random[4] and with equal probability frame 6, 7, 8, 9, or 10 as the central location for the frame-erasure pattern. In the case where only one frame is erased, this central frame (frame *c*) is the only frame erased. When *n* = 2, 3, …, 11 consecutive frames are to be erased, the number of the first frame erased is

$$c - \left(\left\lceil \frac{n}{2} \right\rceil - 1\right), \tag{1}$$

and the number of the final frame erased is

$$c + \left\lfloor \frac{n}{2} \right\rfloor, \tag{2}$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are the ceiling and floor functions respectively. Figure 1 shows the relationships between these patterns for the cases of 1, 2, 3, and 4 consecutive frames erased.

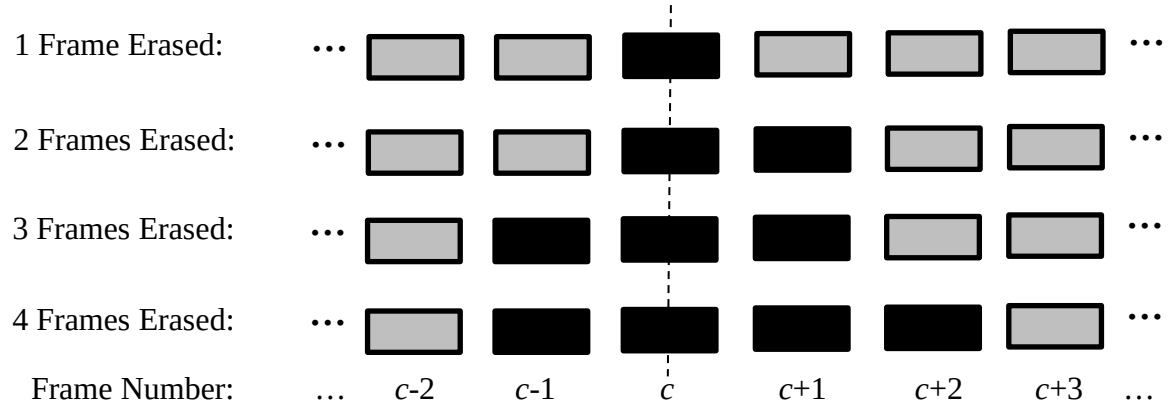

Figure 1. Example patterns of consecutive erased frames used in Test 1. Black indicates erased frame.

By this procedure the initial erased frame is never before frame 1 and the final erased frame is never after frame 15. Thus all frame-erasure patterns are successfully located inside the desired 15 frame (300 ms) window.

The procedure for erasing frames at the final consonant is analogous, but at the end of the keyword. When *N* is the number of the final frame in the keyword, our code picks at random and with equal probability frame *N*-12, *N*-11, …, or *N*-5 as the central location for the frame-erasure pattern. In the case where only one frame is erased, this central frame (frame *c*) is the only frame erased. When *n* = 2, 3, …, 11 consecutive frames are to be erased, the numbers of the first and last erased frames are again given by (1) and (2) respectively. By this procedure the initial erased frame is never before frame *N*-17 and the final erased frame is never after frame *N*. Thus all frame-erasure patterns are successfully located inside the desired 18 frame (360 ms) window.

We selected one central location for each of the 480 MRT source recordings resulting in 480 central locations. We used this central location consistently across all frame-erasure lengths for

[4] This report contains numerous references to "random" selections, orders, etc. Of course the more precise adjective would be "pseudo-random" but we use the term "random" for conciseness.

all five codec modes. Our motivation is to make the different codec modes be the only source of variation.

2.4 Background Noise

We can mix a noise signal with the MRT source speech to simulate background noises at a transmitting location. This test is focused on frame erasures and we seek to minimize variation due to other sources. Thus we selected a single noise mixture for this test. We selected a coffee shop noise recording that contains a broad mixture of sources including multiple moving talkers, music, and the various sounds of coffee making processes and machines.

We do not want the presence of noise to place a significant upper limit on intelligibly in this test. Thus, based on prior experience, we selected a moderate SNR of 20 dB. For each of the 480 MRT source recordings, we excerpted at random an appropriate length segment (1.8 seconds on average) from the full noise recording (328 seconds). We then mixed the excerpted noise with the speech to achieve 20 dB SNR.[5] We used the resulting mixture signal consistently with all frame-erasure patterns and all codec modes. That is, we used the same noise recording segment with any given MRT source recording for each codec mode and frame-erasure pattern.

2.5 Operating Environments

In this test an "operating environment" is defined by a combination of frame-erasure factor and the background-noise factor. The operating environments are shown in Table 2. Each of the 12 operating environments is applied to all 5 codec modes to produce a total of 60 conditions to be tested.

Table 2. Twelve operating environments used in Test 1. Each environment is applied to all 5 codec modes to produce a total of 60 conditions.

Speech Codec Frame Erasure	Erasure Duration	Background Noise
None	—	Coffee shop noise, 20 dB SNR
1 Frame erased	20 ms	Coffee shop noise, 20 dB SNR
2 Frames erased	40 ms	Coffee shop noise, 20 dB SNR
3 Frames erased	60 ms	Coffee shop noise, 20 dB SNR
4 Frames erased	80 ms	Coffee shop noise, 20 dB SNR
5 Frames erased	100 ms	Coffee shop noise, 20 dB SNR
6 Frames erased	120 ms	Coffee shop noise, 20 dB SNR
7 Frames erased	140 ms	Coffee shop noise, 20 dB SNR
8 Frames erased	160 ms	Coffee shop noise, 20 dB SNR
9 Frames erased	180 ms	Coffee shop noise, 20 dB SNR
10 Frames erased	200 ms	Coffee shop noise, 20 dB SNR
11 Frames erased	220 ms	Coffee shop noise, 20 dB SNR

[5]All SNRs in this report are based on A-weighted level measurements of speech and noise.

2.6 Modified Rhyme Test Execution

We tested each of the 60 conditions (5 codec modes crossed with 12 operating environments) in this test by means of 480 MRT trials (20 word groups with six words per group crossed with four talkers). This is a total of 28,800 MRT trials. We conducted these trials using our recently-developed Crowdsourced MRT (CMRT) paradigm [16]. CMRT offers the trials to listeners as microwork using the Mechanical Turk platform offered by Amazon Web Services. Registered workers can accept the work, complete the trials, and receive payment from their location of choice using a web browser, an Internet connection, and speakers or headphones. In the case of CMRT, these workers are called "listeners." We cannot control the conditions under which the listeners perform the trials, but we are able to motivate a serious effort through a bonus payment system. We have found that our CMRT protocols produce results that are equivalent to or better than those that we gather from laboratory-based MRT work [16]. A key factor in the success of CMRT is that it affords much larger listener sample sizes than laboratory MRT does.

Our CMRT protocols include a data processing step [16] in order to provide high-quality results that are commensurate with those obtained by laboratory MRT. This step reduces the data by a factor of two, so in order to obtain 28,800 final trails, we first used CMRT to produce 57,600 raw trials. We packaged the trials into tasks that contained 60 trials, one from each condition. This produced 480 tasks. This packaging used constrained randomization and the trials were unlabeled and presented in a random order. The estimated maximum time to complete a task was 5 minutes.

We made 240 tasks available at 11:00 AM MDT on March 21, 2017. We required two different listeners to complete each task. By 12:04 PM each task had been claimed by two different listeners and all results were submitted shortly thereafter. We made the remaining 240 tasks available at 11:00 AM MDT on March 22, 2017. We again required two different listeners to complete each task. By 12:28 PM each task had been claimed by two different listeners and all results were submitted shortly thereafter.

Of the 57,600 raw trials conducted only 35 (0.06%) produced invalid results (no word was selected). This is a very low failure rate and driving it to zero would be difficult in light of the vast variation in web browsers, playback mechanisms, and human factors. On average, 232 distinct listeners participated in each group of 240 tasks. Since two different listeners completed each task, the average number of tasks completed by each listener is $2 \times 240/232 = 2.1$.

2.7 Results

The final CMRT results consist of 480 trials for each of the 60 conditions. Each trial produces either success or failure. The number of successes and the success rates are tabulated in Table 3. The success rate provides the basis for reporting MRT intelligibility. Because the MRT offers six word choices, a listener can make correct selections one-sixth of the time even with the speech signal turned off (clearly a case of zero intelligibility). Thus [15] specifies a transformation that maps a success rate of one-sixth to zero intelligibility. It also maps a success rate of one to an intelligibly of one:

$$Intelligibility = \frac{6}{5}\left(Success\ Rate - \frac{1}{6}\right). \quad (3)$$

The final column of Table 3 gives the MRT intelligibility as defined in (3).

Table 3. Results of Test 1 (480 trials per condition).

Codec Mode	**Number of Frames Erased**	**Number of Successes**	**Success Rate**	**Intelligibility**
AMR-WB	0	459	0.956	0.948
	1	457	0.952	0.943
	2	449	0.935	0.923
	3	427	0.890	0.868
	4	423	0.881	0.858
	5	410	0.854	0.825
	6	405	0.844	0.813
	7	372	0.775	0.730
	8	370	0.771	0.725
	9	320	0.667	0.600
	10	327	0.681	0.618
	11	269	0.560	0.473
AMR-WB/G.718	0	461	0.960	0.953
	1	466	0.971	0.965
	2	453	0.944	0.933
	3	442	0.921	0.905
	4	420	0.875	0.850
	5	426	0.888	0.865
	6	387	0.806	0.768
	7	379	0.790	0.748
	8	373	0.777	0.733
	9	313	0.652	0.583
	10	315	0.656	0.588
	11	264	0.550	0.460

Codec Mode	Number of Frames Erased	Number of Successes	Success Rate	Intelligibility
EVS-WB	0	462	0.963	0.955
	1	459	0.956	0.948
	2	447	0.931	0.918
	3	443	0.923	0.908
	4	424	0.883	0.860
	5	407	0.848	0.818
	6	391	0.815	0.778
	7	377	0.785	0.743
	8	358	0.746	0.695
	9	324	0.675	0.610
	10	313	0.652	0.583
	11	275	0.573	0.488
EVS-WB CA	0	464	0.967	0.960
	1	454	0.946	0.935
	2	454	0.946	0.935
	3	442	0.921	0.905
	4	440	0.917	0.900
	5	426	0.888	0.865
	6	409	0.852	0.823
	7	399	0.831	0.798
	8	385	0.802	0.763
	9	356	0.742	0.690
	10	328	0.683	0.620
	11	285	0.594	0.513
EVS-SWB CA	0	467	0.973	0.968
	1	456	0.950	0.940
	2	455	0.948	0.938
	3	448	0.933	0.920
	4	426	0.888	0.865
	5	424	0.883	0.860
	6	429	0.894	0.873
	7	392	0.817	0.780
	8	381	0.794	0.753
	9	346	0.721	0.665
	10	328	0.683	0.620
	11	282	0.588	0.505

We have used the exact same MRT source recordings, background-noise excerpts, frame-erasure patterns, and number of trials for each of the five codec modes. This allows us to directly compare results across codec modes at each operating environment. To facilitate this comparison

Table 4 shows the intelligibly results of Table 3 reorganized with one codec mode in each column.

Table 4. Test 1 intelligibility results for five codec modes.

Number of Frames Erased	AMR-WB	AMR-WB/ G.718	EVS-WB	EVS-WB CA	EVS-SWB CA
0	0.948	0.953	0.955	0.960	0.968
1	0.943	0.965	0.948	0.935	0.940
2	0.923	0.933	0.918	0.935	0.938
3	0.868	0.905	0.908	0.905	0.920
4	0.858	0.850	0.860	0.900	0.865
5	0.825	0.865	0.818	0.865	0.860
6	0.813	0.768	0.778	0.823	0.873
7	0.730	0.748	0.743	0.798	0.780
8	0.725	0.733	0.695	0.763	0.753
9	0.600	0.583	0.610	0.690	0.665
10	0.618	0.588	0.583	0.620	0.620
11	0.473	0.460	0.488	0.513	0.505

A graphical presentation is given in Figure 2.

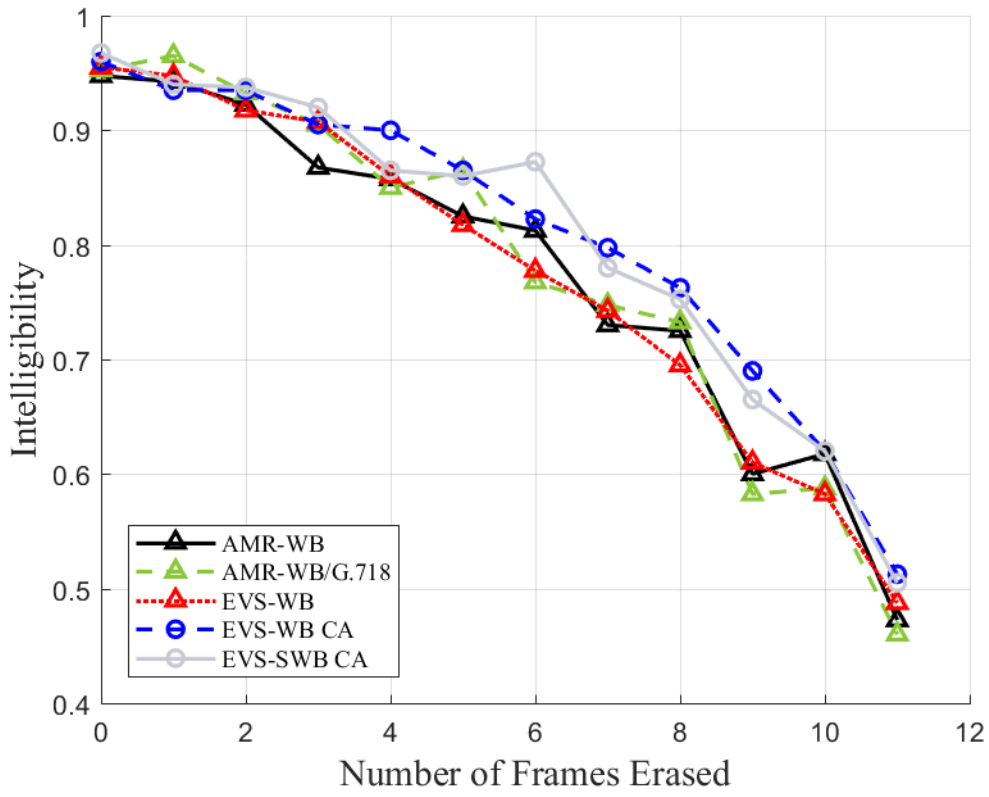

Figure 2. Intelligibility results for five codec modes and twelve frame-erasure conditions.

Figure 2 shows that intelligibility is very high for the case of no frame erasures, ranging from 0.95 to 0.97. As expected, intelligibility drops (mostly monotonically) as the number of consecutive frames erased at the critical consonant is increased to eleven. This limiting case produces intelligibility values ranging from 0.46 to 0.51.

Each of the calculated intelligibility values is based on a large but finite number of Bernoulli trials and thus has some inherent uncertainty. As with any such testing, we therefore must consider the question of statistical significance. Our approach is to state a null hypothesis and perform statistical tests to determine when we should reject the null hypothesis.

We now provide the mathematical basis for these necessary statistical tests. For any two codec modes X and Y we can tabulate successes and failures for each mode as shown Table 5.

Table 5. Example table comparing results for codec mode X with codec mode Y.

	Number of Successes	**Number of Failures**	**Total**
Codec Mode X	S_X	F_X	N_X
Codec Mode Y	S_Y	F_Y	N_Y
Total	$S_X + S_Y$	$F_X + F_Y$	$N_X + N_Y$

We can apply the chi-squared test for independence of categorical data [17]–[19] to test for independence of these numbers of successes with respect to the row variable (codec modes X vs. Y). Thus the null hypothesis is "the success rates defined by the two rows are independent of the labeling of the rows." In other words, codec modes X and Y do not have statistically significantly different success rates.

To apply the chi-squared test for independence of categorical data we form the chi-squared statistic from the normalized squared deviations between the observed results and the expected results under the null hypothesis. Since our results are balanced we have, $N_X = N_Y = N$ and this simplifies the expressions that follow. The expected results are easily extracted from the totals given in the Table 5:

$$S_{NULL} = \frac{S_X + S_Y}{N_X + N_Y} N = \frac{S_X + S_Y}{2}, F_{NULL} = N - S_{NULL}. \tag{4}$$

Next we form the chi-squared (χ^2) statistic associated with the two-by-two core of Table 5:

$$\begin{aligned} \chi^2 &= \frac{(S_X - S_{NULL})^2}{S_{NULL}} + \frac{(S_Y - S_{NULL})^2}{S_{NULL}} + \frac{(F_X - F_{NULL})^2}{F_{NULL}} + \frac{(F_Y - F_{NULL})^2}{F_{NULL}} \\ &= 2\left(\frac{(S_X - S_{NULL})^2}{S_{NULL}} + \frac{(F_X - F_{NULL})^2}{F_{NULL}}\right) = 2\left(\frac{(S_Y - S_{NULL})^2}{S_{NULL}} + \frac{(F_Y - F_{NULL})^2}{F_{NULL}}\right). \end{aligned} \tag{5}$$

Equation (5) shows that the χ^2 statistic is the sum of normalized-squared deviations from the null values. Because the null values are centered between those of X and Y, either X or Y results can

be used to find the deviations. This χ^2 statistic has one degree-of-freedom because $(number\ of\ rows - 1) \times (number\ of\ columns - 1\ ) = 1$.

The chi-squared statistic measures the deviation of the outcomes for codec mode X or Y from the outcome that is expected under the null hypothesis. It goes to zero as outcomes for codec modes X and Y converge and it gets larger as they diverge. The cumulative distribution function of this statistic is well-characterized and thus it is known that when the null-hypothesis is true the statistic will exceed 3.841 less than 5% of the time [17]–[19]. Across many disciplines in science and engineering it is common to reject the null hypothesis when the probability of rejecting it erroneously is less than 5%. This can be described as a 95% significance or confidence level. We adopt this practice in this work and when $3.841 < \chi 2$ we reject the null hypothesis.

Significance testing is performed in the domain of counts (success, failures, and total trials) but its results may be most easily appreciated in the intelligibility domain. The two domains are linked by (3) and the result is shown in Figure 3. This figure shows the upward and downward deviations in intelligibility that produce $\chi 2 = 3.841$ and thus form the boundary between deviations that are significant and deviations that are not significant at the 95% level. These thresholds for significance can be called "resolution functions" because they show what differences can be resolved across the intelligibility scale.

An example reading of Figure 3 follows. Suppose we measure the intelligibility of codec mode X using 480 MRT trials and obtain 0.80. This is the reference intelligibility. Figure 3 shows that for this reference value the upward and downward threshold deviations are 0.053 and 0.060 respectively. This means a second intelligibility measurement must exceed 0.800 + 0.053=0.853 in order to show significantly higher intelligibility than the reference measurement. Conversely a second measurement must be smaller than 0.800 − 0.060 = 0.740 in order to show significantly lower intelligibility than the reference measurement.

We now apply significance testing to the MRT results that we have obtained. Each row of Table 4 contains five entries and these produce ten possible pairs. We applied the test for statistically significant differences to each of these pairs. This is a total of 120 tests and 11 of them resulted in a difference in intelligibility that is significant at the 95% level. All eleven significant improvements in intelligibility are associated with the two CA codec modes. This is a clear result in terms of categories of codec modes (CA and non-CA). But we naturally seek a more detailed or nuanced view of these improvements. Toward that end the eleven improvements are fully described in Tables 6 and 7. The two tables present the same information, first with text and then using symbols to aid in visual grouping.

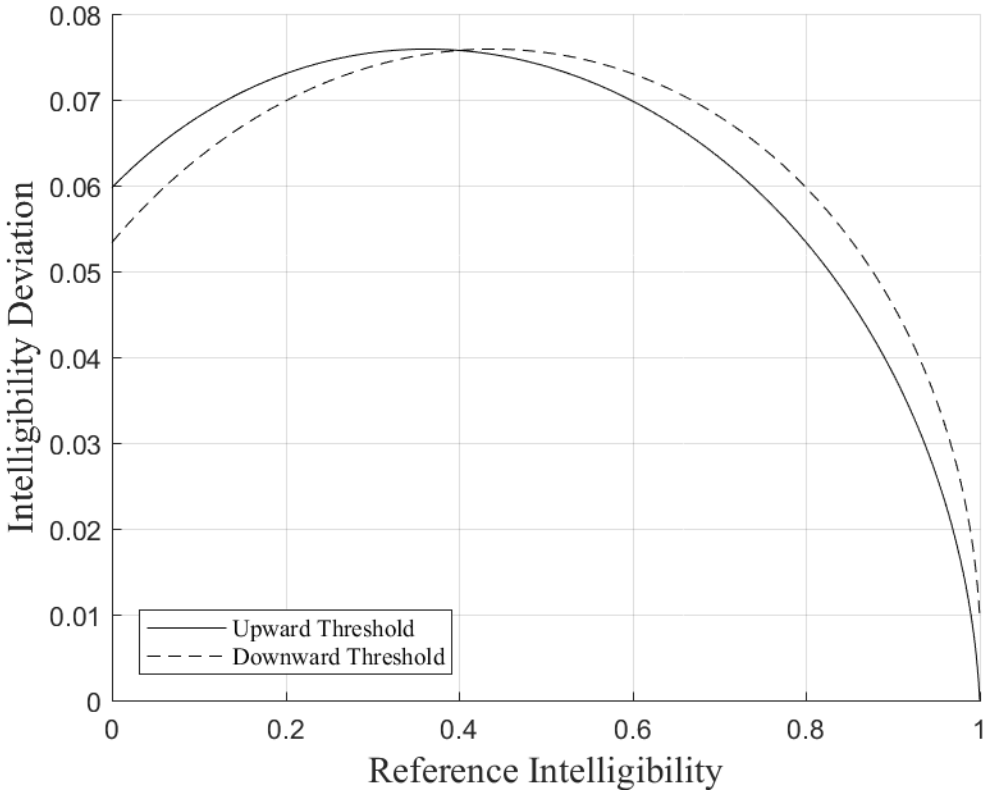

Figure 3. Thresholds for upward and downward intelligibly deviations for 95% significance when 480 MRT trials are used.

Table 6. Increases in intelligibility that are significant at the 95% level.

Number of Frames Erased	**EVS-WB CA Intelligibility is Higher Than**	**EVS-SWB CA Intelligibility is Higher Than**
0	None	None
1	None	None
2	None	None
3	None	AMR-WB
4	AMR-WB/G.718	None
5	None	None
6	None	AMR-WB AMR-WB/G.718 EVS-WB
7	AMR-WB	None
8	EVS-WB	None
9	AMR-WB AMR-WB/G.718 EVS-WB	AMR-WB/G.718
10	None	None
11	None	None

Table 7. Increases in intelligibility that are significant at the 95% level, presented with symbols to aid in visual grouping.

Number of Frames Erased	EVS-WB CA Intelligibility is Higher Than	EVS-SWB CA Intelligibility is Higher Than
0		
1		
2		
3		▲
4	●	
5		
6		▲ ● ♦
7	▲	
8	♦	
9	▲ ● ♦	●
10		
11		
Key: ▲ **AMR-WB,** ● **AMR-WB/G.718,** ♦ **EVS-WB**		

The tables show that there are six cases where EVS-WB CA has significantly higher intelligibility than another codec mode. There are five cases where EVS-SWB CA has significantly higher intelligibility than another codec mode. Thus the eleven improvements are divided as evenly as possible between the two modes and none of them show a CA mode outperforming the other CA mode. While CA modes as a group can be associated with improved intelligibility, neither individual CA mode shows an advantage over the other.

We can also look for trends with respect to the operating environments. EVS-WB CA provides a significant increase in intelligibility in four different frame-erasure conditions. EVS-SWB CA provides a significant increase in intelligibility in three different frame-erasure conditions. Only one of these (9 frames erased) is in common between the two codec modes. All of these improvements occur in the range of three to nine frames erased, inclusive. While no specific frame-erasure condition stands out, this range of conditions appears to be associated with improved intelligibility. This is consistent with our expectations. When frame erasures are minimal, there is no intelligibility issue for CA to mitigate so CA and non-CA codec modes can offer equivalent intelligibility. When frame erasures are extreme, there is a serious intelligibility issue but there is no possibility of mitigating it. So again the CA and non-CA codec modes offer equivalent intelligibility. Between these two limiting cases there is a region where CA codec modes can offer some improvement in intelligibility. In this test that range appears to be 3 to 9 frames erased, but performance is not consistent across this range.

We can also ask which codec modes the CA modes tend to improve on most often. The intelligibility of AMR-WB is significantly exceeded by a CA codec mode in four cases. The intelligibility of AMR-WB/G.718 is significantly exceeded by a CA codec mode in four cases. The intelligibility of EVS-WB is significantly exceeded by a CA codec mode in three cases.

Once again these results are as evenly distributed as possible and no specific trend can be identified.

For convenience we next summarize the results obtained so far:

- The only improvements in this test are associated with CA modes.
- Neither individual CA mode shows a consistent advantage over the other.
- The CA improvements appear in the range where 3 to 9 frames have been erased, but performance is not consistent across this range.
- There are no consistent trends regarding which non-CA codec modes are most often improved upon.

The lack of specificity regarding individual codec types in these results motivates one final analysis. We grouped the two CA code modes and the three non-CA codec modes to generate $480 \times 2 = 960$ and $480 \times 3 = 1440$ trials per group respectively. Increasing the number of trials has the benefit that smaller differences in intelligibility will become significant. The result of this analysis is shown in Figure 4.

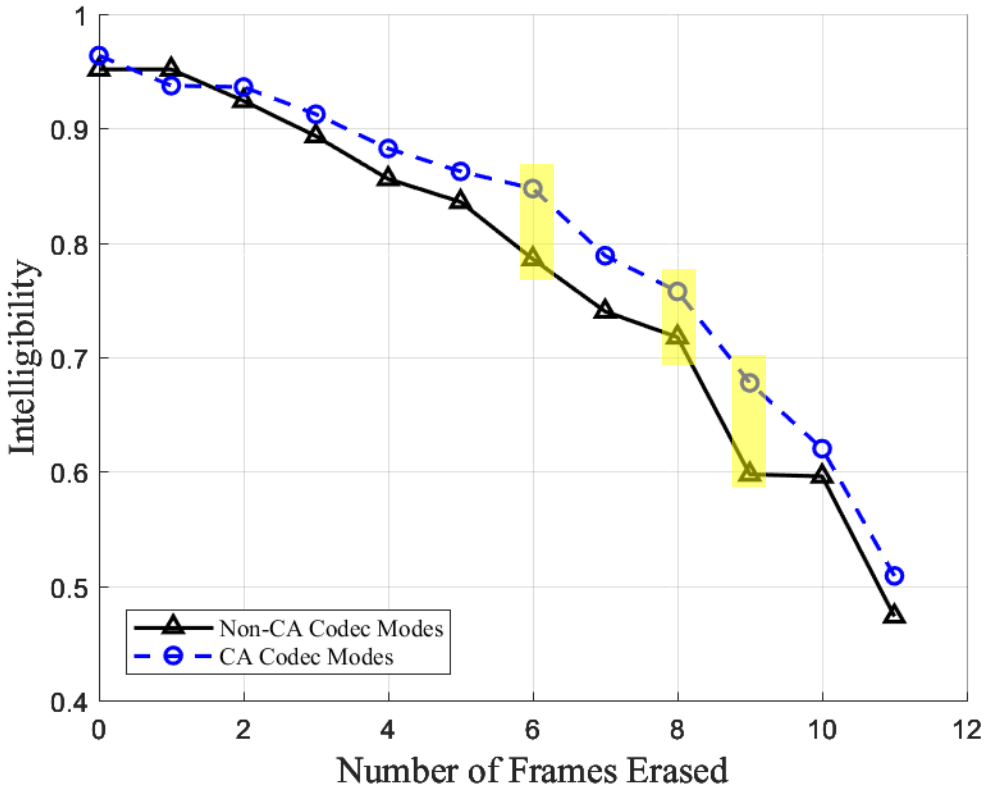

Figure 4. Intelligibility results averaged over three non-CA codec modes and two CA codec modes for twelve fame erasure conditions. Statistically significant differences are marked with yellow.

This figure confirms visually and numerically that the CA modes as a group offer an intelligibility advantage over the non-CA modes as a group. That advantage begins with the case of two frames erased and persists as the number of frames erased is increased to eleven. We can also see the additional trend that the intelligibility advantage is roughly increasing and then roughly decreasing as the number of frames erased is increased through that range. This rough trend is consistent with our previous discussion of the relationship of frame-erasure levels and the potential for CA modes to offer an intelligibility advantage.

Aggregating results into these two groups does provide more distinct numeric and visual trends. And increasing the number of trials has allowed smaller differences in intelligibility to become significant. But Figure 2 makes it clear that the original data was not clearly clustered with respect to CA vs non-CA codec modes. Because of this, aggregating the data does not produce differences that are consistently statistically significant. In fact, when aggregated into these two groups, the CA modes show significant improvement over the non-CA modes for just three operating environments: 6, 8 and 9 frames erased. These cases are marked in yellow in Figure 4. They fall inside the range previously identified (3 to 9 frames erased) but this analysis does not show consistent improvements across that range. In short this analysis is consistent with the previous but it does not refine the range, nor does it provide stronger evidence for the range.

In summary, Test 1 shows that CA modes show intelligibility improvements over non-CA modes in numerous cases when the number of erased frames is in the range of 3 to 9, inclusive. Beyond that, we cannot observe any other clear consistent trends.

3. INTELLIGIBILITY TEST 2 — RANDOM FRAME ERASURES AND BACKGROUND NOISE

In Test 2 we move from highly controlled and targeted frame-erasure patterns to randomly occurring frame-erasure patterns. This section presents each of the key factors involved in Test 2, then presents and discusses the results produced by Test 2.

The approach of Test 1was adopted to create maximum-sensitivity in order to reveal small differences to the extent possible. The approach of Test 2 is more representative of how frame erasures occur in actual operating environments. The results of Test 1 allow us to set parameters in Test 2 to identify the region where differences between the codec modes can be observed. Test 1 also informs the number of trials in Test 2. From Test 1 we observed that 480 trials and even two or three times 480 trials is a marginal number in terms of revealing differences in this situation. Thus Test 2 uses 2160 trials for each condition.

Test 2 uses the same codec modes as Test 1 (see Table 1) with the sole exception of the EVS software version. Because a new version was released, and because we wished to use the newest available version, we switched from version 13.3.0 to version 13.4.0.

Our adoption of 2160 trials per condition was the result of considering numerous practical factors in the structure of the MRT protocol and our implementation of the CMRT. We achieved the 2160 trials by using 45 groups of 6 words from the MRT speech recordings. Using recordings from 2 female and 2 male talkers yields a total of $45 \times 6 \times 4 = 1080$ MRT recordings. We used each to produce two trials, resulting in a total of $1080 \times 2 = 2160$ trials.

In Test 2 we again added coffee shop noise to the MRT speech recordings to simulate background noise at the transmit location. In addition to the 20 dB SNR case, we added the 15, 10, 5, 0, and -5 dB SNR cases. As in Test 1 we used a unique random excerpt with an appropriate length (1.8 seconds on average) taken from the full noise recording (328 seconds) for each of the 1080 MRT recordings. We kept the SNR factor independent of the frame-erasure factor as described below. Our goal was to separately quantify these two factors that can reduce intelligibility.

3.1 Frame Erasures and Operating Environments

In this test we use the two-state Gauss-Markov model shown in Figure 5 to produce random frame-erasure patterns [20]. We record the sequence of states that the model passes through and this record becomes the frame-erasure pattern. A zero value in the pattern means the frame is unaltered and a value of one means that the frame is erased. The transition probabilities p and q allow us to control the frame-erasure rate (FER) and the mean erasure length (MEL). To prevent either state from permanently capturing the model, we require $0 < p \leq 1$ and $0 \leq q < 1$.

FER is simply the ratio of erased frames to total frames in given frame-erasure pattern and we report this as a percentage. To calculate MEL we locate all runs of frame erasures in the pattern. A run may have a length of one frame (an isolated frame erasure), two frames, three frames, etc. If a pattern has N runs and their lengths are $L_1, L_2, \ldots, L_N$, then MEL is the mean of those N lengths.

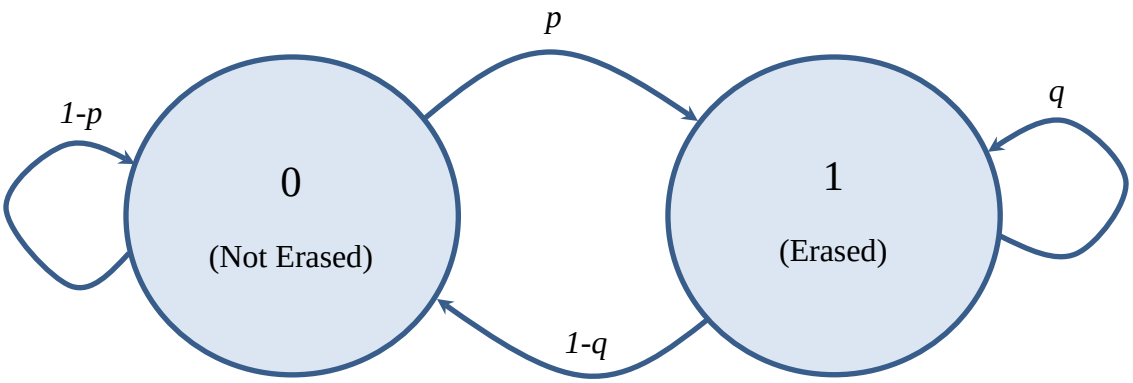

Figure 5. Two-state Gauss-Markov model for generating frame-erasure patterns.

Because it has two free parameters, we can configure the model to produce the MEL and FER values we require. Given a target value of MEL the required transition probability q can be found by

$$q = 1 - \frac{1}{\text{MEL}}, 1 < \text{MEL}. \tag{6}$$

Then q and the target FER produce the transition probability p according to

$$p = \text{FER} \times \frac{1 - q}{1 - \text{FER}}, 0 < \text{FER} < 1. \tag{7}$$

Each of these follows from the results in [20]. Note that when $p = q$, the probability that frame *n* is erased is independent of the fate of frame *n-1*. This is the case of independent frame erasures.

Recall that in Test 1 we deterministically controlled the frame-erasure length and found that frame-erasure patterns from three to nine frames long formed a region of interest. In Test 2 we set parameters in the Gauss-Markov model to insure that the *average* frame-erasure lengths (the MEL values) align with this region of interest. After consideration of our larger test design, and the multiple trade-offs therein, we elected to use MEL values of 4, 5, 6, 7, 8, and 9 frames paired with FER values 5, 10, 15, 20, 25, and 30%.

Consistent with the example set in [21] we generated frame-erasure patterns for 40 ms radio transport frames each of which would contain two speech codec data frames. This means the actual target MEL values were 2, 2.5, 3, 3.5, 4, and 4.5 radio transport frames.

MRT source files have an average duration near 1.8 seconds which is 90 speech codec data frames or 45 radio transport frames. This means that the available FER and MEL values for any single MRT source file are coarsely quantized. Consider the case where 50 frames are available. If the target FER is 5% the closest achievable FERs are 4% (2 erased frames) and 6% (3 erased frames). If two frames are erased then the only possible erasure lengths are 1 and 2 and the only achievable MEL values are 1 and 2. If three frames are erased then the only possible erasure lengths are 1, 2, and 3 and the only achievable MEL values are 1, 1.5, and 3. Because of these effects the p and q values produced by (6) and (7) may need adjustments in order to achieve

some target values of FER and MEL in the case of short (e.g., 45 frames) patterns. The relations given in (6) and (7) do become exact as patterns grow long.

In order to obtain results that are as close to the target FER and MEL as possible, we experimentally tuned the p and q values and we also iterated until a random pattern produced the achievable FER that is closest to the target FER value. That is, when N frames were available we iterated until N_T frames were erased where N_T is given by:

$$N_T = \text{Round}\left(\frac{\text{FER}}{100} \times N\right). \tag{8}$$

Table 8 shows the achieved FER and MEL values for each of the six frame-erasure environments, averaged over all 1080 of the frame-erasure patterns.

Table 8. Properties of twelve operating environments. Each environment is applied to each of the five codec modes to produce a total of 60 conditions.

Environment Name	**Noise**	**Frame Erasures**		
		Mean FER	Mean MEL[6] (speech codec frames)	Mean MEL (ms)
0% FER	Coffee shop noise, 20 dB SNR	0%	—	—
5% FER	Coffee shop noise, 20 dB SNR	4.9%	4.0	80
10% FER	Coffee shop noise, 20 dB SNR	10.0%	5.0	100
15% FER	Coffee shop noise, 20 dB SNR	15.1%	6.0	120
20% FER	Coffee shop noise, 20 dB SNR	20.0%	7.0	140
25% FER	Coffee shop noise, 20 dB SNR	25.3%	8.0	160
30% FER	Coffee shop noise, 20 dB SNR	30.1%	9.0	180
15 dB SNR	Coffee shop noise, 15 dB SNR	0%	—	—
10 dB SNR	Coffee shop noise, 10 dB SNR	0%	—	—
5 dB SNR	Coffee shop noise, 5 dB SNR	0%	—	—
0 dB SNR	Coffee shop noise, 0 dB SNR	0%	—	—
-5 dB SNR	Coffee shop noise, -5 dB SNR	0%	—	—

[6] MEL is mean erasure length and that mean is taken over all erasures within a single MRT source. Mean MEL indicates the average of these MEL values calculated across all 1080 MRT sources.

Given the statistical nature of these patterns, we also document here the range of variation encountered. Table 9 expands on Table 8 by giving the measured minimum, maximum, median, and mean values of FER for each environment across the 1080 MRT recordings.

Table 9. Measured FER statistics across 1080 MRT recordings.

Environment Name	Minimum	Maximum	Median	Mean
5% FER	0.04	0.06	0.048	0.049
10% FER	0.09	0.11	0.100	0.100
15% FER	0.14	0.16	0.151	0.151
20% FER	0.19	0.21	0.200	0.200
25% FER	0.24	0.26	0.255	0.253
30% FER	0.29	0.31	0.300	0.301

Similarly, Table 10 provides the measured minimum, maximum, median, and mean values of MEL for each environment across the 1080 MRT source files. The final column shows twice the mean MEL which is the mean MEL in terms of speech codec frames.

Table 10. Measured MEL statistics across 1080 MRT recordings.

Environment Name	Minimum	Maximum	Median	Mean	2 × Mean
5% FER	1.0000	3	2.0	2.0	4.0
10% FER	1.0000	6	2.0	2.5	5.0
15% FER	1.1667	8	2.7	3.0	6.0
20% FER	1.3333	11	3.0	3.5	7.0
25% FER	1.5714	14	3.7	4.0	8.0
30% FER	1.8571	16	4.0	4.5	9.0

Table 8 shows the twelve operating environments used in this test. The first seven environments manipulate the frame-erasure scenario while background noise remains at a favorable 20 dB SNR (as in Test 1). The remaining five conditions manipulate the SNR while the frame-erasure scenario remains constant — no frame erasures. Thus we have a set of worsening frame-erasure conditions with noise remaining favorable and we have a set of worsening noise conditions with frame erasures remaining favorable.

3.2 Modified Rhyme Test Execution

We tested each of the 60 conditions in this test by means of 2160 MRT trials (two repetitions of 1080 distinct trials). We again employed CMRT [16]. We packaged the trials into tasks that contained 60 trials, one from each condition. This produced 1080 tasks. This packaging used constrained randomization, and the trials were unlabeled and presented in a random order. The estimated maximum time to complete a task was 5 minutes.

We then grouped these tasks into 3 groups of 360 and each was assigned to willing listeners twice, creating an effective total of 6 batches. These six batches were made available in sequence

at 10:00 AM and 1:00 PM MDT on August 16, 2017, 10:00 AM and 1:00 PM MDT on August 17, 10:00 AM on August 22, and 1:00 PM MDT on August 23. On average the work in each batch was claimed in about 80 minutes, and results returned shortly thereafter.

We required two distinct workers to complete each task. Thus the total number of raw MRT trials collected was $60 \times 360 \times 6 \times 2 = 259{,}200$. The CMRT data processing stage reduced these to 129,600 trials, which is indeed 2160 trials for each of 60 conditions. Of the 259,200 raw trials conducted, only 45 (0.02%) produced invalid results (no word was selected). On average, 300 distinct listeners participated in each group of 360 tasks. Conversely $2 \times 360/300 = 2.4$ is the average number of tasks performed by each listener.

3.3 Results

The final CMRT results consist of 2160 trials for each of the 60 conditions. Each trial produces either success or failure. The numbers of successes, the success rates, and the corresponding intelligibility values found via (3), are tabulated in Table 11.

Table 11. Results of Test 2 (2160 trials per condition).

Codec Mode	Environment	Number of Successes	Success Rate	Intelligibility
AMR-WB	0% FER	2049	0.9486	0.9383
	5% FER	2024	0.9370	0.9244
	10% FER	1960	0.9074	0.8889
	15% FER	1886	0.8731	0.8478
	20% FER	1809	0.8375	0.8050
	25% FER	1712	0.7926	0.7511
	30% FER	1636	0.7574	0.7089
	15 dB SNR	2008	0.9296	0.9156
	10 dB SNR	1943	0.8995	0.8794
	5 dB SNR	1758	0.8139	0.7767
	0 dB SNR	1469	0.6801	0.6161
	-5 dB SNR	1028	0.4759	0.3711
AMR-WB/G.718	0% FER	2057	0.9523	0.9428
	5% FER	2002	0.9269	0.9122
	10% FER	1979	0.9162	0.8994
	15% FER	1896	0.8778	0.8533
	20% FER	1871	0.8662	0.8394
	25% FER	1732	0.8019	0.7622
	30% FER	1641	0.7597	0.7117
	15 dB SNR	2014	0.9324	0.9189
	10 dB SNR	1953	0.9042	0.8850
	5 dB SNR	1767	0.8181	0.7817
	0 dB SNR	1458	0.6750	0.6100
	-5 dB SNR	990	0.4583	0.3500

Codec Mode	Environment	Number of Successes	Success Rate	Intelligibility
EVS-WB	0% FER	2054	0.9509	0.9411
	5% FER	2013	0.9319	0.9183
	10% FER	1981	0.9171	0.9006
	15% FER	1859	0.8606	0.8328
	20% FER	1830	0.8472	0.8167
	25% FER	1733	0.8023	0.7628
	30% FER	1611	0.7458	0.6950
	15 dB SNR	2034	0.9417	0.9300
	10 dB SNR	1956	0.9056	0.8867
	5 dB SNR	1763	0.8162	0.7794
	0 dB SNR	1453	0.6727	0.6072
	-5 dB SNR	1000	0.4630	0.3556
EVS-WB CA	0% FER	2050	0.9491	0.9389
	5% FER	2038	0.9435	0.9322
	10% FER	2013	0.9319	0.9183
	15% FER	1926	0.8917	0.8700
	20% FER	1886	0.8731	0.8478
	25% FER	1767	0.8181	0.7817
	30% FER	1712	0.7926	0.7511
	15 dB SNR	1999	0.9255	0.9106
	10 dB SNR	1905	0.8819	0.8583
	5 dB SNR	1764	0.8167	0.7800
	0 dB SNR	1454	0.6731	0.6078
	-5 dB SNR	1011	0.4681	0.3617
EVS-SWB CA	0% FER	2048	0.9481	0.9378
	5% FER	2041	0.9449	0.9339
	10% FER	2010	0.9306	0.9167
	15% FER	1928	0.8926	0.8711
	20% FER	1883	0.8718	0.8461
	25% FER	1770	0.8194	0.7833
	30% FER	1700	0.7870	0.7444
	15 dB SNR	2019	0.9347	0.9217
	10 dB SNR	1946	0.9009	0.8811
	5 dB SNR	1773	0.8208	0.7850
	0 dB SNR	1498	0.6935	0.6322
	-5 dB SNR	1065	0.4931	0.3917

We have used the exact same MRT source recordings, background-noise excerpts, frame-erasure patterns, and number of trials for each of the five codec modes. This allows us to directly compare results across codec modes at each operating environment. To facilitate this comparison Table 12 shows the intelligibly results of Table 11 reorganized with one codec mode in each column.

Table 12. Test 2 intelligibility results for five codec modes.

Environment	AMR-WB	AMR-WB/ G.718	EVS-WB	EVS-WB CA	EVS-SWB CA
0% FER	0.9383	0.9428	0.9411	0.9389	0.9378
5% FER	0.9244	0.9122	0.9183	0.9322	0.9339
10% FER	0.8889	0.8994	0.9006	0.9183	0.9167
15% FER	0.8478	0.8533	0.8328	0.8700	0.8711
20% FER	0.8050	0.8394	0.8167	0.8478	0.8461
25% FER	0.7511	0.7622	0.7628	0.7817	0.7833
30% FER	0.7089	0.7117	0.6950	0.7511	0.7444
15 dB SNR	0.9156	0.9189	0.9300	0.9106	0.9217
10 dB SNR	0.8794	0.8850	0.8867	0.8583	0.8811
5 dB SNR	0.7767	0.7817	0.7794	0.7800	0.7850
0 dB SNR	0.6161	0.6100	0.6072	0.6078	0.6322
-5 dB SNR	0.3711	0.3500	0.3556	0.3617	0.3917

3.3.1 Frame-Erasure Results

The first seven rows of Table 12 address the frame-erasure environments. These results are also shown graphically in Figure 6.

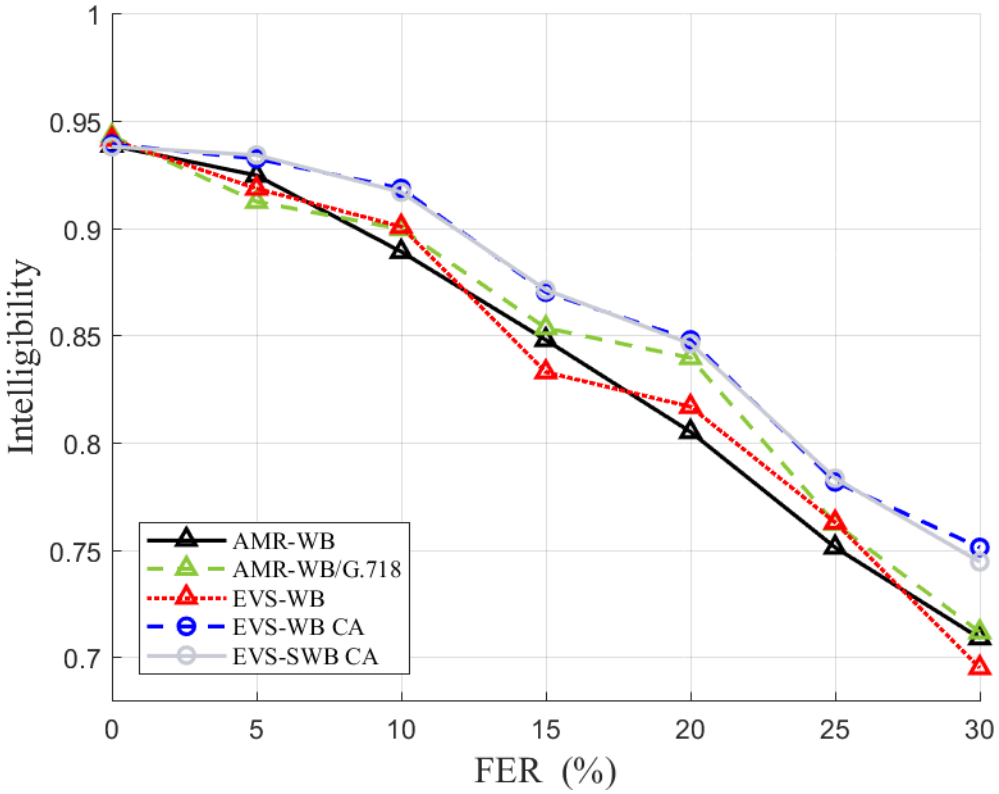

Figure 6. Intelligibility versus FER for five codec modes.

The figure shows the expected trend of decreasing intelligibility as FER (and MEL) are increased. Intelligibility is very high (0.94) for the case of no erased frames and it drops monotonically as the FER increases ending with the range of 0.70-0.75 when FER = 30% (and MEL = 9 frames).

In Test 1 we targeted erasures at critical consonants and in Test 2 frame-erasure locations are unconstrained. We can compare the Test 1 environments where 4, 5, 6, 7, 8, and 9 consecutive frames are erased with the Test 2 environments where the MEL is 4, 5, 6, 7, 8, and 9 frames respectively. We find that the Test 1 environments produce intelligibility results that range from 0.03 to 0.09 lower than those of Test 2. Targeting critical consonants did in fact produce greater intelligibly reductions as intended.

Figure 6 shows that as FER increases the two CA codec modes begin to favorably differentiate themselves from all three of the non-CA modes in a fairly consistent way, with the single exception being the case of AMR-WB/G.718 at 20% frame-erasure rate.

Next we move to testing for statistically significant differences. Test 2 has 2160 trials per condition compared with the 480 trials per condition used in Test 1. Figure 7 shows the resulting improved resolution curves for Test 2, along with the resolution curves of Test 1 for reference purposes. Over much of the intelligibility range the resolution is improved by a factor that is near 2.1 which is $\sqrt{2160/480}$ and is consistent with the mathematics of this situation.

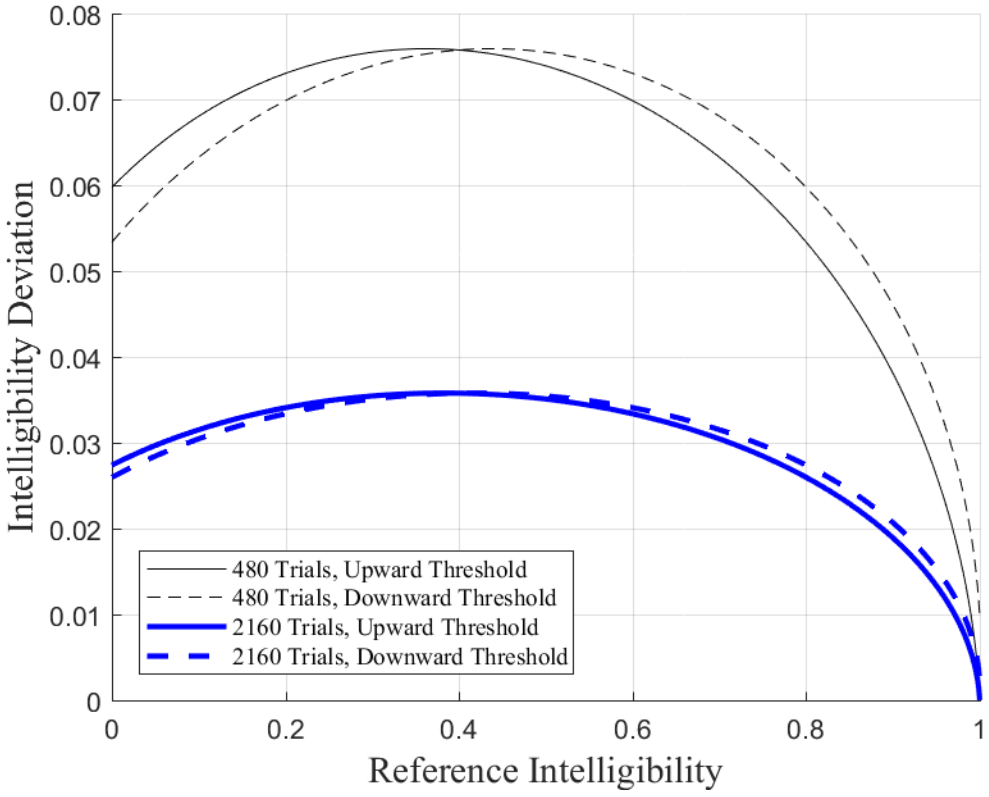

Figure 7. Thresholds for upward and downward intelligibly deviations for 95% significance when 480 and 2160 MRT trials are used.

Each row of Table 12 contains five entries and these produce ten possible pairs. We applied the test for statistically significant differences (the same as described for Test 1 in Section 2.7) to each of these pairs. When we consider the 7 operating environments associated with frame erasures this produces a total of 70 tests and 20 of them reported a difference in intelligibility that is significant at the 95% level. Nineteen of these twenty improvements are associated with CA modes. These improvements are fully described in Tables 13 and 14. The two tables present the same information, first using text and then using symbols to aid in visual grouping.

Table 13. Increases in intelligibility that are significant at the 95% level, frame-erasure environments only.

Environment	**AMR-WB is Higher Than**	**AMR-WB/G.718 is Higher Than**	**EVS-WB is Higher Than**	**EVS-WB CA is Higher Than**	**EVS-SWB CA is Higher Than**
0 % FER	None	None	None	None	None
5 % FER	None	None	None	AMR-WB/G.718	AMR-WB/G.718
10 % FER	None	None	None	AMR-WB	AMR-WB
15 % FER	None	None	None	EVS-WB	AMR-WB EVS-WB
20 % FER	None	AMR-WB	None	AMR-WB EVS-WB	AMR-WB EVS-WB
25 % FER	None	None	None	AMR-WB	AMR-WB
30 % FER	None	None	None	AMR-WB AMR-WB/G.718 EVS-WB	AMR-WB AMR-WB/G.718 EVS-WB

Table 14. Increases in intelligibility that are significant at the 95% level, frame-erasure environments only, presented with symbols to aid in visual grouping.

Environment	**AMR-WB is Higher Than**	**AMR-WB/G.718 is Higher Than**	**EVS-WB is Higher Than**	**EVS-WB CA is Higher Than**	**EVS-SWB CA is Higher Than**
0 % FER					
5 % FER				●	●
10 % FER				▲	▲
15 % FER				♦	▲ ♦
20 % FER		▲		▲ ♦	▲ ♦
25 % FER				▲	▲
30 % FER				▲ ● ♦	▲ ● ♦
Key: ▲ **AMR-WB,** ● **AMR-WB/G.718,** ♦ **EVS-WB**					

As expected, the five codec modes show no significant differences in intelligibility when no frames are erased. But for every one of the environments with erased frames, both CA modes offer a significant intelligibly improvement over one or more of the non-CA modes.

EVS-WB CA provides significant intelligibly improvements in nine different comparisons and EVS-SWB CA does so for ten different comparisons. In terms of intelligibility, CA modes outperform AMR-WB in nine cases, EVS-WB in six cases, and AMR-WB/G.718 in just four

cases. The improvement over AMR-WB and EVS-WB is somewhat consistent from the 10% FER environment up through the 30% FER environment. A counter-intuitive result is that the improvement over AMR-WB/G.718 occurs only at the extreme ends of the range — 5% and 30% FER. We consider AMR-WB/G.718 to be the second most robust option after the CA modes. AMR-WB/G.718 is less often improved upon by CA, and AMR-WB/G.718 also shows an improvement over AMR-WB in the 20% FER case.

So far we have reported differences in intelligibility at specified FER values. The perspective of robustness leads us to invert this discussion and consider how the choice of a more robust codec mode can allow higher FER while providing fixed intelligibility. In other words we can report FER increases tolerated, rather that intelligibility gains at a specified FER.

Figure 8 provides a visual example of one approach. Because the two CA modes show nearly identical intelligibility results (see Figure 6), we have averaged them to produce the results shown in blue in Figure 8. To compare this mean CA result with AMR-WB, we show the AMR-WB results in black. The intelligibility produced by AMR-WB at 2% FER is the same as that of CA at 5.1 % FER, as shown by the red double-arrow. But 2% FER is not of much importance when intelligibility is considered because the intelligibility is nearly perfect at that point. In addition, the low slopes in this area make such comparisons rather sensitive to minor intelligibility variations.

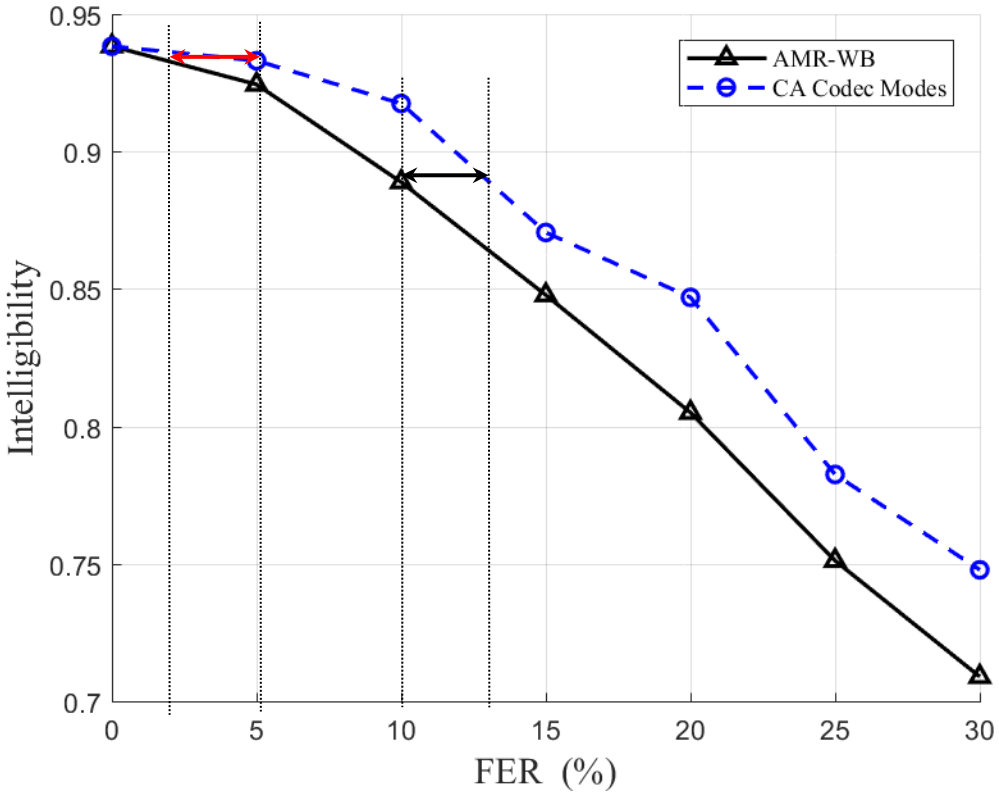

Figure 8. Intelligibility versus FER for AMR-WB and two CA modes averaged. Red arrow shows that AMR-WB 2% FER intelligibility matches CA 5.1% intelligibility. Black arrow shows that AMR-WB 10% FER intelligibility matches CA 12.9% intelligibility.

Thus we turn to the intelligibility results at FERs of 5, 10, 15, 20, and 25% and equate them to the CA intelligibility results. The black double arrow in Figure 8 provides an example at 10% FER. The five measured FER increases range from 2.9% to 5.1%, they show no particular trend, and the mean value of these FER increases is 3.8%. We performed analogous analyses using AMR-WB/G.718 and EVS-WB as references. The results are summarized in Table 15.

Table 15. Average FER increase tolerated by CA relative to non-CA.

	Reference Codec Mode		
	AMR-WB	AMR-WB/G.718	EVS-WB
FER with CA codec modes that gives same intelligibility as 2% FER with reference codec mode	5.1%	5.9%	5.6%
FER increase for CA modes (row 1 minus 2%)	3.1%	3.9%	3.6%
Mean FER increase for CA codec modes[7]	3.8%	2.9%	3.6%

The analysis described above uses linear interpolation between data points. An alternative approach would be to fit individual smooth curves to each set of six data points, and then proceed to equate intelligibility values and find FER increases at arbitrary FER values. The two approaches have their advantages and disadvantages, and they cannot produce dramatically different results.

The rankings for the mean FER increases relative to AMR-WB, AMR-WB/G.718, and EVS-WB are consistent with conclusions we can draw from Table 13. That is, the CA codec modes improve upon AMR-WB the most (3.8%). They improve upon AMR-WB/G.718 the least (2.9%). And the improvement over EVS-WB (3.6%) falls between these two. Note that this analysis approach is somewhat simplified in that it is not based on statistically significant differences. Instead the analysis simply equates the measured intelligibility values without accounting for any uncertainty in those values.

We can compare the speech *intelligibility measurement* results in Table 15 with a related analysis given in [21]. That analysis is based on speech *quality estimates* generated by the POLQA algorithm [22]. Based on those quality estimates the analysis concludes that when the AMR-WB codec mode encounters a 2% FER it produces the same estimated speech quality as the EVS-WB CA codec mode with 8% FER. This is an FER increase of 6%, *based on constant estimated speech quality*. This can be compared with our result in row two of Table 15 above. We found an FER increase of 3.1% for the CA modes relative to AMR-WB, *based on constant measured speech intelligibility*.

"Speech quality" and "speech intelligibility" are different things. "Speech quality" describes how pleasant or unpleasant speech sounds. The POLQA algorithm produces results along a speech quality continuum that uses descriptors "excellent," "good," "fair," "poor," and "bad." "Speech intelligibility" is a measure of how well or poorly a speech signal transfers information. Speech intelligibility results tell what fraction of the information has been successfully transferred. It may well be that when speech quality is "poor" or "bad" some information is not transferred. But

[7] Calculated using intelligibility values for the reference codec mode at FER=5, 10, 15, 20, and 25%. Figure 8 shows an example at 10% FER when AMR-WB is the reference codec mode.

speech quality does not measure this information loss, instead it measures how pleasing the speech sounds are.

Speech quality is indeed important and it can influence the ability to efficiently and correctly complete tasks. If instructions for performing a task arrive with lower speech quality (yet are still fully intelligible), that lower speech quality can increase listening effort or the "cognitive loading" associated with listening to and understanding the instructions. This increased cognitive loading can in turn reduce the ability to efficiently and correctly complete the task. But this multi-step relationship between speech quality and task performance is rather indirect and the connection can be weak.

The link between speech intelligibility and task performance is much more direct. If a telecommunication system causes information to be missing from the received instructions, the task performance will be directly impaired.

Speech quality and intelligibly are also very different from the perspective of a robust speech decoder. When frames of data are erased the robust decoder must still produce some output waveform. To maintain speech quality, the decoder must simply produce a waveform that sounds pleasing to the ear. But to maintain speech intelligibility, the decoder must produce a waveform that contains the correct information. When extended erasures occur, pleasing the ear (preserving quality) is not easy, but producing the proper information (preserving intelligibility) is even harder.

Because many of our public safety stakeholders regularly perform critical tasks based on instructions communicated by radio, we argue that speech intelligibility is the important parameter to measure. If intelligibility is high and steady for all systems of interest, then speech quality can be used to further differentiate between them. From Table 15 we see that on average the CA codec modes can hold speech intelligibility constant when FER increases 2.9 to 3.8%.

Note that this improved robustness does come with an increase in end-to-end delay. Recall that the results we have reported here were obtained with a CA FEC offset value of three frames. This means that when a frame erasure occurs, the CA decoder must have access to three frames (60 ms duration) that follow the current frame.

3.3.2 Noise Results

The final five rows of Table 12, along with the first row, provided results for the operating environments associated with noise. Those results are also shown in Figure 9.

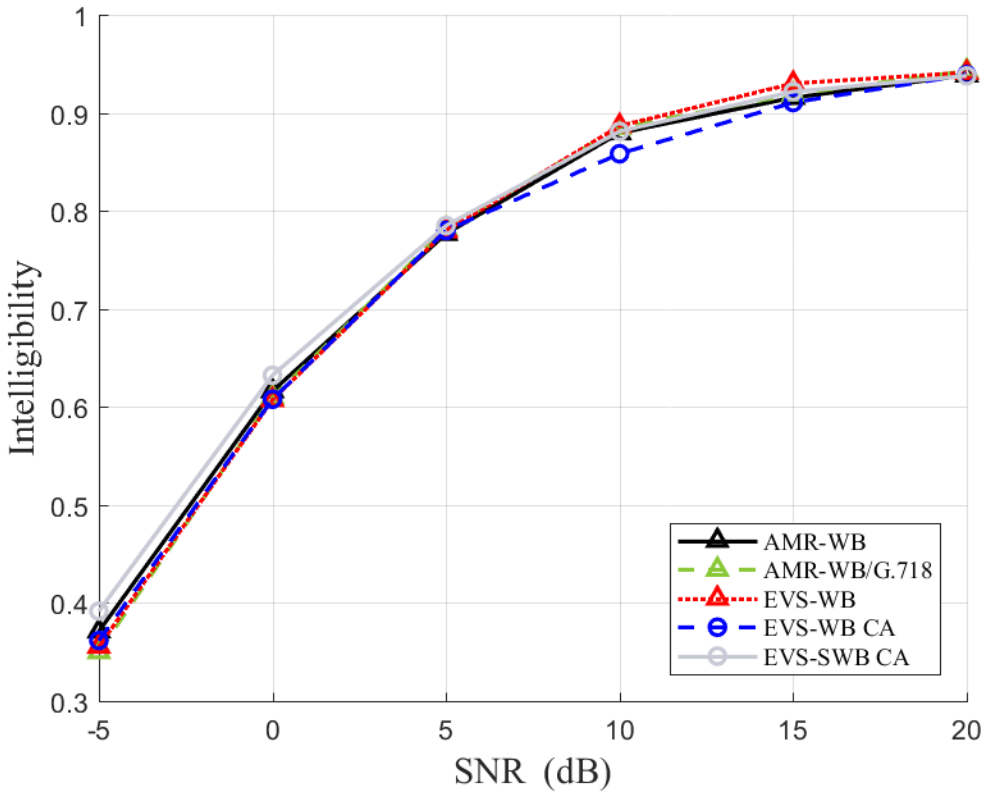

Figure 9. Intelligibility versus SNR for five codec modes.

As expected, the intelligibility falls as the SNR decreases. More specifically intelligibility drops from 0.94 to the range 0.35-0.39 as SNR is decreased from 20 to -5 dB. It is clear that the range of noise in this test causes much greater variation in intelligibility than the range of frame erasures does. At each SNR level we tested all possible pairs of codec modes to find differences that are significant at the 95% level. Tables 16 and 17 show all six significant improvements, first using text and then using symbols to aid in visual grouping.

Table 16. Increases in intelligibility that are significant at the 95% level, noise environments only.

Environment Name	**AMR-WB is Higher Than**	**AMR-WB/G.718 is Higher Than**	**EVS-WB is Higher Than**	**EVS-WB CA is Higher Than**	**EVS-SWB CA is Higher Than**
20 dB SNR	None	None	None	None	None
15 dB SNR	None	None	EVS-WB CA	None	None
10 dB SNR	None	EVS-WB CA	EVS-WB CA	None	EVS-WB CA
5 dB SNR	None	None	None	None	None
0 dB SNR	None	None	None	None	None
-5 dB SNR	None	None	None	None	AMR-WB/G.718 EVS-WB

Table 17. Increases in intelligibility that are significant at the 95% level, noise environments only, presented with symbols to aid in visual grouping.

Environment Name	AMR-WB is Higher Than	AMR-WB/G.718 is Higher Than	EVS-WB is Higher Than	EVS-WB CA is Higher Than	EVS-SWB CA is Higher Than
20 dB SNR					
15 dB SNR			■		
10 dB SNR		■	■		■
5 dB SNR					
0 dB SNR					
-5 dB SNR					● ♦
Key: ● **AMR-WB/G.718,** ♦ **EVS-WB,** ■ **EVS-WB CA**					

All significant differences are associated with CA codec modes. The tables show that EVS-WB CA has significantly *lower* intelligibility than another codec mode in four cases. On the other hand EVS-SWB CA has significantly *higher* intelligibility than some other codec mode in three cases.

Recall that the CA modes do not increase the bit rate but rather they selectively enforce a minor reduction in the bits available for the primary coding so that some bits can be available for the redundant coding. It may be that this minor rate reduction has reduced EVS-WB CA intelligibly in some noise environments. It also appears that this disadvantage is more than compensated by the additional audio bandwidth (nominally the band from 7 to 16 kHz) of EVS-SWB CA, especially at the lowest SNR value.

Note that CMRT implementation provides realistic results in the sense that a wide variety of listeners are using a range of commonly available speakers, headphones, and earbuds in a range of listening environments. These are real-world conditions not laboratory conditions and we consider them to be highly relevant and representative. It is possible however, that one might be able to configure a laboratory test that shows greater differences between WB and SWB under highly-tuned conditions. Recall that SWB extends WB by adding the band from 7 and 16 kHz. Thus one might select speakers or headsets for their high-frequency response and select listeners for their sensitivity to these frequencies. But no matter what measures are taken to insure the reproduction and reception of this band, its positive or negative effect on intelligibility will always hinge on the relative contributions of the speech and the background noise in that band.

3.3.3 General Results

The results of Test 2 afford the opportunity to equate FER and SNR in terms of the intelligibility they produce. More specifically we can do so separately for the non-CA codec modes and the CA codec modes. For the noise environments we averaged over all five codec modes to produce a single reference relationship between SNR and intelligibility. For the frame-erasure environments we averaged over the three non-CA codec modes and the two CA codec modes separately to produce two relationships between FER and intelligibility. For each of these, we then equated intelligibility in order to relate FER to SNR.

These relationships are far from general. They are specific to the coffee shop noise, the frame-erasure pattern parameters, and the MRT paradigm for intelligibility measurement. Nonetheless in Table 18 we report the SNRs that produce equivalent intelligibly for each of the FER values. The process is similar to that shown in Figure 8 and we again use linear interpolation. The results in the table follow the expected trend and they provide another view of the CA intelligibility advantage. That is, across the FERs considered, the CA intelligibly advantage is akin to improving the input audio SNR at the transmitting location by 1 to 3 dB.

Table 18. Equivalences between FER and SNR based on intelligibility.

FER	Equivalent SNR for non-CA codec modes	Equivalent SNR for CA codec modes
5%	15 dB	18 dB
10%	12 dB	15 dB
15%	8 dB	9 dB
20%	7 dB	8 dB
25%	4 dB	5 dB
30%	3 dB	4 dB

Our second general result pertains to test repeatability and sources of variance. There are five conditions that are similar between Test 1 and Test 2. These are the 20 dB SNR, zero frame-erasure environments with the five codec modes. Test 1 used 480 of the 1200 available MRT sources and Test 2 used 1080 of the 1200. Analysis shows that the two tests have 460 MRT sources in common. This is 96% of the sources used in Test 1 but only 43% of the sources used in Test 2. They use the same noise signal (328 second duration) but different randomly selected portions (average length 1.8 seconds) were paired with each MRT source. The interaction between the exact content of the non-stationary noise signal and the critical consonant can be critical. One might argue that the effects could average out and that the two tests could give the same result. Finally, note that the EVS codec modes use different software versions between the two tests.

Tables 4 and 12 provide the results for these five similar conditions. In spite of differences listed in the paragraph above, Tests 1 and 2 produced statistically equivalent results (at the 95% level) for four of the five codec modes. The exception is EVS-SWB CA where Test 1 produced an intelligibility score that is 0.030 greater than the result produced by Test 2. This difference is significant at the 95% level, but it is not significant at the 95.1% level. Thus it might be described as borderline significant at the 95% level. From this we conclude that the sources of variation listed produce insignificant or barely significant differences in intelligibility.

We now summarize the key results reported in this section:

- The CA codec modes provide small but statistically significant intelligibly improvements in multiple cases across the frame-erasure environments considered in this test.

- The CA codec modes show statistically significant intelligibly improvements over AMR-WB in nine of the twelve non-zero FER comparisons.

- The CA codec modes show statistically significant intelligibly improvements over AMR-WB/G.718 in just four of the twelve non-zero FER comparisons.

- The CA codec modes show statistically significant intelligibly improvements over EVS-WB in six of the twelve non-zero FER comparisons.

- The CA codec modes tolerate FER increases ranging from 2.9% to 3.8% (over the FER for non-CA modes) while maintaining the same intelligibility as the non-CA modes.

- EVS-WB CA has statistically significantly lower intelligibility than another codec mode in 4 of the 24 noise environment comparisons.

- EVS-SWB CA has statistically significantly higher intelligibility than another codec mode in 3 of the 24 noise environment comparisons.

4. SUMMARY

We have performed two distinct but related speech intelligibility tests on five speech codec operating modes. We followed the MRT paradigm and crowdsourced the tests by offering balanced groups of MRT trials as micro-work. This approach allowed us to efficiently collect 316,800 raw MRT trials which then become 158,400 final MRT trials.

The tests included deterministic targeted frame erasures of various lengths, random frame erasures of various average lengths, and background noise applied at various levels. In all, 24 different operating environments were applied to the 5 different codec modes for a total of 120 different conditions.

We performed statistical tests on the MRT results to find codec modes that have statistically significantly higher intelligibility than other codec modes. In many frame-erasure conditions we found that the Channel Aware (CA) codec modes offer small but statistically significant intelligibility advantages. This is expected because these modes selectively apply redundant coding to provide higher robustness to erased frames. This redundant coding does not increase the bit rate but it does increase the end-to-end delay.

The greatest number of significant increases is found when comparing CA modes to AMR-WB. The second greatest is the case where CA modes are compared to EVS-WB. Comparing CA modes to AMR-WB/G.718 gives the smallest number of significant increases. In addition to comparing intelligibility at fixed FER values, we also interpolate in order to compare FER values at fixed intelligibility levels. For a fixed reference intelligibility set by AMR-WB, the CA modes can attain that same intelligibility at FERs that are, on average, increased by 3.8%. For EVS-WB this FER increase is 3.6%, and for AMR-WB/G.718 it is 2.9%. Thus from both the intelligibility perspective and the FER perspective we conclude that the CA advantage is greatest with respect to AMR-WB, smallest with respect to AMR-WB/G.718, and the CA advantage with respect to EVS-WB falls between these two.

Our tests did not reveal any significant differences between EVS-WB CA and EVS-SWB CA in the frame-erasure environments and only one difference in the noise environments. Considering all noise environment results together, it appears that EVS-SWB CA has a minor intelligibility advantage over EVS-WB CA in the noise environments.

Our measurements have produced results that can guide development, deployment, and tuning of mission-critical voice applications that use these codec modes, and can also set expectations for the intelligibility they will produce in operating environments.

5. REFERENCES

[1] NPSTC, "Mission critical voice communications requirements for public safety," Littleton, CO, 2011.

[2] 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 https://www.its.bldrdoc.gov/publications/2627.aspx.

[3] 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 https://www.its.bldrdoc.gov/publications/2626.aspx.

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

[5] D. Atkinson and A. Catellier, "Intelligibility of selected radio systems in the presence of fireground noise: Test plan and results," NTIA Report 08-453, Washington D.C., 2008. Available https://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, Washington D.C., 2013. Available https://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, Washington D.C., 2013. Available https://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, Washington D.C., 2015. Available https://www.its.bldrdoc.gov/publications/2811.aspx.

[9] A. Catellier and S. Voran, "Intelligibility of selected speech codecs in frame-erasure conditions," NTIA Report 17-522, Washington D.C., 2106. Available https://www.its.bldrdoc.gov/publications/3165.aspx.

[10] 3GPP, *TS 26.442: Codec for Enhanced Voice Services (EVS); ANSI C code (fixed-point)*, ETSI.

[11] S. Voran, "Listener ratings of speech passbands," in *Proc. 1997 IEEE Workshop on Speech Coding for Telecommunications*, Pocono Manor, PA, 1997. Available https://www.its.bldrdoc.gov/publications/2648.aspx.

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

[14] V. Atti, *et. al*., "Improved error resilience for VoLTE and VoIP with 3GPP EVS channel aware coding," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, 2015.

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

[16] S. Voran and A. Catellier, "A crowdsourced speech intelligibility test that agrees with, has higher repeatability than, lab tests," NTIA Technical Memo 17-523, Washington D.C., 2017. Available https://www.its.bldrdoc.gov/publications/3168.aspx.

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

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

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

[20] E.N Gilbert, "Capacity of a burst-noise channel," *Bell System Technical Journal*, 39: 1253-1266, 1960.

[21] 3GPP TSG-SA4 Meeting #86, Document S4-151313, "MCPTT: Codec performance over MCPTT bearers," October 2015.

[22] ITU-T Recommendation P.863, "*Perceptual objective listening quality assessment*," Geneva, 2014.

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 foundation set by the late DJ Atkinson in his earlier PSCR speech intelligibility studies. The present work would not have been possible without DJ's vision, leadership, and hard work. We are deeply indebted to DJ and seek to honor his memory in our work. We also recognize the technical reviewers who provided essential input to this report and we extend deep gratitude to ITS Publications Officer Lilli Segre for her tireless and thorough editorial revisions that have produced this final product.

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-18-529		

4. TITLE AND SUBTITLE	5. Publication Date
Intelligibility Robustness of Five Speech Codec Modes in Frame-Erasure and Background Noise-Environments	December 2017
	6. Performin Organization Code NTIA/ITS.P

7. AUTHOR(S)	9. Project/Task/Work Unit N
Stephen D. Voran and Andrew A. Catellier	6818000-300

8. PERFORMING ORGANIZATION NAME AND ADDRESS	10. Contract/Grant Number.
Institute for Telecommunication Sciences National Telecommunications & Information Administration U.S. Department of Commerce 325 Broadway Boulder, CO 80305	

11. Sponsoring Organization Name and Address	12. Type of Report and Period Covered
Science and Technology Directorate Department of Homeland Security (DHS) Washington, DC	

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.)

Frame erasures and background noise are two factors that can interact with speech coding to reduce speech intelligibility and thus impair public safety mission-critical voice communications. We conducted two tests of intelligibility in the face of these factors. The tests covered five adaptive multi-rate (AMR) and enhanced voice services (EVS) speech coding modes, each using a bit rate near 13 kb/s. Two EVS Channel Aware (CA) modes were included. Both tests use the Modified Rhyme Test (MRT) protocol and together they comprise over 150,000 trials. The first test used frame erasures targeted at critical consonants for maximum sensitivity and the second used frame erasures generated at random by a two-state Gauss-Markov model. By using these large numbers of MRT trials we found that the CA codec modes offer small but statistically significant speech intelligibility improvements in numerous frame-erasure environments.

16. Key Words (Alphabetical order, separated by semicolons)
AMR, EVS, channel aware, frame erasure, frame loss, MRT, noise, packet loss, speech coding, speech intelligibility, speech quality

17. AVAILABILITY STATEMENT	18. Security Class. (This report)	20. Number of pages
☒ UNLIMITED.	Unclassified	55
☐ FOR OFFICIAL DISTRIBUTION.	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.

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.