NIST Technical Note
NIST TN 2259

Mission Critical Voice Quality of Experience Probability of Successful Delivery Measurement Methods

Jaden Pieper
Jesse Frey
Gary Howarth

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2259

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Technical Note
NIST TN 2259

Mission Critical Voice Quality of Experience Probability of Successful Delivery Measurement Methods

Jaden Pieper
Jesse Frey
formerly, Public Safety Communications Research Division
Communications Technology Laboratory

Gary Howarth
Public Safety Communications Research Division
Communications Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2259

August 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

NIST TN 2259
August 2023

Publication History
Approved by the NIST Editorial Review Board on 2023-07-31

How to cite this NIST Technical Series Publication:
Jaden Pieper, Jesse Frey, Gary Howarth (2023) Mission Critical Voice Quality of Experience Probability of Successful Delivery Measurement Methods. (National Institute of Standards and Technology, Gaithersburg, MD), NIST TN 2259. https://doi.org/10.6028/NIST.TN.2259

NIST Author ORCID iDs
Jaden Pieper: 0000-0001-8061-6216
Jesse Frey: 0000-0002-0354-1377
Gary Howarth: 0000-0002-3587-0546

Contact Information
pscr@nist.gov

Abstract

This paper describes the probability of successful delivery (PSuD), a novel measurement method for mission-critical voice (MCV) communication systems. The PSuD method evaluates two MCV contexts, every word critical (EWC) and average message intelligibility (AMI). These contexts represent successful message transmission in scenarios where context is unavailable and where context is available, respectively. The PSuD measurement is a component of the MCV quality of experience (QoE) system developed by the National Institute of Standards and Technology (NIST) Public Safety Communications Research Division (PSCR). In this paper, we present the "probabilityiser" (PBI), a tool that simulates an unreliable audio channel for purposes of channel access and retention. Using the PBI we validate the PSuD method and demonstrate the measurement's utility on audio encoded with various audio codecs.

Keywords

Mission-critical voice (MCV), quality of experience (QoE), every word critical (EWC), average message intelligibility (AMI), public safety communications.

Table of Contents

List of Tables

List of Figures

Acronyms

ABC-MRT16 Articulation Band Correlation Modified Rhyme Test 16. 1, 2, 4, 6, 13

AMI average message intelligibility. i, 3, 7, 8, 10, 13

AMR-NB adaptive multi-rate narrow-band. 5

AMR-WB adaptive multi-rate wide-band. 5

EWC every word critical. i, 2, 3, 7, 8, 10, 13

FPR false positive rate. 6

KPI key performance indicator. 1, 5, 13

LMR land mobile radio. 1, 13

LTE long-term evolution. 1, 13

M2E mouth-to-ear. 1, 4, 5

MCV mission-critical voice. i, 1, 2, 13

MRT Modified Rhyme Test. 1, 2, 4–6, 8, 13

NIST National Institute of Standards and Technology. i, 1

PBI "probabilityiser". i, ii, 3–5, 7–10, 13

PSCR Public Safety Communications Research Division. i, 1, 2, 13

PSuD probability of successful delivery. i, 1–10, 13

PTT push-to-talk. 1, 2, 6, 13

QoE quality of experience. i, 1, 2, 13

QoS quality of service. 1

ROC receiver operating characteristic. ii, 6

SNR signal-to-noise ratio. 5

SUT system under test. 3, 6

TPR true positive rate. 6

WUT word under test. 2

List of Symbols

c length of time for PBI state changes. 3, 8

G non-transmitting state. 3, 8

H transmitting state. 3, 8

I_t intelligibility threshold. 7

P_A probability of transitioning from a non-transmitting state to a transmitting state. 3, 5, 8

P_R probability of remaining in a transmitting state. 3, 5, 8

ρ_0 correlation coefficient. ii, 5, 6

1. Introduction

The National Institute of Standards and Technology (NIST) Public Safety Communications Research Division (PSCR) has developed a measurement system[1] to evaluate the quality of experience (QoE) of mission-critical voice (MCV) technologies. The QoE system comprises a set of key performance indicators (KPIs), software, and a hardware interface to measure those KPIs. The QoE system is compatible with hardware or applications implementing voice push-to-talk (PTT) protocols, regardless of the communications interface, allowing for the examination of numerous mission critical communication implementations, including land mobile radio (LMR) and long-term evolution (LTE).

This paper describes the probability of successful delivery (PSuD) KPI, which estimates the likelihood that a message will be successfully received under a particular test condition. The PSuD measurement builds from the same framework as previously described KPIs, end-to-end access time [2–4] and mouth-to-ear (M2E) latency [5].

The PSuD measurement evaluates the likelihood that an English audio message of ten seconds or less is successfully received under particular test conditions. We determine successful delivery by evaluating the intelligibility of each of the transmitted words in the message. We use the Articulation Band Correlation Modified Rhyme Test 16 (ABC-MRT16) [6], an objective computational algorithm, to determine the intelligibility of spoken words.

2. Background

PSCR held a roundtable event in March 2017 with industry and public safety representatives to identify expectations and metrics that would enable PSCR to understand, measure, monitor, and predict MCV QoE across LMR, LTE, and future technologies. Quantifying QoE is a departure from traditional quality of service (QoS) metrics, which focus on network and device performance. Instead, QoE focuses on the end users and their experience with the communications system. The four KPIs identified in the roundtable event were M2E latency, end-to-end access time, audio quality/intelligibility, and the probability of channel access and retention.

3. Measurement Definition

We define the PSuD, $P_S(T)$, as the probability of successfully transmitting and receiving a message of length T. Here, success means that the received message satisfied some intelligibility constraints. This definition is closely tied to the notion of probability of access and retention. Namely, the PSuD encompasses and quantifies both the likelihood that a user was granted access to a channel to begin their transmission, and also that they retained access to that channel for enough time to transmit their message intelligibly.

Previous MCV QoE measurement methods have dealt with intelligibility strictly within the context of Modified Rhyme Test (MRT) intelligibility. In an MRT, intelligibility of a

keyword is measured with no context. In particular, an MRT trial consists of a user hearing the carrier phrase, "please select the word", followed by the word under test (WUT). So a full MRT trial might be "please select the word west". Here the carrier phrase provides no context for what the WUT is, as the carrier phrase is always the same and is presented in each trial.

The definition of PSuD necessitates expanding our notion of intelligibility: we must now consider intelligibility of messages rather than intelligibility of individual words. This introduces the notion of context within a message, where components of a message either do or do not contribute to the understanding of other parts of the message. In other words, part of a message may be unintelligible to the listener, but the meaning of the full message is still understood correctly. In most day-to-day speech, there is a lot of context; however, there are certainly circumstances where there is no context in a message. For example, in public safety scenarios, an individual may need to relay a license plate number or a phone number. If one character is missed, the validity of the entire message is compromised. Thus, PSuD must be able to account for different levels of context to fairly measure a PTT communication system.

For all other MCV QoE measurements, PSCR has used the objective intelligibility estimation algorithm, ABC-MRT16. This algorithm allows for on-demand intelligibility estimations of MRT keywords. Because this algorithm has been successfully incorporated in our previous measurement systems, it was decided to also use it for the measurement of the PSuD. In particular, for this measurement, messages were constructed via strings of MRT keywords, with no carrier phrase present. This means that our messages inherently lack context.[1] This implies that our messages are well suited for measuring the PSuD of messages with no context; however, we needed to consider some way of simulating context within our messages to make our measurement of the PSuD more robust. Next, we explicitly describe the two message context measures we consider.

The first case we consider is that of every word critical (EWC) messages. In these messages, there is no context, so one must understand every word in order to understand the entire message. If a single word in the message is not intelligible, then the entire message is not understood. Examples of such messages are phone numbers, license plate numbers, addresses, and the phrase "don't shoot". If you miss a digit of a phone number, the rest of the message is essentially useless. Our messages of MRT keyword strings are very well suited for the EWC case because they lack context by design.

We next considered how to represent the intelligibility of a message with context. To do this we first took the best-case scenario for context in a message, where every word provides context to every other word in a message. This case, where context is uniformly distributed

[1]While MRT lists have structure (words within a list rhyme in some sense) that may provide context, the way our messages were constructed took no requirements about lists into account beyond ensuring no words from the same batch were neighbors in a message. Thus, it is safe to say the messages are truly without context.

throughout the message, represents the ideal context scenario. We call messages of this type average message intelligibility (AMI) messages, where the average intelligibility of each word within the message accurately describes the overall context of the entire message. The message, "both suspects turned right," roughly meets this case. Consider a case whether all words but one come in at perfect intelligibility (1), but one comes in completely unintelligibly (0). Any three words that are intelligible help a user understand most of the message and provide enough information that the message is likely understood. In particular, the average intelligibility across the whole message was $3/4$. Both the EWC and AMI PSuD cases provide important data points. Between the two scenarios, a broad understanding of the performance of a particular system under test (SUT) can be determined.

4. Measurement Implementation

4.1. Simulating Unreliable Channels

The PSuD measurement system is intended to assess the reliability of a communications system. In particular, it aims to capture whether or not a user can access a channel and maintain it long enough to successfully communicate. However, in a lab setting, channel conditions are typically very good. In general, it is difficult to reliably maintain known poor conditions in order to test/verify the PSuD measurement.

The "probabilityiser" (PBI) is designed to address this challenge. It is a simple way to model and simulate conditions where a channel completely drops audio. It is implemented in both hardware and software. The PBI operates from two states: a non-transmitting state, G, and a transmitting state, H. The transmitting state lets audio through, while the non-transmitting state completely blocks audio. The PBI is configurable through three parameters. The first, P_A, represents the probability of transitioning from the non-transmitting state, G, to the transmitting state, H. The second, P_R, represents the probability of remaining in the transmitting state, H. The last, c, represents how often the PBI updates its state. For a detailed discussion of how the PBI model parameters affect communication channel error rates and how model parameters can be derived from specific error statistics, see reference [7]. Fig. 1 shows the Markov chain that represents the PBI.

The PBI also allows the user to record the history of its states. This provides a record of the true state of the communications system in simulations and yields an important reference point for design choices in the development of the PSuD measurement system.

4.2. Audio Clips and Audio Alignment

The PSuD measurement system is intended to help characterize the reliability of a communications channel. This means it must be able to handle both good and bad channel conditions. In bad channel conditions, it is likely that significant portions of the transmitted audio will not be received. Complete loss of audio is the most challenging impairment to deal with because it severely impacts the ability to make accurate latency estimations.

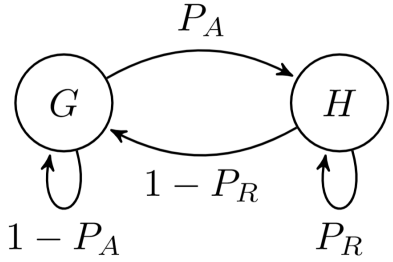

Fig. 1. Markov chain representation of PBI state transition where G and H are the non-transmitting and transmitting states, respectively.

Latency estimations are a critical step in processing received audio because only by accounting for latency can we make measurements about the behavior of specific portions of our messages—namely, the MRT keywords. This concern was primarily addressed in two ways. First, we designed the audio clips with a specific structure to aid in latency estimation. Second, we performed a study on the value of the correlation coefficient calculated during latency estimations to establish a correlation threshold to label successful or unsuccessful latency estimates. If latency estimations are not deemed successful, we can make a more constrained latency estimate relying on audio clip structure; however, that estimate relies on more general assumptions about system behavior because it requires the communications system M2E latency to be less than one second.

4.2.1. Audio Clip Design

The audio clips designed for the measurement of the PSuD are composed of strings of MRT keywords. The MRT includes 300 English-language keywords organized into 50 lists of rhyming words, with six words in each list (e.g., “kit,” “bit,” “fit,” “hit,” “wit,” and “sit”). In the MRT audio library used [8], each MRT set is spoken by a single talker, and the entire set is replicated with four total talkers. The PSuD audio clips are formed by drawing a series of words from the MRT from a single talker, with each word being drawn from a separate list (i.e., none of the PSuD clips contain rhyming words). A single MRT keyword is played each second. Because most MRT keywords in the ABC-MRT16 keyword data set are approximately 500 ms long, each keyword is followed by some amount of “quiet noise,” which is essentially silence, to fill out the complete second. As discussed in [3], variable-length silence can aid in the estimation of latency in the received audio. Thus, two goals are achieved by having a single MRT keyword per second: the message is structured with plentiful information to aid in the latency estimation, and in the event of a bad latency estimation, the likelihood of sending erroneous audio to ABC-MRT16 is minimized as a result of the non-speech audio surrounding the MRT keywords.

The audio clips were designed to be 10 s long, so each contains 10 MRT keywords. Because there are 300 MRT keywords, each talker has 30 clips. With four talkers, this means 120 audio clips exist.

However, we want to minimize potential vocoder dependency/memory effects. Specifically, the behavior of most codecs as it encodes/decodes a word is dependent on the audio it has previously seen. Therefore, the ordering of words in our clips may have an effect on the intelligibility of specific keywords. We have multiple sets of clips, all with different, randomized selections of keywords to attempt to negate this effect. Four sets of clips exist, each with 120 clips. Thus, in order to run a full measurement of PSuD, four sessions of 120 clips each must be performed, so 480 audio clips are used for final results of the PSuD.

4.2.2. Audio Alignment Study

M2E latency estimates rely on a cross-correlation calculation. This means that whenever latency is estimated, we can observe the correlation coefficient, ρ_0, associated with estimated latency. In the case where the received audio is simply offset from the transmit audio with no other impairments, this coefficient will be 1. As the received audio becomes more and more impaired, this coefficient will decrease (i.e., the received audio looks less and less like the transmitted audio to the estimation algorithm, and thus will have a $\rho_0 < 1$). This means the correlation coefficient can be viewed as a confidence measure of the latency estimate. When the correlation coefficient is high, it is likely that the estimate is accurate; when it is low, it may not be accurate.

By using the PBI and recording its state history for a series of transmissions, we created a database of audio with varying degrees of words missing that we could easily label. Using the previously implemented M2E KPI measurement [5], we ran a series of simulations and a variety of PBI values impairing the audio. We use these data to calculate both the delay estimate and the corresponding correlation coefficient, ρ_0, and we identify a threshold for the correlation coefficient that leads to accurate latency estimates.

We have simulations using four audio-encoding scenarios including adaptive multi-rate wide-band (AMR-WB)[9], adaptive multi-rate narrow-band (AMR-NB)[9], simulated frequency-modulated analog radio (analog) [10], and clean audio (no codec). We repeat each scenario with the following set of PBI settings, $P_A = P_R \in \{0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 1\}$. The PBI interval is set to 0.2 seconds, and signal-to-noise ratio (SNR) levels are set to 80 and 0 dB.

A series of 10 MRT phrases spoken by four different talkers, two male and two female, for a total of 40 audio clips. Finally, each set of conditions is run 15 times to get a variety of outcomes from the PBI for the same settings. Thus, in total, 14 440 trials are performed.

To determine a threshold for the correlation coefficient, simulation trials are labeled as successes or failures based on the accuracy of the latency estimate. In particular, because we control the true latency in the simulations, we simply measure the difference between

the true latency and the estimated latency for each trial. We use an error tolerance of 15 ms around the true latency to label successes. Any latency estimates that exceed 15 ms difference from the true latency are labeled as failures. We then plot a receiver operating characteristic (ROC) of the labeled data for a variety of ρ_0 thresholds. In particular, we plot the true positive rate (TPR) against the false positive rate (FPR) for ρ_0 threshold values between 0.1 and 1. TPR measures the percentage of successes that meet the ρ_0 criteria. While FPR measures the percentage of actual failures that would be counted as positive at a specified ρ_0 threshold. The goal is to select a value of ρ_0 that has a high TPR and a suitably low FPR. For the case of measuring PSuD, the cost of false positives is much greater than the cost of dismissing additional true positives. Thus, an FPR of 1% is deemed acceptable. Fig. 2 shows the ROC plot with a latency error tolerance of 15 ms, and shows the ρ_0 value associated with an FPR near 1%. We found this corresponds to a ρ_0 threshold of 0.76. Therefore, we accept any latency estimate where ρ_0 is measured to be at least 0.76. Otherwise, the initial estimate is discarded and the algorithm repeats the estimate with a forced constraint of latency values between 0 s and 1 s. While this constraint may be limiting for certain PTT communications technologies, it allows us to leverage the audio clip structure to achieve better latency estimates in hard-to-estimate audio.

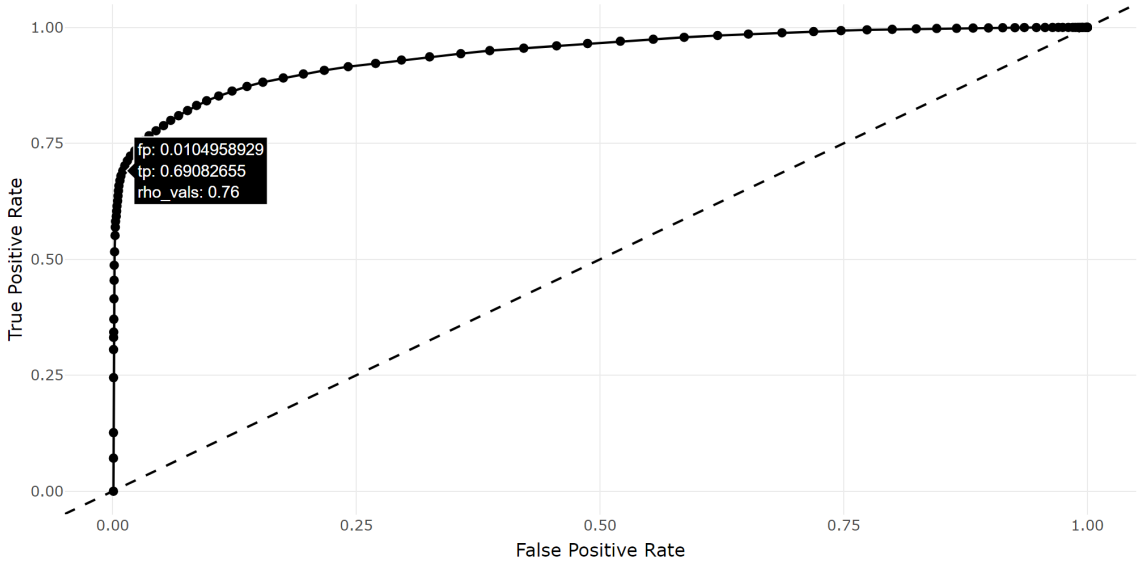

Fig. 2. ρ_0 ROC curve for a error tolerance of 15 ms.

4.3. Measurement Procedure

Measurements of the PSuD consist of sending audio through a PTT communications system, the SUT. One session consists of 120 trials, where each trial sends a single audio clip of ten MRT keywords through the SUT. For a full PSuD measurement, four sessions are run—one for each set (talkers) of PSuD audio. Within a trial, the intelligibility of each of the 10 MRT keywords in the audio clips is estimated using ABC-MRT16. Because the

PSuD is defined for a message of length T, the maximum length message that can be evaluated with this implementation of PSuD is 10 s. It would be trivial to construct longer audio clips for the measurement of the PSuD; however, messages longer than 10 s do not seem common for public safety [11]. Note that because our measurement of the PSuD relies on the intelligibility of keywords, and keywords occur every second, our measurements are essentially discrete in the time domain. Thus, as a function of T, PSuD behaves as a discrete ceiling function—e.g., $P_S(T) = P_S(\lceil T \rceil)$.

As discussed earlier, we consider two different speech contexts for the measurement of the PSuD: EWC and AMI. Next, we specifically detail how these measures of the PSuD are calculated during a measurement.

4.3.1. Every Word Critical

To calculate the EWC PSuD, we specify an intelligibility threshold, I_t. We determine the sequence length of the transmitted message, N, using the cut-points of the audio, which we have designed to be one per second each. Therefore, $N = \lceil T \rceil - 1$, where T is the length of the transmitted message in seconds. We then determine the length (k) of the sequence of words received above I_t from a transmitted N word message. If $N = k$, the trial is labeled a success. If $k < N$, the trial is labeled a failure. The final measured PSuD value is the success rate (success per trail) over the entire set of sequences along with a 95 % confidence interval. Uncertainty for all PSuD measurements is calculated using bootstrap resampling on the success/failure data, regardless of the context metric.

4.3.2. Average Message Intelligibility

For AMI PSuD, the average message intelligibility is calculated. In particular, we calculate

$$I_k = \frac{1}{k} \sum_{j=1}^{k} w_j$$

The intelligibility of a received word is w_j ($j \in \{1, 2 \ldots, N\}$), the transmitted message length is T seconds, and the sequence of received words is k ($k = \lceil T \rceil - 1$). If $I_k \geq I_t$, the trial is labeled a success; otherwise, it is labeled a failure. The final PSuD value is the mean of all trials.

5. Measurement Validation

To validate the PSuD measurement system, we run a series of simulations across different technologies and conditions using the PBI. For all simulations, the PBI is set to change states every second; this means that entire words either received or not.

5.1. Expected PSuD with the PBI

The PBI always starts in state G, the non-transmitting state. Accordingly, for the EWC PSuD case, to send a k-word long message, the PBI must transition from G to H and remain there $k-1$ iterations, when the PBI update interval, c, is equal to 1. More generally, the EWC PSuD when using the PBI is

$$P_S(T) = P_A \cdot P_R^{(\lceil T \rceil / c)-1} \tag{1}$$

which means that given values of P_A and P_R, we can predict the expected EWC PSuD of a simulation.

For the AMI PSuD, we are not aware of a simple closed form expression to estimate the expected result. However, since our PBI update time is 1 s and our maximum message length is 10 s, it is possible to create every possible combination to solve for the expected average intelligibility for a T-second-long message given values of P_A and P_R.

5.2. Simulation Experiment Design

For the simulation tests, we want to study the success of the measurement system across a variety of PSuD values. The PBI can achieve the same PSuD a multitude of ways, so we select four (P_A, P_R) combinations per PSuD value. PSuD values of $[0.1, 0.2, ..., 0.9]$ are studied. For all the following, we fix $k = 3$, (i.e., we studied PSuD values focused on messages of length 3). Given a PSuD value of P_S, we can solve for P_A as

$$P_A = \frac{P_S}{P_R^{k-1}}.$$

For P_A to be a valid probability, $P_S \leq P_R^{(k-1)}$ must hold, or equivalently, $P_S^{1/(k-1)} \leq P_R$. Thus, the minimum P_R value is set to be $P_{R,\text{min}} = P_S^{1/(k-1)}$. The maximum P_R is set as $P_{R,\text{max}} = 0.99$. A step size of $s = (P_{R,\text{max}} - P_{R,\text{min}})/3$ is used to select four equidistant P_R values from $P_{R,\text{min}}$, such that $P_{R,j} = P_{R,\text{min}} + s * j$ for $j = 0, 1, 2, 3$. Then $P_{A,j} = \frac{P_S}{P_{R,j}^{k-1}}$, and we had four unique sets of PBI parameters that all achieve the PSuD value P_S.

The simulation then iterates through each (P_A, P_R) pair and runs a PSuD test. Each PSuD test involves four sessions, one for each keyword clip ordering. Each keyword clip ordering contains 120 audio clips, each containing ten MRT keywords. So, in total, 480 trials are run per test. Results are aggregated across all four sessions in a test.

5.3. Evaluating Success

We say a simulation is successful if the 95 % confidence interval for the measured PSuD contains the expected PSuD. By the definition of a 95 % confidence interval, this means we expect roughly 5 % of our tests to fail; however, even in those instances, it is probable that

the confidence interval is close to the expected value. The expectation is that if there is a systemic error in the PSuD measurement, we will see a significant amount of measurement results that vary greatly from the expected result.

5.4. Test Corrections

The following two corrections are used for evaluating the success of all tests.

5.4.1. Machine Error Correction

In some tests, there are small numerical errors where the expected PSuD is almost 1 but is off by machine precision, tol $= 2.22e-16$. In such cases, the measured PSuD evaluates identically to 1, so the 95 % confidence interval (calculated via resampling) contains only 1, and those trials are labeled as failures. This correction involves labeling any tests that satisfy $|P_m - P_e| \leq$ tol, as a success, where P_m is the measured PSuD, P_e is the expected PSuD, and tol is machine precision.

5.4.2. Limitation of Bootstrap Resampling Statistics

All confidence intervals for PSuD are calculated using bootstrap resampling. The nature of a resampled technique is to take the sample distribution and substitute it for the true population distribution. In this instance, we know the population distribution is extremely unlikely to emit a failure in the relatively small number of samples we take, so the sample distribution is underprepared to account for this. This is a limitation of resampled statistics, but in this instance, since we know the true population distribution, we can account for the measured outcome and say it is "successful enough" and note that with a much larger sample size, we would likely better approximate the true population distribution.

In particular, we say that if $P_m = 1$ and $|1 - P_e| \leq 1/480$, then that simulation is a success. In such cases, the number of trials required to likely see a single failure is higher than the number of trials we run, so we measure a PSuD that is identically 1, regardless of what the PBI may have output, we run many more trials.

To ground this in reality, our assumption is that it is unlikely that first responders will notice a significantly different performance in a SUT if its PSuD is > 0.99, and that these simulations are not demonstrating a meaningful failure of the measurement system.

5.5. Results

Results for a clean audio channel with PBI settings described in Section 5.2 are shown in Figs. 3 and summarized in Table 1. Uncorrected successes refer to the number of successes without the two test corrections described in Sec. 5.4.

Note that all results rely on accurate predictions of the PSuD. Predictions are only intended for the clean case, where no vocoder is used and no impairments exist outside of the PBI.

Table 1. Clean Channel Simulation Results

Technology	Context	Uncorrected Success	Corrected Success	Rate of Success
Clean	EWC	129	135	93.8%
Clean	AMI	110	121	84%

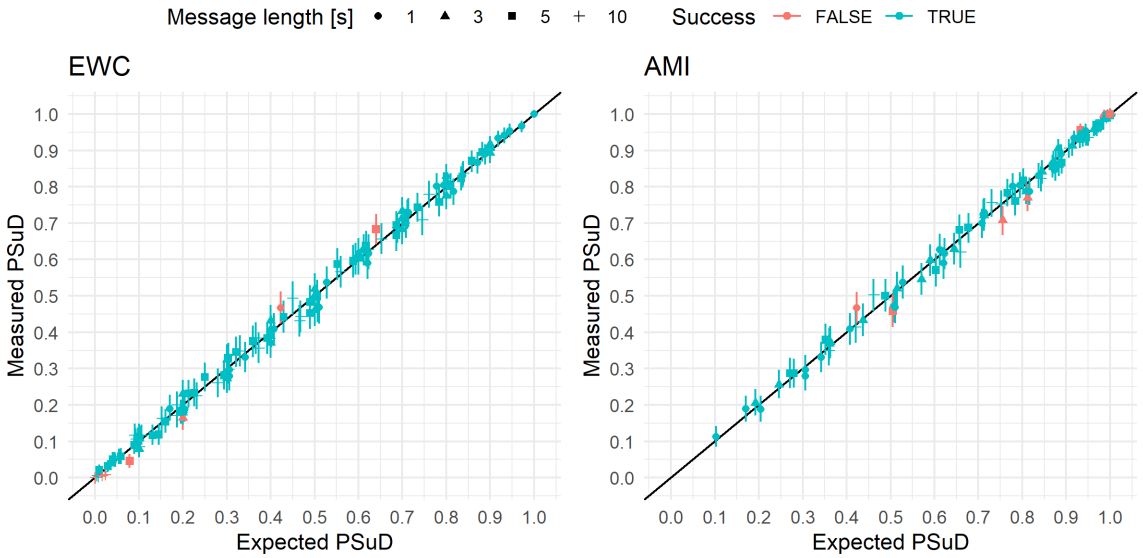

Fig. 3. Clean simulation results

Thus, it is reasonable to expect the measured PSuD to be lower than the expected PSuD for technologies that have a more severe impact on intelligibility.

6. Example Measurements

We perform the PSuD measurement technique simulating an unreliable channel using the PBI settings described in 5.2 with three audio codecs, the Adaptive Multi-Rate - Wideband (AMR-WB) speech codec, the Adaptive Multi-Rate - Narrowband (AMR-NB) speech codec, and analog codec. These measurements are displayed in Fig. 4 through Fig. 6 and summarized in Table 2.

The results suggest that that measurements across simulated channels agree closely with predicted PSuD values for both EWC and AMI models.

7. Discussion

The AMI model sees a higher failure rate, as we define failure, for every test scenario. However, the AMI failures tend to be clustered at high PSuD values.

Table 2. Codec Simulation Results

Technology	Context	Uncorrected Success	Corrected Success	Rate of Success
AMR-WB	EWC	132	138	95.8%
AMR-WB	AMI	116	127	88.2%
AMR-NB	EWC	126	132	91.7%
AMR-NB	AMI	113	124	86.1%
Analog	EWC	97	103	71.5%
Analog	AMI	106	117	81.2%

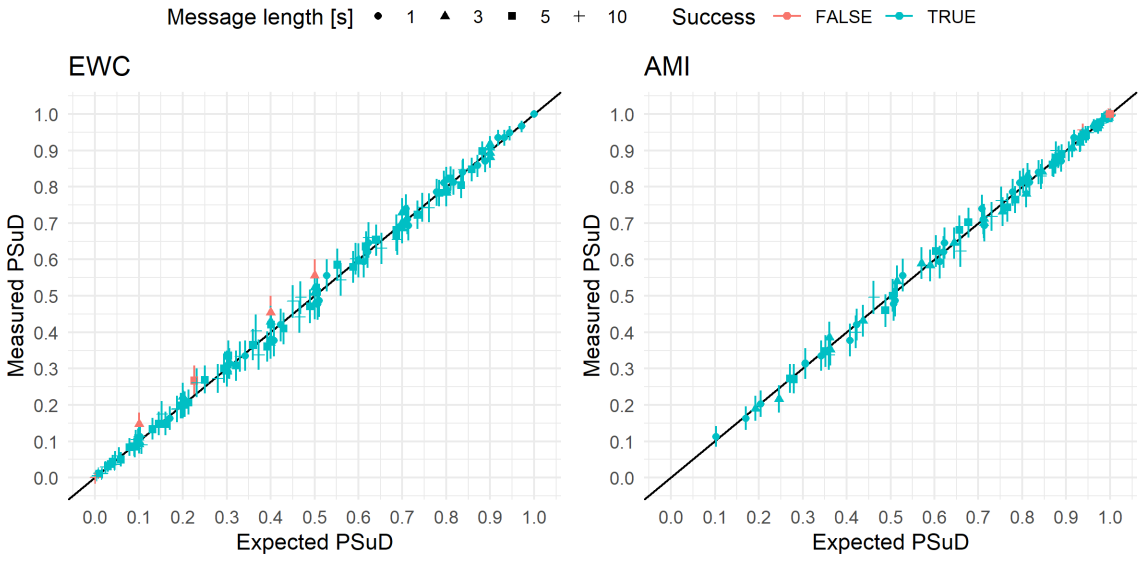

Fig. 4. AMR-WB simulation results

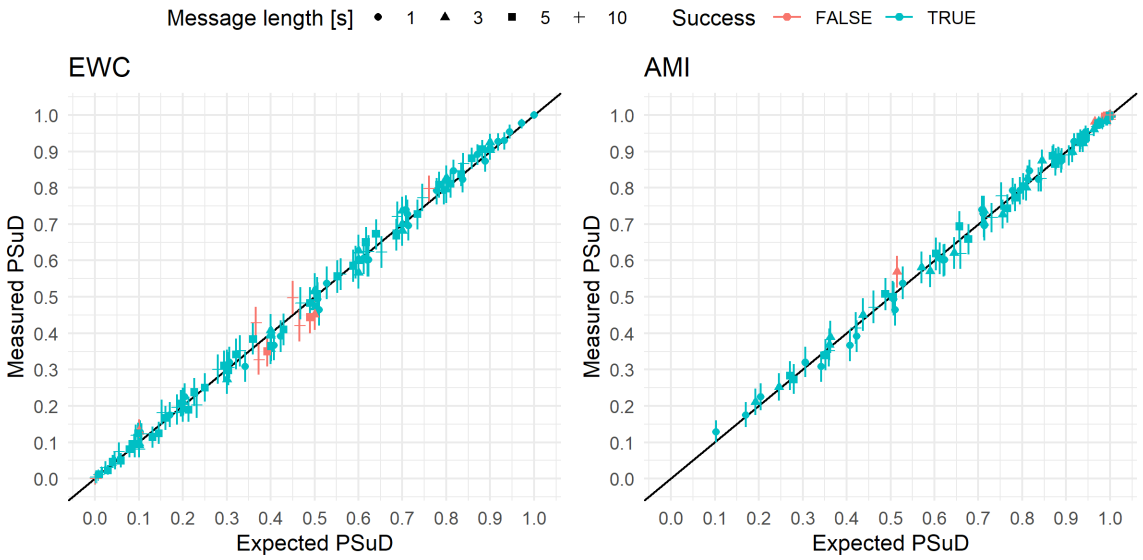

Fig. 5. AMR-NB simulation results

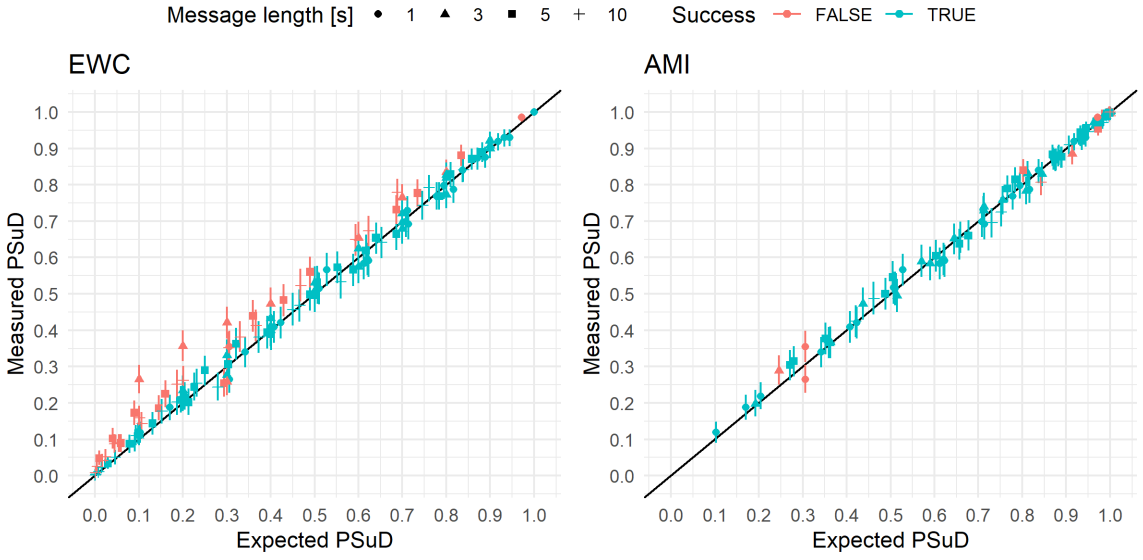

Fig. 6. Analog simulation results

The measured data at high PSuD values are highly replicable and experience a ceiling effect (as measurements cannot exceed a value of one), with both factors leading to tight errors. Therefore even a slight variation from the expected PSuD is out of the 95 % confidence interval. Yet, the absolute error of these points is very low.

From the nature of the models, there are more ways for AMI to obtain high expected PSuD values, so the AMI sampling tends to be skewed toward higher expected PSuD values. This and the relative error effects discussed directly above contribute to worse performance of AMI compared to EWC measurements across all evaluations.

For the FM analog radio simulated data, we see that the expected PSuD values are systematically lower than the measured value for the EWC context, especially at values less than 0.6. At low values of the expected PSuD there are significant periods of silence, which is not what the the ABC-MRT16 algorithm is designed to handle. We hypothesize that silence is leading to a slight overestimation of the measured PSuD.

The acceptable PSuD for MCV is likely much higher than 0.5 for public safety. While our simulations aim to test the full range of PSuD values, it is reasonable to focus on the performance above a functional PSuD threshold such as 0.5. Any communications system that cannot deliver a message reliably more than half of the time will likely never be used nor require stringent testing to demonstrate its poor performance. In the event that such a system is tested, the absolute error in our simulations suggests the measurement system will still characterize the poor performance of the communications system on a practical level, even if it is not as accurate at measuring low PSuD values when compared to high PSuD values.

In other words, it is probable that a PSuD of 0.4 vs 0.5 is functionally identical in a mission critical environment, where a PSuD of 0.8 vs 0.9 would not be.

8. Conclusion

Here we demonstrate the PSuD KPI and validate it using the PBI as an unreliable channel simulator. The PSuD measurement is integrated into the PSCR QoE system, which allows a user to investigate a wide range of PTT technologies, including LMR and LTE platforms. The PSuD measurement is one of the set of KPIs that specifically evaluate systems from a user-experience perspective using an automated testing platform and objective algorithms. The PSuD evaluation is based on the ABC-MRT16 algorithm, which demonstrates a high correlation (0.95) to human-interpreted MRT results when tested across 367 systems. Yet, the ABC-MRT16 algorithm is designed and validated for the use of whole-words. The PBI presented here can simulate interruptions in the beginning, middle, and end of words, as well as multiple interruptions. Real communications channels maybe interrupted at any point. A rigorous study of English language users' ability to interpret interrupted words and correlating this to the ABC-MRT16 could help refine the PSuD measurement.

References

[1] Kahn A (2021) Mission critical voice quality of experience measurement software (National Institute of Standards and Technology), https://doi.org/https://doi.org/10.18434/mds2-2456

[2] Pieper J, Frey J, Greene C, Soetan Z, Thompson T, Bradshaw D, Voran S (2019) Mission critical voice quality of experience access time measurement methods (National Institute of Standards and Technology), https://doi.org/10.6028/NIST.IR.8275. Available at https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8275.pdf

[3] Greene C, Frey J, Soetan Z, Pieper J, Thompson T (2020) Mission Critical Voice Quality of Experience Access Time Measurement Method Addendum (NIST), IR-8328. https://doi.org/10.6028/NIST.IR.8328

[4] Magrogan W, Pieper J, Soetan Z (2021) Mission Critical Voice Start-of-Word Correction for Access Delay Measurement System (National Institute of Standards and Technology, Gaithersburg, MD), https://doi.org/10.6028/NIST.TN.2166. Available at https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.2166.pdf

[5] Frey J, Pieper J, Thompson T (2018) Mission Critical Voice QoE Mouth-to-Ear Latency Measurement Methods (NIST), IR-8206. https://doi.org/10.6028/NIST.IR.8206

[6] Voran S (2013) Using articulation index band correlations to objectively estimate speech intelligibility consistent with the modified rhyme test. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* https://doi.org/10.1109/WASPAA.2013.6701826

[7] Pieper J, Voran S (2023) NTIA Technical Memorandum 23-565: Relationships between Gilbert-Elliot Burst Error Model Parameters and Error Statistics (National Telecommunications and Information Administration, Boulder, Colorado), Available at https://its.ntia.gov/publications/details.aspx?pub=3298.

[8] Institute for Telecommunication Sciences Modified Rhyme Test (MRT) Audio Library. Available at https://its.ntia.gov/research-topics/audio-quality-research/public-safety-audio-quality/mrt-library/.

[9] (2022) Adaptive Multi-Rate - Wideband (AMR-WB) speech codec / G.722.2 / 3GPP Specification: 26.171. Available at https://voiceage.com/AMR-WB.G.722.2.html.

[10] Atkinson DJ, Catellier AA (2013) Intelligibility of Selected Radio Systems in the Presence of Fireground Noise: Test Plan and Results (National Telecommunication and Information Administration, Washington, D.C.), Available at https://its.ntia.gov/umbraco/surface/download/publication?reportNumber=TR-13-495.pdf.

[11] Sharp DS, Cackov N, Laskovic N, Shao Q, Trajkovic L (2004) Analysis of public safety traffic on trunked land mobile radio systems. *IEEE Journal on selected areas in Communications* 22(7):1197–1205.