NISTIR 8328

Mission Critical Voice Quality of Experience Access Time Measurement Method Addendum

Chelsea Greene
Jesse Frey
Zainab Soetan
Jaden Pieper
Tim Thompson

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8328

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8328

Mission Critical Voice Quality of Experience Access Time Measurement Method Addendum

Chelsea Greene
Jesse Frey
Zainab Soetan
Jaden Pieper
Tim Thompson
Communications Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8328

December 2020

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology
Interagency or Internal Report 8328
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8328, 18 pages (December 2020)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8328

Acknowledgments

The authors would like to thank Steve Voran from the National Telecommunications and Information Administration's Institute for Telecommunications Sciences for his superb guidance and willingness to share his in-depth knowledge in developing measurement methods for voice communications systems. The authors would also like to thank Don Bradshaw of the National Institute of Standards and Technology's (NIST) Public Safety Communications Research (PSCR) Division for his leadership and direction in developing measurement systems and properly communicating test results to the public. Finally, the authors would like to acknowledge the mathematical contributions and helpful revisions provided by Cara O'Malley and William Magrogan of NIST PSCR.

Abstract

Access time generally describes the time associated with the establishment of a talk path upon user request to speak and has been identified as a key component of quality of experience (QoE) in voice communications. The National Institute of Standards and Technology's (NIST) Public Safety Communications Research (PSCR) Division has previously developed an access time measurement method in NISTIR 8275 [1]. Improvements were implemented in the access delay measurement method to better handle the idiosyncrasies of the various push-to-talk (PTT) technologies tested. These improvements include a more robust audio alignment procedure, and graceful communications failure detection and handling. This paper covers those improvements to the access delay measurement system, results of testing Project 25 (P25) technologies with and without encryption turned on, as well as some initial Long Term Evolution (LTE) measurements.

Key words

Access delay; Articulation Band Correlation Modified Rhyme Test (ABC-MRT); A-weight; Encryption; Key performance indicator (KPI); Land mobile radio (LMR); Latency; Mission Critical Push-to-Talk (MCPTT); Modified Rhyme Test (MRT); Mouth-to-ear (M2E); Packetized; Project 25 (P25); Public Safety; Push-to-talk (PTT); Quality of experience (QoE); Receive; Streaming; Transmit; Vocoder; Voice.

Table of Contents

List of Tables

List of Figures

Acronyms

ABC-MRT Articulation Band Correlation Modified Rhyme Test. i, 1, 2, 6

ADB Android Debug Bridge. 4, 5

AW A-weighting. 4, 5

dBA A-weighted decibel. 4

E2E end-to-end. 1, 5

KPI key performance indicator. i, 1, 5

LMR land mobile radio. i, 4

LTE Long Term Evolution. i, 1, 4, 6, 8–10

M2E mouth-to-ear. i, 1, 5, 6

MCPTT Mission Critical Push-to-Talk. i, 1, 4, 8, 9

MCV mission critical voice. 1, 10

MRT Modified Rhyme Test. i, 1, 2, 9

P25 Project 25. i, 1, 2, 5, 6, 8

PSCR Public Safety Communications Research. i, 1

PTT push-to-talk. i, 2, 3, 8

QoE quality of experience. i, 1

Symbols

α Intelligibility scaling factor. 6, 8

I_0 Asymptotic intelligibility. 6

λ Logistic parameter, intelligibility curve steepness. 6, 8

P_1 First utterance of MRT keyword. 2, 3

P_2 Second utterance of MRT keyword. 2–4, 10

T Time preceding P_1 and P_2 in audio clips. 3

t_0 Logistic parameter, intelligibility curve midpoint. 6, 8

1. Introduction

The National Institute of Standards and Technology's (NIST) Public Safety Communications Research (PSCR) Division has been developing measurement systems for key performance indicators (KPIs) of mission critical voice (MCV) applications. These quality of experience (QoE) based KPIs for MCV were established by round table discussions with first responders and industry representatives between 2016 and 2017. The PSCR team has developed measurement systems for access delay [1] and mouth-to-ear (M2E) latency [2]. The end-to-end (E2E) access time is the summation of the two parameters, access delay and M2E latency. This paper covers improvements to the access delay measurement system that is described in greater detail in Ref. [1].

Access delay measurements were performed on a few different technologies and results were provided in Ref. [1]. Since then, we have improved the robustness of the measurement system as well as made our first measurements on encrypted Project 25 (P25) and Mission Critical Push-to-Talk (MCPTT) over Long Term Evolution (LTE). In testing of P25 trunked Phase 2 systems, we noticed audio alignment issues. These issues seemed to be caused by the low-rate vocoder used in P25 trunked Phase 2. The audio alignment stage within the measurement system was made to be more robust to contend with audio from low-rate vocoders. For MCPTT over LTE applications, testing was performed on early prototype systems that were available at the time of testing. Extra steps were taken to improve the measurement system for testing of these nascent systems that will likely not be required with more mature MCPTT over LTE systems.

The effects of encryption on E2E access time values were measured on P25 technologies, requiring no modifications of the measurement system. Encryption added 30-50 ms to the access delay values of the P25 technologies. In initial testing of LTE MCPTT for a raw intelligibility of 85%, access delay was 103.9 ± 4.1 ms with a level of confidence of 95%.

Throughout this paper, intelligibility refers specifically to Modified Rhyme Test (MRT) intelligibility [3]. Furthermore, intelligibility estimates were calculated using the objective speech intelligibility algorithm Articulation Band Correlation Modified Rhyme Test (ABC-MRT). Throughout this paper, ABC-MRT refers specifically to ABC-MRT16 [4].

2. Measurement System Modifications

This paper covers modifications to the existing access delay measurement system previously developed by PSCR. To fully understand the topics in this paper, it is highly recommended to read the original paper covering the access delay measurement system which is described in great detail in Ref. [1].

2.1 Audio Alignment Modifications

Accurate received audio latency calculations are critical to the success of an access delay measurement. The latency estimate is used to align received audio with the transmitted

audio so that the MRT keyword played as P_1 and P_2 within a test audio clip can be extracted. If the latency values are inaccurate, the wrong audio will be extracted from the received audio and sent to ABC-MRT. This can cause artificially low intelligibility values to be reported for these trials, which can cause an inaccurate curve fit and also bias the asymptotic intelligibility calculation to be lower than it is in reality.

Received audio latency estimates were erroneous in measurements using a P25 trunked Phase 2 system. The erroneous estimates had larger deviations from truth when the push-to-talk (PTT) times were large. By examining some of the received audio clips where this issue occurred, it was hypothesized that the low-rate vocoder used by the P25 trunked Phase 2 system was causing different words within the filled section of the audio clip to appear similar to the latency estimation algorithm. This issue was compounded by the fact that the trials that have large PTT times are the trials where the least amount of speech is received. Less received audio yields less data for the latency calculation, so the impairments on the envelopes of the words within the filled speech have a bigger impact on the output of the latency estimate.

We have updated the design of the audio clips used for access delay measurements with three modifications to address these issues: add variable spacing to the filled audio section, constrain the latency estimates to be non-negative, and increase the amount of filler speech used in access delay tests.

2.1.1 Variably Spaced Filler Speech

The filled section of access delay audio clips was designed specifically to improve the accuracy of audio latency estimates within an access delay test. The filled section consists of MRT keywords from batches with variable trailing consonants spliced together. The original intent of the filled section was to maximize the density of audio information within the filled section, in order to maximize the amount of information used for latency estimates. As such, the filled keywords were spliced directly after each other, with no silence between them.

The observed behavior of filled speech going through a low-rate vocoder demonstrated that the quality of the speech information for delay estimates was being diminished. While speech is important for accurate delay calculations, it is subject to distortions through the communications process. Silence is additional information that should be more invariant to vocoder effects, and therefore provides additional information for accurate latency estimates. In particular, by varying the amount of silence between MRT keywords, we can add additional structure into the filled audio section. Essentially, this concept can be interpreted as one can identify where in the audio clip they are by using both the keyword and the amount of silence preceding and trailing it.

Both of these sources of information, speech and silence, are important for accurate delay calculations, but they need to be carefully balanced to ensure enough speech is being received in a test. Variable word spacing in the filled section is achieved with the following process. The first word in the filled audio section has 100 ms of silence following it. Each

subsequent filled word has 2/3 the amount of silence following it as the previous filled word. Thus, the sequence of silence lengths between each word in the filled speech is [100 ms, 67 ms, 45 ms, 30 ms, 20 ms, 13 ms, ...]. This sequence continues to decrease until the silence length is 1 sample, at which point a consistent spacing of 1 sample is placed between keywords.

2.1.2 Enforce Non-Negative Latency Estimates

In the case where two different filler words have similar envelopes after being sent through a low-rate vocoder, it is possible that the latency algorithm can over focus on this effect and return an inaccurate delay result. This edge case is most likely when PTT time is large, where PTT is being pressed at the end of P_1. This corresponds with trials that have the least amount of speech in the received audio. In this case, it is possible that this inaccurate delay result is a negative number. However, given the design of the measurement system, it is not possible for negative delays to actually be achieved.

While the latency estimation algorithm was designed to be robust and to measure either positive or negative delays within a signal, this is not a necessary consideration within the access delay measurement system. Since it is impossible for negative delays to occur in a trial, the latency estimation algorithm was modified with the additional constraint that returned delay values had to be non-negative.

The limits are applied just after the cross-correlation in the coarse delay estimation [5] and prevent peaks that are outside the limits from being selected and forcing the selection of a peak that corresponds to a positive delay. It is worth noting that it is still possible for the latency estimation algorithm to return a negative value as the coarse delay estimation is followed by the fine delay estimation that can shift the delay to be a small negative value. This would, however, require that the coarse delay estimate was near zero which often means that there has been an issue that makes this recording unusable.

2.1.3 Additional Filler Speech

Finally, to attempt to ensure that enough filler speech is received even when PTT time is large, the variable T used in example measurements was increased. T describes the amount of silence before P_1 and the amount of filler speech between P_1 and P_2 in the audio clip. In initial access delay measurements, a value of $T = 2$ s was used. In order to ensure more speech would be present in received audio, a value of $T = 2.5$ s was adopted for more recent tests.

2.2 Failure Detection and Mitigation

It is possible that during an access delay measurement, a channel will not be granted and no audio will be transmitted from the transmit device. This can be the result of a poor network connection, something loose in the test setup, or an application crashing. If the channel is lost infrequently, the overall effect on the access delay test could be minimal. However,

if the channel is lost frequently, this can impair the ability to successfully measure access delay, as the integrity of the data is not consistent. In the particular case of an application crash where the channel stops working and does not come back up, this can make the rest of the test worthless. Thus, it is important to detect when connection problems arise and take action to ensure that the integrity of the test data is not compromised. Additionally, keeping track of failed channel detection can provide insight into system reliability.

2.2.1 Detecting Failed Receipt

When the test audio does not transmit as expected, there is a significant decrease in the signal power of P_2 in the received audio. In order to determine the signal speech power specifically, the received audio is filtered using A-weighting (AW) [6]. AW is commonly used to capture the human perception of loudness in signal power measurements. For the purpose of discussing power thresholds, A-weighted decibel (dBA) values are used to denote speech power after received audio is passed through an AW filter.

One can detect when a transmission drops partway through or is not received at all by measuring the dBA of P_2. If the dBA value is above a threshold value, this suggests the intended message was received in full. In a typical P_2 recording, dBA values are between -25 dBA and -30 dBA. A default threshold of -50 dBA was selected to avoid potentially mislabelling a successful test. Trials where the dBA is less than the threshold are flagged as "BAD" and placed in a separate file. A record of sub-optimal trials can help detect a problem in a communications network or may shed light on issues that occur during parts of the day with more network traffic. When a trial is flagged as "BAD", it is run again. The number of retries for a given trial is counted and if it exceeds the retry threshold, defaulted to three, the test is paused to allow the user to troubleshoot the system.

2.2.2 Mitigating Failure in Tests

Once a test reaches the maximum number of retries, the issue causing low dBA values must be addressed. With land mobile radio (LMR) devices, the user must manually continue the test if it stops due to too many retries. In our testing, LMR devices reaching this limit was quite rare. LTE MCPTT applications were still in early stages of their development at the time of testing. Because of this, there were issues with the applications stopping prematurely, impeding test progress. LTE devices with Android operating systems provided the opportunity to automate the retry process using the Android Debug Bridge (ADB).

ADB is a program used to install and debug Android applications as well as run shell commands [7]. The ability to run shell commands on the phone was used to kill and restart the application and simulate screen touches to get the application running again. Because the particulars of restarting each application are quite dependant on which application is being used, an external restart script is used to separate the application-specific code from the main test code.

For the testing of prototype and in-development technologies, we provide an example restart script; however, it is application specific and will not work on any other application.

Interested users can use this example as a template for developing their own restart script. When a restart script is used, the AW check is run as before except now when the number of retries threshold is exceeded, the ADB script is called. The script is given the number of times that it has been called for a given trial and returns whether or not the test should continue. At this stage, a user-created script could attempt various recovery strategies and, when all possibilities are exhausted, pause for the user to fix the problem.

3. Example Measurements of Encrypted P25 Communications

Encryption is important for security and widely used by public safety in sensitive situations. The original example access delay measurements performed in Ref. [1] were conducted without encryption, but it is imperative that our measurement systems can successfully measure encrypted technologies, given their widespread use. It is also informative to study the effect that encryption has on access delay and, in turn, the E2E access time measurements.

The process to enter encrypted communication mode varies by the radio brand; the radios tested have a two-position concentric switch that was programmed to turn on encryption. Once in encrypted communication mode, the test proceeded as usual. Table 1 summarizes the difference encrypted communication mode has on various P25 technologies used during testing. The values in Table 1 represent the access delay required to achieve a raw intelligibility of 85%.

Table 1. Example measurements of encrypted and unencrypted E2E access time results. 2500 ms audio clip access delay values for a raw intelligibility of 85%. Uncertainties are reported as 95% confidence intervals

	Access Delay [ms]	M2E Latency [ms]	E2E Access Time [ms]
P25 Direct Encrypted	174.1 ± 9.7	296.8 ± 1.7	470.9 ± 9.9
P25 Direct Unencrypted	126.9 ± 8.8	243.3 ± 0.3	370.2 ± 8.8
P25 Trunked Phase 1 Encrypted	695.8 ± 7.1	437.4 ± 6.3	1133.2 ± 9.5
P25 Trunked Phase 1 Unencrypted	643.3 ± 6.6	379.6 ± 2.1	1022.8 ± 6.9
P25 Trunked Phase 2 Encrypted	701.2 ± 6.5	636.1 ± 27.3	1337.3 ± 28.0
P25 Trunked Phase 2 Unencrypted	655.1 ± 6.3	601.9 ± 16.1	1257.0 ± 17.3

Because the encryption process effectively inserts additional fixed time processes into the communication chain, it was expected that M2E latency would be the KPI most im-

pacted by the presence of encryption. Our initial expectations were for access delay and intelligibility to be similar regardless of encryption, due to equivalent controlled, low-noise laboratory testing environments. As reflected in the results in Table 1, access delay and M2E latency were both impacted by the transition from clear to encrypted transmissions. For P25 technologies where encryption is turned on, we hypothesize that the delay is due to the loading time of the AES 256 cipher block.

The curve parameters for each test, both encrypted and unencrypted, are shown in Table 2. The column labeled I_0 describes the asymptotic intelligibility for each technology. As expected, encryption did not significantly impact the asymptotic intelligibility for any given technology. The curve parameter related to the steepness of the intelligibility transition, λ, also remained relatively consistent, except for P25 trunked Phase 1. Table 2 shows that encryption has the greatest impact on the intelligibility curve midpoint, t_0, which increases. This, combined with the other curve fit parameters remaining unchanged, demonstrates that access delay is increased by a relatively consistent offset when encryption is used. For most P25 technologies, encryption added approximately 30-50 ms of delay.

This offset can be seen for most α levels. α is an intelligibility scaling factor used in the definition of access delay [1]. Figure 1 shows the visual comparison for the P25 trunked Phase 1 system. For a given α level, the access delay is larger when encrypted communication mode is on. Figure 2 shows the same information represented in Fig. 1, except access delay is shown as a function of the raw intelligibility of the clip from ABC-MRT, rather than relative intelligibility based on α values.

Table 2. Curve parameter results with 95% confidence intervals included in parentheses. I_0 describes the asymptotic intelligibility of the system for the words under test. t_0 describes the time at which intelligibility is 50% of its asymptotic level. And λ describes the inverse of the slope of the intelligibility curve at time t_0.

	I_0	t_0 [s]	λ [s^{-1}]
P25 Direct Encrypted	0.916, (0.914, 0.917)	0.069, (0.065, 0.073)	−0.041, (−0.045, −0.037)
P25 Direct Unencrypted	0.915, (0.914, 0.916)	0.025, (0.021, 0.029)	−0.040, (−0.044, −0.036)
P25 Trunked Phase 1 Encrypted	0.917, (0.916, 0.918)	0.569, (0.566, 0.572)	−0.050, (−0.053, −0.047)
P25 Trunked Phase 1 Unencrypted	0.916, (0.914, 0.917)	0.528, (0.525, 0.531)	−0.045, (−0.048, −0.043)
P25 Trunked Phase 2 Encrypted	0.903, (0.901, 0.905)	0.573, (0.570, 0.575)	−0.046, (−0.049, −0.044)
P25 Trunked Phase 2 Unencrypted	0.901, (0.899, 0.903)	0.525, (0.523, 0.527)	−0.046, (−0.049, −0.044)
LTE	0.996, (0.995, 0.997)	0.033, (0.031, 0.036)	−0.040, (−0.042, −0.038)

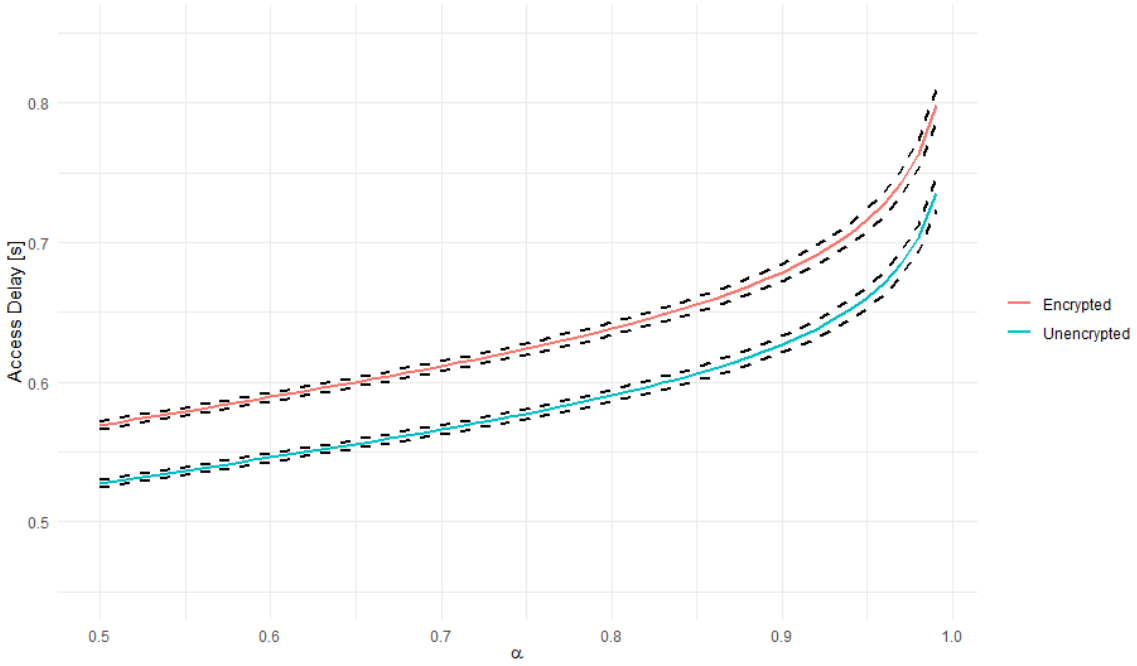

Fig. 1. P25 trunked Phase 1 access delay as a function of α. The dashed black lines represent the 95% confidence intervals.

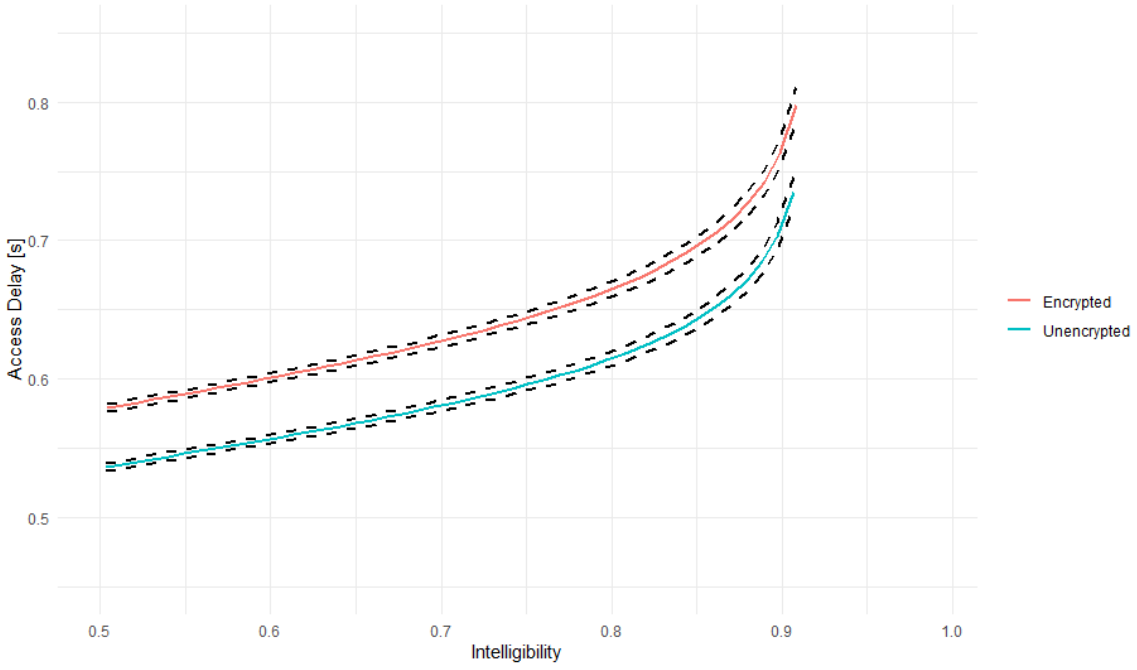

Fig. 2. P25 trunked Phase 1 access delay as a function of raw intelligibility. The dashed black lines represent the 95% confidence intervals.

4. Example Measurements of LTE Communications

LTE MCPTT technology is beginning to emerge on the market. Preliminary measurements were taken using early stages of MCPTT applications on Android devices. All measurements were performed using an RF enclosure. As previously discussed, Table 2 shows the curve parameters for LTE testing; asymptotic intelligibility was very high. In Fig. 3, access delay is shown as a function of α and in Fig. 4 as a function of intelligibility. Note that because the asymptotic intelligibility was nearly 1, these two plots are almost identical. The results for t_0 and λ were most similar to those of unencrypted P25 Direct. For a raw intelligibility of 85%, access delay was measured as 103.9 ± 4.1 ms with a level of confidence of 95%.

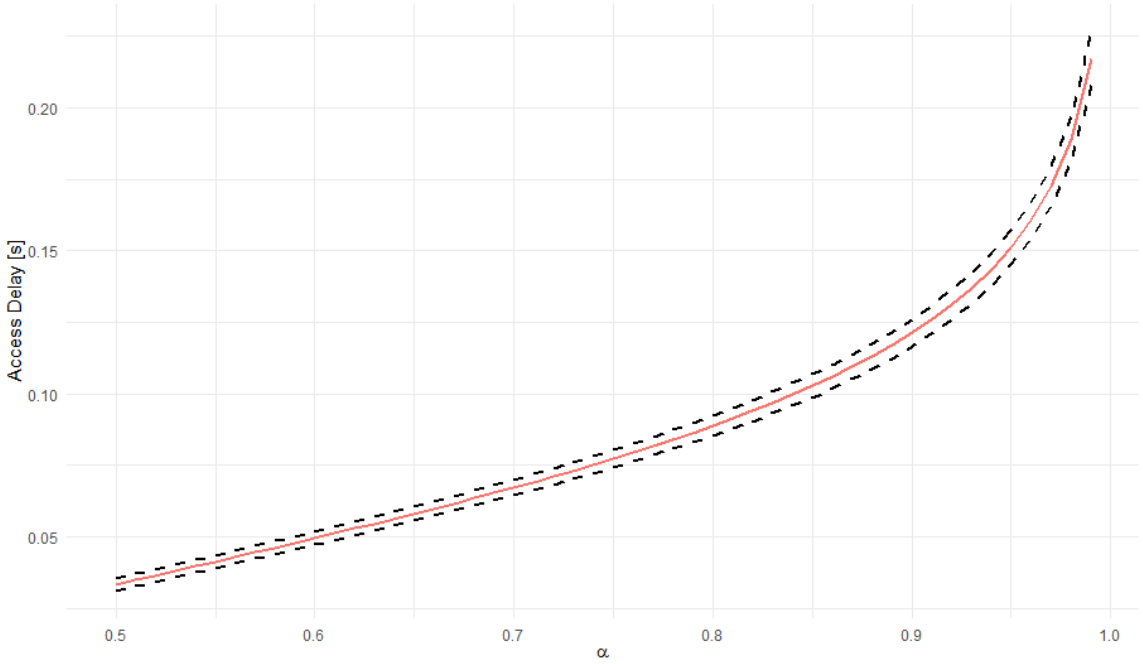

Fig. 3. LTE access delay curve as a function of α. The dashed black lines represent the 95% confidence intervals.

5. Conclusion

This paper describes several improvements made to the access time measurement system [1]. In particular, the audio alignment stage of the measurement process was improved to be more robust to specific impairments observed with low-rate vocoders. Further, the measurement system was made more robust and user-friendly for testing of prototype and in-development PTT applications, which are more prone to application failure in their early development stages. This paper also demonstrates the measurement system can successfully measure encrypted P25 and LTE communication systems. It is shown that encryption

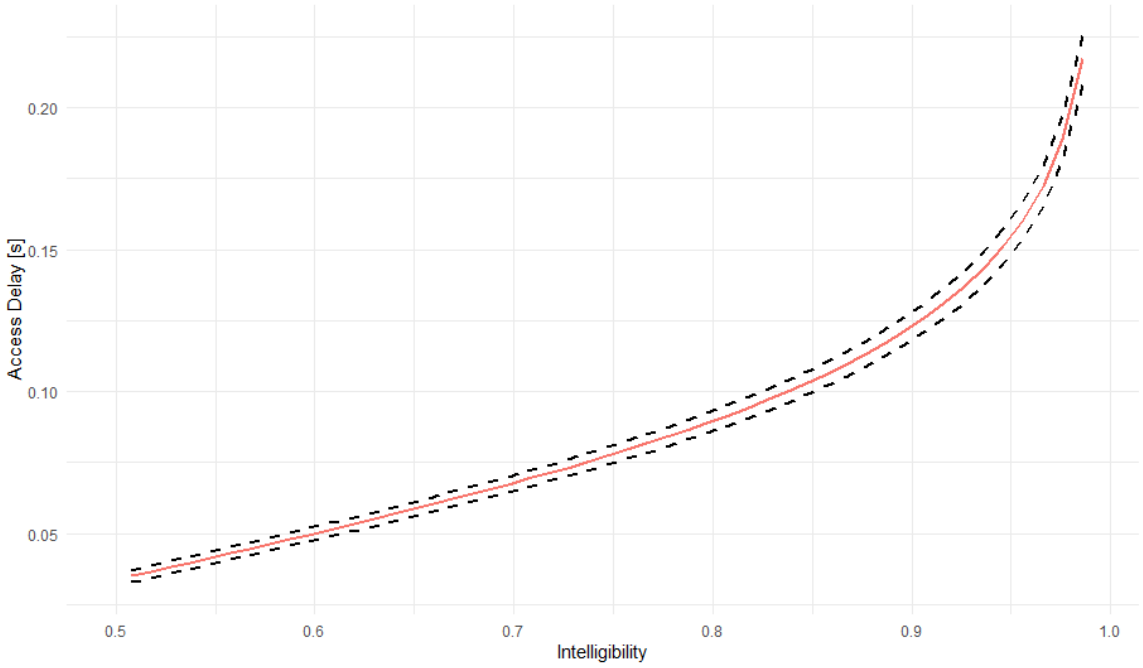

Fig. 4. LTE access delay as a function of raw intelligibility. The dashed black lines represent the 95% confidence intervals.

adds an almost constant amount of access delay to a given communication system. Finally, the measurement system was also able to measure access delay of LTE devices, despite the early stages of development of prototype LTE MCPTT applications.

5.1 Future Work

5.1.1 Curve Fitting Improvements

Limitations were identified with the current procedure for fitting curves to measured intelligibility data and we are currently working on a new and improved curve fitting regime. In particular, we are making the access delay measurements less dependent on characteristics specific to audio clip structure rather than MRT keywords themselves. Access delay currently depends on a subjective determination of when an MRT keyword is spoken within a full MRT phrase. This can potentially add bias to an access delay measurement and cause underestimates of true system access delay. This work is currently ongoing.

5.1.2 Access Focused Intelligibility Research

We also have recognized the need for truth data on the sort of impairments seen in access delay testing. In particular, we are interested in collecting data describing the subjective intelligibility response to partially muted words. Work is currently underway to characterize this response with new MRT intelligibility testing.

5.1.3 Packetized Voice Impairments

Loss, pause, and jump impairments are common in packetized voice and are known to impact speech quality measurements, as described in Ref. [8]. While analyzing LTE device data, a handful of trials with low P_2 intelligibility were noted. The waveforms of such trial recordings contained a gap of varying length in the middle of P_2. Figure 5 shows examples of these impairments: the top waveform shows a trial exhibiting typical behavior, the middle figure contains a pause impairment in a filler word, and the bottom figure shows an example of a pause impairment with P_2 which causes a low estimated intelligibility score. At the time of writing, detecting these pause impairments relies on detecting trials with suboptimal P_2 intelligibility values and reviewing the associated waveform. Further investigation is required to detect, identify, and handle this impairment within the MCV measurement system.

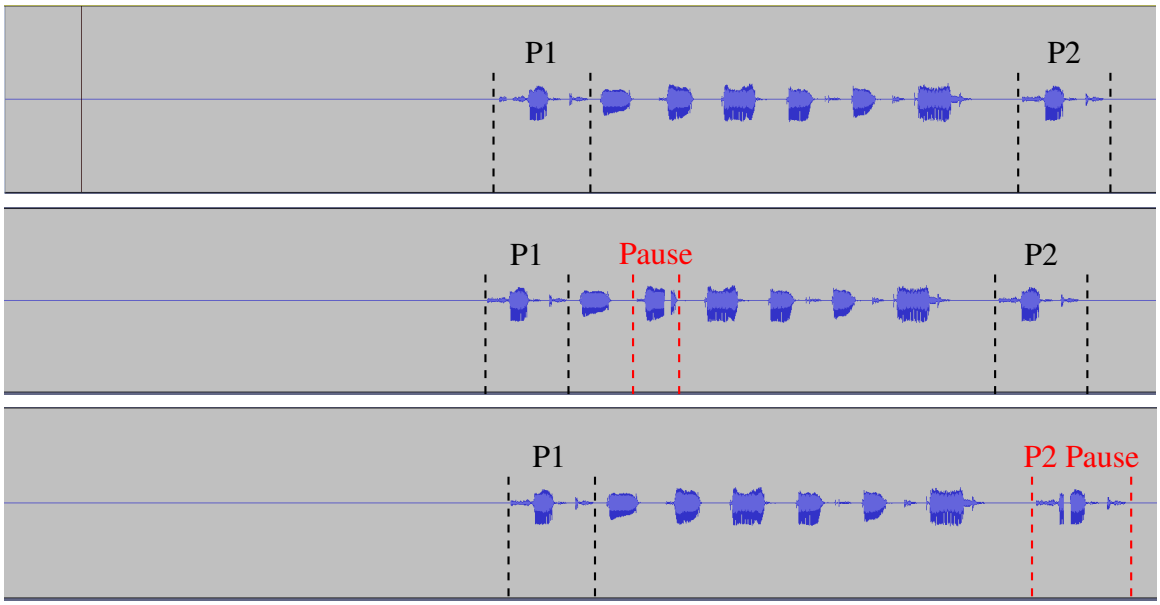

Fig. 5. LTE trial recording waveforms showing pause impairments

References

[1] Pieper J, Frey J, Greene C, Soetan Z, Thompson T, Voran S, Bradshaw D (2019) Mission Critical Voice QoE Access Time Measurement Methods (NIST), IR-8275. doi: 10.6028/NIST.IR.8275

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

[3] ANSI/ASA (2009) ANSI/ASA S3.2-2009 Method for Measuring the Intelligibility of Speech over Communication Systems.

[4] Voran SD (2017) A multiple bandwidth objective speech intelligibility estimator based on articulation index band correlations and attention. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 5100–5104. doi: 10.1109/ICASSP.2017.7953128

[5] Voran SD (2004) A bottom-up algorithm for estimating time-varying delays in coded speech. *Proceedings of the 3rd International Conference on Measurement of Speech and Audio Quality in Networks* .

[6] Fletcher H, Munson WA (1933) Loudness, its definition, measurement and calculation. *The Bell System Technical Journal* 12(4):377–430. doi: 10.1002/j.1538-7305.1933.tb00403.x

[7] (2020) Android Debug Bridge (ADB). URL https://developer.android.com/studio/command-line/adb.

[8] Voran SD (2003) Perception of temporal discontinuity impairments in coded speech – a proposal for objective estimators and some subjective test results. *Proceedings of the 2nd International Conference on Measurement of Speech and Audio Quality in Networks, Prague, Czech Republic, May 2003* .