Prepared for:

Federal Communications Commission

CMS Alliance to Modernize Healthcare
Federally Funded Research and Development Center

Task Order No. FCC15D0002

Internet Protocol Caption Telephone Service (IP CTS) – Summary of Phase 2 Usability Testing Results

Version 0.5

March 23, 2016

Record of Changes

Version	Date	Author / Owner	Description of Change
0.1	November 1, 2015	CAMH	Draft of Phase 2 results summary
0.2	November 24, 2015	CAMH	Complete draft of Phase 2 Report
0.3	December 15, 2015	CAMH	Draft of Phase 2 Report for Sponsor Review
0.4	December 30, 2015	CAMH	Response to Sponsor Review of Draft Phase 2 Report (with comments)
0.5	March 23, 2016	CAMH	De-identified Provider and Vendor names and references to prepare for public dissemination

About the CMS Alliance to Modernize Healthcare

The Centers for Medicare & Medicaid Services (CMS) sponsors the CMS Alliance to Modernize Healthcare (CAMH), the first Federally Funded Research and Development Center (FFRDC) dedicated to strengthening our nation's healthcare system.

The CAMH FFRDC enables CMS, the Department of Health and Human Services (HHS), and other government entities to access unbiased research, advice, guidance, and analysis to solve complex business, policy, technology, and operational challenges in health mission areas. The FFRDC objectively analyzes long-term health system problems, addresses complex technical questions, and generates creative and cost-effective solutions in strategic areas such as quality of care, new payment models, and business transformation.

Formally established under Federal Acquisition Regulation (FAR) Part 35.017, FFRDCs meet special, long-term research and development needs integral to the mission of the sponsoring agency—work that existing in-house or commercial contractor resources cannot fulfill as effectively. FFRDCs operate in the public interest, free from conflicts of interest, and are managed and/or administered by not-for-profit organizations, universities, or industrial firms as separate operating units.

The CAMH FFRDC applies a combination of large-scale enterprise systems engineering and specialized health subject matter expertise to achieve the strategic objectives of CMS, HHS, and other government organizations charged with health-related missions. As a trusted, not-for-profit adviser, the CAMH FFRDC has access, beyond what is allowed in normal contractual relationships, to government and supplier data, including sensitive and proprietary data, and to employees and government facilities and equipment that support health missions.

CMS conducted a competitive acquisition in 2012 and awarded the CAMH FFRDC contract to The MITRE Corporation (MITRE). MITRE operates the CAMH FFRDC in partnership with CMS and HHS, and maintains a collaborative alliance of partners from nonprofits, academia, and industry. This alliance provides specialized expertise, health capabilities, and innovative solutions to transform delivery of the nation's healthcare services. Government organizations and other entities have ready access to this network of partners, including RAND Health, the Brookings Institution, and other leading healthcare organizations. This includes select qualified small and disadvantaged business.

The FFRDC is open to all CMS and HHS Operating Divisions and Staff Divisions. In addition, government entities outside of CMS and HHS can use the FFRDC with permission of CMS, CAMH's primary sponsor.

Executive Summary

The Federal Communications Commission (FCC) requested Centers for Medicare & Medicaid Services (CMS) Alliance to Modernize Healthcare (CAMH) Federally Funded Research and Development Center (FFRDC), operated by The MITRE Corporation (MITRE), to independently assess the quality metrics and associated usability of Internet Protocol Caption Telephone Service (IP CTS) devices and services, as well as alternative technologies that could augment IP CTS services. Given the substantial growth of IP CTS usage in the last two years and advances in speech to text and speech recognition technology, the FCC is interested in understanding whether new technologies or processes can provide improved IP CTS service while continuing to ensure that IP CTS services are appropriately available to individuals who need assistance to obtain equivalent access to telephony services.

This report presents consolidated test results from the Phase 2 usability assessment of alternative speech recognition technologies and provides qualitative and quantitative measures for device and caption performance. Phase 1 captured results from controlled user assessments and established a baseline of usability metrics based on CAMH's assessments of IP CTS devices and services. These activities provide qualitative and quantitative measures for device and caption performance.

What We Found

Phase 2 Usability Assessment and Device Testing Results

The Phase 2 usability assessment and device testing results demonstrate that there is at least one automatic Speech-To-Text (STT) engine that is equivalent or better than three of the four current IP CTS devices at producing an accurate, expedient, and usable service (see Figure ES-1). The current IP CTS providers, Provider 1, Provider 2, Provider 3, and Provider 4, are listed first through fourth, followed by the three STT engines tested in the study.

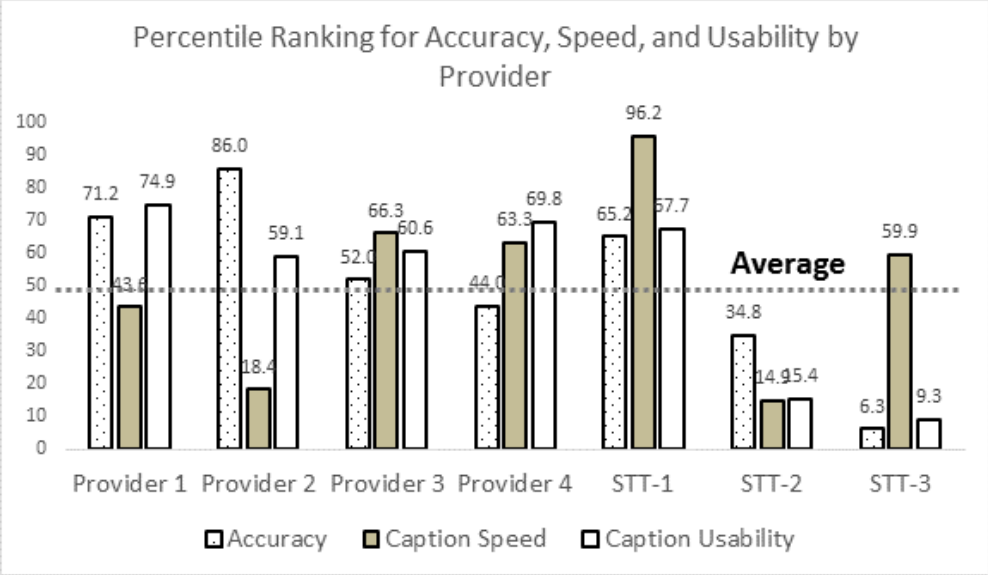

Figure ES-1. Percentile Ranking for Accuracy, Speed, and Usability by Provider

Provider 3 and STT-1 ranked above average in accuracy, speed, and usability (the ease of learning and using a new system). Provider 1 and Provider 2 both ranked above average in accuracy and usability, and below average on caption speed. Provider 4 ranked above average in speed and usability, and below average on accuracy. STT-2 ranked below average on all metrics. STT-3 ranked above average on speed and well below average on accuracy and usability.

Users reported that they preferred accuracy over speed as the more important variable to a successful calling experience, although the associated performance metrics were not directly tested in this study to validate their opinions. CAMH recommends additional testing to determine the point at which slower speeds negatively affect accuracy and where the user preference for accuracy is likely to change.

Recommendations

Based on the results from Phase 1 and Phase 2 Usability Assessment and Device Testing, CAMH recommends that the FCC consider:

1. Research the feasibility of using fully automated STT services in place of existing IP CTS services. At this time, there are no automated STT telephony services. This requires further usability testing to determine if automated STT system prototypes can provide (a) similar levels of usability and call effectiveness as experienced today by current IP CTS users, (b) an overall satisfactory calling experience similar to what is available today, and (c) viable options/alternatives to IP CTS services. This work will assist in developing minimum specifications and requirements for a fully functional automated system that may provide users more options for choosing their IP CTS services.
2. Research the feasibility of using a fully automated STT service in conjunction with an enhanced IP CTS service that uses a skills based Communication Agent (CA) in instances of complex conversation and detailed instruction. Develop and design a feature that allows for real-time, on-demand switching from an automated capability to CA-based services as needed by users. This work will allow users the option of including a third-party CA on the call.
3. Research and determine specifications for IP CTS caption speed and accuracy performance that are acceptable by users. Work with the Telecommunications Relay Services Center of Expertise on the design of the study. This work will assist in determining minimum specifications of accuracy and delay. This information may be used to develop minimum guidelines that all services must meet to operate as a CTS.
4. Continue working with the hard of hearing community to identify areas for service improvements. Engaging and addressing the needs of the community will assist with a smooth transition to future technologies.
5. Establish an independent quality assurance group to test transcription quality. This group will ensure that service providers are meeting minimum CTS standards.
6. Request that IP CTS providers include speakerphone and headphone capabilities in devices. These functions may assist hard-of-hearing individuals gain greater understanding of the conversation.

7. Request that IP CTS providers include visual or tactile feedback for all functions that currently use audio feedback (e.g., dial tone or silence indicators, button pressed indicators). These functions may assist individuals who are deaf or have severe hearing loss with device operation.

Table of Contents

List of Figures

List of Tables

1. Introduction

Hearing loss is an invisible but significant barrier in daily life, including telephone conversations. According to a 2012 study by the United States Census Bureau, 7.6 million (3.1%) of the U.S. population experienced a hearing difficulty—defined as experiencing deafness or having difficulty hearing a normal conversation, even when wearing a hearing aid. Severe hearing loss affects 2.2 million or 0.8% of the population[1]. For those 65 and older, 4.2 million (10.8%) experienced a hearing difficulty, including 1.7 million (4.3%) who reported a severe difficulty hearing. As the U.S. population ages, the number of individuals with hearing loss is projected to increase significantly.

Internet Protocol Caption Telephone Services (IP CTS) is a telecommunications relay service for an individual who can speak, but who has difficulty hearing over the telephone. An individual can use a telephone and an Internet Protocol-enabled device to listen to the other party and simultaneously read captions of the other party's words.

The Federal Communications Commission (FCC) requested the Centers for Medicare & Medicaid Services (CMS) Alliance to Modernize Healthcare (CAMH) Federally Funded Research and Development Center (FFRDC), operated by The MITRE Corporation (MITRE), to independently assess the quality and usability of IP CTS devices and services as well as alternative technologies that could be used in place of IP CTS.

1.1 Document Organization

This document is organized as follows:

Section		Purpose
Section 2:	Study Goals and Assessments for Phase 2	Section 2 provides a description of the tasks performed by CAMH to assess quality and usability of IP CTS devices and automatic STT engines.
Section 3:	Usability Assessment Results	Section 3 provides CAMH's findings from the Phase 2 Usability Assessment of IP CTS devices and automatic STT engines.
Section 4:	Device Performance Testing and Usability Results	Section 4 provides combined results from CAMH's findings from the Phase 2 Device Testing of IP CTS devices and automatic STT engines and Phase 2 Usability Assessment.
Section 5:	Recommendations	Section 5 provides CAMH recommendations for improvements to the FCC IP CTS program.
Appendix A:	Device Testing Quality Metrics and Results	Appendix A provides a description of the Device Testing Quality Metrics, overall results for all scenarios, and detailed results for Phase 2 scenarios.
Appendix B:	Usability Assessment Design and Analysis	Appendix B provides a description of the Phase 2 usability assessment design and analysis.

[1] M. Brault. (2012). "Americans with Disabilities: Household Economic Studies." U.S. Census Bureau. http://www.census.gov/prod/2012pubs/p70-131.pdf

Section	Purpose
Appendix C: Test Call Transcripts	Appendix C provides the scenario transcripts used during the Phase 2 Usability Assessment.
Appendix D: Usability Assessment Questionnaire	Appendix D provides the questionnaires used during the Usability Assessment.
Acronym List	Defines the acronyms used in this document

2. Study Goals and Assessments for Phase 2

The objective of the independent assessment conducted in Phase 2 was to provide the FCC with recommendations on the viability of alternative speech recognition technologies for use in IP CTS environments from the user's perspective via usability feedback and comprehension scoring. The usability assessment and device test results will help inform policy changes regarding performance, quality, and contextually relevant standards for IP CTS service providers.

CAMH and the FCC intend that these activities will help the FCC and other organizations determine if automatic speech to text technologies provide similar or better levels of usability and performance characteristics to current IP CTS services.

2.1 Overall IP CTS Testing Approach

The overall IP CTS study consisted of two phases. The summary results presented in this document address specific findings for Phase 2 Usability Assessment and Device Testing. The phases of testing were as follows:

Phase 1 – Baseline of IP CTS (see Appendix A)

Internet Protocol Caption Telephone Service (IP CTS) Devices: Initial Baseline Test Results

- Established appropriate performance measures and quality standards for IP CTS. The data included capturing baseline technical measures with a focus on accuracy, latency, and completeness.

Baseline Usability Assessment of IP CTS with Users

- Utilized performance measures and quality standards from Part 1 and collected usability feedback, comparing objective performance measures of IP CTS and feedback from users.

Utility of the Service (Survey of Users)

- Attitudinal survey of the hard of hearing community regarding IP CTS use and areas for potential improvement with a focus on usability and user experience.
- Identified and categorized user demographics.

Phase 2 – Testing of Existing Alternative Technologies and Approaches

Usability Assessment (controlled User Testing) of commercially available alternatives to IP CTS Equipment

- Determine if automatic speech to text technologies provide similar or better levels of usability and performance characteristics to current IP CTS services

2.2 Usability Assessment

CAMH conducted a controlled user assessment to compare usability metrics of IP CTS devices and services to automated Speech-to-text (STT) engines. The testing environment included currently available IP CTS systems—four commercial FCC approved providers and devices (IP CTS devices)—and three STT engines. The four providers have been de-identified in this report intended for public dissemination and are noted as Provider 1, Provider 2, Provider 3, and Provider 4. The three STTs are indicated as STT-1, STT-2, and STT-3 in the results summaries.[2] STT-1 and STT-3 are dictation services and not intended for use during telephone calls. MITRE configured the STT engines for testing during telephone calls for the IP CTS usability assessment.

CAMH performed a priori analyses to determine the minimum number of participants required for statistical significance during the Phase 2 Usability Assessment (see Appendix B.7 for a priori analyses). These analyses used the participant usability ratings from the Phase 1 Baseline Usability Assessment, wherein four devices were tested with 5 participants completing six scenarios (120 cases). The analyses determined a minimum of ten participants were required to obtain 120 cases (10 ratings per device) to achieve the required statistical effect size for comparison of device usability ratings. The Phase 2 Usability Assessment included eleven participants for 154 cases.

Participants in the video-captured study included eleven hard of hearing or deaf users who completed the two (2)-hour usability assessment (see Appendix B for Usability Assessment Design and Analysis). Participants watched a pre-recorded video of captions from an audio transcript explaining why a particular teacher should be chosen as speaker (the "Ms. Jackson Transcript"). They also completed a phone call on each device to the Interactive Voice Response (IVR) system that required them to complete a banking transaction (see Appendix C for Test Call Transcripts). After each test case (call and video), participants completed questionnaires and answered open interview questions about their experience (see Appendix D for Usability Assessment Questionnaires).

2.2.1 Usability Metrics

The metrics used in the Usability Assessment were similar to those used in Phase 1, with the addition of the System Usability Scale (SUS).[3] CAMH captured participants' feedback on each IP CTS device and STT engine to identify relationships between system performance and usability feedback.

After each test case, participants filled out a questionnaire based on the scenario and information provided by the captions. The purpose of this questionnaire helped capture user comprehension and caption accuracy. In addition, CAMH queried the participants on the following usability dimensions:

[2] The providers are not identified to avoid any appearance of endorsement or partiality toward a particular STT service.

[3] Brooke, J. (1996). "SUS: a "quick and dirty" usability scale". In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland. Usability Evaluation in Industry. London: Taylor and Francis.

- **Caption Usability –** Overall Caption Usability was the computed average of three metrics
  - **Effectiveness –** The degree to which a user continuously maintains full context and meaning of the conversation.
  - **Efficiency –** The degree to which a user is comfortable with the perceived tempo of the call.
  - **Satisfaction –** The degree to which the user is satisfied with the calling experience, given the test case and use of transcript.
- **System Usability Scale –** This metric focuses on the device's user interface, layout, design, and functions only.
  - **System Usability –** The usability of each device.
  - **Learnability –** The ease of learning to use each device.
- **User Demographics and Preferences –** User's severity of hearing loss, use of IP CTS devices, and preferences in regards to speed, accuracy, and caption agent assistance in calls.

2.3 IP CTS Device Testing

CAMH conducted independent testing of IP CTS devices and services in a controlled environment. This baseline testing provides quantitative measures for key performance characteristics of the IP CTS devices and STT engines (see Appendix A for Device Testing Quality Metrics and Results). The baseline testing does not provide pass/fail criteria or identify acceptable quality standards for the IP CTS service.

2.3.1 Device Testing Metrics

CAMH used the following metrics to compare device and service-related quality to usability ratings:

- **Accuracy –** The percentage of words from the conversation (the IP CTS transcription) that are correctly transcribed on the device screen.
- **Caption Delay (Latency) –** The time elapsed between hearing a voice on the caption phone and the display of captions on the phone's screen.

3. Usability Assessment Results

CAMH collected usability ratings from participants on four IP CTS devices and three automated STT engines (Table 1 has been redacted to de-identify the providers' endpoint devices used in the tests). The objective of this phase was to collect participant usability information on all devices.

Table 1. Usability Assessment IP CTS Devices

Provider	Endpoints
Provider 1	Redacted
Provider 2	Redacted
Provider 3	Redacted
Provider 4	Redacted
STT-1	Windows 8
STT-2	Windows 8
STT-3	Windows 8

CAMH performed the usability assessment in a controlled environment with pre-recorded videos of audio transcribed through each device and the bank IVR utilized in Phase 1 testing.

CAMH confirmed with the IP CTS providers that the setup and initialization of lab equipment was comparable to equipment fielded in users' homes and businesses. During preliminary interviews, Provider 1 indicated that the specific device type under test should have no impact on the captioning service.

Since many STT's available online are touted as 'learning' systems, CAMH ensured the STTs used during the assessment were not learning the test cases by prohibiting user profile updates intended on improving system accuracy and comparing the progression of accuracy between the first and last test cases. Statistical comparisons of the usability and device metrics found that there were no differences between the performance of the first and last test cases for any of the STTs.

3.1 Results

CAMH found that STT-1 scored consistently better than the other STT engines. The objective of this study was to investigate the viability of alternative speech recognition technologies. Due to STT-1's high scoring compared to the other STT engines, CAMH reports its score separately to demonstrate STT-1's comparable performance to current IP CTS devices. Section 3 results show combined averages for IP CTS devices and STT-2/3; STT-1 is unique. Appendix B presents scores for all devices with statistical analysis and results. Section 4 reports percentile rankings discretely at the system level.

3.2 In-Session Comprehension

The in-session comprehension was determined through participants' self-reported answers to scenario-specific information provided by the device captions (see Appendix D). Participants answered three context-specific questions based on the captions presented during the video and call scenarios. Participants were asked to record the answer(s) as presented by the captions. Cut-off points for acceptability are based on participant feedback regarding their ability to comprehend the meaning of a text and previous research.[4] Scores greater than 75% are "acceptable," where participants are confident in their understanding of a text and find the text usable. Scores between 55% and 75% are considered "marginal," indicating participants are able to garner the meaning of the text, but are missing key points. Scores below 55% are "not acceptable," demonstrating participants are confused by the transcript and have difficulty following the meaning of the text. CAMH measured comprehension based on the participant's ability to complete the appropriate information (see Figure 1), such as providing an account balance during the bank scenario.

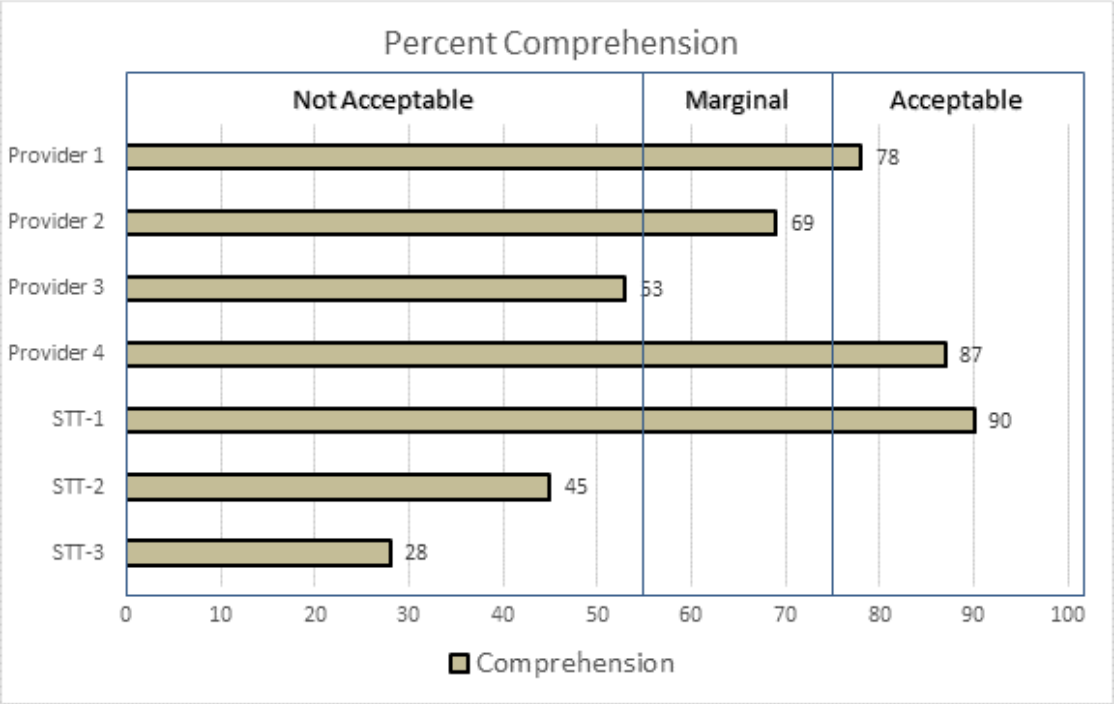

Figure 1. In-Session Comprehension Scores as Shown in Percent by Provider

Comprehension scores are reported as combined averages for IP CTS providers, STT-1, STT-2, and STT-3 (see Table 2). For device-specific statistical results, please see Appendix B.

[4] C. Munteanu, Baecker, R., Penn, G., Toms, E., and James, D. (2006). "The Effect of Speech Recognition Accuracy Rates on the Usefulness and Usability of Webcast Archives." CHI 2006 Proceedings Visualization and Search. Montreal, Quebec, Canada. April, 2006.

Table 2. Percent of Accurate Responses between IP CTS Devices and STT

Provider	Percent Accurate
IP CTS Devices	72
STT-1	90
STT-2 / 3	37

The data show that users were generally able to understand the meaning of the captions when using an IP CTS Device or STT-1. When using STT-2 and STT-3, users were often unable to comprehend the meaning of the captions. STT-1 scored the highest in comprehension; however, there were no significant differences between its score and three of the IP CTS device scores. STT-2 and STT-3 were significantly worse than all IP CTS devices and STT-1. This suggests, for this assessment, that when using STT-1, users' comprehension during calls is as effective for understanding a conversation when compared to current IP CTS devices.

3.3 Usability Scores

This subsection presents a summary of usability scores based on 11 participants in the Phase 2 usability assessment. Participants completed questionnaires after each session (see Appendix D).

3.3.1 Caption Usability

Figure 2 illustrates the caption usability scores for each IP CTS provider and the STT engines. Scores are based on a scale ranging from 1 – 7, where 1 equals Strongly Disagree, 4 equals Neutral, and 7 equals Strongly Agree. After completing a call or viewing a video, participants rated the usability of the captions and their overall satisfaction with the test case experience. Caption usability is based on the user's ranking of the caption effectiveness, efficiency, and how satisfied the user was with the captions. Higher scores indicate a higher degree of usability. Cut-off points for acceptability are based on the scale used during the usability questions. Ratings of 5 through 7 indicate a high level of acceptability. Ratings of 3 through 5 indicate a marginal or neutral acceptability. Ratings below 3 indicate unacceptable usability (see Figure 2).

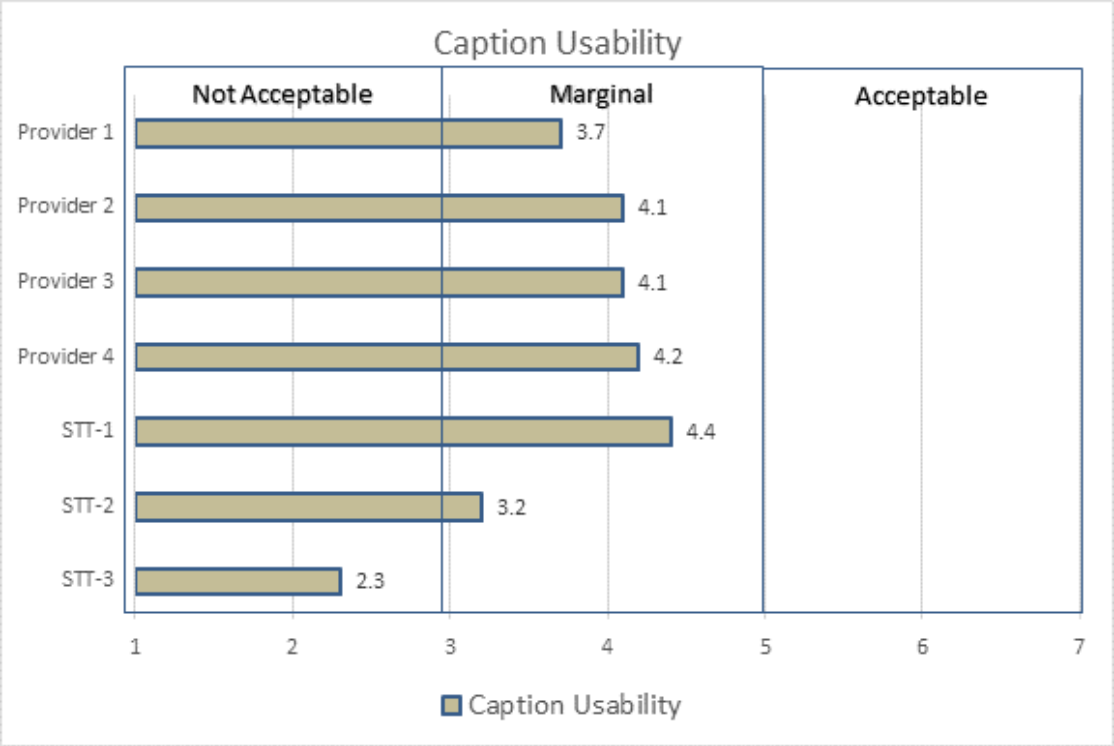

Figure 2. Caption Usability by Provider

Caption usability scores are reported as combined averages for IP CTS providers, STT-1, and STT-2 and STT-3 (see Table 3). For device-specific statistical results, please see Appendix B.

Table 3. Average Caption Usability Scores between IP CTS Devices and STT

Provider	Effectiveness	Efficiency	Satisfaction	Usability
IP CTS Devices	4.7	3.1	4.3	4.0
STT-1	4.7	3.9	4.5	4.4
STT-2 / 3	2.6	3.1	2.5	2.7

The data show that users generally do not believe that captions have a high degree of usability. The average caption usability scores indicate that most users consider the quality of the captions as marginally effective, efficient, and satisfactory. The highest caption usability score was for STT-1 (4.4) followed by IP CTS devices (4.0). STT-2/3 (2.7) scored significantly lower than STT-1 and the IP CTS devices. Although caption usability should be improved for all systems, these results suggest that caption usability of STT-1 is comparable to IP CTS devices.

3.3.2 System Usability

System usability relates to the functionality of the device or system. This metric measures a system's ease of use and users' ability to learn the system. The system usability is not intended to measure the quality of captions. Figure 3 illustrates the system usability scores for each IP CTS

provider and the STT engines. Scores are based on a 0–100 scale where higher scores indicate a higher degree of device usability and learnability. Cut-off points for acceptability are based on previous research.[5] Scores greater than 70% are "acceptable," where participants are confident in their ability to use a system. Scores between 50% and 70% are considered "marginal," where devices have some usability issues that should be of concern. Scores below 50% are "not acceptable," demonstrating that participants find the system overly complex and not user friendly (see Figure 3).

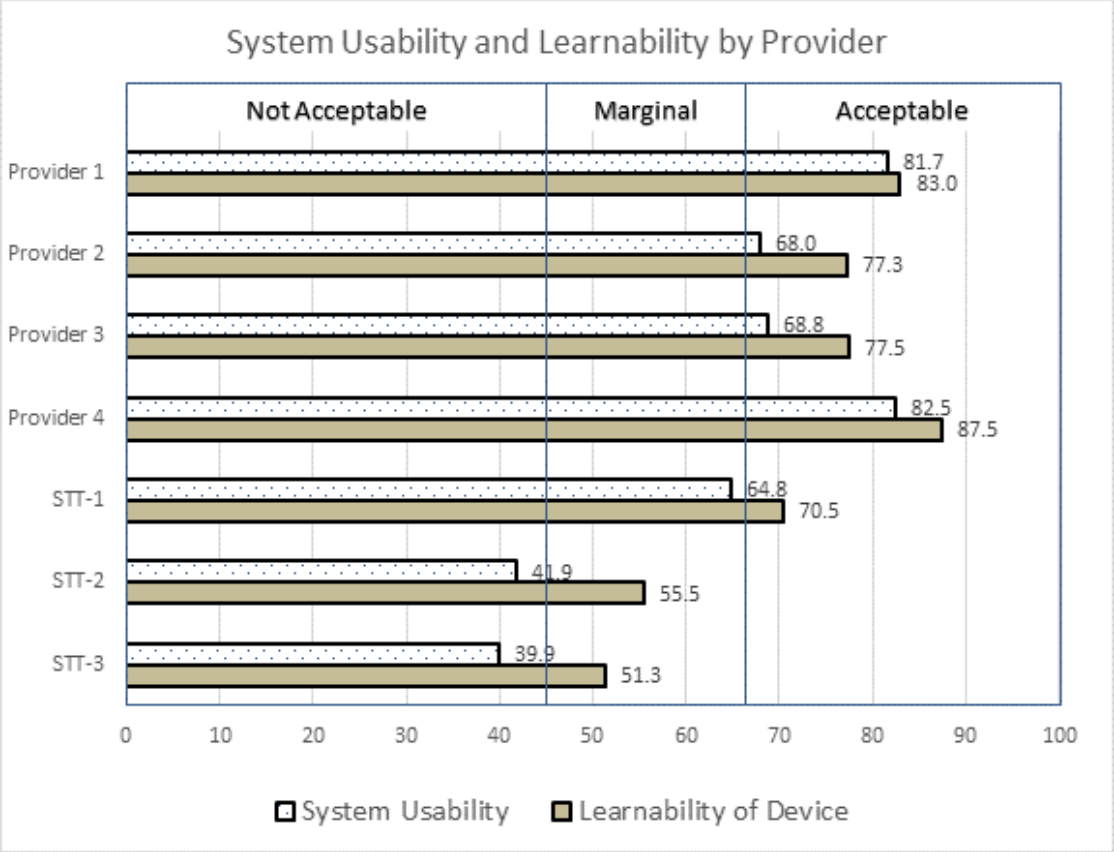

Figure 3. System Usability and Learnability by Provider

System usability scores are reported as combined averages for IP CTS providers, STT-1, and STT-2 and STT-3 (see

[5] A. Bangor, Kortum, P., and Miller, J. (2009). "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale." Journal of Usability Studies, 4-3, pp. 114-123.

Table 4). For device- specific statistical results, please see Appendix B.

Table 4. Average System Usability Scores between IP CTS Devices and STT

Provider	System Usability	Learnability of Device
IP CTS Devices	75.4	81.4
STT-1	64.8	70.5
STT-2 / 3	40.8	53.3

The system usability results indicate that users felt the IP CTS devices (75.4) and STT-1 (64.8) were marginal or better in terms of system usability. There were no significant differences between the IP CTS devices and STT-1. STT-2/3 rated significantly lower in system usability when compared to the IP CTS devices and STT-1, with a system usability score of 40.8. Similar results were found with device learnability. Users felt that IP CTS devices (81.4) were significantly easier to learn than STT-2/3 (53.3). There were no significant differences between learnability for the STT devices.

Further examination of the usability metrics between the IP CTS devices and STT-1 demonstrate that there are no statistically significant differences for any of the usability metrics. This suggests that the system usability of STT-1 is essentially the same as that experienced with an IP CTS device.

3.4 User Preferences

A total of 11 participants took part in the Phase 2 usability assessment. Of those, 82 percent reported the severity of their hearing loss as moderately severe or worse and 81 percent currently use IP CTS.

In terms of caption characteristics collected via post-test questionnaire, 64 percent of participants reported that speed was very important or greater, and 73 percent reported that accuracy was very important or greater. All participants stated that if they could only improve one characteristic, they would choose accuracy over speed. Participants stated that they were willing to accept some delay if it assured more accurate transcripts. In addition, participants expressed that it is more important to receive the correct information the first time and reduce the need to repeat information during the call. All participants would prefer to have a Communication Agent on IP CTS calls because they believe this will improve accuracy. While users believe that CAs provide better accuracy and prefer to have a CA on the call, current findings do not support the premise that accuracy is improved by having a CA on the call. It should be noted that the speed of captions impacts the user's ability to comprehend a conversation, and a 4-second delay in captions can affect comprehension.[6] It is expected that there is a tradeoff between accurate transcription and the speed at which the captions appear. Further research is required to determine the cutoff points for accuracy and speed.

[6] D. Burnham, J. Robert-Ribes, and R. Ellison. (1998). "Why Captions Have to be on Time." Auditory-Visual Speech Processing (ASVP 98). December 4-6, 1998. Sydney, Australia. http://www.isca-speech.org/archive_open/archive_papers/avsp98/av98_153.pdf

4. Device Performance and Usability Comparison

CAMH used metrics from the device testing and usability assessment to compute percentile rankings for accuracy, caption transcription delay, and caption usability (see Table 6). Percentile ranking allows for a clear comparison across providers by assessing scores against a normal curve. The percentile rank indicates the percentage of scores that occur at or below the given score. For example, Provider 4 ranked in the 70th percentile for accuracy (see Table 6), thus indicating 70 percent of the scores were at or below the average percent accuracy across all providers. (An average ranking does not necessarily mean good.) Caption transcription delay was reverse scored to provide a higher percentile ranking for shorter transcription delay (i.e., a higher percentile indicates faster captioning). CAMH renamed this metric as "caption speed" in Table 5 and the following paragraph.

Table 5. Device Testing Percentile Ranking by Provider for Bank Test Calls

Provider	Accuracy (Percent)	Accuracy Percentile	Transcription Delay (Seconds)	Caption Speed Percentile	Caption Usability (Score)	Caption Usability Percentile
Provider 1	80.8	71.2	5.1	43.6	5.2	74.9
Provider 2	88.8	86.0	6.6	18.4	4.5	59.1
Provider 3	73.0	52.0	3.9	66.3	4.5	60.6
Provider 4	70.0	44.0	5.4	63.3	4.9	69.8
STT-1	78.2	65.2	1.1	96.2	4.8	67.7
STT-2	66.3	34.8	6.9	14.9	2.5	15.4
STT-3	48.9	6.3	4.2	59.9	2.0	9.3

Figure 4 summarizes the percentile rankings for accuracy, caption speed, and caption usability. Provider 3 and STT-1 ranked above average in accuracy, speed, and usability. Provider 1 and Provider 2 both ranked above average in accuracy and usability, but ranked below average on caption speed. Provider 4 ranked above average in speed and usability, and ranked below average on accuracy. STT-2 ranked below average on all metrics. STT-3 ranked above average on speed, and well below average on accuracy and usability. These results demonstrate that STT-1 performed better than STT-2 and STT-3 for accuracy; above average, as compared to all providers, for caption speed; and above average for usability. These rankings demonstrate that there is at least one automatic STT engine that is equivalent to or better than current IP CTS devices at producing an accurate, expedient, and usable service.

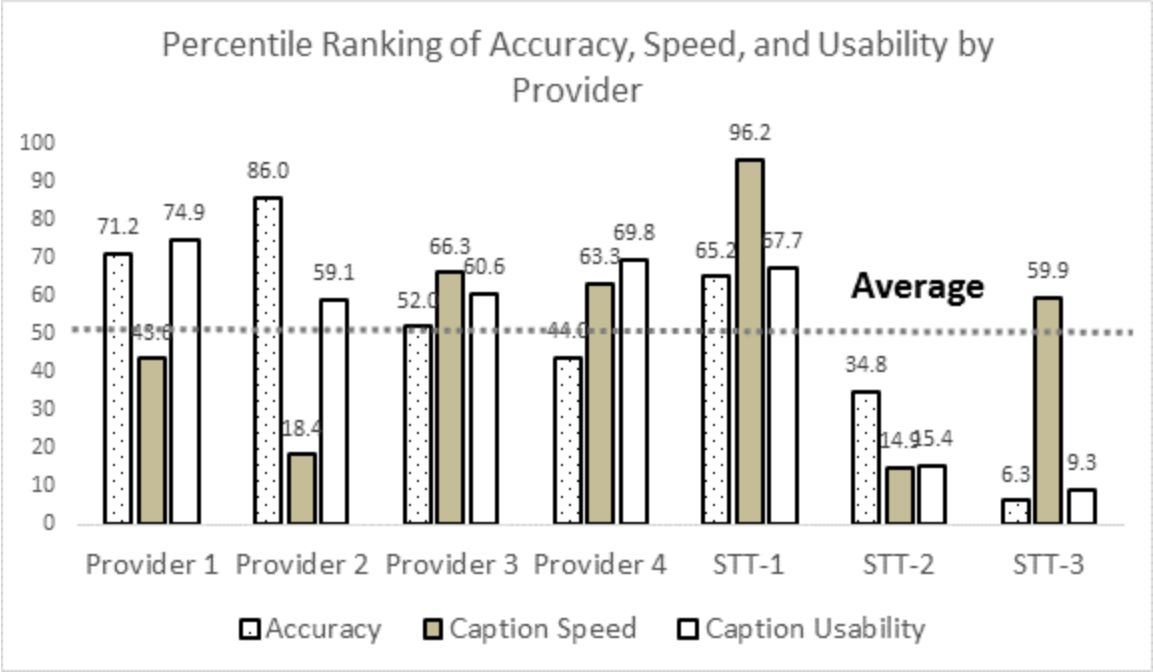

Figure 4. Percentile Ranking of Accuracy, Speed, and Usability by Provider

5. Recommendations

Based on the results from Phase 1 and Phase 2 Usability Assessment and Device Testing, CAMH recommends that the FCC:

1. Research the feasibility of using fully automated STT services in place of existing IP CTS services. At this time, there are no automated STT telephony services. This requires further usability testing to determine if automated STT system prototypes can provide (a) similar levels of usability and call effectiveness as experienced today by current IP CTS users, (b) an overall satisfactory calling experience similar to what is available today, and (c) viable options/alternatives to IP CTS services. This work will assist in developing minimum specifications and requirements for a fully functional automated system that may provide users more options for choosing their IP CTS services.
2. Research the feasibility of using a fully automated STT service in conjunction with an enhanced IP CTS service that uses a skills-based Communication Agent in instances of complex conversation and detailed instructions. Develop and design a feature that allows for real-time, on-demand switching from an automated capability to CA- based services as needed by users. This work will allow users the option of including a third-party CA on the call.
3. Research and determine specifications for IP CTS caption speed and accuracy performance that are acceptable to users. Work with the Telecommunications Relay Services Center of Expertise on the study design. This work will assist in determining minimum specifications of accuracy and delay, and may develop minimum guidelines that all services must meet to operate as a CTS.
4. Continue working with the hard of hearing community to identify areas for service improvements. Engaging and addressing the needs of the community will assist with a smooth transition to future technologies.
5. Establish an independent quality assurance group to test transcription quality. This group will ensure that service providers are meeting minimum CTS standards.
6. Request that IP CTS providers include speakerphone and headphone capabilities in devices. These functions may help hard-of-hearing individuals gain greater understanding of the conversation.
7. Request that IP CTS providers include visual or tactile feedback for all functions that currently use audio feedback (e.g., dial tone or silence indicators and button pressed indicators). These functions may assist individuals who are deaf or have severe hearing loss with device operation.

Appendix A. Device Testing Quality Metrics and Results

A.1 IP TS Device Testing

CAMH conducted independent testing of IP CTS devices and services in a controlled environment. This baseline testing provides quantitative measures for key performance characteristics of the IP CTS service. The baseline testing does not provide pass/fail criteria or identify acceptable quality standards for the IP CTS service. These measures are used to support usability testing and may be used to identify changes in IP CTS service quality. Please see Internet Protocol Caption Telephone Service (IP CTS) Devices: Summary of Phase 1 Activities for the full Device Testing report.

A.2 Device Testing Metrics

1. **Time for a Communication Assistant (CA) to Connect**
    a. Definition: The elapsed time between a user's request for captions (e.g., by pressing the "Captions On" button during a call) and the display of text on the phone's screen, indicating the start of captioning services.
    b. How to measure: From the video recording of each call, CAMH analysts measure and record the elapsed time (in seconds) between the user's request for captions and the notification of CA assigned to the call. Generally, the notification includes a CA-specific number to indicate the call is being captioned. If this identifier is not displayed, CAMH will look for another textual notification of the captioning service.
2. **Accuracy**
    a. Definition: The percentage of words from the conversation (the IP CTS transcription) that are correctly transcribed on the Device under Test (DUT) screen.
    b. How to measure: Using the video recording of each call, MITRE analysts create a text file of the captions displayed on the DUT screen (the IP CTS transcription). CAMH uses the National Institute of Standards and Technology (NIST) SCLITE tool from the Speech Recognition Scoring Toolkit (SCTK)[7] to compare this file to a reference "true transcription". CAMH records the tool's scores and output.

    The SCLITE tool scores accuracy by comparing the total number of words transcribed correctly to the number of words in the reference file. Omitted words are reflected in both the accuracy and completeness metrics.

    There may be more than one correct transcription of spoken words to written words. To provide a consistent, realistic assessment of accuracy, CAMH has identified the following rules for assessing transcribed files:

    - Uppercase/Lowercase and punctuation are not considered in the accuracy calculations. All files are forced to uppercase. All punctuation is removed

[7] http://www.itl.nist.gov/iad/mig/tools/

prior to assessment except for hyphens ("-"), underscores ("_"), and slashes ("/").

- Contractions and expanded phrases are both valid ("that's" and "that is" are considered the same).
- Abbreviations that have spaces or periods between the letters are considered valid ("FCC", "F C C", and "F.C.C" are all considered the same).
- Numbers may be spelled out or numeric ("400", "four hundred", and "4 hundred" are all considered the same).
- Times may be represented as words or numbers ("8:30", "8 30", "830", and "eight thirty" are all considered the same).
- Hyphenated words, non-hyphenated words, and words separated by underscores are all valid ("thank you", "thank-you", and "thank_you" are all considered the same).
- Universal Resource Locators (URL) that contain extra spaces or spell the words "slash" or "dot" are valid ("fcc.gov/smartdevice", "fcc dot gov slash smart device", and "fcc . gov / smartdevice" are all considered valid).
- Disfluencies ("ah", "um", "hmm") may be omitted, but are not counted as errors if included.
- Singular instead of plural, and vice versa, will be counted as incorrect ("hour" is not the same as "hours").
- Homophones will be counted as incorrect ("their", "there", and "they're" are not considered the same; "Press for if" and "press 4 if" are not considered the same).
- Concatenated words are considered correct if the concatenated word has substantially the same meaning as the individual words ("video games" and "videogames" are considered the same, whereas "indecent" and "in decent" are not considered the same).
- Address abbreviations are valid. For example, "South 16th East" and "S. 16 E." are considered the same.

3. Readability

a. Definition: The grade level (based on the U.S. education system) at which a user can understand text.

b. How to measure: CAMH analysts use the reference "true transcription" for each audio file to calculate grade-level readability and comprehension based on the Flesch-Kincaid reading ease formula score with an online tool. Readability is a function of the complexity of the audio file, not the transcription. A conversation with higher complexity should have more transcription errors.

4. Reading Ease

a. Definition: The Flesch reading-ease test uses a formula to rate the ease of readability wherein higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read.

b. How to measure: CAMH analysts use the control for each audio file to calculate reading ease based on the Flesch-Kincaid reading ease formula score with an online tool (https://readability-score.com). Readability is a function of the complexity of the audio file, not the transcription.

   i. Scores > 90 should be easily understood by an 11-year-old student.

   ii. Scores 60 – 70 should be understood by a 13- to15-year-old student.

   iii. Scores < 30 are best understood by university graduates.

5. Caption Delay (Latency)

a. Definition: The time elapsed between hearing a voice on the caption phone and the display of captions on the phone's screen.

b. How to measure: For each audio file, CAMH analysts identify eight words spaced uniformly throughout the recording. Based on the video recording, CAMH measures and records the time between when each word is heard and when it appears in the transcript. If the word does not appear in the transcript, CAMH records the time until the next word appears or, if there is a long gap in conversation before the next word would be expected to appear, CAMH records the time for the prior word to appear. CAMH recorded delay times with single-second resolution. All values less than one second were recorded as one second. CAMH rounded up delay values where the fractional part of the measurement was greater than "0.5" to the next second.

6. Completeness

a. Definition: A measure of the words transcribed correctly or incorrectly as a percentage of the total words in the audio file.

b. How to measure: CAMH analysts use the number of words in the "true transcription" and then compare the IP CTS transcription to determine the number of words not included in the output. Incorrectly transcribed words are considered "included". Completeness is calculated as the number of words transcribed divided by the number of words in the original audio file, expressed as a percentage. From the video recording of each call, CAMH creates a text file of the transcribed audio. CAMH uses the NIST SCLITE tool to compare the IP CTS transcription file to the "true transcription". CAMH records the SCLITE tool scores and output.

A.3 Device Testing Cumulative Results

A.3.1 Time for a CA to Connect

Time for a CA to Connect was the one area where testing revealed differences between providers. As Table 6 shows, the average time for an IP CTS CA to connect varied between 3.6 and 5.9 seconds, with Provider 2 demonstrating longer connect times than other providers. Time to connect was not applicable for automated STTs because these services do not rely on telephony connections and are invoked differently than the IP CTS providers. The average time to connect was based on 30 test calls per provider from device testing.

Table 6. Average Time to Connect by Provider

Provider	Average Time for a CA to Connect (Seconds)
Provider 1	3.7
Provider 2	5.9
Provider 3	3.6
Provider 4	3.9
STT-1	N/A
STT-2	N/A

A.3.2 Transcription Accuracy

Table 7 summarizes the average accuracy per provider across audio samples, excluding usability test calls with added noise. Appendix A includes average accuracy data broken out by scenario (conversation). The STT-1 and STT-2 providers are both fully automated speech recognition engines. SST-2 was not tested against all audio samples and is not included in Table 10.

Table 7. Average Accuracy by Provider

Provider	Average Accuracy (Percentage)
Provider 1	88.3
Provider 2	84.5
Provider 3	82.8
Provider 4	88.7
STT-1	83.0

Provider 4 and Provider 1, while generally having longer delay (see Table 8), also generally had higher accuracy than Provider 2 and Provider 3. Interviews with providers indicated that Provider 1 emphasizes higher accuracy, while Provider 2 and Provider 3 emphasize short delay. Survey respondents and usability study participants indicated that both accuracy and delay were areas where improvement was needed.

For all but one test call type, the average accuracy for STT-1 was higher than at least one IP CTS provider. For two audio samples (Pizza – native English Speaker and IRS IVR), STT-1 achieved higher accuracy than any IP CTS provider. This suggests that transcription using automated STT services may be appropriate for some classes of telephony transcription needs.

CAMH observed that faster speech, background noise, more complex speech, computer-generated voices, and non-native English speakers all have a negative impact on accuracy. One script (Pizza) was executed with both a native English speaker and a non-native English speaker.

For all providers and SSTs, the average accuracy for the non-native speaker sample was lower than for the native speaker.

***Note*:** The Accuracy metric does not account for use of punctuation. Punctuation has an impact on ease of reading not reflected in this metric.

A.3.3 Transcription Delay

Table 8 summarizes the average delay per provider across all test calls, excluding usability test calls with added noise. There is considerable variance in delay between providers, with Provider 4 and Provider 1 generally having higher delay and more variability of delay based on call characteristics. As shown in Appendix C, Conversation #4 – Ordering a Pizza, all providers maintain relatively low captioning delay during some types of calls.

Table 8. Average Transcription Delay by Provider

Provider	Average Transcription Delay (Seconds)
Provider 1	15.8
Provider 2	7.3
Provider 3	4.1
Provider 4	14.6
STT-1	2.2
STT-2	2.1

For all test calls, the delay for Providers 4 and 1 tends to increase as each call progresses, only decreasing when there is a pause in the conversation.

Both automated STTs tested had low and consistent delay. The STT-2 delay was consistently in the one (1) - to three (3) -second range across all calls. The STT-1 delay was in the one (1) second range for the majority of test samples. As shown in Table 4, the STT captioning delay was significantly lower than the captioning delay for IP CTS providers. This is expected because the IP CTS providers use a person to re-voice audio into a speech to text engine.

A.3.4 Conversation #3 – Ms. Jackson

The third audio sample tested was from a conversation in the Virginia Standards of Learning sample test,[8] read by a MITRE employee. During device testing, to mimic a back-and-forth conversation, a total of 80 seconds of "dead air" was inserted into the 185-second conversation. Even though this conversation contained pauses in the conversation, the rate of speech is higher than the first two samples. The "dead air" was removed for the usability assessment. Conversation #3 was rated as grade 6.4 based on the Flesch-Kincaid reading ease formula.

[8] The sample test can be found at: http://www.doe.virginia.gov/testing/sol/released_tests/2010/test10_reading8.pdf

Table 9 summarizes the results for test call #3 by provider. Appendix C contains a transcript of the test call.

Table 9. Conversation #3 (Ms. Jackson) Results Summary

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	84.4	26.3	82.6
Provider 2	79.5	10.9	77.7
Provider 3	62.7	5.1	60.9
Provider 4	87.2	17.7	85.4
STT-1	76.1	1.1	95.5
STT-2	66.5	2.2	63.1
STT-3	N/A	N/A	N/A

A.3.5 Conversation #7 – Bank

CAMH tested a seventh audio sample from a recorded conversation of a script for requesting bank account balance and next payment information from a bank. The bank scenario was scripted by calling a local credit union and transcribing the IVR paths for a specific task.

Conversation #7 was rated as grade 3.2 based on the Flesch-Kincaid reading ease formula.[9] Appendix C contains a transcript of the test call. Table 10 summarize the results from Phase 1 device testing for Conversation #7 by provider. Table 11 summarize the results from Phase 2 device testing for Conversation #7 by provider.

Table 10. Conversation #7 (Bank) Results Summary from Phase 1 Device Testing

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	80.8	5.1	61.9
Provider 2	88.0	6.9	96.1
Provider 3	76.8	3.5	89.0
Provider 4	60.8	6.1	68.5
STT-1	67.5	1.2	88.1

[9] MITRE calculated the reading level using the online calculator at: https://readability-score.com

Table 11. Conversation #7 (Bank) Results Summary from Phase 2 Device Testing (Standard Deviation)

Provider	Accuracy (Percentage)	Average Caption Delay (Seconds)	Average Completeness (Percentage)
Provider 1	80.8 (15.1)	5.1 (0.6)	61.9 (35.7)
Provider 2	88.8 (2.6)	6.6 (1.3)	97.0 (0.8)
Provider 3	73.0 (8.1)	3.9 (1.9)	86.9 (2.8)
Provider 4	70.0 (16.0)	5.4 (1.2)	76.2 (13.3)
STT-1	78.2 (5.0)	1.1 (0.1)	83.3 (5.1)
STT-2	66.3 (15.0)	6.9 (0.2)	85.8 (3.5)
STT-3	48.9 (8.6)	4.2 (1.1)	74.2 (9.0)

Appendix B. Usability Assessment Design and Analysis

B.1 Usability Assessment Procedure

The IP CTS Usability Assessment provided controlled hands-on testing and rigorous capture of usability feedback on varying levels of system performance based on the approved quality test metrics. The output of this assessment is to identify ranges of acceptable system performance based on user feedback.

B.2 Participants

Eleven participants were recruited from the subject pool established during the Phase 1 Baseline Usability Assessment. Three participants were new to Phase 2, having not participated in the Phase 1 usability assessment. Participants ranged in age from 33 to 83, with 65 being the average age. All participants reported seeing an audiologist and identified as hard of hearing (36.4%), having a hearing loss (36.4%), or deaf (27.3%). Nine of the eleven participants (82%) rated their overall hearing loss as moderately-severe or worse. One reported mild hearing loss.

B.3 Testing Scenarios

The design of the study controlled for order effects[10] across all eleven participants by using a counterbalanced block design. To the extent feasible, all audio levels were evenly balanced in the scenarios. Videos of the IP CTS devices (Provider 1, Provider 2, Provider 3, and Provider 4) for test call 3 (Ms. Jackson) were recorded using a tripod mounted camera with audio directly into Moviemaker version 2012 operating on a Windows 7 laptop. The STT engine videos were recorded via screen capture using Share X version 10.3 operating on a Windows 8 laptop. All videos were played for the participants with Windows Media Player 2013. Participants were allowed to adjust the volume and skip through the video to locate information.

The call scenario used scenario 7 (Bank) on the MITRE developed IVR. The participants were instructed to place the phone handset on the desk and interact with the IVR using touch tone digits. The scripted test cases allowed MITRE to measure quality metrics and control against outside variables.

The STT engines tested were not developed for use during phone calls. CAMH engineered the MITRE Usability Lab, a technology testing lab maintained by MITRE, the CAMH FFRDC operator, to simulate how captions may appear if they were developed for such use. This was done by connecting the audio from the handset of a commonly used caption telephone to the Windows 8 3.5 microphone jack. Captions were disabled on the caption telephone. All calls were made using the caption telephone for the STT calls. The audio from the phone was then transcribed by the STT engine on the Windows 8 laptop screen.

[10] Order effects may arise from the order in which treatments are presented. Order effects may be associated with the passage of time (practice effect) and fatigue effect.

B.4 Method

Figure 5 depicts the usability assessment process for assessing usability for multiple devices. Participants watched a video demonstrating the captioning ability for each device (test call 3, Ms. Jackson) and made a call using the device (test call 7, Bank) in the MITRE Usability Lab. Testing was completed for all seven devices. Only one participant performed the usability assessment at a time. CAMH collected participant feedback on their overall impressions of the captions effectiveness, efficiency, and satisfaction—the usability—for all scenarios (seven videos and seven calls). Participant feedback on the system usability was collected after each combination video and call to determine the overall usability of the device.

Figure 5. Usability Assessment Process

The facilitator guided each participant through the following process.

- Participants completed a consent form. The form described the study's purpose and the process for data collection.
- Each participant then completed a short online survey for demographics collection. The facilitator provided the participant with task instructions and the in-session device comprehension questions (Appendix D).
- The participant proceeded to watch the video with audio and collect the information required for the in-session comprehension questionnaire during the video call.
- After each video, the participant completed the caption usability questionnaire and the audio usability questionnaire. The facilitator reviewed the participant's responses and asked follow-up questions to make certain the participant understood the task, and allow the participant to provide more information about the experience.
- Next, using the same device as viewed in the video, participants made a call to the IVR in the MITRE Usability Lab to complete the Bank test call. Participants were instructed to place the telephone handset on the desk and rely on captions to navigate the IVR menu and collect the information required for the in-session comprehension questionnaire during the call.

- After each call, the participant completed the caption usability questionnaire and the system usability questionnaire. The facilitator reviewed the participant's responses and asked follow-up questions to make certain the participant understood the task, and allow the participant to provide more information about the experience.
- After completing all tasks (video and call) for all seven devices, the participant completed the post-test assessment questionnaire and the facilitator asked additional questions regarding the participant's testing experience and any experience with other devices.

B.5 Usability Assessment Metrics

MITRE captured participants' feedback on each IP CTS device and STT engine to identify relationships between system performance and usability feedback. Metrics collected were:

- **User Comprehension and Device Accuracy –** Comprehension of test cases by users based on caption accuracy between providers.
- **Caption Usability –** Mean opinion of caption usability. Overall Caption Usability was the computed average of the three metrics below
  - **Effectiveness –** The degree to which a user continuously maintains full context and meaning of the conversation.
  - **Efficiency –** The degree to which a user is comfortable with the perceived tempo of the call.
  - **Satisfaction –** The degree to which the user is satisfied with the calling experience, given the test case and use of transcript.
- **System Usability Scale (SUS) –** Measures whether a device provides an appropriate fit to the intended purpose of IP CTS calls. This metric focuses on the device user interface layout, design, and functions only; not the transcript output.
  - **System Usability –** The fit to intended purpose of each device.
  - **Learnability –** The ease of learning to use each device.
- **User Demographics and Preferences –** User's severity of hearing loss, use of IP CTS devices, and preferences in regards to speed, accuracy, and caption agent assistance in calls.

B.6 Statistical Results

Statistical reference descriptions provide an explanation of the analyses conducted to examine the impact of captions on the usability of the system and subjective measures. For the purpose of these analyses, results are considered statistically significant if the p-value is less than 0.05 ($p \geq 0.05$); where the p-value (p) helps to determine the significant results. Results reported include the mean score and standard deviation (SD); the mean is the average of all cases and the standard deviation is a measure of the degree of concentration of the data are around the mean. Tests performed during the analyses consist one-way Analysis of Variance (ANOVA), also known as an F-test (F). An ANOVA is used when there are more than two groups of means, and also to determine if there are significant differences between the groups. No data corrections or transformations were performed for the following analyses.

B.7 A priori Analyses

CAMH performed several analyses to determine the minimum number of participants required for the Phase 2 Usability Assessment to obtain significant statistical results. These analyses used the participant usability ratings from the Phase 1 Baseline Usability Assessment, wherein four devices were tested with 5 participants completing six scenarios (n = 120).

This evaluation was performed using four analyses:

1. Minimum Data for Usability Ratings – The Phase 1 Baseline Usability Assessment data was split into random samples to determine the minimum number of participants. Five random samples of five participants' usability ratings were taken from the Baseline Usability Assessment data for a total of 30 (n=30) usability ratings. Participant Usability ratings were found to be statistically significant (analysis- one-sample t-test).
    a. MITRE found that usability ratings were statistically significant with as few as five participants.
2. Minimum Data for Device – To evaluate the minimum sample required for multiple devices, the Baseline Usability Assessment data was split to contain ten usability ratings for three providers (n = 60). This split file is comparable to having ten participants test once on three providers. Participant Usability ratings were found to be statistically significant (two-tailed one-way ANOVA).
    a. MITRE found that multiple devices may be assessed with as few as 10 usability cases.
3. Statistical Effect Size – Effect size is the magnitude of an effect a condition has. The Baseline Usability Assessment data was used to calculate the effect size produced by providers and usability ratings. Using G*Power 3.1.9.2 software the effect size was used to calculate the suggested sample size (analysis – a priori power analysis, Cohen's f^2 - for 2-tailed, repeated measures, within factors, 1-way ANOVA).
    a. G*Power suggested 10 participants with 120 cases (instances of caption usability) to obtain statistical significance

In summary, the analyses performed by MITRE indicates that to compare device usability ratings, a minimum of 10 ratings must be obtained per device and to achieve the statistical effect size 120 cases are suggested.

B.8 User Comprehension and Device Accuracy

The in-session comprehension was determined through participant self-report answers to the accuracy of the captions (see Appendix D). During each video and call scenario for each device, participants were asked to complete three simple questions based on the captioned audio. Participants were asked to record the answer(s) as they appeared in the captions. Comprehension was measured based on the participant's ability to fill-in the appropriate answer information. There were 6 responses possible per session. Caption accuracy was measured by the correctness of these answers. Correct answers for the video questions were given one point, while correct answers for the call questions were assigned two points each. This provided for an overall possible score of 99 for each device (see Table 12).

Table 12. Percent Comprehension between Providers and Average Correct per Session

Provider	Percent Comprehension
Provider 1	78
Provider 2	69
Provider 3	53
Provider 4	87
STT-1	90
STT-2	45
STT-3	28

Analysis of the ability to comprehend the text by device found statistically significant differences. Data are (mean) ± (standard deviation). The comprehension was lowest for STT-3 (1.27 ± 1.70) and highest for STT-1 (4.09 ± 2.09). There was homogeneity of variances, as assessed by Levene's test[11] for equality of variances ($p = .424$). The comprehension was statistically significantly different for devices, $F(6, 147) = 5.931$, $p < .001$. There was a mean decrease of 2.82 in comprehension between STT-1 and STT-3 (95% CI, 0.99 to 4.64). Post hoc analysis revealed that there were no differences between the IP CTS devices or between the IP CTS devices with STT-1. STT-3 performed worse than all other devices ($p < .05$) except for Provider 3 ($p = .508$) and STT-2 ($p = .866$) which was statistically no different in comprehension.

An independent sample t-test was run to determine if there were differences in comprehension between Provider 3 and STT-1. Comprehension was higher for STT-1 (4.1 ± 2.1) than Provider 3 (2.4 ± 2.3), a statistically significant difference of -1.7 (95% CI, -3.01 to -0.36), $t(41.747) = -2.562$, $p = .014$. There were no other statistically significant differences between IP CTS providers and STT-1.

B.9 Usability Scores

This subsection presents a summary of usability scores based on eleven participants in the Phase 2 usability assessment. Participants completed questionnaires after each session (see Appendix D). After completing a video or call, participants rated the usability of the captions, device, and/or audio and their overall satisfaction with the call. Higher scores indicate a higher degree of usability.

B.9.1 Caption Usability

Table 13 illustrates the usability scores for each device. Scores are based on a scale ranging from 1 – 7, where 1 equals Strongly Disagree, 4 equals Neutral, and 7 equals Strongly Agree. After the completion of the video scenario and after calls, participants rated the usability of the

[11] Levene's test is an inferential statistic used to assess the equality of variances (a measurement of the spread of a set of numbers) for a variable calculated for two or more groups. This test is used to identify the appropriate statistical analysis to be performed on a set of data.

captions and their overall satisfaction with the call. Higher scores indicate a higher degree of usability.

Table 13. Average Caption Usability Scores by Providers (Standard Deviation)

Provider	Effectiveness	Efficiency	Satisfaction	Caption Usability
Provider 1	4.4 (2.0)	2.8 (1.3)	3.9 (2.3)	3.7 (1.7)
Provider 2	4.5 (2.0)	3.4 (1.4)	4.4 (1.9)	4.1 (1.4)
Provider 3	4.5 (1.9)	3.6 (1.0)	4.3 (1.9)	4.1 (1.4)
Provider 4	5.3 (1.7)	2.6 (1.3)	4.6 (1.8)	4.2 (1.4)
STT-1	4.7 (1.8)	3.9 (1.2)	4.5 (1.7)	4.4 (1.1)
STT-2	3.3 (1.6)	3.3 (1.6)	2.9 (1.2)	3.2 (1.1)
STT-3	2.0 (1.2)	3.0 (1.4)	2.0 (1.3)	2.3 (1.0)

B.9.1.1 Caption Usability Statistical Reference

Effectiveness

Participants were asked to rate the effectiveness of the captions during the scenario. Analysis of the caption effectiveness by device found statistically significant differences. Data are presented as (mean ± standard deviation). The effectiveness was rated lowest for STT-3 (2.0 ± 1.2) and highest for Provider 4 (5.3 ± 1.7). The assumption of homogeneity of variances was violated, as assessed by Levene's test for equality of variances ($p = .030$). The caption effectiveness was statistically significantly different for devices, Welch's $F(6, 64.105) = 13.066$, $p < .001$. There was a decrease of 3.3 points in effectives between Provider 4 and STT-3 (95% CI, 1.9 to 4.7). Games-Howell post hoc analysis revealed that STT-3 was rated statistically significantly lower than all other devices ($p < .001$), except STT-2 which was statistically the same.

Efficiency

Participants were asked to rate the efficiency of the captions during the scenario. Analysis of the caption efficiency by device found statistically significant differences. Data are presented as (mean ± standard deviation). The efficiency was rated lowest for Provider 4 (2.6 ± 1.3) and highest for STT-1 (3.9 ± 1.2). There was homogeneity of variances, as assessed by Levene's test for equality of variances ($p = .152$). The efficiency was statistically significantly different for devices, $F(6, 145) = 2.517$, $p = .024$. There was a decrease of 1.2 points in efficiency between STT-1 and Provider 4 (95% CI, .06 to 2.4) which was statistically significant ($p = .034$). There were no other statistical differences between providers on efficiency.

Satisfaction

Participants were asked to rate the satisfaction of the captions during the scenario. Analysis of the caption satisfaction by device found statistically significant differences. Data are presented as (mean ± standard deviation). The satisfaction was rated lowest for STT-3 (2.1 ± 1.3) and highest for STT-1 (4.5 ± 1.7). The assumption of homogeneity of variances was violated, as assessed by Levene's test for equality of variances ($p = .001$). The caption satisfaction was statistically significantly different for devices, Welch's $F(6, 64.285) = 8.996$, $p < .001$. There was a decrease of 2.5 points in satisfaction between STT-3 and STT-1 (95% CI, 1.0 to 3.9). Games-Howell post hoc analysis revealed that STT-3 was rated statistically significantly lower than all other devices ($p < .05$), except STT-2 which was statistically the same. STT-1 was not statistically different from the currently available IP CTS devices.

Usability

Caption usability was calculated as the average of participants ranked effectiveness, efficiency, and satisfaction. Analysis of the caption usability by device found statistically significant differences. Data are presented as (mean ± standard deviation). The usability was lowest for STT-3 (2.3 ± 1.0) and highest for STT-1 (4.4 ± 1.1). The assumption of homogeneity of variances was violated, as assessed by Levene's test for equality of variances ($p = .028$). The caption usability was statistically significantly different for devices, Welch's $F(6, 64.238) = 8.852$, $p < .001$. There was a decrease of 2.0 points in caption usability between STT-3 and STT-1 (95% CI, 1.0 to 3.0). Games-Howell post hoc analysis revealed that STT-3 was rated statistically significantly lower than all other devices ($p < .05$), except STT-2 which was statistically the same. STT-1 was not statistically different from the currently available IP CTS devices.

B.9.2 System Usability

Table 14 illustrates the device usability scores for each device. Scores are based on a scale ranging from 1 – 100. After the completion of both scenarios on the device, participants rated the device usability based on usability and learnability of the device. Higher scores indicate a higher degree of usability.

Table 14. Average System Usability Scores between Providers (Standard Deviation)

Provider	Device Usability	Device Learnability
Provider 1	81.7 (12.4)	83.0 (16.1)
Provider 2	68.0 (23.7)	77.3 (22.9)
Provider 3	68.8 (19.3)	77.5 (15.4)
Provider 4	82.5 (13.1)	87.5 (13.7)
STT-1	64.8 (26.0)	70.5 (29.2)

Provider	Device Usability	Device Learnability
STT-2	41.9 (17.8)	55.5 (23.5)
STT-3	39.9 (21.9)	51.3 (26.0)

B.9.2.1 Device Usability Statistical Reference

Device Usability

Analysis of the device usability by device found statistically significant differences. Data are presented as (mean ± standard deviation). The device usability was rated lowest for STT-3 (39.9 ± 21.9) and highest for Provider 4 (82.5 ± 13.1). There was homogeneity of variances, as assessed by Levene's test for equality of variances ($p = .401$). The device usability was statistically significantly different for devices, $F(6, 66) = 7.572$, $p < .001$. There was a decrease of 42.7 in device usability between Provider 4 and STT-3 (95% CI, 16.4 to 68.9). Post hoc analysis revealed that STT-1 was not statistically different from any other device. STT-2 was statistically significantly lower than Provider 1 and Provider 4 ($p < .005$). STT-3 was scored statistically significantly lower than the currently available IP CTS devices ($p < .05$).

Device Learnability

Analysis of the device learnability by device found statistically significant differences. Data are presented as (mean ± standard deviation). The device learnability was rated lowest for STT-3 (51.3 ± 26.0) and highest for Provider 4 (87.5 ± 13.7). There was homogeneity of variances, as assessed by Levene's test for equality of variances ($p = .328$). The device learnability was statistically significantly different for devices, $F(6, 66) = 4.044$, $p < .005$. There was a decrease of 36.3 in device learnability between Provider 4 and STT-3 (95% CI, 7.5 to 65.0). Post hoc analysis revealed that STT-1 was not statistically different from any other device on learnability. STT-2 was statistically worse than Provider 4 ($p = .025$). STT-3 was statistically worse than Provider 1 ($p = .021$) and Provider 4 ($p = .005$).

B.9.3 Audio Usability

Table 15 illustrates the audio usability scores for each provider. Scores are based on a scale ranging from 1 – 7, where 1 equals poor, 6 equals excellent, and 7 equals Not Applicable. Individuals who selected 7 (not applicable) were not included in data analysis. Participants were first asked if they were able to hear the audio. Only those who could hear the audio completed this questionnaire (41.6%). After watching the video, participants rated the usability of the audio. Higher scores indicate a higher degree of usability. There were no statistically significant differences between the providers based on audio usability.

Table 15. Average Audio Usability Scores between Providers

Provider	Effort	Quality	Volume Preference	Audio Usability
Provider 1	2.6 (1.1)	3.6 (1.1)	2.0 (0.7)	2.7 (0.3)
Provider 2	2.7 (1.5)	3.7 (1.2)	2.3 (0.6)	2.9 (0.2)
Provider 3	2.6 (1.1)	4.2 (0.8)	2.8 (0.4)	3.2 (0.4)
Provider 4	2.8 (1.3)	4.2 (0.8)	2.8 (0.4)	3.3 (0.3)
STT-1	2.7 (0.8)	3.8 (0.8)	2.3 (0.8)	2.9 (0.5)
STT-2	2.8 (0.4)	4.0 (0.6)	2.3 (0.5)	3.1 (0.3)
STT-3	3.5 (0.7)	3.5 (0.7)	2.0 (0.0)	3.0 (0.0)

B.10 User Preferences

A total of eleven participants took part in the Phase 2 usability assessment. Of those, 82 percent reported the severity of their hearing loss as moderately severe or worse and 81 percent currently use IP CTS.

In terms of caption characteristics collected via post-test questionnaire, 64 percent of participants reported that speed was very important or greater, and 73 percent reported that accuracy was very important or greater. All participants stated that if they could only improve one, they would choose accuracy over speed. Participants stated that they were willing to accept some delay if assured more accurate transcripts. In addition, participants express that it is more important to receive the correct information the first time and reduce the need to repeat information during the call. Additionally, all participants would prefer to have a Communication Agent on IP CTS calls as they believe this will improve accuracy. It should be noted that the speed of captions impacts the user's ability to comprehend a conversation, a 4 second delay in captions can impact comprehension. It is expected that there is a tradeoff between accurate transcription and the speed at which the captions appear. Further research is required to determine the cutoff points for accuracy and speed.

B.11 Percentile Rankings

MITRE computed the percentile rankings for primary metrics from the device testing and usability assessment; accuracy, caption transcription delay, caption usability, and system usability. Scores were converted to the standard score and probability proportion. The percentile ranking takes into account the mean and standard deviation for each of the scores, providing a more exact examination of metrics by provider and allowing a comparison across providers. Caption transcription was reversed scored to provide a higher percentile ranking for shorter

transcription delay. This metric is renamed as ‘caption speed’ wherein a higher percentile indicates faster captioning.

Appendix C. Test Call Transcripts

This appendix provides descriptions and reference transcripts for each test audio sample for the two (2) conversations used during Phase 2 testing. Words with bracket in bold – **[bold]** – indicate words used to determine latency.

C.1 Conversation #3: Ms. Jackson

The third audio sample tested was from a conversation in the Virginia Standards of Learning sample test.[12] For device testing, to mimic a back and forth conversation, a total of 80 seconds of "dead air" was inserted into the 185-second conversation. For the usability assessment, the "dead air" was removed to reduce the time the participant spent waiting for the audio to continue.

Conversation #3 was rated as grade 6.4 based on the Flesch-Kincaid reading ease formula.[13] CAMH calculated the caption delay using the bracketed (and bolded) words in the following transcription of the call:

> In last week's PTA bulletin, I read that the school is still seeking a speaker for this year's eighth-grade awards ceremony. As an eighth-grade student at Eastwood Middle School, I would like you to please consider asking Ms. Ellen Jackson to give this important **[address]**. I realize that the PTA usually asks a local politician or successful business owner to give this inspirational talk, but I think Ms. Jackson would be better than any of our past **[speakers]**.
>
> Ms. Jackson has many fine qualities that make her an excellent choice to speak at the ceremony. She not only meets the requirement of being a former Eastwood Middle School student, but she is also a longtime teacher at our **[school]**. In fact, Ms. Jackson has been associated with this school for much of her life. During that time she has learned many valuable lessons that she passes on to students whenever she **[can]**.
>
> Eastwood Middle School has many fine teachers, and Ms. Jackson is one of the best. Although she teaches English, she is a genius at social studies and math too. Her tutoring sessions are not limited to assignments she has given in her own classes. Ms. Jackson will help with any assignment for any subject. She has, however, one firm rule: when she helps, students must **[work]**. In other words, Ms. Jackson does not simply supply the answers; she teaches students how to find the answers for themselves. She can make difficult concepts seem easy. She patiently explains complicated formulas or confusing procedures one step at a time. Ms. Jackson's homework sessions last as long as necessary. She never ends a session until all students have been **[helped]**.
>
> Ms. Jackson is more than just a great teacher, though. She supports the students of Eastwood Middle School in everything they do. She attends band, orchestra, and choir concerts, and she can be found cheering the Knights to victory at all the

[12] The sample test can be found at: http://www.doe.virginia.gov/testing/sol/released_tests/2010/test10_reading8.pdf

[13] MITRE calculated the reading level using the online calculator at: https://readability-score.com

school's athletic **[events]**. When Ms. Jackson is absent from one event, it is because she is attending another one. Ms. Jackson also volunteers to chaperone school field trips and outings, including those held during **[summer]** break.

C.2 Conversation #7: Requesting an Account Balance

CAMH tested a seventh sample from a recorded conversation of a script for requesting an account balance from a bank. Conversation #7 was rated as grade 3.2 based on the Flesch-Kincaid reading ease formula. CAMH calculated the caption delay using the bracketed (and bolded) words in the following transcription of the call:

Welcome to bank of **[MITRE]**.

Please enter the last 4 digits of your ATM debit card.

You can also enter your telephone access ID or account number.

{MITRE entered 1234}

Thanks.

Now, please **[enter]** the PIN you use with this ATM debit card.

{MITRE entered 1234}

One moment.

Please hold while I **[locate]** your information.

Your checking account **[balance]** is $8,966 dollars and twenty-five cents.

To hear your balance again please press 1.

For your next payment due please press 2.

{MITRE pressed 2}

Your next **[payment]** is due by August 31st, and the amount due is 478 dollars and ten cents.

This reflects the most current information available on your account.

To end this call simply **[hang]** up

Appendix D. Phase 2 Usability Assessment Questionnaires

D.1 Pre Assessment

This survey is issued at the beginning of the assessment, after the consent form is complete.

Please check the option that most applies:

1. On average, how often do you make and receive telephone calls? *

   O six (6) or more times a day
   O several (3-5) times a day
   O once or twice a day
   O several (3-5) times a week
   O once or twice a week
   O less than once a week
   O I do not use the telephone.

2. Do you use captioned telephone service for any of your calls? *

   O Yes
   O No

3. How much of the conversation do you understand without captions? *

   O 0-40% Very little
   O 41-80% Some
   O 81-99% Most
   O 100% I do not need captions
   O Varies, depending on how fast caller talks, accent, etc.

4. In what year were you born?

5. What is your gender? *

   O Female
   O Male
   O I prefer not to respond
   O Other

6. Which service(s) have you used? (Please choose all that apply) *

   - O CaptionCall by Sorenson Communications
   - O Sprint CapTel
   - O Hamilton CapTel
   - O ClearCaptions by Purple Communications
   - O Federal Relay CTS
   - O Mobile Device Application (apps)
   - O Other
   - O None of the above

7. How do you identify yourself? *

   - O I am hearing. I do not have a hearing loss.
   - O Someone with a hearing loss
   - O Someone with a hearing and vision loss
   - O Hard of Hearing
   - O Deaf

8. How would you rate your overall hearing loss? *

   - O I am hearing. I do not have a hearing loss.
   - O Mild
   - O Moderate
   - O Moderately-Severe
   - O Severe
   - O Profound

9. Have you ever received a hearing test from an audiologist or other hearing health care professional? *

   - O Yes
   - O No
   - O I don't know

D.2 Caption Usability

Survey to be completed after each **Scenario** is complete.

For the next set of questions please give your opinion on the **Captions** you could **See** on the screen. (Circle one) **Video or Call**

1. Throughout the phone call, I was able to maintain full context and meaning of the call using **captions** on the screen.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

2. How much effort was required to understand the meaning of sentences from the **captions**?

1	2	3	4	5	6	7
No effort required	Slight effort	Moderate effort	Considerable effort	Extreme effort	No meaning understood	N/A

3. How was the rate at which the **captions** appeared on the screen?

1	2	3	4	5	6	7
Much Slower than preferred	Slower than preferred	A little slower than preferred	Preferred	A little faster than preferred	Faster than preferred	Much faster than preferred

4. I was satisfied with the overall experience with the **captioned text** of the call.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

D.3 Audio Usability

Survey to be issued after each **Video** is complete.

NOTE: If participant indicates that they are deaf and unable to hear any audio, then only complete once.

For the next set of questions please give your opinion on the **Audio** you could **HEAR** through the handset.

1. Were you able to understand the speaker through the audio?
   - O Yes
   - O No [skip the rest of the questions]
2. How much effort was required to understand the meaning of sentences through the **audio**?*

1	2	3	4	5	6	7
No effort required	Slight effort	Moderate effort	Considerable effort	Extreme effort	No meaning understood	N/A

3. How would you rate the quality of the **audio**?

1	2	3	4	5	6	7
Unacceptable	Bad	Poor	Fair	Good	Excellent	N/A

4. Did you adjust the volume before or during the call?
   - O Yes
   - O No
   - O I wanted to, but didn't know how
5. How was the volume of the **audio**?

1	2	3	4	5	6	7
Much quieter than preferred	Quieter than preferred	A little quieter than preferred	Preferred	A little louder than preferred	Louder than preferred	Much louder than preferred

D.4 System Usability Scale (SUS)

This survey is issued after each **Device** is complete.

Please check/circle the option that most applies for each of the statements:

1. I would like to use this device frequently

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

2. I found this device unnecessarily complex

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

3. I thought this device was easy to use

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

4. I would need the support of a technical person to be able to use this device

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

5. I thought the various features of this device were cohesive

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

6. I thought there was too much inconsistency in this device

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

7. I imagine that most people would learn to use this device very quickly

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

8. I found this device very cumbersome to use

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

9. I felt very confident using this device

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

10. There was a lot to learn before using this device

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree

D.5 Post Assessment

This survey is issued at the **End of the assessment**, after all scenarios are complete.

Please check/circle the option that most applies for each of the statements:

1. How important is the speed of captions to making a successful call? *

1	2	3	4	5	6	7
Extremely unimportant	Very unimportant	Somewhat unimportant	Neutral	Somewhat Important	Very Important	Extremely Important

2. How important is the accuracy of the captions to making a successful call? *

1	2	3	4	5	6	7
Extremely unimportant	Very unimportant	Somewhat unimportant	Neutral	Somewhat Important	Very Important	Extremely Important

3. If you could improve only one, which would you choose? *

O Speed (less delay between the speaker and the captions)

O Accuracy (captions that exactly match what the speaker is saying)

Provide comments on this topic below (optional):

4. Were you aware that IP CTS devices connect to a Communication Agent who hears and repeats, or "revoices," everything the other party says? *

   A voice recognition system automatically transcribes the revoiced words into text and transmits the captioned text directly to your telephone display.

   O Yes
   O No

5. How concerned are you about Communication Agent's hearing one side of your conversation(s)?

1	2	3	4	5	6	7
Extremely unconcerned	Very unconcerned	Somewhat unconcerned	Neutral	Somewhat concerned	Very concerned	Extremely concerned

Provide comments on this topic below (optional):

6. Would you prefer a system that removes the Communication Agent? *

   O Yes
   O No
   O Uncertain

Provide comments on this topic below (optional):

7. If **caption speed** declined, would you prefer a system that had a Communication Agent? *

O Yes
O No
O Uncertain

How important is this to you?

1	2	3	4	5	6	7
Extremely unimportant	Very unimportant	Somewhat unimportant	Neutral	Somewhat Important	Very Important	Extremely Important

8. If **caption accuracy** declined, would you prefer a system that had a Communication Agent? *

O Yes
O No
O Uncertain

How important is this to you?

1	2	3	4	5	6	7
Extremely unimportant	Very unimportant	Somewhat unimportant	Neutral	Somewhat Important	Very Important	Extremely Important

9. Please tell us about any technologies you've found particularly helpful in assisting you with your communication needs.

D.6 In-Session Comprehension Questionnaire

D.6.1 Video Questions

Printed forms containing seven sections were used during each video session of the usability assessment. Participants were asked to write in answers as provided by device captions.

Please answer the following questions based on the information provided in the video.

Please answer the following questions:

Session 1

Device # _____

Where does Ms. Jackson teach? _____

What subject does she primarily teach? _____

When does she chaperone? _____

Comments:

D.6.2 Call Questions

Printed forms containing seven sections were used during each call session of the usability assessment. Participants were asked to write in answers as provided by device captions.

Goal	Current Balance & Next Payment Amount
Dial	9 - 1 -703-436-9339
Select Option	1
Last four of ATM	1234
Pin Number	2222

Press * to repeat any prompt

If no response is provided within 1 minute, the call will disconnect.

Please answer the following questions:

Session 1

Device # _____

Were you able to complete the task? _____

What is your current balance? _____

What is the amount of your next payment?_____

Acronyms

ASR	Automatic Speech Recognition
CA	Communication Agent
CAMH	CMS Alliance to Modernize Healthcare
CMS	Centers for Medicare & Medicaid Services
CTS	Caption Telephone Service
dB	Decibels
DUT	Device under Test
FCC	Federal Communications Commission
FFRDC	Federally Funded Research and Development Center
IP	Internet Protocol
IRS	Internal Revenue Service
IVR	Interactive Voice Response
NIST	National Institute of Standards and Technology
PSTN	Public Switched Telephone Network
SCTK	Speech Recognition Scoring Toolkit
SD	Standard Deviation
STT	Speech to Text
TRS	Telecommunications Relay Services
TTS	Text to Speech
UA	Usability Assessment
URL	Universal Resource Locator
VRS	Video Relay Services
WER	Word Error Rate