August 21, 2018

VIA ECFS

Marlene H. Dortch
Secretary
Federal Communications Commission
445 12th Street, SW
Washington, DC 20554

Re: *Misuse of Internet Protocol (IP) Captioned Telephone Service*, CG Docket No. 13-24; *Telecommunications Relay Services and Speech-to-Speech Services for Individuals with Hearing and Speech Disabilities*, CG Docket No. 03-123

Dear Ms. Dortch,

The undersigned IP CTS providers are jointly submitting to the record the attached recommendations for IP CTS Quality Metrics.

Respectfully,

/s/ Bruce Peterson
Bruce Peterson
Vice President of Marketing
CaptionCall, LLC
bpeterson@captioncall.com

/s/ Cristina Duarte
Cristina Duarte
Director of Regulatory Affairs
Mezmo Corporation (dba InnoCaption)
3930 Pender Dr, Ste 130. Fairfax, VA 22030
(703) 865-5553
cristinaduarte@innocaption.com

/s/ Michael Strecker
Michael Strecker
Vice President of Regulatory and Strategic Policy
ClearCaptions, LLC
3001 Lava Ridge Ct, STE 100
Roseville, CA 95661
(916) 274-8429
mike.strecker@clearcaptions.com

/s/ Dixie Ziegler
Dixie Ziegler
Vice President of Relay
Hamilton Relay, Inc.
1006 12th Street
Aurora, NE 68818
dixie.ziegler@hamiltonrelay.com

/s/ Scott Freiermuth
Scott R. Freiermuth
Counsel, Government Affairs
Federal Regulatory
Sprint Corporation
6450 Sprint Parkway
Overland Park, KS 66251
(913) 315-8521
scott.freiermuth@sprint.com

IP CTS Quality Metrics: Provider Recommendations

21 August 2018

Contents

1. Participants

REPRESENTATIVE	IP CTS PROVIDER	EMAIL
DIXIE ZIEGLER	Hamilton	dixie.ziegler@hamiltonrelay.com
JEFF KNIGHTON	Hamilton	jeff.knighton@hamilton-innovations.com
BRUCE PETERSON	CaptionCall	bpeterson@captioncall.com
MIKE HOLM	CaptionCall	mholm@captioncall.com
JASON DUNN	CaptionCall	jdunn@captioncall.com
MIKE MADDIX	Sorenson Communication	mmaddix@sorenson.com
CRISTINA DUARTE	InnoCaption	cristinaduarte@innocaption.com
MIKE STRECKER	ClearCaptions	mike.strecker@clearcaptions.com
RITA BEIER	ClearCaptions	rita.beier@clearcaptions.com
DENNIS SELZNICK	Sprint	dennis.a.selznick@mail.sprint.com

2. Background

Any successful communication, whether on the phone, in-person, through email, or other media is successful when both parties come together to create shared meaning. The transfer of ideas through speech is at the heart of what happens on the telephone. For individuals experiencing hearing loss, the degraded ability to decipher the words spoken make it difficult and sometimes impossible to engage in successful communication over the phone. For the past 15 years, Captioned Telephones have aided this process by allowing a person to use their residual hearing and to view a transcription of what the caller is saying in near real time. To date, Internet Protocol Caption Telephone Service (IP CTS) providers marry diverse technologies with human capabilities to deliver the most accurate and useful captions as possible. This effort is driven by a mandate within the American Disabilities Act (ADA) to provide a telephone experience that is *functionally equivalent* to what a hearing person would experience. Functional equivalency must be the standard filter through which every TRS program action proposed or taken by the Commission, consumer groups, and TRS providers is assessed. [1] This functional equivalency is at the heart of current measurement systems for IP CTS as well as the recommendations contained herein for new ones and responds to consumer groups' long-standing request for quality standards related to equipment and services[2]. With functional equivalency as the leading goal, individuals with hearing loss are empowered to use the telephone in their daily communications.

3. Purpose

In addition to other mandatory minimum standards established by the FCC, IP CTS providers are guided by two regulatory requirements as it relates to quality: 1) Deliver verbatim captions of what a caller says and 2) Answer 85 percent of all calls within 10 seconds. IP CTS providers have historically relied on their own internally developed quality metric definitions, quality assurance testing and performance evaluation standards to satisfy FCC requirements and to provide IP CTS service to users.

[1] Consumer Groups' TRS Policy Statement, April 12, 2011

[2] Consumer Groups' TRS Policy Statement first of 5 goals stated under III.

To monitor the performance of new technologies, such as the use of fully automated speech recognition (ASR) applied in the industry, it is important to establish a common set of metrics that will allow IP CTS users, regulators and other stakeholders to set meaningful minimum standards by which new approaches to providing IP CTS captions can be measured. This is especially important because, in some calls, safety and personal security are at stake. Failure to deliver, at a minimum, an experience comparable to that provided by legacy IP CTS technology is not acceptable. Ultimately, the establishment of quality metrics and minimum performance standards for all IP CTS providers, regardless of the technology used to create captions, will help improve the quality of captions delivered to IP CTS users. The metrics discussed in this paper apply to all IP CTS providers.

Several steps must be completed in order to achieve the goals of establishing, measuring, and reporting quality metrics. The purpose of this document is to discuss the following areas with the end goal of consistently measuring the quality of current service providers so that metrics can then be established.

- IP CTS Metrics Definition
- Testing and Measurement Procedures
- Establishment of performance standards

Failure to address these three areas together could lead to unintended consequences. For example, what does the establishment of an accuracy requirement of 88% mean in the absence of established testing criteria? Without prescribed testing procedures, a testing method can be created which yields any percentage that is desired. Another consideration is to objectively measure the performance of providers, regardless of technology used, based on comparable and consistent test criteria. The end goal should be to ensure that IP CTS users continue to enjoy functionally equivalent communication as mandated by the ADA.

4. IP CTS Metrics Definitions

Quality metrics should help indicate the efficacy of captions as an aid to eligible individuals to establish shared meaning with another person on the telephone. Together with existing metrics, the IP CTS Providers have agreed upon and are proposing the use of two new metrics to comprise the initial set of common industry metrics: Accuracy and Delay.

Accuracy

Definition:

> *Accuracy for IP CTS is defined as 100% minus the Major Word Error Rate on the final displayed captions, where the Major Word Error Rate is the number of word substitutions, deletions, and insertions that significantly alter, obscure or reverse the meaning of the original speech divided by the total number of words in the original speech. It is expressed as a percentage between 0 and 100%.*

Discussion:

Accuracy is an important way to understand if the words spoken are delivered to the IP CTS user in a way consistent with the current verbatim requirement. Captioned telephone transcription should be as close to an accurate, verbatim representation of the words spoken to the CTS user as possible. Historically, Word Error Rate (WER) has been a commonly used metric for measuring the performance of Automated Speech Recognition systems, along with its three drivers, namely substitutions, deletions, and insertions, as illustrated in the examples below:

VERSION	CONTENT
TRANSCRIPTION WITH NO ERRORS	He knew that he could call Chris even if it was the middle of the **night**. He was a very trusted friend.
SUBSTITUTION ERROR	He knew that he could call Chris even if it was the middle of the **knife**. He was a very trusted friend.
DELETION ERROR	He knew that he could call Chris even if it was the middle of the. He was a very trusted friend.
INSERTION ERROR	He knew that he could call Chris even if it was **not** the middle of the **night.** He was a very trusted friend.

Determining if a variance between spoken words and the captions delivered to an IP CTS user departs from the intended meaning is an important question, as illustrated in the examples above. Differences (in the form of substitutions, deletions and editions) between the spoken words and the captions may or may not have an impact on the comprehension of the IP CTS user depending on which words are missed or changed. Some words are less important to conveying meaning and others are vital. Recognizing that comprehension of the IP CTS user is paramount for functional equivalency, word errors are categorized into major and minor errors. Major errors are any incorrect words that significantly alter, obscure, or reverse the meaning of the original text. Errors in captioning numbers, including the order of numbers, that are part of a conversation are generally considered major errors because of the specific meanings they convey (phone numbers, account numbers, etc.) A minor error is an incorrect word that is not essential to understanding the text or does not alter or obscure the meaning of the text. Major errors are counted as mistakes and are reflected in the accuracy calculation. Minor errors are not calculated as errors.

A more complete manual will need to be developed that clearly spells out how to measure accuracy inside of IP CTS. The industry will continue to work on this in the near-term. This work needs to be done quickly so that measurements can be done and a baseline established so that exact metrics can be established. The industry agrees with this initial list from MITRE and that the following do not constitute an error.

- Uppercase/Lowercase and punctuation are not considered in the accuracy calculations.
- 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.
- Address abbreviations are valid. For example, “South 16th East” and “S. 16 E.” are considered the same.

Following is how an accuracy percentage would be calculated on a particular test call:

1. Start with the final caption transcript after any corrections might have been applied by a CA or other technology and a verbatim transcript that includes each word that was actually spoken by the remote party performing the testing.

2. Calculate a traditional WER (Word Error Rate) between the final caption transcript and the verbatim transcript. This is accomplished by counting each error including word substitutions, deletions, and insertions in the captions transcript as compared to the verbatim transcript.

3. For each of the errors in the list generated by #2 above, classify each error as a "major" error or "minor" error, based on the definition that major errors are errors that significantly alter, obscure or reverse the meaning of the original speech. (See list above that starts to clarify what is a major vs. minor error.)

4. Determine the "Major Word Error Rate" by counting the number of errors that are classified as a "major error" and divide that number by the total number of words in the verbatim transcript.

5. Calculate accuracy percentage by performing the calculation of 100% minus the major word error rate as generated in #4 above.

NOTE: IP CTS providers will continue to develop scoring guidelines for the “scoring manual” described above for review and consideration by all stakeholders. One such guideline will address the amount of time or words that have passed in relation to a CA or ASR engine offering a correction for the error. IP CTS providers agree that if a correction is delayed or displayed later in the captioning text, the correction may become less valuable. While developing these guidelines, the IP CTS providers will gather input from Consumer Groups, the DAC, and the DAC subcommittees.

NOTE: Accuracy encompasses the concept or measure of completeness of a call. Omissions are measured. Use of a separate measure for completeness of a conversation by counting words included may be misleading because a transcript created by an IP CTS provider may, in fact, have the same number of words as were actually spoken and yet have many words omitted, added and inaccurately created. Accuracy is a much more meaningful metric compared to a measure of completeness.

Delay

Definition:

> *Delay is defined as the time that elapses between the utterance of a word by the person on the far-end of an IP CTS user's phone call and the final displayed appearance of that word in the stream of captions on the IP CTS user's primary display. It is expressed in seconds.*

Discussion:

Delivery of the captions in a timely fashion is critical to aid in the construction of shared meaning between individuals on the phone. Capturing, processing, and transporting the voice signal and captions inserts delay into the process. Users can tolerate some delay but, as the delay increases, it can affect comprehension. Measurement of delay should consider the average delay during a call, but should also describe delay changes over the course of a call. Understanding the dynamics of delay during calls should be understood. Tolerance for delay may be particularly small at the beginning of a call when important information that sets the context of the conversation are given including the person calling and purpose of the call. If delay increases throughout the length of call, communication may become more and more difficult as the call progresses. Recommended testing procedures should include measurement of delay at intervals on both long and short calls. In cases where a keyword that is used to measure delay are omitted, the following word may be used to calculate delay.

5. Testing and Measurement Procedures

Regular testing of IP CTS providers, based on measurable, objective methods, would allow for a deeper understanding of how IP CTS services work and the quality of service they deliver. As ASR technology evolves, its suitability for IP CTS needs to be evaluated by adequate testing and comparing performance to human-assisted IP CTS. IP CTS providers have been testing their own products and services for years and have operational experience, which is a key component of the value they deliver to the market.

In order to ensure repeatable, valid measures across providers, Accuracy and Delay metrics should be collected from "Test calls" as opposed to actual phone calls by IP CTS users. Using test calls enables the same calls to be used in evaluating performance across multiple providers and may be played multiple times each in order to ensure reliable and repeatable results are gathered. Speed of Answer metrics are determined on a rolling basis on all live calls. This metric can be automatically generated based on provider records, which will ensure that providers are continually meeting the quality standard set forth by the FCC.

One key to valid and reliable testing is the test call content. IP CTS providers could collaborate to develop and share a library of test call recordings and transcripts. A common library may help facilitate reliable and repeatable testing.

Based on these considerations, the IP CTS providers have documented the following key criteria for testing both current IP CTS providers and prospective IP CTS providers using existing or new technologies. It is crucial that any IP CTS testing be conducted in a reasoned and objective manner to ensure accurate results. The integrity of testing is crucial when the results will be relied upon to justify changes to federal rules and regulations. Anything less than 'apples to apples' comparison of quality of ASR captions vs. CA-assisted IP CTS call types and real-world representative environments would not constitute valid testing. Rule changes based on poor testing designs and unrealistic testing conditions will have a direct negative impact on the quality of life of consumers who are dependent on captions to use the phone effectively. Provider self-testing, as well as independent, objective, and verifiable third party testing, may play a role in ensuring that high-quality captions are delivered to IP CTS Users.

General Guidelines for Testing

1. Size of sample (i.e. the number of test calls) should be calculated to provide reliable information
2. Test calls should mimic the proper use of the service (i.e. two persons should not be in the same room)
3. Test calls should follow the structure of a natural telephone conversation
4. Test calls should not be detectable as "test calls" by Communication Assistants; for example, test calls should not start with a loud DTMF tone followed by live conversation.
5. Raw test result data pertaining to individual IP CTS providers should be readily made available to those providers creating an opportunity for providers to use that data to improve the customer experience
6. Standard quality assessment practice includes calibration, data analysis for service improvement and avoids redundant error counts
7. Audio in test calls used in evaluating the captioning service must mimic the audio encountered by IP CTS users
8. Evaluations of IP CTS must replicate the contexts in which the service is used and include enough data points to allow all measures sufficient statistical significance
9. Due to a wide range of quality issues such as telephone audio (including static, distortion, inaudible or unintelligible conversation, and background noises), various accents and regional dialects, a controlled testing environment works best to measure this standard and replicate the tests with proper sampling
10. Any scripts for test calls used by a third-party performing testing would not be given or identified to relay providers prior to the execution of the tests

Testing to Approximate Customer Experience

All testing should be done with an eye to the actual environment that exists when customers utilize captioning as an aid to their telephone conversations, as well as to the different types of callers with whom they have conversations. Testing that does not reflect this environment or range of callers could result in policy decisions that are harmful to individuals who require accommodation as mandated by the ADA. Below are key elements that must be included to yield valid results. This list is intended to be illustrative, and is not comprehensive, but does provide a reasonable attempt to create a representative test environment that will provide useful information.

Background Noise or Artifacts

All providers agree the following noise or artifacts should be considered when testing:

1. A representative range of background noise, including car noise, background voices, fan, radio, etc.
2. Loud television sets, loud music, dog barking, baby crying, moving vehicle with windows down, subway stations/cars
3. Ambient or background noise
4. Calls where speaker turns their head from the phone intermittently and/or whose volume drops significantly during portions of the call
5. Varying quality of mobile phones and headsets
6. Recordings that include “hold time music”, advertising, etc.
7. Calls including dialogue spoken to the IP CTS user as well as other persons in the background but not directly on the call.

Language Complexity

All providers agree with the following statements/considerations regarding testing and language complexity

1. Approximate complexity of language should be evaluated
2. Idioms and colloquialisms

Speaker and Conversation Considerations

All providers agree the following speaker and conversation considerations should be accounted for when creating testing audio to evaluate IP CTS

1. A range of caller accents, ages, genders
2. Dialect and regional speech and pronunciation patterns mirroring the wide and rich variety of “American English” spoken: regional non-standard vocabulary words, (i.e. ‘you-uns’ in Western PA), rural accents including drawls

Call Content Types

All providers agree the following types of calls are representative of call types that could go through IP CTS

1. Social
2. Business Transactions
3. Medical Calls
4. Interactive Voice Response

Test Call Length

All providers agree the following should be considered as a factor that can affect audio quality of IP CTS Calls.

1. Test call length should be representative of IP CTS calls, both longer and shorter calls should be tested

Connection and Testing

All providers agree the following should be considered as a factor that can affect audio quality of IP CTS Calls.

1. Cellular Service, including use of Bluetooth by the non-IP CTS user while driving a car
2. Range of audio transports should be evaluated (analog-landline, VoIP, mobile 3G, 4G, LTE, etc.)
3. Both fixed and mobile endpoints should be tested

All providers agree that when testing, the following should be considered as a factor that can affect the data connection of IP CTS calls

1. Wi-Fi
2. Weak or unstable mobile Wi-Fi connections
3. Hard-wired internet connection
4. Real-world bandwidth throughputs (High bandwidth, low data latency; Low bandwidth, high data latency; Varied upload/download speeds)

The providers stand ready to build a library of test calls that follow the conditions above in order for the industry and new entrants, including ASR participants, to be consistently evaluated.

Transparency and Third-Party Review

Testing that produces results that may be used to make policy decisions related to IP CTS should incorporate a review of research plans by third parties who can make recommendations and assure integrity of these plans prior to the plans being carried out. This review should be done prior to finalizing program designs and in time to allow feedback to be considered. Review by the third parties discussed below should be conducted in cases where the FCC conducts its own research as well as when MITRE or other companies are involved in carrying out evaluations for the FCC that may be used in its efforts to create sound policy to benefit individuals with hearing loss. Review should be done of all program elements including program purpose, hypotheses, design, questions and surveys, scripts, recordings and other test elements. Sound research design will include an understanding of the applicability of findings and ensure that sample sizes and characteristics of test subjects and design elements are robust enough to ensure the results are representative.   Methodologies and rules for scoring Accuracy and Delay should be published and applied uniformly across test calls.

Third-party review should be conducted by non-biased organizations or individuals, as well as IP CTS providers. Third parties should include at least three groups. Reviewers should be required to disclose their expertise and affiliations related to the IP CTS industry and market, specifically to:

1. The Disability Advisory Committee (DAC) and appropriate DAC Subcommittees.
2. Research Experts. This group is expert in research design, analysis and reporting. Experts with experience related to individuals with hearing loss as well as well as leaders in the field of automated speech recognition should be included.
3. IP CTS providers. Providers should individually review plans and deliver feedback as opposed to having one IP CTS provider represent all providers. Each provider brings a unique perspective that can help ensure operational considerations are included.

Results of this review should be published as part of the public record. Research design and the information extracted will be improved and considered in light of a reasoned and open process. Funding for third-party review should be included as part of any FCC proceeding.

6. Establishment of Performance Standards

Adherence to performance standards can protect IP CTS users from inferior service that fails to meet the ADA mandate and the FCC's established mandatory minimum standards, enhanced as proposed herein. A guiding principle in establishing performance standards must be to ensure that any prospective provider or technology must meet or exceed the quality delivered by currently authorized IP CTS providers.

Establishment of standards should be done methodically through application of the objective testing considerations described above. With baseline testing in hand, metrics can be established. In addition, further research should be completed in order to understand usability (including readability) so that this type of user information can be used to establish the correct performance measures. If the above plan is approved, the providers stand ready to provide baseline accuracy and delay information. With this and any additional research completed, actual performance metrics for accuracy and delay can be established.

With agreement from IP CTS providers on metrics definitions, testing procedures and standards, and the establishment of performance metrics based on measurements following these procedures, it is reasonable that all IP CTS providers be held to standards established through this process. Publication of results should be done in a way that is auditable and available to regulators and the public. A publication process should be established which allows an IP CTS provider to review their performance prior to public distribution of results. In the event that an IP CTS provider does not meet the minimum performance standards, the IP CTS Provider should develop an improvement plan to bring the provider into alignment with the expected performance criteria.