Telephone Caption Quality Measures and Metrics Working Group

FINAL REPORT

June 5, 2024

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Executive Summary

While captioned telephone services have been in use in the U.S. since 2004, well-defined industry standard metrics for evaluating telephone caption quality do not exist. Currently, Internet Protocol (IP) Captioned Telephone Services (CTS) are not measured against standards related to caption accuracy, caption delay, or overall communication effectiveness.

To help address this need, in February 2023, MITRE established the Telephone Caption Quality Measures and Metrics Working Group (the Working Group), comprised of community advocates, IP CTS providers, academia, and subject matter experts (SMEs) from related industries, to identify caption quality measures and metrics related to IP CTS.

The working group met every two weeks to identify information that can inform the FCC's Disability Rights Office and Office of the Managing Director about important considerations for defining quality of service for captioned telephones. According to the charter, the mission and vision of the working group was to:

- Identify and define measures that can be used to quantify and compare caption quality as it relates to effective communication
- Propose methods for assessing IP CTS using these measures
- Identify potential criteria for establishing meaningful thresholds for acceptable caption quality

The Working Group identified recommendations for immediate and longer-term actions that will lead to a more complete understanding of how to measure caption quality and, ultimately, define thresholds for acceptable caption quality. The Working Group identified six recommendations:

- Work with an American National Standards Institute (ANSI)-certified standards developer to initiate a process to formalize caption quality standards
- Continue to refine measures and metrics
- Adopt a more transparent testing framework
- Use the adopted framework to measure characteristics of caption accuracy, caption delay, non-speech information, and puntuation and formatting
- Share research and testing information
- Perform additional research to improve measures, identify appropriate metrics, and establish thresholds for acceptable caption quality

Table of Contents

Executive Summary ..... **i**
1 Introduction **1**
2 Caption Characteristics **3**
2.1 Accuracy 4
2.2 Delay 4
2.3 Non-Speech Information 5
2.3.1 Speaker Identification − Multiple Speakers and Conference Calls 5
2.4 Readability 5
3 Caption Accuracy Measures **5**
3.1 WER 7
3.2 VIME 8
3.3 Linguistic Model and Artificial Intelligence-Based Measures 8
4 Formalize Caption Quality Measures **9**
5 Areas for Further Research **10**
Appendix A Working Group Members **11**
Appendix B Captioned Telephone Test Methodology **13**
Appendix C Accuracy Measure: Word Error Rate **19**
Appendix D Caption Delay Measure **23**
Appendix E Non-Speech Information Survey **24**
Appendix F Punctuation and Formatting Survey **24**
Appendix G User Questionnaires and Satisfaction Surveys **25**
Appendix H Caption Accuracy Measures **25**
Appendix I Statement on Word Error Rate **29**
Acronym List **31**

1 Introduction

While captioned telephone services have been in use in the U.S. since 2004, well-defined industry standard metrics for evaluating telephone caption quality do not exist. Currently, Internet Protocol (IP) Captioned Telephone Services (CTS) are not measured against standards related to caption accuracy, caption delay, or overall communication effectiveness.[1]

To help address this need, in February 2023, MITRE established the Telephone Caption Quality Measures and Metrics Working Group (the Working Group), comprised of community advocates, IP CTS providers, academia, and subject matter experts (SMEs) from related industries to identify caption quality measures and metrics related to IP CTS.

The working group met every two weeks to identify information that can inform the FCC's Disability Rights Office and Office of the managing Director about important considerations for defining quality of service for captioned telephones According to the charter, the mission and vision of the working group was to:

- Identifying and defining measures that can be used to quantify and compare caption quality as it relates to effective communication
- Proposing methods for assessing IP CTS using these measures
- Identifying potential criteria for establishing meaningful thresholds for acceptable caption quality

This final report describes the research and findings of the Working Group. It includes:

- A summary of the characteristics the Working Group considers important for understanding caption quality
- A framework for assessing caption quality,
- Detailed methods for measuring some caption characteristics,
- Potential areas for further research.

The Working Group identified recommendations for immediate and longer-term actions that will lead to a more complete understanding of how to measure caption quality and, ultimately, define thresholds for acceptable caption quality. The Working Group identified six recommendations. Each recommendation is followed by a consensus level[2] among voting members of the Working Group:

[1] https://www.ada.gov/effective-comm.htm

[2] Consensus levels are defined as follows:

- Unanimous consensus
- Rough consensus: a position where a small minority disagrees but most agree
- Strong support: a position where a majority support but there is significant opposition
- No consensus: where there is not strong support for position

1. Immediately take steps to initiate a process to formalize caption quality standards through an American National Standards Institute (ANSI) accredited standards organization, as required under Federal Communications Commission (FCC) rules, ensuring and facilitating stakeholder participation including individuals who are Deaf or Hard-of-Hearing (DHH) and use IP CTS. [Unanimous Consensus]
2. Continue refining metrics and define a process for reviewing and updating metrics as technology improves − recognizing that:
    a. Caption quality is a complex topic, there is no “one size fits all” single measure that reflects caption quality for all CTS users, and that different metrics may be appropriate in the future. [Unanimous Consensus]
    b. By any measure, the caption quality required to provide functional equivalence may not be achievable today. The FCC should consider the distinction between what is feasible today and the requirements for functional equivalence, set metrics accordingly, and continually reassess metrics to minimize functional equivalence gaps. [Unanimous Consensus]
3. Adopt the testing framework described in Appendix B to provide a transparent, repeatable assessment process for caption quality. The proposed framework provides publicly accessible guidance on key testing characteristics including data sharing, test materials, and scoring processes. [Unanimous Consensus]
4. Use the proposed testing framework (described in Section 2) to measure the following caption characteristics to obtain an understanding of the current state of caption quality. This information, coupled with additional research, can inform the decision-making process related to establishing metrics for effective telephone captioning. [Strong Consensus – supported by Zainab Alkebsi, Cristina Duarte, Christopher Engelke, AnnMarie Killian, Jen Schuck, Beth Slough, Neil Snyder, Michael Stinson, David Thomson, and Christian Vogler]
    a. **Word Error Rate (WER)**: Described in Appendix C. **Note**: The Working Group recommends further research on caption accuracy metrics. Alternative metrics have been proposed that account for the severity of errors. As new measures that provide meaningful usability, effectiveness, and relevance to functional equivalency are identified, future working groups may recommend that the FCC consider adopting and standardizing them. The Working Group proposes using WER until better alternatives are identified. A submitted statement related to WER is included as Appendix I.
    b. **Caption Delay**: Described in Appendix D
    c. **Non-Speech Information** (survey): Described in Appendix E
    d. **Punctuation and Formatting** (survey): Described in Appendix F
5. Share research plans and results. The Working Group agrees that more transparency in research plans and results is necessary but did not reach consensus on a recommendation. Two alternative recommendations, each with some support, are provided below.

a. Recommendation Option 1: Develop a public-facing protocol whereby all study and research designs that are directly or indirectly funded by the FCC are made available for public review and comment prior to approval and funding, and that research results are made available to public review no later than 6 months after the research phase is completed regardless of publication status (e.g. in an academic or trade journal). The Working Group notes that special exceptions must be made to this mandate around the publication of confidential information relating to the specifics of provider performance metrics. [No Consensus – supported by Cristina Duarte, Christopher Engelke, Michael Maddix, and Dixie Ziegler]

b. Recommendation Option 2: Develop a public-facing protocol whereby all quality testing methodologies and results related to Telecommunications Relay Services (TRS) directly or indirectly funded by the FCC are made available for public review and comment in a timely manner. In addition, testing results are made available for public review after results are delivered to the FCC within an appropriate timeframe. Results from TRS-related research activities should also be published in a peer-reviewed journal or, if not submitted for publication elsewhere, made available for public review by the FCC. Details regarding timelines and processes for public review of research plans and reports need additional work before they can be specified. This recommendation regarding public sharing of research plans and reports was only considered in the Working Group's last meeting, February 22, 2024, and details need to be given additional thought. Examples of questions that need to be addressed are: How can research findings be shared with the public while protecting needed confidentiality of these reports when under review for publication in academic journals? When should research plans and reports not be shared with the public because the material is sensitive? How can the public provide review and comment on research plans in a manner that does not interfere with timely review and funding of research? [No Consensus – supported by Zainab Alkebsi, Lise Hamlin, AnnMarie Killian**,** Jen Schuck**,** Michael Stinson, Michael Strecker, and Christian Vogler]

6. Perform additional research to improve measures, identify appropriate metrics and establish thresholds for acceptable caption quality that apply equally to all captioning technologies, including automated speech recognition (ASR) and Communications Assistants (CAs). [Unanimous Consensus]

2 Caption Characteristics

The Working Group identified several characteristics important to understanding the overall usability of telephone captions. In many cases, the caption quality that can be achieved today may not be sufficient to achieve functional equivalance. These characteristics are described below.

2.1 Accuracy

Caption errors, at a minimum, impact the call flow and may lead to misunderstandings – potentially significantly altering the intended message. Unlike television captioning, telephone callers need to respond in a timely manner and have opportunities to correct misunderstandings. Because of this interactive nature, measures and metrics for captioned phones are likely to be different from measures for captioned videos. Also, caption accuracy considerations may be different for users with minimal caption needs than for users with greater caption needs. Caption quality measures need to provide meaningful information for callers with differing captioning needs.

Considerations for measuring caption accuracy for telephone calls include:

- Do some errors have minimal impact, cause extra thought, or significantly change the meaning of what was said?
- What types of errors are important to measure? Are phonetic spellings tolerable?
- Communications Assistants (people) and ASR services produce different types of errors. Measures need to represent both effectively.[3]
- How many errors are tolerable?
- How are corrections handled?
- How are repeated words and repetitive speech handled? For repeated speech, is it important to maintain a verbatim transcript, or is it clearer to provide a single instance? How should phrases like “I went to – to the store” be addressed? What should be considered “accurate?”

2.2 Delay

Caption delay measures and metrics need to take cultural norms about turn-taking into consideration. Conversations occur in “turns” where one person conveys information and then another person has a turn. The amount of time between conversation turns varies across cultures, and undue delay can impede conversation flow.[4]

Captions with delays that violate these cultural norms may significantly impact the call flow or arrive too late to be useful. Considerations for measuring caption accuracy for telephone calls include:

[3] Fresno, N. (2021, July). Live Captioning Accuracy in Spanish-Language Newscasts in the United States. In International Conference on Human-Computer Interaction (pp. 255-266). Cham: Springer International Publishing. Link to full text/public access here: https://scholarworks.utrgv.edu/wls_fac/90/

[4] Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M. Heinemann, T., Hoymann, G., Rossano, F., Peter de Ruiter, J., Yoon, K-E., Levinson, S. C. (2009) Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences Jun 2009, 106 (26) 10587-10592; DOI: 10.1073/pnas.0903616106

- For corrections, should delay be measured for the first (incorrect) word, the correction, or both? If both, how?
- Should audio timing be measured from the source or destination end? (Transmission delay varies depending on architecture.)
- Are some timings more important than others? Is timing more important for words at the beginning of a turn or the end of a turn? For example, is it more important to measure the delay between words at the beginning or end of a "turn?"
- Should measurements be based on the initial appearance of the (possibly incorrect) transcribed word or the corrected word? Should each correction be tracked with its own additional measurement?

2.3 Non-Speech Information

Callers often receive information during calls other than just the spoken words. Captions should convey this non-speech information (NSI) for CTS users as well. NSI may include speaker gender, speaker identification for calls with multiple speakers, and external sounds such as a baby crying, ambulance siren, or construction noise.

2.3.1 Speaker Identification − Multiple Speakers and Conference Calls

Conference calls and calls with multiple participants will have additional requirements compared to calls with two participants. For calls with multiple speakers, understanding who is speaking (speaker identification) is important. The Working Group also discussed the need for measures that assess caption quality when two or more people are speaking simultaneously.

2.4 Readability

Display factors that impact caption readability, such as screen size, font type and size, the way text scrolls, punctuation, and how corrections are displayed, impact the effectiveness of captioned phones.

3 Caption Accuracy Measures

The Working Group compiled a list of measures for characterizing telephone caption accuracy, included as Appendix I. One of these measures, or possibly some combination of these measures may be useful for characterizing caption quality. The Working Group discussed three of these measures in detail: WER, Visible, Invisible, Minor, and Essential (VIME), and Automatic-Caption Evaluation (ACE).

> **Note:** The Working Group recommends further research on caption accuracy metrics. Alternative metrics have been proposed that account for the severity of errors. As new measures that provide meaningful usability, effectiveness, and relevance to functional equivalency are identified, future working groups may

recommend that the FCC consider adopting and standardizing them. The Working Group proposes using WER until better alternatives are identified.

A superior caption accuracy measure candidate meets the following criteria:

1. The metric is demonstrated to be more closely aligned with human perception than WER. Word-by-word superiority is insufficient. The metric should award a score that, averaged over the set of test material, is more meaningful than WER.
    a. Example 1: If a new metric is consistently better at predicting quality ratings by users than WER, it may be deemed as superior.
    b. Example 2: If subjects prefer system A over B, WER scores B higher than A, the new metric scores B over A, and these results are demonstrated consistently across multiple and varied test sets, this criterion may be satisfied.
2. The metric is consistent across tests. If a test is repeated with identical captions, the score should be identical.
    a. Consideration: If error severity is determined by a set of human judges, a different panel of judges may assign different penalties unless the error types are strictly defined.
3. The metric is consistent across environments. Some metrics using models or parameters may produce a score that varies depending on the test material. The score should not vary depending on the topic, type of captioning system, or content complexity.
    a. Example 1: If a proposed metric penalizes errors in proportion to the word length, this metric may give a medical conversation containing technical terminology a lower score than a conversation between friends (using mostly short words).
    b. Example 2: If a new metric uses a model trained on business calls, it may not perform well on residential calls. If so, the measure should only be used on business calls.
4. The metric is cost-effective if hundreds of samples are tested. The improvement of the metric, compared to WER, should be large enough to justify any additional cost.
    a. Example: If each error for each test sample for each captioning system must be evaluated by one or more judges, the cost and turnaround time could be prohibitive, especially for large test sets. The cost might be justified if the new metric is shown to be significantly better.
5. The metric is consistent over time (optional). If the metric uses models or parameters that are updated based on further study or new training material, the

metric should provide results that can be compared to previous results. Alternatively, the testing organization may “freeze” the models or parameters so that no updates are allowed.

   a. Example: If future research shows that the model or parameters used by the model needs to be modified, it should be possible to compare new scores to those obtained by the previous version.

6. The metric should be understandable by a lay audience (optional).

   a. Example 1: If WER for a set of captioning systems produces accuracy scores ranging from 85% to 95% and an alternative metric produces scores ranging from 97% to 99%, the new metric could give an inflated picture of the actual captioning quality.

   b. Example 2: If the new metric includes a long list of rules or is based on an AI construct such as language model perplexity, the resulting scores may be difficult to interpret.

3.1 WER

WER accounts for insertions, substitutions, and deletions in captions. It does not account for the importance of specific errors. For example, in the sentence “I am not allergic to penicillin” deleting one word out of six provides a 17% WER. Deleting “am” does not significantly impact the meaning of the sentence, while deleting “not” completely reverses the intended message. WER does not differentiate between these errors.

WER is easily explained and understood and delivers unambiguous results. However, WER is an imperfect measure in that it gives all errors equal weight, regardless of their impact. This drawback is somewhat mitigated by averaging over a large test set.

Analysts may also define errors using different criteria. For example, if someone’s name is “Sarah” and the captioner types “Sara,” should that be counted as an error? There may be good reasons for different responses based on the context of the captions.

National Institutes of Standards and Technology Speech Recognition Scoring Toolkit (SCTK)[5] can be used to evaluate WER.

Weighted word error rates, where different classes of errors are ranked differently are also used. In 2010, the WGBH National Center for Accessible Media proposed one weighting method.[6] VIME (described below) is another.

[5] https://www.nist.gov/itl/iad/mig/tools,
[6] http://ncamftp.wgbh.org/ncam-old-site/file_download/CCM_survey_report_final_Dec_17_2010.pdf

3.2 Visible, Invisible, Minor, Essential Errors (VIME)

VIME is a type of weighted word error rate. VIME assesses each error (insertion, substitution, deletion, or tightly grouped set of errors) and categorizes the error as visible, invisible, minor, or essential. Based on these categories, some errors can be assigned a higher "weight" than others in scoring. VIME does not identify weights for each error category to reflect telephone caption quality most effectively. The components of VIME[7] are:

- **Visible Errors:** The error is visible if it is likely to cause confusion for the user because of being grammatically incorrect and/or having created a significant change in meaning from the original intent in such a way that draws a user's attention to the captions themselves.
- **Invisible Errors:** The error is invisible if it maintains grammatical structure in a complete phrase and creates a change in meaning while maintaining conversational fit. Anything that is dropped or inserted but results in captions that can still be read as grammatically correct and complete but changes their meaning is invisible. Homophones and synonyms do not fall in the invisible category but are considered "minor errors" unless they change the meaning in an invisible way.
- **Minor Errors:** The error is minor if it does not change the meaning or cause undue confusion. Minor errors cannot consist of more than one word in a row in either the truth or the captions, and they must not be adjacent to another error of any kind. A single-word error that would have been categorized as a minor error if not adjacent to other errors should be placed in the same category as the adjacent error.
- **Essential Errors:** The error is essential if it contains materially incorrect information while the typifying or identifiable form of that information is retained and appropriate. Essential errors are invisible to the client; otherwise, they should be placed in the visible error category. Typical instances of essential errors include times, dates, addresses, phone numbers, money amounts, credit card numbers, measurements, and proper nouns (names used for individual persons, places, or organizations).
- **Non-Errors:** Differences of capitalization, contraction/expansion, punctuation, formatting, etc. that produce no change in meaning are not considered as errors.

3.3 Linguistic Model and Artificial Intelligence-Based Measures

Some more recently developed tools for measuring and understanding captions rely on complex linguistic models or artificial intelligence (AI) to asses meaning and the severity of errors. These classes of tools may be providing "type-ahead" suggestions as people

[7] https://www.fcc.gov/ecfs/document/10702135359838/1

write, determining what to display based on user's search queries, or providing a confidence level that captions match a reference transcript.

Automatic-Caption Evaluation (ACE), developed at the Rochester Institute Of Technology/National Technical Institute for the Deaf, is one instance of a class of evaluation tools using linguistic models to assess caption accuracy. ACE uses a pre-defined language model to assess the predictability of a word based on context and the "semantic difference" between the intended word and the captioned word. Using these two factors, ACE applies a score to each error. ACE also provides scores for each sentence or utterance that indicates the difficulty in understanding that sentence. In a study with participants who were DHH, ACE better predicted ratings of the usefulness of captions for understanding sentences than WER.[5] Initial research into ACE[8] suggests a close correlation to WER for caption accuracy, but further research is needed.

Other tools use AI to assign a confidence level or error score to captions as compared to a reference transcript. Some of these tools may provide measures that more closely represent IP CTS user's perception of quality than the measures recommended in this report. One challenge with AI-based tools is that it is often not possible to understand why captions received the score that they did.

Further research in this area is needed. It is possible that some combination of measures, including measures using linguistic models or AI, may be useful in assessing caption accuracy.

4 Formalize Caption Quality Measures

White House Circular OMB-A-119,[9] Federal Participation in the Development and Use of Voluntary Consensus Standards and in Conformity Assessment Activities states: "All federal agencies must use voluntary consensus standards in lieu of government-unique standards in their procurement and regulatory activities, except where inconsistent with law or otherwise impractical." To comply with this mandate, telephone caption quality measures will need to be documented in consensus-based standards through an ANSI-accredited standards developer.

The Working Group recommends that the FCC facilitate this process by:

- Defining the desired scope for proposed standards
- Identifying an ANSI-accredited standards organization, such as International Telecommunication Union (ITU) to manage
- Initiating a request for starting the standards process
- Participating in standard development, using this document as an input, recognizing "loss of control" over the process at this point

[8] https://dl.acm.org/doi/abs/10.1007/978-3-031-08648-9_61
[9] https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-119-1.pdf

- Continuing maintenance/monitoring/updates as needed.

5 Areas for Further Research

The Working Group recommends conducting research in the following areas to inform the FCC's decision-making process related to caption measures and metrics.

1. **Determine the amount of delay** that is generally acceptable to the average IP CTS user and provides a functionally equivalent experience for the consumer by conducting studies, using controlled conditions.
2. **Characterize differences in the functional equivalence of captions** for IP CTS callers compared to non-captioned calls for non-IP CTS callers by performing research and provide findings related to appropriate metrics for IP CTS characteristics. Findings should indicate whether existing services fail to meet, meet, or exceed identified usability criteria for each characteristic.
3. **Identify usability metrics that may vary based on IP CTS user characteristics**. For example, a person with more severe hearing loss may require more accurate captions for an equivalent conversation experience than a person with less severe hearing loss.
4. **Identify the level of caption accuracy that is acceptable for IP CTS users** for effectively communicating in a telephone call, recognizing that some metrics may vary based on IP CTS user characteristics.
5. **Determine how different combinations of caption accuracy and delay impact the level of caption quality that users of IP CTS judge acceptable**.
6. **Identify the factors that determine whether consumers will judge a particular level of caption accuracy or caption delay acceptable**. For example, conversation topic and importance to consumers may affect their standard for what is acceptable caption quality. Consumers may have a different standard for a casual social conversation versus a discussion with a doctor regarding a medical issue.
7. **Determine what evidence is needed to confirm standards established for caption quality are effective** by researching how consumers of IP CTS respond to different levels of performance in proposed measures of caption quality.

Appendix A Working Group Members

The Working Group is comprised advocates for the deaf and hard of hearing community, researchers and academia, Internet Protocol (IP) Captioned Telephone Services (CTS) providers, and related industry subject matter experts (SME). Members are listed below.

Name	Title	Role
Zainab Alkebsi	Policy Counsel, National Association of the Deaf	Voting Member
Yin Bao	Software Systems Engineer, MITRE	SME
Cristina Duarte	Senior Director of Regulatory Affairs, InnoCaption	Voting Member
Christopher Engelke	Vice President, Ultratec	Voting Member
Jon Gray	Business Manager, CaptionMate	Alternate
Lise Hamlin	Director of Public Policy, Hearing Loss Association of America	Voting Member[10]
AnnMarie Killian	Chief Executive Officer, TDIforAccess, Inc. (TDI)	Voting Member
Linda Kozma-Spytek	Consultant and Professional Adviser on Technology, Hearing Loss Association of America	Alternate
Mike Maddix	Director, Government and Regulatory Affairs, Sorenson Communications	Voting Member
Kenny McCann	Vice President of Customer Retention, ClearCaptions	SME
Jim Malloy	Principal Information System Engineer, MITRE	Co-Chair
Kenny McCann	Vice President of Retention, ClearCaptions	Alternate
Brian Meyer	Public Policy Associate, Hearing Loss Association of America	Alternate
Adam Montero	Vice President of Engineering, Captioning, Sorenson Communications	SME
Daniel Muiz	Quality Assurance Manager, ClearCaptions	SME
Mark Pfaff	Principal Cognitive Scientist for Collaboration Systems, MITRE	SME
Christian Vogler	Director, Technology Access Program, Gallaudet University	Voting Member
Genelle Sanders	Director, Programming and Operations, TDIforAccess, Inc. (TDI)	Alternate

[10] Lise Hamlin was replaced as a voting member by Neil Snyder when she retired.

Name	Title	Role
Jen Schuck	Chair, Global Alliance of Speech-to-Text Captioning	Voting Member
Kimberly Shae	Chair, Global Alliance of Speech-to-Text Captioning	Alternate
Neil Snyder	Director of Public Policy, Hearing Loss Association of America	Voting Member[11]
Beth Slough	Director of Account Management and Compliance, Hamilton Relay	Alternate
Michael Stinson	National Technical Institute for the Deaf, Rochester Institute of Technology	Voting Member
Erik Strand	Vice President of Engineering, Sorenson Communications	SME
Michael Strecker	Vice President of Regulatory and Strategic Policy, ClearCaptions	Voting Member
David Thomson	Vice President of Speech Sciences, Sorenson Communications	Alternate
Jan “Yenda” Trmal	Associate Research Scientist, Johns Hopkins University	Co-Chair
Sharon Ward	Senior Applied Operations Researcher, MITRE	SME
Heather York	Vice President Marketing and Government Affairs, VITAC	SME
Dixie Ziegler	Vice President, Hamilton Relay	Voting Member

[11] Neil Snyder replaced Lise Hamlin as a voting member when she retired.

Appendix B Captioned Telephone Test Methodology

Purpose

The test methodology described in this section is intended to provide high quality, repeatable, transparent results for any measures identified for Telephone Caption Quality.The Working Group recommends using this methodology for measuring Word Error Rate (WER) and caption delay, but the methodology can be adopted for other measures.

These procedures and processes should be accessible to the public so that the captioning services being tested and the consumers of these captioning services can understand the testing process. Considerations for these inputs include: testing procedures, criteria for developing test material, documenting test and scoring procedures, and processes for test artifacts and results review.

Inputs

Three key inputs for any measure in this methodology are to develop:

1. Testing procedures
2. Scoring procedures
3. Process for test artifacts and results review

Testing Procedures

Step 1: Determine Representative Test Content

The materials used as test content should reflect the variation in audio and speaker types found in the production service. Test material should be balanced across accents, acoustic environment, topics, and other conditions in approximate proportion to that of typical calls in provision of services. For example, if 75% of calls are residential and 25% are business, the test material should be 75% residential and 25% business. At the same time, it is important for the test battery to include conversations that vary conversation types, speaker characteristics, and audio quality that match the range of conversations IP CTS callers encounter. Note that it may be difficult to accurately determine what constitutes representative test content, because privacy rules around Telecommunications Relay Services (TRS) severely limit the types of analysis that can be perfromed.

Test materials should include participants who know each other in roughly the same proportion as occur in provision of services. For topics such as discussing a birthday party, test audio should be generated using people who know one another. For topics such as making a travel reservation, test audio should be generated using people who do not know each other.

In general, conversation topics should proportionately represent the types of conversations typically encountered by IP CTS callers, For example:

- Conversations with family or friends
- Conversations to schedule appointments or services
- Conversations including medical terminology (for example, a doctor providing test results to a patient)
- Employee conversations at work, including industry-specific terminology (for example, conversations between lawyers or engineers)

Some infrequent types of calls, such as emergency or 911 calls, should be represented more heavily because of their importance.

Ideally, conversations in the test battery should vary in their complexity. For example, a simple medical conversation could consist of making an appointment with a doctor. In contrast, a complex medical conversation between a hearing doctor and a patient who is hard of hearing might involve discussing various issues related to an upcoming medical procedure. Vocabulary is an important factor in conversation complexity. For example, a conversation about a lunch appointment will have a simpler vocabulary than a discussion about a technical, medical, or legal topic.

Test materials should also vary in intelligibility. As used here “intelligibility” refers to how difficult the speech is to understand due to a combination of factors. These factors include accents, background noise, background voices, channel impairments (such as network dropouts and distortion), phone types (such as landline vs. mobile) truncated words (words where the start or end is muted, such as due to half-duplex speakerphones), and parties that do not speak clearly. Note that IP CTS providers may support Wireline devices, mobile devices, and/or browser-based calling. Some factors, such as channel impairments, may not apply equally to all device types.

Test materials should also include speakers of a range of ages, as well as speakers who vary in speech loudness and clarity of pronunciation. This should include, at a minimum, elderly (as determined by a listening test where the advanced age is evident in the voice), children under eight years old, and speakers with deaf accents.

Test content should be neutral in regard to type of provider: automatic speech recognition or communication assistant, landline, Voice over Internet Protocol, or mobile phone.

Determine Feasible Testing Period and Amount of Content for Assessment of Performance

The size and content of the test battery will be restricted by the designated test period. Since it is likely that entity performing testing will be completing the test battery over a multiyear period, it is important that test sessions consist of a manageable amount of test content.

Another important consideration is that the arranging, conducting, and scoring of tests, as well as record keeping and distribution of scores be manageable for the testing organization. There is a trade-off between having a large dataset, which can be representative of more calling scenarios, and a longer time between the release of test results.

Criteria for Developing Test Material

The test development team will establish guidelines for creation of appropriate test materials that have representative test content. Guidelines are needed for:

- Confidentiality of materials
- The process of recording, maintaining, and placing test calls
- Specifying rules for inclusion of material

Confidentiality of materials: It is important for test performance to approximate the level of performance in actual provision of service. Since provision of services typically involves new calls without advance preparation, test material will need to be unfamiliar to the Communications Assistant (CA) or service being tested. For test materials to be unfamiliar, they need to be kept confidential.

Recording, maintaining, and playing test calls: If available recordings are used, the development team will need to establish criteria for determining whether these recordings are acceptable. For example, test audio must not be selected from any source that providers or ASR engines have used for tuning their processes. If the team makes new recordings, the team will need to develop instructions for speakers, including information regarding the purpose of the recordings. Informed consent will be required from speakers.

Natural speech will be collected. Calls will not be scripted, but they could involve participants role playing. Participants may need certain types of backgrounds to do particular types of role-playing, such as a medical conversation. Speech may include disfluencies and ambiguous audio.

All test audio and other testing artifacts necessary to review or reproduce test results should be maintained in a secure repository, with backups to prevent data loss.

Test calls should mimic natural calls, and be placed through the telephone network. For any scenario, the test calls should originate from the same source so that, to the extent possible, the audio path represents a realistic call path and the environment is as similar as possible across tests. Note that Provider's infrastructure varies – calls placed from/to analog phone lines, mobile apps, or browser-based services will have different characteristics. The call path and infrastructure outside of the Provider's networks should be as consistent as possible.

Document Test Procedures

The test procedure describes how a test battery, which consists of a standard set of subtests, will be scheduled, and administered. All CAs or services being tested will complete the same set of subtests. Test batteries will consist of pre-recorded material.

The test development team will specify the test scheduling process. Establishment of this process includes:

- Identification of appropriate frequency of testing so that it is manageable for IP CTS providers and the testing organization
- Determination of when tests may be conducted (e.g., day of week, time of day)
- Procedure for selecting IP CTS providers and related services for testing

The team will also describe the test administration and information collection process. This description will include:

- Instructions to participants during test call audio collection
- Test administration procedure
- The test will be administered in a manner that keeps test material and performance confidential
- Information to be collected during a test session

The test development team will also describe the process for selection of the testing organization. The testing organization will need to possess the required expertise and experience. The test organization will be independent from IP CTS providers and related services being tested. It will also need to publish general, de-identified perfomance information to help the public understand the capabilities of IP CTS and share specific test results with individual IP CTS providers so that individual Providers can understand their results and, possibly, challenge test results.

Document Scoring Procedures

Scoring procedures determine scores for measures specified by the Working Group. For example, the Working Group proposes WER and caption delay time as standard measures, as defined in Appendices C and D. Where possible, procedures should be documented or derived from industry-recognized standards or practices. Development of scoring procedures will be specific to the measure being scored. Using WER as an example, scoring procedures might include:

- Specifications for reference transcripts
- Creation of an equivalency table
- Procedure for modification of reference transcripts

- Scoring of corrections by Communications Assistant (CA) and Automated Speech Recognition (ASR)
- Assessment of reliability of scores

Specification of reference transcripts: For WER, reference transcripts contain text corresponding to the recorded test audio. Reference transcripts define what responses may be considered correct. The transcripts include allowed variations for potentially debatable cases such as filler words, spelling variations (Kathy vs. Cathy), ambiguous (e.g., mispronounced or unclear) words, homophones, background speech, and word fragments.

Modifications in reference transcripts: The test development team will define a process for making modifications in reference transcripts once the test battery has started being used. An IP CTS provider and related service being testedmay produce a hypothesis transcript with an interpretation that the reference text did not anticipate. After looking at results for an IP CTS providers and related services being tested, scorers may realize that the vendor's interpretation is reasonable. For example, a reference transcript starts with "I just want to give you a call" and the Provider returns "I just wanted to give you a call." If the test assessment team concludes that either could have reasonable been heard, the alternative interpretation may be added to the allowed variations for the reference transcript.

Specification of what is and is not captioned: Scoring procedures will also specify what spoken and non-spoken material will not be captioned. Non-speech background noise that will not be captioned includes car noise, fans, and music. Speech background noise that will not be captioned includes crowd noise, unintelligible babble such as in a restaurant, television, and separate conversations. The test development team will decide whether certain conversations (e.g., another person who is participating in the call but is not close to the microphone) will be captioned.

Scoring of corrections: For WER, corrections of word errors (whether by a CA or ASR) may replace the original word. Testing procedure must specify whether the originally displayed word, the final caption, or both, will be assessed.

Scoring reliability: For all measures, the reliability of test scores will be assessed to determine the extent of agreement in independent scoring of the same test performance. High inter-rater reliability provides confidence that the process is consistent and uniformly applied.

Process for Test Artifacts and Results Review

Identifying Process for Sharing and Review of Results

The test development team will describe a process for sharing test results. The Working Group recommends that this process share general test performance information, such

as mean scores and result variability for tests, broken down by Providers, with the public.

There will also be a process for sharing with Providers their own test performance, an opportunity for Providers to communicate with the test organization after reviewing their performance, and a mechanism for challenging the test results where warranted. Adjudication of any challenges should be managed by an independent body, not the organization assessing the test results.

When a test in a battery, including audio, reference text, and scoring procedure, is no longer being used for a given round of testing, it will be made public.

When the testing organization publishes results, they should include confidence intervals. The confidence intervals are determined using standard statistical methods and based on test parameters such the number of test samples, number of words per call, variability of scores across calls, etc. We do not recommend basing decisions or establishing policy based on numbers that fall inside confidence intervals (e.g., based on differences that are not statistically significant).

Develop Multiple Versions of Test Battery

Once the first test battery has been developed and the test development team determines that it is working reasonably well, additional parallel versions of the test battery may need to be developed. Additional versions of the test battery may be developed to prevent practice or familiarity with tests to affect test scores.

Predictive Generalizability of the Test Battery

If questions arise regarding whether the test battery includes test content with sufficient variability to assess IP-CTS caption quality, it is desirable to use an empirical approach to resolve these questions. Once the test battery has been developed and implemented, questions may arise regarding whether the test battery adequately reflects the variations in audio, speaker, and content found in caption production services. For example, production services may report that they regularly provide captioning for speakers with a certain accent discussing a specific topic, but that accent and topic are not covered by the test battery. To address such questions, it may be necessary to conduct one or more studies to determine the extent that performance on the test battery predicts production performance on these specific variations in audio, speaker, and content that are different from those in the test battery. In this way, research will empirically evaluate the generalizability of the test battery.

If performance on the test battery predicts performance on these specific variations not in the test battery, these findings would support continued use of the test battery as has been constructed. If performance on the test battery does not predict performance, and if production services consider these variations not in the test battery important in services, then it may be desirable to make modifications in the test battery.

Appendix C Accuracy Measure: Word Error Rate

Word Error Rate (WER) is used to measure caption accuracy, counting the total number or "errors" (insertions, substitutions, or deletions as a percentage of the total number of words in the conversation). The guidelines below were developed for WER analysis. Other accuracy measures may have use different guidelines. For generating reference transcripts for WER analysis, use the following guidelines for determining what constitutes an error.

Note: The Working Group recommends WER as a caption accuracy measure with the understanding that it has some limitations and further research may identify alternative measures to replace or augment WER. The Working Group recommends additional research related to such measures.

1. Uppercase/lowercase and punctuation are not considered in the accuracy calculations.
    a. Hyphenated words, non-hyphenated words, and words separated by underscores are considered equivalent (e.g. "thank you," "thank-you," and "thank_you" are all considered equivalent).
    b. Abbreviations that have spaces or periods between the letters are considered equivalent ("FCC," "F C C," and "F.C.C" are all considered equivalent).
    c. Universal Resource Locators (URL) that contain extra spaces or spell the words "slash" or "dot" are considered valid ("fcc.gov/smartdevice," "fcc dot gov slash smart device," "fcc. gov/smartdevice," and "fcc dot g o v forward slash smart device," are all considered equivalent).
    d. Words that contain spaces in between the letters are considered equivalent to the word without spaces. For example, "H u m b l i n g" and "humbling" are considered equivalent.
        i. For the purposes of scoring, the word as spelled is counted as a single word consisting of the sum of its letters, not as a series of words with each letter counting as a separate word. Example:
            1. "H U M B L I N G" à "H U M P L I N G" = 1 Error
            2. "H U M B L I N G" à "H U M P P L I G" = 1 Error
    e. Series of numbers that function as a single unit may be spaced or separated in any commonly used form. Example:
        i. For example, the following formats for phone numbers are equivalent:
            1. (123) 456-7890
            2. 1234567890
            3. 123-456-7890
        ii. For the purposes of scoring, the number as spelled is counted as a single word consisting of the sum of its digits, not as a series of words with each letter counting as a separate word. Example:
            1. "(123) 456-7890" à ""(123) 556-7890" = 1 Error

2. “(123) 456-7890” à ““(123) 555-7890” = 1 Error

iii. For example, the phrase “I’m going to count to three. 1 2 3” is considered equivalent to “I’m going to count to three 123”

2. Contractions and expanded phrases are considered interchangeable with exceptions:
   a. Contractions including clitic words are considered equivalent with their expanded form. Examples:
      i. WORD + “-N’T” – “WOULDN’T” and “WOULD NOT”
      ii. WORD + “-S” – “THAT’S” and “THAT IS/HAS”
      iii. “’T-” + WORD – “’TIS” and “IT IS”
      iv. WORD + “-A” − “CUPPA” and “CUP OF”
      v. WORD + “-CHA” – “DON’T” and “DON’T YOU”
      vi. WORD + “-(D)YA – “WHAT’DA” and “WHAT DO/DID YOU”
      vii. GONNA/WOULDA/HADDA/HAFTA/COULDA are considered equivalent to their standard uncontracted form; for example, GOING TO.
   b. However, in the case of Subject-Auxiliary Inversion, hearably contracted words must be represented in their contracted form. Examples:
      i. “DON’T YOU DARE” is not equivalent to “DO NOT YOU DARE”
      ii. “HAVEN’T YOU HEARD” is not equivalent to “HAVE NOT YOU HEARD”
      iii. “SHOULDN’T YOU GO” is not equivalent to “SHOULD NOT YOU GO”
3. Concatenated words are considered correct if the concatenated word has substantially the same meaning as the individual words (“video games” and “videogames” are considered equivalent, whereas “indecent” and “in decent” are not considered equivalent).
   a. In the case of word combinations that include repeated or “soft” leading/trailing consonants and underarticulated consonants at the border, all hearable alternatives are acceptable. For example, “Mike’s stool” and “Mike’s tool.” However, if the audio includes a clear differentiation between the trailing and leading consonant, only the enunciated interpretation is considered correct.
4. Standard abbreviations and typographic symbols are considered equivalent to their spelled-out alternatives *when used within explicit and immediate context of grounding speech.*
   a. For example, “South 16th West” and “S. 16 W.” are considered equivalent, but “Go West young man” and “Go W young man” are not acceptable alternatives.
   b. Additional examples:
      i. “µg” and “ug” are valid alternatives for “microgram” within the context of a measured unit.
         1. e.g., “300 µg of uranium” is acceptable, but “µg is an abbreviation for µg” is not.
      ii. “@” is a valid alternative for “at” within the context of a website or handle.

1. e.g., joe@mitre.org is acceptable, but “it’s @ the end of the street” is not.

iii. “dB” is a valid abbreviation for “decibel” within the context of a measured unit.
1. e.g., “300 dB” is acceptable, but “dB is an abbreviation for dB” is not.

iv. “$” is a valid abbreviation for “dollar” within the context of a measured unit.
1. e.g., “$300” is acceptable but “$ bill” would not.

v. “#” is a valid abbreviation for “pound” and “hashtag” within the context of discussion of a symbol.
1. e.g., “press the # key” is acceptable, but “it weighs 1 #” is not.

vi. “*” is a valid abbreviation for “star” within the context of discussing a telephone keypad symbol.
1. e.g., “press the * key” is acceptable, but “look at the North *” is not.

vii. “&” is a valid abbreviation for “ampersand” and “and.”

viii. “%” is a valid abbreviation for “percent” and “percentage.”

c. Truncated versions of a word are not considered valid alternatives. Example: “exam” is not a valid alternative to “examination” nor “chemo” for “chemotherapy.”

d. Numbers may be spelled out or numeric (“400,” “four hundred,” and “4 hundred” are all considered equivalent).

e. Times may be represented as words or numerals including standard lexical variants.

i. “8:30,” “8 30,” and “eight thirty” are all considered equivalent.

ii. “4:00,” “4 o’clock,” and “four o’clock” are all considered equivalent.

iii. “Quarter to five” and “4:45” are considered equivalent if contextually appropriate.

5. Disfluencies

a. Non-lexical verbal discourse markers may be omitted but are not counted as errors if included.

i. “Ah,” “um,” “hmm” spoken within a larger phrase and without explicit internal reference.
1. “He said uh that he was going” vs. “He said, ‘uh’”

b. Restarts and stutters can be removed or captioned.

c. Lexicalized sounds are counted as errors if excluded.

i. All two-syllable words uh-huh, um-hum, mmm-mmm are considered words rather than disfluencies.

ii. Standalone sounds used in place of words are considered words rather than disfluencies.
1. “hmm” used as a question – e.g., asking for repetition

d. Phonetically similar sounds may be captioned according to any of their hearable alternatives:

        i. “ah” and “a” and “uh” may be equivalent given performance and context.
6. Singular instead of plural, and vice versa, are counted as incorrect where the difference is hearable (“hour” is not the same as “hours”).
7. Heterographs will be counted as incorrect (“their,” “there,” and “they’re” are not considered equivalent; “Press **4** to speak to an operator” and “press **for** to speak to an operator” are not considered equivalent).
    a. Where lack of context would allow any heterograph, all options will be accepted. In the case of the word “four,” if someone simply said the word “four” without any context, “for,” “fore,” “four,” and “4” are considered valid alternatives.
    b. Proper **nouns** transcribed with reasonable phonetic spellings are acceptable.
    c. Spelling alternatives will be counted as correct if the variant presented has more than a million hits on Google or consists of a recognized alternative spelling.
    d. American English and British English spellings with the same pronunciation are considered equivalent. For example, “color” and “colour” are considered equivalent, but “aluminum” and “aluminium” are not.

Appendix D Caption Delay Measure

“Caption delay” is defined as the time between when a word is audible on the Internet Protocol Captioned Telephone Services (IP CTS) phone and the caption for that word being displayed on IP CTS phone. Measurement is typically accomplished by video/audio recording a test call and measuring caption delay based on the recording. There are several factors to consider in performing these calculations. The Working Group recommendations are based on characteristics that are important for telephone calls and may not be appropriate for other types of captioning.

Sample size: It is not always possible or necessary to calculate caption delay for every word in a conversation; for example, since captioning is not perfect, it is likely that there are some words that only occur in the audio or the captions. A minimum of 20 words per call (for calls that contain 20 or more words) and four words per minute for calls longer than two minutes should be assessed for caption delay. These words should be selected prior to beginning analysis and should include words from the beginning, middle, and end of turns and be relatively uniformly distributed throughout the call.

Measurement points: The timing for captioned words should be measured from the time the word is fully displayed on the screen. If a word appears one letter at a time, the measurement should be at the time the last letter is displayed. Similarly, the audio timing should be measured from the time the utterance is completed in the audio. Note that audio timing is measured from the IP CTS user’s experience and relies on the audio and caption timing as experienced on the IP CTS phone, not when spoken by the other caller. All calls have some transmission delay. Caption delay does not consider this delay.

Corrections/substitutions: Caption delay should be calculated from the first captioned word, even if incorrect. For example, if the reference (true) transcript was “I like apples,” and the caption showed “I like oranges,” timing should be calculated based on when apples is heard and when “oranges” appears in the captions. Additional information about corrections, including the number, time required for corrections to appear, and proximity of corrections and errors should be tracked separately.

Omissions: If one of the words selected for caption delay calculation is not included in the captions the next closest word, with a preference for the next word in the transcript, should be used instead. If a word is missing, typically the next word in the transcript would be selected in its place. If the selected word is at the end of a “turn,” with a pause afterwards, then the word prior to the selected word should be used instead. In cases where the captions omit several words in a row, the data point should be skipped.

Timing resolution: Audio and caption delay should be measured with 0.1 second or better resolution.

Summarization: If caption delay data is summarized on a per-call basis, then the number of data points and the median and standard deviation of the collected data points per call should be reported.

Appendix E Non-Speech Information Survey

Non-speech information (NSI) includes all information that can be obtained from the audio of a call except for the conversation being captioned. While the Working Group has not identified quantifiable measures for assessing NSI, Internet Protocol Captioned Telephone Services (IP CTS) users may benefit from knowing which services attempt to provide various classes of NSI. Potential categories include:

- Environmental cues, such as dog barking, baby crying, ambulance, airport announcements, and music
- Verbal cues, such as laughing, coughing, shouting, or crying
- Speaker descriptions, such as male/female voice, speaker 1, speaker 2, or speaker names (if known)
- Audio quality descriptions, such as audio volume low, speaker unclear, or audio cutting out
- Call Status Indicators, such as ringing, dial tone, fast busy
- Silence Indicator, an indicator that there is nothing to caption to allow an IP CTS user to distinguish silence on the call from caption failure.

If NSI is included in captions, it should be clearly marked to indicate that it is not speech.

Appendix F Punctuation and Formatting Survey

Punctuation and formatting impact the level of effort required to understand captions. While the Working Group has not identified quantifiable measures for assessing Punctuation and formatting, IP CTS users may benefit from knowing which services attempt to provide various classes of puncuation and formatting. Potential categories include:

- **Paragraph breaks** − starting a new line after a significant pause in the conversation
- **Capitalization** − upper/lower case letters where appropriate. For example, upper case for the first word of sentences and for proper nouns, lower case for everything else.
- **Periods and question marks** − denoting the end of a sentence, correctly applied based on context

- **In-word punctuation** − includes apostrophes in contractions such as “that’s,” and hyphens and parenthesis in phone numbers, such as (555) 123-4567

Appendix G User Questionnaires and Satisfaction Surveys

The ultimate goal of all the measures and metrics in this document is to ensure that Internet Protocol Captioned Telephone Service (IP CTS) users have the best experience possible and that the highest level of functional equivalency is achieved. To create these recommendations, the Telecommunications Caption Quality Working Group has drawn from consumer organizations, subject matter experts, and IP CTS providers. The working group did not survey a large sample of IP CTS users to obtain their input, which is a significant weakness that should be remedied in the near future.

Without input from the user base, it’s difficult to assess which metrics are most important to measure. We expect that the results of customer questionnaires will point to wide-ranging dissimilarities: different users require different features and have different expectations. For example, an office worker trying to keep up with a conference call in a busy and noisy environment may have very different needs than the grandmother who wants to talk to her grandchild. Or, a person discussing medications with their doctor may have an increased need for accuracy, while one involved in a sales pitch may require faster captions. In addition to surveying the user base in order to determine which metrics are important, we can also establish baselines for user satisfaction.

This could be an important metric for determining the suitability of IP CTS services. Once the Telecommunications Relay Services-User Registration Database (TRS-URD) is established for IP CTS, there will be a centralized repository of users’ contact information. We recommend hiring an independent customer research company to devise and administer a randomized questionnaire for active users, with the primary purpose of discovering both overall satisfaction and which metrics are most important to the users.

We recommend that the questionnaire then be refined, based on results, and administered on a periodic basis.

Appendix H Caption Accuracy Measures

The Working Group reviewed and discussed multiple measures for caption accuracy that are identified in the following sections.

Word Error Rate (WER) and Related Measures

There are several related variations of WER, each with slight differences in prioritization of error types.

Word Error Rate

WER provides a calculation of the number of insertions, deletions, and substitutions as a percentage of words in the utterance being scored.

WER = (Insertions+Deletions+Substitutions)/Total Words:

WER may score >100% if there are a large percentage of insertions. For example, “Hello” captioned as “Hi Ho” has 2 errors/1 word for a 200% WER.

WER can provide consistent scoring across analysts but does not account for word importance (“I **am not** allergic to penicillin,” captioned as “I **not** allergic to penicillin” or “I **am** allergic to penicillin” provides the same score. At high error rates, insertions effectively become weighted more heavily than substitutions and deletions.

Related links:

- Scoring Toolkit: https://github.com/usnistgov/SCTK

Word Correct Rate (WCR)

WCR is like WER but omits insertions.

WCR = (Deletions+Substitutions)/Total Words

WCR, unlike WER, cannot score higher than 100%. For example, “Hello” captioned as “Hi Ho” has 1 substitution error/1 word (extra word not scored) for a 100% WCR.

WCR has benefits and limitations like WER but ignores the impact of inserted words.

Related links:

- Scoring Toolkit: https://github.com/usnistgov/SCTK

Match Error Rate (MER)

MER is like WER, but the score cannot exceed 100%.

*MER = (Insertions+Deletions+Substitutions)/ (Total Words**+Insertions+Deletions+Substitutions**)*

For example, “Hello” captioned as “Hi Ho” has 2 errors/(1 word+2 errors) for a 66% MER.

WCR has benefits and limitations like WER.

Related links:

- https://www.researchgate.net/publication/221478089_From_WER_and_RIL_to_MER_and_WIL_improved_evaluation_measures_for_connected_speech_recognition/link/00b4951f95799284d9000000/download

- Python Library https://github.com/jitsi/jiwer

Weighted Word Error Rate and Related Measures

Weighted Word Error Rate (WWER)

WWER is like WER, but errors are categorized by type, with each category having an importance (weight) assigned.

While like WER, WWER can potentially more accurately reflect human comprehension based on categories and weights (*more research required*), WWER required a subjective assessment of error classes, which may lead to an inconsistent scoring across analysts.

Related Links:

- Caption Accuracy Metrics Project: Research into Automated Error Ranking of Real-time Captions in Live Television News Programs (wgbh.org) (table on page 14)

Automated Caption Evaluation (ACE)

ACE is an example of WWER, with weighting defined by linguistic model. ACE Defines word importance and semantic distance, possibly providing better differentiation of error classes. For example, “I am not allergic to penicillin” scores higher (worse) when “not” is deleted than when “am” is deleted. ACE is more complex to calculate than WER.

Related links:

- Usability evaluation of captions for people who are deaf or hard of hearing: http://www.sigaccess.org/newsletter/2018-10/kafle.html
- Word Importance Modeling to Enhance Captions Generated by Automatic Speech Recognition for Deaf and Hard of Hearing Users”: https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=11438&context=theses https://aict.gallaudet.edu/research/presentations/2021/CaptionsMetrics.pdf
- Kafle, S., & Huenerfauth, M. (2019). Predicting the understandability of imperfect English captions for people who are deaf or hard of hearing. *ACM Transactions on Accessible Computing*, *12*(2), 1-32. https://doi.org/10.1145/3325862

Visible, Invisible, Minor, Essential Errors (VIME)

VIME is another example of WWER, with errors classed by type as likely perceived by the user. VIME Provides a more granular, and potentially more meaningful, score than WER does, but requires subjective assessment of error classes, which may lead to an inconsistent scoring across analysts.

Related Links:

- Appendices A, B, and C in https://ecfsapi.fcc.gov/file/10223293672032/IPCTS%20Group%20Ex%20Parte%20and%20RFP_2.23.2021.pdf
- https://ecfsapi.fcc.gov/file/1115254317141/Hamilton_ex_parte_Nov_13_2019-research_discussion.pdf

Number, Edition Error, and Recognition Error (NER)

NER accounts for "edition" errors, where the captioner omits or abbreviates information and "recognition" errors, where the captioner incorrectly captions information. NER applies a severity (weight) of ¼ for minor errors, ½ for standard errors, and 1 for serious errors.

NER = (N-E-R)/N

Unlike WER, NER considers information loss which may lead to more meaningful scores, but NER requires a more subjective assessment of error classes which may lead to an inconsistent scoring across analysts.

Related Links:

- "Accuracy Rate in Live Subtitling – the NER Model:" https://link.springer.com/chapter/10.1057/9781137552891_3
- "Measuring live subtitling quality:" https://www.ofcom.org.uk/__data/assets/pdf_file/0019/45136/sampling-report.pdf
- https://www.ai-media.tv/the-best-accuracy-measurement-for-captions-yet-the-ner-model/

Word Information Loss (WIL)

WIL is a probabilistic method that attempts to account for information lost due to errors in the captions, as well as avoiding overweighting insertions at high error rates.

WIL is like WER but avoids overweighting insertions at high error rates.

Related links:

- International Conference on Natural Language and Speech Processing, ICNLSP 2015, Automatic Speech Recognition Errors Detection and Correction: A Review § Morris, A. C. (2002). *An information theoretic measure of sequence recognition performance* (No. IDIAP-COM 02-03). Retrieved from IDIAP website: https://infoscience.epfl.ch/record/82766
- **Relative Information Loss (RIL)** − **"***Relative Information Lost (RIL), is based on Mutual Information (I, or MI) [7], which measures the statistical dependence between the input words X and output words Y, and is calculated using the Shannon Entropy H*" (International Conference on Natural Language and Speech

Processing, ICNLSP 2015, *Automatic Speech Recognition Errors Detection and Correction: A Review)*

Domain-Sensitive Error Rate Measures

Domain-Sensitive Error Rate measures are used in cases where the topic or vocabulary is known, and the success criteria is topical accuracy rather than overall accuracy of all words. For example, in a legal conversation, are the legal terms correctly transcribed, or in a medical conversation, are the medical terms correctly transcribed? Many of the models described above can be used for domain sensitive error rates by restricting the assessed words to the domain.

Keyword Error Rate (KER)

KEY is a modified WER measure for domain-specific use where keywords can be known and counted separately from non-keyword content in the text.

KER requires considerable effort to identify and classify keywords, and at error rates less than approximately 25%, WER is a sufficient approximation to KER.

Related Links:

Park, Y., Patwardhan, S., Visweswariah, K., & Gates, S. C. (2008, September). An empirical analysis of word error rate and keyword error rate. In INTERSPEECH (pp. 2070-2073).
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.4118&rep=rep1&type=pdf

Precision and Recall

Precision and Recall refers to terms used from information retrieval to evaluate accuracy over domain specific terms. Recall measures whether the term is reliably recognized (e.g., medical jargon or drug names may be mis-transcribed as phonetically similar words, reducing recall). Precision accounts for false negatives (e.g., when a key term appears in the transcript when it did not actually occur)

Appendix I Statement on Word Error Rate

This statement was approved by the following Group Members: Zainab Alkebsi, **Cristina Duarte,** Lise Hamlin, AnnMarie Killian, Jen Schuck, Neil Snyder, Michael Strecker, and Christian Vogler

The Global Alliance of Speech-to-Text Captioning, while in general agreement with the recommendations in this report, issues the following statement identifying disagreement with some specific points:

Upon review of the Final Report, the Global Alliance of Speech-to-Text Captioning agrees with the Working Group's statement that thresholds for acceptable caption quality needs to be required and there is no "one size fits all" single measure. Communication access is subjective based on an individual's hearing loss level. The lowest level of captioning accuracy able to be achieved should not be used as a minimum requirement. This would leave out millions of Americans who require better accuracy for equal and effective communication on a daily basis.

As stated in Section 1, Recommendation 4a, the evaluation method to be used for IP CTS captioning accuracy is WER and other measures may be adopted. As identified in the report in Section 2, Caption Characteristics, there are many components that lead to equal and effective communication access. These characteristics are not included in the WER scoring. The Global Alliance of Speech-to-Text Captioning opposes the sole use of WER, as identified in Appendix C, as the best method to evaluate accuracy.

As stated in the report, the WER does not account for the importance of specific errors. The specific error of leaving out the word "not" in a sentence could lead to catastrophic results. Formatting of phone numbers as a single unit without identifying features indicating the numbers are a telephone number are not considered an error. A string of numbers without identifying features indicating what the numbers are leads to confusion. Disfluencies are not counted as errors if included. Including them, unless relevant to the conversation, may lead to incomprehension of the conversation.

If the intent of this research and testing, and use of taxpayer dollars, is to ensure that the one out of every five Americans who have hearing loss receives equal and effective communication access when using IP CTS, then the testing score cannot be based on WER alone. The results need to be meaningful and not just a baseline "score" of evaluating words on a screen.

Acronym List

Aconym	Definition
ACE	Automatic Caption Evaluation
ANSI	American National Standards Institute
ASR	Automated Speech Recognition
CA	Communications Assistant
DHH	Deaf and Hard of Hearing
FCC	Federal Communications Commission
IP CTS	Internet Protocol Captioned Telephone Service
KER	Keyword Error Loss
MER	Match Error Rate
NER	Number, Edition Error, Recognition Error
NSI	Non-Speech Information
SCTK	Speech Recognition Scoring Toolkit
SME	Subject Matter Experts
TRS	Telecommunications Relay Services
VIME	Visible, Invisible, Minor, and Essential
WCR	Word Correct Rate
WER	Word Error Rate
WIL	Word Information Loss
WWER	Weighted Word Error Rate

NOTICE