Caption Accuracy Metrics Project

Research into Automated Error Ranking of Real-time Captions in Live Television News Programs

The Carl and Ruth Shapiro Family National Center for Accessible Media at WGBH (NCAM)
By Tom Apone, Brad Botkin, Marcia Brooks and Larry Goldberg
September 2011

Summary

Real-time captioned news is a lifeline service for people who are deaf or hard of hearing, providing critical information about their local communities, national events and emergencies. Captioning mandates designed to level the playing field have resulted in rapid growth of the caption industry, but a shortage of skilled real-time stenocaptioners and industry price sensitivity have made the lack of quality of live captioning on news broadcasts a growing issue.

Disability organizations have filed complaints with the Federal Communications Commission (FCC), reflecting frustration with chronic problems related to live captioning quality, transmission errors, and timely response to their concerns. However, without a common means of measuring accuracy and quality, broadcasters, consumers and regulators have no efficient method of tracking and improving real-time caption accuracy.

The project scope was to:

- develop an industry standard approach to measuring caption quality, and
- use language-processing tools to create an automated caption accuracy assessment tool for real-time captions on live news programming.

These outcomes will greatly improve the ability of the television industry to monitor and maintain the quality of live captioning it offers to viewers who are deaf or hard of hearing. It will also ease the current burden on caption viewers to document and advocate for comprehensible captions to ensure they have equal access to important national and local information. Additionally, it will improve the ability of caption vendors to differentiate their services in the marketplace.

The project built a data set from 47 television news programs. This paper summarizes the research and development WGBH undertook with Nuance Communications, and the resulting automated caption accuracy assessment tool the project produced, ***CC Evaluator™*.**

Part 1: Word Error Rates

Need for consistent error identification

Most closed captioning is done by skilled Court Reporters who have been trained in live, real-time writing. Traditional Court Reporters utilize machine shorthand to transcribe speech. These reporters have the luxury of going back over their shorthand notes and translating those notes into a clean English transcript after the fact. Stenocaptioners do not have such an option; their keystrokes are immediately translated into English by computer and sent out for all to see. Fingering errors or incorrect matches by the computer software result in errors in the captioning. (Corruption of data during transmission often compounds and/or mimics stenocaptioning errors.)

Steno software used by live captioners has built-in error reporting capabilities, but this reporting is based solely on whether a set of keystrokes “translates,” or matches, to an English word. This reflects how large and robust the stenocaptioner’s dictionary is, how well s/he has memorized and practiced the assigned set of keystrokes, and how accurately s/he is fingering the keyboard. However, correctly stroked words that do not have a match in the dictionary will report as an error -- yet may be completely readable. “Mis-strokes” may report as an error (if there is no incorrect match) or they may translate as something wrong (the mis-stroke actually matches something but not the intended word) and be a false positive – an error that doesn’t report as such.

This has engendered a system of calculating errors that does not adequately represent the quality of closed captioning. Some items reported as errors are not, many errors are ignored because they translated with a match (but are actually the wrong word) and anything the captioner has omitted is not reported at all. A primary goal of the project was to develop an error reporting system that captured all differences between the caption transcript and what was actually spoken.

Standard error classes – S, D, I

A basic calculation for Word Error Rate (WER) has been in use for some time, in large part due to the growth of speech recognition technology. In this approach a “hypothesized” transcript (captions, in our case) is compared to a “reference” transcript (an exact transcript of what was actually spoken). This is sometimes called the “ground truth” transcript.

All errors can be then categorized in one of three ways:

1. Substitution – one (incorrect) word in the “hypothesized” transcript has been substituted for a correct word in the “reference” transcript
2. Deletion – one word has been deleted or omitted from the “hypothesized” transcript
3. Insertion – one word has been inserted into the “hypothesized” transcript that was not spoken

The traditional Word Error Rate calculation is then performed:

Word Error Rate = (Sub + Del + Ins) / N

where N is the total number of words in the reference transcript.

While this is a straightforward and reliable metric, it does not adequately reflect the quality of a transcript because it treats all errors the same.

Here are some examples to consider, taken from actual newscasts. The top line is what was spoken; the second line is what appeared in the captions.

(Actual): **THIS PROCESS WILL BE QUICK.**

(Caption): **THIS PROSWILLING QUICK.**

In this case, with the nonsense word “PROSWILLING,” a caption reader would likely be very confused and unable to make much sense of the caption.

To calculate the WER, we would first align the text as best we can, lining up correct words that match:

(Actual): **THIS PROCESS WILL BE QUICK**.

(Caption): **THIS PROSWILLING **** ** QUICK.**
S D D

This identifies the errors more clearly. There is one substitution (“PROSWILLING” has been substituted for “PROCESS”) and two deletions (WILL and BE have been left out).

Using the WER formula, there are 3 errors out of 5 words, or a 60% Word Error Rate.

Here is another example:

(Actual): **SMOKING DEATH RATES HAVE CONTINUED TO INCREASE**

(Caption): **THE SMOKING DEATH RATE HAS INCREASED**

In this case, the caption is understandable and captures the gist of the sentence, though there is clearly more information in the actual text.

The texts are aligned and errors are tagged as follows:

(Actual): ***** SMOKING DEATH RATES HAVE CONTINUED TO INCREASE**

(Caption): **THE SMOKING DEATH RATE HAS ************* ** INCREASED**
I S S D D S

The result is 6 errors out of 7 words, for a WER of 85%. But the substitutions and deletions are minor and, clearly, this caption is more accurate than the previous example, so a better way of calculating error rates is desirable and necessary.

NCAM expanded error classifications

The project examined errors in great detail and, informed by consumer research summarized below, categorized them into 17 different subcategories, all of which remain as members of one of the three main categories (S,D,I). We drew on extensive knowledge of captioning and built upon error types recognized by earlier research, including the NCRA breakdown of error types.[1]

Insertion errors are rare in captioning and usually occur in conjunction with another error. In fact, in many cases, a typical caption error falls into multiple S,D,I categories. For example, the split of a compound word (backyard) into two separate words (“back” and “yard”) is technically two errors – a substitution (“back” for “backyard”) and an insertion (“yard” has been added as a separate word). Yet this kind of error rarely presents a significant impediment to understanding.

Deletion errors, or drops, are much more common in captioning. While stenocaptioners aim for a verbatim transcript, they are trained to produce readable text as a first priority. When speech is exceedingly rapid or the steno

[1] NCRA “What Is an Error?” http://ncraonline.org/NR/rdonlyres/7C0F0BFD-463C-495F-B37D-CA314EAB2736/0/WhatisanError_CRR.pdf

loses his/her place, he/she will paraphrase or omit less critical information, if necessary, in order to produce an understandable text.

Minor drops – asides like “of course” or “well” or “you know” – do not appreciably affect the content. Even longer drops of entire phrases and sentences may not dramatically affect understanding but meaning is inevitably lost as more text is omitted. And, of course, the loss of a single critical word can dramatically affect meaning but these instances are rare.

Substitution errors are common and range from simple punctuation, singular/plural and tense changes to wildly incorrect words that change the entire meaning of a sentence. We broke out general substitution errors into as many detailed categories as we could, in order to test whether viewers distinguished among the different types of errors and whether they rated them as significantly different.

Caption Error Types:

		Substitution	Deletion	Insertion
1	Substitute singular/plural	Yes		
2	Substitute wrong tense	Yes		
3	Substitute pronoun (nominal) for name	Yes		
4	Substitute punctuation	Yes		
5	Split compound word, contraction (Correct words, incorrect segmentation)	Yes		Yes
6	Two words from one (one wrong)	Yes		Yes
7	Duplicate word or insertion			Yes
8	Word order		Yes	Yes
9	Correction by steno			Yes
10	Dropped word - 1 or 2		Yes	
11	Dropped word(s) - 3+		Yes	
12	Homophone	Yes		
13	Substitute wrong word	Yes		
14	Not a valid word	Yes		
15	Random letters (gibberish)	Yes		
16	Word boundary error	Yes		
17	Transmission errors/garbling	Yes		

Consumer survey and results

In spring 2010, NCAM conducted a national online survey to query television news caption viewers about the types of caption errors that impact their ability to understand a live television news program. Survey results informed the definition of error types and criteria for weighting and ranking error types within the prototype automated caption accuracy assessment system.

More than 350 caption viewers from across the U.S. completed the survey. The majority of respondents self-identified as deaf or late-deafened; less than a third indicated they were hard-of-hearing. The survey presented 41 examples drawn from a wide range of major live national broadcast and cable television news programs. These 41 examples represented 17 sub-categories of common caption error types identified by the project team and advisors. Errors in 24 of the 41 examples were rated as severe by at least half the respondents. Severe errors included: garbling caused by transmission problems, nonsense syllables and words caused by stenocaptioner error, and major deletions that impact the meaning of a sentence. The least problematic errors were simple substitutions (such as the wrong tense) and errors in punctuation. The full report, “Caption Viewer Survey: Error Ranking of Real-time Captions in Live Television News Programs” is available on the Caption Accuracy Metrics project Web site.[2]

Error Coefficients and Weighted Word Error Rate

Using the information from the consumer survey, we developed a system for weighting errors in a caption transcript based on the severity of errors. We believe this provides a much more realistic error rate than the traditional S,D,I model. We also believe this is a valuable metric, even as a standalone measurement generated against a clean (ground truth) transcript, as it better reflects how "understandable" the transcript is.

While the S,D,I error rate is certainly appropriate when evaluating the quality of an automated speech recognition (ASR) transcript and the abilities of a particular speech engine (as it was originally intended to be used), it treats all errors the same. The project’s calculated Weighted Word Error Rate (WWER) ranks errors according to their severity and generates a score that is more cognizant and more appropriate for measuring how “intelligible” the text may be to a human reader – in this case, a caption consumer.

---

[2] Caption Accuracy Metrics Survey Report, available at: http://ncam.wgbh.org/invent_build/analog/caption-accuracy-metrics

The basic calculation, which is similar to the approach proposed by Fiscus and Schwartz to remove human error from WER measurements[3], takes the count of each error type and multiplies the numbers of errors for the error type (*errors*$_t$) by a "severity" coefficient (*severity*$_t$) then divides the weighted sum by N the number of words.

$$WWER = \sum_{t=1}^{ErrorTypes} severity_t * errors_t \Big/ N$$

For example, if we have 3 errors of Error Type 1 (singular/plural), we multiply 3 times 0.05 (the weight of error type 1) for a factor of 0.15. Similar subtotals are generated for each of the 17 error types. These subtotals are added together and divided by the total number of words to produce the Weighted Word Error Rate.

This method is similar to the approach proposed by Fiscus and Schwartz to remove human error from WER measurements[4]. This effectively incorporates input on human perception of the severity of an error and its perceived effect on comprehension.

Standard WER v. Weighted WER

The Weighted Word Error Rate correlates well with the traditional WER but better reflects the accuracy of a caption transcript. Many errors can be considered minor and these are given a lower weight in the calculation.

The calculation organically accounts for things like paraphrasing – as long as the key words are accurate, minor changes like reordering of words and phrases, changes in tense, or dropping extra adjectives also end up with lower weights in the calculation.

To illustrate, here is the Weighted WER calculation for the previous example involving smoking death rates:

(Actual) ***** SMOKING DEATH RATES HAVE CONTINUED TO INCREASE**

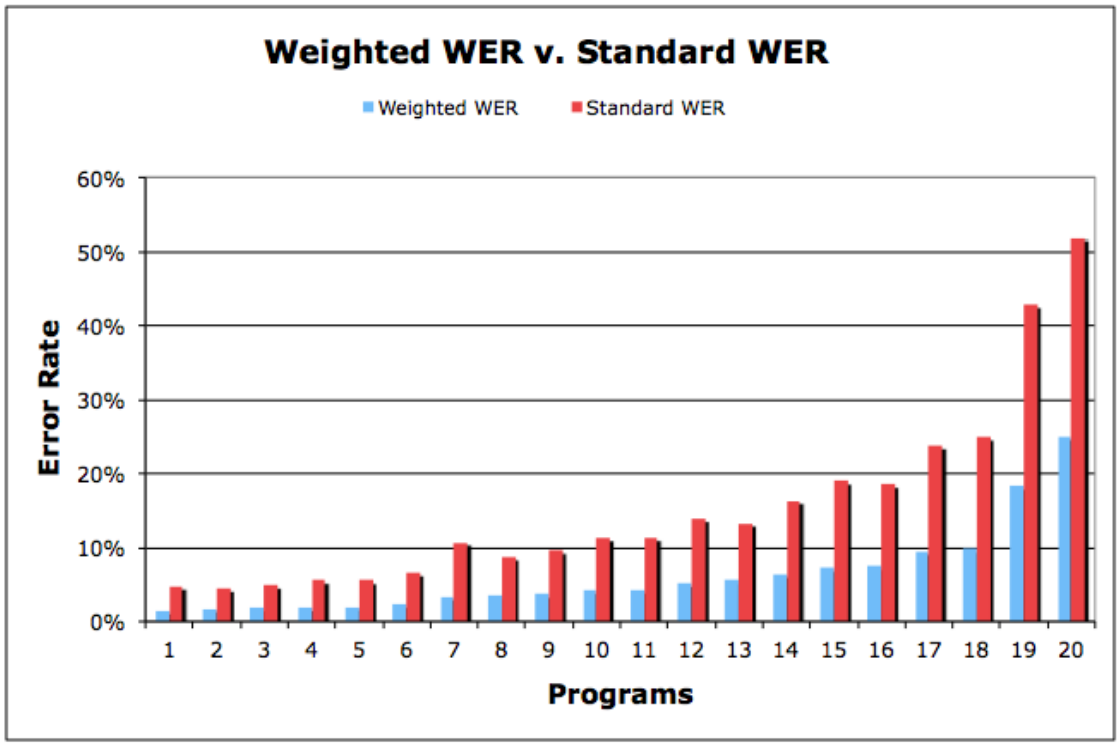
(Caption) **THE SMOKING DEATH RATE HAS ************* ** INCREASED**
I S S D D S

[3] "Analysis of scoring and reference transcription ambiguity", J. Fiscus and R. Schwartz, Proceedings of the 2004 Rich Transcription Workshop
[4] "Analysis of scoring and reference transcription ambiguity", J. Fiscus and R. Schwartz, Proceedings of the 2004 Rich Transcription Workshop

1 insertion – error type 7 = 0.246
1 singular/plural – error type 1 = 0.05
2 wrong tense – error type 2 = 2 * 0.057 = 0.114
2 drops (minor) – error type 10 = 2 * 0.39 = 0.78
WWER = (.246+.05+0.114+.78)/7 = 0.17

The WWER is 17% compared to the standard WER of 85% which better reflects the overall quality of the sentence.

The following graph shows how WWER (Weighted Word Error Rate) compares to the standard or traditional S,D,I word error rate for 20 sample programs, simply labeled 1 through 20 and organized by increasing error rate. Here accuracy is measured against a clean transcript and two different calculations are performed for comparison -- the Weighted WER in blue and the standard WER in red.



These 20 samples were taken from network and cable news programs over a one-month span and clearly show a wide range in error rates, which means a wide range in quality. The first 11 programs have a Weighted WER ranging up to about 5% (and a traditional WER under 10% or 11%). These programs present a relatively high level of accuracy and would be considered acceptable quality captioning by most viewers.

Programs 12 through 18 range from 6% to 10% on our WWER score and are starting to demonstrate some problems for viewers. The last two programs are probably poor or unacceptable quality by any measure – with Weighted WERs near 20% and standard WERs in the 40-50% range. (When error rates get this high in a caption transcript, the vast majority of the errors are deletions, or drops.)

Further study would be necessary to find the exact ranges that most viewers consider acceptable caption quality.

Part 2: Estimating Weighted WER using ASR

Need for automated tool

Calculating error rates is a manually intensive and time-consuming process. Generating a clean "ground truth" transcript alone typically requires several hours of labor; as high as 50 times real-time for research-grade transcripts[5]. Alignment tools exist (from NIST and others) but do not necessarily tally the error totals beyond S,D,I errors. As far as we know, no other tools break down error categories in more detail than Substitutions, Deletions and Insertions.

Rather than use humans to create ground truth, we used technology designed for ASR to infer the error types made by the closed captioning. The ability to reliably estimate caption quality using an ASR transcript and statistical analysis eliminates much of the time and effort involved in the evaluation process. It can, therefore, significantly reduce the costs associated with monitoring caption quality. It is important to note that the project did not use speech recognition technology to create real-time captions.

The use of ASR transcripts in measuring caption accuracy

Transcripts created by automatic speech recognition systems have inherent defects and in most cases are less accurate than a comparable caption transcript. Because the caption transcript was produced by a human operator, there is significant intelligence and judgment on display. For example, a captioner may paraphrase to make a sentence clearer or may drop less critical passages to keep up with a fast-paced discussion. The ASR engine will dutifully churn through every utterance and try to make some sense of it all.

The question becomes: how can you rate the quality of captions against an ASR transcript that might actually be worse that the captions themselves? The answer is that, by carefully analyzing where errors occur in each transcript, we can statistically account for much of this difference.

[5] "Transcription methods for consistency, volume and efficiency", Glenn et. al., Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)

There are several positive factors to note. ASR is particularly good at displaying a word for every utterance, so we find that the overall word count from ASR matches very closely to the actual word count for a given program. Deletions are a major component in the caption error rate, so at a minimum, this allows us to accurately estimate the deletions in the caption transcript.

We also know that insertions are rare, usually less than 1% of the total word count. This leaves us with substitutions as the major unknown. Our initial studies indicate that substitution errors frequently happen at different places in the transcripts – that is, when the caption is wrong, the ASR may be correct and when the ASR is wrong, the caption may be correct.

Through careful analysis and statistical modeling, NCAM and Nuance have developed methods to compare and adjust for these differences and produce a reliable estimate of WWER, using automated ASR tools to stand in for the labor-intensive, manual production of ground truth transcripts.

Nuance partnership

NCAM selected Nuance Communications as its technical partner for its expertise in customizing language processing tools and data analysis software, and for its seasoned professional services staff. WGBH and NCAM commenced work together in November 2009, developing the mechanism for automating the evaluation process, creating reports and shaping an eventual online service for an array of potential clients.

Capture Station and CC Evaluator tool

Components of the Capture Station

To build the data set of television programs for analysis, NCAM assembled a capture station to record complete programs, and then create separate files for each program's caption text files and program audio.

- Digital video recorder (DVR)
- PC running WordMeter and Audacity™
- EEG 241DR (data recovery device)

Steps in the evaluation process

- Obtain a clean (accurate word for word) transcript, if possible
- Gather evaluation elements: mp3 audio, closed caption transcript
- Run CC Evaluator tool

The CC Evaluator produces standard WER and Weighted WER metrics if supplied with the caption transcript and clean text file. Additionally, if an ASR transcript is provided (from an audio file and the Nuance tools), then an Estimated WWER is generated. (It also produces metrics for an ASR-to-clean transcript comparison, if available, which allows us to monitor how well the ASR engine is performing.) An Estimated WWER can also be generated with just the caption text and the ASR transcript – no clean text file is required.

"Adjustment factors" and accuracy estimates

By comparing the error types and totals over a large data set, the project has been able to generate a statistical model for predicting quality based solely on alignment of the caption file to an ASR output.

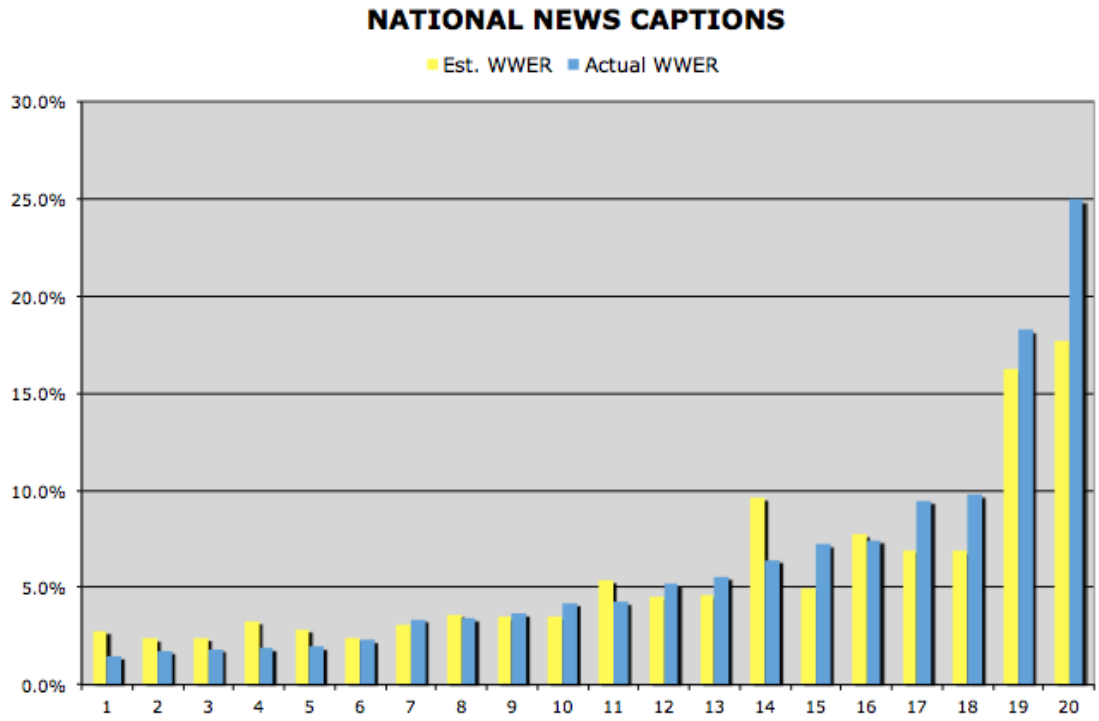
CC Evaluator employs an average difference in each error category and uses this "adjustment factor" to predict an Estimated Weighted Word Error Rate for sampled programs.

As more program samples are accumulated by the evaluation engine, better error estimates will be generated, particularly when applied to specific programs. The use of the ASR tools will also improve as the software is exposed to larger data sets.

Actual WWER v. CC Evaluator (Estimated) WWER

The following chart illustrates how the Estimated WWER (based on an ASR transcript of the program audio) compares to the Actual WWER (based on a clean transcript). Note that the method overestimated slightly for programs that had very low error rates (Programs 1-5) and underestimated for programs with very high error rates (Programs 17-20). Program #14 is the exception to this pattern, given a particularly inaccurate ASR output, largely due to unusual poor voice quality and difficult accents.

Again, both of these calculations are based on the project's Weighted Word Error Rate. The Estimated WWER is accomplished through an alignment of captions to ASR (with some additional statistical factors); the Actual WWER is performed by aligning captions with a clean transcript.



Using ASR and the project's alignment tools, this process demonstrates the ability to reliably rate caption quality. An example of how the rating system might work:

- If the Estimated WWER is less than X (e.g. 4.5%), the caption quality is deemed "acceptable."
- If the Estimated WWER is over X (e.g. 10%,), the caption quality is deemed "unacceptable."
- If the Estimated WWER is between 4.5% and 10%, the caption quality is questionable and should be examined more closely.

With this approach, the first 10 programs would be within the threshold of "acceptable" and the last two would exceed the "unacceptable" limit. Programs 11 – 18 would require closer examination — i.e., manual analysis would have to be employed to review the program captions and/or a clean transcript may need to be created in some cases to measure the WWER more precisely.

This proposed breakdown is based on WGBH's extensive experience with captioned programs but the categories can continue to be refined as the system collects more data from any particular program.

Conclusion

The Caption Accuracy Metrics project's research and development activities demonstrated the proof of concept that text-based data mining and automatic speech recognition technologies can produce meaningful data about stenocaption accuracy, meeting the need for a system of caption performance metrics.

Originally conceived as a tool for local site implementation, e.g., at broadcast facilities, advances in the Software as a Service (SaaS) model (also referred to as "on demand software") have demonstrated the viability and economic advantages of hosted data analysis services such as the ***CC Evaluator.*** NCAM's goal in having conducted the Caption Accuracy Metrics project is to continue to advance opportunities for caption viewers to have equal and reliable access to important national and local information.

For more information and additional reports from the Caption Accuracy Metrics project, visit the Caption Accuracy Metrics Web site at http://ncam.wgbh.org/invent_build/analog/caption-accuracy-metrics.

For more information about the ***CC Evaluator,*** contact:

Larry Goldberg, Director
Carl and Ruth Shapiro Family National Center for Accessible Media at WGBH:
E-mail: larry_goldberg@wgbh.org
Phone: 617.300.3722

About NCAM

The Carl and Ruth Shapiro Family National Center for Accessible Media at WGBH is a research, development and policy development organization that works to make existing and emerging technologies accessible to all audiences. NCAM (ncam.wgbh.org) is part of the Media Access Group at WGBH, which also includes The Caption Center (est. 1972), and Descriptive Video Service® (est. 1990). For more information, visit The Media Access Group at WGBH, (access.wgbh.org).

Acknowledgements

NCAM wishes to thank Nuance Communications, Inc. for its expertise and ongoing support.

NCAM wishes to thank Jonathan Fiscus from the National Institute for Standards and Technology (NIST) for his advice and assistance in applying the results of the error severity study to weighted error weight models.

NCAM wishes to thank its project advisors, including: Gallaudet University/Technology Access Program, and the National Technical Institute for the Deaf (NTID). NCAM's additional thanks go to the participants in the project's technical review panel, which consisted of representatives from: broadcast and cable television networks; caption agencies; the National Court Reporters Association; and project advisors.

Funding was provided by a grant from the U.S. Department of Education under grant #H133G080093.