Recommended Practices for Closed Captioning Quality Compliance

BILL MCLAUGHLIN
EEG Enterprises
Brooklyn, NY

Abstract **-** *In 2014, the United States Federal Communications Commission (FCC) issued a detailed set of new rulings on quality standards for television closed captioning. These rulings defined caption quality in terms of accuracy, synchronicity, placement, and completeness, and provided a set of best practices for improving quality in each of these areas. This paper provides a practical guide to the new FCC framework for judging caption quality, and highlights areas where widely deployed current practices in broadcast technology and workflows may be lacking. A detailed survey of the new generation of state-of-the-art real-time captioning equipment, software, and techniques is provided to assist video providers and distributors seeking to ensure compliance. Specific implementation suggestions are also given for improving word accuracy, reducing caption delay, and managing new completeness requirements for live newscasts that include unscripted or ad-libbed segments.*

Introduction

Since the mid-1990s, most US video programmers have been subject to FCC rules requiring closed captioning on most or all of their content transmitted through over-the-air broadcast, cable, or satellite television. These rules describe in detail the nature of video programs covered by the regulations, schedules for compliance, and the technical standards for caption delivery in both analog and digital television [1]. The FCC begins with consumer complaints submitted online to launch investigations into possible violations. Legal responsibility for a captioning violation falls on “last-mile” video providers, such as local broadcast licensees and multichannel pay TV distributors. These entities in turn typically place a contractual obligation on their programming sources to provide compliant captioning at the original point of ingest.

For years, many consumer captioning complaints have been difficult to assess, because the original FCC rules did not provide any official guidance on how to judge the quality of closed captioning on a program, or on what levels of accuracy and ease-of-reading were required to meet broadcasters’ obligations.

The US DTV transition was finalized in 2009 and led to technical changes in the way closed captioning signals were encoded in video, but did not always improve the service for consumers. The switch to compressed digital video actually increased the delay of captioning on many channels, while the rise of VOIP telephone networks increased the frequency of data gaps and disconnections on the modem systems widely used for live caption text transmission.

In 2014, the FCC acted on several years of consumer complaints on poor caption quality with a series of new rules clarifying “non-technical” requirements of closed captioning [4]. These new rules focus on standards of quality for the overall, end-to-end consumer experience of television closed captioning. The quality requirements also apply to Internet protocol (IP) delivery of video programs that also appear on television. Currently the captioning rules only cover full-length programs viewed on the web [2], but beginning in 2016, clips of television programs will also be covered [3].

The primary focus of this paper will be on developing an understanding of the February 2014 Closed Caption Quality Report and Order (R&O) that set out the new non-technical caption quality requirements, and on highlighting practical implications for widely deployed video production workflows. In some cases, the R&O has provided specific guidance about workflows that no longer meet regulatory muster. Many other workflows and procedures will continue, but must be modified by video programmers, industry vendors, and closed caption service providers to meet new standards. For this reason, it is widely expected that as these new rules phase into effect in 2015 through 2017, industry practices related to the creation of captioning are entering the most significant period of change since the service became widespread over 20 years ago.

FCC Caption Quality Framework – Four Attributes

The February 2014 Closed Caption Quality Report and Order defines a framework for describing closed captioning quality through four independent attributes:

- **Accuracy**: Captioning matches the spoken words or song lyrics, without paraphrasing, and with proper spelling and punctuation. Captioning also covers nonverbal aspects of the audio track such as the identity of off-screen speakers and descriptions of music, sound effects, and audience reactions.
- **Synchronicity**: Timing of caption text display coincides with corresponding spoken words and sounds to the greatest extent possible, and all text displays at a readable speed.
- **Completeness**: Captioning covers all segments of the program from beginning to end, including portions directly abutting any promotional breaks.

- **Placement**: Captions do not block important visual content including faces, news crawls, banners and other textual graphics, and credits.

The R&O specifies that completeness should be 100% of both scripted and unscripted programming, but specific quantitative thresholds for compliance in the other categories are not yet set. An FCC Further Notice of Proposed Rulemaking (FNPRM) issued along with the R&O suggested that quantitative targets may be set in the future for all of these attributes on a per-program or per-channel basis. However, the FNPRM also acknowledged that various industry and consumer comments have shown significant differences in proposed measurement frameworks. Many difficult issues such as the subjective adequacy of music and sound descriptions, or the least obtrusive caption placements, are out of the scope of this paper. With these limitations in mind, Table I provides a simple set of suggested metrics for quantitative comparison of different caption sources.

TABLE I

Quality Attribute	Quantitative Measurement Suggestion
Accuracy	Word error rate
Synchronicity	Average word delay
Completeness	Percentage of program duration transcribed, with inadequately described music or sound effects counted as not transcribed
Placement	Percentage of captions obstructing a speaker's face or any text graphics on the screen

ENT Requirement Changes

In addition to setting up a general framework for defining caption quality, the February 2014 R&O requires changes to several specific closed caption production workflows, particularly for live and near-live programming. The most significant issue for many small and medium-sized broadcasters will be a narrowing of acceptable practices for newscast captioning with teleprompter scripts, known formally as Electronic Newsroom Technique or ENT.

ENT is currently one of the most common sources of closed captioning text for local broadcast news programs aired outside of the top 25 US television markets (the FCC does not allow large market stations to use ENT as a captioning source). ENT provides cost-efficient closed caption coverage by re-purposing text from prompter scripts used by on-air talent. Most teleprompter control systems have the ability to send this text over a serial cable or TCP/IP connection to a hardware closed caption encoder, which then creates a CEA-708 compliant scrolling caption display.

The primary limitation of ENT captioning is that it systematically excludes caption coverage for unscripted segments including field reports, interviews, and weather forecasts. Anchor ad libs are also generally lost. These limitations have led consumer advocacy groups to request that the FCC require a larger number of stations to switch from ENT to captioning techniques based on verbatim real-time transcriptions [5].

In the 2014 R&O, the Commission declined to expand restrictions on the use of ENT, but at the same time emphasized the importance of completeness in all live captioned programs, including newscasts using pre-scripted text. The R&O specifically states that stations may not simply omit captioning for field reports, live interviews, breaking news, and weather, whether or not the on-air talent uses a prompter for these segments, and warns that violators may be subject to an order to replace ENT captioning entirely with verbatim real-time transcriptions.

Compliance with this requirement will require many stations to either place additional segment scripts into the newsroom computing system, or use a verbatim real-time captioning technique for unscripted segments or even the entire newscast. Table II summarizes a variety of approaches that mid- and small-market stations are currently choosing to meet these new requirements. Stations are finding that all options will require some additional capital expenditures, some additional staff attention, or both; decisions are made depending on the amount of unscripted material included in a newscast, the budget and staff time available, and the severity of quality and completeness compliance concerns.

TABLE II

ENT Gap Solution	External Costs	Staffing Effort	Quality Risks
Prepare scripts for all segments before they air	$	$$$$	Scripting not verbatim; risk of late-breaking segments still not being captioned
Automatic Speech Recognition Captioning	$$$	$	Accuracy typically lower and delay higher than other solutions; unusual proper names and non-verbal cues not handled well
In-house voice-recognition "re-speaking" system for unscripted segments	$$	$$$$	Accuracy of transcription varies widely with level of training possessed by specific re-speaker
Real-time captioning for prompter gaps only	$$$	$$	Precise segment timings required in advance, difficult to engage captioners for short segments on short notice
Real-time captioning for entire program	$$$$	$	Low-latency, complete, and verbatim accuracy; best quality but likely most expensive solution

Real-Time Transcription Workflow Implications

In addition to the changes required of broadcasters currently relying on ENT to meet all or part of their captioning obligations, large-market stations and networks are now also re-examining real-time captioning workflows in response to the new regulatory pressure to increase accuracy, offer perfect segment completeness, and reduce caption latency. Many broadcasters are already transitioning away from telephony for live closed caption delivery and moving towards IP-based streaming solutions, a trend that is accelerating due to the compliance advantages that these newer systems offer.

Another widespread discussion has occurred around the possible applications of Automatic Speech Recognition (ASR) technologies. ASR technology is appealing because of its potential to create verbatim live closed captioning without requiring a human stenographic operator. However, speaker-independent ASR captioning systems have achieved limited acceptance for broadcast due to a combination of lower accuracy and higher delay than stenographers. A typical NCRA-certified stenographic captioner can achieve more than 97% accuracy at rates of 180 words per minute or higher, with about 3 seconds of delay. These parameters are difficult for ASR systems to achieve, particularly in combination, because typical language models gain in accuracy only by analyzing more combinations of words over larger block sizes, which leads to increased delay when operating in real-time. Professional captioners are also accustomed to studying programming-specific vocabulary such as athlete names, local places, and neologisms in the news. Even when computerized dictionaries are updated frequently, and contain highly customized models, they will likely not measure up to this level of contextual preparation.

The caption quality R&O has generated new interest in ASR solutions due to their lower operational cost and "always-on" availability, but has also posed significant new problems for these systems. The most intractable issue may be that the R&O defines "accuracy" and "completeness" as including captioning of sounds effects, music, and the identification by name of off-screen speakers during conversations. These capabilities rely on human contextual understanding and are likely to be a long-term weakness of any fully automated caption transcription system, regardless of the level of speech recognition accuracy achieved. An issue with placement also exists - automatic systems generally will not dynamically position caption text, forcing broadcasters to use a fixed screen position for the caption display and keeping graphics out of that area for the entirety of a program.

Real-Time Captioning Communication Systems

Due to the limitations both Automatic Speech Recognition and Electronic Newsroom Technique have in meeting increasingly stringent FCC requirements, it is likely that the use of professional transcriptionists specializing in stenographic typing or voice-writing will not only remain steady but will increase in the short-to-medium term. In the United States, most national television captioning accounts are currently serviced by approximately a dozen large transcription agencies, with hundreds of smaller agencies and independent operators also providing high-quality service on a local or regional basis.

Real-time stenographic captioners typically work remotely, away from the studio or live event venue, with the broadcaster providing a technical infrastructure for reliable communication of the program audio track to the transcriber at their work site, and low-latency return of the corresponding text. Choosing a reliable and accurate service provider, and an appropriate remote communication mechanism, is probably the most fundamental determinant of the level of real-time caption quality that can be achieved. A comparison of the features and requirements of five common communication technologies is shown in Table III (next page).

The fundamental requirement of a real-time captioning communication system is to deliver the audio track reliably to the captioner. Higher quality audio enables better transcription accuracy. The latency of the audio transmission is also critical, because this latency, added with the transcriber's response time, becomes the overall caption delay on the program. This is illustrated in (1) and (2), which demonstrate how a captioner hearing audio from a consumer DTV feed will yield twice the transcription delay of the same service provider using a low-latency streaming or telephone coupler audio service. This may not be acceptable under new FCC Best Practices that require video programmers to "make commercially reasonable efforts to provide captioning vendors with access to a high quality program audio signal to promote accurate transcription and minimize latency."

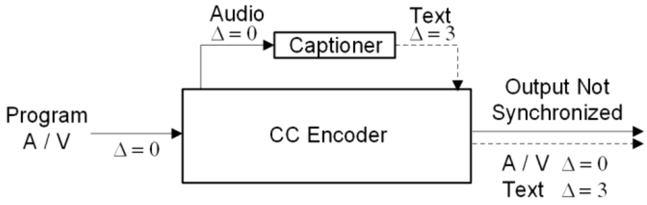
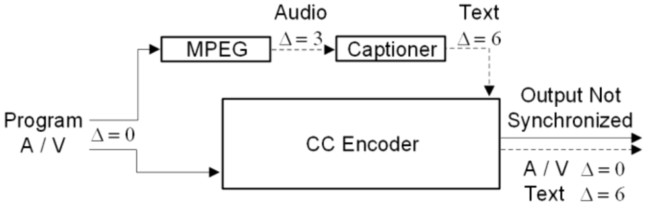


FIG 1: Sending low latency audio through a telephone coupler or iCap to captioner yields 3 seconds audio-to-text delay.
Δ = Delay (Seconds)



+
FIG 2: Captioner monitoring ATSC or cable/satellite transmission with 3 seconds delay returns text delayed approx. 6 seconds
Δ = Delay (Seconds)

Unlike audio, live video feedback to the captioner is not essential for a basic transcription, and does not directly impact caption delay. However, without video feedback, captioners cannot adapt their text positioning to avoid on-screen graphics, nor make use of on-screen graphics to assist Table 3 with proper name spelling and other relevant

TABLE III

Realtime Method	Audio Quality	Video Availability	CC Feedback	Latency	Extra Broadcast Equipment	Extra Captioner Requirements
Phone line	Limited Bandwidth	No	No	0.1 secs	Audio disembedder/mixer, telephone coupler	None
Over the Air	Broadcast Quality	Yes – Full Resolution, Requires separate monitor	Yes	2-4 seconds	None	Captioner must be in local signal area or have appropriate cable package
Satellite	Broadcast Quality	Yes – Full resolution, Requires separate monitor	Yes	2-4 seconds	None	Captioner must work for agency with satellite downlink capability
Streaming	Compressed, but higher clarity than telephone	Yes – Proxy quality	Varies	1.5-6 seconds	Live streaming appliance or PC software	Compatible PC player software (cost/licensing varies)
iCap	Compressed, High dialog clarity	Yes – Proxy quality	Yes	0.3 secs (audio) 1.5 secs (video)	SDI CC encoder or virtualized playout channel with integrated iCap driver	iCap player software (free download)

program context. The Best Practices in the caption quality R&O recommend a video feedback plan, and recognition by broadcasters and transcription agency of the value of integrated video streaming has been a large factor in the ongoing switchover from telephony to IP-based systems.

The final element in providing feedback to the remote provider is a return-path for caption text, either on a consumer set or through another mechanism. This provides a valuable confidence test that the end-to-end system is working correctly – captioners who do not have this feedback will be wholly reliant on broadcast master control operators to alert them of a problem with positioning, timing, or accuracy.

Each system described in Table III has different capabilities, and a different level of equipment investment at both the captioner and broadcast side. The “Extra Broadcast Equipment” column describes those requirements above a few basics: availability of a basic HD-SDI closed captioning encoder at the broadcaster or event venue, and a video program available to captioners through a standard consumer TV hookup. It is also assumed the captioners have a PC with real-time caption creation software, and a choice of a reliable low-latency Internet connection or two long-distance telephone lines, as described in the service provider Best Practices.

Combinations of the approaches in Table III are also common; for example, low-latency audio may be obtained through a phone line while simultaneously, a consumer television with sound muted, is being used for slightly delayed caption and video feedback. Greatly increased redundancy can also be achieved by providing captioners with the ability to connect through both Internet-based systems and telephone systems simultaneously, so that if either link becomes temporarily unreliable, the other can be used exclusively until service is restored.

Systems for Improving Caption Alignment

Since workflows using compressed consumer video feeds for audio are still in use for many live captioned programs, caption synchronicity can often be improved significantly simply through implementation of an IP system with a lower-latency audio feed. This step should enable any video program to reduce audio-to-caption delay to approximately 3 seconds, which is the approximate transcription delay of a typical real-time stenographic writer. Voice-writers and ASR systems vary, but generally have a greater delay. Delay between the audio track and live captioning is significant not only because captions are more easily understood when they are in close synchronization to the on-screen action, but also because any systematic delay in live captioning also causes issues with caption completeness at the end of live program segments. For example, if captions are 5 seconds behind the audio track throughout a segment, then the last 5 seconds of captioning for that segment may display on top of the subsequent commercial block, potentially overwriting captioning provided with a sponsor’s clip. Alternatively, if promotional segments are inserted downstream of the captioning encoder, the last section of text may be lost completely to the viewer. Both of these scenarios introduce completeness violations when two programs that are required to be captioned directly abut.

One solution to this problem described in the R&O and being put in place by some early adopting broadcasters is to provide an “advanced” audio feed to the real-time captioner. Just as delays in the audio feedback chain cause caption text to trail farther behind the audio track, a positive differential between the time that the captioner receives an audio signal and the time that the corresponding video frame is played outward through the caption encoder will decrease the caption delay seen by consumers. For a live program, the equivalent to providing an advanced audio feed can be achieved by introducing a video delay between the audio transmission point and the captioning encoder output. A fully integrated video delay feature is now included in some captioning encoders as shown in (3). Alternatively, an external source of program delay can be used as shown in (4), particularly for programs where a delay of several seconds may already be in place for obscenity filtering. In either case, if previously captioned non-live programming or commercial segments are played out from upstream, it is crucial that the overall system is capable of using metadata or tally automation to avoid adding additional captioning offset to the previously recorded material.

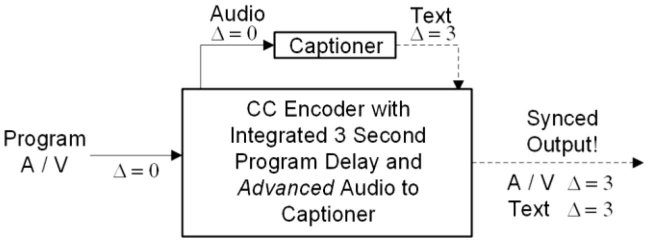

Audio
Δ = 0
Captioner
Text
Δ = 3
Program
A / V
Δ = 0
CC Encoder with Integrated 3 Second Program Delay and *Advanced* Audio to Captioner
Synced Output!
A / V Δ = 3
Text Δ = 3

FIG 3: CC encoders with integrated video delay line and advanced audio feedback capability produce the appearance of perfectly synced live captioning to the consumer.
Δ = Delay (Seconds)

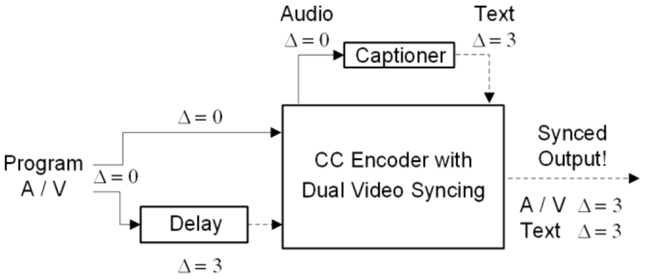


FIG 4: A CC encoder with master-source bridging combined with an external delay device also produces a synced result,
Δ = Delay (Seconds)

Conclusion

The 2014 FCC Caption Quality Report & Order introduces a host of new consideration for broadcasters in meeting caption obligations, particularly for live and near-live programming. Commercially available technical solutions have advanced in response and, compared to previous generation workflows, IP-based live captioning systems can often improve caption performance in all four of the new FCC benchmarks – accuracy, synchronicity, placement, and completeness. Further improvements to synchronicity and completeness of live captioning can be achieved through provision of advanced audio where a small additional video delay is feasible.

Finally, video programmers are advised to open a discussion with their real-time captioning providers about possible process improvements, many of which are suggested in the FCC R&O's Best Practices section. The Best Practices suggest ways to measurably improve captioning quality through simple steps that often have little or no added cost, such as providing advanced scripts where available, tuning equipment correctly for optimal audio feedback quality, and establishing a clear fail-over scheme and alert protocol when technical or operational issues are detected.

Closed captioning is likely to stay near the forefront of video provider's regulatory concerns going forward, as the FCC R&O also contained a "Further Notice of Proposed Rulemaking" (FNPRM) section. The FNPRM requested further input from industry and consumers regarding motivations, feasibility, and measurement procedures for the implementation of official quantitative targets for caption quality. Increased restrictions on ENT also continue to appear on the agenda for consumer groups unsatisfied with the current level of live newscast captioning outside of top markets. With the possibility of an increasingly stringent regulatory environment ahead, accessibility technology and protocols will remain a high priority for broadcast engineering and operations staff, industry vendors, and standards bodies.

REFERENCES

[1] Jones, Graham, "Implementing Closed Captioning for DTV", *NAB Broadcast Engineering Conference Proceedings 2004*

[2] FCC 12-9 – 2012, Report and Order, Closed Captioning of Internet Protocol-Delivered Video Programming, Federal Communications Commission

[3] FCC 14-97 – 2014, Second Further Notice of Proposed Rulemaking, Closed Captioning of Internet Protocol-Delivered Video Clips, Federal Communications Commission

[4] FCC 14-12 – 2014, Report and Order, Declaratory Ruling, and Further Notice of Proposed Rulemaking, Closed Caption Quality, Federal Communications Commission

[5] FCC 05-142 – 2005, Notice of Proposed Rulemaking, Closed Captioning of Video Programming; Telecommunications for the Deaf, Inc. Petition for Rulemaking, Federal Communications Commission