Lexi Automatic Captioning: The Timeline to Primetime

EEG Video

July 2020

Introduction

Closed captioning has been required on nearly all live TV broadcasts in the United States for almost 25 years. In other English-speaking countries, including Canada, Australia, and the United Kingdom, an entire generation of viewers has also grown up expecting and depending on closed captioned news and entertainment. And in recent years, live closed captioning services have been expanded through both regulatory and market-driven measures into streaming video services, public events and religious gatherings, and into enterprise video and the workplace.

Live closed captioning is rapidly being adopted across these applications and more to provide a better viewing experience. Still, many questions persist around how live captioning is created and distributed, such as:

- Why isn't all live video captioned?
- Can't it all be done through familiar consumer-grade voice recognition (speech-to-text) software?
- Why is the understandability or usability of captioning in so many places often poor?

In this paper, we will discuss costs and obstacles that have slowed the diffusion of live captioning in many domains where live video is becoming increasingly prominent and less expensive to produce. We will specifically focus on how EEG's Lexi Automatic Captioning technology can help overcome common obstacles. In addition, we will present an overview of caption quality evaluation frameworks that provides an easy reference to see the results

currently achievable with a well-designed, built-to-purpose Automatic Speech Recognition (ASR) live closed captioning solution.

Live Captioning Landscape

Live closed captioning promotes accessibility, improves language learning, aids retention of information in videos, and is required in many legal jurisdictions. Live television particularly has a well-developed regulatory regime in most countries requiring nearly total caption coverage, and a well-oiled operation flow related to the technical details of caption production. So why doesn't all live video follow the broadcasting path and include closed captioning or subtitling?

Let's review the major roadblocks to comprehensive closed captioning coverage. Key factors include:

Cost

Depending on the quantity of content and access to captioning resources, ensuring that all videos, events, and more are compliant can become quite expensive. This may deter content providers from adding captions until they deem it absolutely critical. Unfortunately, this problem can only be solved by increased reliance on technology solutions for captioning, since traditional skilled human transcriptionists form a supply-limited labor force and there is little evidence to suggest a strong future influx of new professionals into this field.

Workflows

Adding captioning to live video production is a group effort: It requires a strong link between the video engineering teams that plan workflow and capital equipment spending, with the operational teams that will trigger captioning. Even in high-profile broadcasting, captioning operations are often planned on whiteboards and communicated through ad hoc email requests to a vendor. Additionally, software-based automatic systems routinely lack a crucial integration layer to broadcast automation, GPI-driven systems, and other playback controls.

Regulatory Issues

Understanding whether captioning is required by law at all, and what exactly the law asks, is complicated in many uses of live video, especially outside of traditional TV broadcast.

Many types of communications are held to a “reasonable accommodation” standard by the Americans with Disabilities Act (ADA). This standard provides only limited guidance on the specifics of closed captioning service, making it critical to anticipate and measure actual user reaction. What is considered reasonable can vary considerably depending on the size of the sponsoring organization, and in many cases there is no definitive way to resolve the question outside of the context of a specific complaint being brought to the court system.

In the United States, television programs covered by FCC guidelines both benefit and are challenged by a much clearer standard. The FCC generally requires all programs on broadcast and cable/satellite television systems to be live captioned outside of certain limited exceptions for recent startups or very small broadcasters. The style and quality of the closed captions are further constrained by several Report and Order documents describing “Best Practices” for maintaining high caption quality. This paper will give extensive examples on how Lexi helps compliance with these measures.

Some countries, including Canada, now supplement television regulations for high captioning coverage and uniform methods with a results-based quality grading system called “NER.” The NER framework evaluates the accuracy of a transcription based on subtracting points from a perfect score of 100 for each instance where non-verbatim captions impair the reader’s ability to understand the transcribed dialog in whole or in part, or fail to provide smooth readability.

The imperative task of making content accessible through captioning is clearly a challenging one. Today’s increasingly sophisticated automatic captioning services are making it easier to meet that mandate, but content providers must choose their solution wisely: The optimal automatic captioning service should be cost-effective, provide easy automation and workflow support, including hybrid workflows with traditional human captioning, and offer well-documented compliance with US FCC and other international regulatory body standards.

Lexi Automatic Captioning

This paper gives extensive examples on how Lexi helps achieve compliance while keeping costs low, providing compatibility with other captioning methods, and abiding by FCC rules.

Lexi Automatic Captioning is EEG's AI-powered live captioning service that delivers captions with more than 90% accuracy in English, French, and Spanish for many common media types, making the service optimal for improving compliance and accessibility on currently uncovered material. As a cloud service, Lexi is frequently updated with new data for global news and entertainment. Lexi also ingests user-supplied text documents and URLs to absorb and leverage new data to match current on-air media transcriptions.

Broadcasters, government, education, corporate, non-profit, and other media creators use Lexi to caption a wide range of content. This solution has proven to improve compliance and accessibility on material that previously went uncovered by captions. As a cloud-hosted service, Lexi's extreme ease of use has made it optimal to ensure captioning coverage for:

- unscripted broadcast segments (such as weather reports),
- complete accessibility to academic lectures,
- full compliance for municipal meetings (city, county, state),
- and wholly inclusive corporate meetings, events, and presentations,

among a constantly expanding slate of applications where regulations or best practices require closed captioning.

Lexi has consistently made great strides in closed captioning, sparking key advancements to improve accessibility. Lexi is differentiated from other automatic captioning systems, not only through its broad integration and high accuracy, but also through many advanced features. Highlight capabilities include:

- intelligent read speed pacing, which ensures that on-screen text appears at a smooth pace–other ASR systems may suddenly display multiple sentences in groupings too large for the viewer to both read and keep up with the accompanying visual content,

- custom vocabulary Topic Models for unprecedented real-time accuracy,
- and Lexi Vision, a system for processing live video to determine optimal caption positioning and to learn new vocabulary from on-screen text graphics.

Live Captioning Alternatives to AI Solutions

What about other systems besides Lexi? When considering the best methods to make content ADA-compliant, some users consider and explore captioning resources besides AI automatic captioning.

Human Captioning

One alternative to automatic captioning is human captioning. The best human stenographic captioners are capable of achieving accuracy rates of 95% and higher, and some are certified by the National Court Reporters Association or other groups as achieving these results in formal tests.

In practice, though, human caption service customers may experience a variety of results depending on budget and the individual human captioner. Live captioning is a difficult skill to learn and is in high demand, so not all individuals meet the standards of those at the top of the profession. Knowledgeable buyers in the field have reported results anywhere from 98% to as low as 80% from various services they have audited.

Cost and workflow concerns can also be important issues limiting the use of skilled human captioning in new environments.

Teleprompter

Another option for many news broadcasters is teleprompter captioning. With this method, the teleprompter will connect directly to a closed caption encoder and feed text equivalent to what the on-air talent sees. During prompted segments, the accuracy of this approach can exceed 99%. However, very few news programs are entirely prompted, with talent often going off-script

during weather, banter, interviews, and field clips. This can lead to problems with the FCC's "completeness" requirement when text on the teleprompter doesn't align with additional commentary or conversation.

Teleprompter text is rarely available as a solution outside of TV news environments, so this method for captioning has been shrinking as a share of the overall live captioning ecosystem.

Adding Lexi to Existing Workflows

Lexi doesn't have to stand alone–content creators who have already started captioning with other methods can work it into their existing system.

Lexi can be combined with human captioning as a fallback option in case of unexpected scheduling or reliability issues. Lexi is available at a fixed hourly rate, making it a predictable expense, regardless of whether 10 hours or 10 minutes of service are needed consecutively. With its ability to be booted up with as little as 10 seconds of advance notice to air time, Lexi is almost instantly available when other captioning resources are:

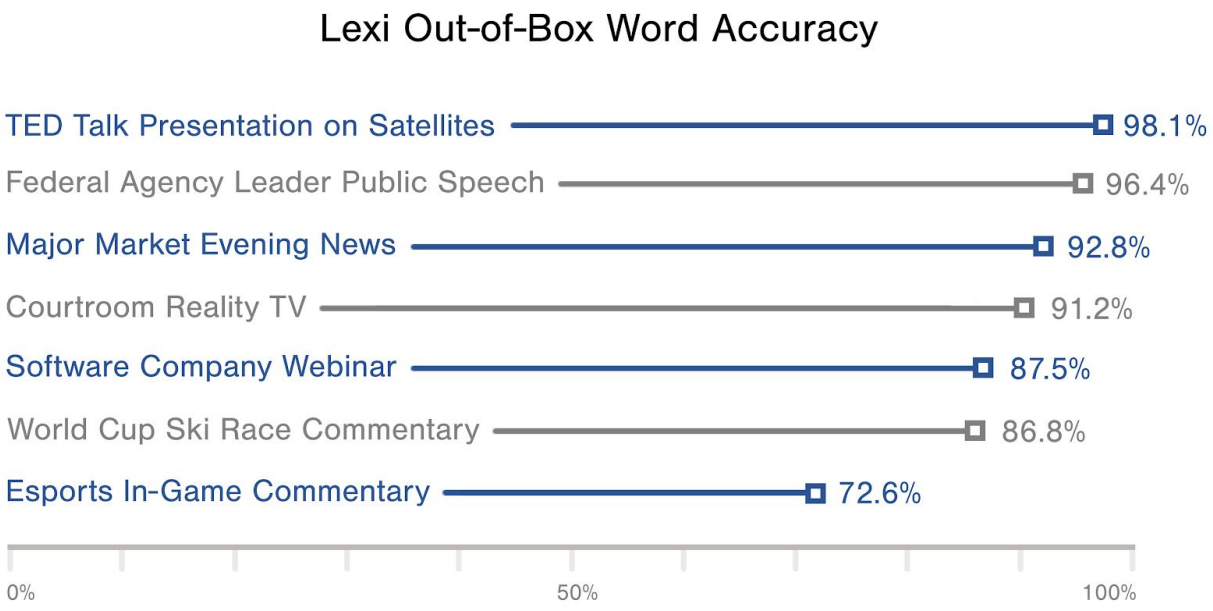
- already booked up,
- notice is too short,
- there is an unexpected staff absence,
- or even when a low-cost source for experimental or testing feeds is needed.

Measuring Automatic Captioning Effectiveness

Lexi Word Accuracy By Sample Program Type

Metrics matter when evaluating captioning, with word accuracy standing as one of the major measurements. Word accuracy in automatic captioning is a function of several characteristics

of the input program. The table below shows word error rates for 5-10 minute samples of live video clips representing various real customer content fed through Lexi. To generate this data, a human analysis was performed to create a perfect transcription of the clip, and an error was marked for every word that was different in a meaning-altering or clarity-impairing way.



These and many other examples underscore the impact of input content on Lexi accuracy. For readers unaccustomed to looking at caption accuracy measurements, we believe that a typical subjective audience response sees captioning as “useful but imperfect” beginning around 75%, “good” at 90%, and “almost perfect” around 95%.

How can you estimate how specific content would score with Lexi? The list below provides some factors that have been observed to significantly affect the out-of-the-box outcomes.

Factors Leading to Higher Lexi Accuracy

- Dialog is clear and forefront; background sounds and music are minimal
- Single speaker, or orderly transitions between different speakers (i.e., limited cross-talk)
- Evenly-paced speech, such as when reading from a prompter or script (i.e., limited disfluencies, including “umm”, “uhh”, etc.)

- Limited use of technical jargon or other unusual words (without Topic Models training)
- Mainstream American English or British English accents

Lexi Word Accuracy Before and After Use of Topic Models

Lexi's Topic Models feature is an important advancement in automatic captioning. It provides powerful tools to improve captioning by using a combination of "base" training provided for popular topics and program types by EEG's in-house expert team, and custom documents uploaded privately by an individual user for their private account. The result is a highly effective system for training vocabulary, phrases, and context that can significantly decrease the captioning Word Error Rate (WER).

For this section, four of the examples from the above "Out-of-Box Word Accuracy" chart were analyzed. EEG's staff then curated a Lexi Topic Model consisting of specific words and phrases that were transcribed inaccurately in the first sample, plus general information sourced from a website related to the video clip. A second 5-minute sample from the same program was then chosen and re-run through Lexi with the newly created and trained Topic Model.

Program Type	Word Accuracy Lexi Out-of-Box	Word Accuracy 2,000-word Topic Model	% Change
Federal Agency Leader Public Speech	96.4%	96.6%	+0.2%
Major Market Evening News	92.8%	93.5%	+0.8%
Software Company Webinar	87.5%	91.9%	+5.0%
World Cup Ski Race Commentary	86.8%	89.2%	+2.8%

In all tested cases, Lexi Topic Models successfully improved the measured accuracy.

Factors Leading to Improved Topic Models Results

- The dialog makes more-than-average use of proper nouns like names, places, and brand or product names, or of jargon and context-specific phrases.
- Starting accuracy is lower than average, providing more room for improvement.

There are instances where Topic Models' impact on improving accuracy will appear to be minimal. These include times when results are already highly accurate, or when most or all of the errors involve misidentification of common words and phrases, rather than proper nouns or jargon. When common words and phrases are misidentified, the source of the problem is more likely to be speaker pacing and clarity, background noise, or factors other than insufficient knowledge of the vocabulary and topic context.

The benefits of Topic Model improvements may also be understated by looking at aggregate Word Error Rates alone. Subjective judgements on the understandability of closed captions tend to value correct recognition of proper nouns, which often form the core subject of a sentence or phrase. Audiences may often value this quality over correct recognition of other words in the sentence, many of which serve only a connecting or subordinate purpose.

Elements of FCC Closed Caption Quality

Accuracy

Accuracy has been the measurement discussed so far in this paper. Although many discussions of caption service quality begin and end with accuracy, it is by no means the only relevant dimension.

Why is that? For instance, imagine captions with very high accuracy that:

- did not appear until 30 seconds after the spoken word;
- appeared on the center of the screen, partially covering the most important visual elements;
- or were sometimes inexplicably absent from entire sections of the program.

It's true that word accuracy is often what users notice first and consider most important. Despite superior word accuracy, however, captions with the above characteristics would by no means be considered high-quality.

Accuracy with Lexi, of course, varies widely between different content types. EEG generally guides customers to expect the 90-95% standard for live news and most types of live event presentations or corporate communications. A thorough Topic Model will often boost aggregate word accuracy another 1-2%, with performance improving most on key proper nouns, followed by common words.

Timing and Latency

They say that timing is everything, and it certainly matters for closed captioning.

Captions are easiest to understand and most enjoyable when the text appears on screen as close as possible to the corresponding video and audio program. With pre-recorded shows, captioning can usually be perfectly timed at the phrase level. But with live shows, captioning is generally somewhat behind the audio track, as all transcription methods require aggregation of some number of syllables or words to produce a coherent output. The more time that passes between audio and captions, the worse that viewers perceive their experience.

Lexi outputs captions with a delay of 3-4 seconds. This is in line with the performance of a good human captioner listening to a properly configured low-latency audio link like iCap, EEG's IP-based closed captioning and subtitle delivery network with global reach.

So, what causes delays between the audio track and captions?

In systems using complex computer models for matching speech to text, higher accuracy can often be obtained through increasing delay. This works up to a point but must be handled with care. TV viewers will generally report dissatisfaction with captions that are more than 5 seconds behind the program content. One recent study estimated that for each additional second that captions were behind, viewers perceived this with similar negativity to a 5 percentage point decrease in accuracy rate (Mike Armstrong, 2013).

When a system using human captioners has higher latency than 3-4 seconds, it is sometimes because the captioner themself is listening to delayed audio, which then is additive to the captioning delay experienced by the viewer.

Another source of highly delayed captions can be human-assisted ASR “re-speaking” systems, or earlier generation fully automatic captioning set with inappropriate amounts of delay to try to hide inadequate accuracy at lower delay levels. Some systems in the field have caption delays of 8-10 seconds, which is very distracting to users–this amount of delay on a news program can often mean an entirely different story is being discussed before captions for the previous story have finished appearing. These highly delayed captions also frequently get cut off by insertion of commercials at the end of segments.

Positioning

Location, location, location–another key factor for quality captioning.

Lexi Vision is a standard module within Lexi that uses image-processing AI to automatically guide caption positioning, for compliance with FCC mandates for optimal positioning of live captioning. Content creators using Lexi Vision can dependably avoid obstructing the faces of on-air speakers, which can be especially important for the many individuals with long-term hearing impairments who partially rely on lip-reading skills for understanding. Lexi Vision also ensures that text graphics such as news crawls, speaker identifiers, and scoreboards air unobstructed by live captioning.

Without Lexi Vision

With Lexi Vision

Most other captioning systems can only comply with these regulations via prior agreement with producers, stating that there is a single area of the screen where captions will always be placed, and that the agreed-upon area will not include elements that must remain unobstructed. Unfortunately, it is very common to find that the mandates are not actually followed consistently throughout the program.

Completeness

"Completeness" refers to the need of the captioning audience to have captioning consistently present throughout the entire program, every day. Common obstacles to completeness include captioning that sometimes does not start in a timely manner due to technical or human performance difficulties, or from workflows such as prompters that may have no data entered for certain program segments.

Lexi shines in the completeness category with a 99.95% SLA guarantee on the cloud service, a sophisticated new scheduling system that is launching in 2020, and an ability to automatically start jobs, even off-schedule, through production GPI and HTTP automation, or even detection of upstream caption outages. Some Lexi customers use prompters or human captioners for much of their work, and keep a Lexi subscription as an affordable backup to ensure uninterrupted compliance.

Looking Forward

Broadcast TV has traditionally been the main market driving live captioning innovation, so the Lexi system has attached great importance to covering all the bases for FCC on-air compliance, additional international regulatory standards such as NER scoring, and practical automation for newsroom workflows.

New industries continue to surface additional high-volume use cases for live captioning. Thoroughly modern technological approaches are critical to accommodate this increased demand, simultaneously rising to audience values of high accuracy and usability, while also providing mission-critical production performance to content producers. Lexi steps up to meet these needs on all counts, with a proven track record in high-profile broadcasting news and

sports productions, business conferences, worship services, live events, and corporate enterprise video.

While competitive solutions may remain unable to modernize or unsuited to scale beyond traditional applications, Lexi boasts full automation support across a very broad range of video insertion and live event display products from EEG and other partner hardware and software vendors. Lexi has an enthusiastic customer base that motivates and supports continuous improvement on the core metrics of caption word accuracy, latency, positioning performance, and usability to get the job done on time, every time.

The capabilities of EEG Video's Lexi provide video and event producers with a powerful, reliable, and always-available tool for promoting accessibility, inclusion, content discovery, and an all-around great viewer experience. These are advances that deliver on the promise of live closed captioning: making it affordable and practical for in-person or remote video to connect with smaller and more impromptu audiences–on any scale.