Understanding Live Automatic Captioning Quality

3PlayMedia SPEECHMATICS

ABOUT THE WHITE PAPER

This white paper, was written by 3Play Media in collaboration with Speechmatics. The brief will highlight the various types of live automatic captioning solutions, and provide data as to how they compare. At the end of this white paper, you will be able to make an informed decision on adding live captions to your video content.

20% of disabled people have canceled a streaming service subscription because of accessibility issues.

The Big Hack

TABLE OF CONTENTS

INTRODUCTION

Online video streaming has been increasing rapidly over the last several years. In the late 2010s it was predicted that **by the end of 2020 live streaming would account for 82% of all internet traffic**. The onset of the global COVID-19 pandemic in early 2020 has only pushed that number higher. New norms have been created in 2020 including remote – well – everything. This sudden and vast shift to a virtual world caused live streaming to skyrocket far beyond what was anticipated.

Between April 2019 and April 2020, the live streaming industry has grown by 99%.

Tech Jury

Live streaming has become what connects us to other people and helps us keep moving forward through a particularly challenging time. Not only does live video help us connect with our friends and family, communities, schools, and coworkers, but it also helps us connect to brands in a number of new and unique ways. The Interactive Advertising Bureau (IAB) found that social platforms are a key source of live video content. Streaming on such platforms share a different type of content. One that is often live, feels authentic, and makes us feel like we have some sort of "special access" to something.

In a world where everything has moved online, individuals are feeling more connected to those on screens than ever before. It's become more clear that online video is and will continue to grow and deliver new uses to audiences.

Brands have found that **82% of people prefer live video over standard social media posts**, and **80% of people would rather watch live video than read a blog**.

The global video streaming market size is expected to reach 184.27 billion USD by 2027.

Grand View Research, Inc.

While these newfound trends bring a lot to the table, there is one piece we must not forget. A critical piece of the conversation around live streaming is making that video content accessible. As technology evolves, and we rely on it more and more, we must be sure to include those with disabilities, including hearing loss, to make this content both accessible and legally compliant. One way to do so is by providing live captioning on streaming video.

Throughout the rest of this white paper we will discuss what live captioning is, the various types of live captioning solutions, and how they compare.

WHAT IS LIVE CAPTIONING?

Live captioning refers to providing captions, or time-synchronized text, in real-time. Live captions can be provided for a number of different mediums including virtual events, meetings, online courses, or performances.

There are several different techniques when it comes to the live captioning process, including live automatic captioning and live human captioning. Let's explore each of these techniques and processes and how they compare.

Live Captioning

Live captioning refers to providing captions, or time-synchronized text, in real-time.

LIVE AUTOMATIC CAPTIONING

Live automatic captions use artificial intelligence (AI) technology that makes it possible for machines to learn from past experience, adjust to new inputs, and perform human-like tasks. AI-based automatic speech recognition (ASR) software transcribes audio speech into a text-based format.

Live automatic captions do not involve a human transcriber, and the outcome of this type of solution can vary greatly. For instance, these captions can lack punctuation, speaker identification, and may require a human to fix such mistakes. The accuracy of the output is largely dependent on the specific speech engine being used and the quality of said engine. In the next section of this white paper we will discuss the many different considerations that impact the quality of a live automatic captioning solution.

LIVE HUMAN CAPTIONING

There are two types of live human captioning we cover in this white paper – CART and Voice writing.

CART

Live captioning, sometimes called real-time captioning, speech-to-text services (STTS) or Communication Access Real-time Translation (CART), is done by stenographers. This captioning technique relies on a skilled transcriber operating a stenotype keyboard in just about real time.

A stenotype differs from a traditional computer keyboard, in that it has fewer keys. It also relies on phonetic forms of words, allowing a skilled user to achieve over 200 words per minute.

Special software is then able to convert those phonemes into the correct words. This technique also allows for quick corrections which is critical when captioning live content.

CART has both pros and cons. This solution produces highly accurate captions.

However, it is more costly than alternative solutions due to the special equipment along with the expertise and training needed for transcribers.

Voice Writing

Another highly accurate solution for live captioning is voice writing. Voice writing is another human-enabled live captioning solution. People tend to be less familiar with voice writing, however it is used by the National Captioning Institute. The process of voice writing is a simple and effective, and consists of several components:

- ➔ The **original speaker**,
- ➔ A **highly trained voice writer**, and
- ➔ A specially tuned **ASR software**.

In the process of voice writing, a highly skilled speaker will repeat the main speaker's words clearly into voice recognition software. This software is specially trained to accurately perceive the voice writer's voice, often referred to as speaker-specific.

Voice writers are also highly trained individuals. They are trained in enunciation so that they can clearly articulate the spoken word. Additionally, voice writers must be well versed and knowledgeable about the topic they are covering.

This allows these individuals to understand and repeat precise language in an effective and accurate manner. By combining a trained voice writer and a high quality, speaker-specific software, you can achieve accuracy rates in the 90s.

WHAT MAKES A HIGH QUALITY LIVE AUTOMATIC CAPTIONING SOLUTION?

Vendor characteristics

Automatically generated captions average at 60%-70% accuracy. Some of the best automatic speech recognition systems can achieve accuracy rates in the '80s and low '90s if all conditions align perfectly. Additionally, there are a number of things that can drastically improve the quality of live automatically generated captions that must be considered. These include:

- The **ASR Engine**
- **Accuracy Enhancements**
- Allowing the Customer to Provide **Wordlists**
- Following **Audio Best Practices**

10 Questions to Ask a Live Automatic Captioning Vendor

★ What is your **average accuracy rate**?
★ What **ASR engine do you use**?
★ What is the **process like to get started with live automatic captioning**?
★ Do you offer an **upgrade path for full captioning** or additional services?
★ Do you allow customers to **provide a wordlist** or glossary of terms?
★ Which **platforms do you integrate** with?
★ What **output formats** are provided after the event?
★ What does the **workflow** look like?
★ What is the **cost for live auto captioning**?
★ Do you offer **customer support**?

READ MORE

The ASR Engine

The accuracy of live automatic captions depends the utmost on the ASR engine being utilized. At 3Play Media we take great care to provide great service by using the best ASR engine compared to other top engines – Speechmatics.

We are continuously testing and evaluating the Speechmatics engine as it compares to other popular engines on the market. Each time we are thrilled to find that our ASR engine outperforms other engines on the market, even in cases of low quality audio.

In our latest test, conducted in 2020, we evaluated the results based on the components of Word Error Rate (WER), including the percent error (ERR), percent correct (CORR), percent substitution error (SUB), percent insertion error (INS), and percent deletion error (DEL).

While WER is certainly an important component of measuring accuracy, it is clearly limited in capturing the requirements of the captioning task. At 3Play Media, we use an additional measure called Formatted Error Rate (FER) to measure accuracy.

FER is the percentage of word errors when formatting elements such as punctuation, grammar, speaker identification, non-speech elements, capitalization, and other notations are taken into account.

Formatting errors are particularly common with ASR technology.

The test data determined that Speechmatics had the lowest overall error rate when looking at WER, at 13.0%, Microsoft at 13.4%, and Rev at 15.3% with the most insertions, according to the McNemar test.

For the most accurate ASR systems, the WER of 13% (with PCOR of > 90% in some cases) is very impressive. For context, 3Play has been measuring WER on this type of text for more than 6 years, and this number has decreased by almost 50% in that period (from ~25% in 2013). Clearly, ASR researchers have been bringing many important innovations into the field, and there continues to be exciting work on applying the latest machine learning techniques to this very difficult problem.

3Play Media's Live Automatic Captioning

3Play Media's live automatic captioning solution integrates with top live streaming video and meeting platforms such as YouTube, Zoom, Brightcove, and Facebook.

Quality Enhancements

One of the main components of AI technology is its ability for it to "learn" and be trained to produce certain outcomes. The more experience and data provided to train the model, the more precise and accurate the engine.

3Play Media's more than 10 years of captioning and transcription experience and vast library of text can drastically improve the outcome of ASR engines. We have transcribed millions of words, which has provided us with critical data used to improve topic-specific learning.

We've invested research and development into increasing the accuracy of our language modeling technology by reducing our WER (word error rate) as much as 10%. In addition, we have implemented learned mappings – or rules about how to modify the ASR transcript – from our human transcript editors, creating a significant distinction among current ASR offerings.

We also work very closely with Speechmatics to help them improve the ASR engine on their end. With our partnership, we are able to provide them with insight into our customers' needs and help them understand where to improve their engine performance. This collaboration allows them to train the ASR engine to perform better for the specific topics and subject areas that our customers use captioning for.

Customer-Provided Information

Another way to dramatically improve the outcome of ASR for live auto captioning is by utilizing wordlists.

Wordlists allow users of Live Auto Captioning to submit a glossary of terms to improve the visual accuracy of the automatic speech recognition (ASR) on those words.

Wordlists consist of a list of words or phrases that may occur in a live streaming event. 3Play Media customers can submit a wordlist to be associated with a 3Play live auto captioning event. Wordlists are then fed to the ASR engine, and in turn the ASR engine is instructed to "listen" to those words or phrases occurring in the live stream. By listening, the ASR engine can then deliver the correct spelling and punctuation of those words or phrases in the captioning output based on the given wordlist.

A wordlist is a custom feature unique to 3Play Media that was created to increase the accuracy of words in a live stream.

The purpose of wordlists is to increase accuracy of the live captioning output by recognizing specific words that may be difficult for ASR to pick up, such as medical or scientific language, as well as proper nouns or acronyms.

A well-formed wordlist will significantly improve the visible accuracy of important terms or words in a live stream by reducing substitution errors (such as ASR choosing the wrong word) for those words in a list. Wordlists minimize the risk of misspelling or misrepresenting proper nouns and complex words or phrases that occur in the live stream. Wordlists also provide more flexibility for handling complex content in live streams, ultimately providing a better viewing experience for the live streaming audience.

Audio Best Practices

While the ASR engine used to produce live captions is going to be critical, the accuracy of live auto captions also depends largely on the environment and specific audio quality.

There are a number of best practices to follow to further improve the output of your captions. These include having:

- ➔ **A Strong Network Connection**: Avoid interruptions or spotty audio.
- ➔ **High Quality Audio**: Choose your microphone carefully.
- ➔ **Little to No Background Noise**: Find a quiet place, or let those around you know you'll be recording.
- ➔ **A Single Speaker**: Have only one speaker talk at once.
- ➔ **Clear Speech and Pronunciation**: Enunciate and talk at an understandable pace.
- ➔ **Wordlists**: 3Play Media allows users to submit a glossary of terms.
- ➔ **Upgrade Paths for Post Processing**: 3Play Media allows an upgrade to full transcription to ensure a fully compliant, 99% accurate recording.

For each of the APIs tested in our research, we examined some of the files with the largest difference in word error rate between that API and Speechmatics.

Generally, the files for which Speechmatics performed much better than any other API had either lower-quality audio, quiet audio, speakers with accents, or field-specific terminology (ex: recordings of lectures).

Learn More About the 7 Best Practices for Live Auto Captioning Quality

Learn how each of these 7 best practices can help you achieve the most accurate live automatic captions for your content.

WHAT ASR SOFTWARE DIFFERENTIATORS ARE CRITICAL TO LIVE?

A key metric for captioning providers is speed. The ability to deliver fast and accurate captions is the overall goal. ASR can be used anywhere and anytime and at scale. ASR can run and process speech-to-text immediately further reducing the effort of human transcribers. The time from speaking to the caption output is crucial and this goes hand-in-hand with deployment flexibility. It is often faster to convert speech-to-text using a real-time appliance on-premises to avoid voice data going to the cloud and back for processing.

Until now, captioning of live content required skilled human subtitlers. With the continual improvement of ASR technology, it is now possible to use it for live captioning – especially for online video content. Accuracy is essential to delivering same-language captions that meet FCC standards.

With features such as Custom Dictionary (Wordlists) – where users can input up to 1,000 words per transcript – content creators and broadcasters can ensure some of the most difficult vocabulary, names and acronyms are never missed. With these kinds of features, along with the general accuracy improvement for ASR technology, it is now possible to utilize it for live captioning, saving time and money.

It is also crucial to utilize an ASR engine that can deal – in the most part – with background noise for live captioning. Background noise is the biggest challenge when it comes to accurate speech-to-text in general, so it is crucial to find an engine that has been trained on noisy data to replicate real-world scenarios to prevent errors in the live captions.

Speechmatics has the best punctuation on the market, offering advanced symbols include exclamations and commas.

Punctuation can make a huge difference to any sentence and using punctuation is key to the readability of captions. When it comes to live captioning, punctuation becomes critical as there is no time to go back and alter the caption, context or error. Speechmatics has the best punctuation on the market, offering advanced symbols include exclamations and commas. This feature was created with live captioning in mind, to improve the readability, understanding and context of live captions where they cannot be altered after the event.

Ultimately, improvements that can be made with added features to uplift the accuracy of live captioning outside of the language pack is going to enhance the experience for the viewers of the content.

STATISTICS AND DATA FROM RESEARCH FINDINGS HIGHLIGHTING DIFFERENT ASR ENGINES

We tested 490 files from both high and low volume customers across diverse subject matters. In total, we tested approximately 65 hours of content, for a total of around 670,000 words.

Our research tested the top APIs:

- ➔ **SMX+** - Speechmatics API V2 with no post-processing
- ➔ **IBM** - IBM
- ➔ **MIC** - Microsoft
- ➔ **GOO** - Google
- ➔ **REV** - Rev.ai
- ➔ **VG** - VoiceGain

All testing used real content, reflective of the most common type and volume that we receive at 3Play Media.

Note: All numbers reflect testing done on recorded content, however, the general conclusions should hold true. We've tested most of these engines with live previously and gotten similar conclusions.

We evaluated the results based on a number of different comparisons. First, we evaluated the different ASR engines based solely on the percent error, percent correct, percent substitution error, percent insertion error, and percent deletion error, the components of Word Error Rate.

Definitions

- ➔ **ERR**: Percentage of incorrect words
- ➔ **CORR**: Percentage of correct words
- ➔ **SUB**: Percentage of substitutions
- ➔ **INS**: Percentage of insertions
- ➔ **DEL**: Percentage of deletions

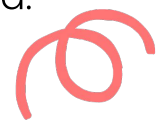
The following table shows the results across the tested APIs.

WER Scoring

WER measurements ignore case and punctuation.

	ERR	CORR	SUB	INS	DEL
SMX+	13.0	90.3	5.93	3.22	3.82
IBM	26.3	80.1	12.8	6.37	7.11
MIC	13.4	90.1	5.83	3.48	4.08
GOO	20.9	85.1	7.13	6.01	7.75
REV	15.3	90.2	5.65	5.50	4.15
VG	15.8	86.6	8.28	2.36	5.12

ASR researchers have been bringing many important innovations into the field, and WER has improved over the years of testing we've conducted. However, particularly for captioning, WER is not the whole picture.



FER Scoring

FER measurements take into account case and punctuation.

	ERR	CORR	SUB	INS	DEL
SMX+	24.66	78.49	17.77	3.15	3.74
MIC	25.67	77.74	18.25	3.41	4.01
IBM*	41.78	64.38	28.71	6.16	6.91
GOO	40.31	62.58	20.88	2.88	16.55
REV	26.39	79.03	16.9	5.42	4.07
VG	28.17	74.11	20.85	2.79	5.05

** IBM doesn't offer automatic punctuation, which is (part of) why they did so poorly here.*

In addition to our measurements of both WER and FER, we did some further in depth analysis with the goal of determining both the general state of speech-to-text technology, and its application to our business requirements. Overall, our findings indicate that the order of quality for the tested engines ranks as follows:

SMX+ Microsoft Rev VoiceGain Google IBM

This confirms that the Speechmatics ASR technology we use continues to be state-of-the-art.

HOW LAC DIFFERS FROM BATCH PROCESSING

Live automatic captions (transcription) are delivered in parallel to input audio being streamed through the ASR engine.

Pre-recorded captions are created after the event. The audio or video file is automatically transcribed to be turned into captions. The transcript can then be edited and tidied to be perfect and turned into captions using approved house styles. This process can be complete efficiently but is not classed as near real time.

Batch processing also allows for the ASR engine time to process the audio or video file and gain all context and understanding of the words before deciding which it believes are the right words to transcribe in the final output. With live captioning, with speed of the essence, the ASR engine only has a split second to decide which is the most probable word that has been spoken which is why ASR has mostly been used for pre-recorded captions until now.

The added features such as Advanced Punctuation and Wordlists are crucial to the success of live automatic captions using ASR because of this challenge.

HOW DO THESE SOLUTIONS COMPARE FOR ACCURACY?

Is Live Automatic Captioning ADA Compliant?

The accuracy, and therefore compliance of live automatic captioning (LAC) solutions can vary depending on the environment, how well the best practices are followed, and the quality of the ASR engine used. Additionally, the accuracy rate depends on how you measure or define accuracy. When measuring the accuracy of LAC, it is imperative to factor in error rate and deletion rate. When following best practices for live auto captioning quality, the accuracy rate should be a minimum of 90%. Additionally, the deletion rate – or the percentage of words spoken that are omitted by the ASR – should be as low as possible. This is critical to the use of ASR for creating captions, as deleting words can change the meaning and convey incorrect information.

Are there Standards for Live Captioning?

Currently, there are no governing bodies for live captioning, but certain states do have standards in place for the quality of live captioning in instances like court reports or sporting events.

Speechmatics continues to be a leading speech recognition technology. When tested, it performed much better than other speech engines.

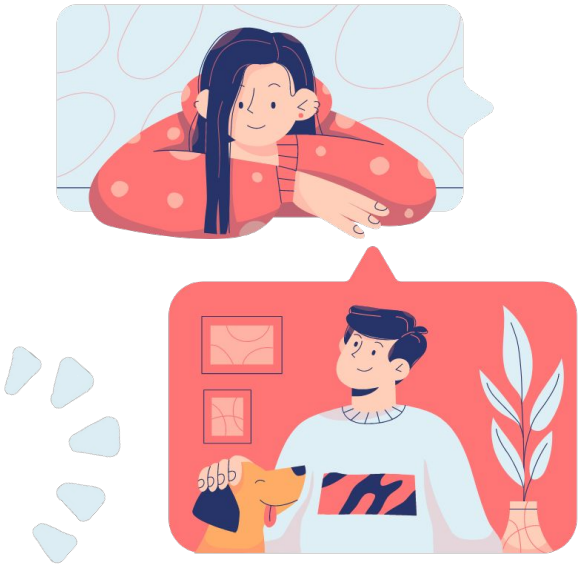
Of the tested ASR engines, Speechmatics had the lowest deletion rate at 3.74% with the next lowest being Microsoft with a deletion rate at 4.01%. A low deletion rate is critical to the use of ASR for creating captions and transcripts, as deleting words can change the meaning and convey incorrect information.

For these reasons, live auto captions are sometimes preferred over human-powered live captions. It's easy to automate scheduling with live automatic captioning and it's competitively priced.

We always recommend that the best practices are followed to achieve 90% + accuracy and that our customers upgrade to a fully accessible solution for post production.

Is Live Auto Captioning Right for You?

Read Top 5 Use Cases for Live Auto Captioning find out when to use live automatic captioning.



How Does Accuracy for CART or Voice Writing Compare to LAC?

CART providers can be certified to type up to 260 words per minute with 98% accuracy and above. Errors that do occur with CART are due to natural human error. For example, the caption writer may mishear a word or hear an unfamiliar word, or even miss a word. Sometimes, there may be an error in the software dictionary. Like with any type of live captioning, CART is susceptible to technical errors that are beyond the control of the caption writer.

Live automatic captioning can range in accuracy, but averages around 80% accurate. When following best practices with 3Play Media's live auto captioning, you can expect an accuracy rate of 90% or higher.

Captioning Latency

Since live captions are generated in real-time, based on the spoken word within the video, there will be a slight latency. This latency ranges depending on the quality of streaming equipment and overall connection, but **averages about 3-5 seconds**.

Why Does Voice Writing Perform Better Than LAC?

Voice writing and live auto captioning both rely on ASR. One of the differentiating factors of using voice writing is the amount of training that goes into it.

Not only do voice writers themselves receive skill-based training, but the ASR models are trained as well. With voice writing, the ASR engines are trained to adapt to and better understand certain speakers. On the contrary, with live auto captioning, the ASR engine is speaker-agnostic.

As world experts in the traditional approach to building language packs, Speechmatics looked for a more efficient solution to overcome training individual models one-by-one. By adding value through improved speed and accuracy, and improving flexibility by using data to continually update all languages automatically. Speechmatics does not rely on linguistic expertise for every language. Instead, when they come across a linguistic problem they devise novel machine learning approaches to make the solution as generic as possible, so when we come across a similar problem in another language they don't need to solve it all over again.

Starting a new language pack can feel like a mountain to climb. So Speechmatics has developed techniques using machine learning and neural networks that mean they can always start from a well-established basecamp. This means learning across builds in different languages. Speechmatics' any-context speech recognition algorithms learn from previous builds, while the Speechmatics experts extract, streamline and improve on any new features to make future builds easier, faster and more accurate.

CONCLUSION

As the world continues to create and engage with live video content, it is important to make sure that it is live captioned. *How* you decide to provide live captioning may vary depending on the type of content you are producing, the resources you have available, the goals you are trying to achieve, and the audience you are hoping to reach.

Now that you have an understanding of the different live captioning processes, including live human captioning and live automatic captioning, you can make an informed decision. Remember – whichever method you choose – the ASR engine and vendor characteristics will play an important role in the accuracy of your solution.

Additional Resources

- YouTube Live Auto Captioning How-to Guides
- 7 Best Practices for Live Auto Captioning Quality
- How to Add Captions to Zoom Video Conference Recordings
- How to Use Wordlists for Live Auto Captioning Quality
- How to Set Up Live Auto Captions for Zoom Meetings
- Introducing 3Play Media's Live Automatic Captioning (Beta) Solution
- Meet 3Play Media's New and Improved Live Auto Captioning Service!
- Tips for Scaling Live Auto Captioning

ABOUT 3PLAY MEDIA

3Play Media is a full-service accessibility partner. We provide closed captioning, transcription, and audio description services to make video accessibility easy.

Born out of MIT 10 years ago, we have a history of being innovative in the video accessibility space. Our measured accuracy rate is 99.6%, the highest in the industry.

We continuously provide dedicated onboarding and support for our customers, and create tons of educational resources and host webinars and conferences throughout the lifespan of a customer.

Follow us on social media.

Follow us for more resources on web and video accessibility.

@3PlayMedia

Drop us a line.

Website: www.3playmedia.com
Email: info@3playmedia.com
Phone: (617) 764-5189

Born and raised in Boston.

77 N Washington Street
Boston, MA 02114

ABOUT SPEECHMATICS

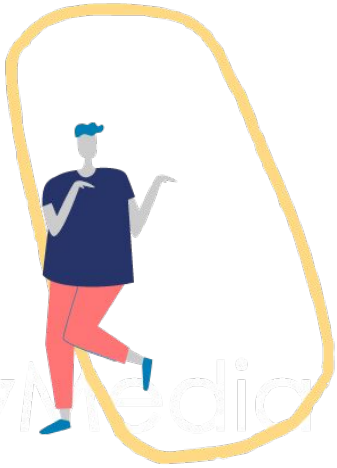
Speechmatics® powers applications that require mission-critical, accurate speech recognition through its any-context speech recognition engine.

Speechmatics' speech recognition technology is used by enterprises in scenarios such as contact centers, CRM, consumer electronics, security, media & entertainment and software. Speechmatics processes millions of hours of transcription worldwide every month in 30+ languages.

Having pioneered machine learning voice engineering, Speechmatics is enabling companies to build applications that detect and transcribe voice in any context and in real-time. Its neural networks consider acoustics, languages, dialects, multiple speakers, punctuation, capitalization, context and implicit meanings.

Follow us on social media.

@Speechmatics
www.linkedin.com/company/speechmatics



Drop us a line.

Website: www.speechmatics.com
Email: hello@speechmatics.com
Phone: +44 (0)1223 907 818 | +1 866 791 8546

Born and raised in the UK.

Unit 296 Cambridge Science Park
Milton Road
Cambridge CB4 0WD
United Kingdom