3PLAYMEDIA

STATE OF ASR 2024

3PLAYMEDIA

Table Of Contents

About The Report

Automatic Speech Recognition (ASR) technology has significantly improved over the years. Despite the developments, the question remains: **Can ASR deliver accurate and accessible captions that revolutionize media accessibility?**

Every year, 3Play Media conducts extensive research to learn how the top ASR engines perform for captioning and transcription. The findings are then published in the *State of ASR* report.

Why is this research important to us?

At 3Play Media we are dedicated to understanding the evolving ASR landscape and its impact on captioning and transcription. Powerful ASR engines provide the foundation of our human-in-the-loop captioning process, allowing our expert editors to focus on creating high-quality and accessible captions faster than ever before.

The report analyzes the advancements and limitations of top ASR engines for real-world captioning applications and offers valuable insights into the evolving ASR industry. These insights help us improve our services and ensure the highest-quality captions for our customers.

P.S. We love to hear your thoughts! Share your insights and connect with us on social media @3PlayMedia.

INTRODUCTION

Introduction

As we explore the current ASR landscape, it is evident that the focus has shifted from revolutionary changes and developments in ASR technology to fine-tuning existing technologies.

The effectiveness of ASR engines hinges on a nuanced interaction between three key factors: error rate measurement, transcription styles, and specific use cases.

The report is structured around these three pillars:

- **Error Rate Measurement:** We'll look more broadly than the basic word error rate (WER) to explore the concept of a Formatted Error Rate (FER). This metric takes into account factors like punctuation, capitalization, and speaker identification, providing a more nuanced picture of performance, particularly for tasks where presentation matters. We'll also discuss the more subjective NER Model, and its shortcomings as an error measurement for captioning.

- **Transcription Style:** We'll explore the two primary styles – clean read and verbatim – and their impact on the final output. Understanding the preference for clarity (clean read) versus capturing every detail (verbatim) is crucial in selecting the right ASR engine.

- **Use Case:** Different use cases have distinct requirements for ASR engines. Captions need to be easy to read and understand, therefore they need to consider formatting and clarity. In contrast, ASR for an automated assistant focuses on understanding the intent behind a user's voice command. Accuracy for every word is less critical than correctly identifying the action the user wants to perform. This section will explore how aligning your use case with the preferred style of an ASR engine results in optimal performance.

By examining these interconnected components, this report aims to equip you with the knowledge to make informed decisions when selecting and utilizing ASR engines for your specific needs. We'll not only shed light on individual aspects but also demonstrate how they interact to influence the overall success of your chosen application.

AN OVERVIEW

An Overview: Measuring Errors

When evaluating the performance of ASR engines, several factors need to be considered. These factors include objective accuracy metrics like WER (Word Error Rate) and FER (Formatted Error Rate), and specific error types such as % ERR (overall error rate), % CORR (correct words), % SUB (substitution errors), % INS (insertion errors), and % DEL (deletion errors).

Word Error Rate (WER)

WER is commonly used to determine quality in ASR and is the best metric for accurately understanding how much content the engine recognizes. Typically, when you see the label “99% accurate captions,” this refers to WER, and those captions have a WER of 1%. WER is a formatting-agnostic measurement, meaning it only measures words, and the scores do not count capitalization, punctuation, or number formatting errors.

A WER-formatted transcript, which only considers the number of correct words, might look like this:

yesterday biden approved nine hundred million dollars in electric vehicle charger funding

Formatted Error Rate (FER)

FER is used to better evaluate the overall experience, readability, accuracy in transcription and captioning, and the amount of additional work needed to make a transcript fully accessible. FER measures word errors, and elements such as punctuation, grammar, speaker identification, non-speech elements, capitalization, number formatting, and other notations.

Formatting errors are widespread in ASR transcription, and some engines prioritize FER more than others.

A FER-formatted transcript, which considers formatting elements, might look like this:

[MUSIC PLAYING] [Speaker 1] Yesterday, Biden approved $900 million in EV charger funding.

Abbreviations and Error Measurements

You may have noticed that in the WER example, the abbreviation "EV" was written as "electric vehicle," whereas in the FER example, it was written as "EV." Expanding abbreviations can be a part of the process to ensure accurate and meaningful comparison for WER measurement. This helps maintain consistency and clarity in the evaluation of transcription accuracy.

NER Model

It’s worth mentioning another evaluation model we have seen growing in popularity – the NER Model. We did not measure accuracy using the NER Model because of its subjectivity.

The NER Model originated in Europe and is often used in Canada. NER Model scoring, which is done manually by a human, emphasizes meaning and how accurately ideas are captured in captions, rather than exact words.

In the U.S., all errors—including spelling, punctuation, grammar, speaker identifications, word substitutions, omissions, and more—are considered to obtain a percentage that measures the average accuracy of the closed captions on a piece of media.

> **Legal Note:** Federal Communications Commission (FCC) closed captioning guidelines require captions to include all words spoken in the order spoken (i.e., no paraphrasing).

The subjectivity of scoring in the NER Model and its inherent risk makes it an unreliable measure of captioning accuracy. However, there are appropriate situations for the NER Model, such as using it to comparatively measure output and evaluate the skill and training of human live captioners.

Customers need to understand the difference between WER and NER, and vendors must be transparent about their models when sharing their accuracy rates.

An Overview: Use Case

When using ASR for captioning and transcription, there are unique challenges to consider compared to other tasks. Captioning deals with:

- Unpredictable audio
- Long-form content
- Human readability

These unique factors require robust ASR capabilities. Unlike interactions with automated assistants, captioning is a one-way process and must handle disfluencies, background noise, and changing audio conditions without the ability to ask clarifying questions. Understanding these challenges is crucial for selecting the right ASR tools for the job.

An Overview: Transcription Style

To better understand the behavior of the tested engines, we considered two common transcript styles: **clean read** and **verbatim**.

- **Clean Read:** Removes disfluencies (like "um" and "uh"), filler words, and false starts, resulting in a smoother read.
- **Verbatim:** Includes everything spoken, providing a complete record.

Understanding these style differences is crucial for interpreting error rates associated with each type.

Not only does the transcript style affect the engine's perceived performance, but the engine itself might also have a “preferred” transcript style based on its training data.

Engines Styles

Engines trained primarily on clean read data might perform better on clean read transcripts, as they're better at interpreting the smoother speech patterns. Conversely, engines trained on verbatim content might be more adept at handling disfluencies and filler words, leading to lower error rates for verbatim transcripts.

RESEARCH

Dataset & Engines

We compared the accuracy of the most popular automatic speech recognition engines (ASR) for captioning and transcribing pre-recorded videos, including:

- Speechmatics (SMX)
- AssemblyAI's Universal 1 model (AssemblyAI)
- OpenAI's Large V2 model (Whisper Large V2)
- OpenAI's Large V3 model (Whisper Large V3)
- Microsoft
- Rev.ai's v2 model (Rev)
- DeepGram's video model (DeepGram)
- Google's model for long-form content (Google Long)
- Google's enhanced video model (Google Video)
- IBM Watson (IBM)

We tested all engines across various industries to see how well they handle different content types.

To ensure a fair comparison, we used a rigorous testing process. First, we created highly accurate transcripts with over 99% accuracy, using a combination of ASR and human review. Then, we tested the ASR engines on a large dataset of real-world videos.

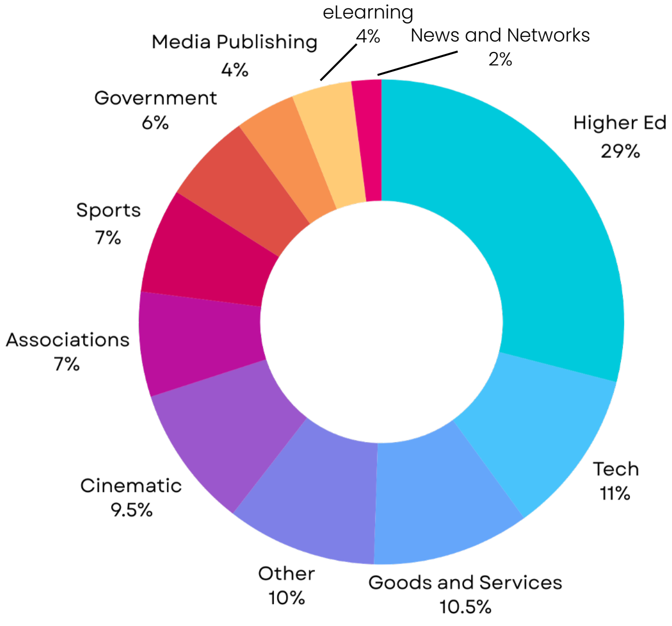
The dataset totaled 700 files, 158 hours, and 1.3 million words. This dataset is significantly larger than previous years, with a 47% increase in content duration and a 43% increase in word count compared to 2023, allowing us to test more broadly than we have in the past.

700
VIDEO FILES
158 HOURS
1.3 MILLION
WORDS

43%
INCREASE IN
WORD COUNT
COMPARED TO
2023
DATA

The videos represent a variety of industries, topics, lengths, speaker numbers, and audio qualities, reflecting the types of content we typically encounter in captioning and transcription workflows.

Content amount by primary market



A DEEPER DIVE

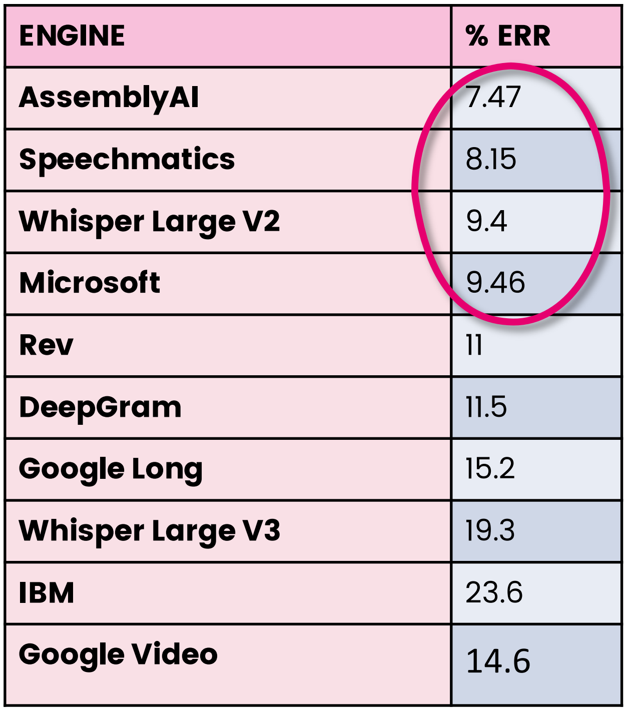
A Deeper Dive: Error Rate Measurement

This section will discuss Word Error Rate and Formatting Error Rate in depth, including the components that make up errors – insertions, deletions, and substitutions.

Word Error Rate

Looking at the overall accuracy rates, AssemblyAI achieved the lowest Word Error Rate (WER) of 7.47%, followed by Speechmatics at 8.15%. Whisper Large V2 had a word error rate of 9.4%, with Microsoft slightly behind at 9.46%.

Keep in mind that you will see a range in performance from these engines. Even with the most accurate engine, which had an average WER of ~7.5%, a given file has a 1 in 5 chance of scoring 10% or worse in WER.



ENGINE	% ERR
AssemblyAI	7.47
Speechmatics	8.15
Whisper Large V2	9.4
Microsoft	9.46
Rev	11
DeepGram	11.5
Google Long	15.2
Whisper Large V3	19.3
IBM	23.6
Google Video	14.6

While the overall error rate is an important indicator of accuracy, it must not be looked at alone, particularly for the use case of captioning and transcription.

Formatted Error Rate

Formatted Error Rate is especially important for 3Play‘s use case of captioning and transcription, as accurate punctuation and non-speech elements make captions more accessible and require less time to edit. FER is also critical to readability and meaning, and an accuracy rate under 90% is extremely noticeable to consumers.

When it comes to FER, AssemblyAI performed the best overall, with a 17.5% error rate. Whisper Large V2 followed closely behind with 17.6%, and then Speechmatics with 19.2%. We’ll take a look at this broken down by error type in the next section.

The table below shows the tested vendors and their overall FER results.

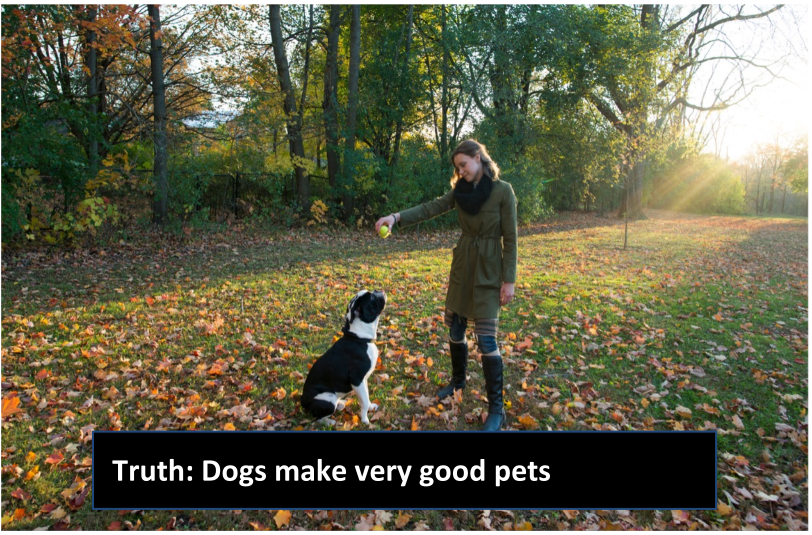
ENGINE	**% ERR**
AssemblyAI	17.5
Whisper Large V2	17.6
Speechmatics	19.2
Microsoft	20.1
DeepGram	20.1
Rev AI	21.6
Whisper Large V3	27.6
Google Long	29.8
Google Video	30
IBM	43.4

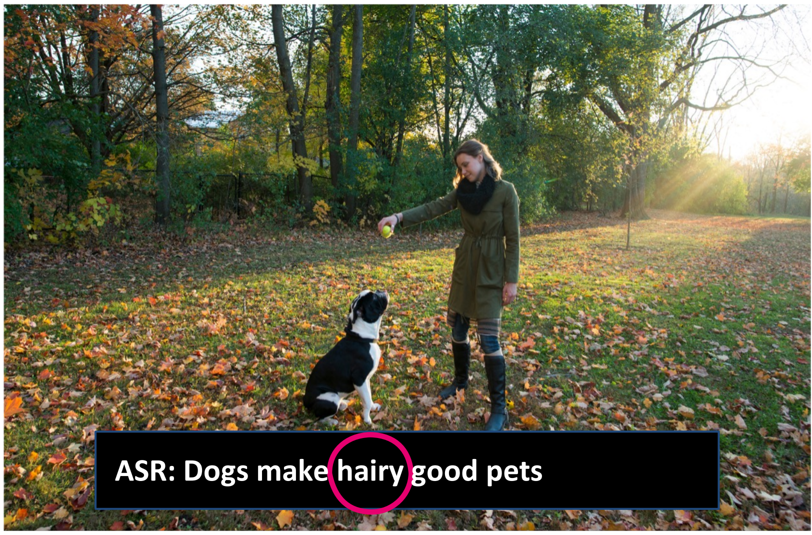
While the most accurate engine averaged a 17% FER, a given file has a 1 in 5 chance of scoring worse than 25% FER.

Error Types

Transcript errors can be categorized into three main types: substitutions, insertions, and deletions. Together these make up the total error rate and apply to both WER and FER measurement.

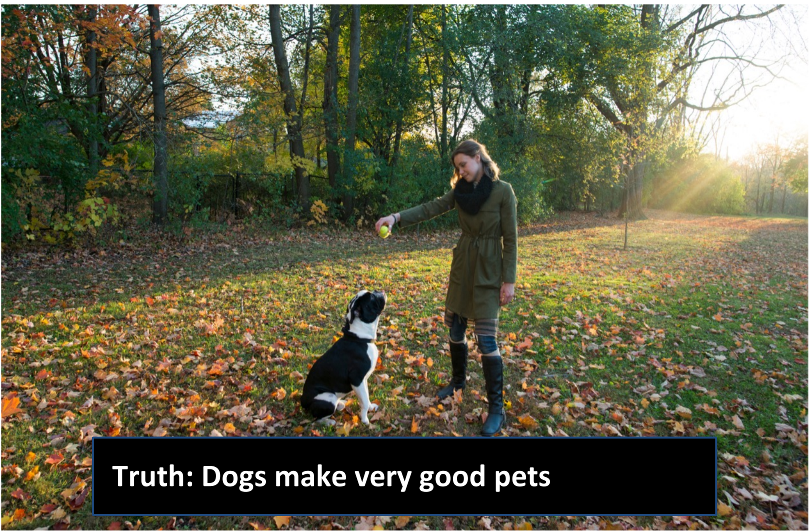
- **Substitutions** occur when the ASR **mishears a word**, for example, recognizing "encyclopedia" instead of "3Play Media."

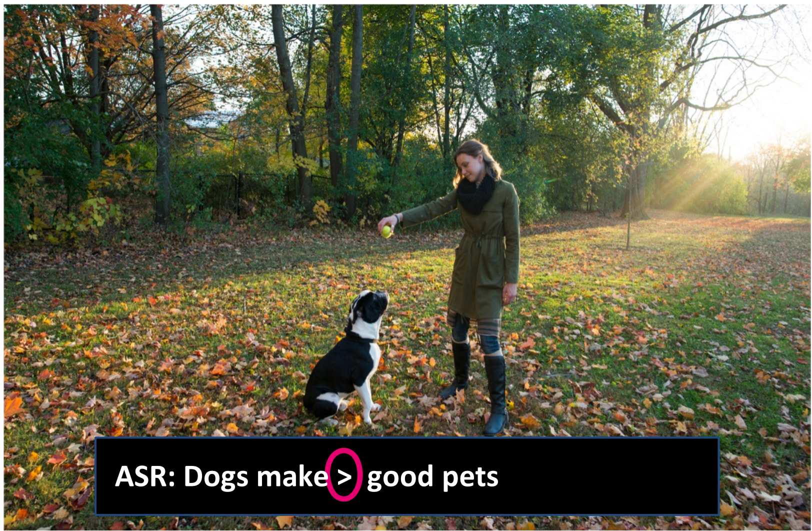




- **Insertions** happen when the ASR **adds extra words**, like misinterpreting background noise as speech and inserting nonsensical words.

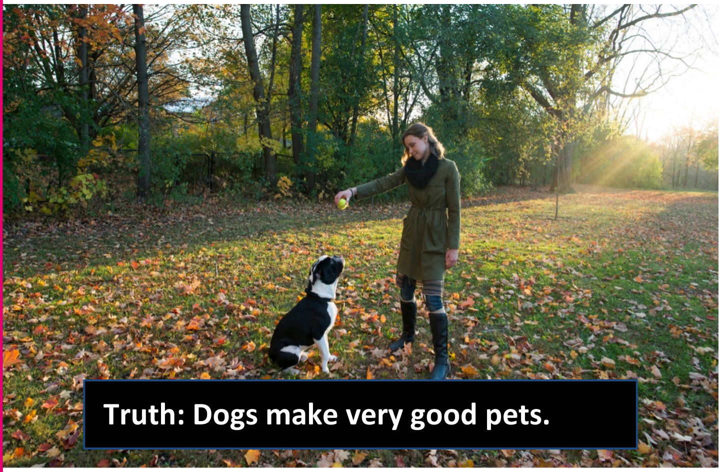
- **Deletions** occur when the ASR **misses words entirely**, leaving no transcript where there should be recognized speech.

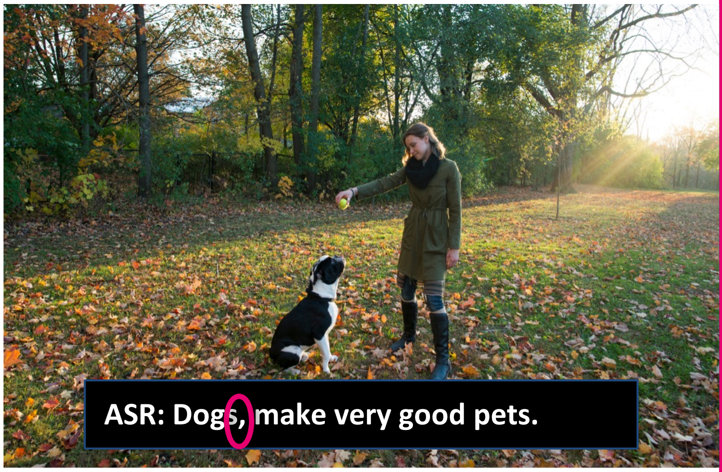




Some errors are easier to correct than others. For example, it's important to consider substitutions, insertions, and deletions when creating captions, as they can affect the timing of the captions.

The below example shows an error that would only be recognized using FER, not WER. However, adding a comma after the noun at the beginning of a sentence can indicate that you are addressing the sentence directly to someone. Thus, the readability with even a seemingly minor error becomes confusing.





How Engines Perform In Terms Of Error Type

Different engines tend to have different strengths and weaknesses when it comes to specific error types. Some error types are also better (meaning easier to correct) than others for certain use cases.

Adding missing words back into a sentence can be more difficult than removing inserted words. Even if the inserted words are incorrect, they can provide valuable information such as timing, which is important for captioning.

Formatting matters for ASR output. Different ASR engines handle formatting in different ways, which can affect how easy it is to understand the transcript.

- **Whisper Large V2:** This engine is trained on a lot of internet caption data, so it tends to format transcripts like captions (short and easy to read).
- **Speechmatics:** This engine is trained on a wider variety of data, including dictation and notes, so its formatting is less like captions and might be more detailed.

The best ASR engine for you depends on what you'll be using it for. If you need captions, Whisper might be a good choice. If you need more detailed transcripts, Speechmatics might be better.

In the future, ASR engines might let you customize formatting. This would be like choosing between "clean read" (fixing errors) and "verbatim" (keeping everything exactly as spoken) transcriptions, but for formatting. The best options will depend on how much control you want and how well the engine's training data matches your needs.

Please see Table 1 and Table 2 in the Appendix for the full breakdown of engine performance by error type.

A Deeper Dive: Transcription Style

When delving into transcription style, it's important to emphasize the interaction between transcription style, engine style, training, and the type of content being transcribed. These factors are all interconnected and influence each other, ultimately affecting the overall accuracy of the captioning output.

The choice between clean read and verbatim styles can influence error rates in speech recognition engines.

Clean read transcripts, while offering a smoother read, are more prone to insertion errors. This happens when removing disfluencies like "um" is mistakenly interpreted as inserting a new word. Conversely, verbatim transcripts, with their inclusion of all spoken content, are more susceptible to deletion errors. These occur when filler words like "uh" are missed during transcription.

Our analysis revealed that clean read content generally has lower error rates compared to verbatim content. This could be attributed to two factors:

- **Market Preferences:** Clean read transcription is more popular in markets where speech tends to be clearer, such as Goods & Services and Higher Education. Engines trained on these clean read datasets might perform better on similar content.

- **Dataset Bias:** Our dataset potentially contained a higher proportion of clean read files compared to verbatim ones. This skews the weighted average error rate, which considers both the number of files and word count of each style.

It's important to consider these factors when interpreting error rates and the impact of transcript styles on engine performance.

Engine Preferences

As noted earlier, transcripts generated by different engines can vary significantly depending on the transcription style the particular engine was trained on.

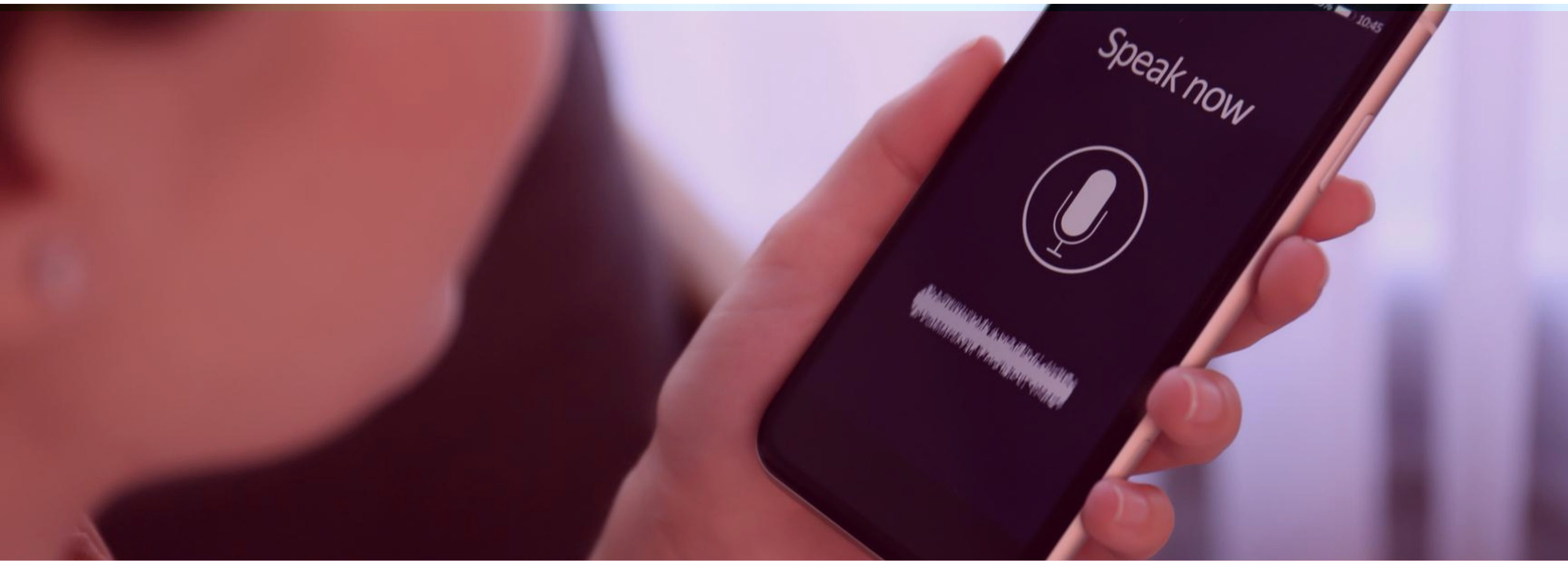
Here's a breakdown of some popular engines and their strengths as related to transcription style:

- **Clean Read:** AssemblyAI and Whisper both excel at clean read content.
- **Verbatim:** Speechmatics and Microsoft outperform others on verbatim content.

While personal preferences may vary given your specific need, verbatim style is generally considered a better starting point for captioning and transcription because it tends to retain more words. This means that less editing needs to be done, and the time-codes will be more accurate.

However, even for captioning and transcription, preferences can vary depending on the market standards and type of content being transcribed.

Speechmatics has recently made an update to their engine that allows them to more accurately capture common disfluencies. By tagging these disfluencies, it's now possible to remove them from clean read files, making the transcription process even more accurate.



Pairs of Engines

While the notion of "more data, better results" is a common belief in machine learning, it's not always the outcome.

We tested multiple models from three leading ASR providers: **Assembly**, **Whisper**, and **Google**. In all three cases, the older, now deprecated models, achieved higher accuracy than their newly released counterparts.

These newer models were built upon advancements made in 2023, which included a strong focus on utilizing more training data.

WER Comparison of Different Models for Each Engine

Assembly AI Engines	**% ERR**	**% CORR**
Older Model	7.13	94.4
Newest Model	7.47	95.1

Whisper Engines	**% ERR**	**% CORR**
Older Model	9.4	94.7
Newest Model	19.3	91.5

Google Engines	**% ERR**	**% CORR**
Older Video Model	14.6	89.3
Newest Video Model	15.2	88.3
Standard Model	28.7	74.8

In 2023, Whisper achieved success with an unprecedented amount of training data. Both Assembly and Whisper have attempted to take that lesson further by training on even more data this year. However, this approach did not prove to be successful, indicating that while the quantity of data is important, the quality and composition of that data also play a crucial role in ASR engine performance.

A model trained on a massive dataset may reach a point of diminishing returns, where the benefit of adding more data plateaus and the additional effort required to collect and process it outweigh the minimal performance gains.

FER Comparison of Different Models for Each Engine

Assembly AI Engines	**% ERR**	**% CORR**
Older Model	17	84.9
Newest Model	17.5	85.2

Whisper Engines	**% ERR**	**% CORR**
Older Model	17.6	86.4
Newest Model	27.6	83.2

Google Engines	**% ERR**	**% CORR**
Older Video Model	30	73.9
Newest Video Model	29.8	73.6
Standard Model	41.2	62.1

Performance By Market

This section will review how the use case, subject matter, and source content significantly affect ASR engine performance.

Markets With The Lowest Error Rates

- **eLearning**
  When we analyze the results based on the primary market, eLearning emerges as the best-performing market. All engines performed well on this content, with the lowest WER and FER at 3.97% and 11.8% respectively.

- **Goods/Services**
  Product demonstrations, training videos, and instructional content may be scripted and produced with high-quality audio, enhancing the ASR output.

- **News/Networks**
  News and Network content has high-quality audio input, controlled recording environments, single-speaker scenarios, and clear dictation from speakers, all enhancing ASR accuracy. Because news content wants to avoid profanities, they may err on the side of caution and include more deletions.

Market	**Average WER of Top 4 Engines**
eLearning	3.97
Goods/Services	5.05
News/Networks	5.25

Higher Education content also performed relatively well, with a WER of 6.38% and FER of 16%, but not as well as eLearning. One notable difference between the two industries is that eLearning usually features **a single speaker recorded in a professional environment, resulting in optimal recording conditions, including no background noise, scripted content, and high audio quality.**

In contrast, classroom content may have a lot of background noise, multiple speakers who do not use microphones, and complex topics that change throughout the class.

Markets With The Highest Error Rates

- **Sports**

  Sports files are typically characterized by a lot of background noise, inaudible speech, or interposing voices, and require extensive research to get the correct names of players.

- **Cinematic**

  Similarly, Cinematic files also require a lot of research as they contain a lot of special formatting, and customers often include extremely specific instructions that may deviate from a typical engine output or transcription standards.

- **Tech**

  Tech's performance can likely be attributed to the fact that Tech customers are providing highly customized services to their customers. This results in extreme variety in content type and highly custom needs, similar to Cinematic customers.

Market	Average WER of Top 4 Engines
Sports	10.2
Cinematic	10.2
Tech	9.96

Please see Table 3 and Table 4 in the Appendix for the full breakdown of how the Top 4 engines performed by market.

Market Preferences and Transcript Style

Our analysis revealed an interesting and important interplay between market demands and preferred transcript style (verbatim or clean read). This impacts ASR engine selection and performance.

- **Content Specificity Matters:** Whisper, despite strong performance in some areas, struggles with copyrighted content due to its open-source training data. This highlights the importance of choosing engines suited to specific markets like Cinematic, where content is subject to copyright law.

- **Verbatim Reigns Supreme in Complex Markets:** Interestingly, verbatim transcripts are favored by challenging markets like Cinematic and Tech. These domains prioritize capturing every detail, despite disfluencies, due to their intricate content and formatting requirements. We see Speechmatics perform really well on Verbatim content.

- **Disfluency Handling Tailored to User Needs:** ASR engines cater to user preferences when it comes to disfluencies (e.g., "um," "like"). Speechmatics and IBM tag them for easy removal, while others like Rev, Microsoft, AssemblyAI, and DeepGram offer configuration options for inclusion or omission. This flexibility allows users to customize transcript style based on their needs.

Please see Tables 5, 6, 7, and 8 in the Appendix for the full breakdown of WER and FER performance by transcript style.

A Deeper Dive: Use Case

When it comes to using ASR for captioning and transcription, there are vastly different demands placed on ASR engines compared to other tasks.

Unpredictable Audio and Dynamic Environments

Captioning faces unpredictable audio and dynamic environments. Imagine a college lecture hall filled with multiple speakers, background noise, with speakers at varying distances from the microphone. Or consider a sports arena with cheering crowds and multiple announcers vying for attention. This vast spectrum of environments, coupled with long-form audio featuring shifting topics and concepts, creates a complex and demanding task for ASR engines.

Captioning also often deals with fluctuating audio quality within a single program. A documentary, for example, might seamlessly switch between interviews, narration, and background music. These constant audio shifts further complicate the challenge of accurate speech recognition for captioning and transcription.

Human Readability vs. Intent Recognition

Other use cases of ASR only require the engine to grasp the intent behind a request. However, captioning success is determined by the human experience. A transcript riddled with grammatical errors, missing punctuation, and no speaker identification would be confusing and difficult to follow. Thus, factors like proper sentence structure, punctuation, and speaker differentiation are crucial for captions. ASR engines often struggle with these nuances of human language that are essential for readability and comprehension.

Limited Interaction vs. Silent Processing

Automated assistants have the luxury of clarification. Captioning, on the other hand, is a silent battle. It grapples with disfluencies (umms and errs), background noise, and poor audio quality – all without the ability to ask clarifying questions. These factors significantly impact the accuracy of captions compared to more controlled interactions.

Understanding these unique challenges allows us to select the right ASR tools for the job.

Hallucinations

ASR (Automatic Speech Recognition) is an application of Artificial Intelligence (AI) that has been around for a while. The latest advancement in AI is generative AI, which focuses on creating new content rather than just observing existing content.

ASR technology has made significant strides however, we've seen that using generative AI for captioning and transcription can pose some issues.

Whisper, for example, has the potential for hallucinations. Hallucinations are instances where the engine generates text that wasn't present in the original audio. This highlights the generative nature of AI, as these engines are not simply transcribing audio; they are, to some extent, creating text based on their understanding of speech patterns.

Whisper has a well-documented tendency to 'hallucinate' or to generate text that has no basis in the audio. This generative aspect can be beneficial for tasks like summarization, but for captioning and transcription, it can pose a significant risk to accessibility and brand image.

Hallucinations

Hallucinations in generative AI refer to the generation of text that has no basis in the input. These hallucinations are often incorrect or misleading and can be caused by various factors, including insufficient training data, incorrect assumptions made by the model, or biases in the data used to train the model.

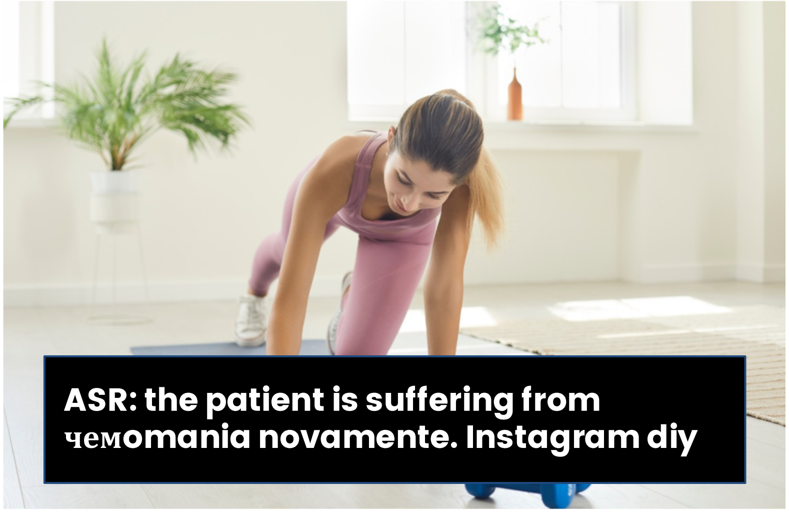
Accessibility

For most insertions, Whisper appears to either get-stuck in a loop predicting the same word or short phrase over and over again, or it inserts a whole piece of correct transcription from a different part of the document.

While there are many non-loop hallucinations from V3, most of them are more nonsensical than the examples seen from V2 last year. Someone viewing this type of error in a captioning setting will probably assume ASR has messed up several words but might not realize the whole section is completely made up.

In rare cases, Whisper generates novel utterances that are not found in the document. These are typically plausible and grammatically correct and could be mistaken for the real audio by someone viewing a video with captions.

Truth	Whisper
the	the
>	patient
>	is
>	suffering
>	from
>	чемоmania
>	novamente.
>	Instagram
>	diy

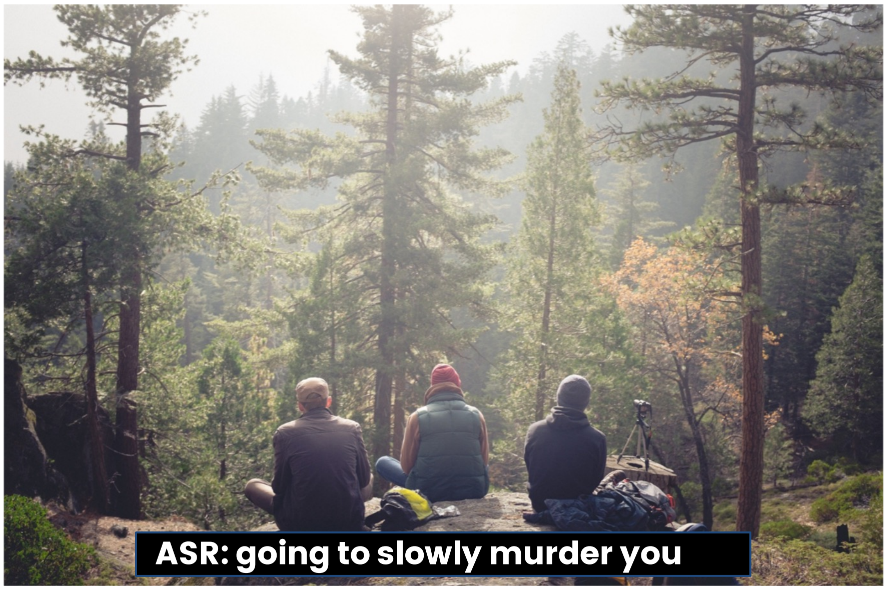


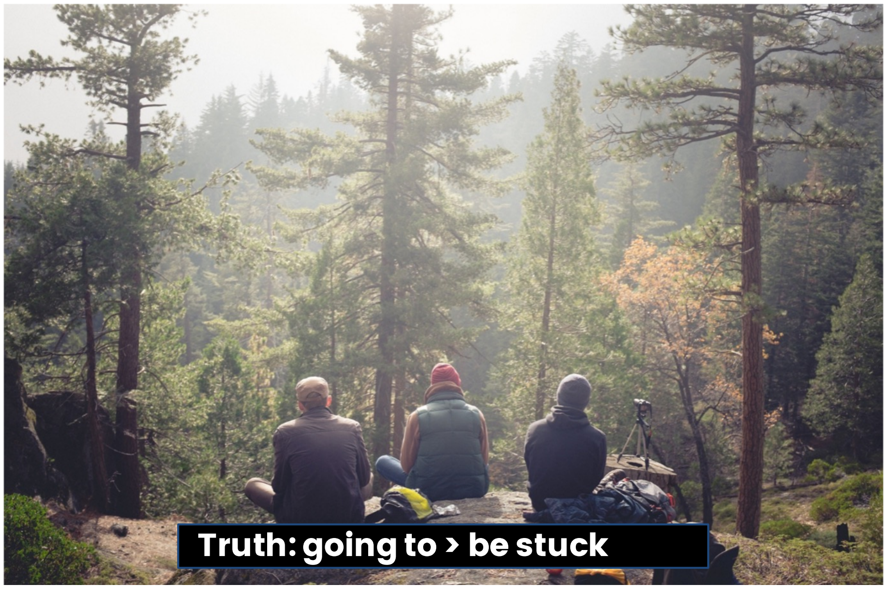
ASR: the patient is suffering from чемоmania novamente. Instagram diy

Brand

There were a few egregious examples of hallucinations discovered that pose a significant risk to the brand. We saw instances of Whisper adding words that were not in the target language, as well as adding inappropriate content, such as talking about murder.

Truth	Whisper
>	Welcome
>	to
>	our
>	house,
>	my
>	darling.
>	I'm
going	going
to	to
>	slowly
be	murder
stuck	you
here	and
for	that's
the	that.
rest	Go
of	get
the	the
night,	<





3PLAYMEDIA

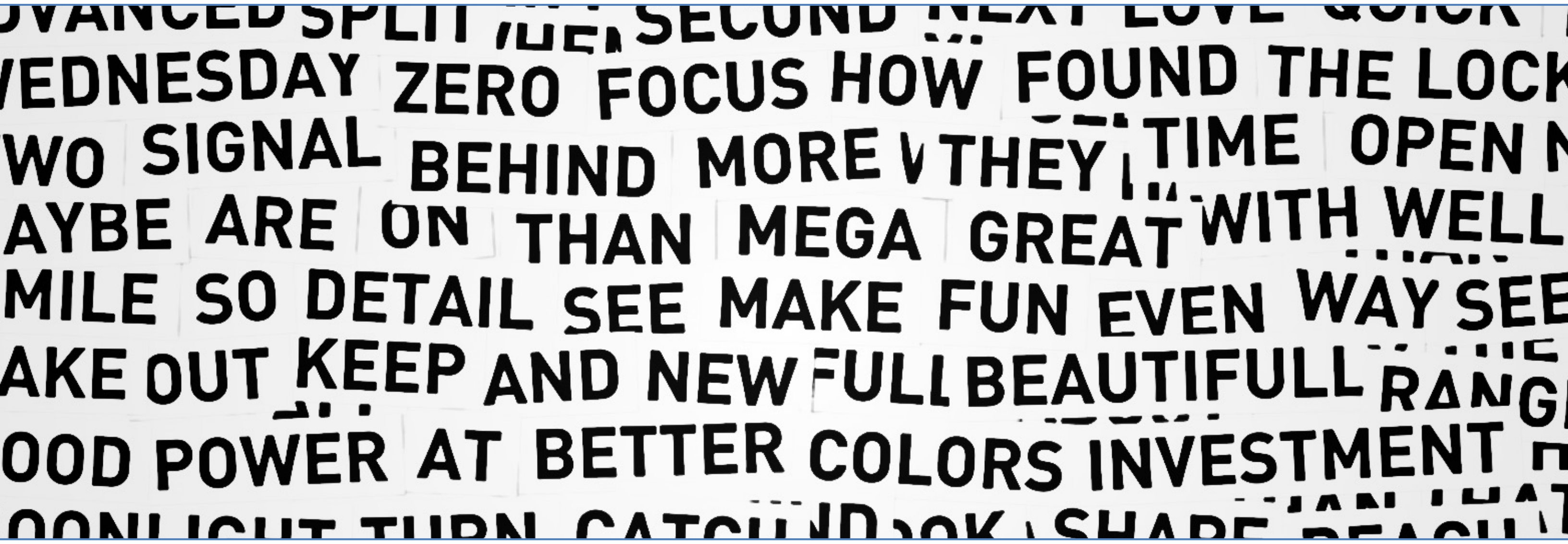
A CASE STUDY

3Play Media as a Case Study

Our proprietary three-step process at 3Play Media begins with verbatim ASR to create time coded transcripts. The next two steps in our process use human editors to edit the transcripts within the existing timecodes, rather than the traditional transcription process of writing out every word as it is heard.

We opt for a verbatim style for our first round of ASR, since it is generally considered a better starting point for captioning and transcription because it tends to retain more words. Some errors are easier to correct than others. Adding missing words back into a sentence can be more difficult than removing inserted words. Even if the inserted words are incorrect, they can provide valuable information such as timing, which is important for captioning.

Therefore, insertions are more useful for our process compared to missing words. In our case, this means our editors are removing/adjusting existing words rather than writing them from scratch or timing the word synchronization from scratch. This **increases the accuracy of the content and synced time-codes** along with **the speed with which we can process and deliver the final caption and transcription products**.


SPLIT SECOND
EDNESDAY ZERO FOCUS HOW FOUND THE LOCK
WO SIGNAL BEHIND MORE THEY TIME OPEN
AYBE ARE ON THAN MEGA GREAT WITH WELL
MILE SO DETAIL SEE MAKE FUN EVEN WAY SEE
AKE OUT KEEP AND NEW FULL BEAUTIFULL RANG
OOD POWER AT BETTER COLORS INVESTMENT

At 3Play Media, **we apply our own post-processing on the ASR engines we use to improve the ASR output further**. We can train mapping models for any input engine and achieve similar results.

We use millions of accurately transcribed words to train models on top of the ASR results, further tuning the accuracy of their initial output. 3Play’s post-processing model is built to deal with the kind of errors that are commonly produced by the engine we use.

This post-processing delivers about a **10% relative improvement in error rate**. We apply this post-processing to the Speechmatics output for our 3Play captioning service, and we expect similar improvements if applied to other engines.

TAKEAWAYS

Takeaways

Last year's State of Automatic Speech Recognition (ASR) report highlighted significant progress in ASR technology. We saw a surge in advancements and the entry of several new players, underscoring the industry's fierce competition and rapid evolution.

This year, the focus has shifted from introducing new engines to optimizing performance metrics and addressing nuanced challenges with existing engines. This shift indicates that ASR technologies are maturing.

More than in previous years, it has become increasingly clear that not all errors are equal, challenging the simplistic interpretation of "accuracy rate" as a standalone metric. This report emphasizes the need for a more nuanced evaluation framework that considers the interaction between three key factors: error rate measurement, transcription style, and use case.

This year, we have identified several key takeaways.

1. Accuracy Models Matter

Accuracy rates can't be taken at face value. Consumers need to be informed and vendors need to be transparent about the model used to define accuracy. Different models imply different strengths.

2. Know the Nuances

Several years ago, Speechmatics was the clear leader. Now we are seeing several top engines with different strengths and weaknesses. The engines are prioritizing different types of content or different styles of transcription. In addition, we're still seeing accuracy depend heavily on the source material - videos with noisy audio and many speakers are still not being transcribed accurately.

Therefore, it's important to keep your use case as well as your source material in mind when evaluating and selecting an ASR engine for your needs.

3. Hallucinations Pose Concerns – Accessibility and Your Brand

Whisper continues to be a competitive engine, though its hallucinations are cause for concern and greater investigation. These hallucinations appear to be more common than initially believed, and the consequences for accessibility – and your brand – could be profound.

4. The Robots Aren’t There Yet

ASR alone is still not sufficient for the captioning use case, especially when it comes to formatting and hallucinations. Human-in-the-loop captioning and transcription workflows remain critical for accuracy, quality, and accessibility.

About 3Play Media

3Play Media provides closed captioning, transcription, and audio description services to make video accessibility easy. We are based in Boston, MA and have been operating since 2007.

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

Made in Boston.

77 N Washington Street
Boston, MA 02114

Also based in:

275 Market Street, Suite 445
Minneapolis, MN 55405

1909 10 Avenue SW
Calgary, AB T3C 0K3
Canada

APPENDIX

Table 1

Formatted Error Rate by Engine

ENGINE	% ERR	% CORR	% SUB	% INS	% DEL
AssemblyAI	17.5	85.2	11.6	2.75	3.16
Whisper Large V2	17.6	86.4	10.4	4.06	3.2
Speechmatics	19.2	84.8	12.7	4.03	2.51
Microsoft	20.1	84	12.8	4.08	3.15
DeepGram	20.1	84.1	10.9	4.16	4.98
Rev AI	21.6	83.1	13.6	4.74	3.29
Whisper Large V3	27.6	83.2	12.4	10.7	4.48
Google Long	29.8	73.6	18.7	3.42	7.64
Google Video	30	73.9	20	3.87	6.14
IBM	43.4	61.8	29	5.2	9.24

Table 2

Word Error Rate by Engine

ENGINE	% ERR	% CORR	% SUB	% INS	% DEL
AssemblyAI	7.47	95.1	2.56	2.53	2.38
Speechmatics	8.15	95.8	2.47	3.98	1.7
Whisper Large V2	9.4	94.7	2.88	4.07	2.44
Microsoft	9.46	94.6	3.07	4.09	2.29
Rev	11	93.8	3.75	4.77	2.48
DeepGram	11.5	92.6	3.27	4.09	4.11
Google Long	15.2	88.3	5.14	3.49	6.59
Whisper Large V3	19.3	91.5	4.49	10.8	4.01
IBM	23.6	81.6	9.98	5.17	8.43
Google Video	14.6	89.3	5.57	3.88	5.13

Table 3

WER of Top 4 ASR Engines by Market

	AssemblyAI	Speechmatics	WhisperV2	Microsoft	Rev	Avg Top 4
eLearning	3.75	4.4	4.6	5.73	7.39	3.97
Goods/Services	4.33	6.01	5.86	7.67	8.65	5.05
News/Networks	5.76	5.03	6.11	6.19	9.13	5.25
Media Publishing	6.36	6.65	6.44	7.97	9.44	6.12
Government	5.82	8.47	8.31	9.44	11.1	6.92
Higher Ed	6.33	7.95	7.94	7.98	10.7	7.16
Associations	7.16	8.59	9.64	10.2	11.3	8.0
Other	7	8.42	11	8.8	9.63	7.77
Sports	9	11	13.1	13.8	15.6	10.2
Cinematic	11.2	9.38	14.8	11.6	12.9	10.2
Tech	9.73	9.73	12.1	10.2	12.6	9.69

Table 4

Average FER of Top 4 ASR Engines by Market

MARKET	AVERAGE TOP 4
eLearning	11.8
Goods/Services	13.4
News/Networks	14.9
Media Publishing	14.9
Government	16.4
Higher Ed	15.2
Associations	16.8
Other	16.2
Sports	20.2
Cinematic	21.2
Tech	19.8

Table 5

Clean Read WER

ENGINE	% ERR	% CORR	% SUB	% INS	% DEL
AssemblyAI	6.87	95.7	2.33	2.6	1.93
Speechmatics	7.74	96.6	2.27	4.35	1.12
Whisper Large V2	8.68	95.5	2.55	4.14	1.99
Microsoft	9.12	95.4	2.9	4.53	1.68
Rev AI	10.5	94.6	3.54	5.09	1.87
DeepGram	10.7	93.7	3.11	4.47	3.16
Google Video	13.4	90.8	5.45	4.28	3.71
Google Long	13.7	90.2	5.09	3.94	4.68
Whisper Large V3	19.2	92.1	4.42	11.3	3.49
IBM	21.7	84	9.79	5.71	6.24

Table 6

Clean Read FER

ENGINE	% ERR	% CORR	% SUB	% INS	% DEL
Whisper Large V2	15.4	90.7	6.04	6.12	3.24
AssemblyAI	15.5	89.6	7.19	5.08	3.21
Speechmatics	17.6	89.5	7.44	7.14	3.05
Microsoft	17.6	88.1	7.61	5.74	4.28
DeepGram	18	89	6.94	6.96	4.09
Rev AI	20	87.9	8.2	7.85	3.92
Google Long	25.7	78.7	11.4	4.35	9.88
Google Video	26	78.6	12.1	4.67	9.27
Whisper Large V3	26.8	86.8	8.19	13.6	4.97
IBM	37.8	65.9	18.8	3.73	15.3

Table 7

Verbatim WER

ENGINE	% ERR	% CORR	% SUB	% INS	% DEL
Speechmatics	9.84	92.2	3.4	2.05	4.4
AssemblyAI U1	10.2	92	3.57	2.19	4.45
Microsoft	10.8	91.1	3.83	1.83	5.09
Whisper Large V2	12.4	91.2	4.35	3.58	4.44
Rev AI	13.1	90	4.65	3.06	5.34
DeepGram N2	14.8	87.6	4.01	2.37	8.43
Google: Video	19.6	82.3	6.11	1.87	11.6
Whisper Large V3	19.8	88.7	4.82	8.44	6.52
Google Long	22.2	79.3	5.4	1.46	15.3
IBM Watson	31.9	70.8	10.8	2.65	18.4

Table 8

Verbatim FER

ENGINE	% ERR	% CORR	% SUB	% INS	% DEL
Speechmatics	20.6	85.8	9.1	6.37	5.09
Microsoft	20.7	84.3	9.17	4.99	6.56
Assembly U1	21.4	86	9.23	7.4	4.8
Whisper Large V2	21.7	85.2	8.5	6.92	6.28
DeepGram N2	23.3	82.8	8.39	6.14	8.8
Rev AI	24.9	82.9	10.3	7.73	6.82
Whisper Large V3	28.6	84.1	8.57	12.7	7.3
Google Video	32.5	70.8	12.6	3.37	16.5
Google Long	34.1	68.7	11.5	2.87	19.7
IBM Watson	48.7	53	19.7	1.67	27.4

Table 9

Overall WER with 3Play

ENGINE	% ERR	% CORR	% SUB	% INS	% DEL
3Play	7.23	94.4	2.24	3.26	1.85
AssemblyAI	7.47	95.1	2.56	2.53	2.38
SMX	8.15	95.8	2.47	3.98	1.7
Whisper Large V2	9.4	94.7	2.88	4.07	2.44
Microsoft	9.46	94.6	3.07	4.09	2.29
Rev	11	93.8	3.75	4.77	2.48
DeepGram	11.5	92.6	3.27	4.09	4.11
Google Video	14.6	89.3	5.57	3.88	5.13
Google Long	15.2	88.3	5.14	3.49	6.59
Whisper Large V3	19.3	91.5	4.49	10.8	4.01
IBM	23.6	81.6	9.98	5.17	8.43

Table 10

Overall FER with 3Play

ENGINE	% ERR	% CORR	% SUB	% INS	% DEL
AssemblyAI	17.5	85.2	11.6	2.75	3.16
Whisper Large V2	17.6	86.4	10.4	4.06	3.2
3Play	18.3	85	12.4	3.08	2.74
SMX	19.2	84.8	12.7	4.03	2.51
Microsoft	20.1	84	12.8	4.08	3.15
DeepGram	20.1	84.1	10.9	4.16	4.98
Rev AI	21.6	83.1	13.6	4.74	3.29
Whisper Large V3	27.6	83.2	12.4	10.7	4.48
Google Long	29.8	73.6	18.7	3.42	7.64
Google Video	30	73.9	20	3.87	6.14
IBM	43.4	61.8	29	5.2	9.24