The 2023 State of Automatic Speech Recognition

An Annual Report

3PLAYMEDIA

ABOUT THE REPORT

Every year, 3Play Media conducts research to learn how the top automatic speech recognition (ASR) engines perform in regard to captioning and transcription. We then publish the results in the State of Automatic Speech Recognition report.

ASR is an integral part of 3Play Media's captioning process, which uses ASR as a first step to create time-coded captions that professional editors then clean up. Better ASR technology leads to a more streamlined process, allowing our editors to work more efficiently.

Therefore, we have a vested interest in understanding the current state of ASR as it pertains to the transcription and captioning use case; we want to make sure we're using the best engine for our use case because it helps us as the first step in producing high quality captions.

We hope you find value in this report. Please let us know your thoughts by connecting with us on social media @3PlayMedia.

TABLE OF CONTENTS

INTRODUCTION

Throughout 2022, there has been undeniable growth in the capabilities of artificial intelligence (AI). From chatbots to autonomous vehicles, AI use cases are varied, and technological advances are improving rapidly—including for captioning and transcription.

This year's State of ASR results were some of the most complex and disruptive we have ever seen. Accuracy is higher, and error rates are lower. With impressive new entrants and improvement across industry leaders, the field is more competitive now than ever.

While we have always been interested in the nuances of different error types in ASR, we have never observed the tested engines perform as close in accuracy as they did this year, especially concerning word error rate. With several players producing competitive word error rates, we are forced to look deeper at this key differentiator and examine the nuances.

Given the tested engines' close performance, differentiation is complex, and the nuances of error types matter. Is one kind of error worse than another for accessibility? Which error types contribute to more time spent editing? These considerations and more factor into our decision on which engine is best for our use case of captioning and transcription—and should factor into your decisions about which engines are best for your unique use cases. We plan to conduct more research to discern which details matter most, as the varied error types affect our efficiencies differently.

The Research

TESTING

Our research tested the performance of the most popular ASR engines across several industries to evaluate the accuracy of each as applied to pre-recorded captioning and transcription. Additionally, we tested how each engine performed across aggregate industries for overall accuracy as well as how each performed on content broken down by the primary market.

Files were transcribed through 3Play's normal, multi-step process, which includes ASR and two rounds of human cleanup and quality review, to create a 99%+ accurate control.

EVALUATING ACCURACY

We measure accuracy in two ways: Word Error Rate (WER) and Formatted Error Rate (FER).

Word Error Rate (WER) is commonly used to determine quality in ASR. If you've ever seen the label "99% accurate captions," then those captions have a WER of 1%. WER is a formatting-agnostic measurement, meaning that WER scores do not count errors in capitalization, punctuation, or number formatting.

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) is used to better evaluate the human experience of accuracy in the context of transcription and captioning and the amount of additional work needed to make a transcript fully accessible. FER is the percentage of word errors when formatting elements such as punctuation, grammar, speaker identification, non-speech elements, capitalization, and other notations are considered. Formatting errors are particularly 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.

DATASET

While the State of ASR always involves a massive amount of effort and time from 3Play Media's data science team, **the 2023 report required even more work** than usual.

To evaluate ASR in the pre-recorded context, we tested 549 files that were uploaded to 3Play Media for English transcription in 2022. Those files include 107 hours of content and 929,795 total words. All files had a word count of at least 100 words.

In terms of audio duration, content increased 57% from last year's 68 hours. Similarly, the word count increased by 56% from last year's 597,675 words.

The files tested are representative of the type of content we transcribe at 3Play Media. They come from multiple markets with diverse subject matters, speaker locales and accents, video lengths, customer upload volumes, audio qualities, number of speakers, and scripted and nonscripted content.

Additionally, we also used two styles of transcription: verbatim and clean read. The majority of our transcripts are done in the clean read style, which omits hesitations such as “uh” and “um,” fillers such as “you know” and “like,” and false starts where the speaker starts over. ASR engines vary on whether they include these disfluencies.

These decisions were made to achieve data with as little bias as possible. 3Play Media has a vested interest in the results being accurate, as we use the results from this research to inform business decisions and improve our process and output.

THE BREAKDOWN

The following APIs/engines were used for testing:

- ★ 3Play Media: Speechmatics with 3Play's proprietary post-processing applied
- ★ Speechmatics: Speechmatics without 3Play post-processing
- ★ AssemblyAI: V9 model
- ★ Whisper Large: Whisper's largest model for out-of-the-box performance
- ★ Whisper Tiny: Whisper's smallest model for out-of-the-box performance
- ★ Microsoft
- ★ Rev.ai: Rev's V2 model
- ★ Google VM: Google's Enhanced Video Model, which is optimized for video
- ★ Google Standard: Standard Model
- ★ IBM Watson

The following markets were evaluated:

- ★ 34% Higher Ed
- ★ 16% Technology
- ★ 15% Consumer Goods
- ★ 9% Cinematic
- ★ 8% Associations
- ★ 7% Sports
- ★ 4% Publishing
- ★ 3% eLearning
- ★ 3% News & Networks

3PLAY'S PROPRIETARY POST-PROCESSING

Throughout this report, 3Play Media is included as its own engine line item. However, it is essential to note the difference between 3Play Media and the other engines tested. 3Play Media is not an engine on its own but rather represents the results of our model and propriety post-processing applied to Speechmatics' output, which results in a 10% relative improvement in transcript accuracy.

From our extensive body of matching ASR and corrected transcripts, we learn what ASR errors are common and what corrections our editors typically need to make. With our proprietary post-processing, we can apply these learned corrections to our ASR outputs pre-emptively to create more accurate transcripts that are easier to understand and edit.

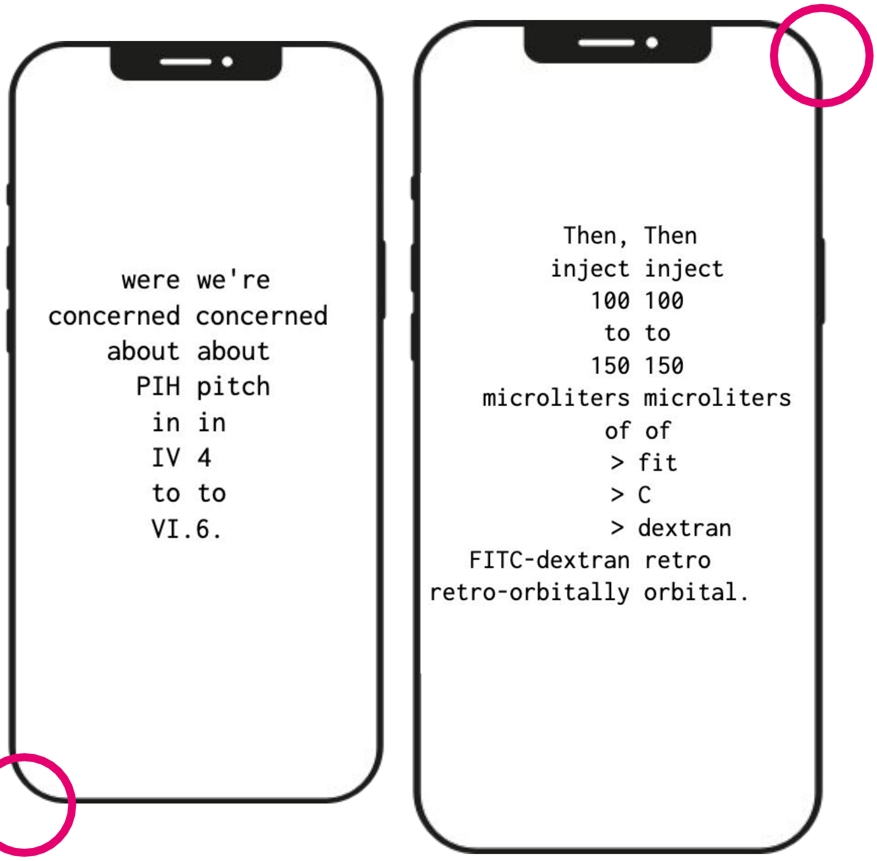
We can adapt this type of post-processing to any engine but currently only have it implemented for our own primary vendor, Speechmatics. We would expect to see the same 10% relative improvement for any engine to which we tuned our modeling and post-processing.

We did not test our post-processing on other vendors for this report because we would need to train our models on a specific engine's output to expect good performance. However, other vendors have shown to be strong competitors, and we plan to explore their advancements further to validate the nuances that are best for our use case.

The Results

RESULTS

When analyzing the testing results, we first look at WER results across all tested engines. Next, we add in FER results, which are a critical component when evaluating accuracy for captioning and transcription, where grammar, punctuation, and speaker labels matter. For our use case, FER provides an extremely helpful marker of how much work our editors must put in to produce an accessible transcript. Last, we take a closer look at how these scores change based on the primary market of tested content.



were	we're
concerned	concerned
about	about
PIH	pitch
in	in
IV	4
to	to
VI.	6.

Then,	Then
inject	inject
100	100
to	to
150	150
microliters	microliters
of	of
>	fit
>	C
>	dextran
FITC-dextran	retro
retro-orbitally	orbital.

These are examples from actual testing that we conducted this year. They tend to focus on names and complex vocabulary that require human expertise and knowledge. In each case, the truth is on the left, and the ASR is on the right.

WORD ERROR RATE

The overall error rate shows that 3Play Media (which includes our modeling and post-processing) had the lowest WER of 6.86%, even lower than last year's 7.96%. 3Play Media was followed by AssemblyAI at 7.5% and Speechmatics alone at 7.56%, two engines with a statically insignificant WER difference. Speechmatics, Microsoft, and Rev all made impactful improvements from last year while Google and IBM lost ground.

The table below shows the tested vendors and their respective WER results.

Vendor	% ERR	% CORR	% SUB	% INS	% DEL
3Play Media	6.86	96.1	2.3	2.95	1.61
AssemblyAI	7.5	93.9	2.98	1.35	3.17
Speechmatics	7.56	96	2.48	3.61	1.48
Whisper: Large	8.42	94.2	2.39	2.57	3.45
Microsoft	9.69	94.1	3.64	3.82	2.23
Rev AI	10.4	94.1	3.86	4.53	2
Google Video	13.5	90.3	5.46	3.78	4.27
Whisper: Tiny	15.1	89	7.48	4.1	3.49
IBM Watson	24.8	80.7	12.6	5.45	6.7
Google Standard	28.1	75.3	9.62	3.42	15.1

QUICK TIPS *WER scores do not count errors in capitalization or punctuation.*

While 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. Substitutions, insertions, and deletions are all essential for the captioning use case.

For example, a low deletion rate is integral to using ASR for creating captions and transcripts, as deleting words can change the meaning and convey incorrect information. At 3Play, we use ASR as the first stage of our process, and having the greatest number of words transcribed (fewer deletions) allows our human editors to more accurately and efficiently edit files. Our results show that Speechmatics and 3Play Media both have the lowest deletion rates, at 1.48% and 1.61%, respectively.

Additionally, rankings change significantly between percent error and percent correct.

Percent error, or error rate, measures the number of errors of any type the ASR made per word in the transcript. This includes errors of any type, such as insertions, deletions, or substitutions.

Percent correct measures the percentage of words in the transcript that were accurately transcribed by ASR. In percent correct, only some error types result in an incorrect transcription of a word. Insertions, which are additive, don't impact percent correct, which is why a high percent correct score does not necessarily translate to few errors.

Different vendors choose to prioritize different error types, and these qualities translate into nuances that impact captioning and transcription at scale.
While AssemblyAI has the lowest error rate of the external vendors, it ranks fifth out of nine in percent correct. AssemblyAI also substitutes the fewest words, whereas Google deletes the highest percentage of words and IBM substitutes the highest percentage of words. Extremely high deletions, such as Google's 15.1%, and substitutions, such as IBM's 12.6%, would pose extreme challenges for captioning and transcription.

FORMATTED ERROR RATE

FER is especially important for our 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 critical to readability and meaning, and an accuracy rate under 85% is extremely noticeable. Our testing shows that even the best-performing engine, Whisper: Large, is still only around 83% accurate, which means that one in six words is incorrect.

Both Whisper and AssemblyAI outperform 3Play's post-processing. However, the percent error is extremely close, at 17.8% for 3Play Media, 17.5% for AssemblyAI, and 17.2% for Whisper: Large, all of which perform far better than the other tested engines. The table below shows the tested vendors and their respective FER results.

Vendor	**% ERR**	**% CORR**	**% SUB**	**% INS**	**% DEL**
Whisper: Large	17.2	85.3	9.86	2.53	4.82
AssemblyAI	17.5	84.1	11.4	1.63	4.46
3Play Media	17.8	85.2	11.9	3.03	2.93
Speechmatics	18.3	85.4	11.8	3.74	2.78
Rev AI	21.5	83.2	13.6	4.67	3.24
Microsoft	22.3	81.6	14.8	3.9	3.56
Whisper: Tiny	25.4	78.6	16.7	3.93	4.71
Google Video	29.8	74.1	20.3	3.85	5.6
Google Standard	41.6	61.7	22	3.34	16.3
IBM Watson	42.5	63.5	29	5.95	7.57

QUICK TIPS *FER accounts for formatting elements such as punctuation, grammar, speaker ID, non-speech elements, and capitalization.*

ADDRESSING BIASES

The majority of our content is transcribed as clean read, so we are biased towards giving higher scores to engines that omit disfluencies altogether.

More “verbatim” engines will have more insertion errors than others, and there is likely some bias in the form of “editing inertia,” in which the final outputs may more closely match Speechmatics, especially anywhere that an editor failed to correct an error. We think that this bias is slightly mediated by the fact that all files originally ran through a different version of Speechmatics ASR than the engine we tested here.

The charts on the next two pages show WER for clean read files and verbatim files.

WER for Clean Read Files

Vendor	% ERR	% CORR	% SUB	% INS	% DEL
AssemblyAI	6.39	95	2.72	1.36	2.31
3Play Media	6.41	96.7	2.11	3.11	1.2
Speechmatics	7.2	96.6	2.28	3.84	1.09
Whisper: Large	8.02	94.6	2.19	2.6	3.23
Microsoft	9.06	95	3.32	4.05	1.69
Rev AI	9.92	94.9	3.57	4.8	1.55
Google: Enhanced	12.3	91.7	5.1	4.01	3.23
Whisper: Tiny	13.8	90.3	6.72	4.09	3.02
IBM Watson	23.2	82.6	12.2	5.78	5.19
Google: Standard	25.9	77.8	9.65	3.68	12.5

Across the board, the engines did much better on clean read files than verbatim files, with AssemblyAI maintaining a statistically significant lead on clean read files. The difference is likely attributable to the relative ease of the task for the markets that make up each group of files.

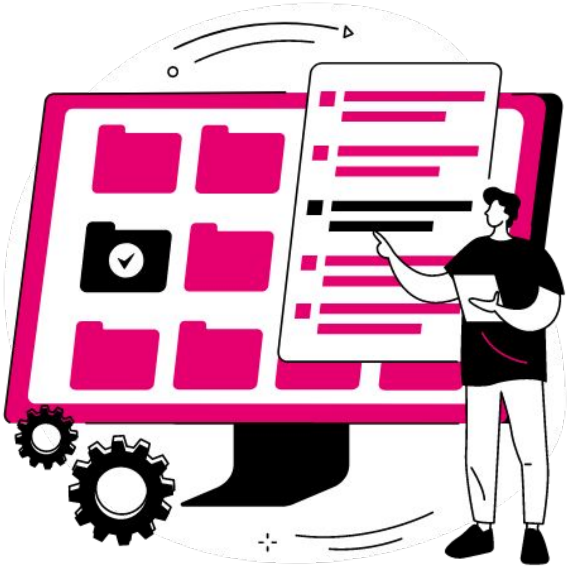
Of note, AssemblyAI drops from first to sixth place on verbatim files. 3Play's post-processing and Speechmatics take a very clear lead. It's also notable that while for clean read files, insertions made up 53% of Speechmatics errors, on verbatim files, insertions make up only 22% of Speechmatics errors. In terms of biases, we recognize that part of AssemblyAI's strong performance is due to their clean read style, whereas part of Speechmatics' lower performance is due to their verbatim style.

WER for Verbatim Files

Vendor	% ERR	% CORR	% SUB	% INS	% DEL
3Play Media	9.56	92.4	3.47	1.97	4.11
Speechmatics	9.74	92.5	3.7	2.2	3.84
Whisper: Large	10.8	91.6	3.64	2.4	4.8
Rev AI	13.2	89.7	5.63	2.89	4.7
Microsoft	13.5	88.9	5.58	2.4	5.53
AssemblyAI	14.2	87.1	4.51	1.3	8.4
Google: Enhanced	20.6	81.8	7.64	2.4	10.5
Whisper: Tiny	22.6	81.6	12.1	4.12	6.37
IBM Watson	34.2	69.2	14.9	3.44	15.9
Google: Standard	41.6	60.2	9.41	1.81	30.4

While we prioritize clean read files, **one style is not inherently better than the other**, so engines should be evaluated with this in mind. For example, customers with scripted content in the media and entertainment industry often prefer a verbatim style, as any disfluencies are intentional. However, customers with non-scripted content often prefer a clean read style, which omits unintentional disfluencies.

2022 VS. 2023



Last year's results showed that Speechmatics maintained its edge across all markets. However, with this year's additions of AssemblyAI and Whisper, Speechmatics is no longer the clear industry leader.

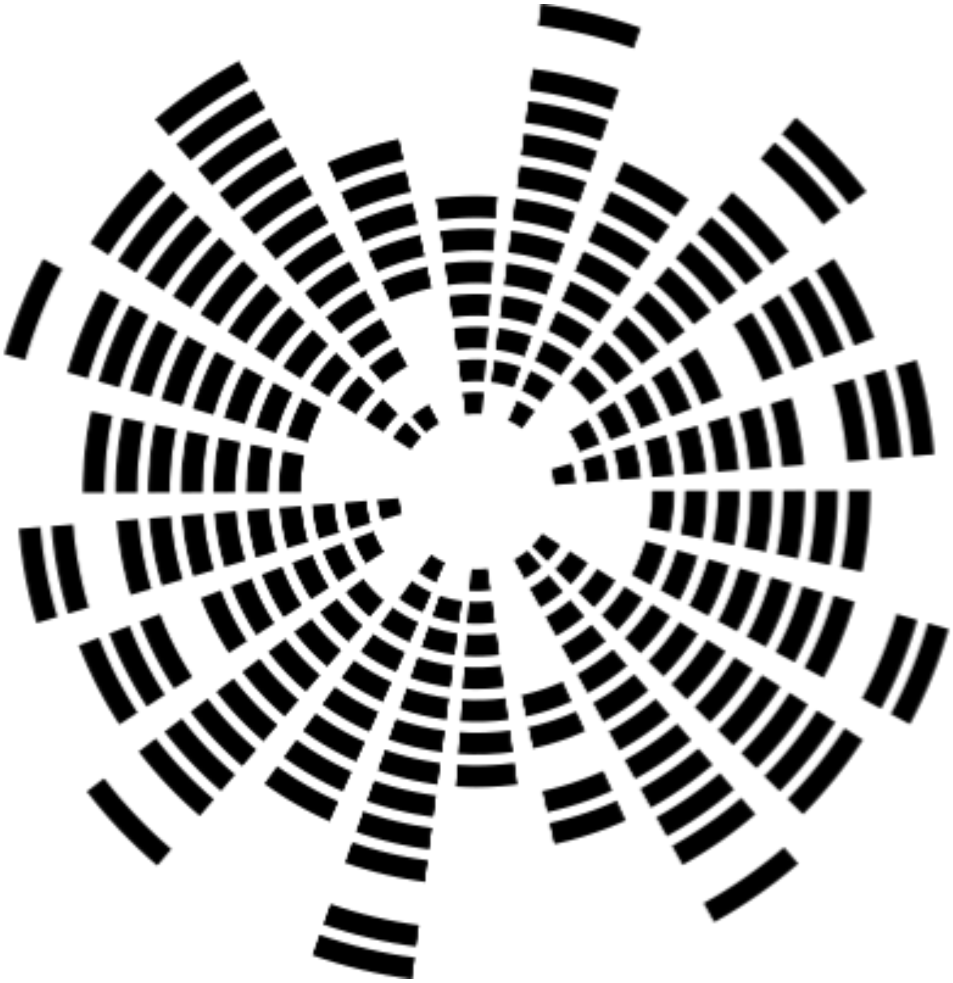
In general, **performance has improved by several absolute percentage points across the board since our last report**, published in 2022. In WER, 3Play Media improved from 7.96% to 6.86%, and Speechmatics alone improved from 8.67% to 7.56%. Rev and Microsoft also improved, but Google VM and Standard declined.

In FER, 3Play Media declined from 17.2% to 17.8% and Speechmatics alone declined from 17.9% to 18.3%. Rev improved from 24.9% to 21.5%, and Google VM and Standard both declined, VM from 27% to 29.8% and Standard from 38.2% to 41.6%.

Our results indicate that **formatting accuracy is starting to plateau** for the highest performing engines. The decline in FER is slight, but given the improvement in WER, the difference might be due to the fact that our dataset in this report is more difficult than last year's dataset.

We are not particularly surprised that formatting accuracy might plateau. Our captioning standards are specific and challenging to meet, and engines are not necessarily working to automatically add caption features given the use cases for which they are trained.

We have also changed the industry breakdown this year, with more content from some of the more difficult markets. This change could be why certain engines, such as Google Standard and IBM, performed worse this year even though their models have probably not changed since last year's testing.



Performance by Primary Market

PERFORMANCE BY PRIMARY MARKET

The tables below show the WER and FER averages, respectively, of the top four vendors (3Play Media, Speechmatics, AssemblyAI, and Whisper) for the different markets.

Market	WER Average of the Top 4 Vendors
Publishing	7.74
Consumer Goods	8.51
News & Networks	11.1
Tech	5.5
Higher Ed	6.38
eLearning	4.07
Sports	9.94
Cinematic	10.8
Associations	6.43

Market	FER Average of the Top 4 Vendors
Publishing	18.2
Consumer Goods	17.7
News & Networks	26.4
Tech	14.5
Higher Ed	16
eLearning	13.4
Sports	21.4
Cinematic	25
Associations	15.9

Extremely important to our use case is how an ASR engine performs in regard to specific markets. To gain deeper insight, we broke the content down into the primary market of the customer who uploaded the file.

Certain markets create unique challenges for ASR. Notably, News & Networks, Cinematic, and Sports are the toughest for ASR to transcribe accurately, as these markets often have content with background music, overlapping speech, and difficult audio. These markets have the highest average error rates for WER and FER, with News & Networks being the least accurate. The WER for News & Networks, Cinematic, and Sports is 11.1%, 10.8%, and 9.94%, respectively, and the FER is 26.4%, 25%, and 21.4%. In these markets, ASR is far from being a good solution on its own.

Whisper performed particularly poorly in Cinematic FER, with a percent error of 32.6%. In comparison, AssemblyAI, 3Play Media, and Speechmatics had Cinematic FERs of 25%, 23.8%, and 23.7%, respectively.

Additionally, some markets have low WERs with corresponding lower-than-average FERs. For example, eLearning is the easiest market to transcribe, with a WER of 4.07% and FER of 13.4%. However, although eLearning appears easiest to transcribe, FER is particularly important in educational content, in which inaccurate captions are detrimental to student learning. Additionally, only 3% of our total tested files were eLearning content, so there may be more validation needed.

Though overall WERs continue to decrease year over year for top vendors, **WER and FER results remain high enough to warrant editor transcription for all markets.**

Hallucinations: A New Kind of Error

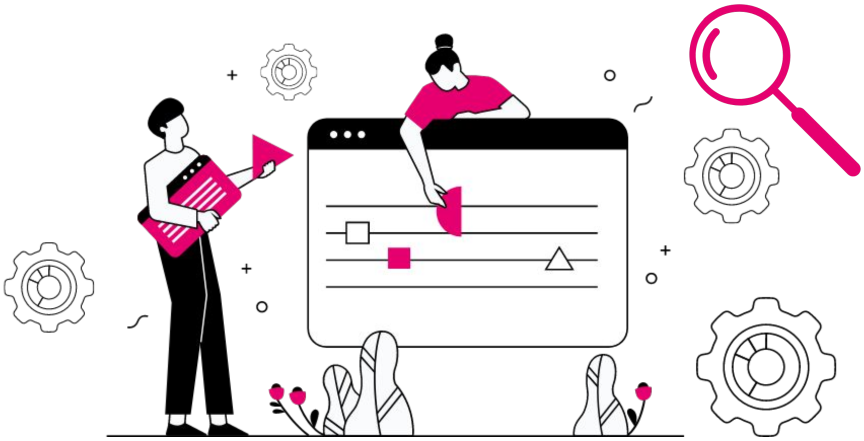
HALLUCINATIONS

Whisper has a well documented tendency to 'hallucinate,' or to generate text that has no basis in the audio. OpenAI has acknowledged this tendency and suggested that it can be mitigated through fine-tuning, saying, "We hypothesize that this happens because, given their general knowledge of language, the models combine trying to predict the next word in audio with trying to transcribe the audio itself."

Fine-tuning would require the process of taking an existing machine learning model, in this case, Whisper, and training it with a specific data set to increase accuracy for your use case. This process is expensive and requires technological expertise, so while Whisper may be useful, working out these quirks would not be an easy task.

Our data showed evidence of these hallucinations, often occurring when the topic shifted. Hallucinations manifested mainly as insertion and substitution errors—insertions if they occurred over silence and substitutions if they occurred during speech.

Some of the hallucinations were significant and could pose issues for the captioning use case in particular. However, hallucinations seemed rare and did not prevent Whisper from performing competitively.



Below is an example of a recorded medical school lecture with hallucinations in the transcript. The hallucinations appear plausible, as if they represent words that could have been said during the audio. However, the audio showed that no one spoke during the brief time when the hallucinations occurred.

Correct Transcript:	**Whisper Transcript:**
Thank you very much. [APPLAUSE}	Thank you very much. This is just one of many we wish you all the luck at your surgery.

Whisper's made-up well wishes—"This is just one of many we wish you all the luck at the end of your surgery"—is a seemingly plausible sentence but is in fact completely inaccurate. This kind of error is different from typical ASR errors because it appears credible, which brings us to an essential point: **Not all errors are created equal.**

From an editing perspective, hallucinations are problematic. An editor would need to delete the entire section, retype all the words, and then correctly sync the text with the audio—a fairly expensive error for a captioning company.

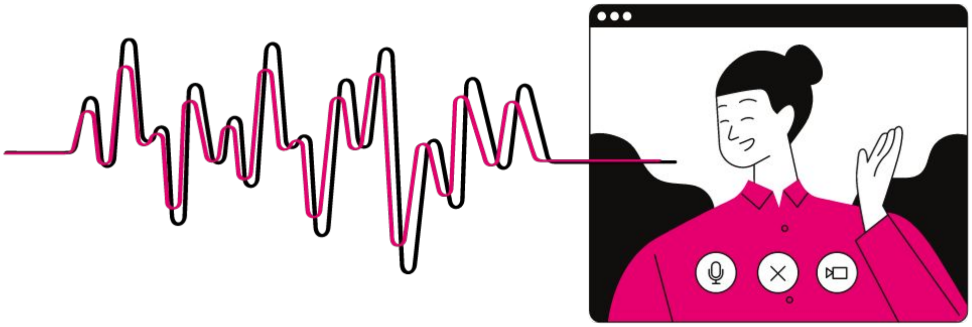
From an accessibility standpoint, hallucinations present an even more egregious problem: The false portrayal of accuracy for deaf and hard-of-hearing viewers. If using Whisper for auto-captions, a deaf or hard-of-hearing viewer might not know that the captions are failing to capture what is being spoken.

While Whisper might perform impressively from the standpoint of well punctuated, grammatical sentences, from an accessibility standpoint, the risks are currently high.

Takeaways

TAKEAWAYS

This year's State of ASR shows great advances in the state of the art, and we will likely continue to see improvements in years to come. Here are several key takeaways we hope you've gained from this report.



New Models Are Emerging

With the exciting addition of two new engines, Assembly and Whisper, Speechmatics, the industry leader for many years, is no longer the clear singular winner. However, all are extremely competitive, relatively equivalent in overall accuracy, and excel in different use cases.

Source Material Matters

While WER has continued to show improvement year over year and is impressively low for certain markets, accuracy results are still heavily dependent on audio quality and content difficulty. Most improvements are driven by training techniques, not changes to technology.

Use Case Matters

While our testing is focused on the transcription and captioning use case, the engines tested are ultimately trained for varied use cases. Understanding your use case and which engine best suits it is critical to producing the highest quality result.

Hallucinations Pose Concerns for Accessibility

Whisper proved to be an extremely competitive engine, but its hallucinations, though rare occurrences, may be cause for concern and greater investigation. While it's possible that these hallucinations would be reduced through fine-tuning, the negative consequences for accessibility could be profound.

Overall, Whisper has extremely competitive performance in the formatting and even captioning feature space. Since it was trained mostly on captions for publicly available online video, it seems well-tuned to the captioning use case and was even able to tag audio features in a handful of instances.

ASR Still Is Not Good Enough

High-quality ASR does not necessarily lead to high-quality captions. For WER, even the best engines only performed around 90% accurately, and for FER, only around 80% accurately, neither of which is sufficient for legal compliance and 99% accuracy, the industry standard for accessibility.

The ASR landscape has undoubtedly evolved, and the captioning industry continues to benefit from the many ways ASR simplifies, expedites, and helps to scale captioning and transcription processes. However, we argue that human editors remain indispensable in producing high-quality captions accessible to our primary end users: people who are deaf and hard-of-hearing.

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