ACCESS - State of ASR

[00:00:00.80] JACLYN LAZZARI: Well, I think we'll go ahead and get started. Thank you so much for joining us today. Welcome to ACCESS 2024. My name is Jaclyn, and I'm on the marketing team here at 3Play Media. My pronouns are she/her, and I'm a young, white woman with dark blonde hair, and I'm wearing a brown cardigan.

[00:00:21.34] So before we dive into the discussion, I'd like to go over just a few housekeeping items. This presentation is being live captioned, and you can view those captions by clicking the CC icon in your control panel. This session also features ASL interpretation courtesy of Deaf Services Unlimited. Please feel free to ask questions throughout the presentation using the Q&A window or the chat.

[00:00:50.87] And without further ado, I'm happy to welcome you all to this session entitled The State of Automatic Speech Recognition. Today, we're joined by Elisa Lewis, senior brand marketing manager at 3Play Media, and Tessa Kettelberger, Senior Data Scientist at 3Play Media. So thank you so much, Elisa and Tessa. I'm excited to pass it off to you both for an exciting presentation.

[00:01:19.81] ELISA LEWIS: Hi, everyone. Thank you so much for joining us. We're really excited to kick off ACCESS today. Hopefully some of you were able to join us this morning for our keynote session. And thank you for joining The State of ASR. We're really excited to dive into this.

[00:01:39.47] So before we get started, I want to do introductions. As Jaclyn mentioned, my name is Elisa Lewis. I'm the senior brand marketing manager here at 3Play Media. My pronouns are she/her. I'm a fair-skinned woman in my early 30s. I have dark brown hair, and I'm wearing a beige sweater with black polka dots. And I will hand it to Tessa to introduce herself.

[00:02:05.81] TESSA KETTELBERGER: I'm Tessa Kettelberger. I am a young, white woman, brown hair, wearing a green shirt. And I am a senior data scientist at 3Play Media. I've worked here for five years, and I did the research backing up this State of ASR Report, as well as many years past.

[00:02:26.12] ELISA LEWIS: Great. Thanks, Tessa. So for an agenda today, we're going to go through a quick overview of ASR, or Automatic Speech Recognition. We want to make sure that the audience is all on the same page around some of the terminology and the definitions that we're going to be using throughout this session.

[00:02:43.55] And then we will go through what goes into our annual State of ASR Report. I'll hand it over to Tessa. She'll talk about the research, the results, and the trends that we saw this year. Then we'll talk about what this means for each of you and how you can actually utilize the information in this report. And we'll finish off with some key takeaways and conclusions.

[00:03:04.50] So really excited-- this is actually the first time that we're presenting our full findings from this year's research. And we're excited to share the written report out in the coming weeks. So stay tuned at the end for the link to that.

[00:03:19.39] So as we get started, like I said, I want to level-set on some definitions here. So ASR, which you'll hear throughout the presentation, stands for Automatic Speech Recognition. It refers to the use of different processes, like machine learning, natural language processing, and artificial intelligence and how these technologies are used to convert speech into text.

[00:03:44.34] So ASR is used in a number of different ways, and many of them, you are probably familiar with in your day-to-day life. For example, you've probably used automated assistant tools like Siri or Alexa. You might use ASR for transcription or captioning. And ASR is improved by using modeling of truth data.

[00:04:07.60] So what this means is that we actually use a fully accurate transcript to point out inaccuracies. And the AI, or the Artificial Intelligence, can actually learn from this. They can learn from the source of truth and improve from their mistakes.

[00:04:25.10] So for the purpose of today's session, we're going to be focusing the discussion on ASR for the use case or the task of transcription and captioning. Of course, we at 3Play Media are a captioning company. So I do want to talk a little bit about what differentiates the use case of ASR when we talk about captioning from some of those different use cases, like automated assistance and some of the things that I mentioned previously.

[00:04:53.68] So we often get this question. People ask us, why is Siri so good and auto captions are so bad? So it's a good question. There are actually a few reasons for this. When we're using an automated assistant, you're usually a single speaker.

[00:05:11.04] So think about when you're asking-- you're speaking into your phone or your Alexa device, and you're asking, what's the weather? It's a single speaker. Your device can actually learn your voice. If you've ever had your smart device address you by your name, it's a little bit startling, but they actually do learn who's speaking to them, and they kind of adapt to your particular voice and your style to understand you better.

[00:05:37.27] And there's usually a high-quality audio. There's not a lot of background noise when you're talking to them. You're usually speaking, like I said, directly into them, no other speakers or things like that. And the microphone is usually pretty close to you.

[00:05:54.23] The other thing to note is that it's usually a constrained task that you're asking. So like I said, you're maybe asking, What's the weather? or you're asking a simple, straightforward question. And Siri can also ask for clarification. It can say, "I didn't quite get that," or "Is this what you asked?" And then it can iterate from there.

[00:06:19.36] It also doesn't have to get everything perfectly correct. It doesn't have to understand everything to get the gist of what it's saying because, like I said, it is only trained on

certain tasks. So if it gets the gist, it can assume, from there, what you're asking and answer accordingly.

[00:06:40.86] Meanwhile, when we think about automatic captions and we think about what's needed for captioning and transcription, there's a lot of factors that need to be considered. So usually there are multiple speakers. The tasks are more open-ended. So we could have an hour-long presentation, like the one that you're in today.

[00:07:04.98] There could be content and topics shifting from one topic to another. It's possible that there could be background noise or poor audio quality if you're on a Zoom session. Hopefully we don't experience this today, but it's possible that you lose frequency.

[00:07:25.77] And most of us don't speak perfectly. And we have disfluencies in our language. We maybe say "um" or "uh." And this is worse when there is a longer-form content.

[00:07:38.80] So if you're just asking, Hey, Siri, what's the weather? you're much more unlikely to say "um" or "uh" or have false starts. But in longer content, where you might be using captioning and transcription, it's more likely that you will have some of those disfluencies.

[00:07:59.31] So I want to talk about the report itself. The State of ASR Report is an annual report. We do this every year. And we typically review about 10 of the top automatic speech recognition engines. So we test how they perform, like I said, for the task of captioning and transcription.

[00:08:23.00] We look at a few different types of error rates. We test both Word Error Rate, or WER-- W-E-R-- and Formatted Error Rate, FER, or F-E-R, and we'll talk more about those a bit later in the session. But I want to mention our goal. The goal of this, since 3Play Media is a captioning and transcription company, is to really understand what engines are-- how they're performing, how the industry looks, what's changed-- how has technology changed?

[00:08:57.57] I do want to mention that people may think that there's some bias, given that, again, we are a captioning company. But we have a really vested interest in understanding an unbiased view of which engines are performing best because we do use automatic speech recognition in our process. It's the first step of our process, and then we have human-corrected-- two rounds of human review and quality to make sure that our captioning is fully accurate.

[00:09:29.31] So using the best engine and identifying what's best for our use case is really important to our business. And it's important that we're able to do that so that we can continue to provide the highest quality captioning that we can.

[00:09:47.19] And we touched on this already briefly, but I want to reiterate a couple ways more specifically that captioning prevents-- or sorry-- presents some unique challenges and a unique use case for ASR. So there are three main categories here that explain this difference.

[00:10:11.28] The first is variety. There's a ton of content in the world that can be captioned, so there's variety in the environment that we're captioning in and the subject matter. So it makes the task a lot more variable than, again, an automated assistant or something of that nature.

[00:10:30.52] There's length. Again, the content is typically a longer form, longer-form audio. For the most part, it's not just those short commands that are commonly used for Siri or Alexa. So there's a ton of content. It's more likely that, again, you'll experience those disfluencies than you would in a shorter content.

[00:10:52.78] And the last note that I think is really interesting is readability. The accuracy of captioning and transcription is really about the human experience of reading. And when we're trying to provide equal access, we want to make sure that we're thinking about things like grammar, punctuation, speaker identification, that we're using proper sentence case.

[00:11:16.81] And this is not necessary when we're talking about automated assistance or other use cases. So this is also an area that ASR is typically really challenged by, so it's definitely something that we're going to be talking about throughout this presentation today.

[00:11:37.26] And then I mentioned this already, but how does the 3Play use ASR? It's the first step of our captioning process, followed by two rounds of human editing and review. So the better the engine, the easier the job is for those humans to get it to a really accurate place.

[00:11:58.09] And then we also do our own post-processing on the ASR engines, so we're able to further improve the ASR output. We have millions of accurately transcribed words that we use to model on top of our ASR to further tune the results. So with that, I will hand it over to Tessa to talk through some of the data.

[00:12:25.09] TESSA KETTELBERGER: Hi, everybody. I'm excited to get into the data. Let's see what we started with. So we started with-- we tested 11 ASR engines this year. That's more than previous years.

[00:12:38.98] We did it on more data than we have ever before. So it was 158 hours of data coming from 700 different videos. It was over 1.3 million words, and it came from 10 different industries. You can go to the next slide.

[00:13:02.10] So the engines we tested this year were Speechmatics, two of AssemblyAI's engines-- their Universal-1 engine and their Conformer-2-- Microsoft, rev.ai, Deepgram's Nova-2 model, IBM, Google's model they've called Latest_long, as well as Google's Enhanced Video Model, Whisper's large-v2 model, and their more recent release, the Whisper large-v3 model.

[00:13:31.53] We pulled data to test on these engines from the following industries. We did higher ed. We have tech, goods and services, cinematic content, associations, sports, government, media publishing, e-learning, and news networks.

[00:13:49.23] And we do this because it's in our interest, as Elisa has described, to get a really general sample of audio by industry, audio quality, different speaking styles, et cetera. We use

this research to make business decisions internally. We want the results to reflect our business needs, which are really general and cover a variety of content. So we try really hard to make sure that that's what's in our test data set.

[00:14:22.49] Yes, all the recordings are in English. We didn't test any other languages this year. So we test two different metrics. The first is Word Error Rate. I will sometimes refer to this as WER, for short.

[00:14:39.89] And word error rate is a ratio. It's the number of errors that occur in a transcription per 100 spoken words. So it looks a lot like a percentage. A transcript that has a 10% word error rate will have 10 errors per every 100 spoken words.

[00:14:58.94] And this is a really common measurement in the industry. You'll see different definitions within the industry about what counts as an error when you see word error rate. But for us and for the purposes of this report, we are only counting word errors for word error rate as differences in the word content. So we aren't counting differences in punctuation, capitalization, or number formatting as errors for word error rate. We are normalizing our transcripts heavily to try to eliminate that.

[00:15:32.96] Then we have Formatted Error Rate, or I will call it FER for short. And this is a variation on word error rate. It's also a ratio of errors to real words based on 100 words. So it looks like a percentage. The methodology there is very much the same.

[00:15:47.90] But for this, we are counting differences in punctuation, in capitalization, in number formatting, all of that as errors. And we're doing very little normalization when we measure formatted error rate. So this measurement is meant to capture, rather than just the content accurately captured by ASR, but also the experience of trying to use this speech recognition output as a caption or as a transcript to follow along with a video.

[00:16:18.29] This speaks to the amount of editing work required to create proper captions from speech recognition output and also the experience of reading and trying to understand the ASR on its own. We think both of these are really important.

[00:16:37.06] So to get into word error rate, these are the results from this year's study in word error rate. I'll read through the table for everybody, and then I'll get into some takeaways. So AssemblyAI's Conformer-2 model had a word error rate of 7.13. AssemblyAI's Universal-1 was 7.47. Speechmatics had 8.15. Whisper's large-v2 had 9.4.

[00:17:03.03] Microsoft had 9.46. rev.ai had 11. Deepgram's Nova-2 model has 11.5. Google's Video model has 14.6. Google's Latest_long model has 15.2. Whisper large-v3 is 19.3, and IBM is 23.6.

[00:17:25.80] So AssemblyAI has really taken first place this year. It's a really exciting result. Speechmatics has come in second, and Whisper and Microsoft had pretty similar results in third and fourth place.

[00:17:39.76] I want to talk about this in greater detail later in the presentation, but I want to acknowledge that you can see we tested some pairs of new and old engines that come from the same provider, and we didn't always see improved performance on the newer engines. You can see that with Whisper, with Assembly, and with Google.

[00:17:59.32] I also want to point out-- and we'll discuss this more later-- that two of these models are going to be deprecated soon and won't be available later. So if you're really excited by the Conformer-2 results, that model is already gone, and you can no longer access it. They very recently deprecated it. And the Google Video model is also going to be deprecated this year. I'm not sure exactly what their timeline is, but they are moving customers over to newer models.

[00:18:30.00] So in the formatting space, I will also read through this table. AssemblyAI's Conformer-2 scored a 17. AssemblyAI's Universal-1 model scored a 17.5. Whisper large-v2 scored 17.6. Speechmatics scored 19.2. Microsoft scored a 20.1.

[00:18:52.54] Deepgram's Nova-2 also scored a 20.1. rev.ai scored a 21.6. Whisper scored 27.6. Google's Latest_long scored 29.8. Google Video scored a 30, and IBM scored a 43.4. IBM does not offer any punctuation, which is part of why they perform so poorly on the formatting measurements.

[00:19:19.00] So in the formatting space, Assembly is still a leader, and Whisper has also pulled forward by a rank once we're in the formatting space. Whisper was trained mostly on publicly available online video because it is an open-source model. And so most publicly available transcripts online are video captions, and we find that Whisper's formatting is very on-captiony, for lack of a better word. And that's part of why it does so well here on our use case.

[00:19:51.78] Speechmatics has come in third. Microsoft and Deepgram are tied for fourth. You can see that all of these error rates are pretty high. The highest-performing engine scored a 17, which is still an error one every six words.

[00:20:08.49] Since we believe formatting is important for readability and for meaning, we find this state of the art still fairly inadequate for captioning. We don't see it blowing us out of the water this year, especially compared to previous years. So next slide.

[00:20:30.78] So I want to talk about the pairs of engines that we tested, new and old engines. We had three examples of this, and in all three, the newer engine was not clearly better. So Assembly's Conformer-2 is an older model, and it performed better than their brand-new Universal-1 model.

[00:20:51.80] The Whisper large-v2 model is better than the new release of Whisper large-v3. And Google's Video model and their Latest_long model, Video model being the older one, are actually very close. But in some aspects, Google's Video model continues to outperform the Latest_long model.

[00:21:13.11] So all of these models were released very recently, and they built upon innovations that we saw last year in 2023. In 2023, Whisper succeeded really well with an unprecedented

amount of training data. And so this year, it seems like Assembly and Whisper have both tried to take that lesson further, training on even more data. Whisper trained on 5 million hours, and Universal-1 was trained on 12.5 million hours of data.

[00:21:43.53] The results of both of these seem a bit underwhelming, given that large size of data and the amount of investment that probably required on both of their parts. Whisper also has been open about the fact that a large part of that data-- about 80% of it-- is synthetic data. And synthetic data is data created by another machine learning model in order to use as training data. So to get the 5 million hours, they had to basically pull in synthetic data, which is known as to be lower-quality and can cause issues when it's used for training.

[00:22:21.84] Whisper and other recent models have been multilingual. So that's one model trained on many languages and able to work for all of them. There's been a suggestion that training on many languages has made the models work better on all languages, or at the very least, that training on multilingual data doesn't hurt any of the individual language performances.

[00:22:45.62] This result is an interesting data point because we see both Universal-1 and Whisper v3 were very, very multilingual this year. And actually, if you are interested in ASR for languages other than English, and particularly underrepresented languages, Whisper v3's purpose was actually to up performance on those underrepresented languages. And so it did succeed, according to their reports, at doing that. But it seems like that happened at the cost of the English performance.

[00:23:23.66] Google's new model uses a conformer architecture, and that architecture is the underlying architecture, also, for Assembly's models. They really pioneered it last year and showed up with incredible performance. So it seems like this is another attempt to build on innovations we saw last year that hasn't necessarily brought astounding success when we tried to reproduce it.

[00:23:48.41] Yeah, it's just all very interesting. It seems like we have entered a part of the cycle of innovation where we're trying to build on our previous successes, and it's really up to a contest in who can interpret those successes well and learn from them correctly.

[00:24:11.79] So now I'm going to get into different types of errors that we see in ASR, and I'm going to break down error types for a couple of engines. So we have three different types of errors in speech recognition. There's substitutions, which is when ASR replaces a word with an incorrect word. So it may mishear "3Play Media" as the word "encyclopedia."

[00:24:34.55] There are also insertions, where an engine will add an extra word. So those extra words could be background noise that it's picking up on. It could be inserting things that were just small sounds in speech that it's trying to create whole words out of. There's a lot of reasons for this. And then there are also deletions, where a engine emits a word entirely, transcribes fewer words than are actually present.

[00:25:07.50] So we can divide the errors that we saw into these categories to try to get a better understanding of the behavior of each model. Here, I've broken them down. I've actually

removed the deprecated models just because it was making the table very long, and you can't access those models anymore.

[00:25:24.57] So you see that Universal-1 has the lowest rate of insertions. It's not picking up extra stuff or over-transcribing certain things. It's not doing-- there just aren't many insertions. Speechmatics has the lowest rate of substitutions, as well as the lowest rate of deletions.

[00:25:45.85] Yeah, Whisper v3's bad performance is really on display here you can see Whisper v3 has a 10% insertion rate, which is far beyond the rates of any of the other engines here, including the ones that performed much worse than Whisper. It's very interesting. I can get into some ideas about that later in the presentation.

[00:26:10.52] I also broke down performance by transcript style. So approaches to captioning can span a spectrum of different styles. Some of them are more clean read. Some are more verbatim. Verbatim transcripts include every spoken word in the video. That includes false starts, disfluencies, and fillers. Clean read transcription is a style that omits those false starts, disfluencies, and filler for clarity and for ease of understanding.

[00:26:39.37] These are appropriate in different settings, and different users have different preferences for them. In general, verbatim may be used for a scripted context, where all of those fillers are intentional and should remain in there. And clean read is used for less scripted settings, most of the time.

[00:27:04.68] So-- sorry-- splitting up the data, you can see that some of the engines are transcribing in a more verbatim style, and some are more clean read. So we used whatever configurations were available to us. Most engines offer some level of configuration about whether they include things like disfluencies in the final transcript. But even then, they still tend to lean into a specific style based on what their training data looked like.

[00:27:32.81] Your preferred engine is going to vary based on what your use case is. Our verbatim content, I should note, has a higher word error rate in general, due to the industries that tend to order it having a more difficult content, overall. So pay more attention to the relative scores in the table between engines. As I said, Speechmatics is the most verbatim. Assembly is the most clean read.

[00:27:56.75] We happen to the verbatim style output because of the way our editing process works. It's usually easier to remove than add back in things, especially things like false starts that otherwise the mind may just skip over for an editor. But either one is appropriate for different use cases.

[00:28:21.43] So now it's time for a poll. Which industry do you think ASR performs the best on? We'll see if that comes up. I'll give everybody-- I'll read the options out. So the options are higher education, goods and services, media publishing, cinematic, e-learning and sports.

[00:28:46.58] So it looks like most people expected higher education to perform the best. People also-- or sorry, e-learning to perform the best. I misread that. People also seem to think that

media publishing and higher education will perform well, with less but still a good number of people thinking goods and services content will perform well. Cinematic got some fewer votes, and sports got very, very few votes for what ASR will perform the best on.

[00:29:21.38] So we can see the answer on the next slide. So it looks like a large portion of you were correct-- that e-learning actually is the best-performing industry, followed-- so these are averages that we have created from the top four engines for both word error rate and formatted error rate to give us a general idea of the performance of each industry.

[00:29:48.86] E-learning averaged 3.97% WER or 11.8% FER. Goods and services averaged a 5.05 word error rate and a 13.4 formatted error rate. News and networks has a 5.25 WER and a 14.9 FER. Publishing has a 6.12 WER and a 14.9 FER. Government has a 6.92 WER and a 16.4 FER.

[00:30:20.26] Higher ed has a 7.16 WER and a 15.2 FER. Our other category, which is very broad, has a 7.77 WER and a 16.2 FER, so very average for the top few engines. Associations has an 8 WER and a 16.8 FER.

[00:30:44.44] Tech has a 9.69 WER and a 19.8 FER. Sports has a 10.2 word error rate and a 20.2 formatted error rate. And cinematic, our worst-performing industry, has also a 10.2 average word error rate and a 21.2 average formatted error rate.

[00:31:06.09] So you can see in common-- in these industries that perform well that they tend to have a single speaker. They have very little crosstalk. So you'll see like, for an e-learning program, you'll have one speaker, probably scripted over either a talking head or over some sort of presentation.

[00:31:32.27] These all have minimal background audio. The high-performing industries are generally recording in professional environments. And they have scripted rather than spontaneous speech, which is a lot easier for speech recognition to handle.

[00:31:48.58] So you see things like sports and cinematic do poorly because they, one, have a lot of needs for very, very specific formatting needs. They also have a lot of research involved. You need to find the names of the players for a sports game. You also might need to look up, if it's a movie, the names of the characters. If it's a fantasy movie, you're going to have to figure out how you want to spell everything. There's a lot more work that goes into creating captions for those industries.

[00:32:26.65] The other notable thing is that sports contains a lot of crosstalk, where you'll hear announcers speaking over each other, speaking over a very loud crowd. And that makes it really difficult for ASR to perform well.

[00:32:40.51] One other note is that I didn't put individual scores in here, but I wanted to point out that Whisper did very poorly on cinematic content. It wasn't included in the top four that created this average. And we think that is partially, at least, because, as an open-source model, Whisper is trained on open and public data, and copyright law would prevent them from

accessing the majority of, say, movies and TV shows. And they just haven't seen that data, and it becomes very clear. So I think that's an interesting constraint we're seeing on open-source models right now.

[00:33:23.97] So finally, the one thing I want to get into is ASR hallucinations. I'm just going to read this quote, and then I will get into some examples and some data that we saw about ASR hallucinations. So this is about Whisper. Whisper's greatest flaw seems to be its tendency to sometimes hallucinate additional speech that doesn't appear in the original audio sample. Hallucinations sometimes look very credible if you aren't listening to the audio.

[00:33:52.34] They are sensible and on-topic, grammatically correct sentences. This would make viewing the captions as a deaf or hard-of-hearing user really confusing. If auto captions are nonsensical, it's usually clear to a user that they're making a mistake. But with these, you could easily assume that the mistakes are what's actually being said.

[00:34:11.45] Whisper's scores don't adequately penalize the hallucinations, in my opinion. Hallucinations show up as errors, but in an area where the text was completely invented, it may still get a score as low as a 50% error rate, rather than the deserved 100% because of common pronouns, function words, and punctuation lining up with the real text.

[00:34:37.70] So here's an example of this. It's quite short. But for to give context, this is from a broadcast where they were discussing the weather and then switched to discussing news. And Whisper did not handle the change in topic very well and continued to invent a weather forecast on its own that was completely made up, and it looked very plausible. So if you were watching this, you may think you need to bring an umbrella if you didn't need to.

[00:35:12.01] So what you see here-- just this small section-- the words "the," "of," "the," and "a" all appear in these texts because they're both completely plausible English sentences, and those are very common words. So Whisper would score as 50% accurate in this section, even though it's just making stuff up off the cuff.

[00:35:38.14] So besides that type of hallucination, we actually see a couple of different types. The most common is repeating a word tens to hundreds of times. The word "thank you" is a common one. You'll see "thank you" repeated sometimes 100 times in a transcript when it wasn't spoken or was only spoken once.

[00:35:57.43] It will also tend to reproduce sometimes very large chunks of the transcript that happened elsewhere in the video, sometimes minutes away. It will just plop those in the middle of a completely separate part of the video.

[00:36:13.52] It will also do what I just showed an example of, which is create plausible but completely made up speech. Sometimes this will be an addition to the actual spoken words. But it will also sometimes completely overwrite the actually spoken words. So you're not only getting extra content but losing the real spoken content.

[00:36:37.32] And then finally, what we have seen this year, as well, is that there are grammatically reasonable but, in content, pretty meaningless sentences. So you'll see something that has a subject, and a verb, and an object, but together, all of the words don't make much sense. One example that we saw was "We ate from Amsterdam to Amsterdam."

[00:37:01.10] So it has a-- grammatically correct sentence. It makes no sense. And a viewer may still be pretty confused by this because not only are they missing the actual speech that was supposed to be there, but it may look maybe one or two words were incorrect, when actually the entire thing was wrong.

[00:37:20.07] So here on the side, I have an example of a really common Whisper hallucination. Since it was trained on online video, it seems like it's consumed a lot of YouTube videos, and so it will sometimes add "Thanks for watching. Please subscribe to my channel" at the end of a video that was not a YouTube video and where there's no channel.

[00:37:46.64] So we decided to, at least heuristically, try to estimate how many hallucinations were in our data. So to find hallucinations, we found sequences of insertion or substitution errors which were at least four words long where the truth transcript didn't indicate that there were lyrics or another language spoken in the area. That can also cause behavior where you see extra words transcribed.

[00:38:16.46] We made sure that the substituted words shared very few letters or sounds with the truth transcript in the area, so we were certain they weren't just mishears. And then we also ran engines we were confident did not have any hallucinations and made sure that they did not have a similar number of errors in the area.

[00:38:36.83] After that, we ran other, nonhallucinating engines through and created an estimate of the false positives that we would get from this methodology, and we subtracted that number of false positives when we ran other engines through that we suspected had hallucinations. So first, we tested Assembly's Universal-1. Their announcement implied that they might have a small number of hallucinations because they explained that they had fewer hallucinations than Whisper v3.

[00:39:09.44] We actually didn't find evidence that they hallucinated at all in our data, which is really good for them. Whisper v2-- we think we actually way underestimated the prevalence of this problem last year. We estimated that Whisper v2 hallucinated on 20% of the files that we fed to it-- some very small hallucinations, some really bad. And we found that Whisper v3 hallucinated on a whopping 57% of files.

[00:39:43.82] So to get into some key findings from all of this, one of the big takeaways is that your use case really does matter. So no matter who you are, if you're considering using one of these engines, it's for your own use case. We have done investigation for what's best for ours, and we've tried to split out the data into a couple of different views to help you make an informed decision about what is best for yours.

[00:40:11.47] We've taken away from this year that more data doesn't guarantee success. So it was a really exciting result in 2023 that Whisper trained on such a huge volume of data, was able to do so well. But it looks like our attempts to recreate that don't always scale up linearly. And it also-- in the quest for more data, which is very expensive and time-consuming, researchers are being forced to turn to lower-quality and synthetic data, which creates more hallucinations, as well as just poorer-quality results, in general.

[00:40:47.60] We believe hallucinations pose a real problem. Whisper's hallucinations are a concern for accessibility, even though their accuracy puts them sometimes in second or third place in our measurements. We also think that they pose a concern for things like brand safety, but that's a whole other world.

[00:41:08.02] We believe that the source material still matters. So it's clear that good ASR results still depend on audio quality and content difficulty. We still see sports performing poorly, and we don't expect that to change anytime soon.

[00:41:23.02] It's also clear that ASR is still not quite good enough for the 99% accuracy required to provide an equal experience. ASR just isn't captions. Sometimes in a pinch, it's helpful, but it's not where it needs to be.

[00:41:45.03] ELISA LEWIS: Great. Thank you, Tessa. And thank you to everyone who's asking really great questions in the chat and Q&A. Definitely keep those coming. So I want to talk now about what this means for you. So now that everyone has a really good sense of the state of automatic speech recognition across these engines, let's talk about what this means, what this indicates for you.

[00:42:12.04] So as Tessa said, while technology is continuing to improve, there's still a significant leap to real accuracy, even with the best engines. So I'm going to talk through some of the causes for this, and then we'll dive into some examples of what this ends up looking like.

[00:42:31.79] So first of all, causes of ASR errors-- for word errors, you'll see-- yeah, word errors-- you'll see more of these when you have any of the following conditions. You have multiple speakers or overlapping speech-- the example that Tessa mentioned was in sports video, when there's a lot of talking over each other-- if there's background noise, poor audio quality, false starts, any acoustic errors, and function words, like "can" versus "can't," particularly when these are not enunciated.

[00:43:10.33] So any time that these conditions are introduced, the word error rate will increase. And with formatting errors, you're looking at things like speaker labels, punctuation, grammar, how numbers are represented-- so if you take, for example, a math class, these should be numerical, versus if you're talking in a different setting where the numbers maybe are written out, that's going to be a big difference in how the viewer is interpreting these and consuming the captions-- and then other elements, like "inaudible" tags.

[00:43:53.79] So punctuation and capitalization are really crucial to relaying the correct message, and incorrect punctuation can make it really difficult to comprehend a file. So for example, if

you're following along, and you want to know who is speaking, and they're lacking those speaker tags, it's going to be really difficult to find out when this speaker is changing and follow along what's going on.

[00:44:25.18] So for these reasons, it's really important to measure accuracy rates that include punctuation as a factor. And I have a little comic on the screen. It shows a graphic of an old woman, and she's saying, "What?" And at the top, it says, "Let's eat Grandma." And then the sentence under it says, "Let's eat, comma, Grandma." And at the bottom, it says, "Punctuation saves lives."

[00:44:54.62] And this is a silly example, but it's a really good portrayal of how one tiny, little comma that's seemingly small can actually change the meaning of a sentence entirely. So the top of the sentence, it's insinuating that they're going to eat Grandma for dinner. And the sentence underneath that doesn't have that comma is indicating that they're asking their grandmother to come eat with them.

[00:45:24.94] And then similarly, we have another example here. It's a picture of Rachael Ray and her dog on the cover of a magazine, and it says, "Rachael Ray finds inspiration in cooking her family and her dog." There are no commas, making the meaning entirely different.

[00:45:44.77] We assume, and we hope that Rachael Ray finds inspiration in all of the following things-- cooking, as well as her family, as well as her dog, not cooking her family and her dog. So again, it's a kind of silly example. But if you think about real-world use cases and individuals who are relying on captions for consuming content, it's really critical that the meaning that is intended is portrayed in those captions.

[00:46:19.48] Function words are another area where ASR typically fails. The example here that I have on the screen says, "I can't attend the meeting," versus "I can attend the meeting." This type of error occurs typically when there's either background noise or maybe a speaker de-emphasizes the second syllable of the key word "can't."

[00:46:47.14] It's really important to note that, again, even though it's a really seemingly small error or small mistake, the meaning is actually completely reversed. And this is a really good example of a type of error that a human is much less likely to make, especially a trained editor, because they can use context and understand from those context clues what is actually meant here.

[00:47:17.33] And they also can be a little bit more aware and more discerning of people who maybe, again, are dropping off and de-emphasizing that second syllable because they know that that's a common speech pattern. So these, again, are just some examples of how errors may show up in the real world and the impact that they have.

[00:47:45.15] And then complex vocabulary is another area that typically requires human expertise and knowledge-- things like names. Again, I know there was some conversation in the chat around sports content. But there are some really specific vocabulary and some really

specific names when we think about sports content. So this is another area where we see ASR not doing so well.

[00:48:17.91] I have an example on the screen here where we see the word-- or the name "Ehrhardt" transcribed as "air." We see "Bowen" transcribed as "bone," "Loyola" as "loyal," "precarity" as "precariously," and "precariat" as "prokaryote."

[00:48:40.02] So these all sound similar, but if you know that the full sentence here reads, "Picked up really well by Ehrhardt, quick pass in front. Bowen slaps it home-- Virginia 1, Loyola 0," it's clearly about a sporting event. And so we know that we're not talking about prokaryotes. So really important to understand, again, context here.

[00:49:09.20] And I want to point out that errors add up quickly. At 85% accuracy, this means that 1 in 7 words is incorrect. So again, it's really important. Even seemingly small errors add up very quickly.

[00:49:28.38] And I know that we touched on brand, and this is something that is important. Of course, the accessibility piece is top of mind for most of us, but it's also important to realize that your brand is at stake when relying on some of these ASR engines that can fail.

[00:49:53.40] So here's an example on the screen of-- it's a screenshot of a video of automatic captions on a JetBlue video. This ad surfaced right at the beginning of the pandemic, and the captions say, "Hello. My name is Joanna Geraghty, and I'm Jeff Blue's President."

[00:50:14.21] Again, a fairly small error-- one word is wrong, but it happens to be the name of the brand JetBlue that they called Jeff Blue's, and it's really meaningful from a brand perspective. So you want to make sure that important words are being transcribed correctly. And I will hand it back to Tessa to recap.

[00:50:41.86] TESSA KETTELBERGER: Thank you. So to recap-- do we have a next slide? AssemblyAI has really pulled ahead as a leader this year. Last year, it was tied with Speechmatics for first. This year, it seems to have pulled ahead.

[00:51:02.24] Speechmatics continues to have very high performance, especially for our use case. And while we see Whisper as very impressive, especially in the formatting space, its outputs contain hallucinations that are quite risky for use in the accessibility world.

[00:51:22.83] Last year was a huge year for ASR innovation. We saw new players performing in the top couple of engines. In 2024, further improvements have depended on our ability to learn and grow from these successes and interpret them in a way that tells us where we should go next. And it doesn't seem like we have succeeded at that yet as an industry.

[00:51:50.94] The best engines can achieve up to 93% percent accuracy, which is quite good. That is for non-specialized content with great audio quality. And after that, you're taking the quality. There's still a long way to go to replace humans.

[00:52:17.33] And if you would like to get a hold of our 2024 report, the QR code is currently on your screen. I'll also read a URL you can visit. There, you'll be asked to fill out a form, and you'll receive an email linking to you to the report when it's published. So the URL is go.3playmedia.com/rs-2024-asr. And now I think it's time for questions.

[00:52:46.17] JACLYN LAZZARI: All right, guys, thank you so much. We have just a few minutes for questions. So we had a bunch of questions come in from attendees. Thank you so much, everyone. Let's dive into a few. Someone asked, does the report notate the accuracy in regards to the accent, speech pattern, or speed of the speaker?

[00:53:11.13] TESSA KETTELBERGER: I can answer that. It does not. We don't have good tagging for things like the accent of the speaker. Actually, the rate of speech is an interesting one. We may have the information to dig deeper on that one already.

[00:53:28.68] But for things like accent, it requires us to probably bring in an outside provider to help us analyze our data and separate it out into groups of accents, groups of speakers that make sense to be grouped together. We just haven't had the availability to do that so far.

[00:53:51.71] JACLYN LAZZARI: Thanks, Tessa. And someone asked, is the Google Video model different from the captions that automatically show up on YouTube videos when they're uploaded?

[00:54:04.09] TESSA KETTELBERGER: So I believe, at least until the release of the Latest_long model, that what they were using was some form of the Google Enhanced Video model. But it's not totally transparent to us. I can see that the results are incredibly similar between the Google Video model and YouTube captions, but they aren't always 100% the same. So they may be doing other post-processing or using a slightly different version on YouTube.

[00:54:35.22] JACLYN LAZZARI: Thanks. And then maybe just one or two more questions. Maybe to round things out, how do you envision the future evolution of ASR technology in terms of accuracy and performance, and particularly in challenging environments or with diverse accents and languages?

[00:54:57.28] TESSA KETTELBERGER: I guess I should do that one. I think that the multilingual model that we saw earlier, while I called it into question a little bit when I was talking about the recent performance of Whisper, I think that all models are essentially going to have a variety of data in them, whether it's multiple languages or, on a smaller scale, multiple accents, multiple abilities, different ways of speaking in general.

[00:55:30.62] And so the future of that just has to be more training data. I don't think we have created a world-- a process yet where we can avoid having real, high-quality training data on speakers with diverse accents and abilities, if we want these models to perform well on people who speak in a variety of ways.

[00:55:57.44] You can see this in the performance of Whisper's models, where they have some languages that do quite poorly but that are in there that have a small amount of training data. And

the approach of bootstrapping it up with other languages, other similar languages, doesn't appear to have played out yet. I think that we just need to work on-- commit to working on that training data being there.

[00:56:27.93] JACLYN LAZZARI: Great. Thanks for your answers, Tessa. And thank you both for your presentation and to everyone in our audience for your engagement, and for asking questions, and just for being here.