ELISA LEWIS: So hello, everyone. Thank you for joining today's webinar, The State of Automatic Speech Recognition. My name is Elisa Lewis from 3Play Media, and I'll be presenting today alongside my colleague Tessa Kettelberger. So with that, let's get started.

As I said, my name is Elisa Lewis. I'm a content marketing manager at 3Play Media. I'm super passionate about web accessibility. And outside of work, I love dogs and I love all things crafting and DIY. And I'll let Tessa quickly introduce herself as well.

TESSA KETTELBERGER: Hi, I'm Tessa. I'm a research and development engineer here at 3Play, which basically means that I work on using machine learning and data science to improve our products and our processes. That has a lot to do with speech recognition and how we use that, which is why I'm really excited to be here.

I forgot to write it here, but my hobby lately has been houseplants and, now that it's getting a little warmer, hopefully gardening. And I'm here to talk about our research and the testing we've done as part of evaluating the state of ASR as a captioning tool.

ELISA LEWIS: Awesome. So quickly, to walk you through the agenda of what you can expect over the next hour, we'll start out with an introduction, talking about what the state of ASR, or Automatic Speech Recognition is, why we conduct this research, our focus on innovation, and then we'll walk you through the research and testing process. Of course, what you're mostly here for is sharing those research findings.

And then, we'll shift gears a little bit, and we'll provide some examples of ASR, kind of what this looks like in real life context. And then we will share some key takeaways for you and your business. And then, like I said, we'll move into questions and answers at the end.

So The State of ASR is a report that we've been publishing annually. And the goal is to provide the most up-to-date findings on the current state of automatic speech recognition, which I'll be referring to as ASR technologies. And we're looking at it as it specifically applies to

captioning and transcription. So in a few slides, we'll talk about why that distinction is important, but that's kind of the context, the lens, that we're going to be looking at it and talking about it from today.

So over the last year, we all saw a really big increase in video with shifting to a remote world. And in fact, between April 2019 and April 2020, the live streaming industry grew by 99%. So this is important because, with the increase in streaming video, we saw a really big increase in captioning and transcription as well. And of course, a lot of that has to do with automatic speech recognition and ASR captioning. So we want to continue to evaluate this technology, how it's improving, and, again, how it applies to captioning and, of course, our process at 3Play Media.

So I want to share why this is so important to us at 3Play. So I think, really, to further understand why we care about this research and why we're taking the time to conduct it and to share it is, we have to understand 3Play's relationship with innovation. So innovation has always been at the core of what we do. We have 11 patents on our processes. And we use machine learning and Artificial Intelligence, or AI, really have heavily in everything we do. We have historically combined these tools with humans to provide a really disruptive solution for video accessibility.

So looking at the research and the changing technology is really important for, again, our process, and also staying kind of true to our core and what we've built our video accessibility solutions on. So that being said, in order to stay on top of this, we do need to understand how machine learning and artificial intelligence are changing. So I will hand it off to Tessa for the next few slides to get into the details of the research and what we found.

TESSA KETTELBERGER: Thanks. So with any sort of evaluation that you see of automatic speech recognition technology, it really matters what data went into that analysis and the use case that they're expecting for the ASR technology that they're evaluating. So the testing and the results I'm

presenting today are focused on using ASR as a tool for captioning, specifically. And that has a really big impact on the meaning of the results and the quality of the scores that we're able to assign for accuracy of ASR.

So we're definitely aware that, out there, there are reports that, in certain situations, automatic speech recognition can reach near-human accuracy, or near 100% accuracy. And those reports are rigorous, and they're academic reports, but they don't necessarily pan out in the captioning space.

ASR has a lot of uses. You're probably really familiar with things like voice assistants like Siri or Alexa. Or if you call a helpline, and it's a robot, and you're talking to a machine, and it's transcribing the audio that you're giving to it, or for things like note dictation, a lot of people use ASR for that. It does really well on those cases, especially Siri and Alexa.

And the challenges you see for something like Siri and Alexa is it's a very different thing to evaluate. So for example, voice assistants-- you know that you are in a house or in an office, or it's on your phone and it's very close to your face. You can do a lot of controlling for the environment and the background noise. For captioning speech recognition, we really can't constrain the environment that we know the audio is being recorded in.

Voice assistants also have a really limited vocabulary. They perform really limited tasks. They know the answers to a limited set of questions. Captioning happens on general video content, which means you can really have any topic. People could be really saying anything. You can't constrain or prefer certain vocabulary in order to make things more accurate.

And as I mentioned before, since you're in a controlled environment, you can adjust it when you're using a voice assistant like Siri or Alexa. If you know that it's not working, as you try to talk to Siri, you can turn

off your music, or you can repeat yourself. Siri can ask for clarification.

And captioning is just obviously not like that. What we call our situation is the long-form transcription problem. That's anything that's more than a minute or two, we would consider the long-form transcription problem. It's a totally different data set, a totally different style of speech and structured speech.

And then, on top of that, humans are the consumers of these transcripts, not machines. So that means that things like formatting matter. Things like punctuation matter. Disfluencies matter, knowing when someone has corrected themselves matters, and, on top of that, all of it needs to be time-aligned, which means that certain types of mistakes and certain errors have greater consequences for captioning than you would see elsewhere. And we have to evaluate these ASR technologies with all of these things in mind.

Next slide. So we tested the six most relevant APIs available. One of them is Speechmatics, which we tested two available versions that they have. They have two APIs. And since that is our provider, we tested one of the versions of their API with the post-processing that we use in our process in our ASR services.

So after Speechmatics, we have IBM Watson, we have Microsoft, we have Google, we have Rev, which you may also know is Rev AI or Temi. And there is also VoiceGain.

So we tested with a lot of data here. That's 490 files. That's 65 hours of content, 670,000 words. We really worked hard to attempt to get a representative sample of 3Play content. The main way we did that was by sampling across different industries that we know our customers are coming from. And we attempted to keep diversity of other qualities as well. So we're trying to get a variety of video durations, a variety of speaker count, audio quality, speaking style, including scripted versus spontaneous.

We tried to make sure this was varied in addition to industry so that, if one of the major industries represented here tends to just produce one type of content, we don't want that type of content drastically overrepresented, and we tried to keep variety even within those industries.

So what we ended up with is a lot of education content. That's 28% of education. That is really representative of how much of stuff has gone online in the past year. Education's moved up a lot in this list since last year. Online video follows at 16%. 15% of the content is entertainment or media. 13% is other. A lot of it's hard to categorize-- eLearning, corporate, government, fitness, societies and associations, and faith societies as well.

So we tested two main metrics. We're testing word error rate and formatting error rate. I don't know if these are terms that everyone is familiar with. But based on the name, I'd love to do a poll to see what you think is important for your captioning solution.

There it is. Yes, so you can see, if the poll is popping up, you should be able to participate now.

Here, let me read out the poll. Which do you think should be taken into consideration when evaluating the accuracy of the captions, Word Error Rate, or WER, or Formatting Error Rate, FER, either, or both? So those are the four options listed-- word error rate, formatting error rate, neither of those, or both of them.

Great. So I see a lot of answers here. OK. So yeah, it looks like a good number of people have said that word error rate is important to them only. A small number of people have said that formatting error rate is what's important to them. Nobody has said neither, which makes sense. And most people, 77% of people, said both WER and FER, or word error rate and formatting error rate, are important to them. And I would say that I agree that both word error rate and formatting error rate are really important for evaluating your ability to use ASR as a

captioning tool.

So to talk a little bit about word error rate and formatting error rate, word error rate is what you will see in mostly academic papers. That's the standard term that they're using. And basically, what that means is, what percent of words did the ASR correctly recognize taking absolutely nothing else into account?

Here, at 3Play, it's something we absolutely need to look at as well as formatting error rate, which includes punctuation, capitalisation, number formatting, audio tags, and speaker changes. This is stuff that humans care about. It's stuff that has bearing on the meaning. And it's really important for us to analyze from a captioning perspective.

So to start by looking at word error rates for all of the APIs that we evaluated, all of these numbers you see here are percents measured per words in the reference or the truth or the correct transcript. So the first column you see here is error rate. In the first row, you can see that number 13.1. That means 13.1 errors were incorrect words per 100 words in the corrected transcript.

The second column is the percent correct. You'll notice that the first and second columns do not add up to 100. That's because you can have additional errors in the case that ASR is inserting words that are not there in the truth transcript. And that's a pretty common error.

So you can see the next three columns are substitution, insertion, and deletion errors. Those are the three types of errors that should add up to the error rate you see in the ERR column.

So 3PM-- we have them abbreviated here-- 3PM is 3Play Media. That's us. Speechmatics-- so 3PM is our post-processing applied to SMX, which is Speechmatics, the next row. Then Speechmatics Plus is the newer, updated available API from Speechmatics. IBM is IBM Watson. MIC is Microsoft. GOO is Google. REV is Rev, and VG is VoiceGain.

To be very transparent about this, we had a lot of difficulties testing the Google API, which we haven't had in the past. We don't know if we caught them. So SMX stands for Speechmatics.

So the Google API, we had some difficulties with this year that we haven't seen in the past. And we don't want to completely blame Google for that. But we ended up having to test their live API instead of their batch API, where you can feed the whole file in at once. That is a much more difficult problem. So their score suffered because of that. I would expect live results to perform about 20% worse than batch results.

So if you're someone who uses Google for this, and you think this looks a lot higher than what you expect, I think if we are able to retest with their batch API, I would expect them to perform much closer to Rev or VoiceGain in that 15 range, again, since live tends to do about 20% worse.

Formatting error rates are also something that's important to us. You can see that the front runners basically remain the same here. But these are much, much different numbers than you saw. We have not been able to break 80% correct in any of these, including the ones that are performing the most.

And you can really look at the difference between these numbers and the word error rate numbers. And you can see the gap between great speech recognition and great captions. Great speech recognition doesn't necessarily translate automatically to great captions if you're missing things like formatting.

So in the best case scenario here, you're seeing about 1/4 of the words, both on Speechmatics Plus and 3Play Media, are still underformatted somehow. So just take a minute to look. 24.9 and 24.7 were the best we could do. Those are not statistically significantly differentiated. And after that, you can see Microsoft and Rev doing pretty well here as well.

Can we jump to the next slide? Thank you.

So really, how good ASR can be is a difficult question to answer. The numbers that I've been given are calculated across tons of files. But obviously, not every file is going to hit that exact accuracy number. Some are going to hit above. Some are going to hit below. Consistency really matters for captioning, especially because we're in a variety of environments with a variety of speaker numbers and topics.

So you can see the actual consistency in these distributions. These are the two highest performing on the word error rate scale. I just pulled them out as sort of best-case scenario examples here. These are per file, so you can see these bars are the distribution of the word error rate for each file.

The first column, or the first bar you see, is the number of files that fell into that 0% to about 5% error rate. And it goes up from there. Keep in mind that the error rate goes up if we we're measuring formatting error rate in these graphs. So for accessibility purposes, you really, really want something in either the first bar-- even the second bar is going to be really difficult for someone to use.

In these first two bars, the files that are falling in these really high-performing areas are files with high audio quality. They're files without music. They're often scripted files with professional speakers. The speech is very clear. The content is not technical. And more often than not, the files falling into these high-performing bars are extremely short files with very low word counts. That's how you really end up over there is if there's not a lot of opportunity for error.

You can tell that they're short files because, if you're going by word count, the average is actually somewhere within that third bar. You can see it sort of skewed towards the front here. You can also see those long tails where WER can go even over 100%. You can see that, for Microsoft, we had a file get up to 105%.

And I think this is demonstrating that you're going to have consistency issues sometimes with ASR. Depending on the kind of content you have, it's something you need to consider even with great breakthroughs that we've been seeing. So if you're somewhere in that 20%, 30%, you can actually see quite a few files are over there. That's something that matters.

So just some key findings-- this report shows a lot of exciting advances. A lot of APIs performed a lot better than they did in previous years. I'd like to shout out Rev and Microsoft there. Most of that improvement is incremental. So it's coming from better data, it's coming from broader data, and from more efficient training methods.

So there's not huge, monumental, ASR technology breakthroughs that are driving these. I'm not going to sit around hoping that, in the next year, we get this huge breakthrough and we jump up to 100% accuracy. That's just not the kind of innovation that's happening in ASR right now.

The audio quality and the type of content you have really impact the quality that you get out of a file. And on the worse end of that, especially, it's just not usable really at all.

We're confident that we're using state-of-the-art technology. We have a really vested interest in making sure that the research we do here is accurate and objective because, if we find out that the technology that we're using is not the best available technology, it really hurts our process, and having the best one really helps us.

So we used these results to confirm that putting in the work to move to Speechmatics's new API, which is what you saw as Speechmatics Plus or SMX Plus there was a priority for us. And we've now done that, so that's available. And we're using that with our mappings technology, which is the jump you saw between Speechmatics and the 3PM labels. We're able to get that jump with basically any of these technologies.

The other thing that we're really getting out of this is that, if we're measuring and formatting error rate, no one is providing an output that's close to sufficient. And this isn't surprising. It's not surprising that ASR isn't giving us audio tagging, diarization. Even punctuation is a really difficult problem within ASR. And they're very specific to the captioning use case. So ASR, it's not captioning, and this isn't super surprising.

I'll let Elisa elaborate.

ELISA LEWIS: Yes. Thanks, Tessa. So a couple of people have asked in the chat as well, kind of looking for more information on word error rate and formatting error rate. And what we're going to dive into next is hopefully going to be helpful for some people who are maybe overwhelmed by the data and numbers or want some more clarification. So we're going to be showing some examples and talk about what this research, what our findings, mean for you.

So what this really does mean is that, while technology is improving, and it continues to do so, there is still a significant leap to real accuracy from even the best speech recognition engine. So humans are really still a critical piece of the captioning process.

[VIDEO PLAYBACK]

- One of the most challenging aspects of choosing a career is simply determining--

[END PLAYBACK]

ELISA LEWIS: Sorry about that. So I want to show an example of a video that was captioned with automatic speech recognition. And I would love to get everyone to participate again and type in the chat window some of the errors that you notice when you're listening and watching this video. So I'll play a couple seconds here.

[VIDEO PLAYBACK]

--determining where our interests lie. And one common characteristic we saw in a majority of people we interviewed was a powerful connection with a childhood interest.

- For me, part of the reason why I work here is, when I was five years old growing up in Boston, I went to the New England Aquarium. And I picked up a horseshoe crab, and I touched a horseshoe crab. And I still remember that. And I still-- I love those types of engaging experiences that really register with you and stick with you.

- As a child, my grandfather was a forester. And my childhood playground was 3,600 acres of trees and wildlife that he had introduced there. So my entire childhood was around wildlife and wild places.

[END PLAYBACK]

ELISA LEWIS: Great. So a lot of you are noticing many errors here, including run-on sentences, no identification of new speakers, missing or incorrect punctuation. Yep, a lot of incorrect words that maybe sound like they make sense, but they don't, like four story instead of forester, new wing of the Quran instead of New England Aquarium. Someone said, yikes. I totally agree. Great. Yeah, these are really great observations that everyone made.

So as everyone pointed out, there are certainly a lot of incorrect

words in this clip. So another thing we notice is that hesitation words are not removed from the transcript, so they spill over into other words and cause a lot of additional inaccuracies. People pointed out the lack of punctuation. This is really difficult for people to follow. There's no way to know who's speaking because of the lack of speaker IDs and lack of notation for speaker changes.

And this is a really good example of, of course, there are word errors here, but there are also formatting errors. And so you can get a bit of an idea of why we consider things like formatting errors or that lack of punctuation, lack of formatting, why that is so critical to captioning accuracy. And we'll look into this a little bit further as well.

So to continue the conversation on these errors that we just saw, there are a lot of really common ASR errors, many of which were in this clip. And many happen all the time. So just to go through the list, we saw a lack of speaker labels. We saw punctuation and grammar errors. We oftentimes see relevant non-speech elements left out.

So for example, if you're watching a video, and maybe there's a car honking, trying to signify something to the audience, if that's not included, that can really change the context. No Inaudible tags for words that are missed or can't be understood.

And then the two that I really want to point out are acoustic errors and function words. They're a little bit less self-explanatory. So acoustic errors are when the audio maybe sounds OK to the ear, but, linguistically, it doesn't make sense. So that's that example of four story versus forester. And a human transcriptionist wouldn't make these types of mistakes because they can use logic and reasoning, and they can rely on context. Obviously, the ASR technology isn't able to consider those things.

And then the, other piece that I want to mention is function words. So for example, words like can versus can't. These are in the area where ASR typically fails. It can happen if a speaker deemphasizes the

second syllable. And although that 't is maybe a small mistake, it actually has pretty big implications because it completely reverses the meaning of the sentence.

This is also really rare for a human to make because, again, they can rely on context to determine the correct meaning. And even in cases of background noise where maybe they don't hear it, they can tell by the other words and by the context.

And then, we talked a lot about formatting error rate. And I have kind of a silly example here, but it really does speak to the importance of punctuation. So we saw this in the example that I just played that punctuation makes it really hard to follow what's going on and know who's speaking.

And then I have an example on the screen that says, let's eat grandma. And then underneath, it says let's eat, comma, grandma. And the meme says, "Punctuation saves lives." indicating that we can either eat grandma as a meal or eat with grandma. Yeah, someone else says, "eats shoots and leaves," another great example of why commas and punctuation are important for relaying the meaning.

And then, I also wanted to mention a couple more causes of ASR errors. I have a list here that says multiple speakers or overlapping speech, background noise, poor audio quality, false starts or hesitations, and then complex vocabulary. And I have another example on the screen that I'll quickly read out side by side.

So on the one hand, we have what the audio actually says. It says, "picked up really well by Ehrhardt. Quick pass in front. Bowen slaps it home-- Virginia 1, Loyola nothing. And then, the ASR said, picked up really well by air. Quick passing front bone slaps at home, Virginia 1, loyal, nothing-- really doesn't make any sense, a lot of missing punctuation-- just really doesn't convey the right meaning at all.

And then, the second example is continuing on. "This week, you will focus on identifying who, primarily, experiences precarity, who makes

up the growing precariat." And the ASR says, this week, period. You will focus on identifying who primarily experiences, precariously, who makes up the growing prokaryotes-- so really kind of a big jumble there.

And I think another thing to point out about these examples is that, with complex vocabulary, it's possible that the ASR can pick up on some of these. But really, it's often necessary to have a human with expertise or knowledge in this area, the subject area.

So I do want to look at last year versus this year and what we have found, what changes we've noticed. So while we can't compare the exact word error rate year over year because, of course, we use different files, we use different data, we have been able to identify overall patterns and several noteworthy findings around the overall state of automatic speech recognition in 2020 and as it compared to 2019.

So a couple of these noteworthy takeaways-- the McNemar Test that we used confirmed that Speechmatics Plus, that new version, that new API, performed the best compared with the other speech engines. And it had an accuracy rate of 90.3%.

We also measured the combined accuracy of Speechmatics V1, which you saw as SMX, paired with our 3Play Media post-processing, which Tessa mentioned. And we do this to better reflect 3Play's process. And then, again, we saw that Speechmatics showed a 7% reduction in word error rate from their Version 1, indicated as SMX, to their Version 2, SMX+, which, again, led us to move to Speechmatics Plus for our own process.

And then, before we get into questions and what's next, I want to leave everyone with a few key takeaways. So we talked about this throughout the presentation. But the application of captioning is really unique in regard to artificial intelligence, AI, and automatic speech recognition, or ASR.

We've seen several improvements in technology and training capabilities in the last year. The best ASR systems can achieve accuracy rates into the high 80s and in the low 90s if all conditions align perfectly in their favor. So again-- one single speaker, clear audio, no hesitation in words or false starts, no background noise. And even with accuracy rates of 80% to 90%, this isn't really sufficient for captioning. And we saw that in a lot of the errors. It's really important in the captioning space that we're getting 99% or more accuracy.

And then, again, when it comes to formatting error rate, or FER, which is a critical measure for captioning, none of the solutions that we tested are sufficient alone. There will really need to be some fundamental advances in machine learning and in language processing in order to replicate that intelligence that professional editors have and that context piece that we talked about that's really important in some of those other types of errors.

So before we do dive into questions, I want to share a few different things that are coming up. We have the State of Captioning Report coming out in the next few weeks. We also have another webinar coming up March 18, next week, The Global Outlook for Accessibility Compliance.

And I also wanted to share some exciting news that 3Play Media has a podcast launching in the next month, probably mid-April. It's called *Allied,* and we're really excited. So I will share some links in the chat window for everyone to take a look at both the podcast and next week's webinar. So I will share those now.

And then, I also do want to encourage everyone, as we move into Q&A, to keep the questions coming. We're going to take a look at what's come in so far. And yeah, keep the great questions coming in.

TESSA KETTELBERGER: So I can answer some of the ones we've already gotten.

ELISA LEWIS: Yeah, go ahead. If you want to just read the question out loud and

then go ahead and answer.

TESSA KETTELBERGER: So someone asked, how do these results relate to our live captioning solution, CART or real time? And that's a great question. So we tested batch transcription on all of these APIs with the exception of Google, as I think I mentioned earlier, who, we used their live captioning solution.

So live captioning, which is what's used also as a part of CART, or real-time captioning, is a more difficult problem because, as the ASR receives speech, it doesn't have the context of future speech, which actually helps it a lot if you give it an entire file to work off of. And as a result, live captioning tends to do about 20% worse. Just in general, for any of these engines, I would expect them to work about 20% worse in the live captioning problem than they do batch transcription, which is what we tested.

Someone asked, did the speakers have any speech impediments, cerebral palsy, deaf speech, et cetera? This is a good question. These are things that can really impact the accuracy of ASR because training data is not often considered with this type of speech. And as a result, a lot of engines aren't necessarily trained to recognize this type of speech.

We used all real-customer data for this. We know that some of our customers have speakers who do things like deaf speech or who have speech imediments. Or I think someone asked about accents earlier. We have lots of people with different accents in our customer base.

That said, I did not go through and watch all 500 hours to identify what percentage of the content had this type of speech in it. And I am willing to bet that number was not very high. So we didn't test that specifically, although, we know it was at least a small part of our training data and our testing data.

Relatedly, would accent errors fall under word error rate? Unfortunately, ASR seems to work best with English speakers only.

Yeah, I would agree that ASR seems to work best with people with standard American English accents and British English accents. We have, again, a lot of the customer content that has a variety of accents in it.

So we tested all kinds of speech, although, probably primarily American English speech. Any error that causes an incorrect word, no matter the reason that that word was incorrect, would be a word error. So that would fall under the word error rate.

Is there progress in captioning other languages? So I don't know off the top of my head what the APIs we tested, what languages they each have available. I know Speechmatics, because that's what we work with, has ASR available in many, many languages besides English. I'm not positive about the rest of them. I think IBM has several available, but they didn't perform terribly well on this.

ELISA LEWIS: Well, I can jump in on a couple of other questions here. We have a question about formatting error rates and if they factor into legal standards. So the legal requirements around captions don't specify accuracy standards. They often use language like providing equal access. So based on the examples that you saw and the research that we found, you can tell that formatting errors really do have a big impact on the accuracy of captions and in being able to provide that equal experience.

Another question that we have is around ASR, and if that's all that you have, either for budget reasons or buy-in, if we have any tips to make the most of it. I would definitely say getting the best audio possible. So trying things like writing a script ahead of time so that there aren't hesitation words or restarts, making sure that there is a really solid microphone, and strong internet connection if it's something that's going to be over the internet, making sure that there's a single speaker, not interrupting, having multiple people talking at the same time, and things like that to really make the most of the ASR.

Another question that we have-- Tessa, this one's probably for you, if you are able to answer it. Have you looked at Otter.ai?

TESSA KETTELBERGER: So we haven't looked at Otter. They're just difficult for us to test because of the way you have to use it. We can't just integrate easily with an API the way we were for most of these. So Otter is, I believe, what is built into Zoom. Am I correct about that, Elisa, do you know?

ELISA LEWIS: I believe it's in Zoom at least partially. I'm not sure if they just released something that's additional.

TESSA KETTELBERGER: So at the time we did this testing, the easiest way for us to test Otter was to take all of the 500 files that we had, and put them into-- like, share-screen and audio into a Zoom call, and have Otter transcribed through there and download the transcript. It wasn't a very practical way to do things, although, I am not sure if they've released a way to integrate with them more easily since so we could think about testing it next year.

We did look at them-- we tested a couple of files this way just to see how compelling it looked. And we did observe that they were reasonable, although not perfect, like the rest. But we can't give very good numbers on them.

ELISA LEWIS: Someone else is asking, if you're using ASR, it sounds like it's important for a person to do quality assurance to correct the errors, especially formatting errors. That's correct. At 3Play, that's why our process uses both ASR from Speechmatics and two rounds of human editing. It's really important to go through and have a human look at those different pieces to really check for that quality assurance, check for some of those errors that are really common, like function words and acoustic errors, lack of punctuation, and things like that. So yes, it's definitely important to have, if you're talking about captions and making an accurate caption file, it's important to have a human go through and look at it as well.

TESSA KETTELBERGER: I think somebody asked what YouTube uses and how YouTube would score here. YouTube is owned by Google and uses the Google speech recognition. Like I said, the numbers we had up there for Google were more accurate, probably, too, if you're using their live speech recognition. And the stuff that they offer on YouTube videos, I believe, is batch. So it's probably more accurate than the number we showed you there.

We actually tested their batch transcript and then uploaded the same videos to YouTube. And we went and looked at the captions to see if they were the same. And it seems that they have their settings configured differently for YouTube videos, maybe to optimize for YouTube. So it's hard to give an exact number there. But I would probably put it in the range of what you saw for Rev or VoiceGain, around that 15%.

ELISA LEWIS: Great. I'm just looking at a couple more questions coming through now. Someone is asking if you recommend speaker self-identifying when speaker changes. I think that's a really interesting question. It can certainly help, particularly if you haven't had an opportunity to kind of provide that information in some sort of glossary or word list or something like that, and the ASR isn't already trained for that, it can certainly help.

I think also, if there are multiple speakers, if there's going to be a panel with many speakers, it can definitely help to kind of self-identify as well.

TESSA KETTELBERGER: Something that self-identification would help with as well is forcing speakers to allow other speakers to finish. It's much easier to caption accurately if there are no overlapping voices. So that's also something that I'd keep in mind.

Somebody asked why AWS Transcribe wasn't included, and that's also a great question. AWS Transcribe, or Amazon speech recognition technology is something that we have tested internally. But they

require, as part of their terms for use, that, if you publish accuracy results about their technologies, you need to provide all of the data publicly that you used to test that, which makes sense. They want to be able to corroborate the work that you did.

But since we're using real customer data, it would really limit the amount of representative sample that we're able to grab because not all of our data is publicly publishable. So this allowed us, probably, to get a more robust report.

Yeah. I can't give you an exact number on AWS. But they have not traditionally performed as well as the top performers. They've usually been about middle of the pack. We saw that again in our smaller tests that we've done internally more recently.

ELISA LEWIS: Someone else is mentioning about racial or ethnic bias in ASR. I think this is really an interesting topic. We did talk a little bit about accents and how ASR can struggle with non-native English speakers and things like that. Certainly something to look more into in future research as well.

Cool. I think that is pretty much-- we've covered a lot. I know a lot of questions came in about Otter AI and about Zoom, which we talked about. So I think that, unless anyone has additional questions, we can wrap it up.

Someone is asking what captioning we're using today. We're using live captioning from A La CARTe. So thank you to all A La CARTe for captioning this webinar. We appreciate it.

Great. And with that, I just want to thank everyone, again, for joining the webinar today. As a reminder, you will receive an email with a link to the recording and slide deck that will be captioned as well. And thank you for all the great participation and the questions. And we hope that everyone has a great rest of the afternoon.