REBECCA KLEIN: Hi, everyone, and welcome to our presentation on the State of Automatic Speech Recognition. My name is Rebecca. And I'm a content marketing specialist at 3Play.

I'll be going over some of our high-level findings today and what the takeaways mean for you and your business. And just so you're all aware, I stutter. So when you hear those pauses, that's all that's happening. I'm joined today by my colleague, Tessa.

TESSA KETTELBERGER: Hi, I'm Tessa Kettelberger. I do data science here at 3Play Media. And I did the data selection and research to create this year's State of ASR Report.

REBECCA KLEIN: Great. Here's our agenda for today. I'll give an introduction on our annual ASR Report and why we conduct this study. And then I'll talk about 3Play's focus on innovation.

After, Tessa will dive into the research and testing process and the results. And I'll then go over what our findings mean for you and your business, some ASR examples, and key takeaways and conclusions. And then at the end, we'll have time for a Q&A.

The State of ASR is a report that we publish 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.

And we're looking specifically at ASR as it applies to captioning and transcription. And in a few slides, Tessa will talk about why this distinction is important. So over the last few years, we've seen a huge increase in video with the world shifting to virtual and hybrid environments due to the pandemic.

With the increase in streaming, we also saw huge growth in captioning and transcription. And a lot of this growth has to do with ASR captioning. And so with our report, we want to continue to evaluate this technology, how it's improving, and how it applies to captioning in our own process at 3Play. And our findings help us to improve our own captioning and transcription processes.

And now I want to share why this research is so important to us at 3Play. 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 heavily in everything we do.

Historically, we've combined these tools with human editors to provide a disruptive solution for video and media accessibility. And so looking at the research and the changing technology is so important for our own process and ensuring that we're using the best possible technologies for our services.

And to stay on top of technology, we have to understand how technologies are changing. And so now I'll hand it off to Tessa for the next few slides. And she'll get into the details of our research and results.

TESSA KETTELBERGER:

Thanks, Rebecca. So yeah, as Rebecca mentioned, we do this study for internal purposes, as well as for public consumption. We do this because we want to know that reason, the best technology. It saves us time, money. It makes our final product better.

And in that regard, we're focused on the use case of ASR for captioning or to assist in the creation of captions. So there are other ASR use cases. You'll see speech recognition used for voice assistants, like Alexa or Mycroft or Google Home. You might see ASR used for personal dictation. Many people use Dragon Naturally Speaking.

For our use case, we need to explore the accuracy of systems presented with a variety of audio environments, including background noise, multiple speakers. We also need to test on a variety of subject domains. We need to test the long-form transcription task, so not sort of short, little spoken sections like you might say to something like Siri or Alexa, but actually long post-production files or live events. And then we need to consider the readability and the human usability of these texts when we're using them as captions instead of using them for any of the other possible use cases that ASR has.

So when we do this testing, we know that there's going to be content with background noise, overlapping speakers. We aren't using technology that can adapt to a single user's voice. There are a lot of technologies that can do that. But that's not applicable to the captioning scenario. There are always a variety of novel speakers.

Captions and accessibility are agnostic to the subject as well. You're going to have to make all kinds of content accessible, as opposed to something like a voice assistant which might have the opportunity to optimize for frequent commands that it's going to hear often and which don't need to have a very large vocabulary.

Crucially, a lot of ASR is used for short-form transcription. And these are interactive technologies some of the time that the user has the ability to correct and clarify errors. As they interact with the technology, they can go back and type over something that was misheard or they can try speaking again.

That's not the case with live captioning or with ASR for captioning. There is no such opportunity. And since captions are meant for human consumption, they're going to be readable. They need to be formatted. They need to be punctuated. They need to be accurately time aligned.

And so the results we are going to present are accurate for this use case. So in other words, we acknowledge that maybe the best dictation tool and the best post-production captioning tool might be different. And we've tailored our methods and the data that we use to test to teach us about the captioning use case in particular.

I can show more, yep. So we did two types of testing. First is batch engine testing. That's asynchronous, post-production ASR-captioned generation where we would upload or request to see transcribed an entire file at once. This is 501 files or 68 hours of content. We had a total of 600,000 words in our sample.

All of this content that we tested with is pulled from our real system to make this test as realistic as possible. So we strategically sampled from our content uploaded by real customers to avoid bias towards our extremely high-volume customers and industries. We intentionally tried to select from a variety of industries and a variety of customers with different upload volumes.

The intent was to make the sample representative across different industries, file lengths, scripted versus unscripted content, multiple accents, many topics, et cetera. So we have 28% of the sample coming from the education space, 20% coming from general online video, 17% coming from the entertainment industry. 14% is corporate internal content. 12% is e-learning. About 2% comes from other sort of unidentified customers, but those include government customers, publishing. And then, finally, we have 1% coming from society, and 1% coming from e-commerce customers.

We tested eight different engines. We tested Speechmatics. These eight engines are all speech recognition vendors. We tested Speechmatics, which I will abbreviate sometimes as SMX. We additionally tested Speechmatics with 3Play Media's post-processing proprietary technology, just because Speechmatics is the vendor that we're currently using and we have the ability to do that.

We tested Microsoft. We tested two different Google engines-- both their standard engine and their more expensive enhanced video model, which you'll see written as VM-- VoiceGain. We also tested Rev, Deepgram, and IBM.

The second kind of testing we did, and which I'll present, is that for the first time that we've done this study, we tested realtime ASR engines as well. Realtime ASR is the kind that you will see in a live event where it might feed small chunks of the video at a time so that it can get back text from ASR with very low latency and display that to a user. This is live automatic captioning that we would be testing here. So this is, by its nature, less accurate. It doesn't have the full context of the entire audio.

And we only tested two engines here. So we tested Speechmatics and we also tested Otter.ai, which are the default Zoom captions if you have a Zoom business account that you'll see sometimes. So this was a smaller data set. It had 10 hours of content, or 83,000 words.

And we selected only from government, corporate, education, and e-learning accounts. This is because these are the types of accounts that we typically get requests for live captions from. And a lot of other content is difficult or impractical. There's not a lot of desire for it on something like entertainment. So just as in the last sample, we saw a variety of different durations, number of speakers, audio qualities, and speaking styles in these files.

We tested two metrics. We tested both word error rate and formatted error rate. You'll see these written-- Word Error Rate as W-E-R. I'll sometimes call it WER. And Formatted Error Rate, or F-E-R, is sometimes called FER.

These are both ratios. They're ratios of mistakes made in the ASR to the actual spoken words in the content. So the output of these metrics will look like percentages-- basically, the number of mistakes made per 100 words. So if there's 10 mistakes per 100 words, you'll see about a 10% word error rate.

Word error rate is sort of academic. It only considers which words are correct. So most reports you see, even in the captioning industry, use this. It's basically the standard, but it only takes into account the individual words.

Formatted error rate is more focused on usability and user experience. And it's more specific to the captioning use case. So formatted error rate is also taking into account incorrect words, but it will also take into account errors in punctuation, capitalization, number formatting, speaker labels, tagged music or audio, all the stuff that is required to make a text human readable and accessible.

We will show scores for both WER and FER for all of the engines that we tested. I'll go over live first, and then we can discuss the batch results.

So the realtime or live WER error rates we got from our tests, I will just go over the five measurements that you see here for each engine. You can see ERR, or E-R-R-- this is the number of errors per 100 words. CORR, or C-O-R-R, is the correct words per 100 spoken words.

And then the next three are breakdowns of the different error types that make up that error rate. The first is substitution errors where a word is misheard as a different word. The second one is insertion errors where an extra word is added which wasn't spoken in the real text, and then deletion errors where ASR missed or omitted a word entirely.

So here, Speechmatics won out against Otter. Otter actually did OK, but it inserted a lot of extra words which can be misleading to viewers of live captions. So you can see their insertion rate is quite a bit higher-- yeah, 11.3% error rate for Speechmatics and Otter at 12.3% error rate.

If we move to formatted error rates, which will take into account punctuation, capitalization, et cetera, Speechmatics had a formatted error rate of 20.1% and Otter had an error rate of 24.1%. So here you can see that the margin by which Speechmatics came ahead was much higher, basically indicating that Speechmatics is punctuation and formatting their numbers probably at a higher rate and definitely more accurately than Otter is.

Yeah, next, we can go over batch. So a lot bigger of a graph here for the batch results. We tested eight engines. We've listed them in order if you're able to see the screen. But I'll read out the error rates for each one.

So 3Play, which is Speechmatics with 3Play's proprietary post-processing applied, had a word error rate of 7.96, followed by Speechmatics without 3Play's post-processing technology at 8.67%. Microsoft is at 10.6% error rate, in third, followed by the Google Enhanced Video Model at 12.8%. VoiceGain had a 13.1% error rate, Rev, 13.8, Deepgram, 17.5, IBM, 23.3, and then Google, the standard model at 26.1.

Once you take formatting into account, the formatting error rates, 3Play-- the Speechmatics with 3Play post-processing-- scored 17.2% formatted error rate, Speechmatics at 17.9%, Microsoft at 22.4%, Rev at 24.9%-- so pulling ahead of Google and VoiceGain when it comes to punctuation and formatting, which is interesting.

VoiceGain was at 26%, the Google Video Model at 27%, Deepgram at 29.4%, the standard Google model at 38.2%, and IBM at 38.6%. It's worth noting-- IBM does not put punctuation in any of their ASR. So they typically are at the bottom of the formatted error rate scores, just because that's not something they offer.

So this year, we also tested across multiple industries. We separated these out into education, online video, entertainment, corporate, e-learning, and then grouped the rest together that weren't statistically significant on their own. So that would include government, publishing, faith content, fitness content, and e-commerce.

We are going to put up a poll. And I will again read out the options. But which industry do you think performed best? The choices again are education, online video, entertainment, corporate, e-learning, or other more general category?

OK, so it looks like most people voted for the entertainment industry to perform the best, which I think is really interesting, followed by education and e-learning-- makes sense that those are similarly voted for. They're similar industries-- corporate, online video, and other.

So I can show you the real results if we go to the next slide. So these are our top three performing vendors. The ranking of the different industries falls the same no matter basically which vendor you go through. So I just put the top three here for readability. But the best performing industry we saw was online video, with our best score coming in at 6.18% word error rate. The worst performing industry was the entertainment industry at 11.5%.

So I thought it was very interesting that the entertainment industry, many people voted for it to perform the best. So entertainment tends to be a more difficult industry to transcribe. Even though it's scripted often, it has a lot of overlapping speech. There's a lot more music and sound effects that ASR is just never going to get.

So in general, I think it may have more complicated requirements for captions and sometimes more difficult audio situations than some of these other still professionally produced video situations. Other did quite well, e-learning did quite well. But you can see a pretty steep drop on the entertainment industry.

So some of the key findings, from the at least data science perspective here, is that we're seeing a lot of vendors improve from our previous studies from past years that we've done this. All vendors have improved from our last test. And that doesn't always happen.

From a tech perspective, that's really exciting to see. If you're someone who uses ASR in your workflow, that's always great to see. Despite the great improvements we're seeing here, ASR services still just generate text basically. So they're not captions, it's text. And formatted error rate, in particular, it's just still we're not seeing it rise to like an incredibly impressive level. And I don't really expect that to change unless the ASR industry changes to really focus on use cases like this.

Most progress from any of the vendors you see here is coming from innovations in their ability to utilize new data, to train on larger volumes of data, to work more efficiently with the data that they have. And the language models themselves are not changing hugely. The general approach to ASR is not what's gaining us big wins here.

So from that angle, the fact that we got this much improvement is really exciting. But again, I don't expect to see something like FER or something captioning-specific to jump up to incredibly user-facing usability levels anytime soon. And with that, I think I'll pass it back to Rebecca.

REBECCA KLEIN: Great. Thank you, Tessa. Now that everyone has an understanding of our testing process and results, I'm going to cover what our findings mean for you.

Now the biggest takeaway that we want to highlight is that, while technology continues to improve year over year, there is still a significant leap to real accuracy from even the best speech recognition engines. This means that humans are still a crucial part of creating accurate and legally compliant captions, and that we can't rely on ASR alone for accuracy and accessibility.

Now, you might be wondering if ASR is ever good enough on its own. And to this question, I'll say no, it's not usually. ASR is not accurate enough to comply with legal standards. And it's not accurate enough to provide an equitable viewing experience for people who are deaf or hard of hearing.

Now the one use case that comes to mind for ASR on its own is if you're using an ASR transcript for an internal process, such as editing a podcast episode, which is a very specific use case, and means that you're using ASR transcription internally and not for your audience.

So in this context, you may be able to use an ASR-generated transcript for editing, provided that your audio quality is extremely good. However, if you want to publish this transcript for your audience, which podcasters should be doing, then the ASR transcript is no longer accurate enough. And so human editors are essential.

So now the question emerges as to why ASR on its own is not enough for accuracy and what this looks like in practice. So I'm going to cover some examples of common ASR errors and how we take these errors into account when calculating accuracy rates. So I am going to play an example here of an ASR-generated transcript. If you're able to see the screen, please read along as the video plays.

And I'm hoping that everyone is up to participating a bit. So please type in the chat any errors that you notice. And for those who were unable to read the screen, I'll share out in a moment some of the errors that occurred.

[VIDEO PLAYBACK]

- One of the most challenging aspects of choosing a career is simply determining where our interests lie. Now one common characteristic we saw in the majority of people we interviewed was a powerful connection with a childhood interest.

[MUSIC PLAYING]

- 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 the horseshoe crab. And I still remember that and 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 in wild places. It just clicked.

- When I was a kid, all the cousins would use my grandparents' drive--

[END PLAYBACK]

REBECCA KLEIN: OK, so some of the errors in that ASR transcripts were probably pretty evident to many of you. There was no punctuation, no indication of speaker changes, no indication of music. And often words got mixed up. So for example, "forester" became "four story."

And so these are errors that can completely change the meaning of a sentence and make it really difficult for someone to follow along. They have a huge impact on accuracy and comprehension. And that's why human editors are essential.

So on this slide, I have a list of common causes of ASR errors. When we're looking at word errors, which Tessa touched on earlier, some common causes include having multiple speakers or overlapping speech, background noise, poor audio quality, false starts, acoustic errors, and function words which, for example, are "can" versus "can't."

So a human will be able to use context clues to get function words correct. But ASR will often mix up these words if there's background noise or if the audio quality isn't great. And these mistakes can end up completely changing the meaning of a sentence.

And as far as formatting errors go, you have things like speaker labels, punctuation, grammar, numbers, relevant non-speech elements, such as a door slamming or a bee buzzing, and no inaudible tags to indicate that speech is not meant to be understood.

I think that for many people it's perhaps easier to understand the importance of word errors and maybe harder to grasp the impact of formatting errors. And of course, one of the best ways to understand this is with another example.

And so on this slide, I have a silly example that hits home why punctuation, a common formatting error, is so important. And there's an image on the screen that says, "let's eat grandma." And then there's a cartoon of an old woman saying, "what?" And then below that is the sentence, "let's eat, comma, grandma." And it says, "punctuation saves lives," because the comma means that grandma doesn't get eaten.

And so this example is silly, but it hits home that the comma really changes the meaning of the sentence. And if we think about this in other contexts, maybe students in a classroom using captions to learn or someone watching a presentation, captions are essential for understanding the content and for creating an equitable experience. Formatting errors can have huge negative implications on whether people who need captions are able to comprehend content.

And now I want to show a few other examples of ASR errors. And the examples I have here are actually from the tests that we did for this year's report. So, as I mentioned before, some common causes of errors can be multiple speakers or overlapping speech, background noise, poor audio quality, false starts, and complex vocabulary.

And on this slide, I have two examples. And the first one shows some complex vocabulary. So the left side is the human-edited transcript. And the right side is the ASR transcript. And so the words "photomechanical" and "mechanical" are confused, "vitronectin" becomes "vibrant actin and fiber," and then "fibronectin" becomes "actin," and "results" becomes "resolved."

And then in the second example, the word "lat" becomes "lap." And there's also an instance of a function word error. So "can" becomes "can't." And the sentence goes from "I can even work" to "I can't even work," which means the complete opposite. And so it's really important that these words are accurate.

And now going back to our report findings, I'd like to compare 2021 to 2022. So we have a few notable findings. Speechmatics maintained its edge across all markets and was the top engine when evaluating ASR alone. Speechmatics also improved significantly by 33.3%. And in general, performance has improved for all engines across the board since our last report published in January 2021.

And to wrap up here, I'd like to reiterate our key takeaways. There were several improvements in technology and training capabilities in the past year. And some of the best ASR systems can achieve accuracy rates in the high 80's and low 90's. However, even with these technological improvements, even the best ASR engines are still a long way from being accurate enough for accessibility and for legal compliance, making humans a crucial part of creating accurate captions.

Last, we have some other exciting webinars coming up that you might be interested in attending. On June 16, we'll present our Quick Start to Captioning webinar on the basics of closed captioning. On June 23, Lainey Feingold will present on a Legal Analysis on the Future of Podcast Accessibility. And on August 11, we'll present on our Intro to Audio Description.

So thank you all for joining us today. And we'll now open it up to questions. The first question is, "Did you use the enhanced models for Deepgram and Rev, or just the default models?"

TESSA KETTELBERGER: These weren't like specialized models. I know some places offer them. The only place we use the specialized video model was with Google. The model we used for Rev was their v2 model that they, I think, in April published a lot of accuracy numbers about. We made sure that we were using that model. And with Deepgram, we used their newest default model as well.

REBECCA KLEIN: Great. "And is 3Play Media post-processing a technology? Or is it human post-editing?"

TESSA KETTELBERGER: So we have two different post-editing, post-processing processes-- that's a mouthful. When we either do ASR-only services, which we offer, or when we're going to present an ASR transcript to an editor to create captions, for both of those cases we will do some additional processing on the Speechmatics transcript to try to improve the accuracy a little bit.

This is something we're able to do. Because we have so much text data, we're able to optimize a little bit for captions and for our customers. And that'll improve a bit.

And then after that, there is additionally editors who will get-- and the editors, we believe, should be getting to 99% accuracy. So the numbers presented here with the proprietary post-processing have not been edited by human editors. We would expect a much lower error rate for that.

REBECCA KLEIN: Someone asked, "Are these results based on English language only?"

TESSA KETTELBERGER: Yes, they are.

REBECCA KLEIN: "Is AWS recognition going to be a part of a future comparison?"

TESSA KETTELBERGER: We would like it to be. AWS recognition has a terms of service on their speech recognition that basically says, if you are going to publish accuracy measurements on our outputs, you need to publish that data that went into the measurement publicly.

It's a fair idea. It would be cool if we could. We would not be able to publish our customers' data publicly. That would be massive data breach.

We actually did internally test AWS recognition. We're not allowed to publicly release the results. So I can kind of speak about them a little bit. They did quite respectably. They were sort of middle, front middle of the pack.

But we aren't switching to AWS recognition. They did not beat out any of the engines that we're currently working with. That's probably the most I can say.

REBECCA KLEIN: "Do you have any hypotheses as to why speech recognition tech has improved significantly over the past year?"

TESSA KETTELBERGER: So I don't build speech recognition technology. I have read plenty of papers about it, but I don't know that I could get into it in the short time that we have. But I will say, first of all, there's just a lot of motivation from these companies as, first of all, accessibility is at the forefront, but also as we've gone remote and there's been a lot of demand for these technologies.

The other thing is just more efficient use of data, the ability to train on data that is not pre-transcribed, that's completely unlabeled data, that's where we're seeing a lot of improvements. We only tested English here, but those improvements we're seeing are making it a lot easier to do speech recognition on multiple languages as well, especially in languages where there is less training data to be had, which is really cool.

REBECCA KLEIN: "Is there a best practice for speaker identification in a multiple-speaker setting? I've seen examples as Speaker A and Speaker B. Is that adequate?"

TESSA KETTELBERGER: It is always best if your speakers are-- and Rebecca might have more to say on this one-- but it is always best if your speakers are known by name to the participants or the viewers, that they are speaker labeled by name if you're going to use speaker labels.

Sometimes, it's all context-dependent. So if the name is not known, Speaker 1, Speaker 2 is probably fine. It's very context-dependent.

REBECCA KLEIN: Yeah, I'd say if a hearing person, or someone who can hear the audio, will know what this speaker's name is, then that should be included in the transcript or the captions.

"Based on your review of the current technology, what is your best guess as to ASR reaching a threshold of 95% accuracy?"

TESSA KETTELBERGER: So I don't know when it will reach 95% accuracy. But we did go through on our best engines and check to see by industry how many were getting above a 95% correct rate, or a below a 5% word error rate.

And we were seeing up to 40%, 50% of files in some of the better performing industries we showed earlier. So if you are coming with very high-quality audio without overlapping speakers, if you're in an industry that the training data is good for, that doesn't have a lot of really technical language, you could aim for that now I would say. Obviously, the circumstances where you can reach that accuracy currently are very specific. But I'd bet it applies to at least one person of the webinar.

REBECCA KLEIN: Someone asked if the captions available for this webinar are Zoom's ASR captions through Otter.ai or if they're our captions. And I can answer this quickly. They are our own live professional captions.

The next question, "Do you have demographic data as well as industry data for the samples we received? ASR often has difficulty with nonwhite and/or accented speakers. So I'm curious if your tests saw differences among speaker demographics."

TESSA KETTELBERGER: Yeah, we weren't able under the time constraints to label the data for speaker accent. There's a lot of difficulty labeling for demographic information like that. The person doing the labeling has to know a lot about it, the decision on which labels make sense. And how to section off into categories is very difficult.

With something like this, it's best to go with a vendor. And we know there are vendors out there. A lot of them have pre-existing data they've labeled themselves that they would like you to use, which means that it would have to be something we do in addition to our study instead of or in place of this study. It's not something we've done. We know we have many accents and a lot of diversity in the speakers in this data. But we weren't able to code for it and measure it.

REBECCA KLEIN: "Do you have data about accuracy rates for specialized language, such as science or more technical terms?"

TESSA KETTELBERGER: We also do not have that. Another thing that we would love to have-- the industries where we see a lot of technical language are education and e-learning. We have science and math, very specialized medical texts in there. And those industries did quite well. But that's the best way we have to view that in aggregate rather than spot checking files.

REBECCA KLEIN: "If a ASR is all you have, maybe for budget reasons, how should you make the most of it?"

TESSA KETTELBERGER: I just clicked away from my notes-- one second. Yeah, so, I can talk about this. Rebecca might also have ideas. But basically, if you only have ASR, you have to take steps to ensure that the ASR you're getting is the highest quality result for the situation you're in.

That means good audio quality, trying to eliminate background noise, trying to eliminate things-- if you can, it's not always realistic-- things like high-frequency noise from fans, air blowing around in the room, something that might not stand out to your ears but is going to cut off the higher frequency of the speech that is part of what a lot of engines will be trained on.

The other thing that could cut off high-frequency noise is talking through a cell phone. Cell phones, if you're calling into a meeting or something, tend to cut off the higher frequencies. And it's something that the human ear and the human mind is able to adjust to. It's not something that ASR adjusts to as well as the human ear. So avoiding phone calls, phone calling into live meetings and stuff is always a good thing.

Try to have one speaker at a time. Try not to have overlapping speech. Try not to play music in the background of the video that you're sending to be ASR recognized, because it will interfere. If you can have ASR done before you start editing in things like audio, additional audio effects, or music, that will always be better.

And then try to take steps to make up for captioning features that ASR doesn't offer. So if you're changing speakers, it's always nice to say, oh, I'll hand it off or thanks for letting me introduce myself or say like I'm speaking now if you're in a meeting. That way, it's easier for people to follow the conversation.

And note audibly or note out loud in a way that speech recognition can understand, things like changes in audio that are relevant. So it's always nice to say out loud, like, we're about to start playing a song if you're going to do that. And then when it's over, note that it's over, what song it was, things like that that ASR isn't going to capture but that you can with your voice.

REBECCA KLEIN: And I'll also just add, make sure that you build into your workflow a way of editing your ASR captions.

So someone says, "I'm a deaf person working in theater. Our recent attempts to use AI for live shows were not successful. I'm curious about the lag of the entertainment industry behind government, education, e-learning, et cetera. The dialogue is already written. So does this mean that the industry priority is finding a way to convert thousands of scripted lines into a subtitle file versus relying on speech-to-text AI?"

TESSA KETTELBERGER: Let me read this a little more closely.

REBECCA KLEIN: Yeah, it sounds like--

TESSA KETTELBERGER: There's a lot of--

REBECCA KLEIN: Yeah, there's a lot there. It sounds like they're generally curious as to why the entertainment industry is lagging.

TESSA KETTELBERGER: It's got to be in part that of the professionally produced content out there, the entertainment industry has the most background noise. Audio quality and background noise make a huge, huge difference. And there just won't be music in the background of a government meeting, for example.

There is technology that can align transcripts. So if you have pre-existing scripts written, you can align using speech recognition-- that text-- to the recording. The only problem is you can't do that live. So live accessibility is a completely different constraint there.

If you were to do alignment, you would still likely need to split it up into shorter sections. It's a process that could take computers a very long time to do if the video is very long and has a lot of noise in it. And then it would require probably some editing afterwards as well to make sure that it did everything the way you expected it to do.

REBECCA KLEIN: And then I think we have time for one last question. "So when you say 95% accuracy, what do you mean?"

TESSA KETTELBERGER: 95% accuracy?

REBECCA KLEIN: Yes.

TESSA KETTELBERGER: So here when I was mentioning it in a previous question, I meant under a 5% word error rate, so not considering formatting. I could in other parts of this have said accuracy, referencing formatted error as well. Generally, 95% accuracy means that 95% of the words that are present in the audio were accurately transcribed, whether or not we were including formatting. Depends on what slide we were on, but.

REBECCA KLEIN: OK. So that's all that we have time for today. Thank you all for being here. And please look out for the email with a link to the recording. I hope you have a great rest of your day.