2025 State of ASR

[00:00:00.56] ELISA LEWIS: Thank you everyone for joining us today for the session, 2025 State of Automatic Speech Recognition. My name is Elisa Lewis. I'm on the marketing team here at 3Play Media, and I will be moderating today's webinar. I will actually be presenting today as well alongside my colleague, Mike Chalson. Mike is the VP of Data Science here at 3Play.

[00:00:29.18] So today's agenda-- first going to start with a little bit of level setting, defining some key terms that we're going to be using throughout. Then we will talk about automatic speech recognition and how it differs for different use cases, specifically some of the nuances of using this technology for captioning and transcription. Then we'll talk a little bit about the annual report that we do and that this webinar is covering. And then I'll hand it off to Mike, who will dive deep into the data, talk about the different engines that are tested, the findings, some key takeaways, and then again, we will have time at the end for Q&A.

[00:01:14.43] So first I want to answer the question, what is ASR? We are here today for the State of ASR webinar. This is a first look at our State of ASR annual report. So what is ASR? ASR stands for automatic speech recognition. It refers to the use of machine learning, which you may see as the acronym ML, natural language processing-- again, you might know as NLP-- and artificial intelligence, or AI technologies, and how they convert speech into text.

[00:01:51.40] ASR is used in a number of different ways, many of which you're probably familiar with in your daily life. I have a few examples of how ASR is used on the screen, so voice assistants, asking Siri or Alexa for something-- maybe what's the weather, or setting timers. Dictation, using voice to text to send or write text messages, emails, take notes. We're seeing a lot more virtual customer service pop up. You might hear, in a few words, tell me about what you're calling about.

[00:02:30.72] Search and commands-- you can use your TV remote to get you to a specific channel. Smart appliances-- you can use ASR to use your smart appliances throughout your house, maybe turning on the oven to a certain temperature, et cetera. And then, of course, transcription and captioning, and that's really captioning the spoken word on a video or a live event or a meeting.

[00:03:00.29] And so for the purposes of our report and this webinar, we're really going to be honing in on ASR for the task of captioning and transcription. And the reason that it's important to make that distinction is because there are a lot of nuances in how ASR is used for captioning and transcription. People sometimes ask, well, why is Siri or Alexa so good, but then my auto-captioning is so bad? And captioning presents some unique challenges that are not present in those other instances that I just mentioned.

[00:03:42.23] So there are three main categories that explain why captioning is challenging for ASR and why it's so unique. Those three categories are variety, length, and readability.

[00:03:56.83] So variety, there's a ton of content in the world that can be captioned. It can be anything from a TV show to a math lecture and anything there in between. But when you're

asking, for example, to set your oven to a certain temperature, it only has a few tasks that it's trained on. It only has a few things that it will understand. Same thing with asking Alexa what's the weather. Sometimes you ask a question and she just says, I didn't understand that. But if it's something that needs to be captioned, there's a huge variety of content.

[00:04:39.19] The next is length. When you're captioning or transcribing content, it's usually long form There's a lot of room for utterances, shifting contexts, all of that in a longer form audio.

[00:04:57.78] And then readability-- Captions are consumed by humans, so they actually need to be understandable. Formatting is much more important. Really capturing every single detail is really critical, where it may not be for asking Siri a question. They just need to get the gist of it in that case.

[00:05:23.91] So that sets up some good context for talking a little bit more about the report itself. The State of ASR Report is a report that we have conducted at 3Play annually for the last several years, with the goal of really understanding the evolving capabilities and challenges of ASR technology.

[00:05:49.47] And we want to understand this not just in theory but in real-world applications. So we really conduct this study to benchmark the top ASR engines, again, understand their real-world content across industries, and measure the accuracy in both word error rate and formatted error rate, which we'll get into in the next few slides, and really ensure that we at 3Play are also using the best performing tech. So we use ASR as a critical step in our human edited process, so this research is really important for us as well to understand and make sure that we're staying on top of trends. So we're deeply invested in understanding this year over year.

[00:06:44.57] So I mentioned industries on the last slide, and I want to explain a little bit more about why that distinction matters. Different industries create different types of content, and they each have their own challenges, their own preferences, and ways of evaluating what good looks like. So just kind of running through some of the industries that we look at. Goods and services, so think Fortune 500 companies. This could be anything like online product videos, associations, e-learning. E-learning videos are really high production value online courses.

[00:07:30.49] News and networks-- that's really talking about live news coverage, media publishing, online news sites, government, technology, so promotional training and product demos, higher ed classroom style lectures. Cinematic, which is more of those streaming platforms, movies, anything that are really dramatic, pacing is important, highly scripted content. And then sports, so live events, interviews, sports clips. And we'll talk more about these nuances as we get into the data.

[00:08:19.47] So again, a little bit of education and level setting before looking at the data. I want to make sure that, in order for these findings to be really clear and meaningful, we understand the scoring metrics that are used. So we're going to talk about word error rate and formatted error rate, which I alluded to a little bit earlier.

[00:08:44.64] And word error rate or WER, W-E-R, is the most common and traditional way of measuring ASR accuracy. It tells us how many mistakes the ASR engine made compared to a human generated transcript. And WER is the total of three different types of errors-- substitutions, insertions, and deletions.

[00:09:12.78] So a substitution is when the engine maybe hears the wrong word, for example, if it transcribed cat, C-A-T, the animal, instead of cap, C-A-P, like a hat. This is something that is more likely in ASR because it can't take into account context like a human would be able to.

[00:09:40.26] Insertions are when extra words are added that were not actually spoken. So perhaps ASR inserts a random "the" where there wasn't one. And this could be due to background noise. Maybe there's somebody speaking in the back, or just shuffling noises or whatever it may be, and that gets picked up as a word. And then deletions are when a word that was said or was spoken gets completely left out.

[00:10:18.80] And then there's formatted error rate, or FER, F-E-R. And FER looks at errors after applying formatting rules like punctuation, capitalization, speaker labels. And this is really important because a transcript may have the correct words but have errors in punctuation or structure that can make it really difficult to read or comprehend.

[00:10:51.73] So I have an example on the screen that says, "I enjoy cooking my family and my dog." And then underneath that it says, I enjoy cooking, comma, my family, comma, and my dog. And it says punctuation matters. So again, that really changes the meaning, even though it's just a difference of two tiny little commas. So this is a good example of why formatting really counts.

[00:11:27.30] We touched on this briefly, but I want to mention some common causes of ASR errors. For word errors, this may be something like multiple speakers have overlapping speech. Again, maybe there's background noise. It could be because there's poor quality audio, false starts, Acoustic errors, so that would be things that sound similar. Maybe saying 3Play Media, if you say that quickly, could be transcribed as encyclopedia if there's no context there-- and then function words.

[00:12:14.57] And for formatted errors, we often see these ASR errors with speaker labels, so leaving out a label of who's speaking, again, leaving out punctuation, numbers written incorrectly. This is really important when we think about educational content in a math lecture. Non-speech elements, and then those inaudible tags. So these are just some common causes or common examples of where ASR will fail, again, because it doesn't have the ability to take into account context or understanding.

[00:13:03.38] And then this is another example of where formatting-- or sorry, this is not formatting. This is another example of just a common ASR error where a small change, or a small substitution in this case, changes the meaning completely. So the example says, "the contract was valid" versus "the contract is valid." And this is something that, depending on the audio quality or the speaker clarity, was or is, they're both one syllable. They can sound very similar, but it really changes legal implications in this case. It implies that something is in force

or was no longer in force and can be really critical in the outcome to understanding or conveying that information. This is something that a trained editor, human editor, would be less likely to make, again, because of context and different things.

[00:14:19.63] And then this is an example. I mentioned sports as an industry that we tested, and we'll get into this a little bit more again when we get into the data. But this is an example where the complex vocabulary was very difficult for the ASR engine.

[00:14:41.13] The example that I have here, I'll read through it in pieces. It should say "picked up really well by Ehrhardt," as in a name, but it says "picked up really well by air." And then the next sentence reads "quick pass in front. Bowen slaps it home." But the ASR transcribed "quick," and it left a blank space, "passing front bone slaps at home." There's no punctuation there. Then we have "Virginia one, comma, Loyola nothing." It transcribed the words correctly, sort of. It actually messed up Loyola and said "loyal nothing."

[00:15:26.11] So really, if you were using the transcript here to try to get that equal access, you have no idea what's going on. It's not even captioning the correct names. The grammar, the punctuation is all off, so they're not complete sentences. It would be very, very difficult to follow. And again, this is something that a human would be able to prevent some of these errors.

[00:15:55.94] And then another component that I want to level set on before we get into really evaluating accuracy is the transcript style. So there are two different transcript styles. There's clean read and verbatim. And this, whether we're evaluating based on a clean read or verbatim standards, is typically a preference either by industry or by an individual customer.

[00:16:28.49] So a clean read, for example, omits filler words and false starts. It's really focusing on the core message and generally will show lower error rates. So this is something that is preferred in educational and corporate content pretty often. And an example of this would be where the raw audio says, "so, um, like I was saying, the-- the project timeline." And the clean read would just say, "the project timeline." So it would remove some of those false starts, those utterances, and just, again, focus on getting the core message clarity, whereas verbatim retains more of the speech elements, and it removes some major disfluencies. But it still, for the most part, is really grabbing every word. It shows higher error rates across engines, and it's important for certain markets like sports or entertainment.

[00:17:41.77] And the example of this would be if the raw audio says, "so, um, like I was saying, the-- the project timeline." The verbatim would say, "So like I was saying, the project timeline." So it got rid of some of the ums and it got rid of some of the false starts, but it still kept the majority of what was being said.

[00:18:08.52] So with that, hopefully that gives a good understanding of what we'll be looking at and the nuances when we dive into some of the numbers here. But I will pass it off to Mike to get into the data.

[00:18:25.68] MIKE CHALSON: Thank you very much, Elisa. I hope my audio is coming through OK. So let's just jump into it. For 2025, we have new data and new engines. First, I'd

like to talk through what's new about the data. So this year, we tested 1,067 files containing 205 hours of content and 1.7 million total words across those files. In terms of duration, this was a 30% increase from 2024, when we did 158 hours.

[00:19:07.05] Like previous years, the files selected for testing contain real-world content and are intended to accurately reflect our business needs. The sample represents a variety of industries and subjects, as well as a broad range of video durations, numbers of speakers, and audio quality levels. All files contained English language speech. And let's go to the next slide.

[00:19:38.98] So we're really excited about some of the changes we made to the list of engines we tested this year. We tested eight conventional ASR engines, and we also tested a multimodal LLM prompted to perform the speech transcription task. I'll give a little more detail on each engine starting now.

[00:20:00.71] So first, Speechmatics regularly releases updates to their Ursa-2 model. We tested the version that was released in August 2024, and we used their enhanced operating point. AssemblyAI also regularly releases updates. We tested their Universal-2 model, which was released in October 2024. Like last year, we tested the official open source library for OpenAI's Whisper Large V2 and Large V3 three models. Those models haven't been retrained since last year, but the library that invokes the models had some very minor updates.

[00:20:42.48] Just for a little historical context, I want to give a reminder that Whisper was released in December 2022 and got a lot of attention because it was a breakthrough, open source, generative AI approach to speech recognition. It was trained on 20 times more data than any previous model of its type, and when it came out, it showed state of the art error rates, despite some occasional and very concerning errors that people tend to call hallucinations.

[00:21:13.41] Last year, we were aware of another open source library called WhisperX, but didn't systematically test it until this year. So this is one of the major changes this year is including WhisperX. WhisperX combines the official Whisper models with more traditional phoneme-based ASR models and more advanced techniques for invoking the models. It was released in early 2023. Like last year, we tested Microsoft's Azure Speech To Text service, which we believe was updated in January 2025.

[00:21:48.32] We again tested Rev AI's V2 model. As far as we're aware, it had no publicized updates since January 2024. From Google, we tested their new Chirp 2 model, which was released in October 2024, and from what we could understand, was their recommended premier ASR model. Lastly, and this is another one of the major updates we're excited about, we tested Google's Gemini 2.0 Flash multimodal LLM, which I'll discuss further in the next slide.

[00:22:32.50] So first, I'd like to talk about why we chose to test Gemini. Generative AI and large language models like those underlying ChatGPT have received a lot of attention over the last two years. We've seen significant success using LLMs to enable several of our other accessibility and globalization services, and we knew that Gemini was one of the few models that claimed to support audio transcription. So we thought it was a good idea to see how it really compared to the leading ASR engines in the industry.

[00:23:03.22] What we found was that using Gemini as an ASR solution required significantly more technical skill and vigilance than the conventional ASR engines. This was true because you may have to do some prompt engineering experiments to increase its reliability, and even then, you may need to implement systems to detect failures.

[00:23:25.25] In general, it was a little bit more brittle than we expected it to be. On some of the longer files, Gemini failed to transcribe the full file and returned a max tokens response code. There's known max token limits for models like these, but in the case of using it for ASR, you wouldn't know how much audio duration would actually reach the max token limit.

[00:23:48.87] With some experimentation, you could conservatively split up your audio files and piece together the transcripts that resulted, but that's more technical effort. The more concerning observation was that sometimes the transcripts clearly didn't cover the full files, but Gemini did not return the max token status. We were able to detect the transcript gaps because we had referenced transcripts. If you were using Gemini as your only ASR solution, you would presumably have no way to detect that the transcript ended earlier than the real speech.

[00:24:25.92] So ultimately, Gemini failed to complete a large enough number of files that we decided not to include its performance metrics in the main results of the study. Since we wanted all engines to be scored on the same files, we felt it biased the results unfairly to filter the full data set to only include ones that Gemini was able to complete.

[00:24:48.45] However, in a separate set of analysis, we did compile the results for just this filtered data set and found that Gemini's error rates were nearly identical to those of Google Chirp 2. Gemini always had marginally lower error rates, but generally ranked just above Chirp 2 compared to other engines. We think this suggests the current version of Gemini at the time we tested was probably processing ASR-related prompts by invoking Chirp 2, then perhaps it was making some small improvements to the transcripts with the main Gemini LLM before returning final results.

[00:25:31.01] OK. It's finally time to look at some of the results. So first I'll talk through the top level WER and FER results for each engine. Then I'll talk through a few segmentations of these results. In the interest of accessibility, I'm going to voice over all of the numbers in the tables. First, I'll read through the WER and FER tables one column at a time. Each table is sorted by ascending error rate or descending accuracy. I'll say the vendor's name, then their error rate in that category.

[00:26:07.34] First, for word error rate. AssemblyAI had a word error rate of 7.3%. Speechmatics had 8.3%. WhisperX had 8.5%. Microsoft had 9.7%. Whisper Large V2 had 10.1%. Rev AI had 11.1%. Google Chirp 2 had 13%. And Whisper Large V3 had 19.6%.

[00:26:36.97] For formatted error rate, WhisperX had 14.8%. AssemblyAI had 15.6%. Speechmatics had 16.8%. Whisper Large V2 had 17.2%. Microsoft had 17.8%. Google Chirp 2 had 19.6%. Rev AI had 20.7%. And Whisper Large V3 had 27%.

[00:27:05.74] There are a couple takeaways from this top level data, and we want to highlight them. So for one, WhisperX performs very differently from the original Whisper models. What's not obvious from these error rates is that WhisperX showed no sign of the hallucination behavior that was the biggest problem with the original Whisper models. The other key takeaway is that both WER and FER for some of the engines still have very high error rates that would substantially hinder understandability of the transcripts. Let's go to the next slide.

[00:27:49.26] So now we're going to look at those numbers segmented separately for clean read versus verbatim files. Before I read the tables, I want to highlight some of the other observations we made about the different types of errors which Elisa described earlier Among the top performing engines, the ranking of substitution errors largely aligned with the overall error rate.

[00:28:12.45] However, for insertion and deletion errors, Speechmatics' numbers were significantly out of trend. Their deletion error percent was by far the lowest out of all engines. Their insertion error percent was much higher than the other top engines. This is consistent with what we measured last year.

[00:28:34.43] When we investigated those numbers by inspecting individual files, we observed that Speechmatics transcripts often included barely audible background speech and disfluencies beyond what is typically desired in captioning outputs, even by verbatim standards. So there could be other use cases where that level of sensitivity is preferred, such as some of the ASR use cases Elisa mentioned at the start of the presentation. But for the captioning use case, we felt it was appropriate to score those as errors.

[00:29:07.70] Now on to the clean read and verbatim segmented results-- These tables are each sorted by the clean read columns in ascending error rate. I'll read through the WER and FER tables one column at a time. I'll say the vendor's name and then their error rate in that category.

[00:29:25.24] So first for word error rate, clean read-- Assembly scored 6.3%. WhisperX scored 6.5%. Speechmatics scored 6.6%. Whisper Large V2 scored 7.6%. Microsoft scored 8.5%. Rev AI scored 9.8%. Google Chirp 2 scored 11.3%. And Whisper Large V3 scored 17.3%

[00:29:53.92] For word error rate on the verbatim segment, AssemblyAI scored 10.7%. WhisperX scored 15.2%. Speechmatics scored 13.9%. Whisper Large V2 scored 18.7%. Microsoft scored 14%. Rev AI scored 15.7%. Google Chirp 2 scored 18.7%. And Whisper Large V3 scored 27.2%.

[00:30:23.28] For formatted error rate, clean read, WhisperX scored 12.7%. Assembly AI scored 14.5%. Whisper Large V2 scored 14.6%. Speechmatics scored 14.8%. Microsoft scored 16.4%. Google Chirp scored 18%. Rev AI scored 19.1%. Whisper Large V3 scored 24.5%.

[00:30:52.26] And then for the verbatim segment of formatted error rate, WhisperX scored 22%. AssemblyAI scored 19.4%. Whisper Large V2 scored 26.3%. Speechmatics scored 23.6%. Microsoft scored 22.7%. Google Chirp 2 scored 25.3%. Rev AI scored 26.2%. And Whisper Large V3 scored 35.6%.

[00:31:28.61] Yeah. Let's move on to the industry breakdown. So the last data I want to show is the summary of performance for different industries. This table shows the average WER and FER for files from each of our industry categories. The numbers are the average across the top four performing engines only, which are Assembly, Speechmatics, WhisperX, and Whisper Large V2.

[00:31:56.36] So first, word error rate-- For goods and services, the word error rate was 4.3% For associations, it was 4.9%. For e-learning, it was 5.1%. For news and networks, it was 5.4%. Media and publishing, 5.5%. Government, 6.1%. Other, 6.1%. Tech, 6.3%. Higher ed, 6.4%. Cinematic, 7.2%. Sports, 14.9%.

[00:32:31.54] For formatted error rate, goods and services had 10.4%. Associations, 10.9%. E-learning, 12.6%. News and networks, 12.0%. Media publishing, 12.5%. Government, 12.8%. Other, 12.4%. Tech, 13%. Higher ed, 13.6%. Cinematic, 15.5%. And sports, 24%.

[00:33:03.10] So one of the key takeaways from this table we want to highlight is how much higher error rates are on sports content. We typically find that sports content is so hard because of a combination of complicated noise environments, unscripted speech, many player and coach names that may not be well represented in ASR training data, a lot of numerical information like scores that have unique phrasing conventions.

[00:33:31.47] In particular this year, we noticed a lot of content from settings like ad-hoc locker room interviews, which often had a lot of false starts and interruptions from secondary speakers. For instance, in the middle of an interview question, the main speaker might be speaking and a secondary speaker might interrupt them to say congratulations or something like that. Next slide, please.

[00:33:59.94] So that was a lot of detail and a lot of data. Let's pop back up to the key takeaways from this year's study. Firstly, based on error rates. Assembly's Universal-2 model and WhisperX seem to stand slightly ahead of Speechmatics, and the three of them stand substantially ahead of all other engines. Beyond error rates, some use cases might benefit from Speechmatics' bias towards transcribing more.

[00:34:29.60] WhisperX, we would say, is not a viable option for users who don't want to self-host their own infrastructure or who require some of the secondary features of Assembly and Speechmatics, like word level confidence scores. Secondly, based on the full as-read experience, formatted error rate, the best available ASR engines still fall far short of accessibility requirements.

[00:34:57.53] Thirdly, as we saw in previous years, ASR accuracy varied greatly among different industries. We believe this is a combination of inherent differences in difficulty of the audio and speech, and also differences in the availability of model training data and evaluation data from different industries. Fourth, large language models can do many amazing things. They're not yet a viable replacement for dedicated ASR engines. Next slide.

[00:35:35.23] So we just want to come back around from statistics to a real-world sense of scale. 5% to 10% word error rate or 10% to 20% formatted error rate may sound pretty low, but remember that means one out of every 10 to 20 words, or 1 out of every 5 to 10 words or punctuations would be incorrect. That's going to have a big impact on understandability. With that, I'd like to hand things back to Elisa.

[00:36:05.92] ELISA LEWIS: Thanks, Mike. So we are getting ready to move into Q&A. We've had a few questions come in, but please feel free to keep them coming and we'll get to as many as we can in the next approximately 15 minutes. The first question we had is, "WER is scored versus a human captioner. What if the human isn't so good and makes a lot of errors? I've experienced some awful human captioners. Do you average multiply humans or just rely on a single instance of a captioner?"

[00:36:52.47] MIKE CHALSON: I think I'll take that one. In general, for a given file, we do not average across multiple human evaluations, but we do overall evaluate on a very large number of human evaluators and a large number of files, and we also take measures to refine the data set to filter out lower quality reference transcripts that we were able to detect.

[00:37:23.52] ELISA LEWIS: Great. Thank you. I think most of these questions are probably going to be for you, Mike. The next question is, "WER and FER provide insight into specific performance functions. How do you calculate overall scoring, so i.e., the bottom line performance?"

[00:37:47.52] MIKE CHALSON: I suppose I'm not quite sure what the bottom line performance would be referring to. In general, when we do word error rate and formatted error rate, we also look a little bit deeper into the rates of substitution errors, insertion errors, deletion errors. And of course, there's a whole host most of ASR engine performance criteria other than error rate and accuracy that a user might care about. But we didn't study those systematically enough that it's appropriate for us to really comment on them in the report.

[00:38:27.46] ELISA LEWIS: That makes sense. I'm guessing they're kind of looking for which is overall the best, which you answered somewhat in the key takeaways. So hopefully that helped answer that question.

[00:38:42.49] The next question we have is, "these tests are done on entire files equivalent to captioning a video on-demand. Live captioning has unique characteristics like less audio context, lower latency needs, et cetera. Any data on how these models perform on a live captioning or in a live captioning setting?"

[00:39:08.82] MIKE CHALSON: So that's a really good question. Those are really good points. Totally agree. Very different ballpark. We don't have systematic evaluation of that to a point where we should be commenting on it, but 100% agree that live introduces a whole host of additional criteria.

[00:39:29.19] ELISA LEWIS: Yeah, and I think that's a great reminder and clarification that this report was focused on prerecorded content types and the captioning for prerecorded content, not live captioning.

[00:39:46.98] Let me see here. The next question that we have-- just moving this window here. The next question we have-- "Someone asked or someone said, I may have missed the mention of accents. Is there guidance for using ASR with individuals whose native language is not English?"

[00:40:15.93] MIKE CHALSON: I would say that generally speaking, the top tier engines are very robust for different accents. We definitely didn't study it directly to know which ones might be better at certain accents. But I would say you could probably bet that overall, recordings where the speaker has a strong accent are going to have a higher error rate. But the leading engines probably do better than you might expect.

[00:40:47.07] ELISA LEWIS: Thank you. Another question we have-- Someone is asking, "do you find that certain engines are better for government or legislative settings?"

[00:41:04.52] MIKE CHALSON: Yes, I think so. I don't have the data handy. It might be in the final report that's going to follow up on the webinar. We certainly have done the analysis.

[00:41:16.46] ELISA LEWIS: Great. Thank you. And that's a good reminder. It is on the slides right now, but there is a QR code to get on the list to receive the full report. The full State of Automatic Speech Recognition Report will be out in the next week or so. This was the first look into this year's data, but as Mike said, there's more information, more examples, tables, just lots more data in the full report. So as we're going through some more questions, feel free to use that QR code that is up.

[00:41:59.45] The next question that we have, someone is asking, "do you have a sense of what ASR engines need to improve to get closer to accessibility requirements?"

[00:42:16.27] MIKE CHALSON: Well, I think clearly from the tables, there's a much, much larger gap in the formatted error rate than in the word error rate, and that shouldn't be underappreciated. As Elisa illustrated with some examples, formatting can dramatically change the interpretation of what's there. And that does seem to be the next frontier where there's still pretty significantly high error rates.

[00:42:48.37] I do think in general, within the realm of word error rates, we still see a lot of errors with what we would call polarity, which tends to be related to the function words, but where can is transcribed as can't or vice-versa, or some of those function words are just completely backwards, gives completely the opposite intent of the original speech. That seems to occur as an error higher than other kinds of word error rates. That's pretty significant, for instance.

[00:43:29.07] ELISA LEWIS: Thank you. Someone else is asking, "did you provide dictionaries for a particular category? For example, sports, the names and jargon can be fed into the ASR engine to improve accuracy."

[00:43:45.30] MIKE CHALSON: So essentially, no. So I think the more general response to that is some engines offer a variety of ways to provide extra information outside the context of the audio file itself to help the engine perform better, given that prior knowledge. That's an example of features that not all engines offer.

[00:44:13.97] We did not use every engine with every feature possible to get the best possible performance from each one. We sort of editorially chose a mainstream usage that a typical user would probably be able to do and tried to use a similar kind of usage of every engine. And so we didn't use custom glossaries for any of the engines.

[00:44:42.44] ELISA LEWIS: Thank you. We have another question. Someone said, "ASR is getting better every year. I would love to see if it holds captions up longer when possible, as opposed to end a caption block as soon as dialogue ends. This allows the viewer a bit more time to read the caption. Do the ASR engines account for this or are they working on this as far as you know?"

[00:45:09.44] MIKE CHALSON: I would say as far as I know, that tends to be something separate from the ASR transcripts themselves. The ASR engines do provide time codes. But often the translation from time-coded transcript to actually segmented captions that show up and stay on frame for a certain period of time have line breaks in a certain place, appear in the image at a certain location, tends to be a separate technical problem with separate solutions.

[00:45:50.05] ELISA LEWIS: Great. Thank you. I think we have time for one or two more. Another question we have, someone's asking, "for ASR, does 3Play use in-house models, third-party models, or a combination of both?"

[00:46:12.09] MIKE CHALSON: I suppose it depends how technically precise you want the answer to be. I'd say we generally use third-party ASR models and overlay on top of that a substantial accuracy improvement from our own proprietary technology, which we've been building as long as the company has been around. So I think it's both. We have a set of NLP-type technology that adds value on top of third-party ASR models.

[00:46:49.41] ELISA LEWIS: Great. And we have time for one more. The question is, "are you doing any research on improving the ASR on videos with the same speaker? Thinking, for example, higher ed lectures where you feed corrections back into the LLM."

[00:47:10.22] MIKE CHALSON: Multi-part question, it implies the use of an LLM in addition to an ASR engine. I'd say that we're generally doing research all around that area, and we already have some technology closely related to that. It's part of our platform.

[00:47:31.28] ELISA LEWIS: Awesome. Thank you. We have some final reminders, so we will end with that for the sake of time. Thank you, Mike, for presenting and answering all of those

questions. And I want to thank our audience for joining this afternoon, asking great questions, and I hope you enjoy the rest of your day.