JOSH MILLER: Hey everyone. Welcome to today's webinar. My name is Josh Miller. I'm one of the co-founders of 3PlayMedia. Thank you for joining us.

So today, we have two experts on this transcript alignment service that we recently launched. We have Roger Zimmerman, who is our VP of research and development here at 3Play, and David Zylber, who is our manager of customer happiness. So the two of them are going to go through a number of great details on the alignment service. So Dave, why don't you go ahead and get us started by telling us what we're going to cover.

DAVID ZYLBER: Thanks Josh. Today, we're going to cover everything you need to know about the transcript alignment service. We're covering submitting transcripts and media files, formatting your transcripts, and most importantly, the best practices to adhere to when submitting your transcripts for alignment. We'll also explore how this automated process works. For anyone who's familiar with the 3Play account system, this all should look fairly straightforward.

JOSH MILLER: Great. So why don't we start by explaining exactly what this transcript alignment service is and how it's different from the standard captioning and transcription service that a number of people are used to seeing from us.

DAVID ZYLBER: Sure. Well, the alignment service is a great option for our customers who already have transcripts for their media content, whereas you'd want to use the standard transcription and captioning service if you didn't already have a transcript for your video. The transcript alignment service offers 3Play account holders a faster and less expensive way to create and use our interactive plug-ins by synchronizing your non-3Play created transcript with your media files. This differs from the default 3Play Media service transcription and captioning, where we create the core text document from either a video or audio file that is uploaded to the system.

And then once your alignment is complete, you'll have access to all the products and services that come with our standard transcription and captioning service. This means you'll be able to download captions and transcripts in a variety of formats, create translations and multilingual subtitles, and use the interactive transcript and captions plug-in. You'll also be able to use the video search and clipping tools, captions and subtitles editors, and any other features, basically, that you would normally have access to.

JOSH MILLER: Great. So in terms of process, Roger, can you explain what some of the major differences are between this alignment service and the transcription and captioning that we normally do?

ROGER ZIMMERMAN: Sure Josh. Hi everybody. This is Roger. I'm happy to be here to discuss this exciting service and the technology behind it.

A major difference between the default transcription and captioning service and the transcript alignment service is that the transcription and captioning service combines an automated process, followed by human cleanup by our team of expert transcription editors and QA specialists. The alignment service, on the other hand, is essentially 100% automated. And that's very important to remember as we go through this webinar.

So just to elaborate a little bit more on those differences. So when using the standard service-- And those of you who have used that service will understand that, we, of course, have no transcript. So the media file that you upload initially goes through ASR, which is the acronym for automatic speech recognition, to create a transcript that is about 70% to 80% accurate.

Now, as everybody can probably understand, that is not good enough for most realistic applications. So 3Play's transcription and captioning service uses these human editors and QA specialists to ensure a nearly 100% accurate transcript. And again, the ASR output goes through two rounds of human cleanup in order to get to that accuracy.

The alignment process, on the other hand, is 100% automated. Doesn't go through any cleanup service. That means that when you submit your transcript to be aligned with a media file, it needs to be formatted and its content needs to be massaged, if you will, if you want to achieve the best results. And we'll get back to that shortly.

JOSH MILLER: Great. All right. So, how about the technology itself? Can we walk through what is actually happening in the background?

ROGER ZIMMERMAN: Sure. So, the first step in this technology is actually one in which the customer interacts directly with the technology. And this is where you are uploading your transcript to the system and associating it with the appropriate media file. Dave will go into some of the mechanics of that later.

But the key concept here is that the automated system requires that the transcript be entirely in pure ASCII text. Now, some of the transcripts that you have will already be in ASCIIs, but

others may not. And in particular, if your transcript document was originally created in MS-Word, you will need to export it to text format prior to uploading to our system.

Now, you should recognize, however, that Word does not use standard US ASCII format. It uses an extended form of ASCII. And as such, the technology needs to convert any non-ASCII character produced by Word to the nearest ASCII equivalent. If it succeeds, you will see the text in the interface looks normal, with quotes and dashes successfully converted.

If however it does not succeed, you will see question marks in the interface. And this may occur for transcripts produced in other word processing programs or by older versions of Microsoft Word. If this occurs, you will need to either manually edit your transcript or work with our support department to get your document into the right format. So Dave, why don't you take it from there?

DAVID ZYLBER: Sure. So I'm going to explain to you guys how to upload a media file. And we're going to do a little demo. So everyone should just be aware that for the alignment service, we offer all of the same upload options that you would normally have access to using the standard service. And we also accept all the same media file formats, as well.

Alignment services can be uploaded through your secure account, directly from your computer, submitting links to your videos, a YouTube video or a link that points towards a downloadable file. Or you can even upload content for alignment from a video platform like Brightcove or YouTube, but from directly within the 3Play account. And we also offer for batch uploads, someone that has a lot of files that they would like aligned, you can use the FTP or API to submit the files.

And we're going to explain a little bit about this later on. And if you're unfamiliar with any of these upload processes, we have this all documented on our support site. So feel free to visit that.

So now I'm going to just show you here uploading a file for alignment directly from the account system. So after you login to your 3Play account, you'll be directed to this beautiful landing page. And what you're going to want to do is hover your cursor over Upload. And you'll see from the drop down menu, we have our upload options here.

And we're going to do the Direct Upload, which is taking a file directly from your computer and uploading it to the system. And as I said before, these other methods are also possible ways

of submitting files for alignment. So we're going to click Direct Upload.

Next, we're going to choose a file from our computer. And we are going to pick our Fundraiser_project.mp4. Click Upload.

Next, we're going to select a folder. You can create a new folder. Or we're just going to select a folder that already exists in our account. We're going to do Alignment Tests. Seems appropriate. Click Next.

Now, here you'll see on this menu, Select Service. And you'll see, by default, that Transcription and Captioning is selected. And you'll also see you'll have options for turnaround times. Now, when you select Alignment Only, which is what you're going to want to do when submitting files for alignment, you'll notice that the turnaround time disappears. And basically, you don't have an option for turnaround time when using the alignment service.

The turnaround depends on the duration of the file and the capacity all across our system at the time upon upload. And most likely, you will get the file back in a couple of hours. If the file is over an hour, you might see the file returned the next business day. And also, please keep in mind too files submitted for alignment should be no longer than two hours per file.

So as you can see, we've selected Alignment Only. And as I said, the turnaround time disappears. We're clicking Next. We're directed now to the Confirm Upload page. You'll see what we've done here.

We have one file that we are going to upload. The Service is Alignment. The Folder we selected, Alignment Test. And by default, every time you choose Alignment, it'll just say, Turnaround Time, Standard. So we're good to go. Click Upload My Files. You'll see, boom, the file is now in the system.

If by any chance you click on the My Files page, and we see Fundraiser_project here now. And we see this pending icon right here. Don't worry. The file isn't actually being processed right now. It's actually still waiting for you, the user, to submit the corresponding transcript. So if you do go to the My Files page, don't worry if you see this. This is normal.

So what you're going to want to do after submitting the media file is go back again up to the Upload tab. And here we see Transcriptions for Alignment, 1 file waiting. This is a holding area, essentially. We're clicking it.

And what you're going to see here is we have a box here, where it says, Enter Transcript Text. We see the media file name, Fundraiser_project. I'm going to just show you guys the file, the actual transcript here. So here we are.

One thing that's interesting about this is you'll notice the speaker ID. We have Senator, Senator, and Narrator are in lower cases, followed by a colon. And as we'll explain in best practices, this should actually be capitalized. But I'll show you something cool here.

So it is lower case, which does not adhere to the best practices. So what you're going to want to do, if you check this box here, it'll actually automatically format it when you drag and drop the text file in. You can cut and paste the file in. But when you have the option to drag and drop, again, mentioning this in best practices, drag and drop is the preferred method.

So watch this. We click the box. And We're dropping in our transcript. And you'll notice that it has actually automatically formatted the speaker IDs, adhering to our standards. So it's all capitals now.

So next, you'll want to Save the transcript. If by chance you notice that there is something you would like to change, you can just click Edit Transcript. make your edit, save it again. But we want to do now once you're all set is click Submit For Alignment. And we're good to go.

You can go back to My Files. The file will still be pending. The next stage you'll see is in progress. And you'll know that the file is on its way to being aligned.

So next, we're going to want to just go over some differences that you might encounter when using FTP. There are some differences that you should be aware of. So FTP, if some of you have used this before, you'll know that if you want to change the service level to rush, you have to create a folder that will give us a way to know that you want rush service.

So similarly for alignment, you're going to want to create a folder in your FTP application called for_alignment. Then what you're going to want to do is first add the media file, for example, Casablanco.mp4. You would want to drag and drop into that folder.

And then secondly, what you would want to do is then add the plain text transcript to the for_alignment folder. And that would be called Casablanco.txt. And the text file must have the same name as the media file. So as you'll see on this slide here, it's Casablanco.mp4 and the transcript is Casablanca.txt.

And if you're going to do a batch upload, which is essentially when you'd want to use the FTP option, you can actually first drag and drop all of your media files followed by the corresponding transcripts. So you don't keep going back and forth, back and forth. This just makes it a little easier for you, the user. So now, we're going to keep on going with the technology part here.

ROGER ZIMMERMAN: OK, yeah sure Dave. Actually, before we get into more of the technology, let's go over the best practices. And just to keep those in the forefront there. And later, when we do go deeper into the technology, hopefully you'll see some of the motivation for these best practices. But let's put first things first. And the key best practice, of course, is that this technology assumes.

And if I may be so strong as to say requires, that the text corresponds to the audio. The assumption is that the transcript you are uploading is a faithful rendition of the audio in the media file. So what are some of the common problems we've seen thus far as regards this correspondence?

Well, as Dave just showed you, we do want to be able to know that the speaker IDs are just labels and not spoken text. So if they conform to our standards, which is they occur at the beginning of the line. And they're all caps followed by colon. The technology will understand that they are not spoken. And in fact, we'll use the fact that a new speaker's speaking to improve the output. So getting those speaker IDs to conforming format is important.

Another thing to be aware of is that any wrapped text, any time you have a hard carriage return in the text, the technology assumes that will be a paragraph. Doesn't matter if it's single line feed or double line feed. That's just an assumption made by the technology. Now, when you export a text document in Word, it doesn't do this. It keeps all paragraph together in one line. But you just need to be aware that when you're exporting from other formats or if your document came from some other source.

We've seen cases where the document itself is a script. And it includes instructions, screen directions, scene headings, headers, footers from the document, page numbers. All those things are assumed to be spoken by the technology. And of course, will lead it astray. So I want to emphasize that those kinds of extraneous metadata will definitely interfere with the performance of the technology.

If they occur only sporadically, the technology can usually recover. But that kind of thing will definitely interfere. And I might want to mention in that context, any sort of timing information is

also text that's assumed to be spoken and should optimally be removed.

Interpretation is a little bit of a fuzzy area. But basically any time a transcript tries to reinterpret what a speaker is saying in spontaneous speech, that can be problematic. Or that could be an area of misalignment. Usually, the biggest problem there is when speakers restart, when they hesitate and restart, We can have some issues with that.

Now, it's going to be very rare for a transcript to cover those kind of restarts literally. But we just want to make you aware of that. Any time the audio has overlapping speakers, people speaking at the same time, of course it's really difficult to represent that textually.

And so that's another area where alignment will have a problem. Again, if it's not a prevalent feature of the media file, it shouldn't be too bad. But again, something to be aware of.

And finally, just as in with our captioning and transcription service, the overall audio quality is a kind of correspondence with the text. If we have trouble discerning the speech because of the audio quality, be it background noise, frequency characteristics and such, that will impact the alignment. So it's all pretty straightforward. But definitely things to be aware of in terms of keeping the text in correspondence with the audio. So Dave, why don't you take it from there on the best practices?

DAVID ZYLBER: Sure. And as Roger was talking, you might have seen that I was pulling up a poorly formatted version of our screenplay *Casablanco.* And this is clearly something you would not want to submit here. There's a lot of extraneous text. It even tells you the play that we based this amazing screenplay off of.

So this is the kind of information we wouldn't want to see. We'd want to see it more like this, which as you'll notice, is a lot more simplified. It just removes all of that extraneous information. It just has speaker identification and the spoken words. We do have a couple of instances here where we call out some sound effects. And Roger, I believe-- That's OK right?

ROGER ZIMMERMAN: Yeah, it's OK, because it's a fairly small amount of text, the technology will insert it at the best timing points where it can fit it. And it will move on from there.

DAVID ZYLBER: Yeah. So use that within reason. And I would say before formatting, we have these articles up on our support site. These are great references to refresh your memory about the best practices before uploading. And if you adhere to these, the results will be far better had you

not.

So we're going to go back to talking about the technology.

ROGER ZIMMERMAN: Oh sure. Yeah, so let's get into the details of the process now. And you've been given a lot of background here. And I'm going to get into the details of the alignment technology. And the point here is not to wow you with our technological prowess, but to hope to give you an appreciation for some of the nuances. And this is for general educational purposes. And also, because these details may help to motivate the best practices we've been discussing.

So, we're just going to go step by step through the process here. Step 1 was the uploading of the text. And now we're on to step 2, which from here on in, I can't emphasize enough, is completely automated, completely done by computer algorithm.

So first, the program infers verbalization from the text. Text is filled with representations of verbalization, which need, essentially, to be interpreted. So we talked about speaker labels a lot. And just want to point out here that they are used for adaptation. The technology tries to adapt its models as it's going along to different speakers. And if we can note where those speakers are changing, technology can do a better job of that.

Punctuation is removed, because most of the time punctuation is not verbalized. But sentence boundaries are replaced with an optional pause, just as we have at speaker labels. A big feature, significant feature, of text is numeric tokens. And numeric tokens are, of course, shorthand representations of how those numbers are verbalized.

And so what we need to do is we need to interpret those. So I've given a few examples there, a date, a regular number, a height. And you can see that there are at least two possible verbalizations for each of those. In some cases, many more. And so the technology needs to understand that.

Finally, we have the acronym and abbreviation expansion in this interpretation step. And you can imagine that different kinds of acronyms and abbreviations will need to be interpreted differently. The St. can be either Saint or street. ABC is spoken as an acronym, whereas NASDAQ is spoken as a word.

So from that, we can build what's called a language model. And it's a biased language model. So a little bit of technical detail here. A language model is a statistical representation of the words that are spoken in a language. And that's one of the important models used in

automatic speech recognition technology. The other model being the acoustic model.

But the language model is key here because what we have is a text which can be used to bias the language model. In standard transcription and captioning, we need to use a general language model that allows virtually any sentence or sequence of words in the English language. But given the text that has been uploaded, we can bias the language model toward that text. And what we do there is we apply all these interpretive steps and we expand the language to have alternatives. That's what the bracketed items are there.

And you can see the speaker being replaced with metadata and the punctuation being replaced with metadata as well. So that's how we bias the language model towards the text. And this helps the speech recognition process. And that is the first pass through the audio is speech recognition, regular automatic speech recognition with biased language model.

So here's some potential output from that, just a sample. What we have is the words that were actually recognized. Notice that all the words there are included in the biased language model and the time points at which those words occur. So we can move onto the next slide, which is the next step, which is we need to realign the ASR output with the original text.

So in this slide you see on the left side the ASR output, on the middle column being the original text, and then the time points. And the key alignment step is we need to find anchors, islands of reliability, between the original text and the ASR output, which allow us to then align all of the text. And you'll notice that in cases where there is a word that occurs in the ASR output that doesn't correspond to the original text, we need to infer the time points from all of the ASR output or vice versa. In the case of "we'll", we need to interpolate the time points for we will in the original text. So these are just small issues that need to happen.

Now, step 6 is the really tricky one, which is that in any case where the ASR output cannot be aligned with the original text, we basically just need to fill in all of that time with the original text that isn't taking up by this alignment process, by the original alignment process. And to the extent that we need to do this over large amounts of original text, where, of course, not being very accurate with the time points. And this is where any problems that we've seen, most of them fall into this category, is a region of text that simply has inaccurate time points.

So at the same time, as we are in parallel, if you will, because this is a computer that we're reconstructing this text, we can also compute confidence from the ASR output and also take

into account the number and length of these gaps that we need to fill in. And those of you who have used our captioning and transcription service may be familiar with the audio quality bars that we associate with every media file. And in the case of the transcription service, we use the audio quality bars to take into account both the ASR output quality and this alignment quality. So that's what you'll see represented in those audio quality bars.

And finally, this was alluded to previously. But the final step in the automated process is to create all the outputs that we normally would create if we were doing standard transcription. And so at that point, the file in the account system looks like it had been transcribed by a human, but you, the user, know it was not. So that's the story. So at this point, it's time to hand it back to Josh to drive the rest of the webinar.

JOSH MILLER: Great. Thank you very much Roger and Dave for going through all of that. We're just going to take one minute to aggregate some of the questions. I've seen a few come in. I want to make sure we address them all. So I will right back in one minute.

DAVID ZYLBER: So we're going to answer a couple of questions that have come up here.

JOSH MILLER: Great. So first, one question is about languages. And we'll just quickly answer that. Right now, the only language we can process for the alignment process is English. There may be some other languages being added in the future. But as of right now, really it's just an English transcript alignment process. But another question along the lines of language is how the alignment process can deal with accents. So Roger, I'm going to let you handle that one.

ROGER ZIMMERMAN: Sure. Absolutely. So it's very similar to the answer we give about our captioning and transcription service, which is that accented, non-native speakers of English actually work quite well in speech recognition, as long as the audio quality is good.

And the reason for that, among them, is because of this speaker adaptation pass that I mentioned that's part of the automatic speech recognition service. The automatic speech recognition pass, excuse me, where inside the speech recognition, there's actually a full repeated pass that the technology goes through after it has adapted to the speaker characteristics. Now in alignment, there's the additional advantage of the biased language model. So as long as the audio quality's good, short answer is, we don't expect to see any problems if there's good audio quality and good correspondence between the text and the audio.

I just wanted to throw in one more thing regarding the first question. I just wanted to remind everybody that once you have the transcript in English, the assets work exactly the same as with the captioning and transcription service. And therefore, they can go through our translation service with our translation partner.

JOSH MILLER: Great. Thanks for adding that Roger. That's exactly right. All right. So great question about what a completed transcript actually looks like. So we'd love to show you that. Unfortunately, we can't show you some live examples, just because we have customer privacy issues. And we don't want to show any of that without their permission.

That being said, we are open to doing demos. So if you are interested, get in touch with us offline. We'll be happy to walk you through that. A question came up. And this is a good one. What happens when there's music? And how can that be handled well? So Roger, I'll let you take that.

ROGER ZIMMERMAN: Yeah. So that is a great question. And that certainly happens frequently in the media files we process, both in the standard service and the alignment service. And the answer probably won't surprise you based on what I've been talking about, which is that if the music is separated from the speech consistently and or if the music that occurs during a speech has been modulated in its volume so that it's very low compared to the speech, things should work pretty good.

It's when the music overlaps the speech and is comparable in loudness-- And we usually use a rule of thumb of about 15 decibel signal to noise ratio as being the floor at which things are good. So if you want to measure that, you can. If not, you can send it to us and we could try it. And if it's a problem, it'll show up in the audio quality bars.

DAVID ZYLBER: This is Dave speaking. I come from a video production background. And since you're submitting the spoken words that already exist in a file, that's basically assuming that there will be no changes made to the media file.

So if you do have access to your video project in a non-linear editing system, one suggestion might be even to strip out the music and make a copy of your sequence. Submit the media file without the music. And then after processing it and getting your captions, you could actually add the music back in post-production.

ROGER Right. That's pretty cool. And finally, annotations about music occurring in the transcript, again,

ZIMMERMAN: as long as they're short snippets of music playing, background music that occur infrequently in the transcript, that shouldn't throw off the alignment process too much.

JOSH MILLER: Great. Along the same lines, what if there is content in the transcript that doesn't exist in the video and possibly the other way around? What's going to happen?

ROGER ZIMMERMAN: Yeah. So the simple answer is any text which does not occur in the audio track of the media file gets forcibly aligned. We assume all of that text is there. So the longer that text is, the more problematic it will be for the alignment process to the extent where it can actually throw off the alignment for the text, which is represented in the media file.

So the recommendation is to go through and remove those chunks of text. And then the vice versa is also a problem, where you're missing text. And again, the text that is there is assumed to occur in sequence. And very often, that will throw off the alignment, because essentially words will begin at points in the audio where other words should have been. And that will throw off the process. And you'll get really long words is one of the things.

JOSH MILLER: And then, again, an extension of that. What if the transcript is accurate, but there's just a big gap between what's being spoken and the content?

DAVID ZYLBER: I'm not quite sure I understand that.

JOSH MILLER: So maybe there's a long pause. That's what the question is asking.

ROGER ZIMMERMAN: Yeah. So long pauses should not be problematic. That as long as there's not too much noise during that pause, that the process thinks it's speech. So gaps are handled quite elegantly by the technology.

JOSH MILLER: Great. So what about the case where a transcript is submitted and it's actually UK English?

ROGER ZIMMERMAN: So UK in the spelling. OK, so we've got that very well covered. Basically, the language model will have the UK words in it. And as long as those words are represented with that spelling in the speech recognition dictionary, they will be recognized.

Those words, the spelling changes there are relatively infrequent across the document. So even if there are words in the transcript which are not represented in the dictionary, the speech recognition won't get that particular word that realigns staff where the speech recognition is aligned with the original transcript will force those words into the appropriate

place. So as long as it's not a long sequence of spelling changes, then it shouldn't be a problem.

JOSH MILLER: Great. Thanks Roger. Question's come up about pricing and just how much the service costs. Everything is duration based, both our transcription and captioning services, as well as the alignment service is all duration-based. We do have volume discounts.

But the full breakdown is available on our website. So definitely we encourage you to take a look at that. It's probably easier to take a look at that full schedule than to go through it here. Another question is with regards to once the file has been aligned, can we use tools like the interactive transcript? Dave, do you want to jump in on that?

DAVID ZYLBER: Yeah. You'll be able to use all of the same features, tools as you would had you used the standard service, including translation. So you'll be able to build a captions plug-in, interactive transcript. You'll be able to submit the file for translation. You'll be able to download your captions and transcript files as many times as you want. There's no limit to that.

Once you pay to have a file aligned, you can do as you wish with it. And there's really no difference. It's more just the process of uploading.

JOSH MILLER: Great. Thanks. We are going to wrap it up with that. We've gone a little bit over time. So thanks everyone for your questions. I really appreciate your time today. We are going to put an archived version of this up on our website. And you'll also be emailed with a link to that. So feel free to reach out to us with other questions. And I look forward to speaking with you.