M+E JOURNAL

It's Showtime!

Innovation explodes across every workflow as technology emerges from the pandemic.

Where are you in this accelerated evolution?

DIVERSITY & INCLUSION
In the office, behind the camera, and on the screen, diversity is crucial

SECURITY
Remote productions create new security concerns, with assets under siege

SMART CONTENT
Artificial intelligence and machine learning are being applied in new, exciting ways

NEW WORKFLOWS
The cloud is delivering on its promise, powering the future of productions

21.01

IS AUTOMATIC CAPTIONING GOOD ENOUGH TO REPLACE TRADITIONAL CAPTIONING SERVICES?

The future of automatic speech recognition technology is bright, despite its current drawbacks

BREAKING NEWS
TENSIONS KEEP RISING, AS THE TWO NATIONS
CONTINUE THESE HIGHLY CONTENTIOUS TRADE NEGOTIATIONS
BREAKING NEWS
LIVE NEWS
CRITICISM GROWS OVER TRADE DEAL
Arms deal raises safety

ABSTRACT: Automatic captioning is on the rise. But is it really good enough to replace traditional captioning services? A look at the ways Red Bee Media has embraced caption automation, and why a little experience goes a long way when it comes to getting the most out of automatic speech recognition.

By Juliet Gauthier, Strategic Product Manager, Access Services, Red Bee Media

I've been watching "Humans" on Netflix. A couple of years late, admittedly, but if lockdown has been good for anything it's catching up on TV shows you didn't notice the first time around. The premise is a good one: humans have created synthetic versions of people to do various jobs in society to make the lives of the real humans easier and more efficient. Because this is TV, stuff happens that asks the viewer to question how capable we are of managing the rise of automation in our day to day lives.

YOU'RE NEVER GOING TO get the best out of automatic speech recognition unless you combine the latest technology advances with every bit of your human expertise.

Automation is something that I deal with every day as a product manager in access services at Red Bee. Access services is a collective term for anything that provides viewers with greater access to media content. Think captioning for hard-of-hearing and deaf viewers, or audio description for partially sighted and blind audiences. We've been providing these services for decades to many of the major global broadcasters in our industry and, like any company that's been around for a while, we've observed and adopted new approaches to our service delivery by taking advantage of technological improvements that emerge. In recent years, the big trend has been the automation of speech-to-text solutions for captioning workflows.

We've used speech-to-text systems for 15-plus years to produce captions for broadcast content. The process is quite simple: broadcaster audio comes in, gets translated into text, text is converted into captions, and captions are sent back to the broadcaster for transmission. In the past, people would assume that you could just plug the audio into a computer, and it would generate captions automatically for you. Yet it was never quite that simple.

Computers weren't great at knowing people's names, or when to punctuate at the end of a sentence so, traditionally, a trained live captioner would need to act as a kind of interpreter, repeating the audio into a speech recognition system trained on their voice, punctuating as they went, and speaking in a kind of robotic monotone to ensure the computer transcribed each word correctly, often live on air. It's incredibly skillful, intensive work. The limited group of people who can do this job well are genuine experts in handling speech-to-text systems.

In the last few years, though, there's been an explosion of automatic speech recognition technology. You can ask Alexa to tell you a joke, or get Siri to tell you the weather, and these advances in speech-to-text automation have benefited captioning workflows as well. Suddenly, we have automatic speech recognition (ASR) engines that know when to put a question mark at the end of a sentence, or how to use phrasal commas. They can transcribe newsreaders almost flawlessly. They're trained on millions of hours of audio and recognize terms like "COVID-19." Combine the best ASR engines on the market with the best captioning toolsets developed by Red Bee, and you have an automatic captioning solution that's accurate enough that you can actually put it on air without needing that trained live captioner to act as an interpreter.

Before we meekly accept the rise of our automatic captioning overlords, it's worth noting that there are still weaknesses in ASR. While it's much better now at understanding how to punctuate, it can be pretty terrible at working out when a new person is speaking. This makes watching an interview quite hard work if you need to use captions. ASR engines also don't do a great job yet identifying non-verbal sounds, like music or applause. They work best on content where speakers say things in full sentences, with limited background noise, because this gives the engine a lot of context to understand the structure of a sentence and more chance of accurately transcribing it. So, using ASR for captioning news programs works great, but when applied to sports content — where sentence fragments, like saying a player's name, are common, and where there is a lot of background noise — the result is usually far from excellent.

Finally, ASR engines are fundamentally still quite dumb. They know what they've been taught, and they don't learn without some sort of human intervention. So, you must teach them new terms if you want them to transcribe them accurately. For news content, with new names and places cropping up every day, several times a day, this is a risk that a captioner just wouldn't experience.

***Juliet Gauthier** is strategic product manager of access services for Red Bee Media, and has worked in accessibility services for the media industry for more than a decade. She started out as a live captioner and operations manager in the UK, before moving into global program delivery, where she established Red Bee's first operational excellence team and flagship U.S. site.*
juliet.gauthier@ericsson.com @RedBeeMedia

Our approach to ASR technology at Red Bee is to combine the best engines with our in-house captioning expertise. We don't develop our own ASR technology (we leave that to the R&D teams at companies like Speechmatics, Amazon and Google) but we do keep track of all the major ASR technologies, and how they perform for captioning use cases, on a regular basis.

When delivering our service, we use the best ASR engine currently available as a foundation, and then apply our own technology, expertise and experience to maximize the accuracy. First, we ask our speech-to-text experts in the captioning teams to train the engines regularly with new terms and vocabulary. There's an art to this; it's not as simple as uploading a list of words because text and audio often don't follow logical pronunciation rules. For example, my surname Gauthier is pronounced "go-tee-ay" but I would guess an English-language trained ASR engine would follow English pronunciation rules and expect it, wrongly, to be pronounced "gaw-theer." Our captioning teams know all of this and know how to get the best results by optimizing vocabulary training to ensure the best chance of an accurate transcription. Second, we apply tens of thousands of bespoke house styles, built by our teams, to improve readability. ASR engines tend to format everything as text: "COVID nineteen," "ten thousand five hundred pounds," "twelve forty-five pm." Using house styles created by our teams, you get "VOID-19"; "£10,500" and "12:45 p.m." A much easier reading experience.

It's fair to say we've embraced automation at Red Bee. Using third-party ASR engines and our captioning experience, we've built a fully-automated live captioning service called ARC that reduces the price of a live captioning service by at least 50 percent, and it's proving increasingly popular for broadcasters on certain types of content in English and Spanish.

Automatic captioning is on the rise, and our priority is to handle that transition responsibly. When I was watching "Humans," it struck me that almost all of the characters are just trying to work out how to combine their human experience with what seems like the inevitability of automation. It's not too far removed from our approach to automatic captioning: you're never going to get the best out of it unless you combine the latest technology advances with every bit of your human expertise. ⊞