Accuracy

SURVEILLANCE EPISODE 1 – CHALLENGES OF VOICE SURVEILLANCE AND THE WIDER ISSUES OF VOICE ANALYTICS

TECH

FEBRUARY 2019 | 4 MINUTE READ

Robert Weston
Partner

1. THE MOST DEMANDING OF COMMUNICATIONS

Monitoring phone calls for traders is the most technically demanding of surveillance activities – for a myriad of reasons: the calls are complex, the quality of the recordings is poor, and, most frustratingly, the technology involved rarely works.

The net result is that regulated industries have, for a long time, avoided deploying voice monitoring solutions, despite the fact that voice communications provide an excellent insight into misconduct. Phone calls are more free-flowing and consequently more incriminating than emails[1].

From a risk perspective, it makes little sense to monitor emails but not phone calls; from a budget perspective, this used to make sense due to the costs involved in monitoring audio. However, technology is now changing and costs are dropping, fixing one issue and creating another!

[1] *Experience in working across numerous investigations has shown that more issues are found on calls than in emails.*

2. WHY IS VOICE TECHNOLOGY SO BAD? ALEXA WORKS!

Many people are, quite rightly, frustrated that technology used to convert voice to text within financial services is so poor. Amazon's Alexa can be bought for £40 or less and is seen as more accurate than a £1m implementation – this seems to be unfair, at best.

The reasons for this are largely historical.

Much of the technology deployed within banks to monitor traders has been repurposed from technology that was designed for the retail industry. The monitored calls in retail banking are generally scripted and predictable conversations, recorded on high-quality headsets in a quiet call centre.

Trader calls are the polar opposite: the calls are unscripted, unpredictable and recorded on low-quality microphones with a very low "bit rate". Most systems are recording at around 8-bit. This is four times lower than a podcast and sixteen times lower than the BBC broadcast standards.

The poor recording quality, combined with the poor microphones and a noisy background, means that it is hard for the technology to discern the different sounds and words, thereby creating a low-quality transcription. In addition, many of the technologies being deployed are using "on-premises" solutions rather than cloud-based platforms. Amazon's Alexa is leveraging not just the massive computing power of AWS but also all of the data that it can access. Alexa constantly updates and improves the platform, supported by millions of users around the world that help drive its deep learning technology. This computing and data power, combined with the fact that Alexa devices have seven microphones in them to ensure high-quality recording and separation of conversations, means that Alexa's capability far exceeds that of the standard, static, on-premises solution.

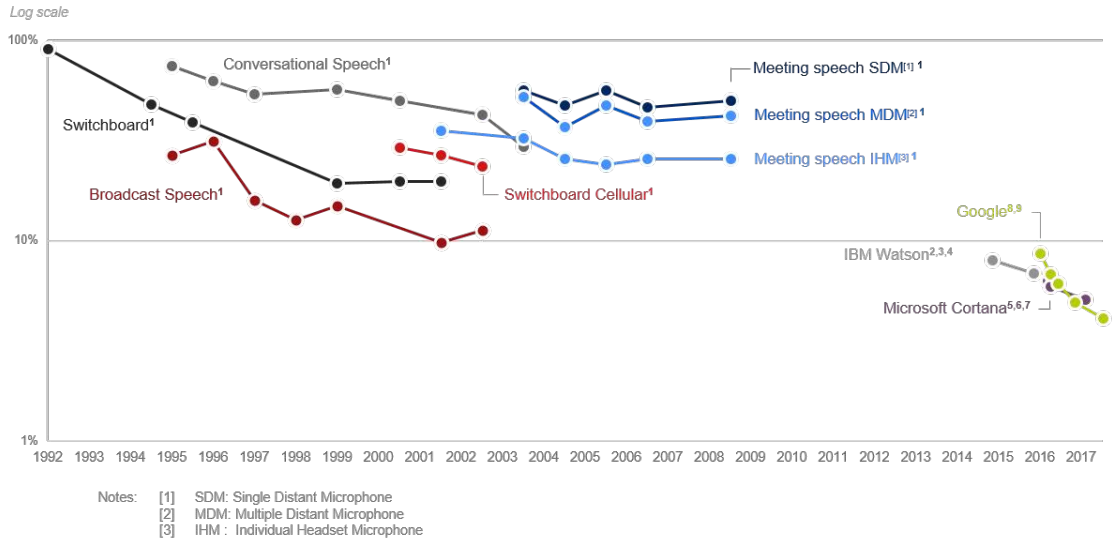
3. IS IT GETTING BETTER?

The accuracy levels for speech-to-text technology have increased dramatically, after being almost static for nearly a decade. From 2000 to 2010, most technology was around 70% accurate at best (see chart below); from 2013 to 2016, this improved radically and hit the 94.1% mark. This is a key milestone as it matches the benchmark of human accuracy.

Since 2016, this leading-edge technology has started to become mainstream, with stable and readily available access to the public.

In short, the real world technology has dramatically improved and can be deployed by banks and insurance firms alike.

Speech-recognition word-error rate, selected benchmarks, %



Notes: [1] SDM: Single Distant Microphone
[2] MDM: Multiple Distant Microphone
[3] IHM : Individual Headset Microphone

4. SO IS IT ALL FIXED THEN?

The accuracy of voice-to-text technology is now the same as – or even exceeds – that of humans. However, this does not mean that the problem has been resolved. Far from it, this is due to two reasons: quality and volume.

The tests and benchmarks are based on a standard “switchboard” data set, not trader calls. The test data is therefore far higher quality than real world trader calls. This difference in call quality means that simply connecting the data to the new voice platforms will not resolve the issue. There will need to be a process to select the most effective platform and then tune it for the specific data set.

The second issue is the volume of data. Even if there were 100% accuracy in speech-to-text transcription, there would also be the same problems of false positives that exist within eComms; all the speech-to-text process will do is provide large amounts of text for analysis. This problem can be addressed through other methods such as cognitive computing to drive the understanding of the subject matter and machine learning to reduce false positives and find more relevant risks. These areas are beyond the scope of this article and are covered in other publications by Accuracy.

5. CONCLUSION

Voice analytics has historically been very poor with low-quality transcription rates and high numbers of false positives. There has been a rapid increase in the ability to transcribe voice to text in recent years with ever improving technology, but this will not solve the problem of voice surveillance.

Voice surveillance requires an understanding of what is being said and why. It’s not just a transcription or search for keywords: it’s about understanding behaviour.

[2] *National Institute of Standards and Technology (2009). "Rich Transcription Evaluation"*
[3] *IBM, George Saon, Hong-Kwang J. Kuo, Steven Rennie and Michael Picheny (2015). "The IBM 2015 English Conversational Telephone Speech Recognition System"*
[4] *IBM, George Saon, Tom Sercu, Steven Rennie and Hong-Kwang J. Kuo (2016). "The IBM 2016 English Conversational Telephone Speech Recognition System"*
[5] *IBM (2017). "English Conversational Telephone Speech Recognition by Humans and Machines"*
[6] *Microsoft, W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig (2016)."The Microsoft 2016 Conversational Speech Recognition System"*
[7] *Microsoft, W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig (2016)."Achieving Human Parity in Conversational Speech Recognition"*
[8] *Microsoft, W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke (2017)."The Microsoft 2017 Conversational Speech Recognition System"*
[9] *Google I/O Keynote (2017)*
[10] *Google (2017). "State-of-the-art Speech Recognition With Sequence-to-Sequence Models"*