Hello? Why can’t you understand me: the challenge of tech pivoting to audio

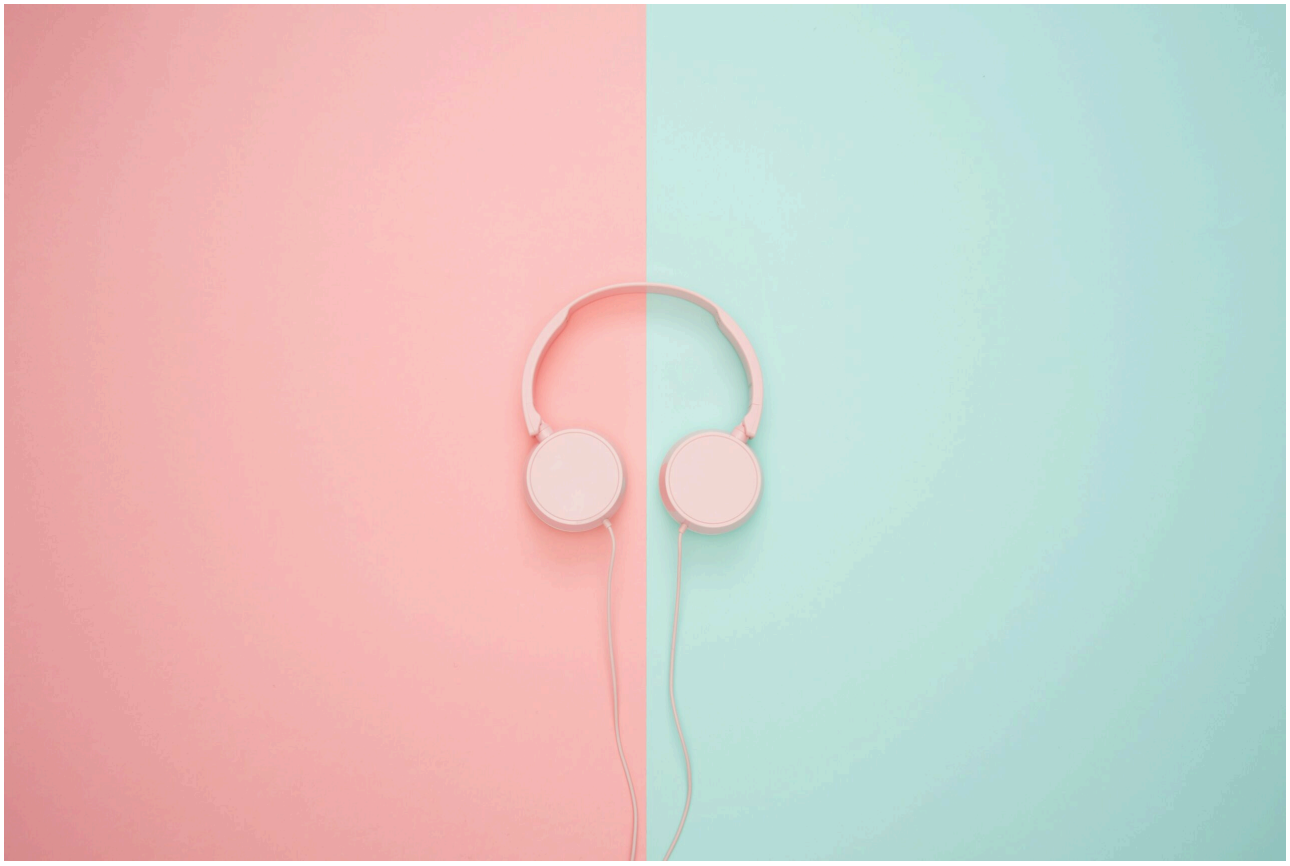
The global pandemic has changed a lot of things about life - the most prominent being the huge reduction in face-to-face contact. We rushed to our Zoom quizzes, constant video calls, and online delivery services to try and keep our lives ticking along. In the process of this, something else changed: we switched to audio.

Temps de lecture : minute

20 July 2021

As traditional ways of connecting to others were lost, we shifted to the most personable form of communication - our voices. As we became inundated with text-based communication – both at work and in our personal lives – audio became critical in conveying that which text can’t always achieve, such as emotion and nuance – the most human aspects of communication. Social media channels and streaming services saw this change in real time, with a massive uptick is consumption of video and audio services. As we were prevented from connecting with each other in the usual way, typing and texting alone were no longer going to cut it.

Look at ClubHouse that was (almost) an overnight sensation - it went from an unknown startup to a tech giant. As of April 2021, the company was valued at over $4B and has had recent appearances from major figures of the tech industry like Elon Musk and Mark Zukerberg. Other social media platforms followed a similar voice-first route: Facebook offering Soundbites, LinkedIn came out with its own version of ‘stories’, and then of course there is TikTok - used by masses of people to share short videos.



Read also
Audio branding: heard any great businesses lately?

This ‘pivot to audio’ has been great for many, but it also raises questions around inclusion. Does technology really understand everyone? Could it possibly? Why do I have to shout at my Alexa to be understood when someone else only has to simply ask? What makes one voice more understandable to technology than another?

So, what’s the problem?

For the most part, the machine learning technology which converts speech to text (speech recognition) is trained on a very narrow dataset that doesn’t account for a variety of pitches, accents, languages and background noise. As a result, huge swathes of the population around the world will be misunderstood because the technology simply wasn’t trained to recognise their existence.

This presents a major challenge to the industry - and for everyone that uses any form of speech, be it calling into a contact centre, trying to watch their favourite Netflix series with captions on, or simply enjoying endless hours on TikTok.

Part of this challenge is that speech data is incredibly rich. Unlike other forms of data, enormous amounts of information can be understood by not only what we say, but how we say it, as well as the pitch and tone we use. In order for this technology to be truly ethical and inclusive, significant changes are required to the way we train the machine learning behind these engines.

To understand why that is so difficult, let's look at the difference between sarcasm and sincerity. Each culture uses language cues to flag these in different ways - it is, for everyone, a learned behaviour. Some people are better at employing and understanding it, others not so much. Now translate that instantaneous understanding your brain does when it hears a phrase, or word, or tone and try and make a machine understand that too. How do you even explain to the machine what is a sarcastic retort and what isn't? Doesn't sound so easy, now does it.

Also, no two people have the same voice - each is a unique biometric, just like a fingerprint. Moreover, not all speech comes with perfect diction and in an accent that is immediately intelligible.

We also need to be aware of the fact that many in society have difficulties with speech - whether it be the hard of hearing, or those with speech impediments. *1% of the world's population (70 million people*) has a stutter, for example. To truly make this technology accessible for everyone, we need speech recognition technology to accommodate everyone, despite accent, way of speaking, tone, pitch, or a myriad of other speech related markers, and provide workable solutions to a wide spectrum of users, not just a majority.

The murky waters of data sovereignty

As anyone working in technology knows, one of the dangers of scale is poor product performance under strain. Being able to include every one of those voices into machine learning algorithms will require diverse datasets.

The tech giants have the resources for this, but lack the processing power. They have a real opportunity to solve the bias problem in machine learning models through democratising access to the vast amounts of public data they hold.

Of course, for companies that monetise data this would seem an illogical move; but there is a middle way, where they could use a version of a data clean room to give the insights a level of anonymity, ensuring privacy but still diversifying the outputs of models.

However, data sovereignty is a murky area: very few consumers know exactly whether their data is being kept privately by a company, is in the public domain or is even being sold to third parties - look at the furore around the proposed NHS data store. Voice is no different and as it continues to become an increasingly coveted source of data, we need to ensure privacy is maintained.

Rules around privacy should be the building blocks of any machine learning algorithms operating in this space, and by balancing these with open-source data sharing, we can safely achieve a compromise where speech recognition becomes more inclusive.

If we all want to be able to ride this ‘pivot to audio’ wave then we need the technology to truly be inclusive for all. This is in no small part due to the lack of understanding around the diversity of voices required for speech recognition models to be effective. ASR companies are striving to

combat these problems, such as the creation of algorithms which do not require all datasets to be labelled: massively increasing the number of sources data can be pulled from and thus the diversity of the data. If the companies operating in this space dedicate more time to including all voices, the impact will be enormous and not just for altruistic reasons – for their bottom lines too.

The pivot to audio is an exciting chapter in tech and society, but we need to ensure every voice is of equal value. We all know how frustrating it can be to increasingly have to raise your voice at Alexa in the hopes that somehow, if you just speak a little louder, it'll finally understand you. Let's make it possible that everyone is understood no matter what.

David Keene is CMO of Speechmatics, he has an extensive track record in B2B SaaS, fintech, and the cloud, and has held several senior roles for global B2B software and service vendors such as Google, Salesforce, and SAP to agile scale-ups such as Deloitte Fast 50 ecommerce marketplace.

Article by David Keene