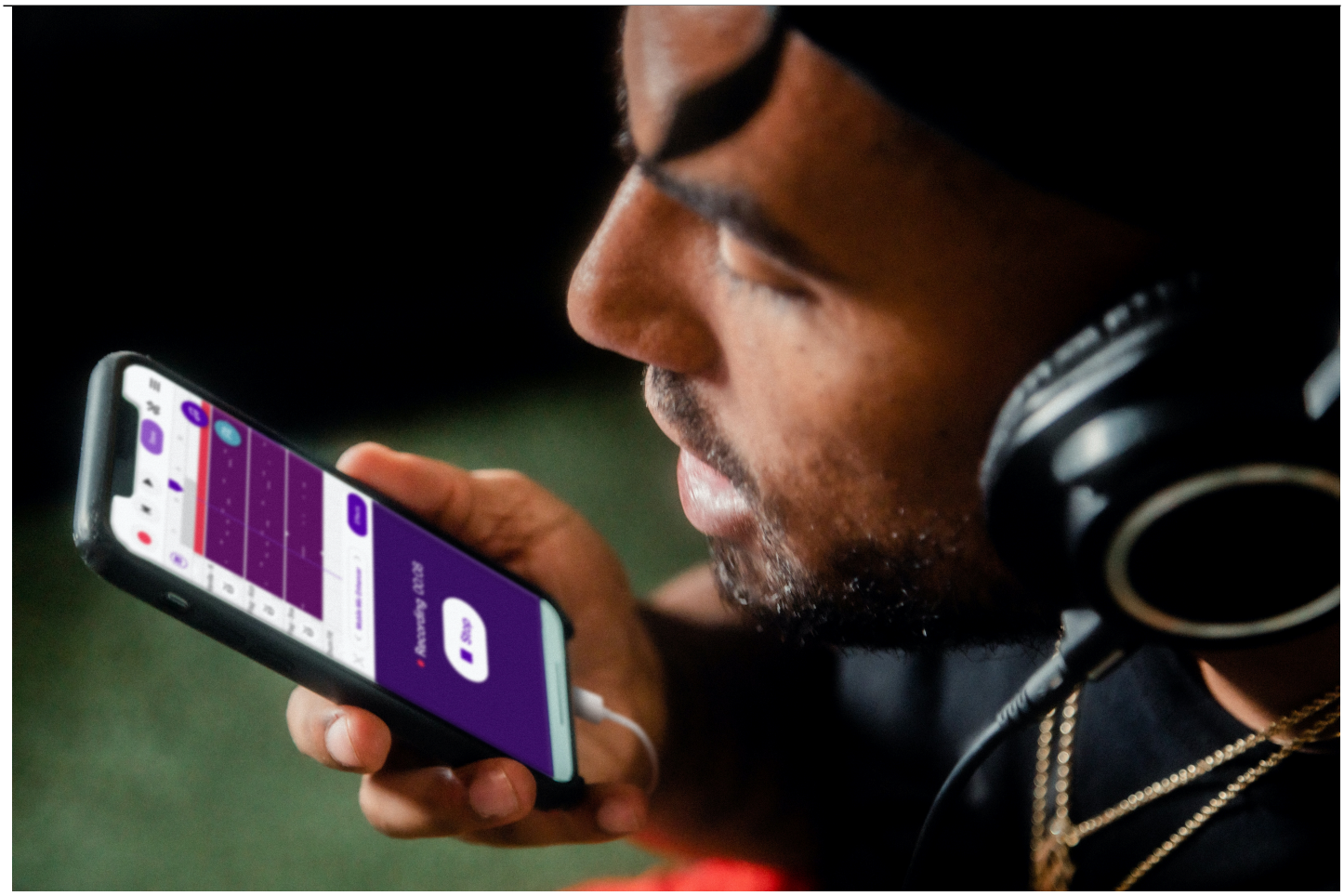
??????????????

??????????????

???Sergio Bruccoleri



Synthetic voice is evolving rapidly. Over the past few years, businesses learned how to use artificial intelligence (AI) to generate synthetic voices for applications such as corporate videos, digital assistants, and video-game characters. Oftentimes, businesses have relied on text-to-speech (TTS) services that can convey written text to spoken word with the right tone and voice to reflect a brand's personality. The goal is to create spoken content such as narration faster and less expensively by relying on machines instead of actors. But one of the challenges of TTS has been creating synthetic speech that re-creates nuances in how human beings speak, such as tone. But as a recently published MIT Technology Review article discusses, advances in AI are making those limitations go away.

Synthetic Voice in Context

Until recently, businesses were satisfied with TTS applications in which the voice sounded largely robotic. So long as the TTS application did what it was supposed to do, nuances such as the tone of

voice did not matter or required extensive editing work through standardized markup languages such as SSML to improve results. But advances in AI made it possible for voices to sound more human with all the nuances of speech that we associate with how real people talk. This branch of TTS is called neural text to speech.

Businesses are applying TTS in some categories where traditionally real human voices were used, such as in tutorial and advertisements. For example, as I discussed in a 2019 blog post, in 2017 KFC celebrated National Fried Chicken Day by reinventing the drive-through experience with a simulation of KFC's international icon, Colonel Sanders.

During the campaign, a voice-based Col. Sanders head gave drive-through customers a humorous experience of ordering from Col. Sanders himself. The experience used speech recognition, AI, and TTS and to make a KFC drive-through operator's voice sound as though Col. Sanders were speaking in a southern drawl evoking KFC's Kentucky roots. Here, TTS helped inject personality and humor into a global brand by enabling a playful experience.

As the MIT article points out, between 2017 and 2022, synthetic voice has made even more impressive strides. Improvements with deep learning have made it possible for synthetic voice to convey many of the subtleties of human speech. Voices pause and breathe when a listener would expect them to, and they change their style or emotion. Just as impressively, synthetic voices (unlike a recording of human voice actor) can update their script in real time, which creates opportunities for personalizing the spoken word for different audiences and applications.

Notes MIT Technology Review, "This opens up the possibility of adapting ads on streaming platforms depending on who is listening, changing not just the characteristics of the voice but also the words being spoken. A beer ad could tell a listener to stop by a different pub depending on whether it's playing in New York or Toronto."

The ability to collect data faster and make sense of it is crucial to the development of TTS. Because AI-based voice applications can collect more data, it is now easier TTS to evolve to *neural* TTS, or one-of-a-kind customized synthetic voice for your applications. Amazon and Google are among the leaders moving toward neural TTS that produces high-quality voices. Neural TTS uses real data collected about emotions are expressed in different contexts in order to mimic those emotions very closely in a synthetic voice application. The synthetic voice industry is also moving beyond TTS to rely on voice cloning – or using voices to generate voices.

However a business develops synthetic voice, context is key. It's possible now for a virtual journalist to report the news by adapting their tone depending on the nature of the story they are reporting – funny, sad, urgent, or somewhere in between.

As neural TTS gets better at adapting to context, neural TTS can more readily understand which emotions need to be conveyed. In other words, a machine can learn from the pattern of how human beings speak (such as a human news caster) to adapt speaking patterns on the fly.

Synthetic Data Sets

Another type of approach for developing synthetic voice, synthetic data sets, has been gaining momentum. Synthetic data sets involve the use of synthetic data for audio, images, and text, which

people use to help train voice recognition AI, optical character recognition, and natural language processing models. All this makes it possible for an AI application to learn faster and more accurately.

Synthetic data mimics what real data would look like, but smart engineering and/or AI with humans in the loop are used to make the data instead, either starting from “hints” of good data from AI, or already collected, and then engineering it to obtain the expected result. For example, Centific recently helped a client rely on synthetic data sets for a voice application to learn about new and evolving concepts such as electric vehicles to provide more relevant results for a voice assistant that answers search queries.

Synthetic Voice Applied

Synthetic voices matter to businesses for many reasons. One of them is the growing importance of sonic branding, or differentiating a brand through sound. One category of sonic branding involves businesses creating chimes or musical interstitials, such as the little ta-dum sound that any Netflix viewer recognizes instantly when they stream Netflix. In addition, KFC’s Harlan Sanders character is an example of sonic branding via voice. For years, brands have hired famous actors to provide narration for ads, which creates a sense of familiarity and imparts a desired tone. But actors can be expensive, and their voices usually have a limited shelf life for commercial use. A synthetic voice offers an alternative.

“If I’m Pizza Hut, I certainly can’t sound like Domino’s, and I certainly can’t sound like Papa John’s,” according to Rupal Patel, a professor at Northeastern University. “These brands have thought about their colors. They’ve thought about their fonts. Now they’ve got to start thinking about the way their voice sounds as well.”

The popularity of virtual people is also stoking a demand for synthetic voices. Virtual people are being used for commercial applications such as introducing products at events and acting as virtual influencers with brands in social media. The growth of the metaverse -- virtual worlds such as Fortnite that rely on avatars for people to interact with each other (and businesses) -- presents a whole new frontier for synthetic voice.

For the metaverse to really grow, someone needs to connect those virtual worlds ranging from Roblox to Fortnite. Meanwhile, synthetic voice can make virtual living through avatars more accessible. If you consider the metaverse as a virtual world where everyone can participate, imagine someone who cannot move from their bed being able to have dynamic and exciting interactions with other avatars. Synthetic voice is important for that experience to evolve in a meaningful way especially when avatars can convey nuances of human speech.

AI Localization

Synthetic voice has a long way to go. One of the biggest challenges for synthetic voice is AI localization, or making AI-based products more inclusive with different cultures. Emotions are local. It’s one thing for a voice application to speak different languages – but quite another to adapt to local ways of talking.

Even with language translation, synthetic voice is in the early stages of learning – an example being mastering different dialects in Spain, Italy, the United States, the United Kingdom, and so on.

In addition, synthetic voice needs to adapt to each language and culture simultaneously. Voice apps such as Alexa are still learning how to do that. Only by collecting data and how voices sound to local people can businesses make synthetic voice more inclusive globally, and you need someone with a knowledge of local nuances – not just the words you use but how you convey those words in the right tone.

This is where people come into play. A global pool of diverse people is the only way to effectively train neural TTS to adapt to different cultures. Humans understand the emotional intent of content, cadence of speaking, and nuances of local cultures.

As an Italian, I can teach an Italian model how to speak better than an English-speaking person can. The job requires more than translation. You need a native language speaker to go beyond translation and convey emotion as well as slang. There might be a slang in Italian that comes across as joking, but if you say that slang in another country, it may convey the wrong emotion or meaning. Humans need to be in the loop to do that.

How to Get Started

Many businesses are interested in synthetic voice but are not quite sure how to get started. They're hearing about other companies succeeding with synthetic voice, understand intuitively its value, but are not sure where synthetic voice can play a role. In branding? Employee training? And where else?

These are understandable questions. Tools such as design sprints can help. A design sprint consist of four-day test-and-learn process in which a team identifies a business problem with no clear-cut and easy solution and develops a prototype for a solution. For example, a business might ask, "How might we improve customer loyalty with AI technologies such as voice?" We use design sprints as part of our FUEL methodology for unlocking innovation.

Many other businesses have moved past the "how do we get started?" and are asking "How do we get better?" For those businesses, we also tap into our global AI expertise – including AI localization -- to help businesses improve. We combine both a diverse, global team of people, techniques such as training with synthetic data, and a platform (OneForma) to scale AI-based applications such as synthetic voice.

Contact us for more insight on how we can help you.

Photo by Soundtrap on Unsplash

-
-

-
-
-