Natural Language Engineering (2022), **28**, pp. 401–408
doi:10.1017/S1351324922000146

CAMBRIDGE UNIVERSITY PRESS

INDUSTRY WATCH

The voice synthesis business: 2022 update

Robert Dale*

Language Technology Group, Sydney, Australia
*Corresponding author. E-mail: rdale@language-technology.com

(Received 9 August 2021)

Abstract

In the past few years, high-quality automated text-to-speech synthesis has effectively become a commodity, with easy access to cloud-based APIs provided by a number of major players. At the same time, developments in deep learning have broadened the scope of voice synthesis functionalities that can be delivered, leading to a growth in the range of commercially viable use cases. We take a look at the technology features and use cases that have attracted attention and investment in the past few years, identifying the major players and recent start-ups in the space.

1. Introduction

Humans have been fascinated by the idea of making machines sound like humans for quite a long time, going at least as far back as Wolfgang von Kempelen's mechanical experiments in the second half of the 18th century.

In the modern era, early attempts at computer-based speech synthesis were already appearing in the 1960s and 1970s, and the 1980s saw the arrival of the DECtalk system, familiar to many as the voice of Stephen Hawking.

The outputs from early applications based on formant synthesis sounded too artificial to be mistaken for human speech and were generally criticised as sounding "robotic." Subsequent products based on unit concatenation dramatically increased the naturalness of the synthesized speech—but still not enough to make it indistinguishable from real human speech, especially when uttering more than a few sentences in sequence. However, by the early 2000s, it was good enough to field telephony-based spoken language dialog systems whose conversational contributions weren't particularly offensive to the ears.

Things stepped up a notch with DeepMind's 2016 introduction of WaveNet, the first of the deep-learning based approaches to speech synthesis. The years since have seen the development of a wide range of deep-learning architectures for speech synthesis. As well as providing a noticeable increase in the quality and naturalness of the voice output that can be produced, these have opened the door to a variety of new voice synthesis applications built on deep-learning techniques.

So, given the advances made over the past few years, what does the commercial voice synthesis market look like today? In this post, we look at the applications of the technology that are enticing investors and aiming to generate revenue, and identify the companies that are making news.

2. Text-to-speech as a commodity

2.1 The majors

If you're a developer building an application that needs a text-to-speech capability, you've never had it so good. Most of the big tech players who offer a suite of cloud-based NLP APIs include

text-to-speech in their portfolios. Today, the state-of-the-art is exemplified by synthesis using deep-learning based models, also commonly referred to as neural TTS, with output that is characterized by natural-sounding changes to pitch, rate, pronunciation, and inflection. Many vendors also provide what they refer to as "standard" voices that use lesser-quality concatenative synthesis at a lower cost.

For example, at the time of writing, Amazon's Polly will generate synthetic speech for you at a cost of US$16 per million input characters for neural voices; you can also use standard concatenative voices at US$4 per million input characters. Google price-matches with WaveNet voices at US$16 per million characters and standard voices at US$4 per million characters. Having deprecated their concatenative voices, Microsoft and IBM now appear to offer only neural TTS, at US$16 per million characters and US$20 per million characters, respectively, although IBM's charging structure actually makes it cheaper than the others at lower volumes. In each case, unit costs drop as volume of usage increases, and all the vendors offer a fairly generous free tier before pricing kicks in.

These numbers might make the services sound cheap to use. But, as an example, suppose that you're building an email reader: bear in mind that an average email message is around 3000 characters in length, so reading just one email using neural TTS will cost about five cents. That's going to mount up quickly if you get fond of having your entire mailbox read to you.

2.2 Choosing a voice

If you're in the market for this technology, how do you choose between the various offerings? The lack of significant price variation noted above suggests that cost isn't likely to be a major factor in making a choice between vendors.

An important consideration, of course, is the quality of the output, so you'll want to try out each company's voices to see what they actually sound like. All the vendors have easy-to-access demos (some require you to create a free account) where you can paste in text and hear what it sounds like across a variety of voices and languages, typically with control over parameters like speed and pitch.

Other considerations that might factor into your choice are the number and range of voices and languages supported by each vendor. At the time of writing, Microsoft appears to be in the lead with 270 neural voices across 119 languages, and Google in second place with 220+ voices (90 of which are neural) across 40+ languages and variants. But every vendor's portfolio is growing over time, so those numbers may well be out of date by the time you read this, and someone else might be in the lead.

You'll also want a voice that is a good fit for your application. For many languages, vendors will typically offer at least both a male and a female voice. Not surprisingly, the number of voices on offer depends on the size of the market. US English is particularly well-supported: Google offers 20 different US English voices, Microsoft offers 15, Amazon offers nine (of which three are child voices), and IBM offers seven.

Selecting a gender for your application's voice may carry the risk of reinforcing stereotypes. Currently, none of the big players offer what might be considered nonbinary or gender-indistinct voices, but this is likely to change. Back in 2019, a team of Danish researchers developed a gender-neutral voice called Q; in 2020, Accenture collaborated with CereProc to create Sam, a nonbinary voice that was subsequently made available open source; and earlier this year, Apple announced that Siri would have a less gendered voice option in iOS 15.4. So we might expect to see explicitly gender-neutral offerings in the majors' catalogues before too long.

The specific characteristics of each voice will determine what fits best for any given context of use. Both Amazon and Microsoft offer voices that are optimized for news reading, reflecting the popularity of this use case. Microsoft's description of its Chinese voices helps you along by noting that some voices are optimized for particular scenarios, such as customer service or story

https://doi.org/10.1017/S1351324922000146 Published online by Cambridge University Press

narrating. For some voices, you can even choose between a speaking style that is cheerful vs. one that sounds more empathetic.

2.3 Other players

I've focussed so far on TTS APIs from the big tech players with well-known cloud-based NLP services, but there are of course other companies active in this space.

In particular, if you need your app to speak Chinese, iFlyTek and Baidu have TTS APIs.

And there are also, of course, many smaller companies whose sole focus is speech technology. Among the more established, those offering developer-oriented APIs or SDKs for integrating their TTS solutions include Acapela (founded 1997), Animo (Japanese; founded 1994),CereProc (founded 2005), iSpeech (founded 2007), ReadSpeaker (founded 1999), Vocalware (founded 2012), and Vonage (founded 2001). We'll introduce many of the start-ups in this space further below when we discuss particular applications and use cases, but relatively new companies offering TTS APIs worth a look are Aflorithmic (founded 2019, US$2.2m in funding to date),[a] Coqui (founded 2021, provides an open-source TTS library), and WellSaid Labs (founded 2018, US$10m in funding to date).

3. Making a voice

With all that choice, it shouldn't be hard to find a voice that sounds right for your use case. Still, if you are Domino's Pizza, you don't want to take the risk that your voice-driven automated pizza ordering service might sound exactly like the mom-and-pop pizza store down the street because you coincidentally chose the same stock voice; your voice is part of your brand, after all. So what do you do? The solution, of course, is to create your own synthesized voice, guaranteed to be unique.

The big tech players will help you with that too, fine-tuning their existing models with audio data that you provide so that you can create your very own voice clone. The services on offer vary largely in terms of how much support is included in the process. Microsoft appears to have the most developed self-service offering here: it provides a set of custom voice creation tools that allow anyone to create voice samples and train a model for demonstration or evaluation purposes for around US$50. Creating a higher-quality production-ready voice requires more data and more compute time, for a cost of up to around US$5k; you'll also have to commit to a number of conditions around the responsible use of AI.

Google also offers a Custom Voice service, whereby you can use your own high-quality audio recordings to create a unique voice. The company uses a review process to ensure that your use case is aligned with the company's AI principles and that you can demonstrate voice actor consent. The website notes that it takes Google several weeks to train and evaluate a custom voice model.

Amazon's Brand Voice feature appears to be a higher-end service that involves working directly with the company's researchers and linguists; at the time of writing, the Amazon website identifies Kentucky Fried Chicken (KFC) Canada and National Australia Bank (NAB) as happy customers.

IBM also lists custom voices as a "premium capability," but I found it hard to locate much information on the product; in its current form it appears to have more of the flavor of Amazon's bespoke service, rather than the more self-service approach taken by Microsoft.

Of course, to clone a voice, you need training data, the acquisition of which is not trivial. A typical estimate is that you need 1–2 hours of professionally recorded audio. That's a lot, given that periods of silence and pauses don't count. Consequently, there's a bit of competition around the amount and quality of data you need to produce an acceptable voice clone.

[a] I've attempted to provide the date of founding and amount of funding received so far for each of the start-ups mentioned, where that information is publicly available; in most cases, this data was gathered from Crunchbase in March 2022.

https://doi.org/10.1017/S1351324922000146 Published online by Cambridge University Press

In mid-2019, Microsoft researchers claimed that 20–30 minutes' worth of data could be enough to generate a realistic voice, although the results reportedly sound a bit robotic. Late last year, Baidu announced that its Deep Voice service could clone a voice with just 3.7 seconds of audio samples, but the resulting output sounds a bit on the warbly side.

Most recently, Resemble AI claims to have used just 3 minutes and 12 seconds of recordings of Andy Warhol's voice from the 1970s and 80s to produce the synthetic voice narration used in *The Andy Warhol Diaries* Netflix docu-series, although an unspecified amount of manual tuning was involved.

Which brings us to celebrity voices. For many areas of business, using the instantly recognized voice of someone famous is much more valuable than using the voice of some unrecognized individual—think George Clooney as the brand ambassador for Nespresso. But you can only get a voice actor to do so much. If you need to record phone trees for IVR systems or scripts for corporate training videos, that gets tedious, time-consuming, and expensive, even when using a noncelebrity voice, and prerecording the full complement of responses is out of the question for many open-ended conversational AI scenarios.

Unless, that is, you have a digital twin to do the work instead. We've yet to see a surge of celebrities developing their voices as licensable assets, but it's surely only a matter of time. Demonstrations of what's possible abound: in 2019, Amazon introduced Samuel L. Jackson as its first celebrity voice replacement for Alexa, and Dessa (founded 2016; US$9m funding, acquired by Square in 2020) created a Joe Rogan voice; in 2020, Samsung added three South Korean celebrities as Bixby voices; in 2021, MSCHF used a synthesized version of rapper Gucci Mane's voice to read a range of classic texts from *Little Women* to *Beowulf*, Amazon introduced the voice of Bollywood star Amitabh Bachchan as an alternative to Alexa in India, and Disney added the voices of characters from the Disney universe to its TikTok TTS feature.

It won't be too long before you'll be able to have a conversation with George on a coffee machine near you.

4. Use cases

These improvements in voice synthesis capabilities have led to better versions of existing applications, but also new applications that previously hadn't been thought of. Here are some of the applications areas that have been enlivened by technology developments in the past two or three years.

4.1 Long-form reading

The most obvious use case where neural TTS overcomes previously troublesome quality issues is in reading long texts, where long might mean anything more than three or four sentences. Earlier approaches resulted in flat and repetitive-sounding text, but the newer approaches are much more convincing: compare the Read Out Aloud feature in Adobe Acrobat Reader to the Amazon newcaster samples in this article in *The Verge* to see how things have improved.

As a result, we've seen lots of applications where TTS is used for what we might call long-form reading. In late 2019, Microsoft introduced a "Play My Emails" functionality in its Outlook for iOS and Android apps; within a year, long-form reading was everywhere. In 2020, Google announced a "read this page" capability in Assistant, enabling the reading of web pages; Cerence introduced Cerence Reader, focussed on reading you news in the car, with reading styles adapted to content type; BBC Global News released a synthetic voice, built in collaboration with Microsoft, to read articles to visitors to the BBC website; and *The Washington Post* announced that it would use natural language generation and TTS to provide audio updates for the 2020 US election.

https://doi.org/10.1017/S1351324922000146 Published online by Cambridge University Press

Today, the technology is within the reach of every blogger: both Amazon and BeyondWords (founded 2016; US$196k funding), for example, offer WordPress TTS plug-ins so your visitors can listen to your posts rather than read them.

4.2 Audiobooks

Some usage scenarios call for what we might think of as real-time on-demand long-form reading, where the synthesized result is consumed immediately upon production, with no scope for correcting the occasional glitch in the TTS engine's output. That's going to be the case, for example, if your application is intended to deliver voiced versions of breaking news in a timely fashion. But we can contrast this with offline voice synthesis, where the production timeline means that it's possible to interject a human in the loop, who can tidy up and fine-tune the results of voice synthesis when it goes wrong, or is just less than optimal.

The ability to combine varying degrees of human support with automated TTS opens the door to new business models. Suppose that you've written a book and would like to make available an audiobook version. Previously, this would have been cost prohibitive for most authors, but that's no longer the case. Amai.io (founded 2019; US$600k funding) advertises an audiobook publication services where completely automatic audiobook creation costs US$99, or you can pay US$498 if you want some editorial assistance—exactly what this amounts to isn't specified, but from listening to the "before" and "after" samples on the website, it sounds like it involves some degree of cleaning up the synthesized output.

DeepZen (founded 2017; £2.8m funding) also operates in this space, offering a more fleshed-out Publisher Portal that provides audiobook publication services. Their automated service at US$69 per finished hour involves a pronunciation check and returns your book's audio in 3–5 days. A managed service, which provides a bit more human review, is US$129 per finished hour and takes 1–2 weeks. The company also provides access to a distribution network at additional cost.

Other companies providing audiobook synthesis include Speechki (founded 2019; US$750k funding), who offer 340+ synthetic voices across 77 dialects and languages by aggregating voices from the majors, with an average production cost of US$800 per book; and VoCoCraft, whose SyntheticAudiobook service will deliver an MP3 file from your epub book for US$299 in three days or less. VoCoCraft also offers a range of services for turning your book into a podcast. Google itself has an experimental auto-narration for audiobooks service that includes publishing your material on Google Play Books.

4.3 Voiceovers

There are a number of application scenarios where, for one reason or another, you want to be able to edit and polish the output of a voice synthesis engine yourself, rather than by using the services of a third party. For example, you might be producing a voiceover for a video or a slide presentation based on a written script. To meet this need, a number of companies combine a script-driven voice synthesis capability with an audio editing tool.

LOVO's (founded 2016; US$4.9m funding) Studio editing tool is a good example of the feature sets on offer: you choose from a library of 180 voices in 33 languages, upload your script into the editor's workspace to synthesize the audio, and then fine-tune the results by using a simple editor that lets you change the speed of the delivery and add emphasis and pauses. The company also offers a Voice Marketplace, where voice actors can produce and market a synthetic version of their voice to interested businesses and organizations.

Other companies offering TTS combined with a voiceover editing tool include the already mentioned Amai.io, along with Murf (founded 2020; US$1.5m funding), Replica Studios (founded 2018; US$5.4m funding), and WellSaid Labs (founded 2018; US$10m funding).

https://doi.org/10.1017/S1351324922000146 Published online by Cambridge University Press

But why limit yourself to a synthesized audio presenter when you can combine this with a synthesized video presenter? Synthesia's (founded 2017; US$66.6m funding) video generation platform lets you select one of more than 40 photo-realistic video avatars; you then provide a script in one of 60+ languages, and a lip-synced video is generated within a few minutes. You can try this out via a demo facility on the company's website. Production access to the service starts at US$30 per month; if you don't want to use one of the stock avatars, you can also create a custom video avatar of yourself or someone else, provided you have consent, for a further US$1k per year.

Hour One (founded 2019; US$5m funding) offers a similar video generation capability, along with a facility that lets you become a licensable character in their virtual actors portfolio—effectively a video avatar version of LOVO's Voice Marketplace.

In the cases just mentioned, the provision of editing support for TTS output is the key feature that makes still-not-perfect TTS usable for voiceovers. But that same capability can be used to edit real human voiceovers.

Particularly impressive here is the feature set offered by Descript (founded 2017; US$50m funding), which provides a text-editing interface to audio and video recordings: the tool uses speech recognition to generate a text representation of the audio which you can edit directly, with the changes you make being automatically reflected back in the audio source. So, for example, if you delete a word or phrase from the text transcript, the corresponding audio and video content is removed. The interface provides a number of clever features: for example, you can remove all the instances of a range of fillers like "um," "uh," and "you know" with a single click. And if you need to add new material to the recording, or replace existing content, you can also deepfake your own voice. Then, if you make a mid-sentence change to a recording by editing the script, the company's Overdub tool inserts the appropriate audio, matching the tonal characteristics on both sides of the edit.

4.4 Dubbing

Which brings us to dubbing proper, by which we mean the replacement of an audio track on a video by that audio translated into another language. Research carried out by Netflix a few years ago came to the conclusion that people prefer dubbed programs to their subtitled equivalents; if that's the case, having good quality dubbing, and lots of it, is a competitive advantage in the streaming world.

But dubbing using real voice actors is not cheap: costs for a 90-minute program range from US$30k to US$100k. To address this challenge, a number of companies have built services that combine speech recognition and machine translation with speech synthesis in a pipeline to produce synthesized dubbing.

Papercup's (founded 2017; £10m funding) partially automated video localization service provides a good example here, once again with a hybrid automation/human review process. You upload your video, specify a target language, and the company's teams of native speakers quality-check and fine-tune the automatically synthesized speech before returning you a video with a translated voiceover. An important feature is that each translated voice sounds like its original speaker. Deepdub (founded 2019; US$20m funding) appears to offer a similar set of capabilities, although their website is short on detail.

Some people find dubbed video content just plain irritating to watch because the lip movements you see on screen don't match up with the translated audio. This problem too is set to go away: in 2020, researchers at Deep Mind demonstrated dubbing technology that not only translates the speech in videos but also translates the lip movements to match the translated audio; you can view examples on YouTube. The technology has already escaped from the research labs: impressive results have been demonstrated here by Flawless (founded 2020), a UK company that is clearly targeted at the big media producers rather than your average YouTuber or podcaster. Watch the showreel on their website to see what can be achieved using their TrueSync process.

https://doi.org/10.1017/S1351324922000146 Published online by Cambridge University Press

You can think of the translation of spoken material from one language into another as a form of voice transformation. There are also other forms of voice transformation enabled by deep learning. Respeecher (founded 2018; US$1.5m funding), for example, carries out speech-to-speech translation that carries across vocal qualities such as whispering or singing, and Sanas (founded 2020; US$5.5m funding) will give your voice a different accent.

And sometimes you might want to use voice transformation to completely change your voice. In mid-2021, Twitter Spaces was reported to be experimenting with a vocal modulator called Voice Transformer, which would allow you change how you sound when talking on the social audio platform. Modulate (founded 2017; US$6m funding), whose key focus is on dealing with toxicity and harrassment in the gaming world, introduced its VoiceWear service, whereby you can replace your voice using a different-sounding voice skin that retains emotion and prosody; the company emphasizes how voice skins can help trans gamers avoid being hassled online.

5. Summing up: The good and the bad

In 2015, actor Val Kilmer lost his voice while being treated for throat cancer. In 2021, Sonantic (founded 2018, €2.3M funding) created a model of his voice using a limited amount of data, thus enabling him to speak again.

Earlier this year, Intel, Dell, and Rolls-Royce introduced a digital tool whose purpose is to preserve the voices of people with motor neuron disease by cloning them. At the I Will Always Be Me website, you provide training data by reading a book, and SpeakUnique's voice banking technology uses the data to create a synthesized voice that can be deployed on an assistive speech device. Google's Parrotron, part of its broader Project Euphoria, helps those with a speech impairment by using end-to-end speech conversion that aims to reproduce the user's intended speech.

These kinds of applications are attention-grabbing, but text-to-speech is a technology that doesn't need to be massively sophisticated in order to produce something that is truly useful. Earlier this year, CVS Pharmacy introduced Spoken Rx, a phone app that will read out prescription labels for those with impaired vision. OrCam won a CES innovation award for its glasses-mounted OrCam MyEye attachment, which reads out printed and digital text, recognizes people, and identifies products for those who are blind or visually impaired.

It's great to see the technology being used in these ways. But as the technology improves, so the potential negatives grow as well. Some of the issues here were brought to mainstream attention with the production last year of *Roadrunner*, a documentary film about Anthony Bourdain, the American celebrity chef who died in 2018. The makers of the documentary used a synthetic version of Bourdain's voice to get him to say three lines that he had written but never spoken. Concerned debate ensued: contrary to the director's claims, Bourdain's wife insisted she hadn't given permission for Bourdain's voice to be used in this way.

Around the same time, voice actor Beverly Standing sued ByteDance E-Commerce, the parent of TikTok, for using her voice without permission.

Voice synthesis companies responded quickly by introducing guidelines around consent. If you visit any of the websites of the companies mentioned above, there's a good chance you'll find material on ethical aspects of the use of the technology, and many are incorporating a means of ensuring that the voice owner's consent to be synthesized has been obtained. BeyondWords and the Open Voice Network have made available an open contract whose purpose is to protect the interests of voice actors: the Voice Services Agreement Template covers a range of commercial and contractual areas, encouraging voice actors to maintain control over their voice IP and earn royalties from its usage.

The biggest worry with high-quality voice synthesis is the potential the technology delivers for the production of convincing deepfakes. In 2019, criminals used a voice clone of a German company's chief executive to demand the fraudulent transfer of €220k from the CEO of the company's

https://doi.org/10.1017/S1351324922000146 Published online by Cambridge University Press

British subsidiary on a telephone call; the British executive said that he had recognized his boss' slight German accent and the melody of his voice. The parties responsible are unlikely to pay heed to ethical guidelines around voice cloning.

In response to these kinds of concerns, Resemble AI has developed a tool that attempts to detect deepfakes, and Google's Version 2 of Translatotron, which translates a voice into another language, removes the previously available ability to generate speech in a different voice. But these initiatives are unlikely to keep bad actors at bay.

When audio is combined with video, deepfakes can be even more convincing. Just take a look at MIT's Richard Nixon, Jordan Peele's Barack Obama or Chris Ume's Tom Cruise. From one point of view these are entertaining, if just a little bit creepy. But they point to the inevitability that we will see increasingly plausible deepfakes of influential people saying consequential things.

In a different but just as worrying direction, PodBot.ai now offers a service that uses GPT-2 to produce the content for completely automatically generated podcasts: type in an episode title, get back a podcast complete with 'made up content, opening music summary and cover photo of someone who doesn't exist'. Sounds like the perfect technology for anyone who wants to advance Steve Bannon's 'flood the zone with shit' strategy.

Scary times ahead. Just don't believe everything you hear.

Cite this article: Robert Dale (2022). The voice synthesis business: 2022 update. *Natural Language Engineering* **28**, 401–408. https://doi.org/10.1017/S1351324922000146