SPEECHMATICS

INDUSTRY REPORT

TRENDS AND PREDICTIONS FOR VOICE TECHNOLOGY IN 2020

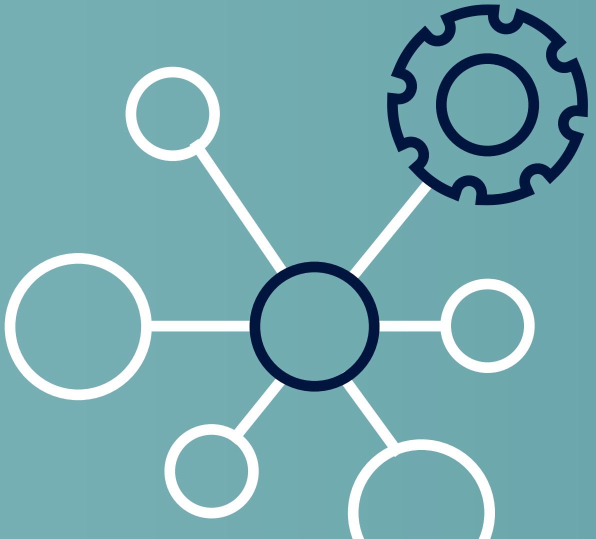
January 2020

www.speechmatics.com

CONTENTS

EXECUTIVE SUMMARY

For years, artificial intelligence and machine learning have been considered as futuristic technology poised to revolutionize how businesses and individuals work, interact and learn. Whereas previously these concepts have been written in PowerPoint slides, documents, PDFs and in WordPress, now they are written in Python and other coding languages. The promises of machine learning and artificial intelligence are now a reality.



These technologies have seen a significant ramp in the quality and value that they can offer to people and businesses. In 2019, more companies and organizations than ever began to understand the value of these solutions and actively looked to integrate artificial intelligence (AI) and machine learning (ML) into their offerings to enhance their customer's experiences and optimize return on investment (ROI). Last year demonstrated that the period of 'hype' around these technologies was over. AI and ML-based solutions were validated by the increased adoption by businesses of all sizes across a wide range of industries.

Machine learning and AI are no longer futuristic technologies, they are here now, and to stay. They have become present in our everyday lives and are already making an impact on the world. We have already seen the emergence, and now adoption of these technologies and it is expected that 2020 will see the maturity of these technologies.

This report will explore the history, development, expectations and trends when it comes to machine learning solutions like voice technology.

It contains key insights from industry experts, product specialists and machine learning engineers at the bleeding edge of these technologies. They reveal their opinions and expectations of voice technology, the markets it will influence, the benefits it will have and the capabilities it will enable.

The report will also look into how machine learning and AI specifically influence the automatic speech recognition (ASR) market. It will address the value and benefits that voice technology can deliver as it continues to evolve and mature, as well as the drivers and motivations behind its adoption. The report will explore the challenges of adoption and key elements that are required to see continued growth.

FOREWORD

VOICE DECADES AGO

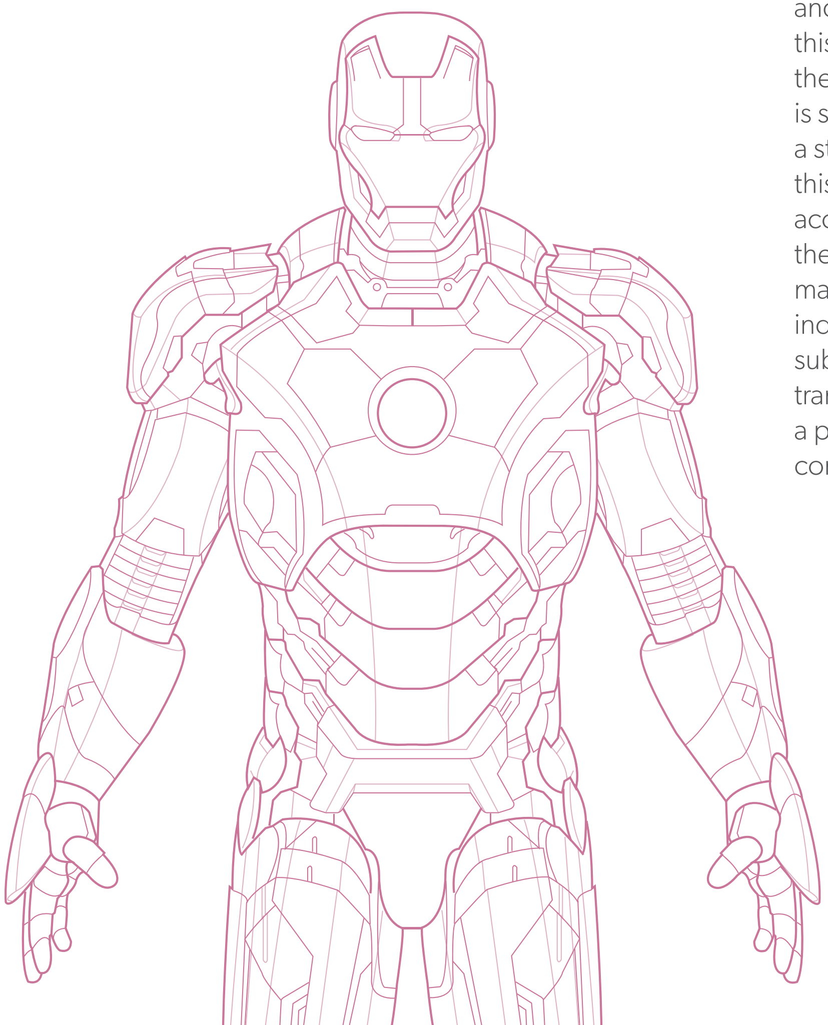
In the last 20 years, voice technology has gone through a complete transformation, not only in the capability of the technology but in its accessibility, diversity and adoption. In its early stages, speech recognition technology had a relatively small number of uses. It delivered little value and was available from a limited number of providers. This could not be further from how the market is today which has penetrated many aspects of our lives. Voice technology is being used by businesses for both consumer use through voice assistants to automate elements of our lives, and for businesses looking to create efficiencies, drive down costs and deliver better customer experiences. This general adoption of voice technology demonstrates the belief and trust that organizations have to place important and secure information in the hands of these solutions.

Speech recognition technology is now in a prominent position in our lives, but this technology is not new. The Voicebot.ai Short History of the Voice Revolution captures all major events in ASR from 1961 to the end of 2019. It shows the length of time that speech recognition has been available (although in very limited ways since the early 60s). In the early days and to some extent even to this day, suppliers of ASR technology were specialized and therefore limited. Speech recognition technology had limited capability which meant that there was little demand from the market and so innovation was slow.

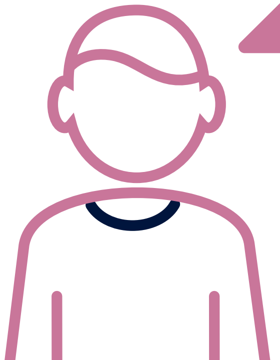
From 1961 to 2008, the real-world capabilities and applications of ASR were limited, ASR was no mystery and it was recognized as a valuable future technology. Automatic speech recognition was prominent in popular science fiction movies and TV shows. The vision of hybrid man and machine solutions and even virtual assistants were commonplace in these mediums, setting the scene for the future. The grounding of ASR in science fiction likely meant that machine learning, AI and voice technology would, even to this day, be seen as futuristic technologies even with its huge adoption today.

In 2008, Marvel released the first Ironman movie (Iron Man Directed by J. Favreau). The film featured a dedicated user interface (UI) called J.A.R.V.I.S. – said to be an acronym for "Just A Rather Very Intelligent System". This was 2 years before the launch of Siri and almost 10 years before the Amazon Echo using Alexa virtual assistant technology. Even though the technology was not present in 2008, the vision of virtual assistants was a very real phenomenon and in 2020 this level of personalization is well on its way to becoming a normal part of everyday life. How these technologies could be used was in scope but even in 2008, the opinion was that these were still firmly rooted as futuristic technologies.

For many, the deployment of Siri on the iPhone in 2010 was a turning point and a realization of the real-world applications of speech recognition technology. However, its release wasn't without its issues. In lots of instances, users would be required to speak slowly and clearly, even adopting neutral accents to be understood. Where previously, languages and recognition models were limited to thousands of words, compromising accuracy, now the success or failure of ASR solutions are firmly rooted in the accuracy that it provides, its ability to understand the user speaking to it and making the correct decision and actions off the back of this. It is no longer acceptable to compromise on accuracy.



So, what changed? Influential figures within speech recognition like Professor Steve Young and Dr Tony Robinson pioneered the approach of applying neural networks to speech recognition in the 1980s. The approach demonstrated that neural networks greatly outperformed traditional systems. Today's computing power, along with the rise of graphics processing and cloud computing, made the huge potential of this approach a reality and got the technology to the point where computers truly understand what is said. The introduction of neural networks was a step-change in ASR technology. The advent of this new approach meant that ASR became more accurate and reliable. Businesses started to realize the value that it could offer and that it could genuinely make an impact on their businesses. One such industry is media broadcast for use in captioning and subtitling. Historically, this has been reliant on human transcribers, however, ASR had been identified as a potential method to increase the speed in which content could be transcribed and consumed.



"Automatic speech recognition as a technology and as a possibility for captioning has been around for a long time without really reaching a credibility threshold. We were aware because we tried to keep up to date with these technologies. Things like deep neural networks, the amount of processing power and the amount of data that can be processed, had seen a bit of a quantum leap in terms of the capabilities of machine learning generally, and these things were having a significant impact on ASR.

"We knew that this technology would disrupt our current model and so as an existing provider of captioning we wanted to take a tactical R&D approach. If this technology is out there, we needed to start getting a sense of it and work out how to build it into our products. So, to a certain extent, we actively looked to disrupt ourselves."

Tom Wootton, Head of Product, Access Services, Red Bee Media

WHERE IS VOICE NOW?

From humble beginnings, voice has seen a significant upward trajectory not only in adoption but also in capability. The ability to deliver quality speech recognition enables organizations to innovate with voice by leveraging speech elements within their business. Voice is being used to add value to more use cases than ever before, from smart speakers and voice bots to machine interfaces in contact centers.

Organizations are looking to more intuitive and engaging methods of interacting with customers while reducing the cost of scaling out their workforce. Until recently, only humans could really understand other humans. However, with advancements in natural language understanding (NLU), the understanding gap is being closed as machines are beginning to understand not just words but intent, meaning and emotion within voice. The deployment of increasingly sophisticated voice-based solutions enables organizations to enhance the experiences of their customers by providing numerous options for customers to communicate with a brand. This method of communication can be dealt with by machines to enhance human's workflows, providing a level of augmentation to save time, reduce complexity and hand off repetitive actions to intelligent automated systems.

When it comes to looking at the most popular uses of voice technology there is no denying the smart speaker market. The rapid adoption of command and control products and smart speakers such as Siri, Alexa and Google Home has driven innovation in the market and increased end-user demand for voice in a way that was previously impossible. The extent of the increased demand is that in 2019, according to Juniper, in an article by Voicebot.ai, there were over 3 billion devices in operation globally.

ASR is being adopted by businesses as a key tool to transform and digitize the billions of hours of captured voice and enable its assimilation into machine learning and AI solutions. Some examples include:

- Call center analytics
- Digital transformation of media content for enhanced searchability
- Building knowledge bases from real-life interactions
- Create best practices

These examples are just some of the ways that organizations are using voice to focus on both real-time engagements and to take owned information and automatically extract a new layer of data, insight and value.

2019 saw significant pressure being put on organizations to protect customer data. Several high-profile examples where this was not done so well were also exposed. For this reason, compliance has never been so important. Voice-based solutions are nothing new in markets like contact centers. For the longest time, calls have been recorded for compliance purposes. Script adherence, protection for customers and the call centers themselves were vital. The advent of solutions like ASR has enabled these organizations to significantly optimize services like compliance. For contact centers, the advancements in ASR technology not only means that they can improve and optimize services that they provide but also add new services that haven't been seen before.

Compliance is an example application for voice. ASR enables the detection of events such as PII data that might slip through call center agent's 'pause and resume' function and be included in a transcript. Automatic detection of this data and appropriate action to remove it not only saves huge amounts of time for compliance and quality assurance staff but also mitigates against potentially huge fines if this breaches data protection legislation.

Not only can organizations understand what was said and who said it, but they can also extract the meaning and sentiment behind voice engagements. Understanding characteristics within speech provides additional value, enabling a better understanding of customers for AI and analytics. It also offers opportunities for real-time scenarios to improve customer service and identify areas for concern. With a shift to automation and machine-based interactions with humans the ability to identify distress, confidence and truth become possible.

WHAT DOES THE FUTURE HOLD FOR VOICE?

Advances in machine learning have led to an extensive acceleration in ASR features, capability and application. Continued improvements in ML has contributed to the creation of more sophisticated and independent machines that can process huge amounts of data and learn on their own with less demand on humans. In turn, enhancements in machine learning help to improve products and services that are powered by ML, such as ASR. Leveraging these ML capabilities enables solutions to deliver more intelligent outcomes. In the case of ASR, the technology offers better accuracy, better language support, identification of elements within voice including intent, emotional cues, and non-voice-based audio (like sounds, music, clapping). These all help to extract more information from voice and deliver relevant and personalized experiences, something that will be seen much more in 2020.

Future applications are not limited to the consumer side of the market. In addition to the massive adoption of smart speakers as mentioned earlier, and voice services being included on almost every smart device sold, businesses are also continuing to invest and rely on solutions powered by machine learning. At the end of 2019, Gartner published its latest hype cycle report (featured in Forbes) within the area of speech. While there are elements of ASR that are still to reach their potential (in both the short and long term), they noted that ASR in its current form has transcended the hype of recent times and now is mature enough that it can be trusted and depended on to deliver real business value.

The ASR market has evolved rapidly in recent years. For example, voice is incredibly important within the media broadcast industry. Government agencies like the FCC in the US have created and enforced legislation around the accessibility of media content online. This means organizations have a requirement to deliver captions for all media content. The need to create and apply these mandatory captions drives demand for ASR improvements in accuracy, speed and language diversity. Similarly, voice is the lifeblood of the contact center. Technologies that can capture the voice of both contact center staff and customers have become vital. Capturing this voice data helps to improve customer experiences, drive down the cost of dealing with the huge volumes of calls and also secures the data of customers without relying on humans.

ASR accuracy is a key consideration for organizations as the technology is adopted more broadly across different markets and industries. Word error rate (WER) is the primary metric to measure accuracy. But with the diversity of languages, accents, dialects and specialist terminology, WER is slowly becoming a false representation of ASR quality and application in the real world. Organizations have their own KPIs to measure success and ASR needs to not only deliver low word error rates but also maintain these rates no matter what language, speaker, accent or subject is being transcribed.

Advancements in ASR mean that speech-to-text output is not limited to just words. More organizations are focusing on human readability of the transcripts and so place importance on non-word-based elements. Punctuation, speaker change and who is speaking are all unrelated to WER but play a massive part in transcript readability and accuracy in the real world.

HOW DOES AI AND ML FIT IN?

The application of deep neural networks to speech recognition made a huge impact on the creation of modern ASR applications. It removed the skepticism surrounding ASR and identified it as a credible solution that can add significant value. 2019 can be seen as a turning point when Gartner called out that speech services like ASR transcended the hype that had been associated with the technology since its conception. Like many technologies that appear on hype cycle reports, ASR is created through the utilization of machine learning processes. While ASR has leveraged advancements in ML to deliver value to a huge range of industries, other technologies that also leverage ML remain unproven. Chatbots are one of these applications and voice assistants another, to a certain extent. Both of these applications are marketed towards the AI space utilizing many separate technologies together to create a solution greater than the sum of its parts. While AI is a growing trend in the technology space, organizations are sometimes keen to use 'AI' as a label to ride engagement when really their level of intelligence is low. At a recent visit to CES 2020, Jeremy Horwitz observed the general misuse of the term AI. In his blog for VentureBeat he commented, **"I can't think of a bigger example of 'charlatan AI' at CES this year than an entire large booth dedicated to fake AI assistants, but there wasn't any shortage of smaller examples of the misuse or dilution of 'AI' as a concept."**

Speech services rely on machine learning for development. In the beginning, this was used primarily to get a voice service that worked in English. Now, as ASR becomes more widely adopted and the technology is validated, there is a growing appetite for increased accuracy, features like the addition of punctuation, recognition of non-voice-based elements, the identification of the speaker and so on. Not only are these features required but the ability to apply them to a greater number of languages, accents, dialects and specific use cases becomes a requirement too. Data is the key to ensuring the continual evolution of ASR and speech services that it contributes to. This will be impossible without machine learning and processes will have to adapt to keep pace with the demand for languages, features, and accuracy.

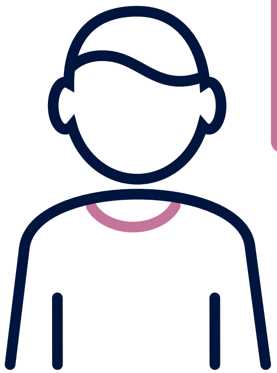
Traditional labeling of data as it is ingested into the machine learning system, known as supervised learning, will soon become unscalable. This will be especially prevalent as the number of labels increases to call out additional elements that are required within the learning process. So, what's the solution? Data is a key factor in the success of a machine learning-powered solution. Having as much data as possible is, therefore, one way to try and ensure success. However, if more data is used, this will require more labeling, which is expensive and time-consuming. The ability for machine learning to rely on self-supervised models and remove the dependence on labeling has the potential to make a huge difference to the time and effort required to further develop speech solutions. If our models can learn unsupervised, this could be a huge step-change.

In a recent interview for VentureBeat, Celeste Kidd, Developmental Psychologist at the University of California, Berkeley, commented that **"Human babies don't get tagged data sets, yet they manage just fine, and it's important for us to understand how that happens."** Moving to methods of machine learning that reflect how children learn without the requirement to label every element that they interact with means that speech models can continue to develop rapidly. Offerings will become more tailored for specific use cases using a relatively small amount of representational data.

As also referenced in the VentureBeat article, Anima Anandkumar, Nvidia Machine Learning Research Director also expects to see progress in the year ahead from iterative algorithms, self-supervision, and self-training methods of training models. These are the kinds of models that can improve through self-training with unlabeled data.

"Artificial emotional intelligence/emotion recognition software using video will start to become more of a theme, this will enable brands to detect emotion from facial expressions, enabling them to customize and target content to users."

Eric Henderson, Product Technologist

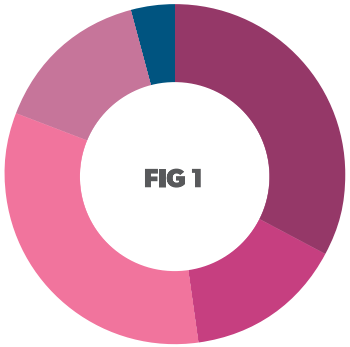


METHODOLOGY AND DEMOGRAPHICS

To write this report, Speechmatics collated data points from Owners/Executives/C-Level, Senior Management, Middle Management, Intermediate and Entry Level professionals from a range of industries and use cases. These people work in industry verticals across the globe including the UK, Europe, United States, Asia and Australia.

FIG 1. ROLES OF PEOPLE FROM WHICH DATA WAS COLLECTED

⬢	**33%**	Owner/Executives/C-Level
⬢	**15%**	Senior Management
⬢	**33%**	Middle Management
⬢	**15%**	Intermediate
⬢	**4%**	Other



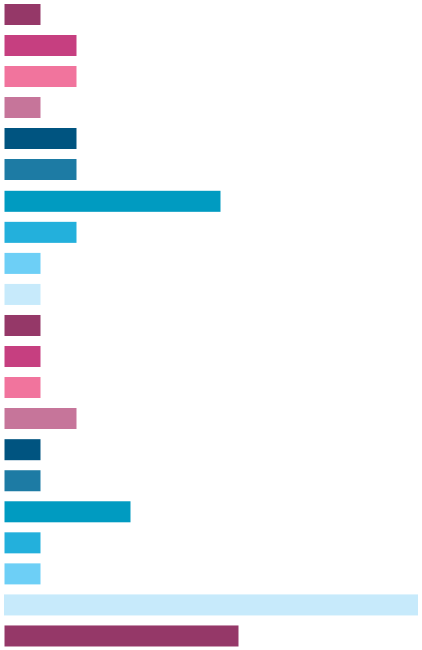
The respondents described their job roles as CEO, COO, CTO, Content Manager, Managing Director, Project Manager, Product Manager, VP Product, Software Engineer, Engineering Manager, Business Development, Journalist, amongst others. The respondent pool included a variety of organizations who operate across industries including contact center, finance, healthcare, legal, education, technology hardware, technology software, telecoms, customer experience and media broadcast.

The collated data encompasses a range of organizations, from large enterprises to smaller startups. 26% of organizations surveyed employ over 1000 people, 5% employ 501-1000 people, 2% employ 251-500 people, with the remaining 67% employing less than 250 people.

60% of these organizations are business-to-business, 13% are business-to-consumer, 22% a combination of the two and the remaining 5% government-to-consumer.

FIG 2. LOCATIONS OF RESPONDENT

Argentina	2%
Australia	4%
Belgium	4%
Canada	2%
Finland	4%
France	4%
Germany	12%
India	4%
Italy	2%
Japan	2%
Malaysia	2%
Mexico	2%
New Zealand	2%
Norway	4%
Pakistan	2%
Russian Federation	2%
Spain	7%
Sweden	2%
Switzerland	2%
United Kingdom	23%
United States	13%



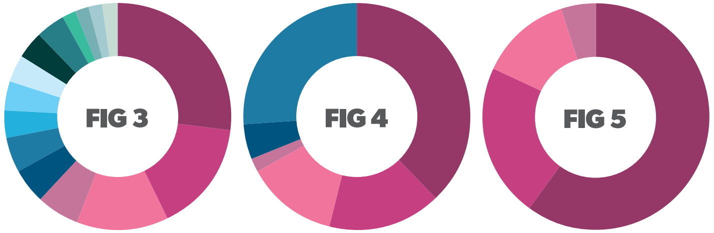


FIG 3. INDUSTRY

27%	Media broadcast
16%	Technology software
13%	Other
6%	Education
5%	Customer experience (CX) and analytics
5%	Healthcare
4%	Contact center solutions
4%	Market research
4%	Telecoms
4%	Transcription solutions
4%	Translation
2%	Finance
2%	Government
2%	Legal
2%	Technology hardware

FIG 4. ORGANIZATION SIZE

38%	1-50
16%	51-100
13%	101-250
2%	251-500
5%	501-1000
26%	More than 1,000

FIG 5. ORGANIZATION TYPE

60%	Business-to-business (B2B)
22%	Mixture of B2B and B2C
13%	Business-to-consumer (B2C)
5%	Government-to-consumer (G2C)

CONTENT STATISTICS

Voice technology is one of the biggest trends of recent times and sees no sign of slowing down. The technology, once depicted as futuristic in movies and comic books, is not only now within the grasp of consumers but is already wildly adopted within a huge range of businesses. The value of ASR through the extraction of voice data from recordings brings new capabilities to early adopters of the technology such as captioning and contact center solution providers. Voice technology has not only proven itself but has driven the adoption of voice-related interfaces, on everyday items within our homes and lives. Organizations (both business-to-business and business-to-consumer) are looking to innovate with voice, enabling their users to leverage their voice to interface in any environment, language or application.

The impact of this surge in adoption is increased pressure on organizations to provide voice support for their solutions and products with increased levels of accuracy, for any language and with the ability to deliver the best possible levels of recognition even in poor audio quality situations.

Markets are not just using voice technology as interfaces to their products. The ability to interact handsfree, to control the world around you with your voice and have your virtual assistant perform tasks for you delivers significant value to consumers. Organizations are also looking to this technology to deliver solutions with better compliance, data security processes and digital transformation.

More organizations that previously had no interest in voice technology are now looking to it as a route to new product, interfaces and points of differentiation. The roll-out and availability of high-quality ASR solutions provide access to rich amounts of voice data. Natural language processing (NLP) tools have grown in sophistication alongside ASR technologies. Text-based bots are commonplace on a huge number of websites. While some are far more sophisticated than others, the technology used to support and drive innovation in this area has seen great advancements in recent years. The combination of these NLP tools and ASR capabilities mean that organizations can not only transform speech-to-text in a huge range of languages but also extract insight from these interactions to optimize their knowledge, better understand their customers or optimize internal business practice efficiencies.

Capturing and transcribing voice not only helps end-users and consumers but has the potential to create a big impact on value for businesses. The cost of miss-communication is high. It’s easy for some level of information to be missed, lost or forgotten – even for internal communications. For this reason, the ability to record interactions can be valuable. Recording, however, does have its limitations. Discoverability of audio data is notoriously difficult. This makes it hard for these files to be indexed, searched and used. Listening to an audio file (if you can find it) in its entirety takes time, concentration and inevitably requires writing or typing elements of that media file. Using an ASR solution enables audio or video files to be transcribed making content easier to find, store and interrogate. It also helps with tracking of data and potentially mitigates the risk of data loss in a simplified format.

AUTHENTIC EXPERIENCES ARE REWARDED AND EXPECTED BY THIS [GEN Z] DEMOGRAPHIC AND SHOULD BE A FOCAL POINT OF YOUR CX STRATEGY. THEY PRIZE TECHNOLOGY.

IN 2020, GEN Z WILL MAKE UP 40% OF U.S AND HAVE A CURRENT BUYING POWER OF

$143 BILLION

IN 2019, AN ESTIMATED 35% OF U.S. HOUSEHOLDS WERE EQUIPPED WITH AT LEAST ONE SMART SPEAKER AND BY 2025 FORECAST SUGGESTS THAT THIS PENETRATION RATE WILL INCREASE TO AROUND 75%

CAPGEMINI RESEARCH INSTITUTE CLAIMS THAT

73%

OF DRIVERS WILL USE AN IN-CAR VOICE ASSISTANT BY 2022

THE SMART AUDIO REPORT CLAIMS

54%

OF U.S. ADULTS HAVE USED VOICE COMMANDS AND 24% USE THEM DAILY

EMARKETER PREDICTS THAT OVER A THIRD OF THE US POPULATION (111.8 MILLION PEOPLE) WILL USE A VOICE ASSISTANT MONTHLY IN 2019, UP 9.5% FROM 2018.

76%

OF BUSINESSES REPORT MEASURABLE BENEFITS FROM INTEGRATING VOICE AND CHAT ASSISTANTS INTO THEIR FEATURES

OVER

58%

OF ONLINE U.S. ADULTS (52.1% OF ALL U.S. ADULTS) SAY THEY HAVE USED VOICE SEARCH AND 33% INDICATE THEY ARE DOING IT AT LEAST MONTHLY

22%

KEEP A SMART SPEAKER ASSISTANT IN THEIR KITCHEN (VIA GOOGLE)

COMSCORE PREDICTS THAT BY 2020,

50%

OF ALL ONLINE SEARCHES WILL BE PERFORMED WITH VOICE SEARCH.

A GARTNER STUDY PREDICTS THAT

30%

OF ALL BROWSING SESSIONS WILL INCLUDE VOICE SEARCH BY 2020.

25%

OF SMART SPEAKER OWNERS KEEP THEM IN THEIR BEDROOM (VIA GOOGLE)

1 OF EVERY 4 AMERICAN HOMES EQUIPPED WITH WI-FI OWNED A SMART SPEAKER IN 2018, ACCORDING TO NIELSEN

52%

OF SMART SPEAKER OWNERS KEEP THEM IN A COMMON ROOM SUCH AS A LIVING ROOM (VIA GOOGLE)

MOST COMMON VOICE SEARCHES ON SMART SPEAKERS ARE ASKING FOR MUSIC (70%) AND THE WEATHER FORECAST (64%), FOLLOWED BY FUN QUESTIONS (53%), ONLINE SEARCH (47%), NEWS (46%), AND ASKING DIRECTIONS (34%)

A STUDY FROM STATISTA.COM SHOWS THAT IN 2018, OVER 34 MILLION SMART SPEAKER DEVICES WERE SOLD IN THE U.S, WITH 2019 SALES PROJECTED AT 36 MILLION UNITS

KEY FINDINGS

1

Some of the **current applications of voice technology** include voice bots, transcription, voice assistants, home appliances, wearables, automotive, captioning, compliance and analytics.

2

The **future of voice** will include conversational interfaces, work assistants, human-robot interaction, wearables and home appliances.

3

73% of respondents said **accuracy is the biggest barrier** when it comes to adopting voice technology within their business.

4

The top two voice applications that will have the **largest commercial impact in 2020** are customer experience (36%) and home automation (34%).

5

Data security, privacy, AI-biases and commercial monopolies are the **biggest risks to speech technology** in the future.

6

93% of respondents indicated that **data privacy** will continue to be a concern in the future.

7

Companies will **overcome the challenges** with data security by using on-premises deployment options and by using solutions that can be operated entirely offline.

8

70% of respondents said that voice is likely to be considered in their **business' 5-year strategy.**

9

The Asia Pacific region will have the **largest growth in the adoption of voice** recognition technology.

10

It is predicted that **voice recognition accuracy will continue to improve** up to 95%, after which time it will not be a commercial priority to improve further.

11

Professionals consider accuracy to mean more than word error rate. People also look at speaker change indicated, intent recognition, punctuation and quick turnaround time for transcription when evaluating providers.

12

73% of respondents expect to see **more robustness when it comes to transcribing voice in noisy environments.**

13

It is expected that **model sizes will be reduced** significantly, enabling more on-device deployments.

14

Larger cloud-based companies will offer **more deployment options.**

15

The future of voice technology development will use **unsupervised learning.**

OVERVIEW OF VOICE TECHNOLOGY USE CASES

From which industries do surveyed respondents come from?

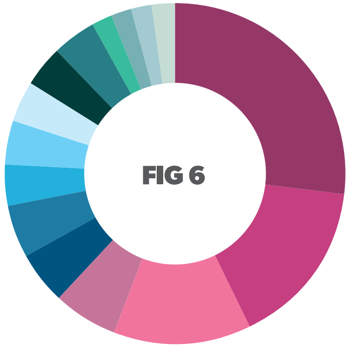


FIG 6. INDUSTRY

⬢	**27%**	Media broadcast
⬢	**16%**	Technology software
⬢	**13%**	Other
⬢	**6%**	Education
⬢	**5%**	Customer experience (CX) and analytics
⬢	**5%**	Healthcare
⬢	**4%**	Contact center solutions
⬢	**4%**	Market research
⬢	**4%**	Telecoms
⬢	**4%**	Transcription solutions
⬢	**4%**	Translation
⬢	**2%**	Finance
⬢	**2%**	Government
⬢	**2%**	Legal
⬢	**2%**	Technology hardware

Voice is becoming widely adopted to support business growth strategies and digital transformation projects. In recent years, the technology has become widespread and consumable, led primarily by the increased adoption of digital voice assistants like Alexa, Siri and Google Home. However, as the technology has matured, it has been adopted more widely by businesses to improve their efficiencies and revenues. Some industries that voice has benefitted include media broadcast, contact centers and transcription solutions.

MEDIA BROADCAST

As spoken content becomes available across more channels, capturing this data becomes pivotal for media companies to manage and monitor valuable voice data. Media broadcast companies are getting ahead of their competition by using automatic speech recognition technology. The technology can be used in media monitoring for setting live triggers on chosen keywords. It is also used in media asset management to transform audio recordings into searchable and indexed transcriptions. It can also be used to form real-time or pre-recorded captioning for use in broadcast scenarios.

Some key drivers and motivations for media companies adopting voice technology include reduced costs, productivity improvements, support/ assistance for human tasks, developing better insights, operational efficiencies, generating competitive advantages and product enhancements.

CONTACT CENTER

With increased expectations from customers comes an increased number of calls to contact centers. There has never been a more important time to think about the customer experience. The contact center market is readily adopting automatic speech recognition technology to improve customer experience through voice data analytics and other strategies. Contact centers can provide unrivalled customer experience using speech recognition to transform customer calls into valuable insights to help with practices such as issue resolution and providing an agent knowledge base.

Turning the customer voice into text enables contact centers to analyze their call content and understand the mood, tone and overall sentiment of customers. This supports continuous improvements in customer experience. The technology helps contact centers with increased pressures surrounding compliance regulation and ensuring the data security of all recordings. Automatic speech recognition capabilities enable contact centers to easily locate and replay stored recordings for a range of applications including best practices, compliance, quality management and event reconstruction.

Some key drivers and motivations for contact centers adopting voice technology include improving the productivity of support staff, increasing customer satisfaction, reducing staffing needs, improving agent related KPIs and knowledge, unlocking insights to empower and drive change.

TRANSCRIPTION SOLUTIONS

More video and audio content is being created than ever before. The need for this content to be transformed into text for SEO purposes, captioning capabilities, or simply to provide text versions of the content becomes ever more useful. Companies are providing transcription services to generate searchable, editable transcripts for their customers quicker and more accurately than ever before. Features such as speaker identification, highlight and comment functionality, adjustable timestamps and a custom dictionary make this process streamlined and efficient. The technology enhances the speed and accuracy at which transcripts are created, taking the heavy lifting away from manual transcribers and enabling them to add value where only humans can.

Some key drivers and motivations for transcription solutions adopting voice technology include taking heavy lifting and mundane tasks away from manual workers, improving efficiency, saving time, reducing cost and speeding up their service.

CURRENT USE CASES FOR VOICE TECHNOLOGY

Industry professionals involved in the research all use voice in numerous ways to deliver business objectives and make day-to-day tasks more efficient, including;

- **VOICE BOTS & COMMAND AND CONTROL**
- **TRANSCRIPTION**
- **VOICE ASSISTANTS**
- **HOME APPLIANCES**
- **CONTACT CENTER ANALYTICS**
- **CAPTIONING**
- **COMPLIANCE**
- **AUTOMOTIVE**
- **AUTOMATIC SPEECH RECOGNITION**
- **ACCESSIBILITY**

VOICE BOTS & COMMAND AND CONTROL

Voice services are used by a diverse number of businesses and organizations to create specific business cases. Voice bots for instance, while similar to virtual assistants, are more focused on delivering specific answers to questions rather than triggering complex and personalized call flows from requests. These are usually referred to as command and control style gestures.

The voice bot market is still relatively new and remains very fragmented. Additionally, the terminology is also yet to be completely locked down and so often the lines between chatbots, voice bots, text bots and virtual assistants are blurred. Some large cloud providers offer the full stack of technologies such as Amazon Lex and Google DialogFlow. Groups of smaller, more specialist providers leverage a range of providers to create bots for market-specific use cases.

Data privacy and the ability to deploy on-premises are key differentiators for providers in the voice bot market and are typically exploited by the smaller providers who have the flexibility to deploy where the large cloud providers cannot.

From 2020 onwards, there is an expectation that there will be a level of consolidation in this market. For example, leading provider Snips.ai was recently acquired by device manufacturer Sonos. There is also the expectation that the large cloud providers will make offerings available that support on-premises deployment.

For the smaller providers, the components of the bots are created from many sources. For example, a bot might be architected from several products from different providers to deliver the desired outcome and functionality. To further complicate things, these bots might leverage parts of multiple, similar products that might perform the same function or deliver the same desired outcome for their specific use case. An example of this might be that a bot is made up of an ASR that also contains the ability to perform an NLP function and a separate NLP product which is preferred.

Smart speakers, like those provided by Sonos but more specifically Google, Amazon and Microsoft have taken the market by storm and are the main companies driving the market forward. Smart speaker sales have seen huge consumer adoption over the past few years. Voicebot.ai call out the scale in which these solutions are available to customers.

Smart speakers that leverage voice bots assistants are used by nearly 20% of U.S. adults.

400 million consumers have access to Google Assistant through Android smartphones.

400 million consumers can access Microsoft Cortana on PCs.

500 million consumers can use Siri through Apple's iOS devices.

The rollout of these services and the integration of voice influences the next generation of younger consumers. The ability to leverage voice to 'command and control' a device like a smartphone makes tasks and content more accessible for younger audiences without the dexterity or written language skills to search using a touchpad. This level of exposure at a young age can establish voice as a key interface further driving the requirement for other devices in the home to have similar interaction capabilities.

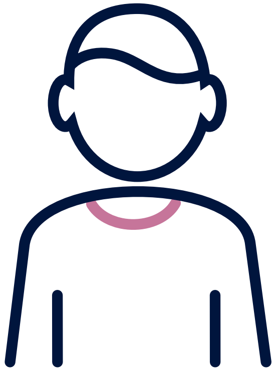
The mass adoption of these technologies and the penetration of smart speakers into the homes of consumers has driven other manufacturers to look to assimilate voice services into their workflow and product. For example, automotive 'infotainment' systems enabling drivers to have better control over their environment while remaining both hands and eyes-free.

VOICE COMMERCE

The use of voice bots is extending also to voice commerce. The links between one of the largest online marketplaces and the manufacturer of one of the most popular smart speakers cannot be overlooked as a significant factor in the growth of the voice commerce market. A marketplace that has almost any product you can think of, synced with personal information like bank details, a smart speaker and next day delivery are a powerful collection of features to create a compelling workflow. The last few years coupled with the growth in popularity of smart speakers as a method to interface with buying products has increased by huge volumes.

> **"One of the keys to success in online retail is truly understanding your customer, from their past behavior to future trends. An example of this is mobile commerce and the importance of smartphones to how shoppers now make purchasing decisions and learn more about brands and services.**
>
> **"By 2023 there will be an estimated 500 million smart speakers globally, 18 billion IoT devices and more than 50% of users believing that digital assistants will help them make purchases totaling $40bn. Use of voice commands to make calls, search the internet, check the weather or play music has quickly become the new normal which presents a massive opportunity for retailers to develop this consumer behavior into revenue via voice commerce (v-commerce).**
>
> **"From telling your smartphone to order the weekly shop, asking your TV to recommend a movie to rent on Friday night or instructing your watch to switch on the central heating, the potential to monetize voice technology is unlimited.**
>
> **"Retailers should embrace v-commerce as they did e-commerce, understanding how it can improve their user experience and how it can support their core business. The user journey through content discovery, recommendations and checkout experience should be reimagined to develop trust with consumers and unlock the potential of v-commerce.**
>
> **Simon Homent**, eCommerce Consultant/ Specialist

The shift to voice-enabled purchasing has implications that run far deeper than just the ability to request a product and expect it the next day. It is likely, in part, due to the full end-to-end offering that is expected to contribute to the $40bn business projected for 2023. In the busy lives of consumers, the ability to order products on the fly with no delivery cost, next day delivery with all billing and delivery information synced provides convenience. If any of these elements were not provided it could have impacted the adoption rate. Additional considerations to the adoption of voice are the impact to visual merchandising and organizations, products and monetization of these services.



COMMAND AND CONTROL

Command and control models within ASR utilize intent to understand the content of a request, and trigger workflows from them. It is expected that specific use cases will require specific intent models customized to facilitate the use of voice for specific use cases. James Page, Product Owner at Speechmatics states that **"The development of intent models will become more automated and more commoditized, with less reliance on hand-crafting of dialogs. Machine learning techniques will be leveraged to generate grammars from real-world data sets"**.

While the big cloud provides like Google and Amazon are well established in the command and control and voice bot markets, the expectation is that organizations looking to integrate voice-enabled and control technology will have specific needs and requirements to enable it to fit into their existing ecosystems and solutions. This will require customization at a relatively deep level to ensure voice capabilities work with the existing solution. Eric Henderson, Product Technologist expects that **"Brands will invest more into developing a voice UX for their customers. They will hit constraints based on the capabilities of Google Assistant and Alexa such as the way the assistant needs to understand the many unique ways of articulating grammatical structures such as dates, financial data or other industry specific terms."** It is for this reason that organizations like Sonos, while continuing to work with big providers like Amazon, are looking to acquire their voice partners and bring them in-house to enable development in specific use cases. It is expected that there will be a surge in organizations leveraging voice and while a large number might be serviced by the big cloud providers, it is impossible for them to tailor their offering for the unique requirements of all providers.

VOICE ASSISTANTS

Voice assistants have been a growing trend in recent years with major technology providers looking to incorporate this technology into existing and new products to enable customers and users to engage with products via voice. Taking engagement beyond touchpads, remote controls and even apps on smartphones or other devices. The expectation is that there is no sign that virtual assistant applications will slow down in market adoption. The assumption is that voice assistants will become even more widespread than before, with consumers seeing their inclusion into household appliances, vehicles, wearables and beyond.

Virtual assistants are essentially the natural evolution of the voice bot, moving from performing specific and simple tasks like ordering an item from Amazon, responding to a question such as "what's the weather like today?", to taking a more proactive approach to interactions. An example would be "Hey Dave, the weather forecast has changed since you asked, I'm now recommending you take an umbrella".

In their current form, both chatbots and virtual assistants both rely on wake words to trigger activation. There is an inherent security issue around this requirement for assistants to be constantly listening for these wake words, potentially recording everything that they hear. While nothing has currently been done to solve this issue, now end-users and organizations are aware of the problem and are being informed of the risks around personal data being captured.

The expectation is that virtual assistants will continue to evolve to better reflect more natural human interactions. This means that the wake words will most likely become less prevalent, further differentiating the chatbot-type command and control model from virtual assistants.

Artificial intelligence continues to remain a popular term within the virtual assistant space. There is, however, a danger that this term will become – or already has become – overused. This could harm future, genuine advancements in AI.

In his blog for VentureBeat about CES 2020, Jeremy Horwitz commented "I found it painfully obvious that tech companies — at least some of them — didn't get the message ["tech companies should spare the world overblown or fabricated pitches of what their AI can do."]. Once again, there were plenty of glaring examples of AI BS on the show floor, some standing out like sore thumbs while others blended into the massive event's crowded event halls".

HOME APPLIANCES

Voice has the potential to be included in almost anything including home appliances and other objects within the home that could be automated. The buzz around IoT has opened the doors to connect anything and everything around us. For example, enabling your fridge to order your favorite beverage whenever the quantity gets low (without pushing a button). With home appliances now able to connect to the internet, this provides the potential for integration with the Amazon Echo with an Alexa skill, with interactions with your washing machine being no more complex than simply saying "wash on 30".

The addition of voice to these products enables more conversational interfaces for mundane tasks and enables tasks to be commanded, elements to be controlled and tasks to be completed by the power of voice. It is expected that **"The ability for B2C enterprises to provide a branded voice experience across a range of devices and situations will become more important, and these organizations will increasingly be concerned with ownership of the voice interaction with their customers"**, according to James Page, Product Owner at Speechmatics.

Back in 2017, furniture giant IKEA announced via its blog that it was voice enabling one of its lights through an Alexa skill. "Earlier this year [2017] IKEA announced the initiative to add functionality to its smart lighting range by enabling people to voice control their lighting with others on the market. From November 1st you can steer IKEA Smart lighting with Amazon Alexa and Apple's Home app."

It is these simple interactions that make sense when leveraging an Alexa skill. However, home technology such as Sonos, demonstrates more complex voice requests. The ever-changing landscape of musicals, songs and albums require quick content development to ensure requests are always understood and the customer experience maintained.

ASR models are trained on natural language data. Like any machine learning system, the engine uses examples of what it has been trained on to deliver the output that it feels best fits the audio that it has heard. While this results in accurate transcriptions, for the most part, uncommon or bespoke vocabulary are unlikely to be included in the training data. For example, brand names, people's names and acronyms are often miss-transcribed. Specifically for music, artist names, song titles, album titles etc. are less likely to be included within standard ASR training data and are constantly changing. Custom dictionary approaches offer the option of adding bespoke new words to uplift the accuracy of the output. This is a particularly good strategy when it comes to voice technology within home appliances as they require the most up-to-date vocabulary to work.

Due to the challenges around the complexity of some specific but high-value use cases, some organizations will no longer be able to rely on the big cloud providers. An Alexa skill will only get these use cases so far and will require constant updates and specialist skills to not only make changes within these interfaces but also in the markets that these services are a part of. The requirements of these use cases might also surpass the capability of writing an Alexa skill and organizations may require direct support from the cloud providers themselves which can be a significant barrier to adoption. Organizations that require more complex voice interfaces, will have to look to more agile providers to facilitate their unique needs. In the future, it will not only be vital to delivering best-in-class speech-to-text accuracy but also to have flexible deployment options, effective support capabilities, and the knowledge and agility of the technology to understand, facilitate and deliver adaptations quickly and effectively to accelerate their partners' and customers' time to market.

AUTOMATIC SPEECH RECOGNITION

According to a Research and Markets January 2020 report into 'Speech-to-text API – Global Market Outlook (2018-2027)', the global speech-to-text API market accounted for $1.32 billion in 2018 and is expected to reach $6.63 billion by 2027 growing at a compound annual growth rate of 20%.

Speech recognition has come of age in recent times and as mentioned in the Research and Markets report, this is in part due to the adoption of smart speakers that have forced consumers to realize that speech recognition has improved significantly. This is no surprise to industry insiders. ASR providers and organizations that have already adopted voice technology to deliver significant value, reduce cost and augment workforces through the introduction of machine processes, enhanced customer experience and customer analytics.

Recent improvements in speech recognition technology fuelled by adaptations and advancements in machine learning, mean that organizations can now innovate with voice. As voice becomes more popular as a means of interfacing, more organizations are looking to voice technology to add layers of sophistication to their products and solutions. Speech recognition is no longer the thing of old sci-fi and has transcended the hype that was previously associated with it. The days of having to speak very slowly into voice products are far behind us. Extensive language diversity also means that speech recognition can be used in more geographies to power global enterprises. By no means is speech recognition a perfect technology, however, it does provide the opportunity to significantly change how organizations can leverage their voice data.

CONTACT CENTER ANALYTICS

MarketsandMarkets predict that the contact center analytics market is expected to grow from USD 634.3 million in 2016 to USD 1,483.6 million by 2022, at a compound annual growth rate of 15.9% from 2017 to 2022.

Contact centers have been at the frontline of adopting new developments when it comes to voice. As an industry and market where this is the primary source of interaction and data, contact centers are motivated by leveraging voice technologies to optimize their businesses. By capturing, structuring, and analyzing data they can understand patterns in data and even predict future outcomes. This comes in a range of forms and the attitude to voice has shifted as consumer expectations have changed and legislation has brought in new rules that must be respected for the protection of consumer data.

Originally, call recording was enough for contact centers to keep track of interactions and ensure compliance. While this provided a solution in the short-term, recordings of audio conversations are hard to index, search and interrogate especially when the calls need to be investigated quickly in a dispute situation. Calls in audio format also require significant storage space. With customer experience being a core factor of measurement in the contact center and unhappy customers resulting in loss of revenue and loyalty for the represented organization, analysis needs to quick and cost-effective.

The ability for contact centers to transcribe their content provides many advantages. Interactions have become easier to index and search if they need to be found quickly. Agents are empowered to significantly reduce the time taken to resolve disputes and the contact center can innovate their solutions by transforming the audio from calls to a text-based format. When in text, call recordings can be added into natural language processing tools that already exist in the contact center to gain insight from omnichannel approaches like text bots, instant messaging and email interactions with customers. The archives of existing call recordings in contact centers are a potential gold mine of data that voice technology can transform into key insights.

CURRENT MARKET ADOPTION OF VOICE TECHNOLOGY

DOES YOUR COMPANY HAVE A VOICE STRATEGY?

50% **of respondents said that their company currently has a voice strategy.**

Surprising only 50% of respondents currently have a voice strategy but as you'll later discover, 70% of respondents will consider a voice strategy in the next 5-years. This shows that the inclusion of voice technology in a business' technology strategy is currently between the early adopters and early majority stage of the product adoption lifecycle. It demonstrates how progressive voice technology is and how much more it is intended to develop and improve. With voice being the easiest form of communication, over the next 5 years as businesses start to fully integrate voice technology into their workflows, there will continue to be a major shift in communication for enterprise.

CURRENT APPLICATIONS FOR VOICE

Respondents indicated that there are currently lots of applications for voice. They are listed below.

- **VOICE BOTS**
- **TRANSCRIPTION**
- **VOICE ASSISTANTS**
- **HOME APPLIANCES**
- **CONTACT CENTER ANALYTICS**
- **CAPTIONING**
- **COMPLIANCE**
- **AUTOMOTIVE**
- **SPEECH-TO-TEXT**
- **ACCESSIBILITY**

It's no surprise that many of the current applications for voice are consumer-based. Consumer voice applications have put voice technology on the map over the past few years. With the likes of Siri, Alexa and Google Home bringing the possibilities of voice communication to millions of devices. Voice technology in consumer devices is currently restrictive and only enabled via short utterances or commands.

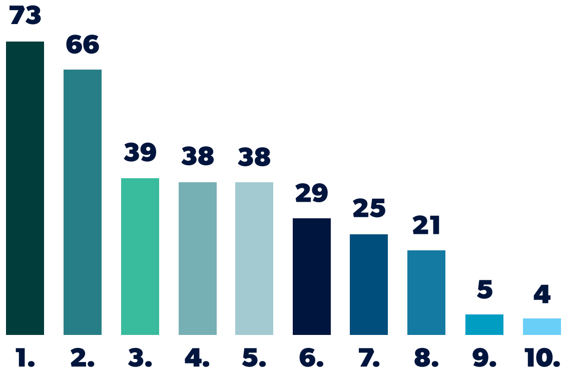
There are endless opportunities for consumer devices to adopt voice, but the vast amount of business applications rapidly adopting voice technology shouldn't be overlooked. There is an opportunity for consumer devices to adopt conversational voice technology like business applications use, but use cases such as captioning, contact center analytics and automotive have already been applying voice technology into their technology stacks and innovations for many years.

BARRIERS TO ADOPTING VOICE TECHNOLOGY

Barriers to entry are commonly the reason why businesses are late to adopt innovative technology. When it comes to voice technology, it has been a common misconception for many years that voice technology isn't good enough to adopt as an integral part of a workflow and technology stack. This just simply is not true anymore. Over the past few years, voice technology has improved to the point at which the output for the most spoken languages in the world such as English, French, Spanish and German is highly accurate in terms of word error rate. It's at this stage where other challenges and factors affect the rate of adoption.

As businesses have started to overcome this misconception and consider voice technology in their 5-year innovation and technology strategies, what other barriers to adoption are proving challenging? And why is accuracy still a problem?

FIG 7. SOME OF THE BARRIERS TO ADOPTION OF VOICE TECHNOLOGY ARE INCLUDED BELOW.



⬢	**1.**	**73%**	Accuracy
⬢	**2.**	**66%**	Accent or dialect related recognition issues
⬢	**3.**	**39%**	Cost
⬢	**4.**	**38%**	Language coverage
⬢	**5.**	**38%**	End-user expectation
⬢	**6.**	**29%**	Complexity of deployment and integration
⬢	**7.**	**25%**	Flexibility to meet the use case
⬢	**8.**	**21%**	Operational support
⬢	**9.**	**5%**	Data security and privacy
⬢	**10.**	**4%**	Other

ACCURACY

73% of respondents said accuracy is the biggest barrier when it comes to adopting voice technology within their business. Accuracy means more than just the accuracy of the word output which is known as word error rate (WER). With the most spoken languages in the world at a consistently low level for WER, there are often many other factors (listed below) that can affect the level of accuracy on a case by case basis. These factors are often unique to a use case or a business' needs.

FACTORS AFFECTING THE LEVEL OF ACCURACY INCLUDE:

- BACKGROUND NOISE
- PUNCTUATION PLACEMENT
- CAPITALIZATION
- CORRECT FORMATTING
- TIMING OF WORDS
- DOMAIN-SPECIFIC TERMINOLOGY
- SPEAKER IDENTIFICATION

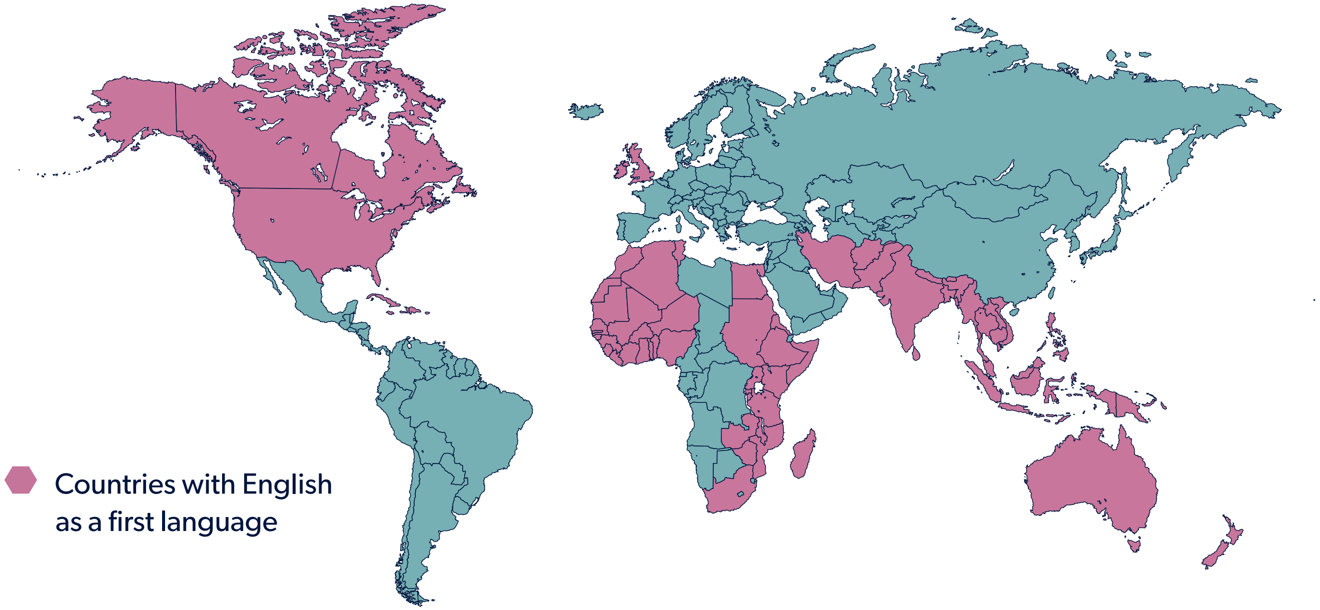
DEPLOYMENT

Deployments and integrations need to be simple. However, **29% of respondents said the complexity of deploying and integrating voice technology is a barrier to adoption.** Whether a business requires deployment on-premises or in the cloud, integration needs to be simple and secure. With many voice technology providers, deployments are far too complex, with little support and documentation offered to make the technology consumable. This makes the process of integrating voice technology within a company's technology stack time consuming, expensive and therefore a significant barrier to adoption.

LANGUAGE COVERAGE

38% of respondents identified language coverage as a key barrier to entry. Many of the leading voice technology providers have a gap when it comes to language coverage. All providers cover English but when global businesses are looking to adopt voice technology into their strategy, the lack of language coverage provides a big problem and barrier to adoption.

Even with some providers offering languages other than English, it is then a question of how accurate are these other languages? The development of accurate language coverage from providers will help businesses to adopt voice technology into their technology strategies.



FUTURE APPLICATIONS OF VOICE

The future applications of voice are potentially limitless. Recent innovation and market adoption of smart speakers have not only pushed voice as an interface into the spotlight but has also engaged the imagination of the consumer. Now, voice is a usable interface for interaction and the market demand is increasing. Consumers are looking to voice to simplify their lives and enable additional multitasking in a hands and eyes-free manner.

Consumers are also becoming more trusting of voice technology applications within businesses. Voice-based IVRs are becoming more common. Better transcription quality fosters better responses and enhanced customer experiences. These better experiences drive adoption, enabling businesses to maintain the trust of their customers while deploying solutions that improve ROI and drive down cost.

Businesses are also coming around to the value that voice technology can offer them when it comes to important compliance and regulatory tasks.

RESPONDENTS THINK THAT THE FUTURE OF VOICE TECHNOLOGY WILL INCLUDE:

- Legal/compliance
- Real-time court reporting
- Voice assistants
- Home assistants
- Automotive
- Work assistants
- Conversational interfaces
- Customer service/experience
- Meeting conferencing
- Human-robot interaction
- Wearables

THE FUTURE OF VOICE APPLICATIONS

Responses show a large diversity of opinions as to where the future of voice can potentially exist. Some industries and sectors such as media broadcast and captioning (who are in part driven by legislation) were early adopters of ASR technology seeing the value of augmenting their human transcription capability and leveraging new technologies to maintain market position and capability. In contrast, some organizations have waited until the ASR technology has matured to ensure good enough accuracy. These organizations limit the potential risk of implementing new technologies into established products or solutions. However, those organizations that view their voice strategies as a long-term, strategic initiative have the advantage of being positioned as an innovator in their field. While speech technology is nothing new, there are still many advantages for organizations that might have never considered a voice strategy before. Organizations can diversify their offering and attract new customers through the adoption of new capabilities delivered through voice.

Increased levels of accuracy across more languages enable businesses to innovate with voice. It gives them the confidence that they can exceed their customers' expectations in any language or use case.

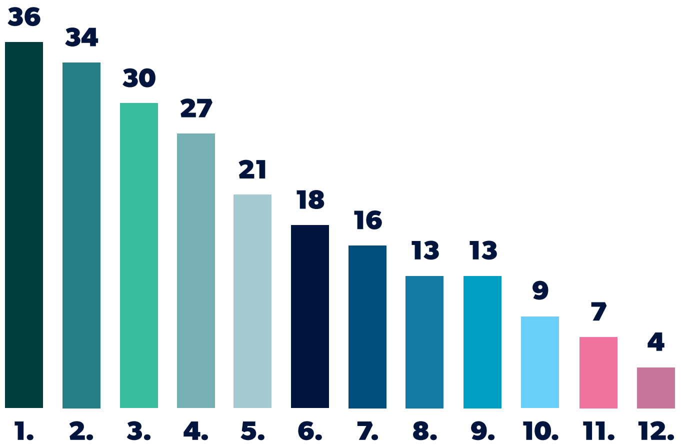
Connectivity also has a role to play in the adoption of ASR for consumer applications, especially outside the home or business environment. Automotive and wearables not only depend on high-quality ASR but also the ability to connect to the internet to process speech, especially in real-time. With the rollout of more stable and faster internet connectivity through 5G technology, there will be no limit to where these services can be accessed.

Legal/compliance and real-time court reporting demonstrate an appetite for speech technology to be used not only in consumer-facing applications but to innovate highly regulated use cases such as the court system. The benefits of the digital transformation of audio are critical for legal professionals to accelerate the discovery of information.

> **"There are 6,000 to 7,000 languages spoken in the world. There should be systems created so that people can use voice recognition systems in their daily lives at home, in the car, in schools, at work, in the supermarket, in transport, and wherever else is necessary. The voice recognition should be easy to use and feel natural. Companies and governments should work together to create open systems so that speakers of any language can create voice recognition systems in many languages and that these systems can then be used in as many ways as possible."**
>
> **Seanán Ó Coistín**, Translator

VOICE APPLICATIONS THAT WILL HAVE THE LARGEST COMMERCIAL IMPACT IN 2020

FIG 8. RESPONDENTS THINK THAT THE VOICE USE CASES THAT WILL HAVE THE LARGEST COMMERCIAL IMPACT WILL BE:



⬢	1.	36%	Customer experience
⬢	2.	34%	Home automation
⬢	3.	30%	Contact center
⬢	4.	27%	Automatic captioning
⬢	5.	21%	Voice bots
⬢	6.	18%	Home assistants
⬢	7.	16%	Automotive
⬢	8.	13%	Wearables
⬢	9.	13%	Meeting conferencing
⬢	10.	9%	Work assistants
⬢	11.	7%	Legal/compliance
⬢	12.	4%	Other

CUSTOMER EXPERIENCE

36% of respondents said that they believed that voice will have the largest commercial impact on customer experience. Customer experience is a vital metric in almost every application of product and technology. Simply put, if a customer has a bad experience of a product or service, they won't buy it again. Even worse, with social media it's never been easier to publicly share a bad – or good – experience and influence others even if they have never interacted with the product or service directly. Voice enables a simplified point for humans to interact with machines in an increasingly natural way. The removal of physical user interfaces like keyboards makes it easier than ever to make requests on the go.

HOME AUTOMATION

Home automation is a big business and 34% of respondents said that voice will have the largest commercial impact in this market. MarketsandMarkets forecasted that the home automation system market will grow from USD 39.9 billion in 2016 to USD 79.6 billion by 2022, at a compound annual growth rate of 11.3% during the forecast period. This significant uplift in market size is in part due to the increased popularity of the internet of things (IoT) devices in the home. Use cases include controlling the blinds, lights and heating in the home via a smartphone app. These use cases have enabled consumers to personalize their environment with enhanced levels of ease. The market continues to diversify and more elements in the home can be interacted with. Buyer trends for enhanced levels of security and home monitoring are further changing the market and adding more products that can be voice-enabled. Parks Associates forecasted close to four million unit sales [of video doorbells] in 2019, up to more than five million by 2023. The increased demand for home automation products has also created competition in the market, driving down the price of these technologies. To remain competitive, providers have had to seek out new ways of differentiation with voice being a potential battleground. The biggest player in this market is undoubtedly Ring owned by one of the largest cloud providers in the world.

CONTACT CENTER

The contact center is a hub of innovation especially when it comes to voice. **30% of respondents said that they believed that voice will have the largest commercial impact on contact centers.** The ability to transcribe voice empowers contact center agents and optimizes the customer experience. The ongoing digital transformation of the contact center to leverage tools like ASR enables a better and fully automatic way of capturing interactions. These transcribed interactions can be used in conjunction with sophisticated natural language processing (NLP) tools to extract insights into customer and agent behaviors. This insight enables the contact center to make data-driven decisions to better enhance their business and the experience of the customers.

RISKS FOR SPEECH TECHNOLOGY IN THE NEXT 5-10 YEARS

Respondents outlined some risks for voice technology in the next 5-10 years. Some of these risks are:

- IN THE EU: COMPLIANCE, REGULATION, LACK OF TRAINING AND TEST DATA IS A RISK
- PRIVACY, SECURITY AND IP ISSUES
- TOO MANY BIAS-PRONE APPLICATIONS ON TOP OF INCORRECT TRANSCRIPTIONS
- VOICE TRIGGERED APPLICATIONS COULD GENERATE DANGEROUS MALFUNCTIONS
- PUBLIC SKEPTICISM ABOUT EAVESDROPPING. HIGH PROFILE INSTANCES OF EAVESDROPPING (ACCIDENTAL OR MALICIOUS) COULD SPARK FEARMONGERING
- ANY VOICE SYSTEM NEEDS TO BE ABLE TO CAPTURE AUDIO. HACKING/ INTRUSION OF SUCH SYSTEMS IS ALWAYS A RISK
- THE PROLIFERATION OF VOICE BOTS AND MISUSE OF TECHNOLOGY
- HACKING,LEADING TO FALSE VOICE RECOGNITION
- MINORITY LANGUAGES, ACCENTS AND DIALECT WILL GET LEFT BEHIND

  THIS COULD LEAD TO DISCRIMINATION AGAINST POPULATIONS THAT DON'T HAVE THE SAME ACCESS TO SPEECH TECHNOLOGY SERVICES
- THE TECHNOLOGY WILL NOT MEET CONSUMER EXPECTATIONS
- MANY CONSUMERS WILL EXPECT VOICE TECHNOLOGY TO BE FREE
- THE PUBLIC COULD LOSE TRUST IN VOICE TECHNOLOGY OVER FEARS AROUND DATA COLLECTION, AS WELL AS FEARS AROUND AI
- DOMINANCE BY THE USUAL MONOPOLY PLAYERS IN IT, AND HENCE A POOR COMMERCIAL LANDSCAPE
- GETTING STUCK ON THE "EASY" 80% OF VOICE APPLICATIONS AND NOT TRYING TO MOVE ON TO MORE CHALLENGING FUNCTIONALITY
- CONTRIBUTION TO JOB LOSS MAY LEAD TO A BAD REPUTATION FOR TECHNOLOGY PROVIDERS AMONG THOSE WHO LOSE THEIR LIVELIHOOD TO AUTOMATION
- FAILURE TO IMPROVE ACCURACY AND PERFORMANCE
- TOO MUCH TRUST WILL BE PLACED IN AUTOMATIC SYSTEMS WITH WEAK CONTROL MECHANISMS

> **"Currently, Kaldi is the de-facto open-source ASR leader, but the leader of Kaldi (Dan Povey) was recently fired from his University and has moved to work for Xiaomi in China. Dan has noted that he might be working on a 'Pytorchy Kaldi', which will be interesting. It might go a couple of ways. Current Kaldi may fall into neglect and harm users of Kaldi. New Pytorchy Kaldi might come out and be easier to use and better for the field. Or, anywhere in between. The risks faced with Kaldi are that Dan Povey was so core to Kaldi that such a large change in his circumstances has a high chance of having a genuine impact on the field."**
>
> **Tom Ash**, Machine Learning Engineer, Speechmatics

FUTURE CONCERNS AROUND DATA PRIVACY

93%

of respondents said that data privacy is likely to be a concern in the future.

Unsurprisingly, 93% of all respondents said that data privacy will indeed be a concern in the future. Consumers are voicing their worry over where their data is stored and how. Data collection is a growingly important topic. Providing options for deployment in ASR solutions seems to be key to consumers being at ease with how their data is handled. Consumers are expressing their desire to be able to access or maintain the data themselves and are looking to have complete transparency in the process. These concerns are an increasing trend given the recent news about Google and Amazon's unlicensed use of gathered data, making users wary of the usage of personal data and how it could impact them. Consumers are also becoming savvier and wanting assurance to questions about how their data is being collected, stored, owned and utilized in ways that are deemed clear. Those conditions are influencing how users choose ASR providers.

> **"Consumers continue to judge access to their data to be a major concern. To the extent that this is now informing decisions about access to their speech and transcription data. Where the data resides geographically, and who ultimately owns it, have become important topics of discussion and a major influencer of the ASR system that customers choose."**
>
> **James Page**, Product Owner, Speechmatics

> **"We are all becoming so busy and voice technology has a clear role to play. Data security would remain my concern when considering what data I will use any such speech recognition tool for. I need to be confident in its integrity."**
>
> **Jennifer Joi Field**, Creative Director, Cultural Mapping Pty Ltd

> **"Products built by Silicon Valley are already under scrutiny around how they use data. Expect brands to be challenged on how their products benefit a global audience. Global accents will become more important to end-users using mainstream products as the uptake of voice technology increases.**
>
> **"Google and the gang will make more of their transparency around data collection. Customers will need to actively opt-in and will be able to turn off snooping etc. How and whether this impacts the rate of progress, will become a discussion topic."**
>
> **Eric Henderson**, Product Technologist

OVERCOMING THE CHALLENGES OF DATA SECURITY WHEN IT COMES TO VOICE TECHNOLOGY

Data security continues to be a concern across all industries and not just voice technology. However, concerns around voice data have been prominent in the news in 2019 with global brands such as Amazon and Google confirming they are using their home devices to "listen" to conversations for development and improvement to their devices. With such prominent voice data breaches, businesses are planning for the future to overcome these challenges now and ensuring data privacy is front of mind when integrating voice technology into their workflows.

Respondents believe that data privacy will continue to be a concern in the future, but there will be ways to overcome this as displayed in Figure 9.

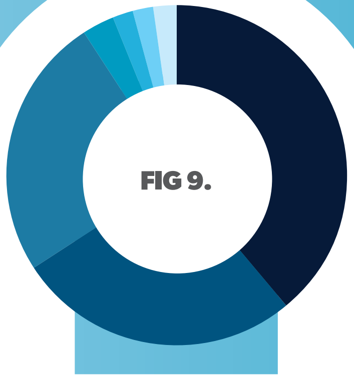


FIG 9.

⬢	1.	39%	On-premises deployment
⬢	2.	27%	Solutions able to operate entirely offline
⬢	3.	25%	Cloud deployment
⬢	4.	3%	Hybrid, depending on business model
⬢	5.	2%	Legal frameworks
⬢	6.	2%	Government cloud
⬢	7.	2%	Not sure

ON-PREMISES DEPLOYMENT

39% of respondents said that on-premises deployment is a great way to overcome preventing issues with data privacy now and in the future. On-premises deployment options enable customers to keep their data securely within their own data centers with no need for customer data to go near the cloud. On-premises deployments for voice technology are often done using appliances or containers so they can seamlessly be deployed into existing technology stacks.

Many industries such as banking and contact centers face compliance and regulation challenges where customer data and voice data cannot leave the premises to 3rd party providers. In these cases, on-premises deployments are the best solution to preventing data breaches and avoiding risks associated with private cloud deployments.

DARK SITE ENVIRONMENTS

Dark site deployment options enable customers to keep their data securely within their own data centers. **27% of respondents said solutions that can operate entirely offline will alleviate the concern around data security now and in the future.** Typically, when deploying an on-premises solution for voice technology, businesses are required to connect to the public internet for licensing. Offline licensing is supported in dark site deployments which means all work is completed within a business' private environment.

Offline licensing enables customers to license and operate the solution without being connected to the public internet. This deployment delivers customers with more robust solutions for compliance and data privacy needs.

Dark site environments are a great solution for many businesses such as the government who need the added level of security and privacy when it comes to voice data.

CLOUD DEPLOYMENT

25% of respondents said cloud deployments satisfy their business need for data security. In most instances, private cloud deployments are secure enough to keep data safe. If cloud deployment security is good enough for the business and use case needs, cloud deployment is often the preferred option due to the reduced operational cost and complexities.

> **"I expect for large cloud providers to start offering on-premises deployments in the future as data privacy continues to be a concern for consumers."**
>
> **James Page**, Product Owner, Speechmatics

> **"I expect that providers will set out improved and clearer measures to ensure the security of voice recordings."**
>
> **David Pye**, Head of Speech, Speechmatics

COMPETITIVE LANDSCAPE

The voice market has never been so competitive with providers in the market as diverse as the application requiring the technology. Prominent tech giants like Google, Amazon and Microsoft continue to extend their solution in long-form speech recognition through the addition of new languages and enhancements to reduce word error rates. The rollout and mass adoption of devices like Google Home and Amazon Alexa have brought new audiences to their technologies now hungry to see the same level of command and control capabilities in other hardware like automotive. With the end-to-end capabilities of the tech giants – especially Amazon with a full-service solution – it has never been easier for organizations looking to add voice capabilities into their products such as creating an Alexa skill or simply to use Amazon's speech-to-text service. While these solutions might offer a faster route to market, entry for voice technology or features, organizations need to understand that data delivered through their interaction will be stored on their cloud through tech giant relationships.

As the complexity of customer use cases increase and requirements become more specific, it is expected that companies might need to look for more flexible solutions than those offered by the large cloud providers. Examples of this have already been established, most famously by Sonos moving away from Amazon and Google to acquire Snips.ai and place the ASR capability under their control. Sonos bought Snips for $37.5 million. At the time, the reason for this acquisition was to enhance ease of use for customers using the Sonos device, however, it later came out (referenced in the article above) that Sonos claimed that Amazon and Google had copied their intellectual property and used this to enhance their offering.

MAJOR PLAYERS IN THE MARKET

GOOGLE

Since Apple brought Siri to the market in 2010, Google has introduced several voice solutions for mobile devices. Google has invested heavily in voice since 2016, creating Dialogflow through the acquisition of API.ai. Google's Assistant is available in over 30 languages. Google has three smart speakers and two smart displays as well as integrating their ASR and virtual assistant technology on thousands of other 3rd party devices. Google was also the first to have a voice assistant available on over one billion devices worldwide.

AMAZON ALEXA

The Amazon Echo was introduced in 2014. Unlike the competition at the time, the Echo and Alexa voice assistants reinvigorated the assistant market that had previously started strongly for both Apple and Google but without defined use cases. While offerings from Apple and Google were fun, many saw them as a gimmick. The new products from Amazon changed the game, offering a fresh channel and interface to do more than just ask about the weather or perform a Google query with voice instead of typing. Amazon educated the market on how voice could be used in a whole new way and effectively created voice commerce for its online retail platform.

SOUNDHOUND

Unlike the tech giants like Google and Amazon that targeted their voice assistants at the consumer market, SoundHound positioned itself to companies to white-label a voice solution. In recent times, the automotive industry has become a key battleground for voice assistants with all the major providers including Amazon, Google, SoundHound and Cerence (a rebrand of Nuance's existing automotive business) looking to gain position for their assistants in as many cars as possible. Historically, the automotive market was owned by Nuance/Cerence who will now need to fight off stiff competition to retain their position in this market.

CERENCE

A rebrand of Nuance's existing automotive business. Nuance is well known for leading the medical transcription market. They are present in some other markets but have spun out the automotive part of their solution to differentiate themselves.

With the introduction of new voice providers, the market remains fragmented with a range of products used to create specific solutions. While organizations can leverage a single provider for all components, others integrate multiple products and solutions such as NLP and ASR to create a larger solution like a voice bot.

When it comes to integrating voice into a solution, organizations need to properly understand their requirements and make the best decision as to the provider/s they use. While the accuracy of transcription will likely be a requirement for all organizations, considerations need to be made for where data is stored, the deployment location of the solution, the flexibility that the solution can offer, the support and agility of the provider to make changes to their solution to ensure they retain a best-in-class offering.

> **"Amazon is regularly releasing new languages and are devoting most R&D effort into improving their transcription language models. Their accuracy rates track Google's quite closely. I expect Amazon to release more languages over the next 12 months as they seek to reach parity with Google's range of languages.**
>
> **"Microsoft is consistently the best overall (and steadily improving), although they do not seem to be interested in expanding their range of languages.**
>
> **"IBM is consistently the worst performer of the major players, and do not appear to be actively improving their models."**
>
> **James Page**, Product Owner, Speechmatics

VOICE TECHNOLOGY CONSIDERATION

WILL VOICE BE CONSIDERED IN YOUR 5-YEAR STRATEGY?

70%

of respondents said that voice is likely to be considered in their business' 5-year strategy.

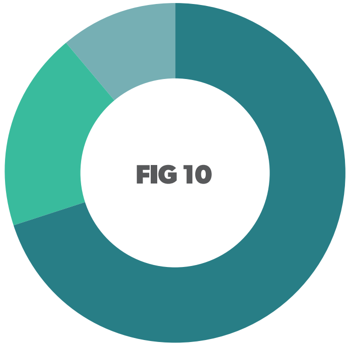


FIG 10.

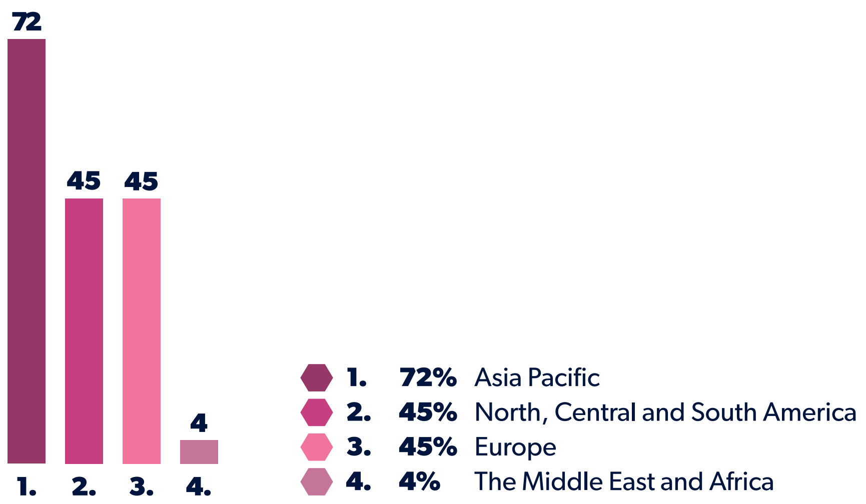
⬢	**70%**	Yes
⬢	**19%**	No
⬢	**11%**	Not sure

70% of respondents indicated that their organization will be considering voice as part of their 5-year strategy. This is a large increase from the 50% that are currently adopting a voice strategy. It's no surprise that there has been an increase since the technology is reaching maturity and is more widespread. Companies are realizing the value that voice technology brings to their businesses, not only to improve efficiencies and streamline processes but to improve customer experiences and ultimately increase revenues. As customer expectations increase, companies are encouraged to improve their products and services, and voice offers unique and innovative ways of doing this.

> **"2019 showed us that more and more companies are realizing that voice technology is good enough to solve real-world problems and support significant improvements to workflows and customer experiences. The markets are waking up to the fact that voice technology is an enabler for them."**
>
> **Ian Firth**, VP Products, Speechmatics

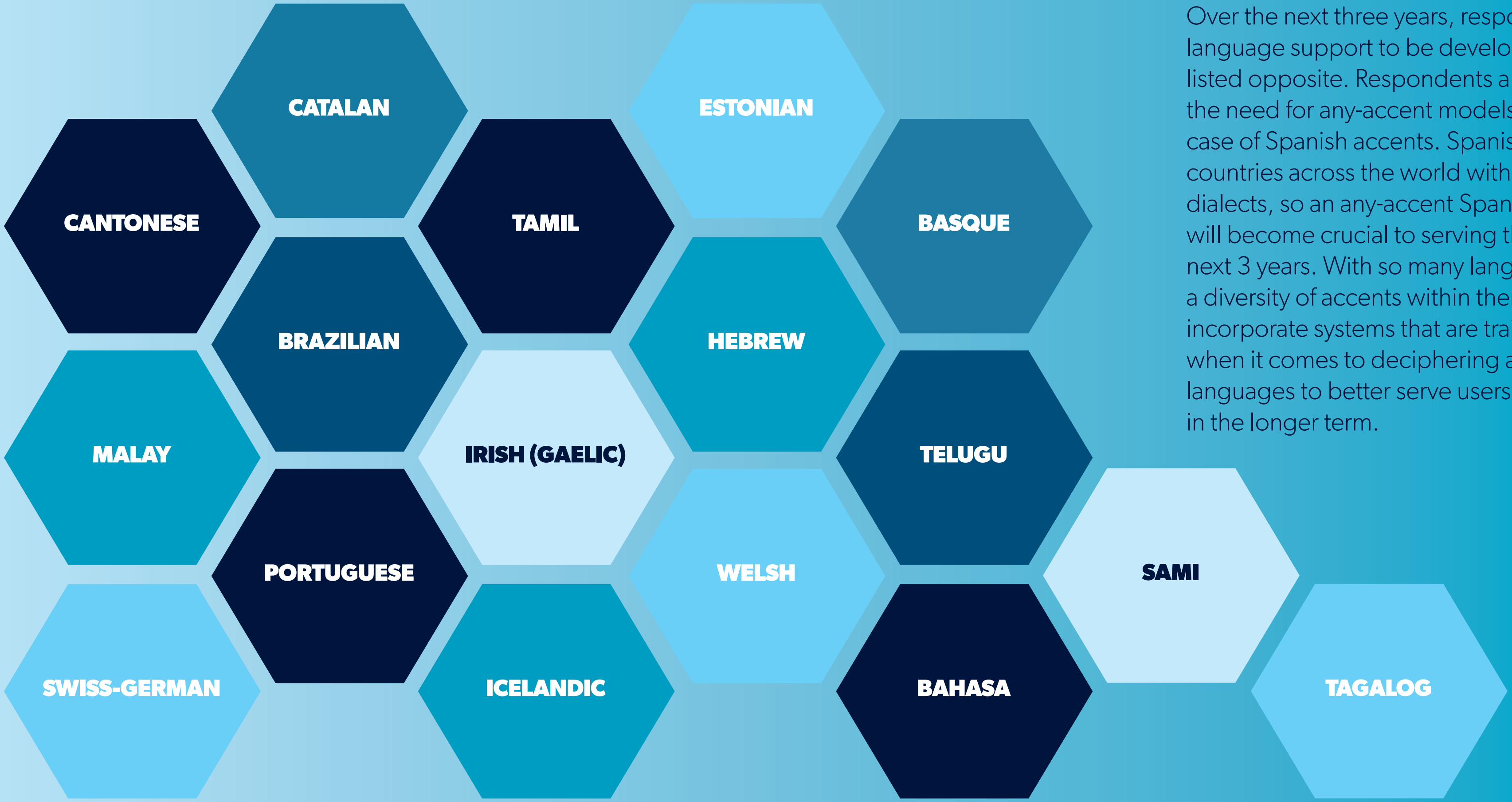
GLOBAL REGIONS THAT WILL HAVE THE LARGEST GROWTH IN THE ADOPTION OF SPEECH RECOGNITION TECHNOLOGY

FIG 11.
(%)



72% of respondents indicated that the Asia Pacific regoin will have the largest growth and need globally for adopting speech recognition technology. This is largely due to the growth of the economy and population which will impact business and consumer trends. Following on from there, North, Central and South America, as well as Europe, are also predicted to adopt voice technology at a rapid rate. Again, this is largely due to economic, social and technological factors, with the value of voice being realized. Conversely, the Middle East and Africa are unlikely to adopt speech recognition technology rapidly as there is not yet need for it in many instances.

LANGUAGE SUPPORT IN THE NEXT THREE YEARS

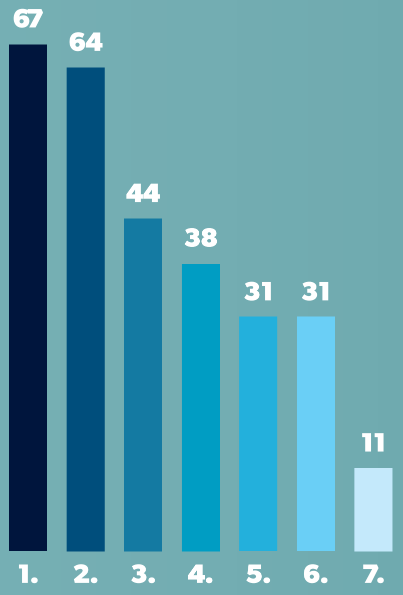


- CANTONESE
- MALAY
- SWISS-GERMAN
- CATALAN
- BRAZILIAN
- PORTUGUESE
- TAMIL
- IRISH (GAELIC)
- ICELANDIC
- ESTONIAN
- HEBREW
- WELSH
- BASQUE
- TELUGU
- BAHASA
- SAMI
- TAGALOG

Over the next three years, respondents expect ASR language support to be developed in the languages listed opposite. Respondents also acknowledge the the need for any-accent models, for example, in the case of Spanish accents. Spanish is spoken in many countries across the world with varying accents and dialects, so an any-accent Spanish language model will become crucial to serving that market over the next 3 years. With so many languages having such a diversity of accents within them, ASR will need to incorporate systems that are trained to be agnostic when it comes to deciphering and transcribing those languages to better serve users and increase accuracy in the longer term.

FEATURE DEVELOPMENT IN THE NEXT THREE YEARS

FIG 12. RESPONDENTS SAID THAT FEATURE DEVELOPMENT WILL BE CRUCIAL OVER THE NEXT THREE YEARS FOR VOICE TECHNOLOGY, THESE FEATURES INCLUDE: (%)



⬢ **1. 67%** Increased word error rate (WER) accuracy
⬢ **2. 64%** Better speaker separation (speaker diarization in mono media files)
⬢ **3. 44%** Language identification
⬢ **4. 38%** Short utterance accuracy
⬢ **5. 31%** More languages offered by more companies
⬢ **6. 31%** Translation
⬢ **7. 11%** Other

67% of respondents said that they would like to see better accuracy over the next three years. WER levels have started to plateau for the most spoken languages in the world, however, rates will still continue to improve in these languages over the next three years. Improvements we will see will be incremental compared to the past few years. Where we will begin to see a bigger change in WER is in languages that are hitting around 70% accuracy at the moment. This is only possible due to better data and smarter algorithms.

WER is the industry standard for measuring the correct output of words for a transcript but other features can help improve the overall accuracy of the transcription output as well.

64% of respondents said they would like to see better speaker separation over the next three years. The ability to detect and label different speakers within the same media file or stream is a powerful feature for many use cases such as call recording and online conferencing software. While speaker separation already exists as a feature for voice technology, it still has a long way to go to deliver the results enterprises need for high-volume, large scale operations.

Language identification is an important voice technology feature for global businesses. **44% of respondents said that they would like to see this feature available for voice technology over the next three years.** Language identification automatically detects what language is being spoken and then transcribes in that language without manual intervention. This feature will increasingly become crucial as enterprises scale voice technology into their global technology stack and workflows. Specifically, language identification will enable contact centers to operate more efficiently and reduce costs.

> **"I'm loving speech recognition improvements when it comes to accents. As someone who uses automatic speech recognition to transcribe phone interviews, the ability to clearly distinguish between speakers would be the most valued improvement for me. It would make it easier to scan through transcriptions in search of good quotes without needing to return to the recording."**
>
> **Adam Turner**, Freelance Journalist and Corporate Writer

SPOKEN LANGUAGE TRANSLATION

Innovation within voice means that industry use cases will continue to evolve with an expectation that speech recognition accuracy will improve, and features and intelligence will also grow around it. Transcription and translation are often spoken about together, however, these are currently separate features of language identification but have the potential to add significant value if used together. The ability to automatically identify a spoken language and trigger transcription can solve use cases where speakers flick between one language and another. This optimizes the accuracy of a specific media file or when transcribing in real-time.

While language translation exists today there is an area range of nuances around how each language is constructed. Audio can be transcribed in one language, translated word for word and then fed into a text-to-speech engine. The output, however, will never reflect a natural output. If this use case becomes more relevant, additional understanding and experimentation will be required with specialist providers to dedicate effort to enabling the delivery of a transcribed, translated and machine spoken output that is near to indiscernible from a natural speaker.

IMPROVED ACCURACY OF VOICE TECHNOLOGY

WER will continue to improve by throwing more data at the problem, however, this approach will see diminishing returns. The leading providers in the ASR space are now delivering WER accuracy around 95% for English. Using data to increase accuracy even further will require a huge amount of data and increasing levels of processing power for single percent increases. For many providers, this will not be worth it or they simply will not have the available hardware to retrain with such huge data sets. With English reaching a point of accuracy which is hard to surpass, ASR providers need to look at the accuracy of other languages. Providers that claim to support ASR in a huge variety of languages will be challenged to ensure that the accuracy of these languages are fit for purpose and able to facilitate the use cases that they are selected for.

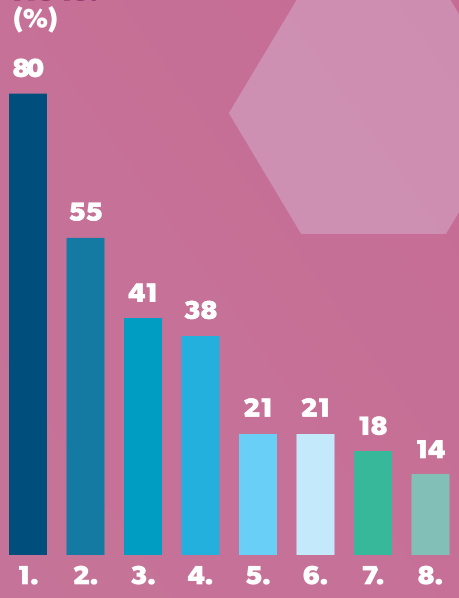
ASR providers will be forced to look to more customer-relevant terms for accuracy, beyond word error rate. Whether that be accurate identification of speakers, accurate placement of punctuation, or more robust accuracy in challenging environments such as noisy locations or audio recorded on low-quality devices.

> **"I expect WER to continue to improve with more data, larger networks and more compute."**
>
> **Tom Ash**, Machine Learning Engineer, Speechmatics

PERCEPTION OF ACCURACY

The perception of accuracy for voice technology is very specific to the use case and business. There are several measures that respondents said are important when it comes to accuracy, these include:

FIG 13.



⬢	1.	80%	Word error rate
⬢	2.	55%	Speaker change indicated
⬢	3.	41%	Intent recognition
⬢	4.	38%	Punctuation
⬢	5.	21%	Quick turnaround time for transcription
⬢	6.	21%	Multiple languages
⬢	7.	18%	Character error rate
⬢	8.	14%	Unified number recognition format

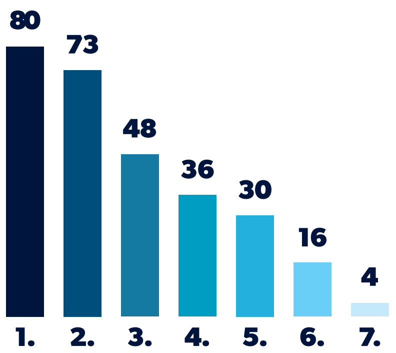
The perception of accuracy for voice technology is often interpreted as the word error rate (WER). **80% of respondents also had the same perception.** WER is the industry standard for measuring the accuracy of word output per transcript and is a great benchmark. However, many businesses and use cases use other metrics to measure the accuracy of the transcription output.

55% of respondents said speaker change being indicated in the transcription is an indicator of accuracy. Understanding the conversation flow provides an added level of accuracy to the transcript other than just how accurate the word output is. Industries such as contact centers and media, entertainment and broadcast find this feature incredibly valuable in real-time scenarios as well as post-processing. **41% of respondents also said intent recognition is an important feature in an accurate transcript output.** Intent recognition provides the context of what has been said and can often change the output of a sentence to ultimately increase the accuracy, but this could affect the WER as a direct measure.

> **"Word error rate (WER) will start to plateau at around 95% accuracy. Achieving the other 5% is either not important for some markets or will require a really deep understanding of the real world to solve, using innovative advances in fundamental machine learning. These advances will enable more to be achieved with less data, supporting lower resourced languages, domains and use cases that will open up the technology to a massive global market. ASR will become a platform onto which new functions are added to support higher-level understanding to be gained."**
>
> **Ian Firth**, VP Products, Speechmatics

THE FUTURE OF SPEECH RECOGNITION

FIG 14.
(%)



⬢	1.	80%	Improved word error rate (WER) (95%+)
⬢	2.	73%	More robustness when it comes to noisy environments
⬢	3.	48%	More focus on any-accent language models
⬢	4.	36%	More on-device deployments through reduced model size
⬢	5.	30%	Less data required to improve models
⬢	6.	16%	Large companies will offer more deployment options
⬢	7.	4%	Other

IMPROVED WORD ERROR RATE (WER) (95%+)

Looking back to the perceptions of accuracy, lots of people regard low word error rates to be the main definition of accuracy. **80% of respondents said that the word error rate is likely to improve in the future.** WER is likely to hit 95% accuracy and plateau due to the remaining 5% not being worth chasing. Although more data and better understanding, improved machine learning techniques etc. is likely to drive word error rates down to 0%, hence 100% accuracy. But this is a way off, as echoed by John Parkinson, CEO at Parkwood Advisors who said that **"Speech recognition has come a long way but there's still a long way to go..."**

MORE FOCUS ON ANY-ACCENT LANGUAGE MODELS

Most ASR providers have multiple accent packs for their languages. Due to this, the expectation from 48% of respondents is that these providers will move towards a single model per language as accents continue to diversify and broaden. Brands will be challenged on how their product benefits a global audience, in addition to balancing the cost of deploying and operating their language packs. As global accents will become more important with wider adoption of voice, organizations and existing ASR providers will be required to factor in how they plan to efficiently deliver speech recognition in global applications.

> **"We need greater accuracy with quick deployment in a low-code environment."**
>
> **John Wood**, Director, C3

Regarding languages used by small populations of people like Irish, Icelandic and Estonian, it is likely that these languages will remain without a dedicated ASR model. For this reason, users in the countries or areas that these languages are spoken will likely be forced to use another language to interact with devices and applications that are voice-enabled.

MORE ROBUSTNESS WHEN IT COMES TO NOISY ENVIRONMENTS

Contemporary ASR solutions are incredibly effective even in noisy environments or when media is recorded on low-quality devices. **73% of respondents expect ASR providers to get even better when dealing with audio recorded in noisy environments.** ASR providers are likely to improve the diversity of their training models. They will feature equally challenging audio profiles and aggressive benchmark testing that identifies key areas for improvement and triggers additional training. They will fortify natural language models to ensure the quality of ASR even in challenging audio environments.

"I expect ASR will improve particularly robustness to difficult audio, acoustic environments, crosstalk, accents and dialects."

David Pye, Head of Speech, Speechmatics

LARGE COMPANIES WILL OFFER MORE DEPLOYMENT OPTIONS

16% of respondents expect large companies to offer more deployment options. End-to-end solutions like Amazon which require limited effort and time to market are an attractive option for organizations of all sizes. Smaller organizations might not have the technical expertise to create their own solution, even if it does mean that they miss out on a better ASR provider. It might be acceptable to trade-off accuracy, language diversity, or features for ease of use. While cloud providers might deliver ease and accelerated time to market, for some providers their models are contained to their cloud. This means that organizations using their service need to understand the implications of their customer data being used by these cloud providers and the potential impact this can have on the brand using these cloud providers to power their solution.

It becomes important for brands to understand the differences in deployment options from the cloud through to on-premises and on-device. Some of these options were touched upon earlier in this report. For technologies like virtual assistants that require huge quantities of data to deliver an effective service, cloud deployments are often the preference. However, this comes at the cost of consumer privacy. In 2019, there were lots of stories around tech giants leaking data. With increased adoption comes increased understanding and awareness of what goes into making a great virtual assistant. Consumers are becoming much more tech-savvy and less forgiving of tech giants exploiting their data.

Brands that are not happy for their data to be shared with large cloud providers might decide that an on-premises option suits them better. This ensures that their customers' data remains securely within their environment. This is already common practice for organizations in the financial and banking sectors who are required to store and secure their customers' data with huge fines and the degradation of their brand depending on how effectively this is done.

As privacy concerns become more important, on-device options may become a deciding factor for organizations and consumers alike. Notably, Siri has fallen behind the curve in terms of the virtual assistant ecosystem, especially when compared with Alexa and Google Home. A potential glimpse into the future of Apple might lie in the recent acquisition of Xnor.ai. Xnor.ai focuses on the development of AI solutions that remain solely on-device with no connection to the cloud, instead, delivering all processing on-device. By investing in on-device, Apple has the potential to differentiate against its competitors. It is a great solution to the issues surrounding personal data, and on-device helps to keep this data in the control of the user.

LESS DATA REQUIRED TO IMPROVE MODELS

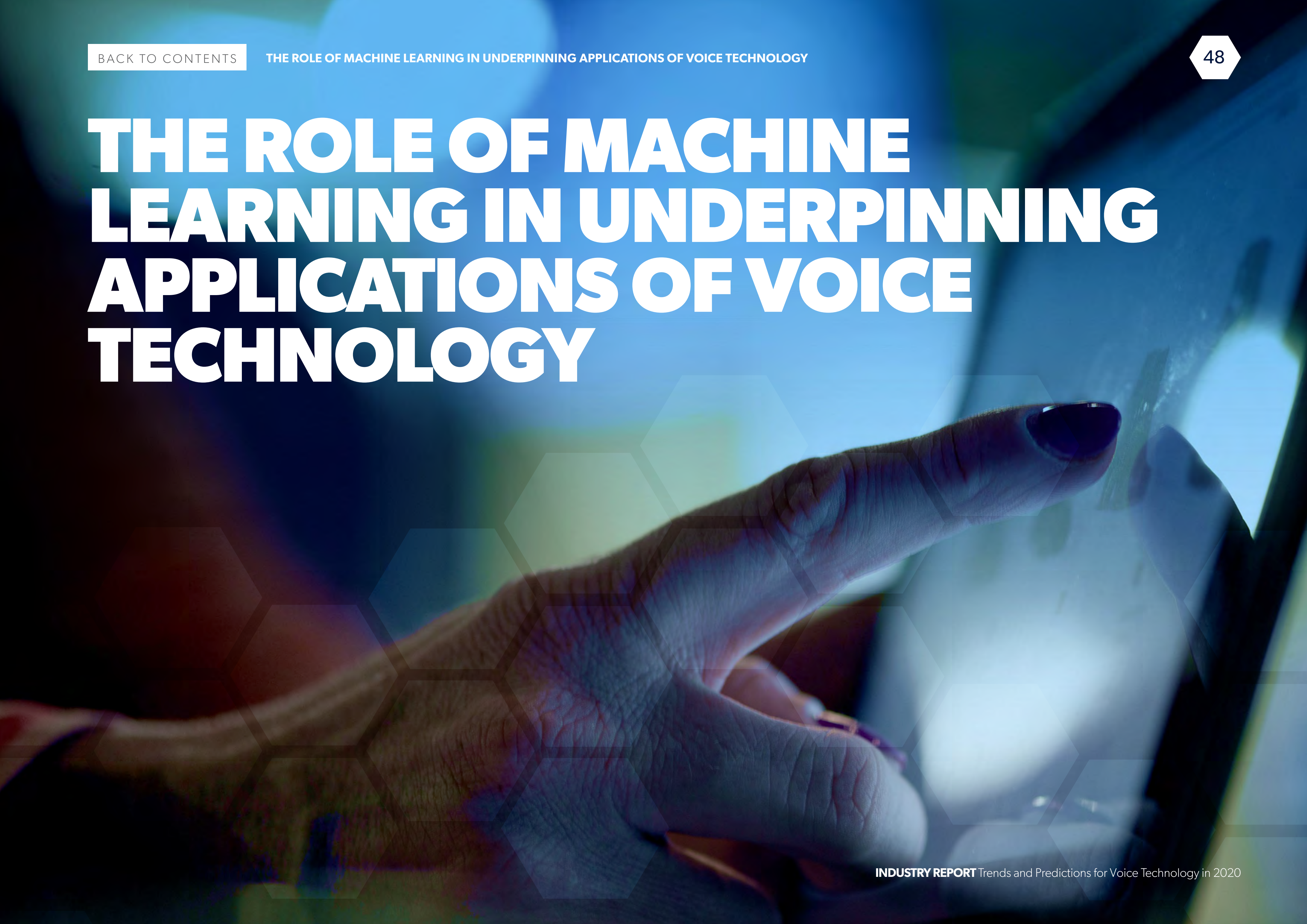
Large cloud providers with extensive CPU resources have the potential to uplift the capabilities of their machine learning solutions by throwing more and more data at them. However, concerning automatic speech recognition, the potential gains of 1-2% improvements to word error rates might be deemed inefficient when considering the cost. In addition to this Tom Ash, Machine Learning Engineer at Speechmatics comments that **"reducing model size will be a priority to allow deployment on-device".** Tom continues that **"low resource approaches to ASR continue to be of interest to the field. Allowing obscure languages to be created with less data".**

It isn't really surprising that in a competitive market that relies on cloud services like the virtual assistant market, there is a focus on edge device-based solutions. These solutions put data security firmly in the hands of the consumer and delivers a different approach to that adopted by the data-hungry solutions available by the tech giants. Also, the ability to deliver increasingly high-performance ML solutions like ASR on edge devices will drive the creation of models that require significantly less data for them to work on low resource devices. The ability to create these high-performance solutions with minimal data that work on edge devices will mean that more obscure languages can be created to support ASR. Improving ASR systems without increasing data usage will mitigate against current challenges and enable providers to support more use cases and enable voice technology to be adopted through broader global geographies.

"Voice is what makes out human interaction. It will be more important in the future to a degree that we cannot imagine today."

Ralf Mühlenhöver, CEO, voiXen

THE ROLE OF MACHINE LEARNING IN UNDERPINNING APPLICATIONS OF VOICE TECHNOLOGY



NEW MACHINE LEARNING METHODS THAT ARE BEING UTILIZED

Machine learning underpins the progress of voice technology. The approach of applying recurrent neural networks to speech recognition was demonstrated in the 1980s, where it outperformed traditional methods. The rise in computing power, graphics processing and cloud computing made the huge potential of this approach a reality.

Advances in machine learning, new techniques and innovation are all contributing towards continuous improvements in speech recognition technology, not only through providing better accuracy but also in feature development and language capabilities. Already, voice providers are using machine learning innovation to provide value to consumers through better language capabilities, for example doing away with specific language models for different English accents i.e. UK, US, AUS, and providing a single English language pack. It is predicted that this method will be extended to Spanish, French, Italian etc. in the future.

> **"Speech recognition technology has been in the making for the last decade and has finally made its quantum leap in the last 2 years. With fast-developing deep machine learning and AI, I do expect it to make its next crucial quantum leap sooner rather than later."**
>
> **Doreen Chong**, Quality Assessor/Trainer, Epiq Global

UNSUPERVISED LEARNING

There has been a lot of progress in the machine learning field with natural language processing (NLP) due to a family of algorithms called unsupervised learning. Companies are being built based on the breakthroughs in unsupervised learning in text and language, but those algorithms can also be applied to speech. So, how can companies label elements in speech without using labeled data? For this to work, unsupervised learning algorithms need to be leveraged to see the same breakthroughs we've seen in NLP but to bring those through to speech recognition. This is a trend that is starting now and we're going to see a lot more of it in 2020 and 2021.

To effectively train machine learning models for use in speech recognition, thousands of hours of labeled data is required. The issue with labeled data is that it is very expensive, hard to get hold of (especially in all the required domains and languages) and is often mislabeled. Unsupervised learning mitigates against these problems and enables machines to learn much more like humans would, with limited labeled data. Sam Ringer, Machine Learning Engineer at Speechmatics says that **"it is at least theoretically possible to come up with strong, or weak, unsupervised solutions to these problems because we [humans] can do it"**. It is, therefore, likely that unsupervised learning will surpass traditional methods of using labeled data in the next year or two.

“If we can leverage how humans learn, we can not only save money (because we don’t need to come up with labeled data), there will be less room for labeling issues, and it will lead to better results.

“One of the most underappreciated reasons as to why machine learning is moving at the pace it is, is because all of the machine learning research is available all of the time and anyone can access and look at it for free.”

Sam Ringer, Machine Learning Engineer, Speechmatics

“End-to-end style approaches that remove the expert knowledge and various steps in the current ASR pipelines will continue to be heavily researched. There is a view that we are currently in a highly over-optimized ‘local minima’ for ASR and that end-to-end systems are going to be the next step-change. However, the bar for entry is very high because of how heavily optimized the current paradigm is. In the long term, this would be quite a large paradigm shift when it comes in. This is likely to be more in the 5-10-year bracket, with people gradually moving over time.

“In the meantime, technical approaches like network distillation, quantization and similar will continue to have a lot of practical interest in reducing model size to allow for deployment on low spec devices (internet of things type approaches).

“Low resource approaches to ASR continue to be of interest to the field and seem to be incrementally slowly improving, allowing more obscure languages to be created with less available data.”

Tom Ash, Machine Learning Engineer, Speechmatics

END-TO-END STYLE APPROACHES

End-to-end style approaches to ASR will continue to be researched to remove the need for expert knowledge specifically in this field. Emphasis will then shift to gaining further in-depth knowledge in ML to keep advancing the field. An end-to-end approach is still a way off and is likely to have an impact in the next 5-10 years. In the meantime, network distillation, quantization and other technical approaches will be used to reduce model size to enable more deployment options, especially on-devices.

SUMMARY

Today, voice technology will continue in its upward trajectory in both popularity and adoption. Voice technologies are no longer the thing of science fiction, they are a valuable asset that show no sign of slowing down in the foreseeable future. From consumer devices in homes like the Amazon Echo and Google Home to the deployment of voice-based solutions within business and enterprise environments, voice is and will continue to be additive to our daily lives. **50% of professionals that engaged with the research said that their company currently has a voice strategy.** Further to this, **70% said that voice is likely to be considered in their business' 5-year strategy.** This emphasizes that voice will continue to be a core technology to drive the advancements in businesses across many industries.

Voice bots and virtual assistants have become common-place and remain the most visible application of machine learning, artificial intelligence and speech recognition technology. With that said, terms like AI are used very broadly and on occasions misused to describe a technology landscape to make it sound impressive. This presents a risk in the enterprise space to stifle adoption of genuinely groundbreaking technology due to a poor experience with a solution that claims lots but delivers little.

In terms of business applicability of voice, customer experience will lead the way with the highest percentage of professionals calling this out as a major application of voice technology. Delivering better experiences for customers enables brands to strengthen their position in the market whilst demonstrating an appetite to adopt new technologies. Consumers will start perceiving these brands as future thinkers and innovative. Voice technology and ASR enables businesses to improve interactions with customers from empowering contact center agents to deliver better information and resolve issues faster to creating fully autonomous IVR-type solutions able to solve basic issues that never need human interactions. Additionally, archived audio recordings can be leveraged by transforming them into text and then adding them as an additional stream of data into existing NLP tools. This means that true digital transformation is available, exposing a huge amount of insight on customers and the wider business.

In the consumer space, home automation is driving the market. Even furniture brands are adopting voice enablement through the simplistic means of interaction with Amazon Alexa skills. However, **93% of respondents said that data privacy is likely to be a concern in the future.** Consumers have previously been prepared to give away personal or intimate data as the price for new and exciting technology like virtual assistants that require data to perform desired tasks. However, more consumers are becoming educated about the cost of these solutions and what they are willing to give up for convenience. For this reason, challenger approaches to intelligent tools will become of interest.

As a result of data privacy concerns, small-footprint approaches to ASR and other machine learning derived solutions are improving. A small footprint enables these technologies to be deployed on end-user devices like smartphones and smart speakers that do not require a connection to the internet and cloud services that consume and store end-user data. These devices have less available resources and so models will be required to be smaller but still deliver comparable performance to cloud services. In terms of ASR, the result of delivering a high-performance solution with small models means that there is an opportunity for more obscure languages to be built using less training data at a standard that can be used in commercial applications.

Voice-enabled devices will become more multilingual so they can be deployed in broader geographies, supporting multiple accents and dialects. Additionally, the expectation is that these voice services will do more and in real-time. For example, if someone speaks Chinese on a call, the mobile phone may not only recognize and transcribe the speech but also translate it and send an English output to your earphones.

Word error rate will continue to be a major metric and progress will continue to be made by throwing more data at the problem with growing networks to compute. There may not be many benefits for end-users as they are too large and potentially too slow to be practical. For this reason, ASR may move past the 'more data is all you need to improve accuracy' argument. Expectations will start to plateau at around 95% accuracy and achieving the last 5% will be either not important for some markets or will require a really deep understanding of the real world to solve.

The evolution in voice technology from language diversity, feature availability like punctuation and non-speech recognition through to WER accuracy, makes it even more useful and widely adopted. The ability to effectively trade-off the capability of these solutions with the data required to enable their function is a more complex issue. Businesses are under increasing pressure to keep their customers' data secure with increasing levels of regulation and compliance. Also, a new breed of tech-savvy consumers interested in asking questions about how their data is being used by technology providers is emerging. The use cases around voice are also becoming increasingly complex dependent on industry, market, use case and consumer base. While cloud providers might provide the easier route to voice-enabled products or services, considerations should be made to understand if that approach is the right decision to meet the consumer in a constantly evolving market.