SPEECHMATICS

INDUSTRY REPORT

TRENDS AND PREDICTIONS FOR VOICE TECHNOLOGY IN THE MEDIA AND ENTERTAINMENT INDUSTRY

MARCH 2021

WWW.SPEECHMATICS.COM

CONTENTS

EXECUTIVE SUMMARY

The rise in AI and machine learning has opened up opportunities for media companies to expand their capabilities beyond expectations in the past few years. The media and entertainment industry was an early adopter of voice technology and as a result, it is now one of the leading technologies providing value to the industry.

2020 changed the world in a way that no one could have ever predicted. The COVID-19 pandemic has forced us to adapt, change and evolve the way we live, work and interact across the world. With this change came huge opportunity for some industries and the media and entertainment industry is one of them. Increased demand on social media, over-the-top (OTT) streaming, broadcasting, digital advertising and marketing has not only delivered financial reward to businesses operating in the industry, but has also forced the industry to adopt future technologies – such as voice technology – faster which has had an impact on related service industries too.

The results from our research found that COVID-19 has changed the use of and demand for media solutions and explains how important future technologies – such as speech technology – are to ensuring the industry continues to grow and adapt with increasing consumer demand. The research was taken from a range of business sizes from 1 to over 10,000, in media and entertainment, education and transcription – all of which believe there is value in adopting voice technology at the core of their media solutions.

We hope that the findings presented in this report will help to advance the progress of media organizations’ voice technology strategies in light of the new rate of adoption of media and entertainment technology.

FOREWORD

THE RISE OF STREAMING SERVICES AND VIDEO

Video, social media and streaming services were already on an upward trajectory across all platforms before 2020. Millions of hours of content were being consumed each day across a multitude of channels from Instagram and YouTube to Netflix and Amazon Prime Video pre-pandemic. But if we look at the current situation, the old adage 'content is king' has never been more relevant.

Statista reported that on the weekend the majority of the world plummeted into lockdown in March 2020, the time spent streaming TV and video grew by between 44% and 7.5% across the world.

Ofcom also reported in their **Media Nations 2020 annual study** that in the UK, the majority of adults signed up to Netflix and Amazon Prime Video if they hadn't already been subscribing, and Disney+ overtook Now TV as the third most popular paid-for streaming platform.

VIEWING HABITS DURING LOCKDOWN

Adults spent nearly **6 and a half hours a day** watching TV and online video (or 45 hours a week)

1 hour 11 minutes per day spent watching streaming services, **double** what it was before the pandemic

12 million customers signed up to new services like Netflix, Amazon Prime Video and Disney+

Viewing figures for video streaming services up **71%** on 2019

Source: Ofcom

THE ROLE OF TECHNOLOGY IN THE MEDIA CONTENT BOOM

Along with the evolution of media solutions to meet increasing demand since the pandemic, technologies have also swiftly evolved to address the challenges of this media content boom. It's no secret that to deliver enhanced media solutions, businesses need to utilize next-generation technologies to manage a huge amount of content that they want to make available to their audiences. It also helps to deliver better insights, enhance the customer experience when interacting with the content, and ensure organizations are conforming to regulatory requirements.

The rise in demand and advancement of artificial intelligence and machine learning has opened up opportunities for media companies to leverage speech technology in their solutions. Speech technology helps to reshape how content is interacted with by the end user. It opens up accessibility and enables more people to consume information whenever they wish.

The purpose of this report is to explore the trends and predictions for voice technology in the media and entertainment industry for 2021. It will also provide insights into the value of voice as determined by professionals within the media and entertainment industry. It will address key expectations of speech providers and discuss how organizations can implement new tools to enhance the core of their media solutions and meet the increasing consumer demand.

METHODOLOGY AND DEMOGRAPHICS

To write this industry report, Speechmatics collated data points from Owners/Executives/C-Level, Senior Management, Middle Management, Intermediate and Entry Level professionals from the media, entertainment and education technology industries. These people work across the globe including the UK, Europe, United States and Asia.

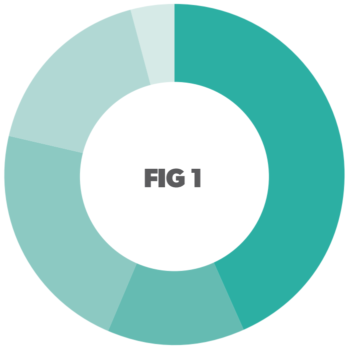


FIG 1. ROLES OF PEOPLE FROM WHICH DATA WAS COLLECTED

⬢	**44%**	Owner/Executives/C-Level
⬢	**13%**	Senior Management
⬢	**22%**	Middle Management
⬢	**17%**	Intermediate
⬢	**4%**	Entry Level

JOB TITLES INCLUDE:

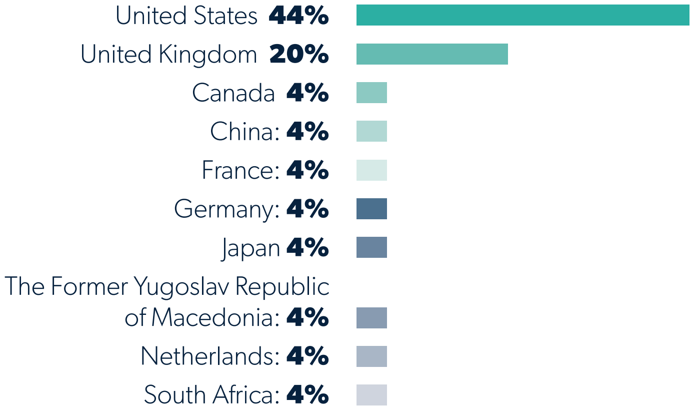
- COO
- CTO
- CEO
- MD
- R&D Director
- Director of Research and Development
- President
- Director
- Data Scientist
- Senior Software Engineer
- Head of Archive
- Media Infrastructure Architect
- Post-Production Engineer
- Production Manager
- Product Manager

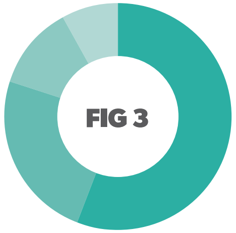
The respondents described their job roles as COO, CTO, CEO, MD, R&D Director, Director, President, Head of Archive, Product Manager, Production Manager, Post-Production Engineer, Media Infrastructure Architect amongst others. The respondent pool included a wide range of organizations who operate within the media and entertainment and education technology industries including but not limited to, editing, media asset management, media monitoring, subtitling and captioning, transcription and e-learning.

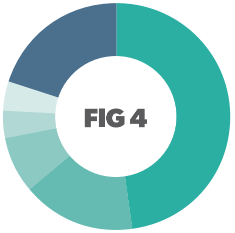
The collated data encompasses a range of organizations, from large enterprises to smaller start-ups. 20% of organizations surveyed employ over 1,000 people, 4% employ 501-1,000 people, 4% employ 251-500 people, with the remaining 72% employing less than 250 people.

56% of these organizations are business-to-business and 44% are business-to-consumer.

FIG 2. LOCATIONS OF RESPONDENT







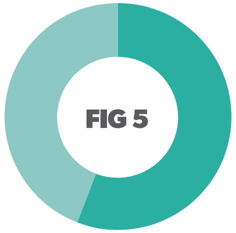


FIG 3. INDUSTRY

⬢	**56%**	Media and Entertainment
⬢	**24%**	Transcription
⬢	**12%**	Other
⬢	**8%**	Education

FIG 4. ORGANIZATION SIZE

⬢	**48%**	1-50
⬢	**16%**	51-100
⬢	**8%**	101-250
⬢	**4%**	251-500
⬢	**4%**	501-1,000
⬢	**20%**	More than 1,000

FIG 5. ORGANIZATION TYPE

⬢	**56%**	Business-to-business (B2B)
⬢	**44%**	Business-to-consumer (B2C)

CONTENT STATISTICS

As the media and entertainment industry continues to grow rapidly, the ability to process the volume of audio and video content increases in parallel. But how large and rapid is this growth?

MORE THAN TWO BILLION LOGGED-IN USERS ACCESS YOUTUBE EVERY MONTH. THE PLATFORM HAS OVER **30** MILLION DAILY USERS WATCHING OVER A BILLION HOURS OF VIDEO EVERY DAY

IN 2020, NETFLIX SPENT OVER **$17** BILLION ON ORIGINAL CONTENT

CISCO HAS PREDICTED THAT BY 2022, VIDEO WILL MAKE UP **82%** OF ALL IP TRAFFIC

DURING THE THIRD QUARTER OF 2020, FACEBOOK STATED THAT **3.3** BILLION PEOPLE WERE USING AT LEAST ONE OF THE COMPANY'S CORE PRODUCTS (FACEBOOK, WHATSAPP, INSTAGRAM, OR MESSENGER) EACH MONTH

REVENUE IN THE VIDEO STREAMING (SVOD) SEGMENT IS PROJECTED TO REACH OVER **$70** BILLION IN 2021 WORLDWIDE

WITH ROUGHLY **2.8** BILLION MONTHLY ACTIVE USERS AS OF THE FOURTH QUARTER OF 2020, FACEBOOK IS THE BIGGEST SOCIAL NETWORK WORLDWIDE

STATISTA REPORTED THAT THE VIDEO STREAMING (SVOD) SEGMENT IS EXPECTED TO SHOW AN ANNUAL GROWTH RATE (CAGR 2021-2025) OF 11.0%, RESULTING IN A PROJECTED MARKET VOLUME OF OVER **$108** BILLION BY 2025

DISNEY+ LAUNCHED IN NOVEMBER 2019 AND SINCE THEN, IT HAS AMASSED AROUND **87** MILLION SUBSCRIBERS WORLDWIDE

IN 2020, OVER **3.6** BILLION PEOPLE WERE USING SOCIAL MEDIA WORLDWIDE, A NUMBER PROJECTED TO INCREASE TO ALMOST **4.41** BILLION IN 2025, REPORTED STATISTA

THE WORLDWIDE E-LEARNING MARKET IS PROJECTED TO BE WORTH **$325** BILLION IN 2025

KEY FINDINGS

1. 1. Surveyed companies think the **consumer electronics (52%) and media and entertainment (52%) industries experienced the biggest positive impact** as a result of COVID-19

2. The overwhelming majority of respondents (72%) think **the main application for voice technology is currently subtitling and closed captioning**

3. Media companies find the key barriers to the adoption of voice technology to be accuracy **(76%), accent/dialect related recognition issues (60%) and language coverage (48%)**

4. **Data privacy matters a lot to 88%** of the media and entertainment industry and is a growing concern

5. **69% of companies surveyed** want to see continued improvement in word error rate (WER) accuracy and **61% want to see improvement of speaker diarization accuracy**

6. Key factors for media companies when evaluating the accuracy of speech technology include **word error rate (62%), punctuation (54%) and speaker change indicated (54%)**

7. 96% of surveyed companies think that **the demand on collaboration tools will increase from 2021 and beyond due to COVID-19**

8. **48% of media companies said that their company currently has a voice strategy** and of those that don't have a voice strategy, **52% said it is something that will be considered in the next 5 years**

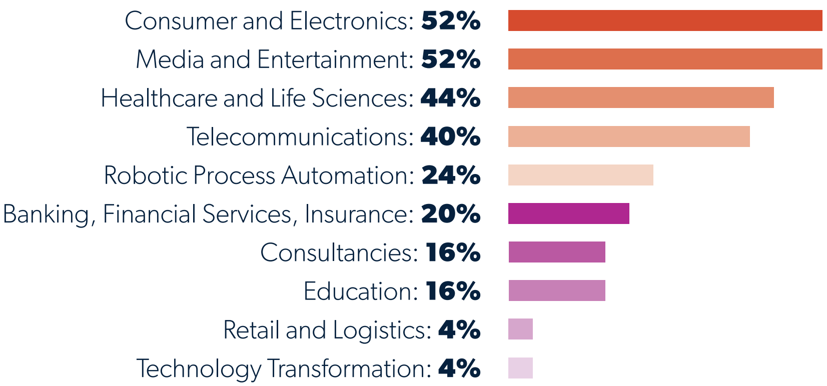
COVID-19 INFLUENCE

There is no hiding the impact COVID-19 has had on the world's economies and within those, specific industries. Some industries have experienced significant negative impact worldwide including travel, high street retail and hospitality. On the other hand, we have seen substantial positive impact on other industries as we have had to change and adapt our ways of working and living to deal with the persistent pandemic.

INDUSTRIES THAT HAVE EXPERIENCED THE BIGGEST POSITIVE IMPACT AS A RESULT OF COVID-19

Respondents think that the industries that have experienced the biggest positive impact as a result of COVID-19 are:

FIG 6.



52% of respondents think the consumer and electronics industries have experienced the biggest positive impact as a result of COVID-19 in 2020. As the rate of media content consumption has increased, so too has the adoption of consumer electronics. Consumers are investing in new technology for personal use to enhance at-home entertainment. New computers, smartphones and televisions can enhance the viewing experience and make home entertainment more enjoyable when public entertainment is inaccessible.

44% of respondents think the healthcare and life sciences industry experienced a positive impact in 2020 with the industry in the spotlight around the world as a result of the pandemic. With continually developing healthcare breakthroughs in general but also the new development of ventilators and vaccinations. **40% of respondents also think the telecommunications industry has also experienced significant positive impact as a result of COVID-19** due to the increased demand on web conferencing and collaboration tools. Due to home isolation, we have had to rely almost solely on the telecommunications industry to work and socialize.

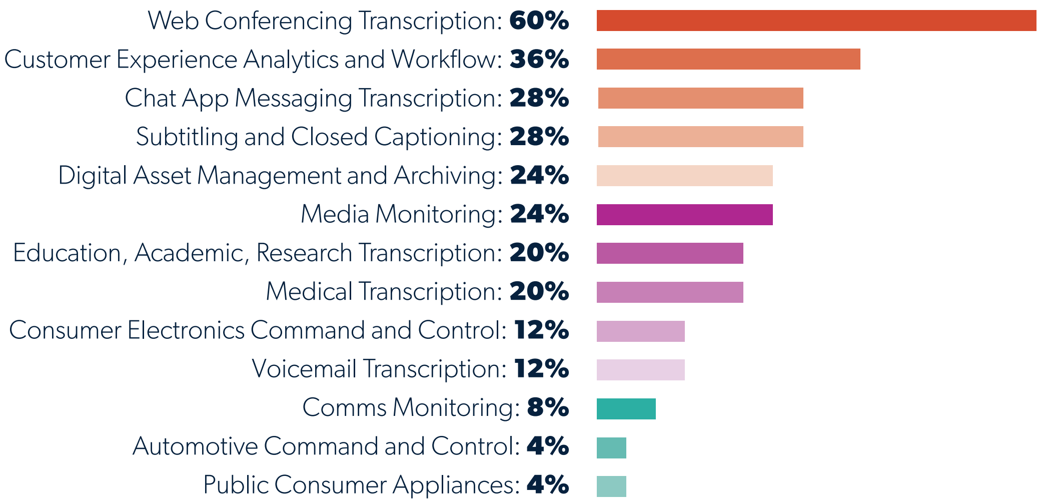
52% of respondents also think the media and entertainment industry has experienced a significant positive impact as a result of COVID-19. The growth in demand for social media, online videos and OTT streaming services since the beginning of 2020 has been substantial and this is expected to continue on the same trajectory into 2021 and beyond.

The impact on both the consumer electronics industry and media and entertainment industry go hand-in-hand with increased pressure to stay entertained at home. The requirement for speech technology in media solution workflows has become essential. Real-time or pre-recorded transcription enables media businesses to enhance their solutions whether it be for accessibility reasons or making processes more efficient to deal with the increased volume of content that needs processing each day.

USE CASES THAT WILL HAVE EXPERIENCED THE BIGGEST POSITIVE IMPACT AS A RESULT OF COVID-19.

Respondents think that the use cases that will have experienced the biggest positive impact as a result of COVID-19 are:

FIG 7.



With the overwhelming majority, **60% of respondents think web conferencing transcription will have experienced the biggest positive impact as a result of COVID-19.** With web conferencing platforms being adopted to suit home working across the world, the subject of captions and transcription has cropped up continuously. Web conferencing has become the best – and in some cases only – way of communication overnight.

Savvy businesses recognized a huge opportunity when it came to accessibility of these video conferencing platforms. The requirement for captions became a must to ensure conversations were accessible and understood by everyone. Transcription enables web conferencing calls to be transformed into text which can then be analyzed, reviewed and stored using metadata, providing a 360-degree view of any interaction or used simply as automatic meeting minutes.

36% of respondents think customer experience and analytics has seen a significant positive impact as a result of COVID-19. Since the pandemic became an international emergency, customer experience has been highlighted as a major issue within contact centers. Not only are agents now dealing with more customer issues by phone, but customers are also more vulnerable than ever and so must be treated sensitively on a case-by-case basis.

HAS COVID-19 HAD PROFOUND EFFECT ON THE MASS ADOPTION OF VOICE TECHNOLOGY INTO BUSINESS AUTOMATION WORKFLOWS?

YES: 40%
NO: 60%

60% of respondents believe COVID-19 hasn't had a profound effect on the mass adoption of voice technology into business automation workflows, while 40% think it has. Looking at the data in this report as a whole, many respondents already had a voice technology strategy implemented and regard voice as an already widely adopted technology. This could be an explanation for the even split.

The respondents that believe COVID-19 has had a profound effect, recognize that many industries and use cases had significant gaps in their voice technology strategies when the pandemic hit. Businesses in these industries – such as web conferencing – assessed the risks of not adopting a voice strategy and opportunities associated with adding voice technology into their workflows and have seen significant growth as a result of wide adoption.

> **"Amongst all the AI-driven workflows that have erupted in the media production industry over the last few years, speech recognition is proving to be one of the most useful and applicable to real-life use cases. Newsrooms, for example, are seeing the time to deliver their stories to their audiences significantly reduced thanks to the ability to quickly locate relevant content by searching for keywords on automatically indexed speech. Hours of properly catalogued archive content can be easily restored based on the requirements of each story.**
>
> **Our production tools combine the power of speech recognition with context-driven content suggestions, while a reporter or producer write their script. Rough cut editing is also possible based on specific keywords as logged by the speech recognition engine. These technologies reduce manual work and accelerate production, as well as improving the overall quality of the stories being told".**
>
> **Bea Alonso**, Chief Market Officer, Dalet

WILL THE INCREASED DEMAND ON COLLABORATION TOOLS FROM COVID-19 CONTINUE INTO 2021 AND BEYOND?

YES: 96%
NO: 4%

The overwhelming majority of **respondents (96%) think that the demand on collaboration tools will increase from 2021 and beyond.** 2020 set a new way of working and living across the world. It taught businesses that many jobs can be done from home with no need for commuting or expensive office leases. This new approach to working – whether that is a hybrid home/office approach or full remote working – has now become the norm and many businesses have stated that they wouldn't return to the way they were operating before the pandemic hit. As a result, businesses will become more reliant on collaboration tools to ensure business stays efficient and productive.

MEDIA AND ENTERTAINMENT INDUSTRY OVERVIEW

WE SURVEYED MEDIA AND ENTERTAINMENT PROFESSIONALS FROM A RANGE OF SECTORS, INCLUDING:

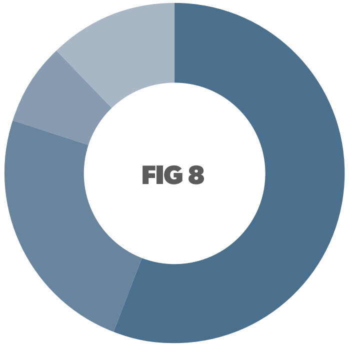


FIG 8. INDUSTRY

⬢	**56%**	Media and Entertainment
⬢	**24%**	Transcription
⬢	**8%**	Education
⬢	**12%**	Other

MEDIA AND ENTERTAINMENT

Media monitoring

According to a **report from MarketsandMarkets**, the global media monitoring software market is expected to generate around $7 billion by 2027, at a compound annual growth rate (CAGR) of around 13.9% between 2019 and 2027.

Subsegments of media monitoring include:

- Broadcast Monitoring
- Social Media Monitoring
- Online Monitoring
- Print Monitoring

Market verticals for media monitoring include:

- IT and Telecommunications
- Retail and Consumer Goods
- Banking, Financial Services, and Insurance (BFSI)
- Media and Entertainment
- Travel and Hospitality

Digital and media asset management

According to a **report by MarketandMarkets** the media asset management market is projected to grow from over $2 billion in 2017 to over $5 billion by 2022, at an expected compound annual growth rate (CAGR) of 18.3%.

Digital & media asset management **solutions** have become increasingly popular due to the increase in the volume of digital assets that are created daily. With a monumental rise of digital data that is created from so many existing and new channels, it's no surprise that organizations are having to seek out new tools and solutions to manage new and existing assets.

Over half the globe (**almost 4.57 billion people**) have access to the Internet. However, the number of devices able to access the web is up to three times as high as the global population. These statistics of growing video adoption highlight the growing need for advanced solutions for broadcasters and media facilities to improve operational efficiency by streamlining their workflows.

> **"Organizations want to manage their media assets from ingest to distribution. An organization has more and more media to ingest and more and more media to distribute and is looking for a solution. A good example used by our customers: a MAM is a kind of a 'factory' of media where stuff gets in and other stuff (transformed) gets out."**
>
> Dalet

> **"DAM involves the creation of an archive, the development of an infrastructure to preserve and manage digital assets and a search functionality that allows end users to identify, locate and retrieve an asset. At its simplest, a DAM is a set of database records. Each database record contains metadata explaining the name of the file, its format and information about its content and usage. Digital asset management software can be used to create and manage the database and help the company to store rich media in a cost-effective manner."**
>
> Tech Target

Captioning and subtitling

According to a **report by MarketWatch**, the global captioning and subtitling market size is expected to reach $370 million by the end of 2025, with a compound annual growth rate of 7.7% during 2019-2025.

Captioning and subtitling comprise encoding, editing, and repurposing of video subtitles and captions for delivery platforms such as web, mobile, and television.

Captioning and subtitling media assets help broadcasting and web media organizations to automate the insertion of captions into a huge amount of audio and video content that is produced daily.

The key driver behind captioning and subtitling is to support accessibility in all forms of communication. As the amount of video content grows, the ability to provide captions for it all becomes more challenging and more costly. On top of this, legislation – in particular, the Federal Communications Commission (FCC) – is moving fast and some tough targets have been put in place for media companies providing captioning for television and online content in the United States. Currently, **29% of the market is using human processing** as their solution to captioning, however, the costs are high and require a great deal of human resource to transcribe, align and position captions.

A key reason for the use of human processing is because the media and broadcast market has a very high accuracy demand. Most companies accept a 0–1% word error rate (WER) for most use cases, and at the moment this is only achievable using technology in rare circumstances. For cases such as broadcast news, 0-1% WER is achievable using voice technology. However, cases with noisy environments, over talk, multiple speakers, singing or other musical elements may require a combined human and machine approach. The ability for speech solutions to deliver a WER of less than 10% and at a much faster rate than humans for pre-recorded and real-time content, provides significant advantages.

> **"Accelerated by the global pandemic, 2021 and beyond will see an unprecedented change in how video is used as well as the more obvious hyper-growth in video. Naturally, there'll be significant challenges and opportunities that companies will face. Compliance to accessibility for companies who have shifted their entire operations to remote working is one such challenge, but with that comes transformational opportunities. Real-time transcription and translation of live events can and will breakdown language barriers, opening up markets and removing barriers between operating territories and teams."**
>
> **James Jameson**, Commercial Director, CaptionHub

Editing

According to a **report by GlobeNewswire**, the video editing software market is projected to grow at a compound annual growth rate (CAGR) of 2.6% to reach $932.7 million by 2025. The growth in video content creation and consumption, as well as increased usage of the Internet and hand-held devices, has fuelled the growth of this market.

Voice technology has been used by media and broadcast companies for some while. Previously, media companies were required to have large teams of editors to edit transcripts and ensure they were accurate. This method was time-consuming especially in the cases where a large number of files required checking and editing in parallel.

The value of speech technology is no different today to what it was when it was first commercially available. However, the rise of artificial intelligence and machine learning has provided a huge step change for voice technology, enabling the delivery of WERs of less than 5%. The uplift in accuracy from voice technology means that there is less to edit but it doesn't make editing teams redundant. Instead, organizations are making editing teams more efficient and their specialist skills are adding value to more files than ever before.

TRANSCRIPTION

The global transcription industry was estimated to be worth just over $3 billion in 2015 and is expected to reach almost $10 billion by 2020.

The industry is often divided into two categories:
1. **Medical Transcription**
2. **Non-medical Transcription** (includes legal transcription, audio visual transcription, interview transcription etc.)

Most transcription providers in the market cover both categories and work across B2B and B2C use cases. With the increased demand for transcription, many providers offer a range of services with varying accuracy levels whether that be pure automatic speech recognition or a combination of machine and human editing.

Automatic speech recognition technology does the heavy lifting by transforming audio into a text-based format. However, humans can pick up on other elements such as laughter, silence, music and other nonverbal elements within a media file which are essential to gaining 100% accuracy and understanding in a transcript.

With an increase in online recordings since the pandemic began, there are now more interactions trapped in audio files that need transcribing. Voice technology is essential to help transcription solutions cope with the increased volume, time and cost that are critical to transcription services.

> **"The COVID-19 pandemic has had an impact on business at TranscribeMe but at the same time, customer expectations for word accuracy and turnaround time has not changed. In order to meet those expectations, TranscribeMe partners with Speechmatics to combine the best of automated speech recognition (ASR) with human correction in a hybrid workflow that delivers text output consistently above 99% accuracy within the customers' expected turnaround times. Working together we lower the cost, scale up our delivery capabilities, and provide our customers with the best quality transcription products."**
>
> **Anthony Ettaro**, CTO, TranscribeMe

EDUCATION

E-learning

According to a research study by **GlobeNewswire**, the global e-learning market was estimated at $144 billion in 2019 and is expected to reach more than $374 billion by 2026.

The global e-learning market is expected to grow at a compound annual growth rate (CAGR) of USD 14.6% from 2019 to 2026.

E-learning or electronic learning is an activity of learning or training through digital resources. The demand for e-learning was always expected to grow due to the growth of the telecommunication and consumer electronics industry, however, the pandemic has dramatically accelerated the market growth.

While it was expected that e-learning would play an important role in strengthening the primary and supplementary education system in the future, it was completely unexpected that e-learning would replace classroom teaching across the world in 2020.

With the improvement of telecommunications and devices in the home, it was possible in most countries to transition to e-learning in a timely manner and utilize the solutions on the market. Speech technology enhances the e-learning experience by providing a transcript alongside a video lesson or by providing captioning to videos so students can listen and read simultaneously.

“At Ai-Media we have experienced growth through our live and recorded captioning services in the education sector, both of which have been used extensively in the past year due to the shift in how people work and educate themselves. What started off as a COVID safe plan has now become the new norm and we continue to see growth as we adapt and scale with the market.”

Shannon Trembath, Sales Manager, Ai-Media

CURRENT MARKET ADOPTION OF VOICE TECHNOLOGY IN THE MEDIA AND ENTERTAINMENT INDUSTRY

DOES YOUR COMPANY HAVE A VOICE STRATEGY?

48% of respondents said that their company currently has a voice strategy.

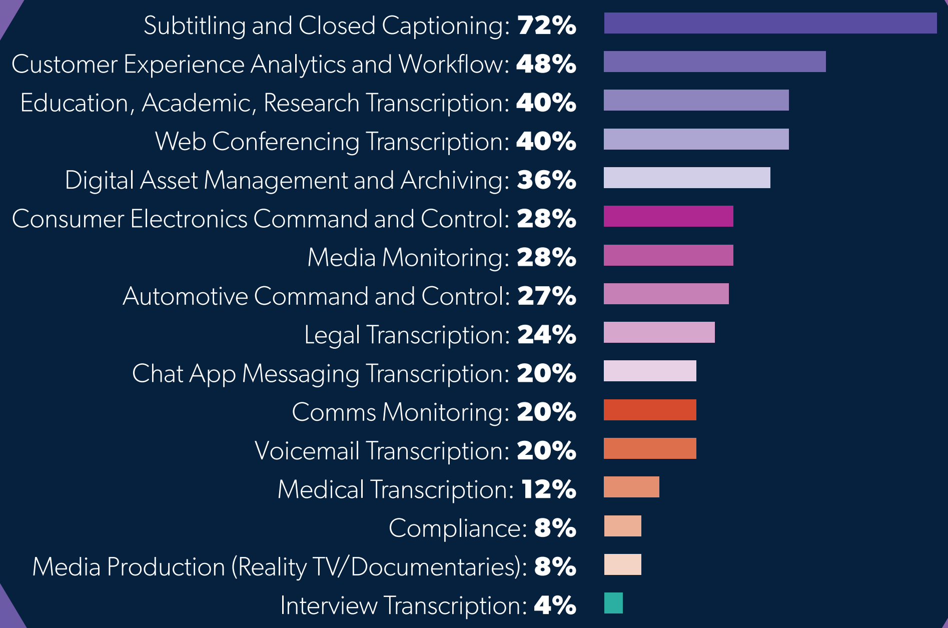
Of those that don’t have a voice strategy, 52% said it is something that will be considered in the next 5 years.

YES: 48%
NO: 52%

CURRENT APPLICATIONS FOR VOICE TECHNOLOGY

Respondents indicated that there are currently many applications for voice technology. They are listed below.

FIG 9.

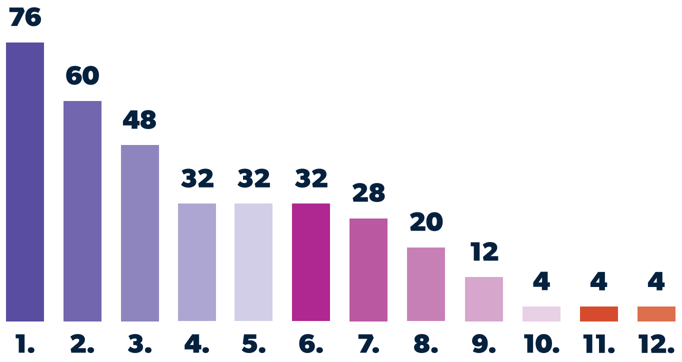


The overwhelming majority of respondents (72%) think the main application for voice technology is currently subtitling and closed captioning. There is no hiding the fact that video content – and the need for subtitling and closed captioning – has grown exponentially since the pandemic hit but it isn't just for pre-recorded video content. Live video content has started to increase in demand since the introduction of Facebook Live and Instagram Live as well as other new channels. **Cisco** predicts that live internet video will account for 17% of internet video traffic by 2022. To ensure accessibility and adherence to FCC regulations, voice technology enables the automation of captions and subtitles at scale.

From applications that have been positively impacted by the pandemic such as customer experience analytics and workflow, education, academic and research transcription, and web conferencing transcription, voice technology brings operational efficiencies and improvements to businesses of all sizes.

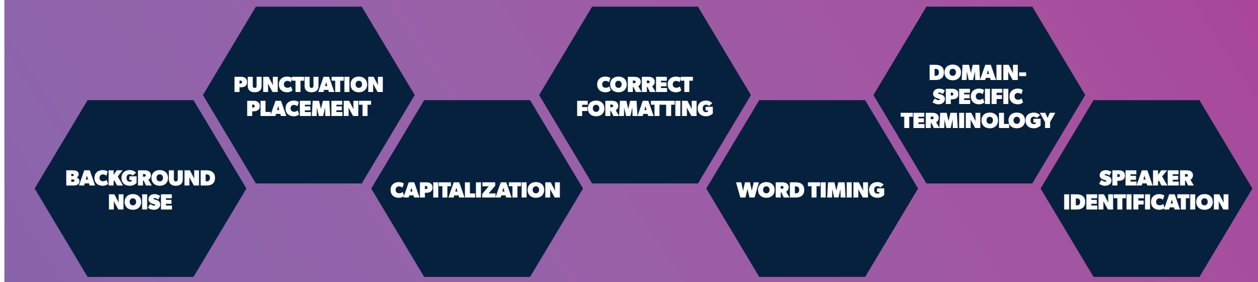
BARRIERS TO ADOPTING VOICE TECHNOLOGY

FIG 10. SOME OF THE BARRIERS TO THE ADOPTION OF VOICE TECHNOLOGY ARE INCLUDED BELOW.



⬢	**1.**	**76%**	Accuracy
⬢	**2.**	**60%**	Accent or dialect related recognition issues
⬢	**3.**	**48%**	Language coverage
⬢	**4.**	**32%**	Cost
⬢	**5.**	**32%**	Data security and privacy
⬢	**6.**	**32%**	End user expectations
⬢	**7.**	**28%**	Complexity of deployment and integration
⬢	**8.**	**20%**	Flexibility to meet the use case
⬢	**9.**	**12%**	Operational support
⬢	**10.**	**4%**	Backlash from humans
⬢	**11.**	**4%**	Establishing norms and agreements for usage
⬢	**12.**	**4%**	Sophistication required to derive meaning from voice compared to conventional UIs

76% of respondents believe that accuracy is the biggest barrier when it comes to adopting voice technology within their business. These days, accuracy represents more than just the accuracy of the word output – known of as word error rate (WER). With the most spoken languages in the world at a consistently low WER, many other factors affect the level of accuracy on a case-by-case basis. These factors are often unique to a use case or a business' needs.


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

For media solutions, it is no surprise that the biggest barriers to adopting voice technology are **accent/dialect related recognition issues (60%) and language coverage (48%)**. These barriers go hand-in-hand with accuracy, however, voice technology vendors can address these as separate challenges.

Language coverage is essential to businesses when operating globally and high accuracy is needed in multiple languages beyond English. A provider may offer English, but which accents and dialects do they support? Providers are often challenged by thick accents, causing issues when it comes to the transcription output. What happens when an American is speaking with a British person, for example? Which accent variation is used?

A **global approach to languages** solves this challenge. By incorporating data drawn from global sources, encompassing a variety of accents, providers can offer global language packs, so users don't have to worry about different accents in their video or audio.

With the continual advancements in artificial intelligence and machine learning, the use of data is becoming more accessible. With this data accessibility comes better, more accurate speech recognition engines better equipped at transcribing multiple accents, dialects and languages.

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 INCLUDED BELOW.

- PRIVACY, IDENTITY PROTECTION, COMPLIANCE
- STRONG COMPETITION FROM THE LARGEST TECH COMPANIES (GOOGLE, IBM, MICROSOFT, AMAZON)
- "LONG TAIL" OF IMPERFECT RECOGNITION DAMPENING CONFIDENCE IN MORE CRITICAL USES
- THE SUPPORT OF REGIONAL ACCENTS STILL NOT BEING GOOD ENOUGH
- ACCURACY THROUGH CONTEXT AND LEARNING THE VOCABULARY OF THE SPECIFIC USER
- NEED FOR IMPROVED SPEAKER IDENTIFICATION
- LACK OF CONSUMER ADOPTION OR RELUCTANCY TO USE THE TECHNOLOGY DUE TO SECURITY AS WELL AS POOR EXPERIENCE (I.E. NOT WORKING AS WELL AS IT SHOULD)
- COPING WITH CODE-SWITCHING AS USED IN LOW RESOURCED MINORITY LANGUAGES
- PARITY BETWEEN DIFFERENT LANGUAGES (ACCURACY, ETC.) AS IT BECOMES THE NORM
- PEOPLE COULD GET FRUSTRATED WHEN SPEECH TECHNOLOGY DOES NOT PERFORM AS EXPECTED
- ADOPTION COULD GET TOO FAR AHEAD OF THE TECHNOLOGY AND ERODE CONFIDENCE
- STAGNATION DUE TO PROBLEMS OF INTEGRATING CONTEXTUAL INFORMATION TO IMPROVE TRANSCRIPTION EVEN FURTHER
- PRIVACY CONCERNS RAISED BY DATA-MINING AUDIO SOURCES FOR TEXT-BASED CONTENT
- LEGAL RISKS AND POLICY
- NOT GETTING NEARER TO 100% WER
- HIGHER BANDWIDTH FOR HCI THAN SPEECH EMERGING FROM WEARABLES OF ONE KIND OR ANOTHER
- LANGUAGE DIVERSITIES (IDIOMS, DIALECTS) AND FAKE NEWS/FAKE MESSAGES
- UNREASONABLE EXPECTATIONS FROM CONSUMERS OF THE TECHNOLOGY (EXPECTING NEAR PERFECTION)
- LITERACY AND NUMERACY SKILLS NOT BEING DEEMED AS IMPORTANT
- A BETTER MANAGEMENT OF SERVER CONSUMPTION IN AN ENVIRONMENTAL SENSE
- PRICE CAN BE THE MAJOR HURDLE FOR WIDER ADOPTION

FUTURE CONCERNS AROUND DATA PRIVACY

YES: 88%
NO: 4%
NOT SURE: 4%

88% of respondents said that data privacy will be a concern in the future. People are concerned over where their data is stored and how. Data collection is a growingly important topic. Providing different deployment options for ASR – including on-premises – seems to be key to consumers being at ease with how their data is handled.

Consumers want to be able to access or maintain their data themselves and are looking to have complete transparency in the process. These concerns are an increasing trend given the track records of the tech giants such as Google and Amazon who freely use consumer's data. This makes users wary of the usage of personal data and how it could impact them. Consumers are also becoming savvier and want assurance and clarity to questions about how their data is being collected, stored, owned and utilized. Those conditions are influencing how users choose ASR providers.

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

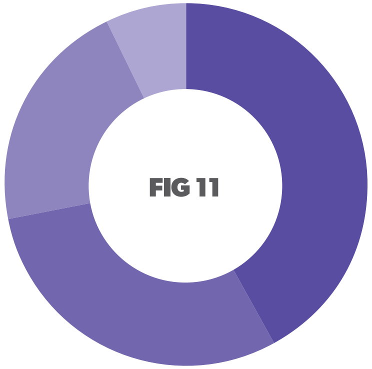
Data security continues to be a concern across all industries. In 2019, concerns around voice data were prominent in the news with global brands such as **Amazon** and **Google** confirming they are using their home devices to "listen" to conversations for development and improvement to their devices.

In 2020, these articles swiftly prompted a **reaction from the tech giants** to communicate the importance of data privacy. With such prominent voice data breaches, businesses are planning to overcome these challenges and ensuring data privacy is front of mind when integrating voice technology into their workflows.

We've established that 88% of respondents believe that data privacy will continue to be a concern in the future, but there will be ways to overcome data security issues.

FIG 11.

⬢	**42%**	On-premises deployment
⬢	**30%**	Cloud deployment
⬢	**21%**	Solutions able to operate entirely offline
⬢	**7%**	Other



On-premises deployment

42% of respondents said that on-premises deployment is a great way to overcome issues with data privacy now and in the future. On-premises deployment options enable users to keep their data secure within their own environments with no need for data to go into the cloud. On-premises deployments for voice technology are often done using virtual appliances or containers so they can be deployed effortlessly into existing technology stacks.

The media and entertainment industry has generally operated its solutions in the cloud up until now. But with growing concerns around data privacy, businesses are increasingly looking to on-premises deployments to ensure they keep data secure within their own environments. It is essential that speech technology vendors can deploy the technology wherever organizations need it.

Secure deployments will also become essential to the e-learning sector given the volume of children's voice data that will be processed. Solution providers utilizing speech technology as part of their e-learning platforms will require full data privacy for transcription of lessons that contain student interactions.

Cloud deployment

30% of respondents said cloud deployments satisfy their business need for data security. Private cloud deployments are secure enough to keep data safe for lots of applications. If cloud deployment security is good enough for the business and use case needs, cloud deployment is often the preferred option due to low operational cost and complexities.

Private cloud deployments are often secure enough for businesses operating in the media and entertainment industry due to the majority of the data that is processed being publicly available or already broadcast.

Dark site environments

Dark site deployment options enable customers to keep their data secure within their own data center environments. **21% of respondents said solutions that can operate entirely offline will alleviate their concerns around data security**. 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 meaning all work is completed within a business' private environment.

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

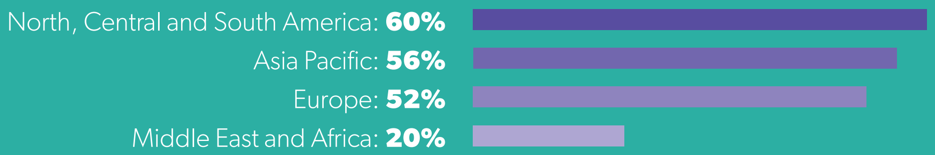
VOICE TECHNOLOGY CONSIDERATION

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

60% of respondents indicated that North, Central and South America 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. The media and entertainment industry is dominant in these regions and it is also highly regulated.

Following on from there, Asia Pacific and 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. In the North, Central and South American regions media and entertainment solutions are widely adopted and advanced in their processes and workflows, whereas in the European and Asia Pacific regions, adoption (until the pandemic hit) was slower. 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.

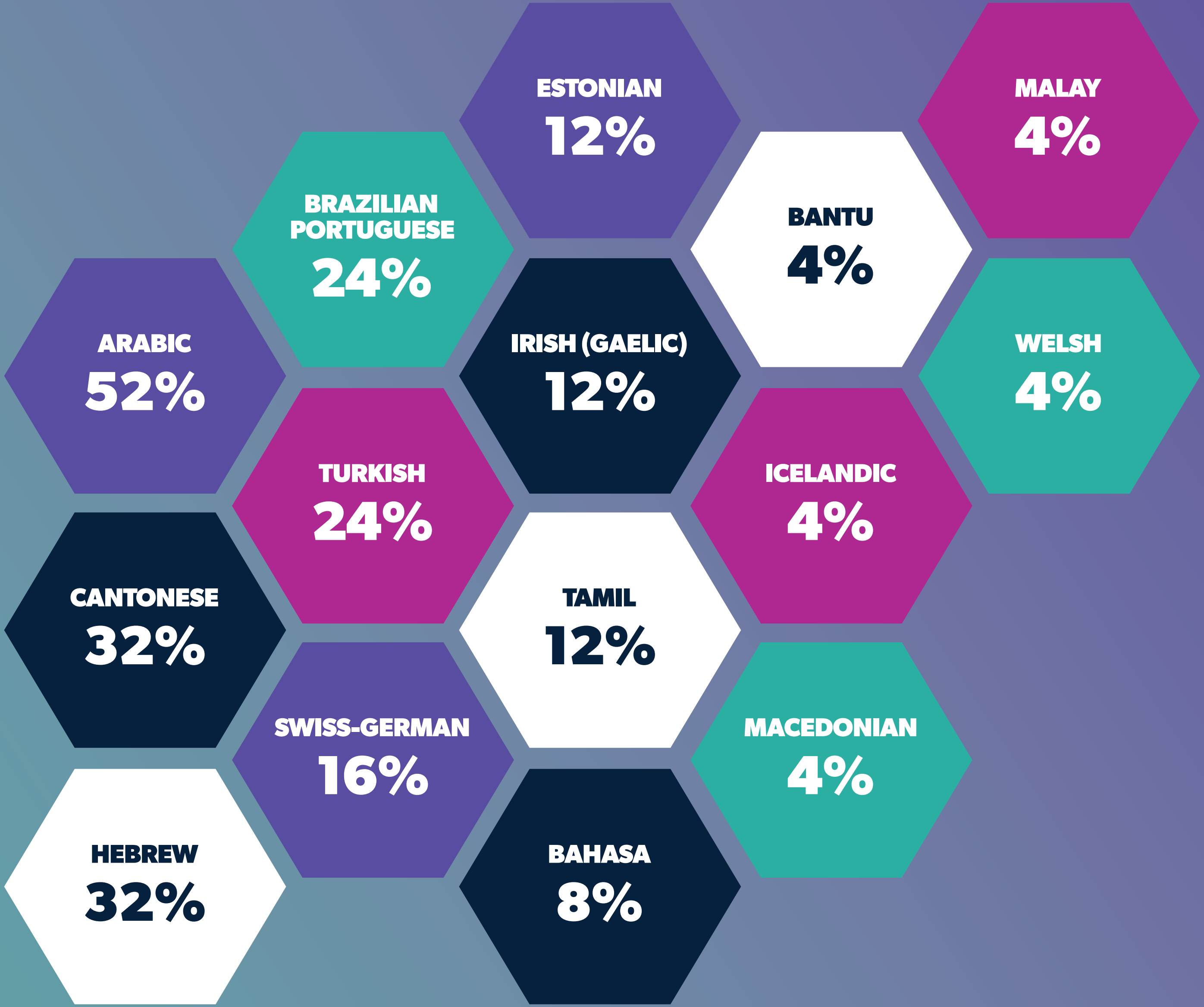
FIG 12.



LANGUAGE SUPPORT IN THE NEXT THREE YEARS

Over the next three years, respondents expect ASR language coverage for media and entertainment solutions to include the languages listed to the right. As well as specific languages, respondents also acknowledge the need for accent-independent language packs, 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 three years.

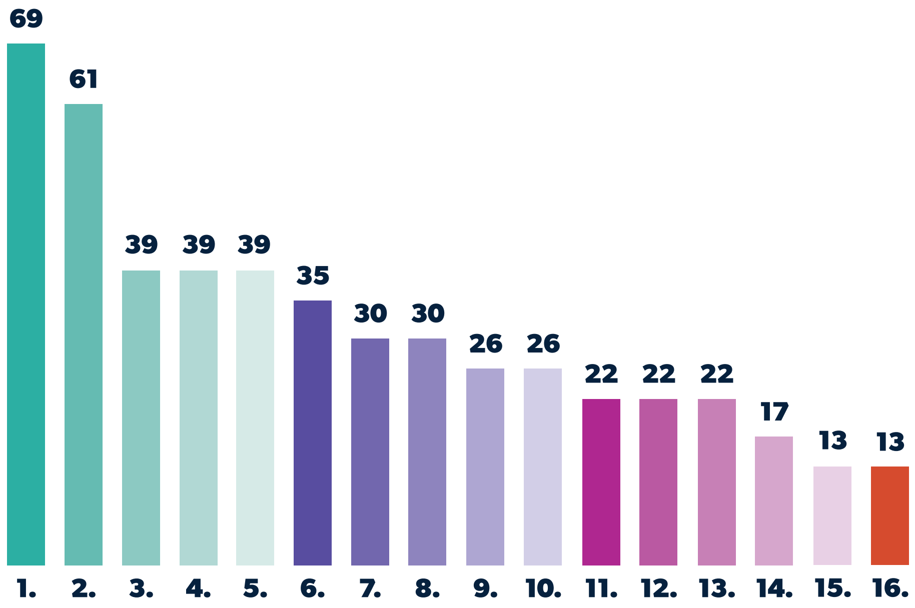
With so many languages having such a diverse range of accents within them, ASR will need to incorporate systems that are trained to be accurate when it comes to deciphering and transcribing those languages to better serve users and increase accuracy in the longer term.



ARABIC	52%
CANTONESE	32%
HEBREW	32%
BRAZILIAN PORTUGUESE	24%
TURKISH	24%
SWISS-GERMAN	16%
ESTONIAN	12%
IRISH (GAELIC)	12%
TAMIL	12%
BAHASA	8%
BANTU	4%
ICELANDIC	4%
MACEDONIAN	4%
MALAY	4%
WELSH	4%

FEATURE DEVELOPMENT IN THE NEXT THREE YEARS

FIG 13. RESPONDENTS SAID THAT VOICE TECHNOLOGY FEATURE DEVELOPMENT WILL BE CRUCIAL OVER THE NEXT THREE YEARS, THESE FEATURES INCLUDE:



⬢	**1.**	**69%**	Increased word error rate (WER) accuracy
⬢	**2.**	**61%**	Increased speaker diarization accuracy
⬢	**3.**	**39%**	Increased accuracy of number recognition
⬢	**4.**	**39%**	Language identification
⬢	**5.**	**39%**	Translation
⬢	**6.**	**35%**	Customer-specific language models trained on customer text data (Language Model Adaptation)
⬢	**7.**	**30%**	Non speech detection (detect sounds, noises, music, disfluencies, hesitation, silence)
⬢	**8.**	**30%**	Real-time transcription from the cloud
⬢	**9.**	**26%**	Customer-specific language models trained on customer acoustic data (Acoustic Model Adaptation)
⬢	**10.**	**26%**	More languages available
⬢	**11.**	**22%**	Audio file quality assessment
⬢	**12.**	**22%**	Increased speed/latency
⬢	**13.**	**22%**	Short utterance accuracy
⬢	**14.**	**17%**	Noise reduction
⬢	**15.**	**13%**	Redaction
⬢	**16.**	**13%**	Word alternatives available in the output

IMPROVED WORD ERROR RATE ACCURACY

69% of respondents said that they would like to see better accuracy over the next three years. WER will continue to improve by throwing more data at the problem, however, this approach will likely see diminishing returns. The leading providers in the ASR space continue to deliver accuracy of 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. Providers can also look to shift focus to delivering supporting features that enhance the quality of output provided to its users.

Features like entity tagging within the audio, identification of the languages spoken and better diarization will all count towards the delivery of a more accurate representation of the audio as part of the files transcribed.

As mentioned previously in this report, all media and entertainment use cases require near perfect or 100% accuracy for speech-to-text. With increasing volume and demand for video and broadcast content, it is essential for these organizations to incorporate ASR technology into their workflows, but it needs to be accurate.

WER improvements will likely be incremental compared to the past few years. However, with issues around bias continuing into 2021, it's likely that improvements can be made across all languages, particularly when it comes to accents and dialects.

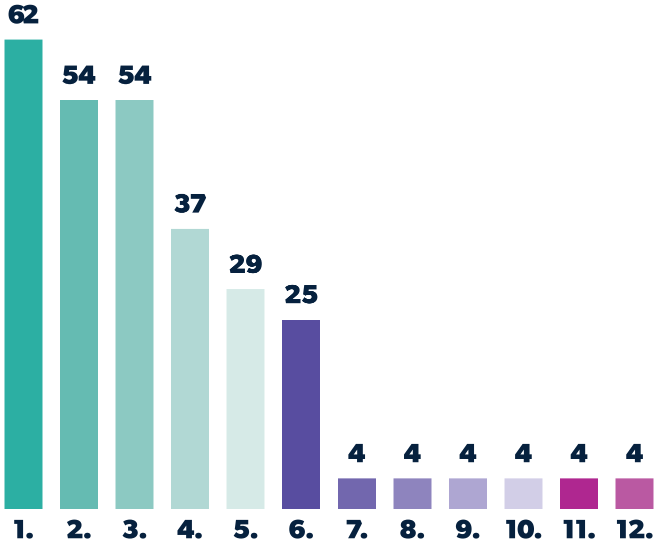
SPEAKER DIARIZATION

61% of respondents said that they would like to see better speaker diarization accuracy over the next three years. Speaker diarization is used to understand which speaker was talking in single-channel media files. It does this by detecting unique speakers and assigning speaker labels to the corresponding portions of text within the transcript. It's one thing to know what was said, but the means to split the transcript by speaker adds additional value to a range of users.

Speaker diarization is one of the most challenging elements of speech recognition and one of the most important features when it comes to media solutions. While speech and other audio characteristics are easy for the human brain to detect, this poses a challenge for automated systems due to the fluctuations in a single speaker's voice depending on their mood, hesitations, word emphasis, noise etc. While speaker diarization exists today, it is still a key challenge that speech providers have not yet mastered.

2021 will likely see increased effort to improve speaker diarization to uplift use cases that benefit from being able to match a speaker with the words spoken.

FIG 14.
(%)



⬢	1.	62%	Word error rate
⬢	2.	54%	Punctuation
⬢	3.	54%	Speaker change indicated
⬢	4.	37%	Intent recognition
⬢	5.	29%	Multiple languages
⬢	6.	25%	Quick turnaround time for transcription
⬢	7.	4%	Background noise
⬢	8.	4%	Character error rate
⬢	9.	4%	Concurrent or agitated speakers
⬢	10.	4%	Dialects
⬢	11.	4%	Good quality audio
⬢	12.	4%	Multiple languages in 1 audio

PERCEPTION OF ACCURACY

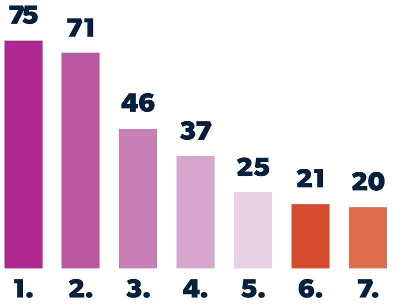
The perception of accuracy for voice technology is very specific to the use case and business. There are several measures that respondents said are important when it comes to accuracy in the media and entertainment industry, these include:

The perception of accuracy for voice technology is often interpreted as the word error rate (WER). **62% 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 use other metrics to measure the accuracy of the transcription output – especially in the media and entertainment industry where other factors are crucial to the accuracy of the output.

Other crucial factors that respondents deemed indicators of accuracy were punctuation (54%) and speaker change indicated (54%). Understanding the conversation flow provides an added level of accuracy to the transcript beyond word output accuracy. The media and entertainment industry finds these features incredibly valuable in real-time scenarios as well as post-processing as they improve the readability of transcripts or captions immediately.

THE FUTURE OF SPEECH RECOGNITION FOR MEDIA AND ENTERTAINMENT INDUSTRY APPLICATIONS

FIG 17. RESPONDENTS SAID THAT THERE ARE LIKELY TO BE CHANGES IN THE SPEECH RECOGNITION MARKET. IMPORTANTLY, THE CHANGES WILL BRING IMPROVEMENTS TO THE TECHNOLOGY THAT WE SEE TODAY. SOME OF THOSE KEY IMPROVEMENTS AND CHANGES ARE OUTLINED BELOW.



⬢	1.	75%	Improved word error rate (WER) (95%+)
⬢	2.	71%	More robustness when it comes to noisy environments
⬢	3.	46%	More focus on any-accent language models
⬢	4.	37%	Less data required to improve models
⬢	5.	25%	More on-device deployments through reduced model size
⬢	6.	21%	Large companies will offer more deployment options
⬢	7.	20%	Enterprise-grade ASR is no longer just speech-to text

75% of respondents said the future of speech recognition is improved word error rate (WER). Looking back to the perceptions of accuracy, low word error rates are regarded to be the main definition of accuracy. As machine learning algorithms continue to evolve, it is likely that WER accuracy will reach 95%+ especially for commonly used languages like English. However, there is still significant work to do due to the range of accents and dialects in all languages and to deliver the same level of accuracy across them all. ASR providers and users need to be objective around how they conduct testing to ascertain WER scores and understand what they mean for them and their customers. For media solutions, other KPIs such as speed of output and punctuation placement are already critical and will become increasingly important.

71% of respondents said the future of speech recognition should see more robustness when it comes to noisy environments. It's no surprise that the ability to deal with noisy environment is a key consideration for the future of speech recognition. Noise is a major factor that impacts accuracy. This is particularly pertinent in the media and entertainment industry where you might have a sports broadcast with lots of background noise for example.

> **"Quality is still the first word on media customers' lips in a conversation about ASR. So improving word error rate, and getting consistently high quality across a range of content and programme genres is still a clear market priority. However, the importance of features beyond transcription is growing rapidly. The most important of these is speaker diarization and differentiation; it will be a critical breakthrough for media applications of ASR, especially subtitling and captioning. There's also a more general need to better interpret the acoustic media space, by identifying music and non-speech sounds for example.**
>
> **ASR is being seen as an increasingly viable tool across the media industry, thanks to the massive improvements in quality we've seen recently. One consequence of that is that customers are getting more demanding and are looking to convert the considerable promise of ASR into reality. At Red Bee, we see our partnership with Speechmatics as being a critical part of bringing the promise of this technology into viable broadcast and online use cases. The regular leaps in quality we've seen from Speechmatics do a lot to reassure both us and our customers that we're only at the beginning of showing the value of ASR and automatic captioning."**
>
> **Tom Wootton**, Head of Access Services Product and Portfolio, Red Bee Media

It was also highlighted as a risk that audio quality may suffer due to the pandemic as the population are required to wear masks and other PPE. If ASR engines are unable to detect the words of a speaker due to background noise, they are unlikely to produce an accurate transcript. The ability to reduce interference or deliver high-quality recognition even in challenging environments, therefore, remains a top priority.

Contemporary ASR solutions are incredibly effective even in noisy environments or when media is recorded on low-quality devices. ASR providers will continue to improve the diversity of their training models to include challenging audio profiles to deliver greater robustness when dealing with background noise.

SUMMARY

It goes without saying that 2020 was a year like no other and we are going to feel its effects for a long time. One thing we can take from the experience is how important it is to ensure we are continually evolving and adapting our products and solutions by embracing new technologies.

Automated workflows and processes will enable media and entertainment companies to scale and meet the increasing consumer demand for their solutions while still delivering exceptional customer experiences. With the rapid change and adoption in 2020 of on-demand services, media and entertainment companies require more sophisticated methods of extracting metadata within the content. This has seen a rise in demand for better indexing and searchability of digital assets.

The rise of artificial intelligence and machine learning has been instrumental in the increased adoption of speech technologies. Accuracy continues to be the key metric when it comes to evaluating speech providers due to the time and cost saving this brings. Continual reductions in word error rate, therefore, opens up more applications for the technology. The media and entertainment industry is considered to be in the late majority phase of adoption of speech technology, so now it is essential that ASR vendors continue to develop and improve the technology beyond word accuracy.

The benefits of speech technology within the media and entertainment industry are undeniable. From accelerated transcription and reduced dependencies on human transcribers to extracting valuable data from within video and audio content, the media and entertainment industry is seeing efficiency gains and opportunities for product enhancements and business expansion. Digital assets are being managed and indexed effortlessly, meaning they are easier to find by the end user, ultimately improving the customer experience. E-learning adoption has become a requirement and has placed demand on high volumes of captioning and subtitling.

While voice technology has already proven to provide value through automatically transcribing audio and video content at scale, people will continue to be an integral part of the media workflow. Advances in voice technology will mean that organizations can automate large quantities of manual transcription work. Organizations can then make their workforce more productive by augmenting their responsibilities outside of the low-skill tasks and add value to activities currently out of scope by automated processes.

The integration of voice technology into media solutions is truly the only way to scale and meet the ever-increasing demand we see today for speech-to-text applications.