verbit

Exploring the Era of Automatic Speech Recognition

A report on the impact, limitations and opportunities of ASR technologies

About the report

Business leaders across industries are experimenting with Automatic Speech Recognition (ASR) technologies. The speech and voice recognition market stands at $13B currently, and is projected to reach **$60B by 2030**, highlighting its significant impact.

While ASR tools were initially implemented for intelligence needs and in the military sphere, the technology quickly evolved to present promise for the business world. Today, ASR technologies are helping leaders across industries with everything from speech analytics for call centers to addressing accessibility needs for individuals with disabilities to offering greater workplace efficiencies and more engaging experiences.

ASR technologies are being implemented across a variety of use cases, and businesses are currently using a mix of free, built-in solutions and paid ASR services to meet their needs. Three product and speech acoustics experts at Verbit composed this report to explore the use of ASR across industries, its potential and its limitations.

Throughout, you'll find data and **insights acquired from a third-party survey commissioned by Verbit**. The survey gathered responses from 200 senior-level professionals responsible for captioning and transcription and/or Diversity, Equity, Inclusion and Accessibility (DEIA) efforts. They represent companies based in US, UK and Canada with 500 to 5,000 employees. The respondents surveyed work across industries, including Media, Banking, Tech, Retail and more.

This report also explores what drove Verbit's team to launch a more advanced, proprietary ASR technology, **Captivate™**, to bring a stronger ASR offering to the market based on exposed shortcomings.

Contributors

Adi Margolin

Director of Product Management, Verbit

Adi oversees product management and the release of new offerings and features to improve user experiences. She is responsible for managing Verbit's advanced proprietary ASR technology, Captivate™. Her knowledge in the hi-tech industry includes previous experience at enterprise B2B and big data marketing analytics companies.

David Landsberg

VP of Product, Verbit

As an enterprise software leader, David leverages experience, foresight, insight and data to address today's and tomorrow's biggest challenges. His software engineering background and experience leading product teams and cross-functional R&D groups enable him to build innovative and disruptive software products that solve real-world problems.

Dr. Irit Opher

VP of Research, Verbit

Irit is a research manager specializing in Machine Learning and Speech Recognition, with over 20 years of experience in high-tech. Her experience includes algorithm development and research in various domains, including speech and speaker recognition, data analysis, text analysis, video analytics, classification and clustering methods and neural networks. Her MSc and PhD theses focused on Computational Neuroscience.

Introduction

Automatic Speech Recognition (ASR) is transforming how we engage in everyday experiences, including how we work and how we learn. Essentially, ASR technology translates spoken language into text and is being used to improve education, reach wider audiences, enhance accessibility for individuals with disabilities, fuel communication, transcribe complex content for records and more.

70% of organizations using ASR opt for free tools.

Our commissioned survey found that **70% of organizations using ASR opt for free tools**, often built into the applications they are using such as web conferencing or video streaming platforms. However, most of these built-in ASR technologies are generic and not designed to handle the variety of use cases and different needs for which they are currently being deployed, Adi Margolin noted.

“The market expects their technologies to be capable of handling diverse and complex scenarios that mirror everyday situations. Whether it is multiple people speaking over each other, background noise or highly specific terms being used, when users enlist these tools, they expect high performance. However, our analysis has shown that generic ASR technologies often fall short in accuracy or in meeting their users’ specific requirements,” she said.

Despite being viewed as an efficiency-driving tool, ASR users frequently find themselves spending valuable time correcting errors in the output of these built-in tools. **40% of survey respondents cited that they spend significant efforts editing and formatting the results** of the ASR outputs they receive. Plus, the stakes are even higher when companies turn to ASR for accessibility purposes or when applied to use cases in the legal and medical fields where every word must be captured accurately. When companies lean on ASR to provide captioning and transcription, certain legal regulations mandate that they must reach very high accuracy levels. Additionally, in many scenarios, timely delivery of accurate results is crucial, leaving little room for manual editing.

"In most of what we do, we simply cannot make an error," said Dr. Irit Opher, VP of Research, Verbit. "Imagine a mistake in legal proceedings' transcription or a politician's name misspelled in a live news broadcast. It's important to leverage **domain expertise** to proactively train the models, enabling them to anticipate and prevent mistakes. We sought to develop a solution that could consistently enhance its performance; one where teaching the engine once yields significant improvements."

Verbit's speech and acoustic experts conducted thorough research and tests on existing ASR over the last two years with these realities in mind. Together, they dug into trends and uncovered existing limitations of generic ASR tools and compiled helpful information for users on how to get more value from their speech and text data.

Exploring the current use cases for ASR

An overwhelming 97% of professionals surveyed across industries are implementing ASR technologies for more than one use case.

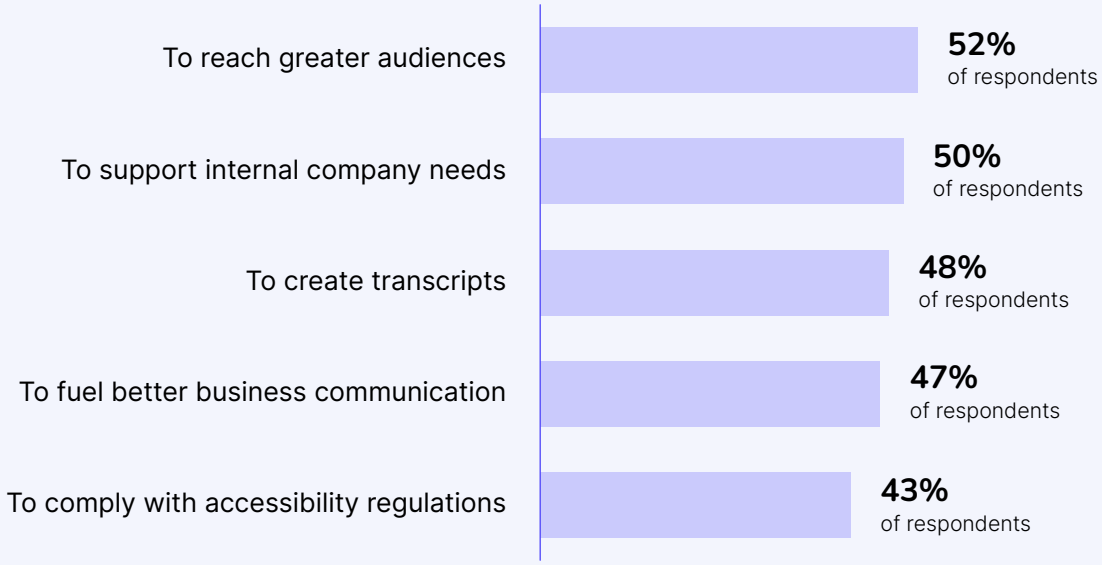
Being able to lean on a solution which can deliver in a variety of environments is proving to be critical. Whether it be to generate live captions for an event or transcripts for efficient record-keeping, professionals need to know that the ASR they're using can address multiple needs.

Some of the most common applications of ASR today include:

- **Live sports and events**
  Displaying live captions on-screen for attendees
- **Media programming and live streams**
  Enabling equitable viewing experiences and increased engagement for viewers
- **Virtual meetings**
  Empowering remote work and global conversations with inclusive communication and documentation
- **Conferences and events**
  Ensuring equity and content comprehension for in-person and remote audiences, plus records of what was shared
- **Legal proceedings**
  Producing transcripts for analysis and documentation of proceedings
- **Medical needs**
  Creating transcripts of doctor-patient conversations, protocols and more
- **Earnings calls**
  Providing transcription of financial discussions for analysts, investors and stakeholders
- **Campuses and online learning environments**
  Enabling inclusive learning experiences in classrooms and online, plus assistive tools across the campus in buildings, for guest lectures, graduation ceremonies and additional events for students, alumni and families
- **Museums and visitor centers**
  Captioning informational content and exhibits to offer inclusive experiences
- **Interactive personal assistants**
  Using speech recognition to understand and respond to user questions

- **Call centers**
  Transcribing customer interactions for speech analytics and quality assurance
- **Marketing and social videos**
  Helping businesses and brands engage effectively online

Top 5 use cases of ASR tools



36% of ASR users find existing tools inaccessible for their needs.

With the use cases for ASR spanning diverse industries and needs, our experts wanted to explore whether off-the-shelf ASR tools – those readily available online or built into existing business platforms - fulfill or miss the mark. They also sought to explore if they enable access to individuals with disabilities and fulfill accessibility requirements mandated by legislation such as the Americans with Disabilities Act (ADA), UK Equality Act and The Accessible Canada Act (ACA), among others. Our commissioned survey found that **36% of ASR users** find existing tools **inaccessible for their needs**.

The limitations of existing ASR tools

In exploring the competitive market and working with 4,000+ customers globally, Verbit's team found that until now, many professionals have not been impressed with existing ASR technologies.

Only 53% of professionals surveyed are satisfied with the accuracy of their current tool's output.

They often encounter inaccuracies and need to spend time manually fixing the output to reach the level of professionalism and accuracy needed. For instance, for live events, they don't have the luxury of editing the output later; they take a risk of errors appearing onscreen during broadcasts and live events.

Despite significant technological progress in speech recognition and AI technologies, our experts and customers continue to find that many traditional speech recognition tools, including free built-in ASR tools still fall short of meeting accessibility benchmarks and the diverse end user needs.

“Auto captions, machine captions, YouTube captions - none of that is compliant. It’s really, really important for institutions to keep that in mind,” said Becky Borello, Office of Instructional Technology and Teaching Development, Chatham University.

While generic ASR technology offers an efficient solution for many small-scale and internal speech-to-text tasks, it often fails in complex environments. Margolin said users should consider a conference call that includes multiple speakers with diverse accents and a large amount of overlapping speech. Built-in tools often struggle to navigate such complexities effectively and cannot properly attribute who said what to whom. They also don’t allow users to provide continuous feedback, meaning their accuracy levels are stunted. Their use is commonly associated with a trade-off between speed and accuracy. When professionals optimize for speed, they often compromise accuracy and vice versa, she said.

On the other side of the spectrum, many often turn to human transcribers when they feel they cannot count on ASR tools to produce reliable results. Human transcribers bring a wealth of expertise and an ability to react to diverse scenarios. They can handle unknown content, search online if needed while transcribing and adapt when a speaker changes. They can also receive notes on specific customer requirements in real time. Yet relying on human transcribers comes with a cost and may entail issues such as the need to schedule in advance, burn out, human bias and error (e.g., skipping words when not sure about spelling or in a rush to catch up). These factors can add to the already premium price and may hinder scalability, particularly for certain objectives.

Advancing the ASR playing field

With the limitations of existing ASR in mind, Verbit's R&D team, speech recognition and machine learning experts researched what was needed to bring a stronger, more customizable automatic solution to the market. They wanted to leverage their research team, the latest technological advancements including Generative AI and years of expertise in transcribing millions of hours of audio and video to consider what makes for the best ASR solution.

The ASR technologies and engines which excel on the market today are those which combine the speed and efficiency of ASR with human precision to achieve higher accuracy levels. However, to improve the user experience and enable ASR users to get the most out of their ASR solution and content, there had to be a missing piece.

"Customers understand what matters to them better than anyone else," said Margolin. "We wanted to enable **collaboration between technology and users**. Our new ASR solution, Captivate™, is therefore powered by customer feedback. Our technology teams use this feedback to continuously refine the model so that it accurately captures the nuances of their unique content world."

Our technology teams use this feedback to continuously refine the model so that it accurately captures the nuances of their unique content world.

What Captivate™ does differently from generic tools is it's putting the customer and their unique, dynamic needs at the center of the offering. This tailored approach allows for customization at every stage of the speech recognition process with the help of Verbit's research and development team.

Adi Margolin
Director of Product Management, Verbit

Meeting market expectations for a stronger ASR offering

ASR solutions which stand out from the rest and can handle the variety of scenarios they're being applied to, cannot be generic. They need to be tuned to individual requirements. Yet to be effective, the solution needs to be able to take in this customization while also facilitating the ability to scale, so that users can receive unlimited coverage when many events are happening at once or when many files need to be captioned or transcribed. Finally, to be competitive, today's solutions must also deliver cost-efficiency, while still maintaining high accuracy. In other words, it's no easy feat to tick all these boxes.

Customizable, industry-specific solutions are needed

> Almost half of the survey respondents (47%) highlighted insufficient accuracy when using ASR.

While they noted their ASR of choice performs well on commonly known terms, it falls short when dealing with specifics which are important to them, including names and industry jargon. **44% of survey respondents** consider **custom vocabulary** a critical and/or important feature, underscoring the importance of offering customizable solutions that cater to the unique requirements of end users.

Through Verbit's extensive collaboration with customers over the years, it became clear that the best way to ensure that their critical content and

formatting requirements are fulfilled is to provide them with a high level of control over the ASR process itself and final outputs.

Verbit's new ASR solution, Captivate™, was therefore designed to utilize sophisticated **term boosting** to enhance the accuracy of the words that matter the most to a specific customer. Verbit utilizes term boosting in multiple ways:

- Users can **provide lists of important terms** such as proprietary acronyms, names, places and other specialized terminology enabling the model to detect and spell them correctly when they are mentioned
- Verbit's experts can **proactively research** terms important to specific subject matters and instantly add them to the model (e.g., when assisting a business with a course on effective management, Verbit's team can leverage similar materials to preload into the engine)
- **Integration** into the customer workflow, including website domains and specific directories where the most up-to-date information is stored, enables the engine to continuously scrape these sources, further

"What we have learned from our customers over the years is that there are always specific words or nuances that are especially close to their heart. Whether it's the name of the CEO, or the correct formatting of scores in a sports game, these specifics matter to them. This is where off-the-shelf tools still fall behind, so Verbit focused on what's needed to get the right words right."

David Landsberg
VP of Product, Verbit

Captivate™ is based on a **continuous learning** model that adapts over time to specific domains as more content is captured. The model's training set comprised of tens of thousands of audio hours and was supervised by professional human captioners. The adaptability of Captivate™ extends across all speech-intensive industries. Whether learning domain-specific content for a business training course or staying current on real-time events for a news channel, the solution utilizes **dynamic domain dictionaries**. They are updated continuously and the update process is domain-dependent, offering different, tailored methods for each domain.

Professionals need to be able to scale their ASR usage

33% of survey respondents noted that **less than half of their audio and video content is currently captioned**. With a lot of untapped content in every organization and the growing need for real-time delivery, users need solutions that can be easily applied to additional use cases and scenarios. Whether utilizing ASR to generate captions for an online professional development course or for a live broadcast, the amount of time and effort required to bring these to life needs to be minimal. Unlike traditional human captioning that requires advance scheduling and may involve unexpected delays, Verbit sought to offer a leaner solution for dynamic business needs.

A notable 39% of surveyed professionals cited difficulties with setup and integration as a primary barrier to increased ASR usage.

In developing Captivate™, Verbit's team dedicated a lot of effort to addressing these challenges and simplifying the user experience. Making it easier to get started with the technology, enabling quick ordering, and

providing flexible deployment options – all this is critical to reduce the turnaround time and eventually bring more use cases to production.

But it goes beyond simply adding new use cases, it's about effectively tackling the complexity of different scenarios as well. **20% of survey respondents** selected enhanced performance in **handling complex audio scenarios** - such as background noise, multiple speakers, various languages, and accents—as the main improvement necessary to drive ASR usage. Captivate™ has significantly enhanced ASR's capacity to handle challenging acoustic scenarios and navigate complex real-world situations. For example, imagine the chaotic environment of a basketball game with cheering crowds, shouting players and rapid commentary. Advanced models behind Captivate™ ensure reliable speech recognition, not possible with generic ASR engines, even in demanding situations.

"An interesting scenario we are working on in our research lab is using Captivate™ to caption a live performance of a national anthem, which is usually sung in loud stadiums and venues. This presents numerous acoustic challenges, including background noise, varying durations of the same syllables, and the inherent complexity of singing as opposed to conventional speech."

Dr. Irit Opher
VP of Research, Verbit

Businesses are looking for cost-effective ASR with human-level accuracy

Budget constraints often hinder companies from expanding their use of speech recognition, as echoed by 33% of business leaders.

This is particularly true when businesses rely heavily on human transcription services, which offer high quality but may come at a high cost. While generic ASR tools are more affordable, their accuracy often falls short for critical use cases. Today's professionals need an ASR offering that is both affordable and reliable, enabling them to expand usage without exceeding budget limitations or compromising quality.

One of the key factors driving the superior performance of Captivate™ is **human curation**, where a dedicated team of experts continuously provides feedback on the output. Over time, this process leads to significant quality improvements, ensuring that no critical mistakes occur and no detail is missed.

"Accuracy is not entirely objective and shouldn't be measured only with one specific metric like the number of word errors. Accuracy also has important implications for downstream usage - for example, if an acronym in a corporate call recording isn't captured correctly, it will make the summary and action items created based on it simply irrelevant". - Adi Margolin, Director of Product Management, Verbit

Another significant benefit of human curation is the ability to grasp **context**. This becomes especially important in transcribing conversations where code words are utilized, such as in law enforcement scenarios. For instance, consider a conversation where a code word, unrelated to the

context, replaces the term “package” to maintain discretion. In such cases, a generic ASR engine lacking pre-session preparation may output a word that appears contextually fitting but is incorrect, potentially compromising the transcript’s utility. This illustrates that domain expertise extends beyond word frequency and requires nuanced contextual understanding.

At the end of the day, the accuracy is in the eye of the beholder. As nicely illustrated by Tenysa Santiago, a Deaf and Hard of Hearing Coordinator in San Francisco State University, combining the best of both worlds – technology and human - ensures that those who rely on accessible content

“The combination of artificial intelligence and the human touch just to verify, just to make sure that everything is as accurate as it can be, that’s what allows us to take all of our content to the level of accuracy and comprehensibility that we need to be able to put in front of our students and say, ‘this is fully accessible for you.’”

Tenysa Santiago

Deaf and Hard of Hearing Coordinator, San Francisco State University

Conducting rigorous testing with real customers

In order to test out what they had built, Verbit team recruited beta users for Captivate™. Verbit conducted a pilot project with one of the largest local broadcast television groups in the US. Captivate™ was used to caption **60+ hours** of content for two different TV stations. **Within just one month**, it achieved a **quality difference of only +/- 0.5%** compared to human captioners, occasionally even surpassing them.

“We rigorously tested our new ASR in live broadcasting with global networks, where precision is critical. These industry leaders expect, or even demand, human-level accuracy in captions. Our new solution is able to deliver exactly that.” - David Landsberg, VP of Product, Verbit

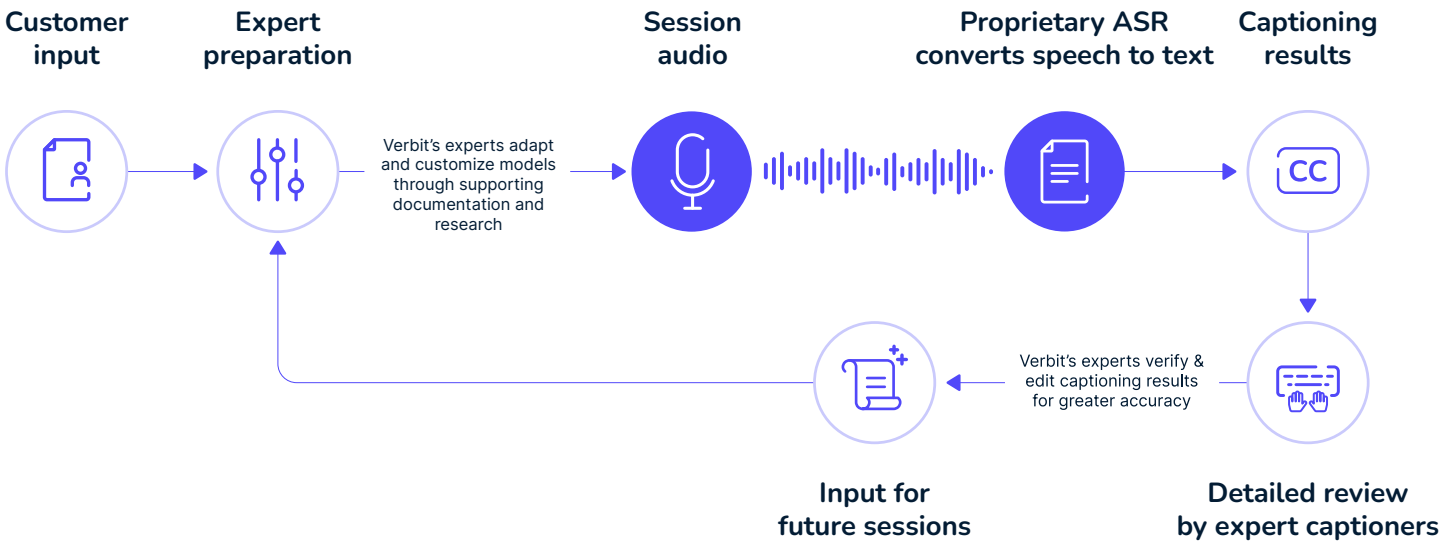
In another significant project, a multinational cable news network tested Captivate™ for its newly launched 24/7 live streaming service and was highly impressed with the quality of the results. Within a few weeks, they made the decision to transition all their live linear content, including the Spanish channel, to Captivate™.

Within the first month in production, over 2,500 hours of captioned content were successfully delivered.

Until now, major media networks and other businesses with 24/7 needs were mostly only ever able to rely on the work of humans over existing ASR solutions. Now, this new ASR alternative offering is viable since it can be effectively trained, specialized and customized to a specific user or network’s needs. Captivate™ has now been deployed to deliver thousands of hours of captioned content.

Breaking down how Captivate™ works

The entire process behind Captivate™ is designed for continuous improvement and maximum accuracy. In the pre-session phase, Verbit's experts utilize customer inputs and thorough research to tailor the model to the specific session. Following the session, they evaluate and adjust the results to refine accuracy for subsequent sessions.



Conclusion

The ongoing exploration and use of ASR technology has shed light on both its vast potential and the pressing need for improvements. As the market progresses towards more user-centric solutions that leverage combined human-machine strength, businesses will be able to apply speech recognition technology for more real-world scenarios.

With enhanced accuracy and the ability to capture nuance, ASR technologies will enable business to leverage their verbal information in previously impossible ways. For example, imagine a corporate training session transcribed in real-time, then converted through generative AI into interactive learning modules for employees on-the-spot. Overall, beyond enhancing accessibility, productivity and efficiency, the transformation of ASR will continue to disrupt how verbal content is utilized in the modern business landscape.