# Voice Revolution

## Win Shih

*Voice allows us to command an army of digital helpers —administrative assistants, . . . advisors, babysitters, librarians, and entertainers.*

*—James Vlahos*[1]

## Voice Assistants

Voice assistants (VAs), such as Amazon's Alexa, Google Assistant, Apple's Siri, and Microsoft's Cortana, are computer programs designed to assist users by answering questions and performing tasks. Sometimes called virtual assistants, digital assistants, or intelligent personal assistants, VAs represent a paradigm shift of human-technology interaction. In the past, we interacted with computers through a keyboard, mouse, monitor, or touch screen. Voice technology now lets us engage the digital world with our speech through conversational user interfaces. The voice is how we communicate with other human beings from the time we first learn how to speak. As a result, interacting with technology through speech comes most naturally and intuitively to us. Further, VAs add a richer dimension to our relationship with technology, providing a different immersive experience and the potential to reduce our efforts in technology use.

For years, we have dreamed about conversing with machines directly. We see intelligent machines and robots in movies, such as the HAL 9000 killer supercomputer in Stanley Kubrick's *2001: A Space Odyssey* and Vox, a holographic sentient librarian, in *The Time Machine*; in written science fiction—Herbie and Robbie in Isaac Asimov's *I, Robot*; and on television—the computer in *Star Trek*. However, it was not until recently that advances in artificial intelligence (AI), natural language processing, machine learning, and computing processing power made voice computing a reality. The voice assistant is the first application that exposes consumers to the power and potential of artificial intelligence, and the spectrum of its capability continues to expand. Now voice assistants can perform multiple tasks in response to one request. Both Alexa and Google Assistant let you define routines or mini-automation to perform a set of tasks with a single voice command. For example, when you say, "Hey Google, good morning," Google Assistant will execute a sequence of predefined tasks, such as adjusting your lights and thermostats; giving you a weather report; estimating your commuting time; looking up your calendar; providing reminders; playing news, radio, or music; or performing whatever tasks you preselected from its action list.[2] Alexa offers a similar feature to let users automate a set of routine tasks. When you say, "Alexa, I am home," Alexa will automatically turn on the lights, set the thermostat, turn on music or a radio station, and announce that you are home across your Echo devices.

Apple's Siri is the first AI-empowered, large-scale virtual assistant program on the consumer market. Siri was originally a stand-alone mobile app for iPhone. Apple acquired Siri and later integrated it into the operating system of the iPhone 4S in October 2011. It was a huge success, with more than four million sales in the first four days of its release.[3] Three years later, Amazon introduced its voice assistant device, Echo, on November 6, 2014. More than one million Echo devices were sold in just two weeks of its introduction.[4] Two years later, in November 2016, Google revealed its version of a virtual assistant device, Google Home.[5] To compete with Echo and Google Home, in 2018 Apple introduced its HomePod smart speaker that runs the Siri voice assistant.[6] Facebook, late in the game, admitted in April 2019 that it was developing its version of a voice assistant that would work on Portal, its video call device.[7]

## Growth of Virtual Assistants and Smart Speakers

More than eight years since the debut of Siri, VA technology has penetrated US households and has become an integral part of many people's daily lives. According to a 2018 survey, 72 percent of 1,000 participants indicated that they have used a virtual assistant. Furthermore, 44 percent of these survey respondents have used VAs to control another smart device at home.[8] A 2019 survey estimated that more than 110 million people in the US use VAs at least monthly, a 9.5 percent jump from 2018.[9]

Smart speakers are now a common household item, with more than 133 million in use in the US and about 26 percent of US consumers owning smart speakers. Amazon is the market leader with 61 percent of market share of smart speakers, while Google takes 24 percent of the market.[10] To expand market share, Amazon and Google also opened up their voice technology to allow third-party manufacturers to include Alexa or Google Assistant on their devices. Smart speakers from premium audio companies, such as Bose, Bang and Olufsen, and Sonos, now have either Alexa, Google Assistant, or both baked in. Alexa is not available only on 150+ Amazon products, such as Echo, Echo Dot, and Echo Show, but is also included in more than 100,000 third-party smart devices, ranging from TVs to microwaves to washing machines, from 9,500 brands.[11] Although Google Assistant was almost two years behind Alexa's release, it is catching up and is now available on more than 10,000 devices from 1,600 brands.[12]

Although voice assistants are usually associated with smart speakers, VAs are also available as mobile apps and even come as default programs of mobile devices. Siri and Google Assistant are integral parts of their corresponding mobile platforms (i.e., Siri for iPhones and Google Assistant for Android phones). Alternatively, Siri, Google Assistant, and Alexa can also be downloaded free from app stores if they do not come with your smartphone. Although Amazon has dominated the smart speaker market with over 100 million Alexa-powered devices, there are more mobile devices with either Google Assistant or Siri installed.[13] The popularity of smart speakers further drives the use of voice assistant apps on smartphones. The mobile app complements the smart devices because it can be used to perform the same tasks as the smart speaker and is also able to customize and manage the speaker.

## Alexa Skills and Google Actions

Both Amazon and Google make their proprietary technology available to third-party developers to learn and build new applications on their virtual assistant platforms. These third-party-developed applications, called *skills* on Alexa's platform and *actions* on Google's platform, let companies build their own voice applications tailored to their products and services, as well as reaching their customers beyond their websites and mobile apps. Need a ride? You can request either Uber or Lyft from Alexa. Hungry for pizza? You can ask Alexa to have your favorite pizza delivered from Domino's or Pizza Hut.

As of September 2019, the number of Amazon Alexa skills has quadrupled since 2017, from 25,000 to more than 100,000 worldwide. Among them, 65,901 skills are available in the US.[14] Games and trivia (21 percent) is the largest category of Alexa skills, followed by education and reference (14 percent). Meanwhile, Google boasts of its 2,253 actions available at the beginning of 2019, a 250 percent jump from the same period in 2018. Education and reference represents the largest category (15 percent) of the Google Assistant programs.[15] Such exponential growth of VA applications confirms a continued interest in and commitment to both platforms from developers.

To encourage third-party developers, Amazon holds a variety of training and promotional events, including Alexa Dev Days, conferences, presentations, workshops, and hackathons at various locations as well as virtually. Amazon also sponsors community meetups for local developers and enthusiasts to meet and share best practices. The Amazon Alexa Fellowship program is another effort designed to inspire interest in conversational AI and Alexa technology among academic institutions. This competitive program is comprised of two components: the Alexa Graduate Fellowship and the Alexa Innovation Fellowship. The former supports doctoral or postdoctoral students in pursuing education or research in conversational AI. The Innovation Fellowship program partners with academic institutions' existing innovation or entrepreneurship centers in helping students develop voice-based start-ups. For example, the Alexa Innovation Fellowship at the University of Southern California is a twelve-month program that supports student teams at all stages of developing their voice technology start-ups.[16] Finally, the Amazon Alexa Prize Socialbot Grand Challenge, an annual competition now in its third year, invites university students to create "socialbots that can converse coherently and engagingly with humans on a range of current events and popular topics such as entertainment, sports, politics, technology, and fashion" for twenty minutes at a time.[17] The winning team will be awarded a $500,000 prize, and its university will receive a $1 million research grant.[18] In 2017, Team Sounding Board from the University of Washington took the top prize, with an average conversational duration of ten minutes and twenty-two seconds.[19] The next year, Team Gunrock from the University of California, Davis, won the prize, with an average duration of nine minutes and fifty-nine seconds.[20]

## Language Support

The market of virtual assistants is language-dependent. The applications developed are closely associated with the language they are intended to support. In the case of Alexa, if a skill is developed in English, it can be made available in other English-speaking countries; Amazon will migrate the skill to its English variants for other English-speaking countries, such as Australia, Canada, India, the UK, and the US. Table 1.1 shows the number of languages each virtual assistant supports. By far, Siri is the most polyglot virtual assistant with twenty-one languages, followed by Google Assistant.

Bilingual support is another feature that allows companies to expand their markets. Both Alexa and Google Assistant support a bilingual mode, which can distinguish between different languages and respond to the question in the language in which it is asked. For example, in the US, Alexa can switch between English and Spanish; in Canada, English and French; in India, English and Hindi. Google Assistant's bilingual mode supports a combination of twelve languages.

## How Smart Are Virtual Assistants?

Surveys have consistently found that asking general questions or searching for a quick answer is the most commonly used feature by consumers of virtual assistants, followed by playing music.[21] How do virtual assistants stack up in understanding questions and providing accurate answers? Several studies have shown that voice assistants have been steadily making progress in comprehending human questions and responding with accurate answers over the last few years.

Loup Ventures, a research-driven venture capital firm, has systematically gauged the performance of leading smart speakers and digital assistants since 2017.[22] In what is called the annual smart speaker and digital assistant IQ Test, Loup's researchers ask digital assistants to answer 800 real-world questions in five categories: local information, commerce, navigation, information, and commands (calling, texting, e-mailing, scheduling, and reminders). Between February 2017 and August 2019, Loup conducted eight tests using either smart speakers or virtual assistant apps (see table 1.2). Among the eight tests conducted so far, Google Assistant consistently outperformed three other digital assistants in understanding the questions and answering them correctly. Siri is the second-smartest VA, followed by Alexa and Cortana. Over a two-and-a-half-year period, both Alexa and Google Assistant have made tremendous improvement in both understanding the questions and providing satisfactory answers. As shown in table 1.2, in correctly understanding voice queries, Google Assistant marked an impressive 23 percent improvement during this short period, compared with Alexa's 5.5 percent. As for providing correct answers, Google Assistant again outperforms Alexa, with 54 percent improvement, versus Alexa's 45 percent.

In another large-scale, structured study, Enge asked four voice assistants (Alexa, Cortana, Google Assistant, and Siri) through their associated devices 5,000 questions annually since 2017. Google Assistant consistently outperformed other competitors in two categories: the number of questions answered and the number of questions answered correctly and fully. Alexa, although showing continual improvement, still fell behind Google Assistant in both categories.[23] Other similar studies have had similar results as well.[24]

## Where Does a Virtual Assistant Get Answers?

Virtual assistants use a variety of sources to get answers. However, it is not always clear what their

**Table 1.1**
Language support (Source: "Language Support in Voice Assistants Compared (2019 Update)," Globalme language & technology, January 27, 2020, https://www.globalme.net/blog/language-support-voice-assistants-compared.)

| Virtual Assistant | No. of Languages | Notes |
|---|---|---|
| Amazon Alexa | 7 | English (5 dialects), French (2 dialects), German, Hindi, Italian, Japanese, Spanish (3 dialects) |
| Apple's Siri | 21 | Arabic, Chinese (2 dialects), Danish, Dutch (2 dialects), English (9 dialects), Finnish, French (4 dialects), German (3 dialects), Hebrew, Italian (2 dialects), Japanese, Korean, Malay, Norwegian, Portuguese, Russian, Spanish (4 dialects), Swedish, Thai, Turkish |
| Google Assistant | 13 | Danish, Dutch, English (6 dialects), French (2 dialects), German (2 dialects), Hindi, Italian, Japanese, Korean, Norwegian, Portuguese, Spanish (3 dialects), Swedish |
| Microsoft Cortana | 6 | Chinese, English (2 dialects), French, German, Italian, Spanish |

**Table 1.2**
Summary of Loup Venture research on virtual assistant IQ Test (This table is a compilation of research conducted by Gene Munster and Will Thompson of Loup Venture between February 2017 and August 2019.)

| **Answered Correctly** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Testing Date** | **August 2019** | **December 2018** | **July 2018** | **December 2017** | **August 2017** | **April 2017** | **February 2017** | **Overall Improvement** |
| **Test via** | **Digital Assistant** | **Smart Speaker** | **Digital Assistant** | **Smart Speaker** | **Smart Speaker** | **Digital Assistant** | **Smart Speaker** | |
| Amazon Alexa | 79.80% | 72.50% | 61.40% | 63.81% | 53.57% | — | 34.40% | 45.40% |
| Google Assistant | **92.90%** | **87.90%** | **85.50%** | **81.10%** | **65.25%** | **74.80%** | **39.10%** | **53.80%** |
| Apple Siri | 83.10% | 74.60% | 78.50% | — | — | 66.10% | — | — |
| Microsoft Cortana | — | 63.40% | 52.54% | 56.38% | — | 48.80% | — | — |
| **Understood Query** | | | | | | | | |
| **Test via** | **Digital Assistant** | **Smart Speaker** | **Digital Assistant** | **Smart Speaker** | **Smart Speaker** | **Digital Assistant** | **Smart Speaker** | **Overall Improvement** |
| Amazon Alexa | 99.90% | 99.00% | 98.00% | 97.87% | **95.88%** | — | **94.40%** | 5.50% |
| Google Assistant | **100%** | **100%** | **100%** | **99.88%** | 94.63% | **99.90%** | 77% | 23.00% |
| Apple Siri | 99.80% | 99.60% | 99.00% | — | — | 94.40% | — | — |
| Microsoft Cortana | — | 99.40% | 98.00% | 98.87% | — | 97.30% | — | — |

**Table 1.3**
Alexa built-in Intents (Source: "Built-in Intent Categories," in "Built-in Intent Library," Amazon Alexa, https://developer.amazon.com/en-US/docs/alexa/custom-skills/built-in-intent-library.html.)

| Category | Description |
|---|---|
| Books | Intents for asking about books and other written works, such as rating books, adding books to reading lists, or navigating through audiobooks. |
| Calendar | Intents for asking about calendars and schedules, such as asking about upcoming events, adding events to a calendar, and looking up events such as birthdays. |
| Cinema Showtimes | Intents for asking about show times and events at movie theaters. |
| General | Intents for requests that don't fall into any of the more specific categories. |
| Local Search | Intents for asking about local businesses and locations, such as operating hours, phone numbers, and travel times to locations and businesses. For example, users can ask when a particular business is open or ask for the phone number of a business. |
| Music | Intents for asking about music, such as asking about songs, playlists, and music libraries. |
| Video | Intents for asking about television and other types of video media. For example, users can ask for information about episodes of a TV show. |
| Weather | Intents for requesting weather reports and forecasts. |
| Standard Built-in Intents | Intents for general actions such as stopping, canceling, and asking for help. |

sources are and how the sources and answers are selected. The Alexa Skills Kit, used by developers to create Alexa skills, offers a library of predefined functionalities, called *built-in intents*. Developers can incorporate these built-in intents in their skills to answer questions. For example, the Weather built-in intent allows users to ask for weather information in a specific location. Alexa organizes its built-in intents into nine categories, each representing a different type of inquiry. Table 1.3 lists these built-in intents with a brief description of each. Each built-in intent, except General, handles a specific type of inquiry.

Amazon provides no details about the information sources used by each Alexa built-in intent. It is also unclear how Amazon prioritizes the use of different sources or what is the process, if any, to ensure the quality of answers. According to the postings on the Alexa Forum by Amazon employees, Alexa "gets her information from a variety of trusted sources such as IMDb, Accuweather, Yelp, Answers.com, Wikipedia, and many others," and the source used to provide answers depends on the type of question asked.[25] Fast Company reported that Amazon has licensing agreements with hundreds of sources that "it deems high-quality" for responding to Alexa users' inquiries.[26] One example of such a business deal is Wikipedia, whose free content is commonly used by Alexa, Google Assistant, and Siri to answer questions. Although Wikipedia makes its content free for anyone to use, Google is by far the largest donor to the Wikipedia Foundation, donating more than $1 million in 2018. Amazon also donated $1 million in 2018 to support Wikipedia's mission of sharing free knowledge worldwide and to ensure its long-term sustainability.[27] As for prioritization, Amazon is said to employ machine learning technology and algorithms to rank answers and determine the best one to use for any given query.[28]

On the other hand, Google Assistant benefits from the company's long-held dominance in the search business and its gigantic corpus of indexed web pages. Google Assistant gets its information from Google's own products—including Google Maps, Search, and Google Photos—as well as third-party services.[29] Google employs "a combination of explicit linguistic knowledge and deep learning solutions" to ensure Google Assistant's audio responses are "grammatical, fluent and concise."[30] Google has released a document used to guide its human evaluators on ranking voice search results into five grades (Fully Meets, Highly Meets, Moderately Meets to Slightly Meets, Slightly Meets, and Fails to Meet).[31] The rubric for rating response quality contains the following key parameters and provides insight on how Google assesses and formulates its answers:

- **Information Satisfaction:** the content of the answer should meet the information needs of the user.
- **Length:** when a displayed answer is too long, users can quickly scan it visually and locate the relevant information. For voice answers, that is not possible. It is much more important to ensure that we provide a helpful amount of information, hopefully not too much or too little. . . .
- **Formulation:** it is much easier to understand a badly formulated written answer than an ungrammatical spoken answer, so more care has to be placed in ensuring grammatical correctness.
- **Elocution:** spoken answers must have proper pronunciation and prosody. Improvements in text-to-speech generation, such as WaveNet and Tacotron 2, are quickly reducing the gap with human performance.[32]

To catch up to Google and close the intellectual gap created by Google Assistant, Amazon has developed a crowdsourcing program called Alexa Answers. Launched in September 2019, Alexa Answers asks the Alexa user community to provide answers not found in Alexa's knowledge repertoire. With the goal to "help make Alexa smarter," touted on the program's website, the program encourages Alexa users to "join the experts and enthusiasts sharing their knowledge with Alexa and the world."[33] Through the Alexa Answers portal, participants can browse through and respond to unresolved questions of their choice. Questions are grouped into twelve categories (Animals, Climate, Film and TV, Food, Geography, History, Literature, Miscellaneous, Music, Science, Sports, and Video Games). Answers are limited to 300 characters. As of January 2020, there are 43,000 unanswered questions listed in the Alexa Answers question bank.[34]

*Alexa Answers portal*
https://alexaanswers.amazon.com/

To encourage participation, Amazon devised a gamification system. Instead of receiving monetary rewards, contributors earn points and badges based on the number of times their answer is used and the quality of their answer. Amazon uses a combination of automated and manual processes to rate contributors' work. Alexa customers can also rate the answer and report incorrect or inappropriate answers. Alexa Answers participants can upvote or downvote contributed answers.[35]

Crowdsourcing for answers is not new. Other community-based question-and-answer services, such as Yahoo Answers, Answers.com, and Quora, have been around for years. However, these question-and-answering services have not always been viewed as a reputable information source. Several evaluators have reported questionable answers from Alexa Answers; the problematic answers identified range from inaccurate to asinine. Some testers also found answers containing advertising, sponsorship, or spam.[36] The Alexa Answers program faces several closely related challenges of its business model, including the following:

- **Quality control:** How reliable is its semiautomated filtering mechanism to ensure the authenticity, quality, and reliability of the contributed facts?
- **Abuse:** How effectively can Amazon detect fake information and prevent malicious attempts to game this system?
- **Data voids:** Not everything has an answer. There are many questions with limited available or nonexistent data.[37]
- **Original sources:** When an answer is from Alexa Answers, it is identified as "according to an Amazon customer" in Alexa's response. However, it is unclear if the contributed answer is copied from another source.
- **One-shot answers:** In web-based question-and-answering services, users can see a list of contributed answers and their ratings by the user community. They can judge which one is more trustworthy and have the option to choose the one they want to use. With virtual assistant, users will get one and only one answer. They do not have the option to scan and get a broad perspective of all possible answers. In this sense, virtual assistants assume a much more powerful and intermediary role than web-based services in deciding what answer customers receive in the search process.

Alexa Answers operates on the assumption of faith in its user community. There are risks as well as benefits to leveraging human intelligence and people's willingness to help. Amazon needs to make the vetting process transparent and the screening mechanism effective to ensure trustworthiness of the information offered. This is especially critical in the evolving voice assistant ecosystem as more and more users, especially children, now view voice assistants as a reliable source of knowledge and information.

## How Voice Assistant Technology Works

Being able to converse and interact with machines has been a long-standing goal of scientists. For years, researchers from computer science, linguistics, cognitive science, and information engineering have worked tirelessly to program computers to process and analyze natural speech data. However, it was not until the advent of artificial intelligence and machine learning that voice technology began to progress by leaps and bounds in the last few years. Now machines can not only transcribe human speech into text but also understand a request and respond with accurate actions.

For each voice inquiry-answer interaction, the process, called natural language processing, mimics human reasoning and communication and typically involves several specific technologies. First, automatic speech recognition converts the sound of a human request into text. Then, natural language understanding technology analyzes the text to make sense of the request. The appropriate response or semantic intents are then converted into readable text through natural language generation. Finally, speech synthesis transforms text responses into audio signals that we can understand.

### Automatic Speech Recognition

Automatic speech recognition (ASR) is the technology that converts spoken sound waves into a corresponding sequence of words.[38] ASR has been a field of research for more than sixty years. However, scientists have made more progress in speech recognition technology in the last thirty months than in the first thirty years of its existence due to the advance of AI and other related technology.[39] In 1952, engineers at Bell Laboratories introduced Audrey, capable of recognizing a voice speaking single digits of zero to nine aloud. In 1962, IBM researchers presented Shoebox, which was capable of understanding sixteen English words. In a video, we can watch a scientist instructing Shoebox to perform simple calculations, including addition, subtraction, total, and subtotal, all with voice commands.[40] In the mid-1960s, an MIT computer scientist developed the first text-based chatbot, Eliza, which could respond to predefined human questions. By the 1970s, with funding from the Department of Defense, Carnegie Mellon University developed Happy, which could recognize more than 1,000 words.[41] In 1997, Dragon's NaturallySpeaking software had the capacity to transcribe human speech at a rate of 100 words per minute.[42] However, it was not until the 2010s that ASR reached a suitably mature and reliable level for practical use. With ASR technology, computers can now "detect patterns in audio waveforms, match them with the sounds in a given language, and ultimately identify which words we spoke."[43]

With the computational power to manage large data sets, ASR technology can now convert speech signals into text within milliseconds. Scientists also are able to reduce the word error rate to a reasonable level. In 2017, Google reported that its ASR technology had achieved an error rate of less than 5 percent, which is close to the average 4 percent error rate of human transcription services.[44] ASR also has gained progress in neural networks—layered mathematical functions modeled after biological neurons and statistical models—to make educated decisions and determine the right word in a situation of homonyms—words having the same pronunciation or spelling but different meanings; for example, "These *two* pastries are *too* delicious *to* resist. While buying some *deer* *meat,* I happened to *meet* a *dear* friend. I ran *four miles for Miles*." Today, ASR has a wide range of applications. The technology is used in learning foreign languages and in helping people with visual impairment or a physical disability. It can even generate closed-captioning for people experiencing hearing issues.

## Natural Language Understanding

Natural language understanding (NLU) is a process of teaching computers to understand and interpret human speech based on grammar and the speech's context. The task involves digesting a text, translating it into computer language, and generating an output that humans can understand. Employing machine learning through past examples, NLU can deduce and disambiguate what people actually mean, not just the words they say. So when you ask, "What is it outside?" NLU will infer that you are actually asking for a weather forecast at your current location.[45] NLU can also learn from historical interactions to tell that inquires such as "Do you have *Wall Street Journal* in the library?" "Where can I find *Wall Street Journal*?" and "How can I access *Wall Street Journal*?" are essentially different versions of the same question.

The performance of NLU will continue to improve with the advancement of machine learning, especially deep learning or learning from examples, combined with the large corpus of historical transactions and cloud computing power. According to Amazon, "the error reduction has been threefold" since Alexa was introduced in November 2014 even though the scope and complexity of user inquiries and the range of responses Alexa can handle has increased tremendously.[46] In addition, NLU will further expand its learning models and strategies to include semi-supervised learning, active learning, and context-aware models.

## Natural Language Generation

Natural language generation (NLG) is the process of software systems transforming structured data into meaningful phrases and sentences that humans can understand.[47] Using techniques from computational linguistics and AI, NLG can process a large amount of text, identify a data set that meets predefined criteria, and automatically generate narratives. Some practical applications of NLG include automatically creating machine-written corporate earnings reports based on a company's earning data, generating weather forecasts based on temperature prediction data, and producing financial portfolio summaries and updates for individual customers based on the performance of their investments.

## Speech Synthesis

The last step in the process of responding to a user's question is to turn the responding text to sound waves. This process is commonly called speech synthesis or text-to-speech. For example, when you ask Alexa to search a library's catalog by keywords, Alexa will read the brief bibliographic information of top results back to you. Converting text into human speech has its own challenges. Written text can be ambiguous. Words or phrases can have different meanings based on the context and thus be pronounced differently. Below are some examples:

- Numbers might have different meanings and thus be pronounced differently. For example, the number 1984 can represent the year in history, a quantity of items, or a code. The voice assistant needs to be smart enough to pronounce it based on the context. When 1984 refers to a year or book title, a VA will pronounce the number "nineteen eighty-four." When the number indicates a price or a quantity, a VA will say "one thousand nine hundred eighty-four." When the numbers is part of a street address, it will say "one nine eight four."
- Homographs are words spelled the same but pronounced differently. They might also have different meanings depending on the context in the sentence. For example, the word *perfect* will sound different in the following two sentences: Your French is perfect; Practice will perfect your French. VAs need to be able to tell the difference and pronounce each according to the appropriate context.
- Proper names, acronyms, special characters, and abbreviations can be difficult to pronounce correctly. For example, a voice assistant might not pronounce an e-mail address such as dml@usc.edu or an acronym such as AACR2 the way a human would say it aloud.
- *Prosody* refers to the patterns of stress and intonation in speech. Voice assistants might sound monotone with limited capability for varying pitch and intonation. To make the artificial utterance sound natural and expressive, speech synthesis needs to be able to perfect the proper use of these linguistic functions, including tone, intonation, stress, pauses, and rhythm.

Voice synthesizing technology employs deep neural network and statistical probability techniques to overcome these ambiguities, as well as to improve the quality of speech. Google's Cloud Text-to-Speech, based on DeepMind's WaveNet technology, can generate high-fidelity speech in more than 180 voices across thirty languages and variants. Users can also specify the pitch of the voice, the speaking rate, and the volume of the speech.[48] Amazon Polly, Amazon's speech synthesis service, now provides twenty-seven synthesized voices across twenty-nine languages and variants in two speaking styles: newscaster and conversational.[49] Using Speech Synthesis Markup Language, Alexa developers can further refine speech by defining the speaking style, emotional expression, length of pause, phonemic or phonetic pronunciation, volume, pitch, and rate of speech. Apple's Siri used to be based on a hybrid neural network system that includes both synthesized audio and human-generated voice clips. With the new iOS 13 released in September 2019, Siri's voice is entirely generated by software. According to Apple, this new "neural text to speech" technology allows Siri to sound much more natural, lifelike, and smoother, especially for longer sentences, and stress syllables more accurately than the older version.[50]

## Notes

1. James Vlahos, *Talk to Me: How Voice Computing Will Transform the Way We Live, Work, and Think* (Boston: Houghton Mifflin Harcourt, 2019), 3–4.
2. "Set Up and Manage Routines," Google Nest Help, accessed January 21, 2020, https://support.google.com/googlenest/answer/7029585?hl=en.
3. "iPhone 4S *First* Weekend Sales Top Four Million," *Apple*, October 7, 2011, https://www.apple.com/newsroom/2011/10/17iPhone-4S-First-Weekend-Sales-Top-Four-Million/.
4. Vlahos, *Talk to Me*, 44–45.
5. Ava Mutchler, "A Timeline of Voice Assistant and Smart Speaker Technology From 1961 to Today," Voicebot.ai, March 28, 2018, https://voicebot.ai/2018/03/28/timeline-voice-assistant-smart-speaker-technology-1961-today/.
6. Vlahos, *Talk to Me*, 280.
7. Nick Statt, "Facebook confirms it's working on an AI voice assistant for Portal and Oculus products," Verge, April 17, 2019, https://www.theverge.com/2019/4/17/18412757/facebook-ai-voice-assistant-portal-oculus-vr-ar-products.
8. PwC, *Consumer Intelligence Series: Prepare for the Voice Revolution* (PwC, 2018), 3, https://www.pwc.com/us/en/advisory-services/publications/consumer-intelligence-series/voice-assistants.pdf.
9. Victoria Petrock, "US Voice Assistant Users 2019: Who, What, When, Where and Why: Executive Summary," eMarketer, July 15, 2019, https://www.emarketer.com/content/us-voice-assistant-users-2019.
10. Bret Kinsella and Ava Mutchler, *Smart Speaker Consumer Adoption Report* (Voicebot.ai, March 2019), https://voicebot.ai/wp-content/uploads/2019/03/smart_speaker_consumer_adoption_report_2019.pdf.
11. "Brands with Works with Alexa Certified Products," Amazon Alexa, accessed January 21, 2020, https://developer.amazon.com/en-US/alexa/connected-devices/compatible.
12. Ron Amadeo, "Google Boasts 1 Billion Assistant Devices—10x Amazon Alexa's Install Base," Ars Technica, January 7, 2019, https://arstechnica.com/gadgets/2019/01/google-assistant-flexes-on-alexa-announces-1-billion-strong-install-base.
13. Christi Olson and Kelli Kemery, *Voice Report: From Answers to Action: Customer Adoption of Voice Technology and Digital Assistants* (Redmond, WA: Microsoft, 2019), 9, https://about.ads.microsoft.com/en-us/insights/2019-voice-report.
14. Bret Kinsella, "Amazon Alexa Has 100k Skills but Momentum Slows Globally," Voicebot.ai, October 1, 2019, https://voicebot.ai/2019/10/01/amazon-alexa-has-100k-skills-but-momentum-slows-globally-here-is-the-breakdown-by-country/.
15. Bret Kinsella, "Google Assistant Actions Total 4,253 in January 2019, Up 2.5x in Past Year but 7.5% the Total Number Alexa Skills in U.S.," Voicebot.ai, February 15, 2019, https://voicebot.ai/2019/02/15/google-assistant-actions-total-4253-in-january-2019-up-2-5x-in-past-year-but-7-5-the-total-number-alexa-skills-in-u-s.
16. Tristan Taush, "USC Viterbi Startup Garage and Amazon Alexa Fund Announce Support for Voice Tech Startups," IncubateUSC, August 30, 2019, https://incubate.usc.edu/viterbi-startup-garage-amazon-alexa/.
17. "Propel AI Forward. Push Yourself Further," Amazon, Alexa Prize, last modified June 5, 2019, https://developer.amazon.com/alexaprize.
18. "Propel AI Forward."
19. "Propel AI Forward."
20. "Propel AI Forward."
21. Kinsella and Mutchler, *Smart Speaker Consumer Adoption Report*; Christi Olson and Kelli Kemery, *Voice Report* (Redmond, WA: Microsoft, 2019), https://advertiseonbing-blob.azureedge.net/blob/bingads/media/insight/whitepapers/2019/04%20apr/voice-report/bingads_2019_voicereport.pdf; PwC, *Consumer Intelligence Series*.
22. Loup Ventures, https://loupventures.com/.
23. Eric Enge, "Rating the Smarts of the Digital Personal Assistants in 2019," Perficient, October 24, 2019, https://www.perficientdigital.com/insights/our-research/digital-personal-assistants-study; see also Eric Enge, "Rating the Smarts of the Digital Personal Assistants in 2018," *Perficient Digital* (blog), May 1, 2018, https://blogs.perficientdigital.com/2018/05/01/2018-digital-personal-assistants-study and Eric Enge, "Rating the Smarts of the Digital Personal Assistants in 2017," *Perficient Digital* (blog), April 27, 2017, https://blogs.perficientdigital.com/2017/04/27/1-rating-the-smarts-of-the-digital-personal-assistants.
24. Dan Moren, "Alexa vs. Google Assistant vs. Siri: Why Google Wins," Tom's Guide, May 23, 2019, https://www.tomsguide.com/us/alexa-vs-siri-vs-google,review-4772.html.
25. Jenn@amazon, reply to "What sources does Alexa use for information?" Amazon Developer, January 19, 2018, https://forums.developer.amazon.com/questions/105495/what-sources-does-alexa-use-for-information.html.
26. Jared Newman, "Exclusive: Amazon Will Let Anyone Answer Your Alexa Questions Now," Fast

Company, September 12, 2019, https://www.fastcompany.com/90402924/exclusive-amazon-will-let-anyone-answer-your-alexa-questions-now.

27. Justin Bariso, “Amazon Just Donated $1 Million to Wikipedia. Here’s Why It Matters,” Inc., September 27, 2018, https://www.inc.com/justin-bariso/amazon-wikimedia-wikipedia-donation-1-million-emotional-intelligence.html.
28. Newman, “Exclusive.”
29. Enrique Alfonseca, “Evaluation of Speech for the Google Assistant,” *Google AI Blog*, December 21, 2017, https://ai.googleblog.com/2017/12/evaluation-of-speech-for-google.html.
30. Alfonseca, “Evaluation of Speech.”
31. Google, *Evaluation of Search Speech—Guidelines*, version 1.0 (Google, December 13, 2017), http://storage.googleapis.com/guidelines-eyesfree/evaluation_of_search_speech_guidelines_v1.0.pdf.
32. Alfonseca, “Evaluation of Speech.”
33. Alexa Answers about page, https://alexaanswers.amazon.com/about.
34. Alexa Answers, https://alexaanswers.amazon.com/questions (requires login).
35. Courtney Linder, “Amazon Is Crowdsourcing Alexa’s Answers, So This Should Be Fun,” *Popular Mechanics*, September 26, 2019, https://www.popularmechanics.com/technology/a29086631/alexa-answers-crowdsourcing.
36. Kyle Wiggers, “Amazon Is Poorly Vetting Alexa’s User-Submitted Answers,” VentureBeat, November 1, 2019, https://venturebeat.com/2019/11/01/amazon-alexa-answers-vetting-user-questions; Jack Morse, “Now Any Idiot off the Street Can Answer Your Dumb Alexa Questions,” Mashable, September 12, 2019, https://mashable.com/article/amazon-alexa-answers.
37. Michael Golebiewski and danah boyd, *Data Voids: Where Missing Data Can Easily Be Exploited* (Data and Society, October 29, 2019), https://datasociety.net/output/data-voids.
38. Jinyu Li, Li Deng, Reinhold Haeb-Umbach, and Yifan Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications* (Waltham, MA: Elsevier, 2015), 1.
39. “A Short History of Speech Recognition,” Sonix, accessed January 21, 2020, https://sonix.ai/history-of-speech-recognition.
40. Hursley Museum, “1961 Shoebox IBM Archives 78 013,” posted February 12, 2018, YouTube video, 2:06, https://youtu.be/rQco1sa9AwU.
41. Ava Mutchler, “A Short History of the Voice Revolution,” Voicebot.ai, July 14, 2017, https://voicebot.ai/2017/07/14/timeline-voice-assistants-short-history-voice-revolution.
42. “A Brief History of Voice Assistants,” Vox Creative, September 13, 2018, https://next.voxcreative.com/ad/17855294/a-brief-history-of-voice-assistants.
43. “What Is Automatic Speech Recognition (ASR)?” Amazon Alexa, accessed January 21, 2020, https://developer.amazon.com/en-US/alexa/alexa-skills-kit/asr.
44. Beth Worthy, “Word Error Rate Mechanism, ASR Transcription and Challenges in Accuracy Measurement,” *GMR Transcription* (blog), November 26, 2019, https://www.gmrtranscription.com/blog/word-error-rate-mechanism-asr-transcription-and-challenges-in-accuracy-measurement.
45. “What Is Natural Language Understanding (NLU)?” Amazon Alexa, accessed January 21, 2020, https://developer.amazon.com/en-US/alexa/alexa-skills-kit/nlu.
46. Rohit Prasad, “Alexa at Five: Looking Back, Looking Forward,” *Amazon Science Blog*, November 6, 2019, https://www.amazon.science/blog/alexa-at-five-looking-back-looking-forward.
47. Ehud Reiter, “Natural Language Generation,” in *Encyclopedia of the Mind* (Sage, 2013), SAGE Knowledge online database.
48. “Cloud Text-to-Speech: Text-to-Speech Conversion Powered by Machine Learning,” Google Cloud, accessed January 21, 2020, https://cloud.google.com/text-to-speech.
49. “Amazon Polly: Turn Text into Lifelike Speech Using Deep Learning,” AWS, Amazon, accessed January 21, 2020, https://aws.amazon.com/polly.
50. Chaim Gartenberg, “Siri Is Getting a New Voice in iOS 13,” Verge, June 3, 2019, https://www.theverge.com/2019/6/3/18650906/siri-new-voice-ios-13-iphone-homepod-neutral-text-to-speech-technology-natural-wwdc-2019.