

How Speech Recognition Works

By: [Ed Grabianowski](#)



There are several software programs that you can purchase for home speech recognition.

IMAGE COURTESY [AMAZON](#)

Today, when we call most large companies, a person doesn't usually answer the phone. Instead, an automated voice recording answers and instructs you to press buttons to move through option menus. Many companies have moved beyond requiring you to press buttons, though. Often you can just speak certain words (again, as instructed by a recording) to get what you need. The system that makes this possible is a type of **speech recognition program** -- an automated phone system.

You can also use speech recognition software in homes and businesses. A range of software products allows users to dictate to their computer and have their words converted to text in a word processing or [e-mail](#) document. You can access function commands, such as opening files and accessing menus, with voice instructions. Some programs are for specific business settings, such as medical or legal transcription.

People with disabilities that prevent them from typing have also adopted speech-recognition systems. If a user has lost the use of his hands, or for visually impaired users when it is not possible or convenient to use a Braille keyboard, the systems allow personal expression through dictation as well as control of many computer tasks. Some programs save users' speech data after every session, allowing people with progressive speech deterioration to continue to dictate to their computers.

Current programs fall into two categories:

Small-vocabulary/many-users

These systems are ideal for automated telephone answering. The users can speak with a great deal of variation in accent and speech patterns, and the system will still understand them most of the time. However, usage is limited to a small number of predetermined commands and inputs, such as basic menu options or numbers.

Large-vocabulary/limited-users

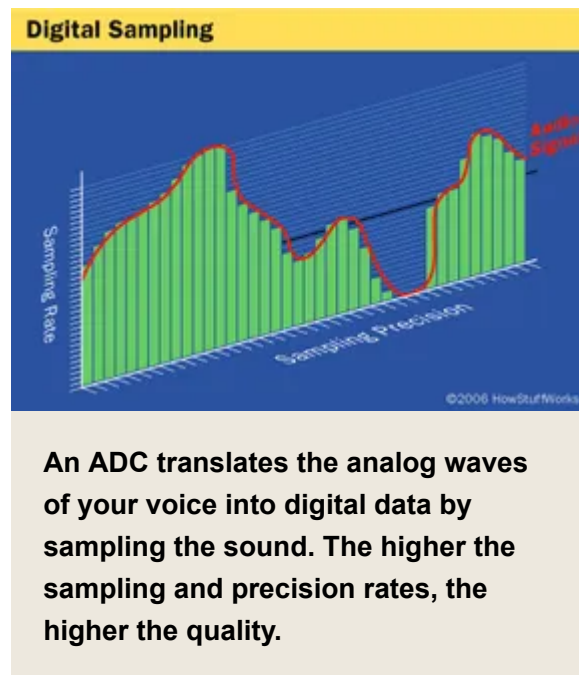
These systems work best in a business environment where a small number of users will work with the program. While these systems work with a good degree of accuracy (85 percent or higher with an expert user) and have vocabularies in the tens of thousands, you must train them to work best with a small number of primary users. The accuracy rate will fall drastically with any other user.

Speech recognition systems made more than 10 years ago also faced a choice between **discrete** and **continuous** speech. It is much easier for the program to understand words when we speak them separately, with a distinct pause between each one. However, most users prefer to speak in a normal, conversational speed. Almost all modern systems are capable of understanding continuous speech.

Thank You

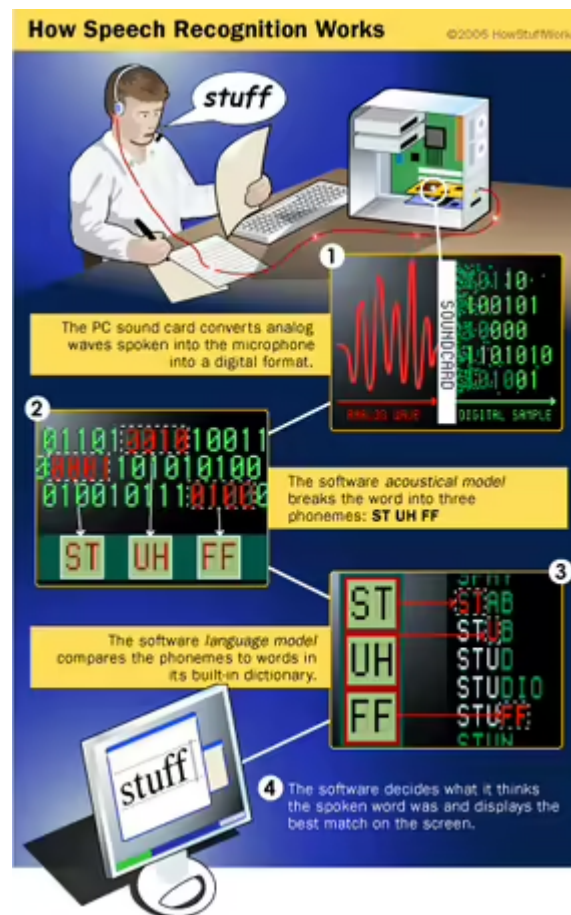
For this article, we spoke with **John Garofolo**, Speech Group Manager at the Information Technology Laboratory of the [National Institute of Standards and Technology](#). We'd also like to thank **Joshua Senecal** for his assistance with this article .

Speech to Data



To convert speech to on-screen text or a computer command, a computer has to go through several complex steps. When you speak, you create vibrations in the air. The **analog-to-digital converter (ADC)** translates this analog wave into digital data that the computer can understand. To do this, it **samples**, or digitizes, the sound by taking precise measurements of the wave at frequent intervals. The system filters the digitized sound to remove unwanted noise, and sometimes to separate it into different bands of **frequency** (frequency is the wavelength of the sound waves, heard by humans as differences in pitch). It also normalizes the sound, or adjusts it to a constant volume level. It may also have to be temporally aligned. People don't always speak at the same speed, so the sound must be adjusted to match the speed of the template sound samples already stored in the system's memory.

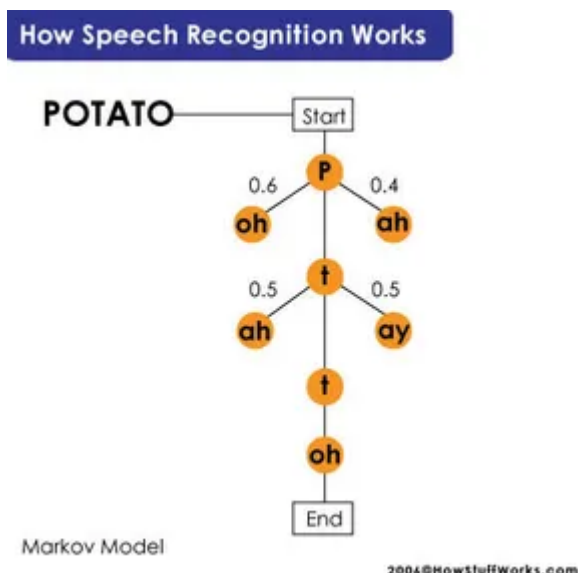
Next the signal is divided into small segments as short as a few hundredths of a second, or even thousandths in the case of **plosive consonant sounds** -- consonant stops produced by obstructing airflow in the vocal tract -- like "p" or "t." The program then matches these segments to known **phonemes** in the appropriate language. A phoneme is the smallest element of a language -- a representation of the sounds we make and put together to form meaningful expressions. There are roughly 40 phonemes in the English language (different linguists have different opinions on the exact number), while other languages have more or fewer phonemes.



The next step seems simple, but it is actually the most difficult to accomplish and is the focus of most speech recognition research. The program examines phonemes in the context of the other phonemes around them. It runs the contextual phoneme plot through a complex statistical model and compares them to a large library of known words, phrases and sentences. The program then determines what the user was probably saying and either outputs it as text or issues a computer command.

We'll take a closer look at exactly how it does this next.

Speech Recognition and Statistical Modeling



Early speech recognition systems tried to apply a set of grammatical and syntactical rules to speech. If the words spoken fit into a certain set of rules, the program could determine what the words were. However, human language has numerous exceptions to its own rules, even when it's spoken consistently. Accents, dialects and mannerisms can vastly change the way certain words or phrases are spoken. Imagine someone from Boston saying the word "barn." He wouldn't pronounce the "r" at all, and the word comes out rhyming with "John." Or consider the sentence, "I'm going to see the ocean." Most people don't enunciate their words very carefully. The result might come out as "I'm goin' da see tha ocean." They run several of the words together with no noticeable break, such as "I'm goin'" and "the ocean." Rules-based systems were unsuccessful because they couldn't handle these variations. This also explains why earlier systems could not handle continuous speech -- you had to speak each word separately, with a brief pause in between them.

Today's speech recognition systems use powerful and complicated **statistical modeling systems**. These systems use probability and mathematical functions to determine the most likely outcome. According to John Garofolo, Speech Group Manager at the

Information Technology Laboratory of the National Institute of Standards and Technology, the two models that dominate the field today are the Hidden Markov Model and neural networks. These methods involve complex mathematical functions, but essentially, they take the information known to the system to figure out the information hidden from it.

The Hidden Markov Model is the most common, so we'll take a closer look at that process. In this model, each phoneme is like a link in a chain, and the completed chain is a word. However, the chain branches off in different directions as the program attempts to match the digital sound with the phoneme that's most likely to come next. During this process, the program assigns a probability score to each phoneme, based on its built-in dictionary and user training.

This process is even more complicated for phrases and sentences -- the system has to figure out where each word stops and starts. The classic example is the phrase "recognize speech," which sounds a lot like "wreck a nice beach" when you say it very quickly. The program has to analyze the phonemes using the phrase that came before it in order to get it right. Here's a breakdown of the two phrases:

r eh k ao g n ay z s p iy ch

"recognize speech"

r eh k ay n ay s b iy ch

"wreck a nice beach"

Why is this so complicated? If a program has a vocabulary of 60,000 words (common in today's programs), a sequence of three words could be any of 216 trillion possibilities. Obviously, even the most powerful computer can't search through all of them without some help.

That help comes in the form of program training. According to John Garofolo:

These statistical systems need lots of exemplary training data to reach their optimal performance -- sometimes on the order of thousands of hours of human-transcribed speech and hundreds of megabytes of text. These training data are used to create acoustic models of words, word lists, and [...] multi-word probability networks. There is some art into how one selects, compiles and prepares this training data for "digestion" by the system and how the system models are "tuned" to a particular application. These details can make the difference between a well-performing system and a poorly-performing system -- even when using the same basic algorithm.

While the software developers who set up the system's initial vocabulary perform much of this training, the end user must also spend some time training it. In a business setting, the primary users of the program must spend some time (sometimes as little as 10 minutes) speaking into the system to train it on their particular speech patterns. They must also train the system to recognize terms and acronyms particular to the company. Special editions of speech recognition programs for medical or legal offices have terms commonly used in those fields already trained into them.

Next, we'll look at some weaknesses and flaws in speech recognition systems.

Speech Recognition: Weaknesses and Flaws



A high-quality noise-canceling microphone can help the accuracy of

your speech recognition system.

IMAGE COURTESY
AMAZON

No speech recognition system is 100 percent perfect; several factors can reduce accuracy. Some of these factors are issues that continue to improve as the technology improves. Others can be lessened -- if not completely corrected -- by the user.

Low signal-to-noise ratio

The program needs to "hear" the words spoken distinctly, and any extra noise introduced into the sound will interfere with this. The noise can come from a number of sources, including loud background noise in an office environment. Users should work in a quiet room with a quality [microphone](#) positioned as close to their mouths as possible. Low-quality [sound cards](#), which provide the input for the microphone to send the signal to the computer, often do not have enough shielding from the electrical signals produced by other computer components. They can introduce hum or hiss into the signal.

Overlapping speech

Current systems have difficulty separating simultaneous speech from multiple users. "If you try to employ recognition technology in conversations or meetings where people frequently interrupt each other or talk over one another, you're likely to get extremely poor results," says John Garofolo.

Intensive use of computer power

Running the statistical models needed for speech recognition requires the computer's processor to do a lot of heavy work. One reason for this is the need to remember each stage of the word-recognition search in case the system needs to backtrack to come up with the right word. The fastest personal computers in use today can still have difficulties with complicated commands or phrases, slowing down the response time significantly.

The vocabularies needed by the programs also take up a large amount of hard drive space. Fortunately, disk storage and processor speed are areas of rapid advancement -- the computers in use 10 years from now will benefit from an exponential increase in both factors.

Homonyms

Homonyms are two words that are spelled differently and have different meanings but sound the same. "There" and "their," "air" and "heir," "be" and "bee" are all examples. There is no way for a speech recognition program to tell the difference between these words based on sound alone. However, extensive training of systems and statistical models that take into account word context have greatly improved their performance.

We'll look at the future of speech recognition programs next.

The Future of Speech Recognition

The first developments in speech recognition predate the invention of the modern computer by more than 50 years. Alexander Graham Bell was inspired to experiment in transmitting speech by his wife, who was deaf. He initially hoped to create a device that would transform audible words into a visible picture that a deaf person could interpret. He did produce spectrographic images of sounds, but his wife was unable to decipher them. That line of research eventually led to his invention of the [telephone](#).

For several decades, scientists developed experimental methods of computerized speech recognition, but the computing power available at the time limited them. Only in the 1990s did computers powerful enough to handle speech recognition become available to the average consumer. Current research could lead to technologies that are currently more familiar in an episode of "Star Trek." The Defense Advanced Research Projects Agency (DARPA) has three teams of researchers working on Global Autonomous Language Exploitation (GALE), a program that will take in streams of information from foreign news broadcasts and newspapers and translate them. It hopes

to create software that can instantly translate two languages with at least 90 percent accuracy. "DARPA is also funding an R&D effort called TRANSTAC to enable our soldiers to communicate more effectively with civilian populations in non-English-speaking countries," said Garofolo, adding that the technology will undoubtedly spin off into civilian applications, including a universal translator.

A universal translator is still far into the future, however -- it's very difficult to build a system that combines automatic translation with voice activation technology. According to a recent CNN article, the GALE project is "'DARPA hard' [meaning] difficult even by the extreme standards" of DARPA. Why? One problem is making a system that can flawlessly handle roadblocks like slang, dialects, accents and background noise. The different grammatical structures used by languages can also pose a problem. For example, Arabic sometimes uses single words to convey ideas that are entire sentences in English.

At some point in the future, speech recognition may become speech understanding. The statistical models that allow computers to decide what a person just said may someday allow them to grasp the meaning behind the words. Although it is a huge leap in terms of computational power and software sophistication, some researchers argue that speech recognition development offers the most direct line from the computers of today to true artificial intelligence. We can talk to our computers today. In 25 years, they may very well talk back.

For lots more information on speech recognition and related topics, check out the links on the next page.

Vista SR Demo

The potential problems with using speech recognition were on public display recently in a Windows Vista demonstration. While the system performed flawlessly at opening programs

and accessing documents, when it came to transcribing text, it wasn't very accurate. The problems likely stemmed from the background noise and echo present in the large auditorium with an audience where the demo took place. A video of the incident soon spread across the Internet, hurting the reputations of Windows Vista and speech recognition in general.

Lots More Information

Related HowStuffWorks Articles

- [How E-mail Works](#)
- [How Analog And Digital Recording Works](#)
- [How Universal Translators Will Work](#)
- [How Speakers Work](#)
- [How Sound Cards Work](#)
- [How do microphones work, and why are there so many different types?](#)

More Great Links

- [National Institutes of Standards and Technology](#)
- [CNN: DARPA challenge - November 6, 2006](#)
- [University of Colorado: The Center for Spoken Dialog Research](#)
- [Microsoft: Speech Technology](#)
- [DARPA: Speech-to-speech Research](#)

Sources

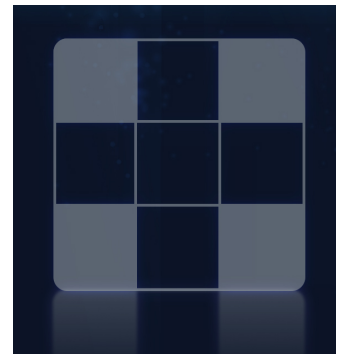
- Jelinek, Frederick. "Statistical Methods for Speech Recognition." The MIT Press, January 16, 1998. ISBN 0262100665.

- Kurzweil, Raymond. "When will HAL understand what we are saying?" in "HAL's Legacy: 2001's Computer as Dream and Reality." edited by David G. Stork, The MIT Press, February 6, 1998, ISBN 0262692112.
- "More About Bell." PBS.org.
<http://www.pbs.org/wgbh/amex/telephone/peopleevents/mabell.html>

Featured



Special Offer on Antivirus Software From HowStuffWorks and TotalAV Security



Try Our Crossword