25

Teaching a Machine to Listen

Sol Lerner

Photo taken by John Ost

Key Concepts
from Previous Chapters

- 7 Systems
- 24 Digital and Analog Signals
- 24 Data Transfer

New technologies have altered the way humans communicate. These include cell phones, faxes, web cameras, instant message programs, and more. Some might argue that these technologies have changed our society for the worse, that people now prefer the company of a computer to real human contact. Personally, I think the technologies of the Information Age have brought all of us closer together. New communications systems help us to exchange ideas and information quickly with each other, regardless of our physical limitations, differences in language and culture, or our geographic location.

Besides, I find it fascinating how machines do it. My name is Sol Lerner, and I work at a company called Scansoft, where I develop speech recognition systems—programs that allow computers to recognize and interpret the words people say.

Computers respond to spoken language by running preprogrammed tasks. Many speech recognition systems display on the screen what a person says into a microphone. Other systems have a conversation with the person, much like a human would, in order to provide a service, such as checking flight information. Scansoft is an industry leader in speech recognition. In fact, we created Dragon Naturally Speaking, which is the number-one-selling speech recognition software available today.

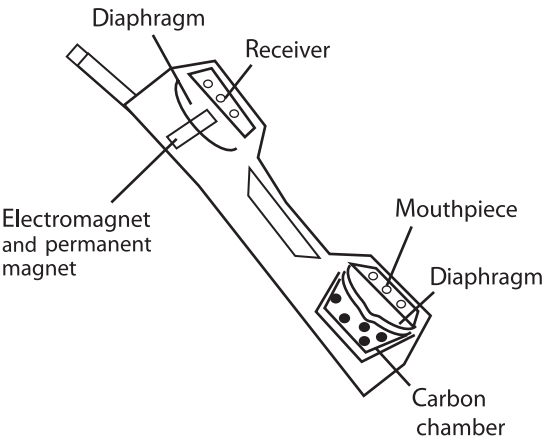
My job is to improve voice recognition systems by attempting to understand the rules humans use to communicate with sounds. I work in the field of artificial intelligence, which makes me part of a long history of scientists and engineers interested in writing programs that allow computers to "think" more like humans. Of course, a computer can't really think. Instead, it follows incredibly intricate sets of rules that allow it to respond to different stimuli in complex ways—much like a human would. But, because spoken language is so complex, a computer must be smart in order to decipher spoken words accurately.

Why are we so interested in doing this? Well, people with injuries to their hands, with chronic arthritis, or who are paralyzed or have impaired vision may have difficulty using a keyboard. Without voice recognition systems like the ones I've worked on, these people would be unable to use a computer to work, play, or communicate.

Scansoft has also developed speech recognition systems for telephone service requests. Now, when calling a business, instead of a person answering the phone, you'll often hear a computerized voice asking, "How can I help you?" When you speak, the computer responds to the key words in your request and routes the call to the correct extension.

Given all that computers can do these days, it may seem that designing such a system would be easy. In fact, it's a real dilemma. Everyone's voice is different. Many people speak with different accents, and even the way we speak can make it difficult for a computer to make sense of what we say. But in the past couple of years we've greatly improved our system's ability to recognize words and sentences. Now, some of our software, which allows a computer to display words as someone speaks, is accurate more than 95 percent of the time, after the software has been programmed to understand the user's voice.

So how do you design and build a communications system that allows humans to use their voice to interact with a machine? First, let's look at a simple communication system: the telephone. When using a telephone, you transmit your message by speaking into a microphone, sometimes called a mouthpiece. Sound waves from your voice strike a diaphragm, making it vibrate. In a simple microphone, carbon granules are compressed between two disks; the diaphragm, which consists of a thin disk; and a backing plate. The vibrating diaphragm compresses and decompresses the carbon granules inside this chamber, changing their resistance. When connected to a low-voltage electric circuit, the changing resistance of the granules also varies the current flowing through the circuit. The microphone translates the sound wave information from your voice into electrical current patterns called a signal. The electrical signal, which encodes the sound waves of your voice, carries the information through wires to a second telephone.



The speaker, or receiver, of the second telephone receives the electrical signal, which the speaker translates or decodes into sound waves that we recognize as speech. The speaker also has a diaphragm, as well as two magnets. One is a permanent magnet that constantly holds the diaphragm near it. The other is an electromagnet, usually made of iron, wrapped with a coil of wire. When an electric current passes through the coil of wire, the iron core becomes magnetized, pulling the diaphragm away from the first magnet toward the iron core. The pull of the magnet duplicates the changes in the electric current pattern, causing the diaphragm to vibrate, which moves the air forward and back and decodes your original message.

A ***signal*** is the part of a communication system that carries information.

Encoding is when a signal is translated into a form that the communications system can transmit.

Decoding is when the electrical signal is converted to a form that is useful to the receiver.

Recognizing Speech

Systems such as the telephone are great at encoding, decoding, and transmitting signals. So getting a computer to "hear" human speech is relatively easy. The electrical signal is transmitted to a computer in the same way as the telephone. The computer converts the ***analog*** signal into binary code, a ***digital*** signal of ones and zeros.

But only recently have we started trying to design systems to do what the human brain does so remarkably well—understand the spoken word. Because a computer system does not have a human brain to interpret what the signals mean, scientists like myself must give the system enough information to do so on its own.

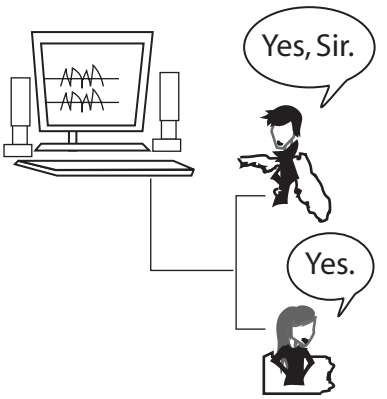
That's what I do every day at Scansoft, and it's an incredibly complex task. The computer needs to be programmed with the intricate set of rules that govern speech. Why not just catalog every word spoken in every dialect of English? If we did that, we would have a database of several hundred thousand words, and because language is constantly changing, the database would never be complete. Fortunately, every language has a much smaller group of distinct sounds. By identifying the frequencies that represent each sound, the computer can scan any cluster of signals that it receives and identify it as a particular word, letter, or sound.

Unfortunately, spoken language is not that straightforward. You may use the same sounds to express any number of ideas. For example, "I know you ate the cake" has the same sounds as "Eye no ewe 8 the cake." Human listeners can tell the difference between the two phrases because our brains use our own set of rules; we interpret the sounds how and when they are used. We know that the word "cake" is more likely to show up next to the word "ate" than the number "8." We also know from our experience that "Eye no ewe 8 the cake" does not make sense.

I spend my time analyzing language and trying to understand and make use of all of the rules that we use when we speak or hear. For example, I might create a rule that the word "bat" is often used with the verb "hit" and the noun "ball" and is often linked with "sports."

As soon as the computer recognizes the signal pattern for "B – A – T" and it recognizes the signals and sound patterns for "THE," it might be able to "understand" the sentence:

T H E/ B A T / _ _ T / T H E / B A _ _/.



By following the rules we created for associating "hit" and "ball" and linked with sports, it can then predict that the sentence is:

T H E/ B A T / H I T / T H E / B A L L/.

Whenever I travel to other parts of the country, I meet with clients who speak English differently than I do. Have you ever noticed how different a word sounds when it is spoken by someone from another part of the country? Southerners may draw out their vowels while people on the west coast clip their consonants, speaking in a punctuated rhythm. If you come from Pennsylvania, you are likely to say "yes" or "no" to a question. But a person in central Florida might answer by saying "Yes, Sir" or "No, Ma'am"—even to a machine. A computer programmed to respond to just "yes" would not be able to process "Yes, Sir." People also may have any combination of different speech patterns, depending on what city, state, or country they have lived in. So there are real challenges to developing a program that allows software to recognize any person's voice.

When it comes to developing voice recognition systems that work across languages and cultures, the challenges are endless. It might be a long time before we can develop a system that truly recognizes speech as well as humans can. It'll be much longer before we can make a computer that "thinks" or even "feels." We have only just started on the right path to solve those problems. Right now, there is no doubt that we can communicate through and with machines more easily than ever. Whether or not you believe that's a good thing, it's the reality of our time.

What's the Story?

1. What reasons does Sol offer for wanting to develop speech recognition systems?

2. Why is it challenging to design a computer system that can recognize and interpret human speech?

Designing with Math and Science

3. What's the difference between a signal and a message?

4. What parts of the telephone communications system are the encoder and the decoder? What is being encoded? Why is encoding necessary?

Connecting the Dots

5. How is electricity used in the two communications systems described in this chapter?

What Do You Think?

6. What are some other ways that humans communicate with or through machines?

7. Sol says that technologies have brought people closer together, but some people disagree. Do you agree? Explain why or why not.