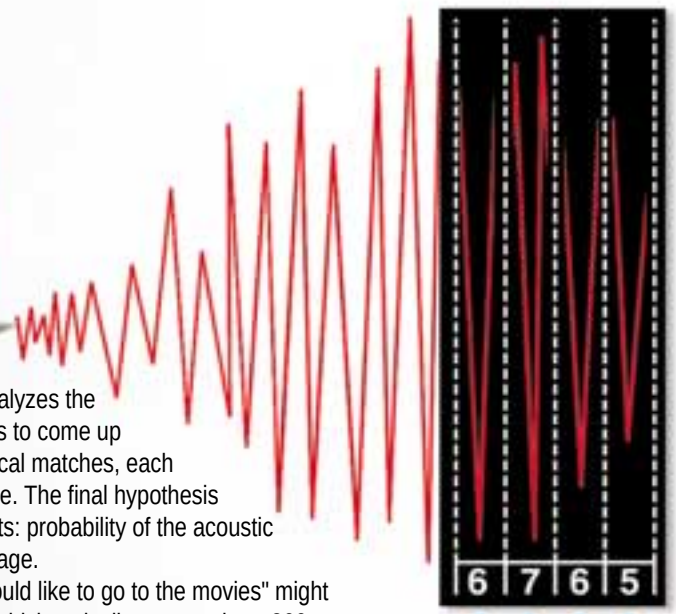
How Speech Recognition Works

Although computers can't "listen" to you and have a conversation, with the right technology, they can "hear" you and perform tasks based on your verbal commands. It may sound very futuristic, but the process breaks down into a statistical system that anticipates which word you will say next and then compares that word to a database of words to see if it finds a match.

1 The speech recognition process begins, obviously, with speech. Specifically, it starts with someone speaking into a microphone that is connected to a computer.

Software



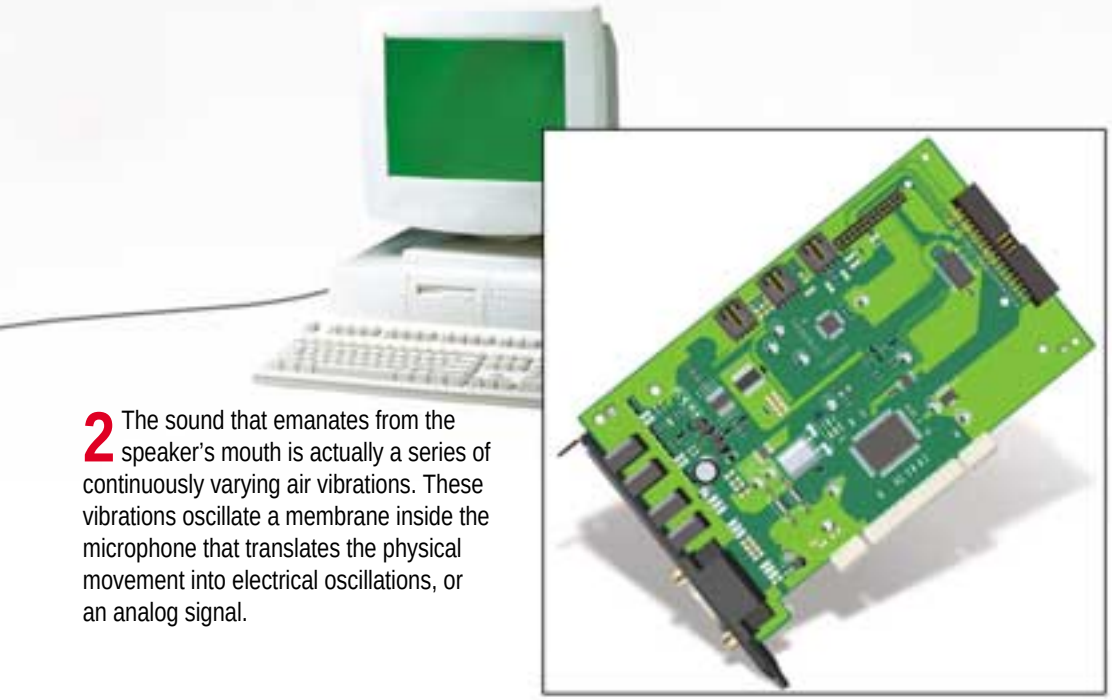
4 The software then reanalyzes the refined data. The goal is to come up with a short list of hypothetical matches, each with its own probability score. The final hypothesis is built on two main elements: probability of the acoustic and probability of the language.

A sentence such as "I would like to go to the movies" might take eight seconds to say, which typically means about 800 samples, or vectors. The vectors are run through various algorithms to clean up background noise and normalize for variations such as dialect, accents, and speech speed.

Acoustic Probability

5 Acoustic probability uses a method known as the HMM (Hidden Markov Model) to determine what the computer "thinks" was said by parsing the sounds into phonemes. This is a complicated technology that attempts to account for different ways that a sound sequence could occur for a given phoneme or a given word.

Data	Phoneme
011010	I
100100	wh
001110	oo
11100	d
101001	Lie
110110	ck



2 The sound that emanates from the speaker's mouth is actually a series of continuously varying air vibrations. These vibrations oscillate a membrane inside the microphone that translates the physical movement into electrical oscillations, or an analog signal.

3 The computer's sound card converts this analog information into the language of computers, digital data, by assigning values to the signal's characteristics at specific points in time. This sampling reduces the sound to a sequence of bits, a bunch of 0s and 1s.

Trigram Analysis

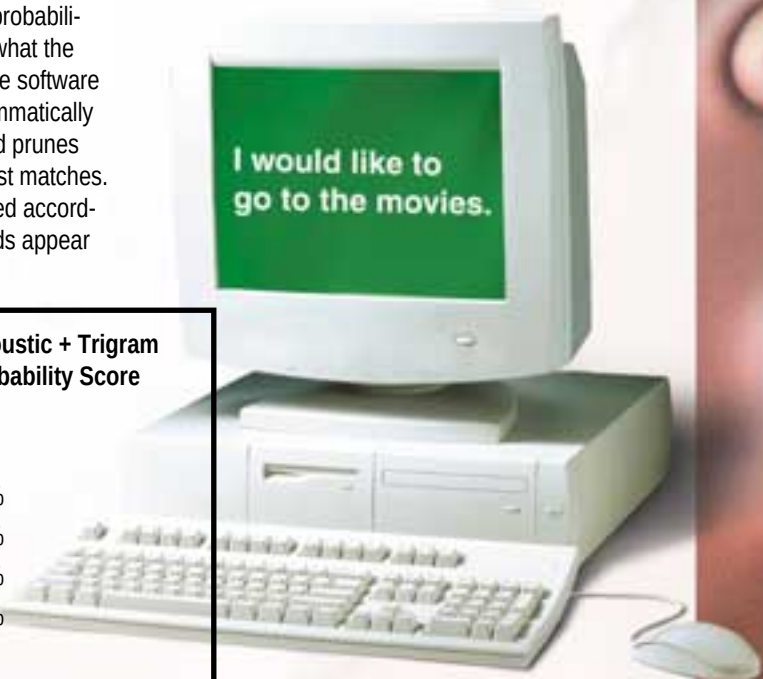
6 Probability of the language is based upon trigram analysis, which looks at words in groups of three. This process determines the probability of the third word based on the first two words. For instance, "like" often follows "I would." This step greatly improves accuracy, especially when speakers use homonyms, which are words that sound the same but are spelled differently and have different meanings.

First Two Words	Probable Match
I would	go
I would	need
I would	like
I would	bike
I would	know
I would	walk
I would	get
I would	leave

Acoustic + Trigram

7 The two analyses are combined to come up with the best probabilities, a sort of top 10 list of what the speaker may have said. The software quickly dispenses with grammatically nonsensical sequences and prunes the choices down to the best matches. When matches are approved according to their scores, the words appear on the computer screen.

Word Combinations	Acoustic + Trigram Probability Score
I would go	5%
I would need	6%
I would like	95%
I would bike	80%
I would know	50%
I would walk	30%
I would get	3%
I would leave	25%



Compiled by Anne Steyer Phelps
Graphics & Design by Lori Garris