STAR Watch

Statewide Technology Assistance Resources Project

A publication of the Western New York Law Center, Inc.

Volume 14 Issue 5 Sept-Oct 2010

Speech Recognition Programs: Steady Improvement

The origins of speech recognition go all the way back to Alexander Graham Bell. Initially, Bell was trying help his deaf wife by creating the phonoautograph, a device that would convert spoken words into visually recognizable patterns. The device actually worked, but no one was able recognize which words the patterns represented. He eventually gave up, using the knowledge gained to patent the telephone in 1876 and actually create a working version some time later.

It wasn't until the 1940's, when Bell Laboratories developed an extremely primitive device that could actually recognize a limited number of spoken words. The error rate for this device was horrendous, but it provided fundamental insights into the issues surrounding speech recognition that took several more decades to solve.

In the 1960's, researchers focused on individual word recognition. The device they created could recognize individual words if carefully pronounced and separated by pauses. The accuracy of this machine was a quantum improvement over the 1940's device. On a good day, it could achieve an accuracy rate of 40 – 60%.

A decade later, work on continuous speech recognition began. The goal was to create a device that did not require pauses between words, allowing the user to speak in a more

In this issue…

natural manner. Fast-forward another decade to the time when researchers came up with a means to solve the continuous speech issue. All that was left to do is wait for a computer to be developed that was powerful enough to perform the required tasks.

The wait was over in the 1990's, and speech recognition software products began to be marketed to the general public. Those initial offerings were somewhat less than perfect, but managed to spark a lot of interest.

How does speech recognition work?

The theory is actually very simple.

- The user starts up his/her favorite speech recognition software and begins to speak into a microphone.
- The software digitizes the spoken words of the user, then breaks the stream of words into small segments called "phonemes". A phoneme is one of the unique sounds used to create spoken English.
- The string of phonemes is then compared to a list of phonetic pronunciations of words stored in the software's built-in dictionary.
- When a word is found that matches all the phonemes in the order presented, the software assumes that it has found the desired word.

Here's a simplified example: The word "stuff"

- The word is broken into phonemes: "st" + "uh" + "ff"
- The dictionary is searched
  "**st"** + "ah" + "b" (stab)
  "st" + "uh" + "b" (stub)
  "st" + "uh" + "ff" (stuff)

That seems easy – or is it?

What's so hard about speech recognition?

In a word: everything.

- The word used in the example above doesn't have alternative pronunciations, but what about the word "potato"? Few if any people pronounce the word correctly (poh – tay – toh). It could be pronounced:
  Puh – tay – tuh
  Poh – tah – toh
  Puh – tay – duh

  How about the word "recognize"? People pronounce it:

  Reh – cog – nize
  Re – con – ize
- The program receives poor quality audio. Have you ever received a call from someone who was at a sporting event or rock concert? Have you ever been at a presentation where the other attendees around you couldn't stop talking to one another? Were you able to easily discern the message from the person that you were trying to listen to?

  The speech recognition program must be able to accurately parse the phonemes from the speaker's words. If the microphone used is of poor quality or the microphone picks up other sounds in the room such as echoes or background noise, the phonemes will not be accurately determined. This, in turn, will cause the wrong word to be selected.
- Speakers run words together. Many times there is no separation between adverbs and verbs, adjectives and nouns, or articles and nouns. Two or more words have just become one.
- People interrupt the speaker, talk over the speaker, or there are multiple speakers.

- The speaker changes the pronunciation of a word, mispronounces a word, or uses a word that does not exist in the language spoken. This could be caused by mental or physical fatigue or change in emotional state of the speaker.

How does speech recognition handle the variations in speech?

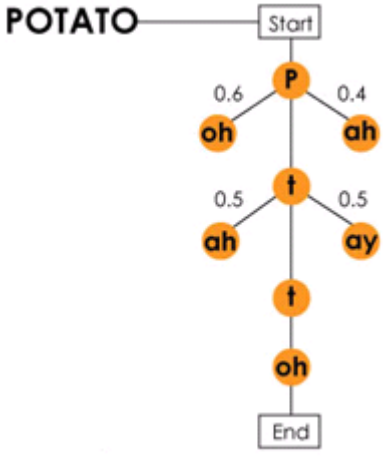
In their infancy, speech recognition programs tried to apply grammar and syntax rules to speech recognition to improve accuracy. It didn't work. The only readily available examples of grammar and syntax were contained in written documents. As any qualified linguist will tell you, for any language:

- The rules for speaking the language and the rules for writing the language are not the same. Using the written examples of the language to create predictions of spoken word usage actually increased error rates.
- Unless the speaker is extremely disciplined, typical spoken communication tends to be off-the-cuff and ad-hoc. If written down exactly as spoken, verbal communication tends to be unclear, rambling, and excessively verbose.
- People only hear about half of what is actually said.
- Spoken communication puts the burden on the listener. Compare the time it takes to process a voicemail versus the time it takes to read an email.

Though humans have been trained from birth to use their voices to communicate, voice communication remains filled with pitfalls and misunderstandings.

The breakthroughs in speech recognition accuracy came when developers started to ignore the findings of linguistic research on humans and treat computerized speech recognition as something completely different. Today's speech recognition systems use highly sophisticated statistical modeling systems to determine the most likely target words and phrases. Additionally, they rely on the user to train them in his/her speaking style and idiosyncrasies. These programs try to understand the speaker instead of assuming that the speaker fits a pre-determined profile. That approach appears to be successful.

Essentially, the voice recognition programs use the information that is known to them (the phonemes in the spoken phrase, the data accumulated from the user training the system, and the database of possible word pronunciations) and try to figure out the information that is not known (the desired words). That sounds easy enough, but is it really? Let's explore just one example:



Using the most common speech recognition model, each phoneme is like a link in a chain. The links are assembled to create a chain that is the equivalent of one word.

As we stated earlier, words have multiple pronunciations. The dictionary in the program contains each word described as a linked set of phonemes, but with a significant difference: It also contains the known alternate pronunciations of the various phonemes in the word and the probability that the phoneme would be used. These are stored as branches on the chain.

In order to decide if this word is the "best" word to match the spoken word, the chain of sounds from the spoken word is compared to the chain of sounds in the dictionary word. For every phoneme matched, the word score is increased.

When branches in the dictionary word chain are found, the sound from the spoken word is compared to the choices in the dictionary word. If there is a match, the score for that phoneme is added to the word score.

If the word score is high enough, the program assumes that the potentially correct word was found. It will then try to find other possible word matches.

But that is only one word. Humans string words together into phrases and sentences. After arriving at an entire array of possible words and scores, the program attempts to fit the word possibilities into a phrase that maintains the order of the spoken phonemes.

Each of the possible phrases is given a score based on the likelihood of the words being used in the manner found in the phrase. The phrase with the highest score is the phrase that is chosen – But is that phrase the correct one? It depends on how well the program was trained to understand the speaker.

Consider the phrase "recognize speech". When asked, we would all say that the correct pronunciation is "reh-cog-nize speech". But, when we are tired or distracted (or consumed some quantity of alcoholic beverage), we might say "reh-con-ize speech" (the "g" is missing). Without any training, the speech recognition program might determine that the spoken phrase is "wreck a nice beach".

To a speech recognition program, training is everything. Why? Because the training causes the program to adjust the assigned probability scores on words and phrases. This increases the likelihood that the words that come out of the program match what the speaker actually said. How long does it take to train a program?

Many of the vendors claim that it can be done in as little as 10 minutes. We will concede the point that 10 minutes of training will create a noticeable improvement over no training at all, but it is hardly adequate for someone who intends to become a power user. It would be more accurate to say that anyone who uses a speech recognition program on a regular basis will periodically set aside time to train and retrain the system. The quality of the output will justify the time spent.

Is it possible for speech recognition programs to achieve 100% accuracy?

When you speak, do people understand you 100% of the time? We sincerely doubt it. When people don't understand what you said, is it possible that you misspoke, or is it completely the fault of the listener? Speech recognition is a two-way street for human-to-human and human-to-computer. Several studies of speech recognition in human-to-human interactions arrived at the conclusion that the error rate is 2-4%. In articles about human-to-computer speech recognition, the error rate can be as high as 20% -- but that was the error rate from years ago. The currently available products are capable of achieving 95% accuracy, depending on the amount of training performed by the user of the program. That isn't quite as good as what humans are capable of, but give it a little more time. It may actually get to be better than a human.

In 2004, a gentleman by the name of Mike Bliss composed a poem about voice recognition. Upon completing the poem, he dictated the poem into the voice recognition software on his computer and recorded the result:

a poem by Mike Bliss

like a baby, it listens
it can't discriminate
it tries to understand
it reflects what it thinks you say
it gets it wrong... Sometimes
sometimes it gets it right.
One day it will grow up,
like a baby, it has potential
will it go to work?
will it turn to crime?
you look at it indulgently.
you can't help loving it, can you?

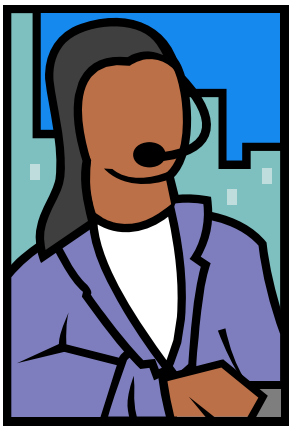
The voice recognition software gave the following result:

a poem by like myth

like a baby, it nuisance
it can't discriminate
it tries to oven
it reflects lot it things you say
it gets it run sometimes
sometimes it gets it right
won't day it will grow bop
Ninth a baby, it has provincial
will it both to look?
will it the two crime?
you move at it inevitably
you can't help loving it, cannot you?

In 2008, he repeated the experiment with a newer voice recognition program. This time, the software correctly recognized all but two of the words in the poem.

Almost three years have passed since the experiment was repeated. Possibly, the error count might be down to just one word – or zero.



WNYLC Web Statistics For September 2010

Total Hits…….....……483,075
Number of Pages Viewed.....…..…...152,559
Total Visitors.....……61,867
Average Hits/Day…….....…..……16,102
Average Pages /Day…….……..……5,085
Top Web Browsers Used:
Internet Explorer 8.x…….…….....27%
Internet Explorer 7.x…….…….....33%
Internet Explorer 6.x……...22%
Firefox.....…..……...……15%
Google Chrome.....1%

Top 5 Operating Systems Used:
Windows 7…….....……9%
Windows Vista.....…….……14%
Windows XP.....……47%
Mac OS.....…….……....….3%
Other.....…….……27%

WHO WE ARE

Joe Kelemen - Attorney
Marisa Villeda - Attorney
Keisha Williams - Attorney
Brad Davidzik - Attorney
Danielle Mayer-Dorociak - Attorney
Graham Leonard - Paralegal
Tom Karkau - Programmer
Sherry Soules - Administrator
Joy McDuffie - Data Analyst

Wnylc@wnylc.com
716-855-0203
www.wnylc.net

Want to know when StarWatch is available?

If you wish to receive an email telling you when the next edition of StarWatch is available, please email us at starwatch@wnylc.com. In the subject area, simply enter the word “Subscribe”. When the next edition of StarWatch is available, we will send you an email that contains a link to the newsletter.

If don’t wish to receive email notifications, send us an email to at starwatch@wnylc.com with the word “Unsubscribe” in the subject area. We will stop sending email notifications to you.

WNYLC values your privacy. If you provide us with your email address, Western New York Law Center will not give the information to any other organization.