Speak Naturally

Speech recognition technology is becoming independent of speakers and languages.

By Tania Hershman

Machine recognition of continuous, real-world speech has been one of the most complex challenges faced by linguists and software developers. However, several new packages coming to market this year, from companies such as Dragon Systems, IBM, Lernout & Hauspie, and Philips, can deal with normal speech, recognizing up to 200 words per minute. "Speaking speed is no longer an issue. How fast can you think?" says Melvyn Hunt, managing director of Dragon Systems U.K.

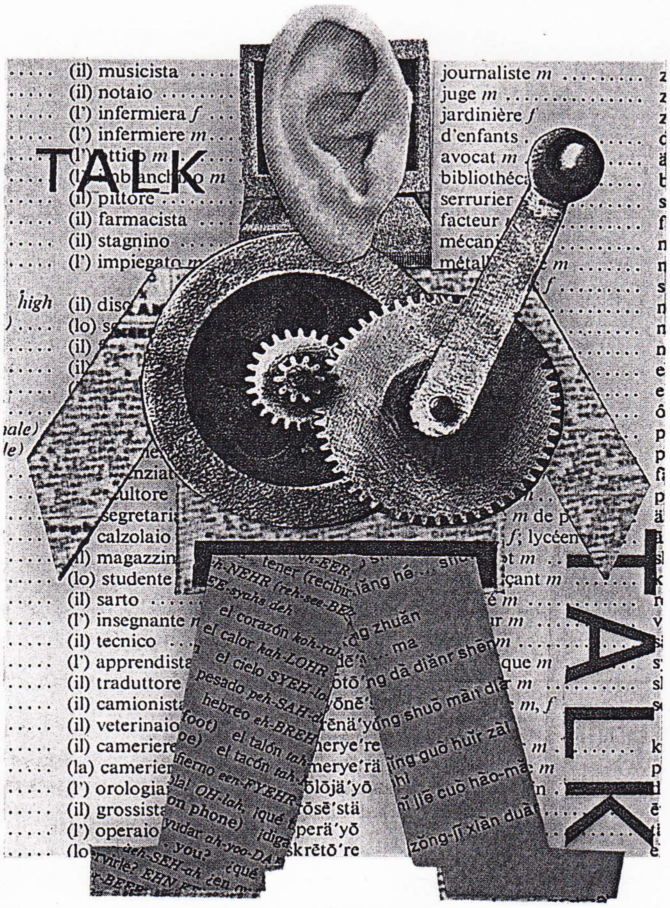


ILLUSTRATION: LAUREN LING © 1997

There will never be an exact match between two spoken words. Even one person doesn't say the same word in the same way twice. Speed, emphasis, and length of the pronunciation all vary depending on context but also on the speaker's emotional situation, making speech recognition a complicated task for developers.

Continuous speech dictation software has been available for the last 18 months but has been limited to around 25,000 words and profession-specific vocabularies, such as for radiologists (e.g., IBM's MedSpeak). Now continuous systems such as Dragon's NaturallySpeaking are starting to replace existing systems that usually work only with discrete speech punctuated with pauses or that are limited in vocabulary.

Simply put, these new systems do the same as humans do, albeit primitively. They separate speech into words or phonemes (the basic building blocks of speech), compare the acoustic patterns of the speech with the patterns stored in a database, and find the most likely word.

General-purpose dictation software such as Dragon Dictate, IBM VoiceType, and Kurzweil AI Voice typically come with up to 60,000-word vocabularies and the ability to add new words. However, they cannot cope with input at natural talking speeds (limited to about 100 words per minute), and they require the user to punctuate sentences with short pauses. They usually "understand" straight out of the box, although they work better when given the chance to adapt to a regular user's speech patterns and learn frequently used words.

Today's dictation software, when adapted to a user's speaking characteristics and optimized for certain contexts, achieves around 95 percent accuracy. However, the ultimate aim is for all systems to be speaker-independent and multilingual.

Dragon's new NaturallySpeaking, one of the first general-purpose continuous speech dictation packages, is an example of software that heralds the next generation in computer dictation. Although the first version doesn't allow the speaker to dictate into other applications, you can paste recognized text into other software. Also, it does not include the command-and-control features that come with some discrete dictation packages for vocally navigating around the computer, opening and closing applications, and even surfing the Net hands-free.

Processing natural speech eats up a lot of computing power, and this is one reason why it has taken until now for viable commercial products to hit the shelves. "When our first system was developed in 1993, the processor power of a PC was not sufficient to run natural continuous speech recognition," says Ralph Preclik, communications director of Philips Speech Processing. "We had to develop a dedicated accelerator board at that time."

With the introduction of Pentium Pro and MMX technology, speech recognition applications are now running straight from the CPU without a dedicated DSP to perform the signal-processing analysis. According to Preclik, the bottlenecks in speech recognition are now related to other factors; for example, the insufficient display speed of word processing applications.

continued

Most new (and also many earlier) speech recognition applications not only require high computing power but also a minimum of 32 MB of RAM. However, in an embedded-system environment such as a mobile phone, algorithms have to get along with much reduced system resources and perform one- or two-word recognition at best.

Speech recognition algorithms that identify your utterances as a sequence of whole words are usually very fast. But they require more training and greater processing power. Therefore, they apply very well to small-vocabulary applications such as command/control or hands-free phone dialing.

On the other hand, algorithms that recognize phonemes, the basic building blocks of spoken language, are usually more compact and flexible. Phoneme-based algorithms allow for the addition of new words to a vocabulary by identifying and combining existing phonemes. (Most languages have between 30 and 60 phonemes, so the number of combinations is huge but manageable.)

An automated directory-inquiry system, which can retrieve, for example, a name without linguistic context, is a typical application of phoneme-based algorithms. Phonetic Systems' Phonetic Database Server, for example, uses such algorithms for speech recognition and rapid searching of very large databases. It can currently handle databases containing 100,000 names, but the company aims to have search capabilities of one million entries by the middle of next year.

Both types of algorithms reinterpret the signal phonetically and match it with its database of acoustic samples by allotting probability scores to possible word matches. Hidden Markov Modeling, based on a two-stage probabilistic process, is currently the most popular statistical modeling technique used for allotting such scores. Alternative models that use neural networks do not perform as well as Hidden Markov Models (HMMs). Says Philips' Ralph Preclik, "Today neural nets can gain acceptable performance only in combination with HMMs."

Acoustic matching produces the most likely phonemes or words, but this is not the end. Words can be spoken in different ways, at different speeds, so intelligence is needed to make the leap from a combination of phonemes into actual words or sentences. This process is called linguis-

The GlobalPhone Project

Tanja Schultz, a German computer scientist, hopes to break down some language barriers when she finishes her Ph.D at the University of Karlsruhe. She is working on a multilingual speech recognition system–called GlobalPhone–that could provide access to information regardless of the speaker's language. Professor Alex Waibel, who leads the speech recognition groups at Karlsruhe and at Carnegie Mellon University in the U.S., supports the project.

"The user will be able to speak to the system in his native language, and the system will decide what language was spoken and recognize the input," Schultz says. After GlobalPhone has identified the language and recognized a user's speech or commands, it will be able to turn the content into text, then translate it or rephrase it as synthesized speech.

The goal of the project is to emerge with a system that recognizes any one of the 12 most widespread languages. English, French, and German linguistic databases are already available. In addition, the GlobalPhone team has collected high-quality databases of samples in Arabic, Chinese, Japanese, Croatian, Korean, Portuguese, Russian, Spanish, and Turkish.

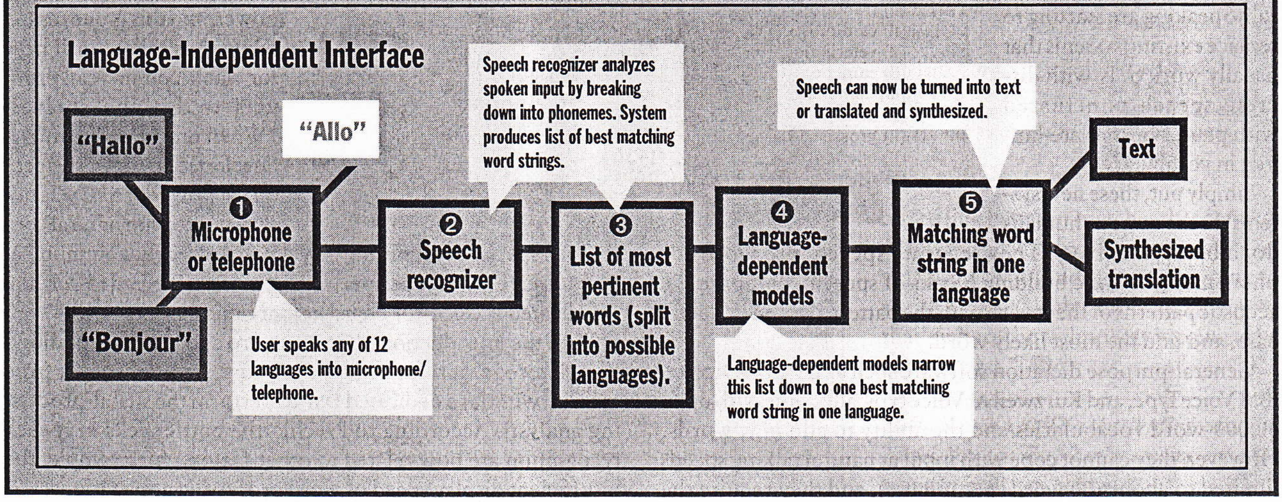
For each language, the GlobalPhone researchers asked about 100 native speakers to read 20 minutes of newspaper text. They recorded each session digitally and characterized the recording session for each person by speaker characteristics and environmental conditions.

"The data collection is now done," says Schultz. The next step will be the training of the recognition engine based on the collected acoustic samples.

The GlobalPhone engine uses a phoneme-based algorithm, and its dictionary contains all known words from each language in a multilingual phoneme set. "Our phonemes are no longer language-specific but shared by several languages," explains Schultz.

When up and running, the GlobalPhone engine will produce a list of the most pertinent word strings separated into different languages. A scoring procedure will then reduce the number of best words and result in a best-matching word string. Schultz expects to have a running version with this functionality next spring.

The number of potential applications is huge. It includes any sorts of multilingual information and ordering systems, automatic telephone operators, or translation services.


Language-Independent Interface
"Hallo"
"Allo"
"Bonjour"
1 Microphone or telephone
User speaks any of 12 languages into microphone/telephone.
Speech recognizer analyzes spoken input by breaking down into phonemes. System produces list of best matching word strings.
2 Speech recognizer
3 List of most pertinent words (split into possible languages).
4 Language-dependent models
Language-dependent models narrow this list down to one best matching word string in one language.
5 Matching word string in one language
Speech can now be turned into text or translated and synthesized.
Text
Synthesized translation

tic matching. The speech engine then emerges with what it considers to be the most likely word that was spoken.

Multiple Languages

The Holy Grail of computer linguists is a language-independent, speaker-independent, continuous speech recognition interface. Lernout & Hauspie's Language Factory, a software development kit (SDK), helps developers move closer to this paradigm. This suite of multilingual, speaker-independent technologies—which includes components for automatic speech recognition, text-to-speech conversion, translation, and digital speech compression—is tailored to small and medium vocabularies. It has already been implemented in a variety of areas, such as language-learning software, voice verification systems, and car navigation.

Lernout & Hauspie's SDK probably has the widest range of supported languages. Its products are available in U.S. English, U.K. English, French, German, Italian, Cantonese, Dutch, Korean, Malay, and Spanish. Japanese, Mandarin, Portuguese, and Russian versions are currently under development.

Building a speech recognition engine in multiple languages requires a lot of resources because you need to collect a large database of speech samples first, including all accents, dialects, and the unique sounds in that language. "This is quite a lengthy process, not least because you must have recordings of several hundred speakers to be able to produce a good model," says Richard Winski, manager of language resources and technology at Vocalis Group. "With access to a suitable database, however, you can normally add a new language in a few weeks."

Each new language presents a unique challenge. "You have to devote a lot of resources to the peculiarities of each language," says Hunt of Dragon Systems U.K. English, for example, is difficult to synthesize because pronunciation is not always obvious from the way a word is spelled. French, on the other hand, is more difficult to recognize. The French verb *appeller* (to call), for example, can be spelled 12 different ways yet pronounced identically. In German, compound words are difficult to deal with, and the various Chinese dialects differ largely in tone, which isn't an issue in European languages. A case in point is the Chinese word *ma*, which can have five different meanings, depending on intonation.

One of the first companies to rise to the Chinese language challenge was Motorola's Lexicus division. Discrete speech recognition software has been very difficult for Chinese because word boundaries are sometimes ambiguous. As a result, speedy recognition in Chinese wasn't possible until continuous systems worked well enough. Motorola's Chinese continuous speech recognition engine, released late last year, can now recognize over 10,000 spoken words running on a standard PC. That's good news for the 20 percent of the world's population that speaks Chinese. B

Tania Hershman is a freelance writer based in Jerusalem. You can contact her by sending e-mail to **t_hersh@netvision.net.il.**

The Ultimate User Interface

Star Trek had it right: Speech is the best user interface. We're starting to see innovative applications that use speech to turn on and off the lights in your house, for example, or to tie all your inboxes together and give you access to their contents from a remote phone, or to replace touch-tone phone commands and menus.

Registry Magic Virtual Operator, from Registry Magic, is a new office automation tool that can answer and direct calls without an operator. A bank can program it to check a customer's balance after it matches a verbal password to their stored voiceprint.

Keyware Technologies recently released a software development kit that will allow system integrators and value-added developers to create software verification applications based on VoiceGuardian software technology. The SDK provides an API for voice verification using a dynamic link library, ActiveX control, or Windows NT service. This API can be used to construct secure stand-alone or client/server applications. It also includes sample programs for a voice-secured Web site.

The company also sells an application that combines both facial and voice verification technologies in a single integrated security system. Called Keyware S^2 Security Server, the system matches facial and vocal input against a centrally stored user profile. In highly sensitive or classified areas, a special input station could prompt a user for a password while capturing a facial image and asking that the user speak an ID into a microphone.

And on the home front, you can now control almost every appliance in your house from anywhere in the world. A program called HAL2000, from Home Automated Living, provides interactive control of your domestic domain through continuous speech recognition technology. You just speak naturally to your appliances. Household appliances are controlled using X-10, RF, or infrared devices.

WHERE TO FIND

Dragon Systems
Bishops Cleeve, Cheltenham, U.K.
+44-1242-678-575
fax: +44-1242-678-301
info@dragonsys.com
http://www.dragonsys.com

GlobalPhone Project
Karlsruhe, Germany
+49-721-608-4735
tanja@ira.uka.de
http://werner.ira.uka.de/~tanja

Home Automated Living
Burtonsville, MD, U.S.
+1-301-879-2305
fax: +1-301-384-8275
info@AutomatedLiving.com
http://www.AutomatedLiving.com

Keyware Technologies
Brussels, Belgium
+32-0-2-721-4574
fax: +32-0-2-721-5015
http://www.keyware.be

Kurzweil AI
Waltham, MA, U.S.
+1-617-893-5151
fax: +1-617-893-6525
Info@kurzweil.com
http://www.kurzweil.com

Lernout & Hauspie
Ieper, Belgium
+32-57-228-888
fax: +32-57-208-489
sales@lhs.be
http://www.lhs.com

Motorola Lexicus Division
Palo Alto, CA, U.S.
+1-650-494-0800
fax: +1-650-494-1141
danab@lexicus.mot.com
http://www.mot.com/MIMS/lexicus

Philips Speech Processing
Vienna, Austria
http://www.speech.be.philips.com

Phonetic Systems
Petach Tikva, Israel
+972-3-921-0905
fax: +972-3-921-0966
jerry_p@phonetic.co.il

Registry Magic
Boca Raton, FL, U.S.
+1-561-367-0408
fax: +1-561-367-0608

Vocalis Group
Cambridge, U.K.
+44-1223-846177
fax: +44-1223-846178
enquiries@vocalis.com
http://www/vocalis.com