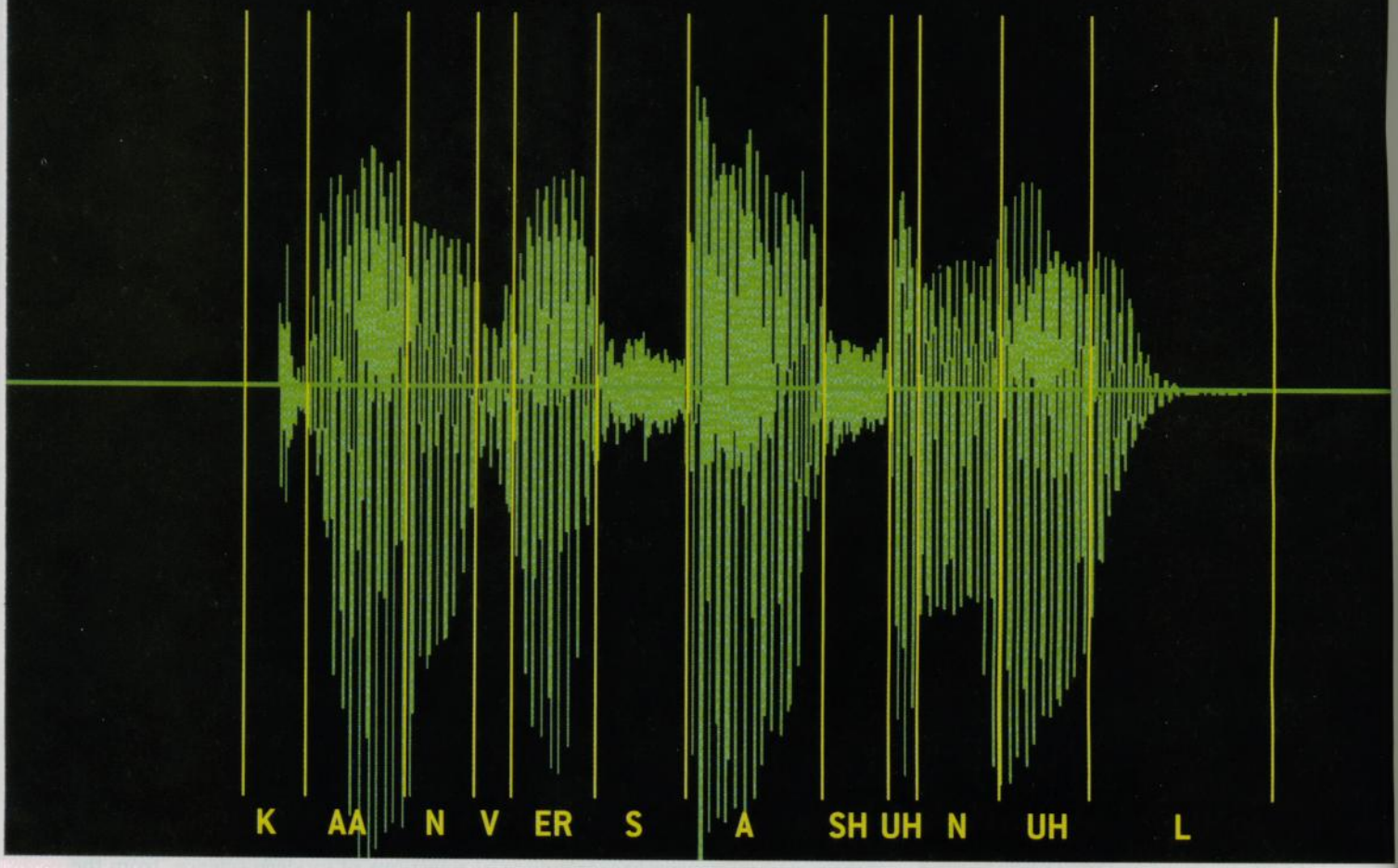
CONVERSATIONAL



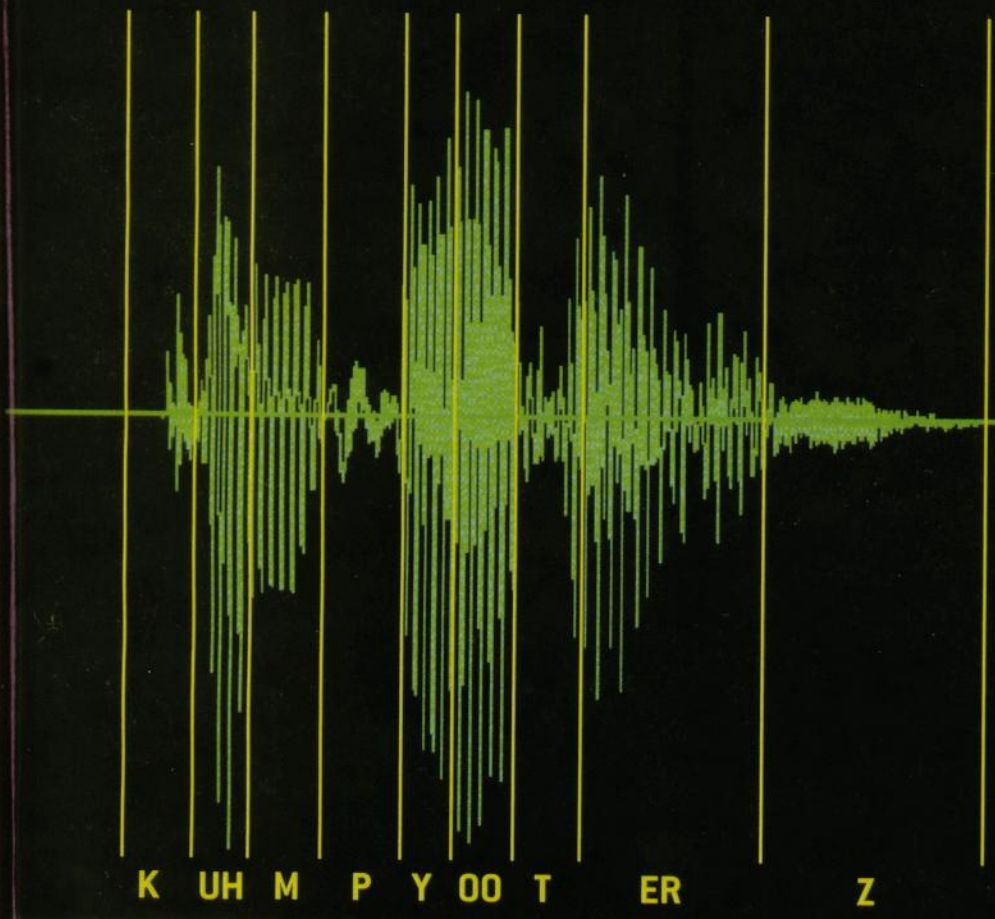
Call a large company these days, and you will probably start by having a conversation with a computer. Until recently, such automated telephone speech systems could string together only prerecorded phrases. Think of the robotic-sounding "The number you have dialed ... 5 ... 5 ... 5 ... 1 ... 2 ... 1 ... 2...." Unfortunately, this stilted computer speech leaves people cold. And because these systems cannot stray from their canned phrases, their abilities are limited.

Computer-generated speech has improved during the past decade, becoming significantly more intelligible and easier to listen to. But researchers now face a more formidable challenge: making synthesized speech closer to that of real humans—by giving it the ability to modulate tone and expression, for example—so that it can better communicate meaning. This elusive goal requires a deep understanding of the components of speech and of the subtle effects of a person's volume, pitch, timing and emphasis. That is the aim of our research group at IBM and those of other U.S. companies, such as AT&T, Nuance, Cepstral and ScanSoft, as well as investigators at institutions including Carnegie Mellon University, the University of California at Los Angeles, the Massachusetts Institute of Technology and the Oregon Graduate Institute. Like earlier phrase-splicing approaches, the latest generation of speech technology—our version is code-named the IBM Natural Expressive Speech Synthesizer, or the NAXPRES Synthesizer—is based on recordings of human speakers and can respond in real time. The difference is that the new systems can say anything at all—including natural-sounding words the recorded speakers never said.

Commercial and institutional enterprises worldwide spend billions of dollars annually on speech-enabled information service centers. The centers rely on a suite of technologies: speech recognition, language understanding, database searching, text generation and, finally, speech synthesis. Synthetic speech, produced by linking words or pieces of words recorded by human speakers, serves as the personification, or human face, of the whole system; hence, people often judge the system by the quality of the voice they hear. An expressive voice, for instance—one whose tone adjusts appropriately both when the system can accommodate your reservation request and when it can-

ANDY AARON AND JOHN F. PITRELLI

COMPUTERS



BY ANDY AARON, ELLEN EIDE AND JOHN F. PITRELLI

EFFORTS TO MAKE COMPUTERS SPEAK NATURALLY WILL LET MACHINES BETTER COMMUNICATE MEANING

not—means a more pleasant and effective experience for callers.

Consumers will soon benefit from a variety of new services made possible by this fast-evolving technology. These services will offer verbal delivery of up-to-the-minute news and weather reports that would otherwise be available only as text. The technology can also tirelessly read aloud written materials for the handicapped or foreign-language students. Other helpful capabilities are vocal human-machine interaction to control automotive functions, including recital of automated driving directions that contain the many millions of different street names in the world and retrieval of e-mail messages over the phone—or from any information system—without need for a visual display.

In time, natural-sounding synthetic speech will give meaningful voice to handheld and household devices. And at some point, the technology will be used to generate lifelike speech for characters in video and computer games and even in motion pictures.

Speaking Machines

SYNTHESIZED SPEECH is both a triumph of technology and the latest incarnation of an old dream. Attempts to simulate human speech date to the late 1700s, when Hungarian scientist Wolfgang von Kempelen built what he called a Speaking Machine, which employed an elaborate set of bellows, reeds, whistles and resonant chambers to produce rudimentary words.

By the 1970s digital computing had enabled the first generation of modern text-to-speech systems to reach fairly wide use. Makers of these systems attempted to model the entire human physiological speech production process directly, using a relatively small number of parametric features. The model typically has an audio source, which takes on the role of a person's larynx, and an audio filter, which acts as the rest of a human vocal tract. The system adjusts physical aspects of the sound (resonance, bandwidth, periodicity and fundamental frequency) continuously to create the sequence of sounds needed to compose speech.

The result was intelligible, though mechanical-sounding, speech. An early example of a mass-market product incorporating this technology was the Speak & Spell toy launched in 1978. Such synthesizers still have a place nowadays because they are simple to make

and can generate intelligible speech at tremendous speeds—up to 600 words a minute. (English is generally spoken at 140 to 190 words a minute.) Those willing to trade natural-sounding speech for speed, such as the visually impaired, find these systems useful.

The advent of faster computers and inexpensive data storage in the late 1990s made possible today's most advanced synthesizers. Researchers, including our team at IBM, base designs on linguistic building blocks called phonemes and arrange sequences of recorded phonemes to create any given word. The word "school," for example, contains four phonemes, which we might call S, K, OO and L. Languages differ in the numbers of phonemes they contain. English makes use of about 40 distinct phonemes, whereas Japanese has about 25 and German 44. Just as typesetters once sequenced letters of metal type in trays to create printed words, current synthesizers fit together chunks of speech to create spoken words. Engineers call these systems concatenative synthesizers because they link together small pieces of sound. We will explain how we build such a system and then describe how the synthesizer produces lifelike speech in real time [*see box on opposite page*].

Build Me a Voice

CONCATENATIVE SPEECH synthesis starts with a human voice, so when our group develops a new system, we audition dozens of candidate speakers. Unless a foreign inflection is required, as for, say, a character in a movie or on a Web site, our preference is usually for speakers free of regional accents; they use the general American English dialect of many television news anchors. A selected speaker then sits in a recording booth and reads aloud more than 10,000 sentences, a task that takes about two weeks. We pick the sentences in part for their relevance to real-world applications and in part for their diverse phoneme content, which ensures that we capture many examples of all the English phonemes in different contexts.

The result is about 15 hours of recorded speech. Ensuring the consistency of the entire 15 hours is no small task, however. Because the recordings are destined to be chopped up and reassembled as needed, a speech sound recorded one day may wind up next to another archived a week later. Thus, a director or voice coach guides speakers and listens for deviations in their speaking rate, emotional tone, overall pitch and loudness, helping them maintain uniformity. At least once an hour the speakers hear a sample sentence recorded on the first day for reference, much as a musician uses a pitch pipe to stay in tune.

The software then converts the words from the text that was spoken into a series of phonemes using a pronunciation dictionary, a reference that lists the phonemes that make up each word. It notes specific features of each phoneme occurrence, such as what phonemes preceded and followed it and whether it begins or ends a word or a sentence. The system also identifies the part of speech of each word in the text.

Once the text has been processed, our software analyzes the audio recordings, measuring them for three characteristics: pitch, timing and loudness—collectively called prosody. Knowing these features for each phoneme recording helps us decide which example to use to synthesize a given phrase.

Next, using techniques borrowed from speech recognition—dictation programs that translate speech into text—the computer code associates each recorded phoneme with its text counterpart. With the audio and the text aligned, our software can analyze each recording and pinpoint the boundaries at which each phoneme begins and ends. This procedure is key, because once the phonemes are located and labeled, the software can then precisely catalogue them for use in a searchable database.

The database for our NAXPRES Synthesizer contains an average of 10,000 recorded samples of each of the 40 or so phonemes in the English language. That might at first appear to be taking redundancy to absurd levels. But when words are combined into sentences, the relative loudness and pitch of each sound change, based on the speaker's mood, what he or she wants to emphasize, and the type of sentence (think about the difference between a question and an exclamation). So the phoneme samples derived from these sentences can vary significantly: some were spoken with different prosodies, others were employed in differing phonemic contexts, and so forth.

Because human speech is so amazingly subtle and complex, experts understand only a few of the many effects that contribute to natural-sounding speech. Hence, we need computers to help do the job. We use our speakers' database to build a statistical model to infer automatically the general properties that govern the rise and fall of pitch, as well as the duration and loudness of each person's speech. The model will apply the properties later to make the system's speech sound more humanlike.

Talk to Me

NOW THAT WE HAVE described the elements of a modern speech synthesizer, let us take a closer look at one in action. IBM's speech synthesizer does all the following processing in milliseconds—fast enough that people can con-

Overview/*Making Machines Speak*

- As computer-generated voice transactions become increasingly common in everyday life, researchers are coming ever closer to synthesizing human-sounding speech.
- Drawing on huge databases of recorded word fragments (phonemes), new speaking machines can modulate tone and expression to better communicate meaning to users.
- These systems are finding applications in mobile electronic devices such as automotive navigation systems. At some point, video and computer games and even films will take advantage of lifelike artificial speech capabilities.

HOW RESEARCHERS GIVE VOICE TO COMPUTERS

The process by which speech-synthesis researchers and engineers make it possible for a computer to speak in a humanlike fashion is a complex one, involving recording a person's voice and then rearranging the component sounds—known as phonemes—to produce words and sentences they never said.

BUILDING A SPEAKING MACHINE

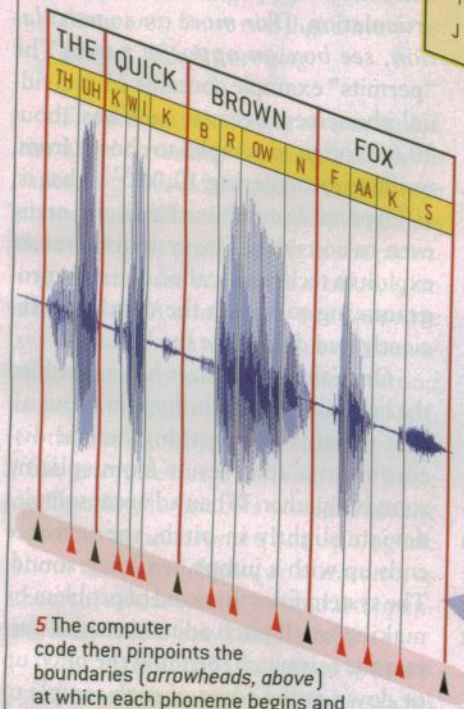
1 Sound engineers record more than 10,000 sample sentences selected for diverse phoneme (component sound) content and relevance to practical applications

- The quick brown fox
- Thank you for calling
- The Dow Jones rose four . . .

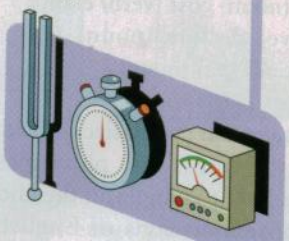
2 The system splits the words into phonemes

PHONETIC DICTIONARY

- TH UH K W I K B R OW N F AA K S
- TH A NG K Y OO F AW R K AU L I NG
- TH UH D OW J O N Z R O Z . . .

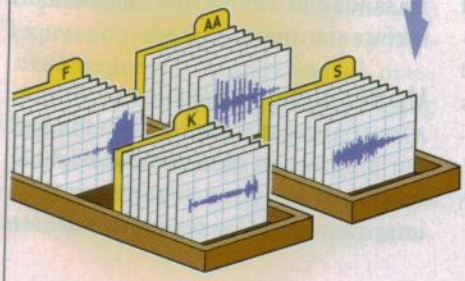


3 Software analyzes the recordings, measuring their pitch, timing and loudness



4 Speech-recognition software correlates each recorded phoneme with its text counterpart

5 The computer code then pinpoints the boundaries (*arrowheads, above*) at which each phoneme begins and ends and catalogues them for use in a searchable database



6 Using the speaker's database, researchers build a statistical model to infer automatically the general properties that govern the rise and fall of pitch, as well as the duration and loudness of each person's speech. The model will apply the properties later to make the system's speech sound nearly human

RUNNING A SPEAKING MACHINE

1 Given text to be converted into speech, the system translates any symbols or abbreviations into words and then analyzes the sentences' grammar as well as any pronunciation ambiguities to produce a string of suitable phonemes

52°, winds 10 mph

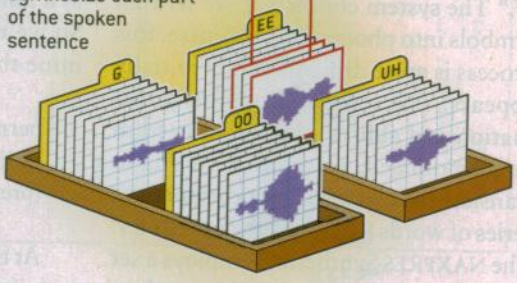
FIFTY TWO DEGREES, WINDS TEN MILES PER HOUR

WINDS?

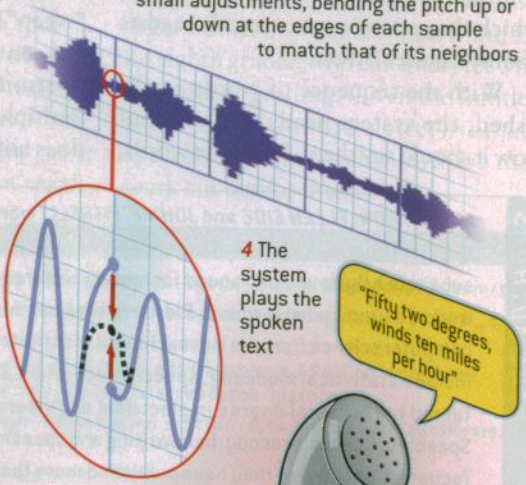
WINDS?

F I F T EE T OO D UH G R EE Z . . .

2 The system selects the phoneme examples with the most appropriate pitch, timing and loudness features to best synthesize each part of the spoken sentence



3 With the best phonemes assembled in a row, the system smoothes out discontinuities that result from splicing sounds together. For example, when adjacent samples deviate slightly in pitch, the software makes small adjustments, bending the pitch up or down at the edges of each sample to match that of its neighbors



4 The system plays the spoken text

"Fifty two degrees, winds ten miles per hour"

INDUSTRIAL TEXT-TO-SPEECH RESEARCH

COMPANY	PRODUCT NAME	WEB SITE
Acapela Group BELGIUM	Acapela TTS	www.brightspeech.com
Advanced Telecommunications Research (ATR) JAPAN	*No name*	www.slt.atr.jp/ss-e
AT&T U.S.	Natural Voices	www.naturalvoices.att.com
Cepstral U.S.	Cepstral Voices	www.cepstral.com
Fonix U.S.	DECtalk	www.fonix.com/page.cfm?name=espeech_dectalk
IBM U.S.	NAXPRES Synthesizer	www.research.ibm.com/tts
iFLYTEK CHINA	InterPhonic	www.iflytek.com/english/products.htm
Infotalk HONG KONG	InfoTalk-Speaker	www.infotalkcorp.com
Loquendo ITALY	Loquendo TTS	www.loquendo.com
Nuance U.S.	Vocalizer	www.nuance.com
Scansoft U.S.	RealSpeak	www.scansoft.com
SVOX SWITZERLAND	SVOX-TTS	www.svox.com
Toshiba JAPAN	*No name*	www.toshiba.co.jp/rdc/mmlab/tech/w21e.htm

verse with the computer in real time. First we will give it something to say in text form, such as "Permits cost $80/yr." The system converts these written symbols into phonemes. Of course, this process is more difficult than it initially appears. The sentence contains punctuation and abbreviations that need to be pronounced, so the first step is to translate the text into the corresponding series of words for the synthesizer to say. The NAXPRES Synthesizer employs a set of rules to clear up ambiguities such as multiple ways of interpreting abbreviations. "St. Charles St.," for example, contains two identical abbreviations, which the system must correctly read as "Saint Charles Street."

With the sequence of words established, the system needs to figure out how it should say them. For some words, pronunciation depends on the part of speech. For instance, the word "permits" is spoken *per*mits when it is used as a noun and per*mits* when it is a verb. So we use a grammar parser to determine the part of speech of each word:

> permits (noun) cost (verb) eighty (adjective) dollars (noun) per (preposition) year (noun)

At this stage, our synthesizer is ready to convert the words into phonemes. Synthesizers have to handle all the idiosyncratic pronunciations of English, such as silent letters (*k* in "knife," *t* in "often"), proper names ("Reagan" does not start like "real") and words like "permits" that can be pronounced in multiple ways. It is a rare sentence that does not contain some verbal anomalies.

THE AUTHORS

ANDY AARON, ELLEN EIDE and *JOHN F. PITRELLI* work on speech synthesis technology at the IBM Thomas J. Watson Research Center. Aaron studied physics at Bard College and subsequently developed sound for motion pictures at Zoetrope Studios, Lucasfilm's Skywalker Sound and elsewhere. Eide received her doctorate in electrical engineering and computer science from the Massachusetts Institute of Technology. Her research interests include statistical modeling, speech recognition and speech synthesis. Pitrelli also obtained his doctoral degree in electrical engineering and computer science from M.I.T. Speech synthesis, prosody, handwriting and speech recognition count among his research focuses. Collectively they have published more than 40 papers and hold 19 patents.

Our system applies rules to convert letters to phonemes, making use of the part-of-speech information when necessary. To get an idea of just how tricky this task can be, think about all the ways that "ough" can be turned into phonemes. Think of "bough" (OW), "cough" (AW F), "dough" (OH), "rough" (UH F) and "through" (OO).

After the software converts the words in the sample sentence into phonemes, it looks like this:

> P ER M I T S / K AW ST /
> AY T EE / D AA L ER Z /
> P ER / Y EE R

Selecting Sounds

DETERMINING WHICH phoneme example should be selected to synthesize each part of the sentence is challenging. Each sound in a sequence varies slightly, depending on the sounds that precede and follow it, a phenomenon called coarticulation. [*For more on coarticulation, see box on opposite page.*] The "permits" example contains 23 individual phonemes. Because each has about 10,000 original samples to choose from, we have a staggering $10{,}000^{23}$ (that is, 10^{92}) possible combinations—too many even to consider. The system therefore exploits a technique called dynamic programming to search the database efficiently and determine the best fit.

Once the synthesizer has assembled the best phoneme recordings in a row, all that remains is smoothing out the discontinuities that result from splicing sounds together. When adjacent samples deviate slightly in pitch, the sentence ends up with a jumpy, warbling sound. The system fixes this kind of problem by making small pitch adjustments to correct the mismatch, bending the pitch up or down at the edges of each sample to fit that of its neighbors, just as a carpenter sands glued joints to create a smooth surface transition.

Up for Discussion

ALTHOUGH WE ARE PLEASED with our progress on our synthesizer, we always keep an eye on how we can make improvements. Our team often debates

what some call the holy grail of text-to-speech technology: Should machine speech be indistinguishable from a human speaker, as in the well-known Turing test for artificial intelligence? Our conclusion: probably not. For one, people are not likely to feel comfortable with the notion that they might be being "tricked" when they dial in to a company's service center. In addition, natural human speech is not the best choice for some situations, such as warning signals for drivers or voices in toys, cartoons, and video and computer games. A better goal for the technology might be a pleasing, expressive voice to which people feel comfortable listening.

Or perhaps the ultimate aim should be a system sufficiently sophisticated to exploit humans' social and communication skills. Consider this example:

> *Caller:* I'd like a flight to Tokyo on Tuesday morning.
> *Computer:* I have two flights available on Tuesday *evening*.

The software's ability to emphasize the word "evening" to contrast it with "morning" would simplify the exchange enormously. The caller then understands that no flights are available in the morning and that the computer is offering an alternative. A completely unexpressive system could cause the caller to assume that the computer had misunderstood the inquiry, requiring it to be repeated. By the same token, if the response were, "I'm sorry, there are no flights available on Tuesday," engineers would want the voice to sound somewhat apologetic about the lack of flights or at least not as cheerful as the system's standard opening line, "How may I help you?"

Our team at IBM has recently developed new prototype systems that can deliver speech incorporating several such expressions. Beyond the basic, neutral expression, the technology can synthesize a sentence to sound cheerful, questioning or apologetic. The developing technologies can also emphasize individual words for effect.

Even though our speech synthesizer and its counterparts already sound astonishingly close to live human speech, truly expressive vocalization is the next big challenge facing this class of technology. After all, the software does not really comprehend what it is saying, so it may lack the nuanced changes in speaking style that one would expect from, say, an eighth grader, who can interpret what he or she is reading. Given the limitless range of the human voice, synthetic speech researchers certainly have their work cut out for them. SA

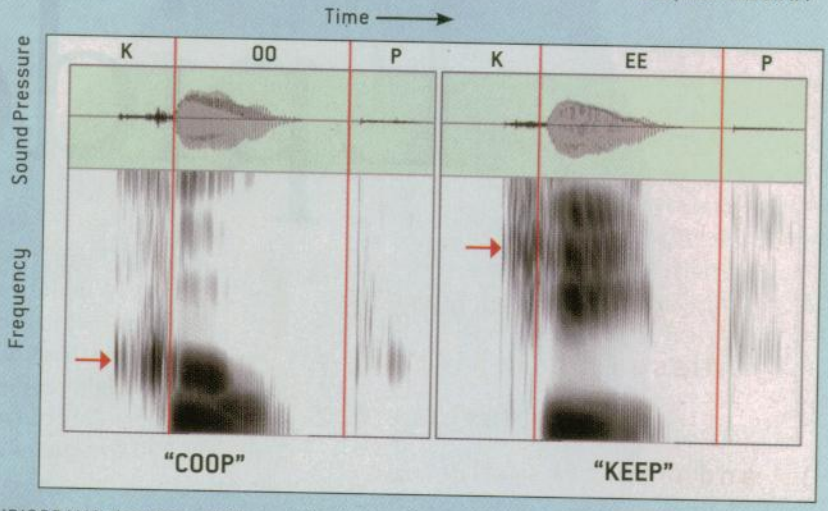
Variations in Phoneme Pronunciation

Machines cannot produce natural-sounding speech unless they become better at sequencing and mixing sounds together the way people do. This coarticulation occurs as the tongue and mouth begin getting into position to make the next sound before the first has ended. As a person pronounces the K sound in "keep," for example, the tongue is already moving forward, anticipating the EE, whereas in "coop" the tongue moves backward, anticipating the OO.

Coarticulation complicates matters for speech synthesizers. English speakers, for instance, hear K, P and T as three distinct sounds. But the K sound in "coop" and the K in "keep" are actually very different from each other; in fact, they are as dissimilar as either of them is from P or T. To hear the differences in the K's, ask someone to start to say "keep" or "coop" but say only the K sound; you will be able to tell which word was begun. People simply learn to hear the K's as equals and the P and T as different. In speech synthesis, engineers have to keep these distinctions straight—each K must be put in the proper place, because the wrong K will not sound right.

Context is not all that can affect phonemes. How phonemes are arranged in syllables and words matters, too. A classic example is "gray train" versus "great rain." Both have the same phonemic sequence, G R AY T R AY N, but we can easily hear the difference between the two. The T in "gray train" has a loud burst of the tongue off the front of the palate, as is typical for a T at the beginning of a word. But the T in "great rain," being at the end of a word, may be produced with no burst at all.

—*A.A., E.E. and J.F.P.*



AUDIOGRAMS show how a "K" sound (*red arrows*) can vary subtly because the mouth moves differently in anticipation of enunciating the following sound.

MORE TO EXPLORE

IBM text-to-speech research (includes a demonstration system): **www.research.ibm.com/tts**

Guidelines for evaluating text-to-speech systems: **www.speechtechmag.com/issues/6_3/cover/88-1.html**

History of text-to-speech systems: **www.cs.indiana.edu/rhythmsp/ASA/Contents.html**
Audio recordings appendix is at **www.cslu.ogi.edu/tts/research/history/**

Technical details on current speech synthesis systems: **http://tcts.fpms.ac.be/synthesis/introtts.html**

TTS Update. TMA Associates Web site: **www.ttsupdate.com/**