fects. In effect, the mission would be a topographic imager that would yield a water map of volumetric gain or loss after each overpass (*14*).

Such a satellite mission would enable hydrologists to move beyond the point-based gauging methods of the past century to measurements of the spatial variability inherent in surface water hydrology. Global coverage would ensure that, despite local economic and logistic problems, all countries could access measurements critical for forecasting floods and droughts, both of which have dramatic economic and human impacts.

References and Notes

1. A. I. Shiklomanov, R. B. Lammers, C. J. Vörösmarty, *EOS Trans. AGU* **83**, 13 (2002).
2. J. Roads *et al.*, *J. Geophys. Res.* **108** (D16), 8609, 10.1029/2002JD002583 (2003).
3. C. Prigent, E. Matthews, F. Aires, W. Rossow, *Geophys. Res. Lett.* **28**, 4631 (2001).
4. C.J. Vörösmarty, P. Green, J. Salisbury, R. B. Lammers, *Science* **289**, 284 (2000).
5. A. Ohmura, M. Wild, *Science* **298**, 1345 (2002).
6. L. C. Smith, *Hydrol. Process.* **11**, 1427 (1997).
7. An outstanding effort to measure inundated areas throughout the world has been constructed by the Global Rain Forest Mapping project using SAR images collected by the Japanese Earth Resources Satellite (JERS-1) (*15*).
8. C. M. Birkett, L. A. K. Mertes, T. Dunne, M. H. Costa, M. J. Jasinski, *J. Geophys. Res.* **107** (D20), 10.1029/2001JD000609 (2002).
9. D. E. Alsdorf *et al.*, *Nature* **404**, 174 (2000).
10. B. Tapley, personal communication.
11. J. Wahr, M. Molenaar, F. Bryan, *J. Geophys. Res.* **103**, 30205 (1998).
12. M. Rodell, J. S. Famiglietti, *Water Resour. Res.* **35**, 2705 (1999).
13. D. Alsdorf, D. Lettenmaier, C. Vörösmarty, the NASA Surface Water Working Group, *EOS Trans. AGU* **84**, 269 (2003).
14. An instrument that matches these requirements has been sketched by E. Rodriguez of NASA's Jet Propulsion Laboratory, based on the Shuttle Radar Topography Mission (SRTM).
15. A. Rosenqvist *et al.*, *Int. J. Remote Sens.* **23**, 1215 (2002).
16. We acknowledge the constructive comments of C. Birkett, T. Dunne, J. Famiglietti, J. Melack, L. Smith, and C. Vörösmarty. D. Lettenmaier is funded by NASA under NAG5-9454 and NAG5-9416. D. Alsdorf is funded by NASA under NAG5-12740, NAG5-13461, and NAG5-10685.

COMPUTER SCIENCE

The Power of Speech

Lawrence Rabiner

In the next few decades, advances in communications will radically change the way we live and work. The concept of "going to work" will change from commuting to a particular place to get things done, to "getting things done" no matter where you are. Life at home will also change radically as communications between individuals become multimodal (using voice, visual, and tactile modes) and multimedia (with sharing of text, data, audio, images, video, and other forms of information). For example, you will be able to control virtually any device in the home—such as the family home entertainment center—by pointing to it with your finger and issuing voice commands such as "find me a good classical music station."

The driving force for these changes is the seamless integration of real-time communications (voice, audio, video, virtual reality) and data (text, images, files) into a single network that can be accessed anywhere, anytime, and by a wide range of devices. Speech and language processing plays a crucial role in this network by enabling enhanced services and providing seamless access to new services (*1*).

Traditional speech and audio coding and compression will remain important even as bandwidth increases dramatically to the home, to the office, and in wireless environments. The need for high-quality, low-delay streaming of voice, CD-quality audio, and HDTV-quality video is a driving force for advanced coding research. Advanced coding and compression technologies enable networks to provide high signal quality at low delays without requiring excessive network resources.

Speech and language processing is also crucial for seamless user access to new and advanced services. As communication devices become ever smaller, the ability to provide and use keyboards and pointing devices (such as the mouse) becomes limited and problematic, and voice access to services becomes an essential component of the user interface. To access services on such devices, we will increasingly rely on speech recognition and speech understanding to command and control machines, and on speech synthesis to respond back to the user.

A third opportunity for speech processing is in user authentication. Speaker verification technology is a convenient and accurate method for authenticating the claimed identity of a user for access to secure or restricted services. It has the potential to be much more robust and reliable than conventional log-ons and passwords.

Finally, the opportunities for speech and language processing in services and operations are almost limitless. Voice commands may be used to access movie schedules or airline schedules or to add new people to a teleconference, whereas text-to-speech synthesis can be used to convert a text message to a voice message. At help desks or in customer care, voice processing can act as a surrogate for an attendant or an operator in handling routine transactions.

The speech dialog circle (see the figure) illustrates the speech-processing technology that enables voice conversations between humans and machines. Its major elements are speech recognition, spoken-language understanding, dialog management, and text-to-speech synthesis. In addition to these basic speech-processing technologies, two other key technologies, speech coding and speaker verification, are used in multimedia communications.

The author is in the Center for Advanced Information Processing, Rutgers University, Piscataway, NJ 08854, USA. E-mail: lrr@caip.rutgers.edu

Speech Coding

Speech coding has existed for more than 60 years, beginning with the classic work of Dudley on the "vocoder" (*2*). The original goal of speech coding was to provide a compression technology that would enable existing copper wires to handle the continual growth in voice traffic without having to continuously add new lines. Recently, the need for speech coding has grown because of the rapid growth in wireless systems and in the transmission of voice signals over data networks, where speech is just one (very important) data type.

The goal of speech coding (*3*) is to compress the speech signal—that is, to reduce the bit rate necessary to accurately represent the speech signal—without distorting it excessively. Two main techniques have been used in speech coding. Waveform coding tries to match waveform characteristics directly, whereas model-based coding tries to match spectral and source-excitation characteristics of speech.

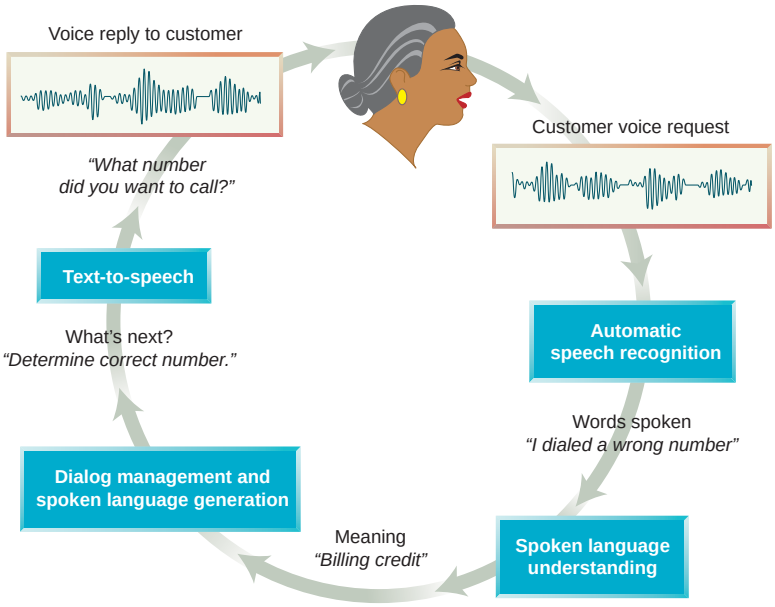
Today, speech can be coded down to bit rates of about 8000 bps, with intelligibility and quality approaching that of telephone-bandwidth speech (which has a bit rate of about 64,000 bps). The challenge for the next few years is to lower the bit rate by a factor of 2 without seriously lowering the quality of the resulting speech. Achieving this goal requires improved signal processing for accurately representing the excitation source and the short-time spectrum properties of the time-varying speech signal.

Text-to-Speech Synthesis

Text-to-speech synthesis aims to convert an ordinary text message into an intelligible, natural-sounding speech utterance, thus giving machines the ability to "speak" (*4*, *5*). Two approaches have been proposed

and studied. Concatenative methods are based on a small vocabulary of about a few thousand stored diphone (half-phoneme) units (a phoneme is a basic unit of speech, such as a vowel or a consonant). These units are pieced together to make up speech; each unit is artificially adapted to match the desired inflection, timbre, and duration in the sentence. In contrast, unit selection synthesis is based on storage of a vast set of diphone units that reflect the range of their variability in the speech of a single speaker.

The quality of unit selection synthesis is rapidly approaching that of natural speech. It far surpasses simple concatenative methods, but at a huge cost in computation and storage. The challenge is to achieve high-quality synthetic speech with a minimal set of stored units, and doing so for any language, any speaker, and any accent. To achieve this goal, advanced data-reduction methods are required to represent the variability in speech without having to use many samples of each spoken diphone.



The speech dialog circle. A sustainable conversation with a machine can be created by following the speech circle.

Automatic Speech Recognition

The goal of automatic speech recognition is to accurately and efficiently convert a speech signal into a text message independent of the speaker or the speaking environment (*6*, *7*). Two main approaches have been studied for speech recognition. In the acoustic segmentation and labeling approach, spoken sentences are first divided into segments, which are then matched to appropriate phonemes. In hidden Markov models, statistical models of each of the sounds of language are derived from an extensive training set of speech units; a matching procedure is used to align a spoken sentence with a set of model speech sounds. The speech recognition achieved with the latter far exceeds that of the former, with word accuracies exceeding 95% for a range of vocabulary sizes and application scenarios.

The challenge is making such speech-recognition systems robust to the acoustic environment (noise, reverberation) so that the performance does not degrade significantly in automobiles or other environments where cellular phones are typically used. A range of signal-processing methods for speech enhancement, noise removal, speaker normalization, and feature normalization have been proposed to solve the problems associated with noisy and reverberant environments.

Spoken-Language Understanding

The goal of spoken-language understanding systems (*8*, *9*) is to interpret the meaning of key words and phrases in a speech string and map them to actions that the system should take. The systems exploit a task grammar and task semantics to restrict the range of meaning associated with the recognized word string, and use high-information word sequences (such as "collect call" in the phrase "I would like to make a collect call to X because I lost my credit card") to determine the appropriate meaning of the sentence.

Current spoken-language understanding systems are crude and can only handle sentences with a small number of possible action outcomes. The system therefore only has to decide which of the finite number of outcomes is most likely. The challenge is to extract meaning from spoken inputs without restricting the range of meanings to one of a small set of possibilities. Artificial intelligence methods that automatically find the phrases in a sentence that determine the desired outcome, and that continue to "learn" new words and phrases used to signify the same set of actions, will be key tools in advancing the technology in this area.

Speaker Verification

Speaker verification systems (*10*) aim to verify a claim of user identity based on voice input. To achieve this goal, speech features of the input utterance are analyzed and compared with those of the claimed speaker. A statistical model is used to determine the likelihood of a match, from which a decision is made to accept or reject the claimed identity.

The performance of simple voice password systems is quite high, with error rates of about 0.5% for spoken digit strings recorded in quiet laboratory conditions. The challenge is to verify a person reliably in real-world environments and with minimal training of the system. Signal-processing algorithms similar to those used for speech recognition are being used to meet this challenge.

Outlook

We carry our voice with us wherever we go, and our ability to convey and receive information via voice commands is virtually unlimited. In the not-so-distant future, we will routinely communicate with machines by simple voice commands. Bandwidth will become seemingly unlimited, both via fiber optics and through wireless connections, enabling instant access—much of it by voice—to information and entertainment of all types.

The key challenge will be to make the user interface between humans and machines as easy to learn and use for advanced services as voice telephony is today. Meeting this challenge will require methods ranging from advanced statistical models to machine learning and adaptation. The power of speech will be manifest in a panoply of advanced services that are available at the sound of your voice.

References

1. R. V. Cox, C. A. Kamm, L. R. Rabiner, J. H. Schroeter, J. G. Wilpon, *Proc. IEEE* **88**, 1314 (2000).
2. R. V. Cox, B. G. Haskell, Y. LeCun, B. Shahraray, L. R. Rabiner, *Proc. IEEE* **86**, 755 (1998).
3. W. B. Kleijn, K. K. Paliwal, Eds., *Speech Coding and Synthesis* (Elsevier, Amsterdam, 1995).
4. A. Hunt, A. Black, *Proc. ICASSP'96*, 373 (1996).
5. M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, A. Syrdal, *J. Acoust. Soc. Am.* **105**, 1030 (1999).
6. L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ, 1993).
7. J. Makhoul, J. Schwartz, in *Voice Communication Between Humans and Machines*, D. Roe, J. Wilpon, Eds. (National Academy Press, Washington, DC, 1994), pp. 165–188.
8. A. L. Gorin, G. Riccardi, J. H. Wright, *Speech Commun.* **23**, 113 (1997).
9. A. L. Gorin, *J. Acoust. Soc. Am.* **97**, 3441 (1995).
10. A. E. Rosenberg, S. Parthasarathy, *Proc. Eurospeech* **97**, 1371 (1997).