By Li Deng and Xuedong Huang

Although progress has been impressive, there are still several hurdles that speech recognition technology must clear before ubiquitous adoption can be realized. R&D in spontaneous and free-flowing speech style is critical to its success.

Challenges in Adopting Speech Recognition

In recent years, significant progress has been made in advancing speech recognition technology, making speech an effective modality in both telephony and multimodal human-machine interaction. Speech recognition systems have been built and deployed for numerous applications. The technology is not only improving at a steady pace, but is also becoming increasingly usable and useful. However, speech recognition technology has not been widely accepted in our society. The current use of speech recognition by enterprises and consumers reflects only a tip of the iceberg of the full power that the technology could potentially offer [3]. To realize such a potential, the industry has yet to bridge the gap between what people want from speech recognition, typically in a multimodal environment, and what the technology can deliver. To make the mainstream use of speech recognition a reality, the industry must deliver robust and high-recognition accuracy close to human-like

performance. To this end, the industry and research communities must collectively address and overcome both technical and business challenges to uncovering the full potential of speech technology in multimodal and intelligent human-machine communication.

Speech recognition technology has made dramatic progress over the past 30 years, triggered by advances in computing devices and architectures, success of algorithm development, and availability of large quantities of data. Despite this progress, some fundamental and practical limitations in the technology have hindered its widespread use. There has been a very large performance gap between human and machine speech recognition. Mainstream adoption of speech recognition would not be possible without the underlying recognition technology that can deliver a sufficiently robust and low-error performance. Reducing the speech recognizer's error rate under all deployment environments

THE industry and research communities must collectively address and overcome both technical and business challenges to uncovering the full potential of speech technology in multimodal and intelligent human-machine communication.

remains one of the greatest challenges to making speech recognition more mainstream.

Many leading researchers in the field understand the fragile nature of the current speech recognition system design. How to overcome this technical challenge has formed a number of core issues for adopting speech recognition technology. These issues include intrinsic error rates of speech recognition engines and how they can be reduced by using multimodal and multisensor system design; user interface designs (including error correction) that improve the user experience given the error rate; application designs that constrain the user input space by multimodal interactions; and reduction of the tremendous efforts required to port the speech technology applications from one application domain to another and from one language to another. Among these issues, recognition error rates, which reflect the fundamental capabilities of the recognition technology, override all other issues. For example, after machine recognition errors drop to a significantly low level, both the user interface and application designs would be easily relaxed to provide the required flexibility and most constraints would be virtually eliminated.

Historical progress on machine speech recognition performance. The goal of reducing the machine

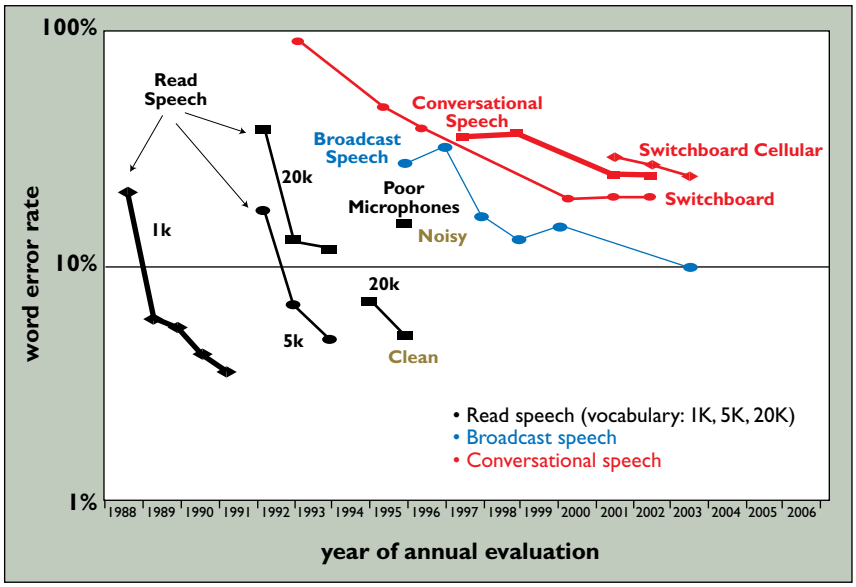

Figure 1. Historical progress on word error rates for speaker-independent speech recognition for increasingly difficult speech data (simplified and redrawn from [1, 2])

speech recognition error rate has been rigorously pursued by researchers for many years. One major driving force for this effort has been the annual speech-to-text evaluation program sponsored by DARPA. Figure 1 shows progressive word error rate reduction achieved by increasingly better speaker-independent systems from 1988 to 2003 [1, 2]. Increasingly difficult speech data was used for the evaluation, often after the error rate for the preceding easier data had been dropped to a satisfactorily low level. Figure 1 illustrates that on average, a relative speech recognition error-reduction rate of about 10% annually has been maintained through most of these years.

Two other noticeable and significant trends can be identified from Figure 1. First, dramatic performance differences exist for noisy (due to acoustic environment distortion) and clean speech data in an otherwise identical task (see the 1995 evaluation results for read-speech in Figure 1). Such differences have also been observed by nearly all speech recognizers used in industrial laboratories, and intensive research is under way to reduce the differences. Second, speech of a conversational and casual style incurs much higher errors than any other types of speech. The difficulties arising from acoustic environment distortion and from the casual nature in conversational speech form two principal technical challenges for the current speech recognition technology.

Overcoming the challenge of making speech recognition systems robust in noisy acoustic environments. One fundamental challenge facing the development of speech recognition technology is to make the systems robust in noisy acoustic environments. Speech recognition systems often work well in quiet settings, but work poorly under noisy conditions. For example, the error rate of a system may be acceptable if one calls from a phone in a quiet office, yet the error rate can be unacceptably high when using a cellular phone in an automobile or an airport lobby. The technical challenge is how to handle the often co-occurring additive noise including interfering speakers (arising from background) and convolutive distortion (arising from a less expensive microphone or other data acquisition devices). The noise robustness problem is especially serious in changing acoustic environments.

Over the past several years, noise-robust speech recognition has been one of the most popular problems addressed by both academic and industrial researchers of speech recognition. The research has recently been placed within a common evaluation framework called Aurora [5, 7]. Much research has been focused on characterizing and estimating the frequently changing acoustic conditions for speech recognizers, and on identifying and compensating for major sources of recognition performance degradation. In general, noise-robust techniques can be classified into two categories: feature-space compensation and model-space compensation [4, 8]. Both have been pursued by diverse groups of researchers. Further, strides have been made in recent years within the multimodal community on major speech recognition robustness gains achieved by integrating visual and acoustic speech information [9].

Overcoming the challenge of creating workable recognition systems for natural, free-style speech. The ultimate technical challenge for speech recognition is to make it indistinguishable from the human's speech perception system. At present, when users interact with any existing speech recognition system, they must be fully aware of the fact their conversation "partner" is a machine. The machine would easily break if the users were to speak in a casual and natural style as if they were talking with a friend. In order to enable mainstream use of speech recognition, naturalness or the free style of speaking on the user's part should not produce so many recognition errors.

As was shown in Figure 1, the annual evaluation of speech recognition technology sponsored by DARPA has in recent years shifted to natural, conversational-style speech. Much of the speech recognition research in Japan has also recently shifted to natural speech [6].

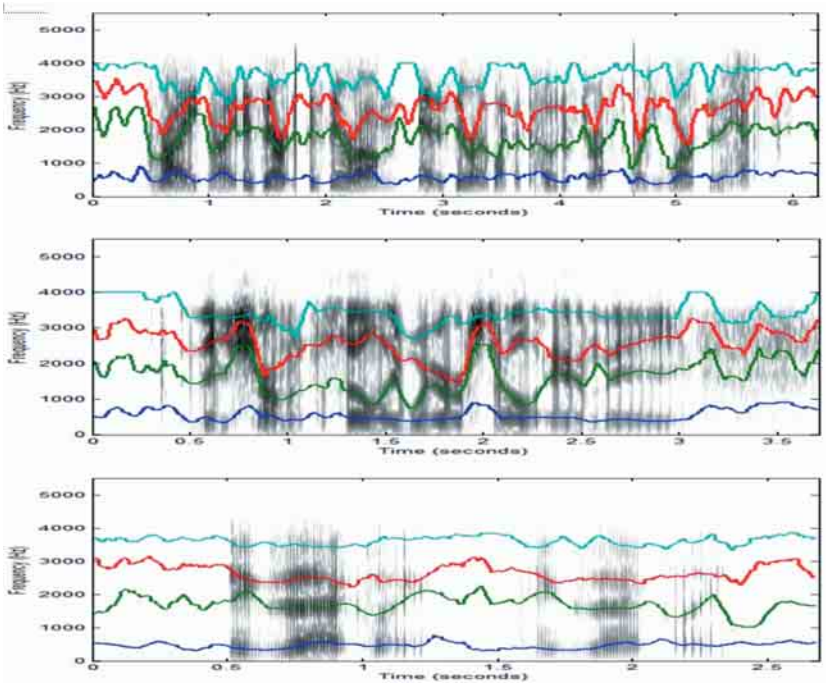

Figure 2. Example trajectories for the lowest four vocal tract resonance frequencies estimated using an automatic learning algorithm on speech waveforms and are superimposed on spectrograms. The examples are from three typical utterances from the Switchboard database, which consists of conversational telephone speech.

In fact, the DARPA speech program called EARS (Effective, Affordable, and Reusable Speech-to-text) established in May 2002 [2] has a specific focus on Conversational Telephone Speech (CTS). The expectation for EARS over the next five years is to reduce the recognition error rate on CTS by over 25% each year so the current error rate of approximately 30–40% will be reduced to the 5%–10% range [1, 2].

Here, we briefly describe the two research programs within the "Novel-Approaches" part of the DARPA's EARS. One program is built on principles of speech and language science and on computational models for some essential aspects of the human verbal communication process responsible for the generation of naturally uttered human speech signals with the unconstrained style. Based on the general principles of computational phonetics and phonology [4], quantitative models have been established in statistical terms, so that advanced algorithms can be developed to automatically and optimally determine the parameters, which are all physically meaningful, in the models from a representative set of training data. The speech recognizer architecture designed in this approach is different from that based on the conventional Hidden Markov Models (HMMs) [8]. Some detailed stages in human speech production, from the distinctive feature-based linguistic units to speech acoustics, which is embedded in the architecture of a novel speech recognizer, are represented explicitly in the table here.

Examples of the automatically estimated vocal tract resonance trajectories (output of Stage III in the table) in typical utterances of the Switchboard database are shown in Figure 2. The main advantage of representing the detailed structure in the human speech process is that a highly compact set of parameters can now be used to capture phonetic context and speaking style variations (including both hyper-articulated and reduced speech) in a common framework. Using this framework, many important subjects in speech science and those in speech recognition previously studied separately by different communities of researchers can now be investigated in a unified fashion.

	Stage I	Stage II	Stage III	Stage IV
Input	Distinctive-feature representation of an utterance	Discrete articulatory state sequence	Segmental target sequence	Articulatory or vocal tract resonance vector
Mediating process	Temporal overlapping mechanism	Symbolic-to- numerical mapping	Explicit or recursive trajectory modeling	Static, numerical-to-numerical, nonlinear mapping
Output	Overlapping articulatory gestures, represented by a set of discrete meta-states	Segmental target sequence, represented as a left-to-right, constrained switching random process	Continuous, smooth, and target-directed trajectories for articulatory or vocal tract resonance variables	Acoustic or auditory feature vector computable or measurable directly from speech waveforms
Domain	Phonology	Interface between phonology and phonetics	Phonetics	Phonetics
Properties	Account for partial or full sound deletion or modification for the pronunciation variation in casual speech; Also account for contextual variation at the pronunciation level	Account for compensatory articulation, or different ways of activating articulators to achieve similar acoustic effects or auditory perception; Targets are used as the control signal directing the dynamic system governing speech articulation	Account for variability of speech due to reduced speaking effort or increased speaking rate (phonetic reduction) and due to increased effort (e.g., Lombard effect); Also account for coarticulation at the physical level due to inertia in articulation	Account for differences in different speakers' speech production organs and the distorting effects due to acoustic environments

Four major stages in the architectural design of a novel speech recognizer.

The other "Novel-Approaches" research effort in the EARS program is based on aspects of human speech perception, which counter the conventional use of the acoustic information of speech contained in frame-by-frame spectral envelopes in all the current major speech recognition systems [1, 2]. The goal of this approach is

to develop new acoustic processing front-ends that escape dependence on the spectral envelopes, and use-multiple front-ends across the entire time and frequency plane.

Other directions for making speech recognition usable. Here, we discuss three additional directions of reducing speech recognition error rates to facilitate the technology's adoption.

There can be a substantial robustness gained by improving the existing *microphone ergonomics.* Use of microphone arrays can improve the signal-to-noise ratio by up to 12dB, significantly reducing the technical challenge in the remaining processing stages of noise reduction or adaptation. Also, use of multiple types of sensors embedded in one or more microphones to detect speech-related sensory signals can provide valuable information to the recognizer to improve a user's experience and to reduce recognition errors.

To equip speech recognizers with the *ability to learn and to correct errors* is another promising direction. Currently, only very limited capabilities are provided. The error correction process in most speech recognizers today is cumbersome and time consuming. This drastically reduces the users' incentive to adopt speech recognition, especially when the same errors are committed that are not appropriately corrected the second time around. Advanced techniques that could automatically learn and correct errors will prove highly desirable. Two main types of errors subject to correction are new words outside the recognizer's lexicon; and errors due to user's pronunciations that are different from those in the lexicon.

BUSINESS benefits derived from adopting speech recognition will be significant in the coming years, and the current technology has already made a positive impact on businesses and consumers.

Finally, *adding semantic knowledge* (that is, meaning) and *pragmatic knowledge* (that is, application context) into speech recognizers will eliminate a key factor responsible for recognition errors—lack of common sense understanding of what is being said. One promising approach is to exploit statistical parsers that automatically learn semantic knowledge from examples.

Business Challenges

Although speech recognition technology still has a long way to go to reach the human-level performance, it is already capable of delivering sufficiently robust and low-error performance for many practical applications. Business benefits derived from adopting

speech recognition will be significant in the coming years, and the current technology has already made a positive impact on businesses and consumers. For example, telecommunications carriers already use speech recognition for directory assistance. Speech recognition is employed in many luxury cars to control useful functions. And PC users can dictate directly into Microsoft Word XP.

Speech technology can be used by enterprise customers to reduce costs, enhance revenues, and improve business agility. The rich experience offered by the technology can also increase customer satisfaction. Importantly, with multiple locations and mobile employees, companies are demanding anytime, anywhere access to information. Enterprises can improve employee productivity and agility via a unified Web and voice portal that enables employees to access information not only through a PC, but a wireless personal digital assistant or telephone. This will be one of the key factors driving the growth of the speech industry, since speech is the only modality that can provide a consistent user interface across all devices.

The speech recognition-based user interface paradigm, as with other existing pervasive paradigms such as the Web, requires key infrastructure enablers in order to reach the status of mainstream adoption. One important aspect of such adoption is speech-enabled Web services; that is, incorporation of speech functionality into Web applications where a large population will effectively and conveniently use speech recognition to enjoy a diverse range of applications built with Web services. A key enabler of making speech mainstream is the establishment of open standards supported by the industry. The importance of establishing standards for the speech paradigm is similar to that of HTML for the Web paradigm. Since the general Web services inevitably touch on a wide range of software, hardware, client/server, PDA, and telephone services, a standard way must be adopted to add speech to Web services and applications. In this way, duplication of development work performed for different environments can be effectively avoided and the investment in deploying speech recognition can be preserved. One recently established standard is Speech Application Language Tags (SALT; www.saltform.org), which extends existing

A key enabler of making speech mainstream is the establishment of open standards supported by the industry. The importance of establishing standards for the speech paradigm is similar to that of HTML for the Web paradigm.

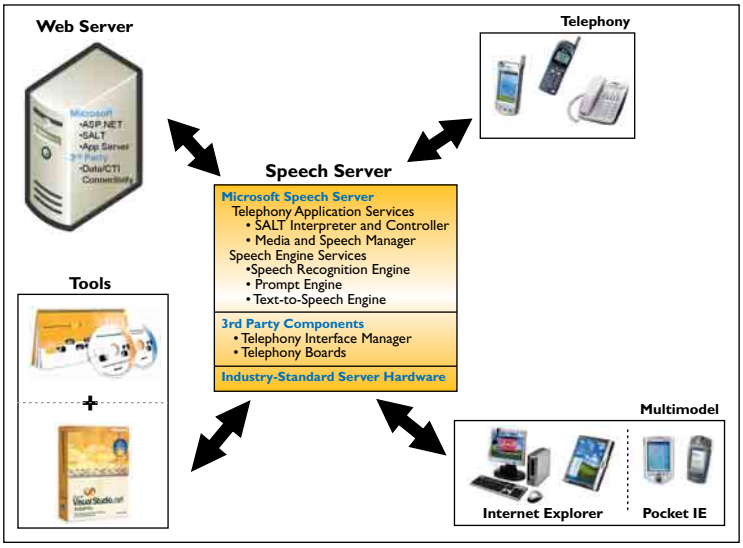

Figure 3. Microsoft speech platform reference architecture.

Web markup languages to enable multimodal (speech plus other modalities) and telephony (speech only) access to the Web. It is thus superior to the earlier standard VoiceXML—a different programming language supporting only telephony interaction for call center speech applications. The recently developed Microsoft speech platform reference architecture based on SALT for both telephony and multimodal applications, shown in Figure 3, leverages the Web server and a set of tools facilitating the use of the platform.

The availability of mobile devices capable of browsing the Web will demand SALT-based services that fully support multimodal interactions. With the growth of the Web, people now interact not only via phones but also by email and the Web. The rise of Web applications, however, will not reduce the phone-based interaction. For most enterprises, the call center remains one of the premier support mechanisms for customer connection. This trend of supporting both phone and multimodal interactions will continue, and people will rely more on mobile devices to access information with the evolution of wireless networks to 2.5G and 3G. Rich user experiences with these mobile devices would be hampered by lack of better input capabilities until speech recognition with desirable performance can be enabled.

Conclusion

Speech recognition systems have been deployed for a wide range of telephony applications, yet a number of fundamental and practical limitations in speech recognition technology have been encountered that hinder ubiquitous adoption of the technology. We must continue to invest in research and development to bridge the recognition performance gap between human and machine, and, in particular, to invest in novel approaches with the potential to create breakthrough progress. Dealing with conversational speech in a natural, free style and making systems robust in all possible practical acoustic environments are two of the most critical technical challenges. Nevertheless, achieving workable solutions or even solutions close to human-level speech recognition in certain highly constrained tasks is on the horizon with carefully designed, limited-domain applications taking advantage of multimodal interactions (for example, [5, 10]).

Further improvement on system robustness can be achieved by using additional sensory information ([11] and see Turk's article in this section). We expect more speech recognition systems will be deployed in the future, especially with the use of multimodality, and that mainstream adoption of speech recognition will become a reality as both the technical and business challenges discussed here are being successfully addressed. C

References

1. DARPA's EARS Conference (Boston, MA, May 21–22, 2003).
2. DARPA's EARS Kickoff Meeting (Vienna, VA, May 9–10, 2002).
3. Datamonitor. *Voice Automation—Past, Present, and Future*. White Paper (July 2003).
4. Deng, L., and O'Shaughnessy, D. *Speech Processing—A Dynamic and Optimization-Oriented Approach*. Marcel Dekker, NY, 2003.
5. Deng, L. Wang, K., Acero, A., Hon, H., Droppo, J., Boulis, C., Wang, Y., Jacoby, D., Mahajan, M., Chelba, C., and Huang, X.D. Distributed speech processing in MiPad's multimodal user interface. *IEEE Transactions on Speech and Audio 10* (2002), 605–619.
6. Furui, S. Recent progress in spontaneous speech recognition and understanding. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing* (Dec. 2002).
7. Hirsch, H., and Pearce, D. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. *ISCA ITRW Workshop on Automatic Speech Recognition* (Paris, 2000).
8. Huang, X.D., Acero, A., and Hon, H. *Spoken Language Processing—A Guide to Theory, Algorithms, and System Development.* Prentice Hall, NY, 2001.
9. Neti, C., Iyengar, G., Potamianos, G., Senior, A., and Maison, B. Perceptual interfaces for information interaction: Joint processing of audio and visual information for human-computer interaction. In the *ICSLP Proceedings 1*. (Beijing, 2000), 11–14.
10. Oviatt, S. Breaking the robustness barrier: Recent progress on the design of robust multimodal systems. *Advances in Computers.* M. Zelkowitz, Ed. Academic Press, 2002, 305–341.
11. Zhang, Y. et al. Air- and bone-conductive integrated microphones for robust speech detection and enhancement. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*. (St. Thomas, U.S. Virgin Islands, Dec, 2003.)

Li Deng (deng@microsoft.com) is a senior researcher at Microsoft Research in Redmond, WA.

Xuedong Huang (xdh@microsoft.com) is the general manager of Microsoft .NET Speech Technologies Group in Redmond, WA.