

Survey of Current Speech Technology

ALEXANDER I. RUDNICKY ■ ALEXANDER G. HAUPTMANN ■ KAI-FU LEE

Speech recognition and speech synthesis are technologies of particular interest for their support of direct communication between humans and computers through a communications mode humans commonly use among themselves and at which they are highly skilled. Both manipulate speech in terms of its information content; recognition transforms human speech into text to be used literally (e.g., for dictation) or interpreted as commands to control applications, and synthesis allows the generation of spoken utterances from text.

Automatic speech recognition (ASR) has gained a significant amount of commercial success due to its demonstrable increase in productivity by greatly assisting human operators or replacing the human element altogether. Several major areas of commercial application of ASR are dictation, personal computer interfaces, automated telephone services, and special purpose industrial applications.

Large vocabulary dictation: "unrestricted" dictation (e.g., business letters or newspaper articles) and structured report generation (e.g., radiology reports or insurance claims) can be speaker dependent or adaptive on the assumption that a single user will interact with the system for an extended period of time. Discrete-word systems are prevalent. Current vocabulary capacities range up to 40,000 words. These systems are meant to be used under benign conditions such as in offices, with head-mounted noise-cancelling microphones. Vendors of dictation technology include: Dragon (Dragon-

Dictate), IBM (Speech Server), and Kurzweil (Kurzweil Voice).

In order to add constraint to an otherwise unbounded problem, unrestricted dictation systems use statistical language models to favor more frequent words and word sequences [4]. Domain-specific systems can further enhance performance by incorporating a structured dialog to generate a full report, although this requires extensive application-specific engineering.

The usability of a dictation system is increased by its ability to adapt to an individual's voice (speaker adaptation), vocabulary (new word learning), and tasks (language model adaptation).

For interfacing with personal computers, the "electronic desktop" market has a number of products available. Computer manufacturers are nevertheless proceeding on the assumption that speech will become an important component of the computer interface. Some of the nearer-term opportunities include:

- Speech as a shortcut. Rather than opening a file by traversing many levels of hierarchy, a user just says "OPEN BUDGET."
- Hands busy/eyes busy. Change the font style while a user is typing, or change the drawing tool while the user is drawing.
- Information retrieval. Graphical user interfaces are awkward for specifying constraint-based retrieval ("find all documents from John received after March").
- Portable applications. As computers shrink in size from desktop to notebook and subnotebooks, keyboards will be more difficult to use or even nonexistent, thereby making speech a competitive alternative.

The main challenge for desktop speech recognition is the current existence of a mature and efficient alternative—the keyboard (and pointer). It is unlikely speech could completely replace these devices. Rather, the future interface will likely combine all of these and will allow the user to pick the input mode or combination of

modes that is ideal for a particular task. Other challenges to the usability of speech in the PC interface include efficient error management, appropriate user feedback, as well as integration into the computing environment.

In the future, speech recognition might be combined with natural language processing in more ambitious tasks, such as information retrieval or delegation interfaces. The proper use of speech in the personal computer will probably require the development of a new interface paradigm, rather than augmentation of the existing graphical user interface.

Telephone-based recognition offers a huge potential, due to the pervasiveness of voice communication and the lack of alternative methods for high-bandwidth input. It is also by far the technically most difficult area of recognition due to the lack of control over the conditions of use. Problems include a large and unpredictable user population, differences in handset microphones, the presence of channel noise and low signal bandwidth. The most successful current systems are ones that limit themselves to very small vocabularies, on the order of 10 to 20 words. Significant functionality does not require large vocabularies; some currently deployed systems provide a vocabulary size of two words ("yes" and "no"). In addition to poor control over signal quality, telephone recognition presents problems due to users' expectation that telephone recognition systems behave much like a human listener. For example, they will talk over system prompts (known as "barge in"), or they will embed their responses in other, out-of-vocabulary speech ("yes, please"). Word-spotting techniques can be used to achieve acceptable recognition accuracies [15].

Operator services involve small vocabularies, interactive dialog, and prompting. Opportunities for extension include: credit card validation, catalog shopping and automated answering machines. These extensions, however, require progressively larger vocabularies and interface intelligence.

While these various domains have the greatest potential for widespread

use, more specialized domains exist and can support commercially viable applications. These latter domains are highly structured and are amenable to a number of constraint-based techniques that increase the success of voice input. Speaker-dependent recognition is sufficient for many applications, since a particular device will be used by a single individual for an extended period such as a work shift. On the other hand, connected speech capabilities are necessary in many applications, since discrete entry would be unacceptably slow. Financial trading and parcel sorting are examples of the latter.

Speech recognition devices (e.g., Verbex) are used as part of simulators, allowing an automatic system to replace a human trainer. ASR has also been included as part of mobile inspection and inventory control systems (e.g., Vocollect) and has been adapted for process control in eyes-busy (or no-eyes) situations such as microscopy and photographic dark-room work.

Further down the line of specialized applications are devices such as recognition-based dialing for car phones (e.g., Motorola). In the future we expect to see more ASR in cars and other eyes-busy environments (such as air traffic control towers), controlling nonessential functions such as radio tuning, information displays, and so forth. It is unlikely that ASR will be used to control critical functions, since automatic systems will make recognition errors and by themselves will not be capable of assessing the correctness of an action.

Current Research in ASR

Current speech recognition research activities are strongest in industry, but with significant academic presence. The main sponsor of speech recognition research in the U.S. has been ARPA through its Spoken Language Technology program [2]. This program has concentrated on two domains, continuous speech recognition based on *Wall Street Journal* text and the Air Travel Information Service (ATIS) spoken language task. The former is meant to extend basic recognition technology along dimensions such as increased vocabulary size and operation in acoustically dif-

icult environments. The ATIS task promotes the integration of speech and natural language processing and focuses on dialog and interface issues. The ARPA program features an annual competitive evaluation of systems on a common test corpus, which has proven to be highly useful in promoting rapid improvements in algorithms. This annual competition has begun to attract international participation. The ARPA program stresses research in robust and accurate front ends, detailed acoustic modeling, and statistical modeling of a language. It has also encouraged the development of natural language processing techniques that address the interpretation of spontaneous speech and focus attention on speech interface design. Neural network-based recognition research is also active, with current emphasis being placed on transitioning the technology to the larger domains studied by the hidden Markov model (HMM) community. Success has been achieved by systems that combine both HMM and neural network technology, using the latter for signal classification and the former for modeling speech over time. For the time being, these two research communities have been functioning more or less independently.

Industrial research outside of the ARPA community has concentrated on different problems. For example, AT&T Bell Labs continues to do significant work on telephone bandwidth recognition, while IBM continues to do work on large vocabulary dictation systems. Industrial research has concentrated on the development of proprietary technology, and consequently its impact on the field as a whole has been indirect.

Table I provides a guide to the increases in recognition capabilities over the past two decades. As a rule of thumb, technology has been thought to be ready for commercialization once the word error rate has been reduced below about 5%. In practice, this threshold will vary as a function of the demands of a particular application. Note also that while research system evaluation often focuses on word error, in practice other metrics may be more significant, such as utterance error (whether an entire sentence was correctly transcribed) or

Synopsis of Speech Recognition

Recognition systems match a transform of incoming speech against a stored representation in a process often described as *decoding* [14]. A recognizer will make use of acoustic models that capture phonetic or word-level properties of speech and often a statistical model that captures the syntactic and semantic regularities of language in a particular domain [3, 4, 6, 16].

All speech systems use some form of spectral representation. Historically, other schemes (e.g., time-domain) have not fared as well. A major difference between current systems is whether they are based on spectral templates or hidden Markov models (HMMs). The latter representation is found in state-of-the-art systems. The former is present in older products with limited capabilities. Figure 1 shows the structure of a typical HMM system.

Recognition systems are classified along a number of standard dimensions. The capabilities of a particular system strongly depend on where a system falls in this space.

- *Speaker dependent (SD) vs. independent (SI)*: whether the system recognizes speech of many individuals without training or has to be trained for a particular voice. Speaker-adaptive systems also exist, which initially function as speaker independent but become tuned to the speech of individual users (with a concomitant increase in recognition accuracy).
- *Discrete (IWR) vs. continuous (CSR)*: whether the user needs to separate individual words by short silences. Isolated-word recognition systems are easier to implement since the system knows the exact extent of each word (and can use this information to improve decoding accuracy).
- *Vocabulary size*: typically task dependent, although this has implications for the choice of recognition algorithm and details of implementation. Other things being equal, small vocabularies are easier to recognize, though the actual difficulty of recognition is better indicated by the perplexity of a task (roughly corresponding to the number of word alternatives the system must consider at any one point in the decoding process).

Automatic speech recognition requires the use of an analog-to-digital (A/D) converter with the remaining computation taking place on a general-purpose computer. Sometimes signal processing is offloaded to a separate processor. Some large systems (e.g., the IBM Speech Server) also include special-purpose processors for search. Recent increases in computer processing power together with advances in algorithm design have allowed software-only recognition on workstation and PC-class computers. BBN's HARK product is an example. Since A/D capabilities are becoming a standard computer feature, the need for custom processing components for speech may be eliminated.

semantic error (whether the user's intent was identified and the correct action taken).

The availability of speech recognition technology has encouraged the

development of prototypes that merge speech capabilities into a general-purpose computer interface [12, 13]. Research on the usefulness of speech recognition for various activi-

ties is also under way. For the most part, work in this area has made use of commercially available devices and has not been able to explore the capabilities of more advanced recognition technology [9].

Synthesis

Speech synthesis offers an output channel in cases where visual displays are either not possible, insufficient, or awkward (such as over the telephone or while driving). The goal for most speech synthesis systems is to produce speech of useful intelligibility; the ultimate goal is to produce speech from any text that is at least as natural and intelligible as human speech [5]. For practical reasons, speech must be produced in real time and from arbitrary text. The reality of commercial and research synthesis systems still falls short of its ultimate goal. While intelligibility is quite high with error rates as low as 3.25% for individual word perception (compared to 0.53% on natural speech) on a standardized test, existing synthesizers still do not sound natural [7]. Text-to-speech systems are noticeably lacking in the quality of intonational characteristics or "prosody." Prosody encompasses the timing, intensity and pitch, as well as some of the coarticulation effects that change the way words sound when spoken together instead of individually.

The low error rate cited here can be misleading. The error rate for the perception of complete sentences is again over five times as large as that of natural human speech (4.7% synthesized vs. 0.8% for natural speech) [11]. Furthermore, these intelligibility studies were conducted in a controlled laboratory setting. Other experiments have shown that the

Table 1. Progress in speech recognition, as expressed by word error rate. Refer to sidebar for an explanation of the acronyms.

TASK	Late-1970s	Mid-1980s	Early-1990s
SI IWR Alphabet	30%	10%	4%
SI CSR Digits	10%	6%	0.4%
SD CSR Query, 1,000 word (perplexity 4)	2%	0.1%	
SI CSR Query, 1,000 word (perplexity 60)	—	60%	3%
SD IWR Dictation, 5,000 word	—	10%	2%
SI CSR Dictation, 5,000 word	—	—	5%
SI CSR Dictation, 20,000 word	—	—	13%

amount of cognitive processing required for synthesized speech is greater and response latencies longer. In general, synthesized speech constitutes an impoverished acoustic signal, making it even more difficult to understand under adverse circumstances. Progress in synthesizers appears to have reached an asymptote; successive small improvements are seemingly difficult to achieve.

Probably the most common use for text-to-speech is in the voicing of arbitrary text. This is also the most difficult application, since text of arbitrary complexity and of any origin may need to be synthesized. An example of this is reading machines for the blind (e.g., Kurzweil). Text-to-speech synthesis is also embedded in a number of electronic gadgets, such as telephones, answering machines, dictionaries, and pocket computers. Text-to-speech synthesis is used as an extension of personal computer and workstation interfaces and is used as an alternate channel of communication with the user. Finally, text-to-speech provides telephone access to database information (such as account information, automated customer name and address synthesis, and so forth). As the applications become more specialized, the full text-to-speech synthesizer is frequently replaced by prerecorded phrases. The advantages in naturalness and intelligibility usually outweigh the large storage requirements for limited amounts of speech data.

Examples of Commercial Systems

The industry standard in synthesis is the DECTalk system [5]. Developed from the earlier MITalk system, DECTalk has consistently outranked other synthesizers for general text synthesis in published evaluations.

Berkeley Speech Technologies Inc. (BST), started with a formant synthesizer similar to DECTalk and has evolved its product into a synthesizer/generator for many languages. Similar to DECTalk, the BST system divides synthesis into general-purpose algorithms and language-specific rules. By creating separate sets of rules, synthesizers have been configured with relative ease for En-

Synthesis Technology

All text-to-speech systems initially transform text input into a sequence of sound symbols, usually either phonemes, diphones, or demi-syllables [1]. The complexity of English requires between 500 and 1,000 mapping rules to derive most pronunciations. Some (limited) success has been achieved in using neural networks to model English pronunciations based on large numbers of examples, but commercial systems generally use a rule-based approach. Even then, high-quality systems include an exception dictionary to cover anomalous pronunciations. Still other systems (such as Orator) include special rules to pronounce loan words from other languages [8]. To further increase the quality of text-to-speech synthesis, limited syntactic analysis is used to determine sentence structure and augment the string of sound symbols with pitch and duration information.

Two synthesis techniques are in common use: mathematical modeling of the waveform generated by the human speech production apparatus (formant synthesis) and splicing prerecorded segments of speech (synthesis by concatenation). Formant synthesizers use the sound symbols to define a sequence of acoustic targets, then interpolate the acoustic signal between these targets to mimic the dynamics of the human voice. Concatenation-based systems achieve the same effect by selecting a sequence of prerecorded elements corresponding to the sound symbols in context and then smoothing the junctures between these elements. Because they are based on actual human voices, concatenation systems tend to produce a richer and more natural-sounding signal, but require more storage space. Concatenation-based systems can be further categorized into two groups: time domain and frequency domain systems. Time domain systems store and concatenate the individual elements using the amplitude of each sample directly, while frequency domain systems transform the signal into a spectral representation which allows easier manipulation of pitch and duration, but with a penalty on natural voice quality.

glish, German, French, Spanish, and Japanese.

Bellcore's Orator system specializes in the pronunciation of names it achieves by a two-step process of first identifying the probable language origin of a name then applying language-specific, text-to-speech rules [8]. In contrast to DECTalk and the BST products, Orator uses a concatenation approach. An inventory of about 1,000 demi-syllables is stored, and the synthetic speech is spliced together from these units. Various computer manufacturers (Apple, IBM, and others) have demonstrated text-to-speech capabilities and are likely to include them in future products. In addition to these systems, a variety of low-end programs and plug-in boards provide low-quality synthesis for PCs and Macintoshes.

In the international arena, Japan is particularly active in speech synthesis research. The Advanced Telecommunications Research Laboratory has achieved state-of-the-art synthesis for Japanese. Within the European Community, CNET in France has pro-

duced an excellent text-to-speech synthesizer based on waveform concatenation. The Belgian company Lernout & Hauspie offers concatenation-based speech synthesis for a large variety of languages. Several other European nations also have commercial text-to-speech synthesizer projects (e.g., Infovox in Sweden), but comparisons between synthesizers for different languages are difficult.

Text-to-speech synthesis technology is available in the form of synthesis chips (e.g., by BST, TI) for use in hand-held pocket dictionaries, answering machines, smart cars, and other electronic gadgets with speech output capability. The most common device is a text-to-speech board for PCs. The current generation of scientific workstations is fast enough to allow high-quality speech synthesis without add-on boards or additional signal-processing hardware.

Speech synthesis quality will improve as more processing power and memory become available. Thus, we expect the most dramatic improve-

Progress in synthesizers appears to have reached an asymptote; successive small improvements are seemingly difficult to achieve.



ments in concatenative synthesis, where large amounts of prerecorded natural data are stored and concatenated together on demand. The formant synthesis approach, as demonstrated by the DECTalk system, seems to have reached a plateau, and subsequent improvements may be difficult to achieve.

New and better prosodic rules based on Pierrehumbert's theory of intonation [10] have been encoded in some synthesizers, and we can anticipate further refinements in prosodic quality in the next generation of synthesizers. While neural network-based synthesis has been demonstrated, these methods so far have failed to approach the quality of formant synthesis or concatenation.

Conclusion

The current state of recognition technology still falls short of human listening capabilities, limiting the application of speech to areas where it provides a quantifiable advantage in productivity. Nevertheless, speech recognition is beginning to have a broad impact; the current introduction of recognition-based services by AT&T, Northern Telecom, NYNEX, and others is a good example. Introduction of speech recognition in additional limited domains should follow. Probably the largest potential area of application is the personal computer interface, both for conventional desktop systems and for the coming generation of small portable devices ("personal digital assistants")

that provide computing and communication functions to the user. The process of introduction will involve both a search for applications where speech adds true value and the development of sophistication on the part of people about the uses and limitations of the technology.

Speech synthesis is currently a stable technology. It is cheap and also usable, although it also falls short of a human standard. Generated speech still sounds unnatural and is less intelligible than natural speech. The incorporation of synthetic speech needs to be well motivated, particularly in situations where more reasonable alternatives are available. Potential uses of synthesis will increase in two domains: the provision of over-

the-phone services and the enhancement of personal computing environments.

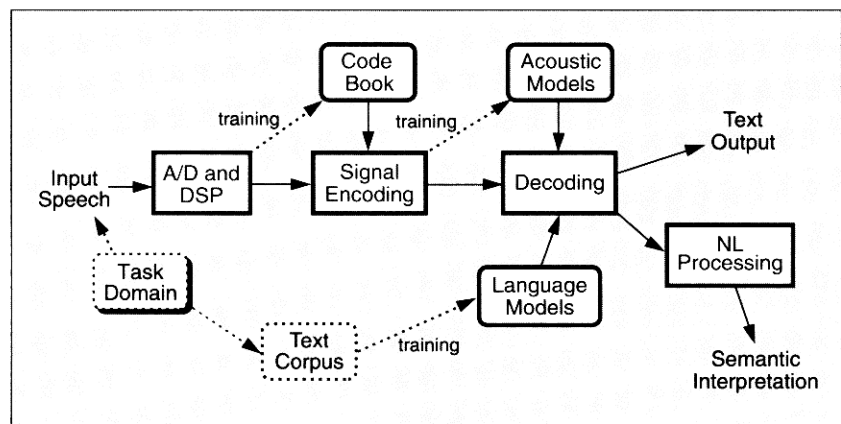
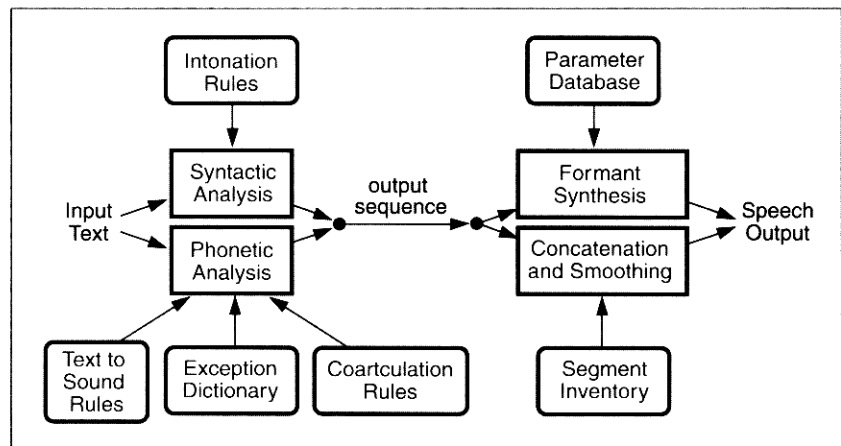
Perhaps the greatest potential lies in the development of systems that combine recognition and synthesis to support conversational interaction between humans and computers in complex task domains. ■

References

1. Allen, J., Hunnicutt, and Klatt, D. *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge, UK, 1987.
2. ARPA. In *Proceedings of the Speech and Natural Language Workshop*. Morgan Kaufmann, San Mateo, Calif., 1992.
3. Huang, X., Alleva, F., Hon, H.-W., Hwang, M.-Y., Lee, K.-F., and Rosenfeld, R. The SPHINX-II speech rec-

Figure 1. Block diagram of an HMM-based recognition system. Square boxes are processes. Rounded boxes are knowledge bases. Dotted items are off-line components.

Figure 2. Block diagram of text-to-speech synthesis. Both major approaches rely on similar pre-processing of text, but differ in the technique used for acoustic realization.



- ognition system: An overview. *Comput. Speech Lang.* 2, 2 (1993), 137-148.
4. Jelinek, F. The development of an experimental discrete dictation recognizer. In *Proceedings of the IEEE 73*, (Nov. 1985), 1616-1624.
 5. Klatt, D. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.* 82, 3 (Sept. 1987), 737-793.
 6. Lee, K.-F., Hon, H.-W. and Reddy, R. An overview of the Sphinx speech recognition system. *IEEE Trans. Acoust. Speech Sig. Proc. ASSP-38*, 1 (Jan. 1990), 35-45.
 7. Logan, J., Greene, B., and Pisoni, D. Segmental intelligibility of synthetic speech produced by rule. *J. Acoust. Soc. Am.* 86, 2 (Aug. 1989), 566-581.
 8. Macchi, M. and Spiegel, M. Using a demi-syllable inventory to synthesize names. In *Speech Tech '90, Official Proceedings, Voice Input/Output Applications*. Vol. 2, Media Dimensions, Inc., New York, 1990, pp. 208-212.
 9. Martin, G. The utility of speech input in user-computer interfaces. *Int. J. Man-Mach. Stud.* 30, 4 (Apr. 1989), 355-376.
 10. Pierrehumbert, J.B. The phonology and phonetics of English intonation. Ph.D. thesis, MIT, Sept. 1980.
 11. Pisoni, D., Nusbaum, H., and Greene, B. Perception of synthetic speech generated by rule. In *Proceedings of the IEEE 73*, (Nov. 1985), pp. 1665-1676.
 12. Rudnicky, A.I., Lunati, J.-M., and Franz, A.M. Spoken language recognition in an office management domain. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (May 1991), pp. 829-832.
 13. Schmandt, C., Ackerman, M.S., and Hindus, D. Augmenting a window system with speech input. *IEEE Comput.* 23, 8 (Aug. 1990), 50-56.
 14. Waibel, A. and Lee, K.-F. *Readings in Speech Recognition*. Morgan Kaufmann, San Mateo, Calif., 1990.
 15. Wilpon, J., Rabiner, L., Lee, C.-H., and Goldman, E. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Trans. Acoust. Speech Sig. Proc. ASSP-38*, (Nov. 1990), pp. 1870-1878.
 16. Zue, V., Glass, J., Phillips, M., and Seneff, S. Acoustic segmentation and phonetic classification in the SUMMIT system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, New York, (1989), pp. 389-392.

search associate at Carnegie Mellon University. His current research interests include speech synthesis, speech recognition and reading instruction.

ALEXANDER RUDNICKY is a systems scientist at Carnegie Mellon University. His current research interests include speech, language and human-computer interaction.

Authors' Present Address: School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213. Email for Rudnicky: air@cs.cmu.edu; email for Hauptmann: alex@cs.cmu.edu

KAI-FU LEE is the manager of the Interactive Media Laboratory at Apple Computer. His current research interests include speech, language, media-rich computing and human-computer interaction. **Author's Present Address:** Apple Computer, Inc., MS 301-36, 1 Infinite Loop, Cupertino, CA 95014; email:kfl@apple.com

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

© ACM 0002-0782/94/0300 \$3.50

About the Authors:

ALEXANDER G. HAUPTMANN is a re-