Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 10046–10051, October 1995
Colloquium Paper

This paper was presented at a colloquium entitled "Human–Machine Communication by Voice," organized by Lawrence R. Rabiner, held by the National Academy of Sciences at The Arnold and Mabel Beckman Center in Irvine, CA, February 8–9, 1993.

Speech technology in 2001: New research directions

BISHNU S. ATAL

AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974

ABSTRACT **Research in speech recognition and synthesis over the past several decades has brought speech technology to a point where it is being used in "real-world" applications. However, despite the progress, the perception remains that the current technology is not flexible enough to allow easy voice communication with machines. The focus of speech research is now on producing systems that are accurate and robust but that do not impose unnecessary constraints on the user. This chapter takes a critical look at the shortcomings of the current speech recognition and synthesis algorithms, discusses the technical challenges facing research, and examines the new directions that research in speech recognition and synthesis must take in order to form the basis of new solutions suitable for supporting a wide range of applications.**

After many years of research, speech recognition and synthesis systems have started moving from the controlled environments of research laboratories to applications in the real world. Voice-processing technology has matured to such a point that many of us wonder why the performance of automatic systems does not approach the quality of human performance and how soon this goal can be reached.

Rapid advances in very-large-scale integrated (VLSI) circuit capabilities are creating a revolution in the world of computers and communications. These advances are creating an increasing demand for sophisticated products and services that are easy to use. Automatic speech recognition and synthesis are considered to be the key technologies that will provide the easy-to-use interface to machines.

The past two decades of research have produced a stream of increasingly sophisticated solutions in speech recognition and synthesis (1). Despite this progress, the perception remains that the current technology is not flexible enough to allow easy voice communication with machines. This chapter reviews the present status of this important technology, including its limitations, and discusses the range of applications that can be supported by our present knowledge. But as we look into the future and ask which speech recognition and synthesis capabilities will be available about 10 years from now, it is important also to discuss the technical challenges we face in realizing our vision of the future and the directions in which new research should proceed to meet these challenges. We will examine these issues in this paper and take a critical look at the shortcomings of the current speech recognition and synthesis algorithms.

Much of the technical knowledge that supports the current speech-processing technology was created in a period when our ability to implement technical solutions on real-time hardware was limited. These limitations are quickly disappearing, and we look to a future at the end of this decade when a single VLSI chip will have a billion transistors to support much higher processing speeds and more ample storage than is now available.

The speech recognition and synthesis algorithms available at present work in limited scenarios. With the availability of fast processors and a large memory, tremendous opportunity exists to push speech recognition technology to a level where it can support a much wider range of applications. Speech databases with utterances recorded from many speakers in a variety of environments have been important in achieving the progress that has been realized so far. But on the negative side, these databases have encouraged speech researchers to rely on trial-and-error methods, leading to solutions that are narrow and that apply to specific applications but do not generalize to other situations. These methods, although fruitful in the early development of the technology, are now a hindrance as we become much more ambitious in seeking solutions to bigger problems. The time has come to set the next stage for the development of speech technology, and it is important to realize that a solid base of scientific understanding is absolutely necessary if we want to move significantly beyond where we are today.

The 1990s will be a decade of rising expectations for speech technology, and speech research will expand to cover many areas, from traditional speech recognition and synthesis to speech understanding and language translation. In some areas we will be just scratching the surface and defining the important issues. But in many others the research community will have to come up with solutions to important and difficult problems in a timely fashion. This paper cannot discuss all the possible new research directions but will be limited to examining the most important problems that must be solved during this decade.

CURRENT CAPABILITIES

Voice communication from one person to another appears to be so easy and simple. Although speech technology has reached a point where it can be useful in certain applications, the prospect of a machine understanding speech with the same flexibility as humans do is still far away. The interest in using speech interface to machines stems from our desire to make machines easy to use. Using human performance as a benchmark for the machine tells us how far we are from that goal. For clean speech, automatic speech recognition algorithms work reasonably well (2, 3) with isolated words or words spoken in grammatical sentences, and the performance is continuing to improve. Fig. 1 shows the word error rate for various test materials and the steady decrease in the error rate achieved from 1980 to 1992. This performance level is not very different from that obtained in intelligibility tests with human listeners. The performance of automatic methods, however, degrades significantly in the presence of noise (or distortion) (4) and for conversational speech.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

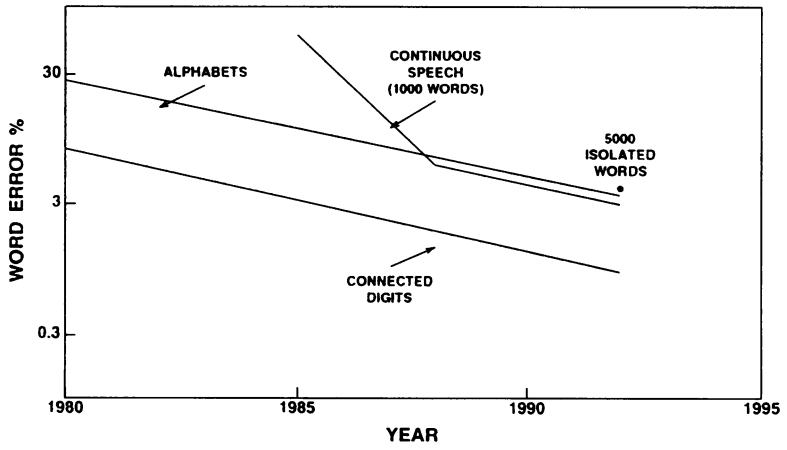

FIG. 1. Reduction in the word error rate for different automatic speech recognition tasks between 1980 and 1992.

There are many factors besides noise that influence the performance of speech recognition systems. The most important of these are the size of the vocabulary and the speaking style. Fig. 2 shows examples of automatic speech recogition tasks that can be handled by automatic methods for different vocabulary sizes and speaking styles. Generally, the number of confused words increases with the vocabulary size. Current systems can properly recognize a vocabulary of as many as a few thousand words, while the speaking style can vary over a wide range, from isolated words to spontaneous speech. The recognition of continuously spoken (fluent) speech is significantly more difficult than that of isolated words. In isolated words, or speech where words are separated by distinct pauses, the beginning and the end of each word are clearly marked. But such boundaries are blurred in fluent speech. The recognition of spontaneous speech, such as is produced by a person talking to a friend on a well-known subject, is even harder.

Examples of speech recognition applications that can be handled by the current technology are shown on the left side of the diagonal line in Fig. 2. These include recognition of voice commands (prompts), names, digit strings, and key-word spotting. New applications in speech recognition are rapidly emerging (5). Commercial products are available for the recognition of isolated words, connected digit strings, and speech with vocabularies of up to several thousand words spoken with pauses between words.

The items on the right of the diagonal line in Fig. 2 are examples of speech recognition tasks that work in laboratory environments but that need more research to become useful for real applications (6). Automatic recognition of fluent speech with a large vocabulary is not feasible unless constraints on the syntax or semantics are introduced. The present knowledge in handling natural languages and in following a dialogue is very much limited because we do not understand how to model the variety of expressions that natural languages use to convey concepts and meanings.

Text-to-speech synthesis systems suffer from much of the same kinds of problems as speech recognition. Present text-to-speech systems can produce speech that is intelligible (although significantly lower intelligibility than natural speech) but not natural sounding. These systems can synthesize only a few voices reading grammatical sentences but cannot capture the nuances of natural speech.

CHALLENGING ISSUES IN SPEECH RESEARCH

For speech technology to be used widely, it is necessary that the major roadblocks faced by the current technology be removed. Some of the key issues that pose major challenges in speech research are listed below:

- *Ease of use.* Unless it is easy to use, speech technology will have limited applications. What restrictions are there on the vocabulary? Can it handle spontaneous speech and natural spoken language?
- *Robust performance.* Can the recognizer work well for different speakers and in the presence of the noise, reverber-

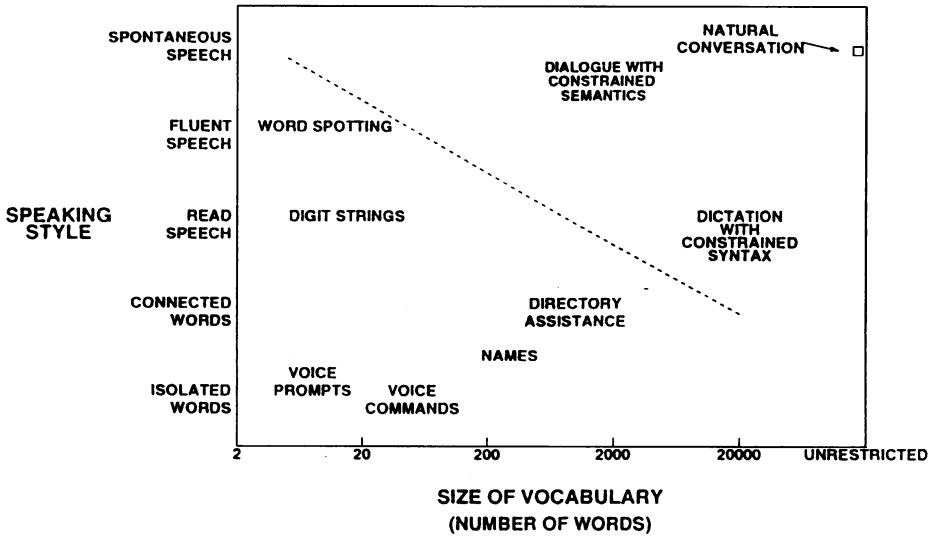

FIG. 2. Different speech recognition tasks shown in a space of two dimensions: speaking style and size of vocabulary.

ation, and spectral distortion that are often present in real communication channels?

- *Automatic learning of new words and sounds.* In real applications the users will often speak words or sounds that are not in the vocabulary of the recognizer. Can it learn to recognize such new words or sounds automatically?
- *Grammar for spoken language.* The grammar for spoken language is quite different from that used in carefully constructed written text. How does the system learn this grammar?
- *Control of synthesized voice quality.* Can text-to-speech synthesis systems use more flexible intonation rules? Can prosody be made dependent on the semantics?
- *Integrated learning for speech recognition and synthesis.* Current speech synthesis systems are based on rules created manually by an experienced linguist. Such systems are constrained in what they can do. Can new automatic methods be developed for the training of the recognizer and synthesizer in an integrated manner?

Some of the issues mentioned above, such as ease of use and robustness, need to be addressed in the near future and resolved. Others, such as automatic learning of new words and sounds or grammar for spoken language, will need major advances in our knowledge. Understanding of spontaneous speech will require tight integration of language and speech processing.

A number of methods have been proposed to deal with the problem of robustness. The proposed methods include signal enhancement, noise compensation, spectral equalization, robust distortion measures, and novel speech representations. These methods provide partial answers valid for specific situations but do not provide a satisfactory answer to the problem. Clean, carefully articulated, fluent speech is highly redundant, with the signal carrying significantly more information than is necessary to recognize words with high accuracy. However, the challenge is to realize the highest possible accuracy when the signal is corrupted with noise or other distortions and part of the information is lost. The performance of human listeners is considered to be very good, but even they do not approach high intelligibility for words in sentences unless the signal-to-noise (S/N) ratio exceeds 18 dB (3).

THE ROBUSTNESS ISSUE

Let us consider the robustness issue in more detail. Current speech recognition algorithms use statistical models of speech that are trained from a prerecorded speech database. In real applications the acoustic characteristics of speech often differ significantly from that of speech in the training database, and this mismatch causes a drop in the recognition accuracy. This is illustrated for noise-contaminated speech in Fig. 3, which shows the recognition accuracy as a function of the S/N ratio for both matched and mismatched training and test conditions (4, 7). These results point to a serious problem in current speech recognition systems: the performance degrades whenever there is a mismatch between levels of noise present in training and test conditions. Similar problems arise with spectral distortion, room reverberation, and telephone transmission channels (8). Achieving robust performance in the presence of noise and spectral distortion has become a major issue for the current speech recognition systems.

Robust performance does not come by chance but has to be designed into the system. Current speech recognition algorithms are designed to maximize performance for the speech data in the training set, and this does not automatically translate to robust performance on speech coming from different user environments. Fig. 4 shows the principal functions of an automatic speech recognition system. The input speech utterance is analyzed in short quasi-stationary segments, typically 10 to 30 ms in duration, to provide a parametric representation at the acoustic level. The parameters from the unknown input utterance are then compared to patterns derived from a large training set of speech utterances collected from many speakers in many different speaking environments. This comparison provides a set of scores representing the similarity between the unknown pattern and each of the prestored patterns. The last step combines these scores together with other knowledge about the speech utterance, such as the grammar and semantics, to provide the best transcription of the speech signal. To achieve robustness, each function shown in the block diagram must be designed to minimize the loss in performance in situations when there is a mismatch between the training and test conditions.

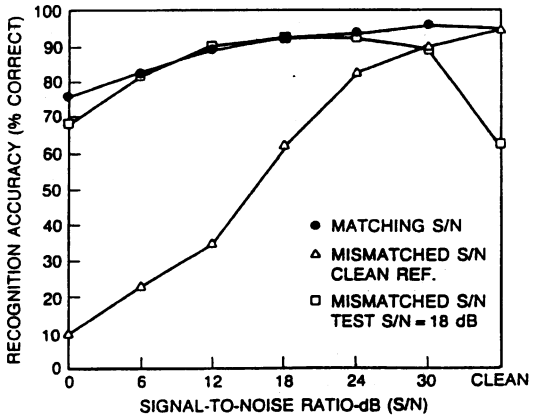

FIG. 3. Speech recognition performances in noisy conditions: ●, training and testing have matched S/N ratios; △, only clean training data are used; □; training and testing S/N ratios are mismatched with test S/N ratio fixed at 18 dB (4).

A speech recognizer can be regarded as a method for compressing speech from a high rate needed to represent individual samples of the waveform to a low phonemic rate to represent speech sounds. Let us look at the information rate (bit rate) at different steps in the block diagram of Fig. 4. The bit rate of the speech signal represented by its waveform at the input of the recognizer is in the range of 16 to 64 kb/s. The rate is reduced to approximately 2 kb/s after acoustic analysis and to a phonemic rate in the range 30 to 50 b/s after pattern matching and selection.

The bit rate at the acoustic parameter level is large, and therefore the pattern-matching procedure must process speech in "frames" whose duration is only a small fraction of the duration of a sound. The scores resulting from such a pattern-matching procedure are unreliable indicators of how close an unknown pattern from the speech signal is to a particular sound. The reliability can be improved by reducing the maximum number of acoustic patterns in the signal (or its bit rate) that are evaluated for pattern matching. The bit rate for representing the speech signal depends on the duration of the time window that is used in the analysis shown in Fig. 5 and is about 200 b/s for a window of 200 ms. Suppose we wish to compute the score for a speech segment 100 ms in duration, which is roughly the average length of a speech sound. The number of acoustic patterns that the pattern-matching step has to sort out is 2^{200} at 2000 b/s, but that number is reduced to only 2^{20} at 200 b/s. This is a reduction of 2^{180} in the number of patterns that the pattern-matching procedure has to handle. The present speech analysis methods generate a static (quasi-stationary) representation of the speech signal. To achieve robust performance, it is important to develop methods that can efficiently represent speech segments extending over a time interval of several hundred milliseconds. An example of a method for representing large speech segments is described in the next section.

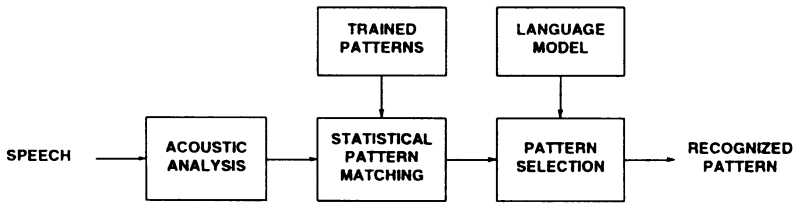

FIG. 4. Principal functions of an automatic speech recognition system.

SPEECH ANALYSIS

The goal of speech analysis is to provide a compact representation of the information content in the speech signal. In general, those representations that eliminate information not pertinent to phonetic differences are effective. The short-time power spectrum of speech, obtained either from a filter bank, Fourier transform, or linear prediction analysis, is still considered the most effective representation for speech recognition (the power spectrum is often converted into the cepstrum to provide a set of 10 to 15 coefficients). However, the power spectrum is affected by additive noise and linear-filtering distortions. We need new representations that go beyond the power spectrum and represent the frequency content of the signal.

The cepstral coefficients are instantaneous (static) features. One of the important advances in the acoustic representation of speech has been the introduction of dynamic features (9), such as first- and second-order derivatives of the cepstrum. Recently, new representations based on human hearing have been proposed (10), but these representations have not yet been found to have significant advantage over the spectral representation. The following is a list of interesting new research directions in speech analysis:

- *Time-frequency and wavelet representations.* Time-frequency representations map a one-dimensional signal into a two-dimensional function of time and frequency (11–13). The traditional Fourier analysis methods divide the time-frequency plane in an inflexible manner not adapted to the needs of the signal. New methods of time-frequency analysis are emerging that allow more general partitioning of the time-frequency plane or tiling that adapts to time as well as frequency as needed (11, 14).
- *Better understanding of auditory processing of signals.* Although auditory models have not yet made a significant impact on automatic speech recognition technology, they exhibit considerable promise. What we need is a better understanding of the principles of signal processing in the auditory periphery that could lead to more robust performance in automatic systems.
- *Articulatory representation.* Models that take advantage of the physiological and physical constraints inherent in the vocal-tract shapes used during speech production can be useful for speech analysis. Significant progress (15) has been made during the past decade in developing articulatory models whose parameters can be estimated from the speech signal.
- *Coarticulation models at the acoustic level.* During speech production, the articulators move continuously in time, thereby creating a considerable overlap in the acoustic realizations of phonemes. Proper modeling of coarticulation effects at the acoustic level can provide better accuracy and higher robustness in speech recognition.

TEMPORAL DECOMPOSITION

We discussed earlier the importance of extending the quasi-stationary static model of speech to a dynamic model that is valid over much longer nonstationary segments. We describe here one such model, known as temporal decomposition (16). The acoustics of the speech signal at any time are influenced not only by the sound being produced at that time but also by neighboring sounds. Temporal decomposition seeks to separate the contributions of the neighboring sounds on the acoustic parameters by using a coarticulation model in which the contributions of sounds are added together with proper weights (16–19).

In the temporal decomposition model the continuous variations of acoustic parameters are represented as the output of a linear time-varying filter excited by a sequence of vector-valued delta functions located at nonuniformly spaced time intervals (17). This is illustrated in Fig. 6, where the linear filter with its impulse response specified by $h(t, \tau)$ (response at time t due to a delta function input at time τ) has the role of smoothing the innovation $\mathbf{x}(t)$ that is assumed to be nonzero only at discrete times corresponding to the discrete nature of

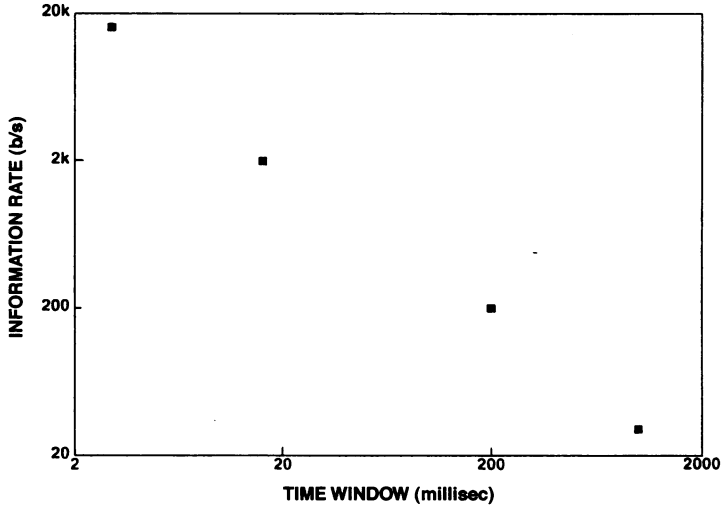

FIG. 5. Information rate (b/s) of speech signal as a function of the length of the time window used in the analysis.

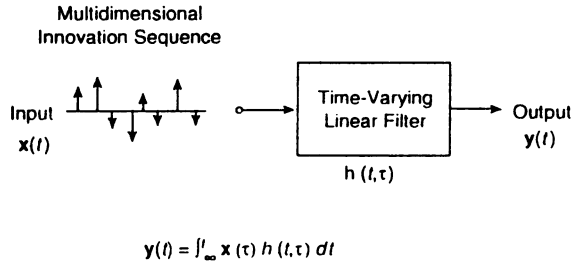

FIG. 6. Temporal decomposition model to represent coarticulation at the acoustic level.

speech events. The number of nonzero components in the innovation in any given time interval is roughly equal to the number of speech events (and silence) contained in that interval of the spoken utterance. Speech analysis techniques have been developed to determine both the innovation and the time-varying impulse response of the filter for any utterance (17–19). Fig. 7 shows an example of this decomposition for the word "four." The three parts of the figure show: (*a*) even components of the linear predictive coding (LPC) line spectral frequencies as a function of time, (*b*) the filter impulse responses for each speech event, and (*c*) the waveform of the word "four." In this example the entire variations in the acoustic parameters over 0.5 s of the utterance for the word "four" can be represented as the sum of five overlapping speech events. We find that the information rate of the innovation signal $\mathbf{x}(t)$ is about 100 b/s, which is much lower than the corresponding rate for the acoustic parameters $\mathbf{y}(t)$.

TRAINING AND PATTERN-MATCHING ISSUES

The application of hidden Markov models (HMMs) has been a major factor behind the progress that has been achieved in automatic speech recognition (1). The HMM framework provides a mathematically tractable approach to the training and classification problems in speech recognition. While the speech recognition algorithms based on the HMM are important at the current state of the technology, these algorithms suffer from fundamental shortcomings (20) that must be overcome.

The HMM method is based on the Bayesian approach to pattern classification, which assumes that the statistical distributions of the HMM states are known or can be estimated. In the HMM, therefore, the problems of training and recognition are transformed to the problem of estimating distributions from the training data. In reality this is a difficult task requiring untested assumptions about the form and the underlying parameters of the distributions. Moreover, the misclassification errors depend on the amount of overlap between the tails of the competing distributions and not on the exact shape of the distributions for the classes. Thus, the emphasis in the HMM approach on distribution estimation is unnecessary; a cost function defined in a suitable fashion is all that is required.

Other approaches to speech recognition based on discriminant functions are being investigated and appear to be promising. Significant progress has been made in formulating the discriminant approach for speech recognition and in developing methods that seek to minimize the misclassification errors (21). The major issues in training and recognition are listed below:

- *Training and generalization.* An important question is whether the trained patterns characterize the speech of only the training set or whether they also generalize to speech that will be present in actual use.
- *Discriminative training.* Although the discriminative training does not require estimation of distributions, they still need knowledge of the discriminant functions. What are the most appropriate discriminant functions for speech patterns?
- *Adaptive learning.* Can the learning of discriminant functions be adaptive?
- *Artificial neural networks.* Despite considerable research, neural networks have not yet shown significantly better performance than HMM algorithms. New research must address the important issue—what is the potential of neural networks in providing improved training and recognition for speech patterns?

ADDITIONAL ISSUES IN SPEECH SYNTHESIS

Much of what has been discussed so far applies to speech synthesis as well. However, there are additional research issues that must be considered. We will discuss some of these issues in this section.

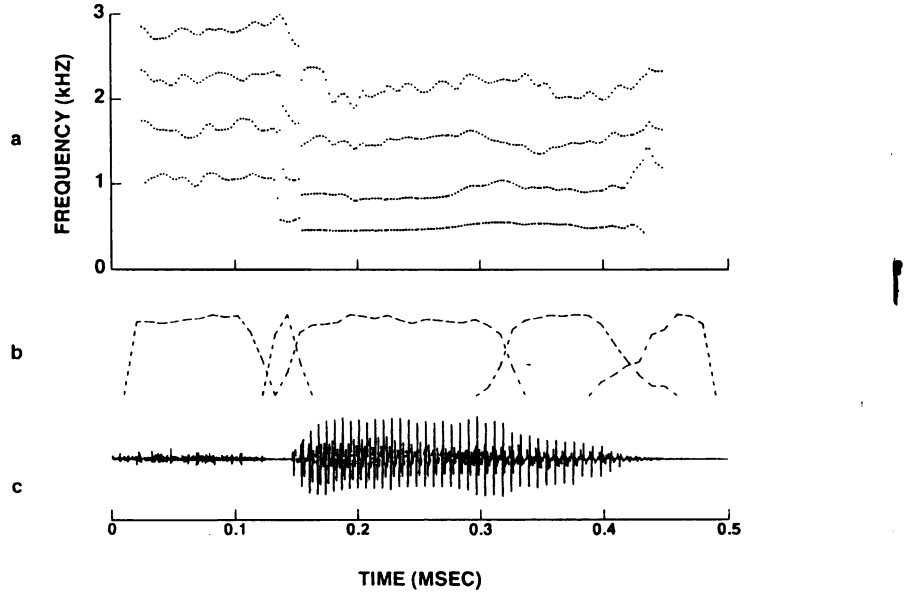

FIG. 7. Temporal decomposition of the spoken word "four": (*a*) LPC line spectral parameters, (*b*) filter impulse responses for the different speech events, and (*c*) speech waveform for the speech utterance.

The core knowledge that forms the basis of current speech recognition and synthesis algorithms is essentially the same. However, there are important differences in the way the two technologies have evolved. Speech synthesis algorithms generate continuous speech by concatenating segments of stored speech patterns, which are selected to minimize discontinuities in the synthesized speech. Segmentation of speech into appropriate units, such as diphones or syllables, was therefore incorporated into the speech synthesis technology at an early stage and required the assistance of trained people (or phoneticians) to learn the segmentation task. Lack of accurate models for representing the coarticulation in speech and the dynamics of parameters at the acoustic or the articulatory level has been the major obstacle in developing automatic methods to carry out the segmentation task. Without automatic methods, it is difficult to process large speech databases and to develop models that represent the enormous variability present in speech due to differences in dialects, prosody, pronunciation, and speaking style. Future progress in synthesizing speech that offers more than minimal intelligibility depends on the development of automatic methods for extracting parameters from speech to represent the important sources of variability in speech in a consistent fashion. Automatic methods for segmentation are also needed in order to develop multilingual capability in speech synthesis.

The primary goal of speech synthesis systems so far has been to synthesize speech from text—a scenario coming out of an earlier interest in "reading machines for the blind." New applications of speech synthesis that do not depend on synthesizing speech from text are rapidly emerging. As we proceed to develop new applications that involve some kind of dialogue between humans and machines, it is essential that the issue of synthesizing speech from concepts be addressed.

CONCLUSIONS

Voice communication holds the promise of making machines easy to use, even as they become more complex and powerful. Speech technology is reaching an important phase in its evolution and is getting ready to support a wide range of applications. This paper discussed some of the important technical challenges in developing speech recognition and synthesis technology for the year 2001 and the new research directions needed to meet these challenges.

Robust performance in speech recognition and more flexibility in synthesizing speech will continue to be major problems that must be solved expeditiously. The solutions will not come by making incremental changes in the current algorithms but rather by seeking new solutions that are radically different from the present.

New speech analysis methods must move beyond quasi-stationary representations of the power spectrum to dynamic representations of speech segments. Solution of the coarticulation problem at the acoustic level remains one of the most important problems in speech recognition and synthesis. Temporal decomposition is a promising method along this direction.

In speech recognition, new training procedures based on discriminant functions show considerable promise and could avoid the limitations of the HMM approach. The discriminant function approach achieves higher performance by using a criterion that minimizes directly the errors due to misclassification. In speech synthesis, articulatory models and automatic methods for determining their parameters offer the best hope of providing the needed flexibility and naturalness in synthesizing a wide range of speech materials.

1. Rabiner, L. R. & Juang, B. H. (1993) *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).
2. Makhoul, J. & Schwartz, R. (1995) *Proc. Natl. Acad. Sci USA* **92,** 9956–9963.
3. Miller, G. A., Heise, G. A. & Lichten, W. (1961) *J. Exp. Psychol.* **41,** 329–335.
4. Juang, B. H. (1991) *Comput. Speech Lang.* **5,** 275–294.
5. Wilpon, J. G. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 9991–9998.
6. Roe, D. B. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10017–10022.
7. Dautrich, B. A., Rabiner, L R. & Martin, T. B. (1983) *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-31,** 793–806.
8. Acero, A. & Stern, R. M. (1990) *Proc. ICASSP-90*, pp. 849–852.
9. Furui, S. (1986) *J. Acoust. Soc. Am.* **80,** 1016–1025.
10. Ghitza, O., (1992) in *Advances in Speech Signal Processing*, eds. Furui, S. and Sondhi, M. M. (Dekker, New York), pp. 453–485.
11. Daubechies, I. (1990) *IEEE Trans. Inf. Theory* **36,** 961–1005.
12. Hlawatsch, F. & Boudreaux-Bartels, G. F. (1992) *IEEE Signal Process. Mag.*, pp. 21–67.
13. Rioul, O. & Vetterli, M. (1991) *IEEE Signal Process. Mag.*, pp. 14–38.
14. Herley, C., *et al.* (1993) *IEEE Trans. Signal Process.*
15. Schroeter, J. & Sondhi, M. M. (1992) in *Advances in Speech Signal Processing*, eds. Furui, S. & Sondhi, M. M. (Dekker, New York), pp. 231–267.
16. Atal, B. S. (1983) *Proc. Int. Conf. IEEE ASSP*, pp. 81–84.
17. Atal, B. S. (1989) in *Towards Robustness in Speech Recognition*, ed. Lea, W. A. (Speech Science Publ., Apple Valley, MN), pp. 209–220.
18. Cheng, Y. M., & O'Shaughnessy, D. (1991) *IEEE Trans. Signal Process.* **39,** 1282–1290.
19. Cheng, Y. M. & O'Shaughnessy, D. (1993) *IEEE Trans. Speech Audio Process.* **1,** 207–220.
20. Juang, B. H. & L. R. Rabiner, L. R. (1991) *Technometrics* **33,** 251–272.
21. Juang, B. H. & Katagiri, S. (1992) *IEEE Trans. Signal Process.* **40,** 3043–3054.