VOICE COMMUNICATION BETWEEN HUMANS AND MACHINES

David B. Roe and Jay G. Wilpon, *Editors*

National Academy of Sciences

National Academy Press
Washington D.C. 1994

Computer Speech Synthesis: Its Status and Prospects

Mark Liberman

SUMMARY

Computer speech synthesis has reached a high level of performance, with increasingly sophisticated models of linguistic structure, low error rates in text analysis, and high intelligibility in synthesis from phonemic input. Mass market applications are beginning to appear. However, the results are still not good enough for the ubiquitous application that such technology will eventually have. A number of alternative directions of current research aim at the ultimate goal of fully natural synthetic speech. One especially promising trend is the systematic optimization of large synthesis systems with respect to formal criteria of evaluation. Speech recognition has progressed rapidly in the past decade through such approaches, and it seems likely that their application in synthesis will produce similar improvements.

Many years ago at Bell Laboratories, Joseph Olive and I sat late one evening at a computer console. We were listening with considerable satisfaction to the synthetic speech produced by a new program. A member of the custodial staff, who had been mopping the floor in the hall outside, stuck his head in the door, furrowed his brow, and asked: "They got dogs in here?"

Things have changed considerably since then. For one thing, computer

consoles no longer exist. Also, speech synthesis has improved a great deal. The best systems—of which the current Bell Labs system is surely an example—are entirely intelligible, not only to their creators but also to the general population, and sometimes they even sound rather natural. Here I will give a personal view of where this technology stands today and where it seems to be headed. This assessment distills contributions from participants in the colloquium presentations and discussion, who deserve the credit for any useful insights. Omissions, mistakes, and false predictions are of course my own.

The ongoing microelectronics revolution has created a striking opportunity for speech synthesis technology. The computer whose console was mentioned earlier could not do real-time synthesis, even though it filled most of a room and cost hundreds of thousands of dollars. Almost every personal computer now made is big and powerful enough to run high-quality speech synthesis in real time, and an increasing number of such machines now come with built-in audio output. Small battery-powered devices can offer the same facility, and multiple channels can be added cheaply to telecommunications equipment.

Why then is the market still so small? Partly, of course, because the software infrastructure has not yet caught up with the hardware. Just as widespread use of graphical user interfaces in applications software had to wait for the proliferation of machines with appropriate system-level support, so widespread use of speech synthesis by applications will depend on common availability of platforms offering synthesis as a standard feature. However, we have to recognize that there are also remaining problems of quality. Today's synthetic speech is good enough to support a wide range of applications, but it is still not enough like natural human speech for the truly universal usage that it ought to have.

If there are also real prospects for significant improvement in synthesis quality, we should consider a redoubled research effort, especially in the United States, where the current level of research in synthesis is low compared to Europe and Japan. Although there is excellent synthesis research in several United States industrial labs, there has been essentially no government-supported synthesis research in the United Sates for some time. This is in sharp contrast to the situation in speech recognition, where the ARPA's Human Language Technology Program has made great strides, and also in contrast to the situation in Europe and Japan. In Europe there have been several national and European Community-level programs with a focus on synthesis, and in Japan the ATR Interpreting Telephony Laboratory has made significant investments in synthesis as well as recognition

technology. There are a number of new ideas at all levels of the problem and also a more general sense that a methodology similar to the one that has worked so well in speech recognition research will also raise speech synthesis quality to a new level.

Before considering in more detail what this might mean, we should consider some of the ways in which speech synthesis research has developed differently from speech recognition research. We will start by exploring what is meant by the term *computer speech synthesis*.

Obviously, this term refers to the creation by computer of human-like speech, but that only tells us what the output of the process is. Synthesized speech output may come from a wide range of processes that differ enormously in the nature of their inputs and the nature of their internal structures and calculations.

The input may be

1. an uninterpreted reference to a previously recorded utterance;
2. a message drawn from a small finite class of texts, such as telephone numbers;
3. a message drawn from a larger or even infinite, but still restricted, class of texts, such as names and addresses;
4. a message drawn from unrestricted digital text, including anything from electronic mail to on-line newspapers to patent or legal texts, novels, or cookbooks;
5. a message composed automatically from nontextual computer data structures (which we might think of as analogous to "concepts" or "meanings"); or
6. a specification of the phonological content of a message, which for most applications must be produced from one of the types of input given previously.

Most commercial applications so far have been of type 1 or 2. Classical "text-to-speech" systems are of type 4 and/or 6. Ultimate human-computer interaction systems are likely to be of type 5, with a bit of 4. Many of the people closely involved in applying speech synthesis technology think that the most promising current opportunities are of type 3. Note that choosing such restricted-domain applications has been crucial to the success of computer speech recognition.

The system-internal structures and processes of "speech synthesis" may involve

1. reproduction of digitally stored human voice, perhaps with compression/expansion;

2. construction of messages by concatenation of digitally stored voice fragments;
3. construction of messages by concatenation of digitally stored voice fragments with modifications of the original timing and pitch;
4. construction of messages by concatenation of digitally stored voice fragments with rule-generated synthetic pitch contours and rule-generated segmental timing values;
5. construction of messages using rule-generated synthetic time functions of acoustic parameters;
6. construction of messages using rule-generated synthetic controls for the kinematics of simplified analogs of human vocal tract; and
7. construction of messages by realistic modeling of the physiological and physical processes of human speech production, including dynamic control of articulation and models of the airflow dynamics in the vocal tract.

The largest scale of commercial activity has been of types 1 and 2, which might be called stored voice. This includes telecommunication intercepts, Texas Instruments' Speak 'N Spell toy, voice-mail prompts, and so forth. Much classical speech synthesis research was of type 5 or 6. Several of the best current systems, and what some consider to be the most promising areas of research, are of types 3 and 4, techniques that are sometimes called *concatenative synthesis.*

These alternative types of computer-spoken messages, and alternative techniques for producing them, seem so different that people often feel that it is unreasonable to use the same phrase to describe them. Despite many efforts to clarify the terminology, however, there is a stubborn tendency to use *speech synthesis* for all of these cases. This tendency is understandable, since there is indeed a kind of continuum of techniques and applications, and as the range of data-intensive synthesis techniques increases, the category boundaries become increasingly blurred. However, it creates considerable confusion, and so we will adopt a more carefully defined terminology.

The present discussion is focused on inputs of types 3 through 6 (i.e., restricted or unrestricted text, or nontextual computer data structures) and synthesis techniques of types 3 through 6 (which involve producing spoken messages from a phonological specification). The process of transforming text into a suitable phonological specification is generally known as *text analysis*, and the process of creating sound from this specification has (confusingly) no common name other than *speech synthesis*, which as we have seen is used for many other things as well. We will refer to it as *speech synthesis proper*, sometimes abbreviated as *speech synthesis* or *synthesis* if the context is clear.

To put the present research situation in perspective, it is useful to present a bit of history. Lawrence Rabiner did his Ph.D. research on a speech synthesis system almost 30 years ago. In using this work as a point of reference, we do not mean to exaggerate its historical role. In a sketch of the intellectual history of speech synthesis, we would cover the early work of Delattre, Cooper, Holmes, Mattingly, Fant, Dixon, and many others, and Rabiner's dissertation would find its place primarily as an influence on the subsequent research of Dennis Klatt. However, in order to make some general points about trends in speech research over the past three decades, Rabiner's work is a particularly useful point of departure.

The results were presented in Rabiner's 1964 MIT dissertation and also described in a 1968 *Bell System Technical Journal* article. This system used a technique of type 5 (rule-generated time functions of acoustic parameters), with a tinge of type 6 (rule-generated articulatory kinematics) in the control of fundamental frequency, based on a concept of subglottal pressure as the crucial variable. Its input was of type 6, consisting of a string of phonemic symbols with stress indications and marks for word boundaries and pauses; thus, it accomplished "speech synthesis proper," with no text analysis component.

The underlying conception for this system is admirably simple: each phoneme is characterized by a single invariant acoustic target, and the observed contextually varied time functions are generated by a smoothing process. In addition to its specified control parameter values, each phoneme defines a specified frequency region around each of the formant values in that vector, indicating tolerance for coarticulatory modification.

The method for creating actual time functions from these tables is general but somewhat subtle. The parameter time functions are generated by critically damped second-degree differential equations whose time constants depend on the parameter and the pair of phonemes involved. The phonemic goals change discretely in time, but the timing of these changes depends on a nonlinear interaction of the phoneme sequence with the computed durations and the specified formant tolerances. A new set of formant targets is not introduced until the formants have reached the tolerance region of the current phoneme, and a durational criterion (only defined for stressed vowels) is also satisfied. Thus, the method could be informally summarized as "move each parameter under the control of phoneme i until all parameters are close enough to their target; then continue for a specified time if the phoneme is a stressed vowel; then switch to the target for phoneme $i + 1$." There are some additional complexities, such as the provision for delaying by a specified amount the change in formant targets for certain formants in a few specified phoneme sequences.

This system was state of the art in 1964, but it was more than a decade earlier than the Bell Labs system that the janitor mistook for a dog, and if we played it alongside one of today's systems, it would be quite clear how far we have come in 30 years.

Enormous progress has been made in the area of text analysis (which of course was outside the scope of Rabiner's dissertation). In the 1960s methods for translating English text into phonological strings did not have very good performance. A high proportion of words were mispronounced, and the assignment of phrasing, phrasal stress, and phrasal melody was ineffective. Today's best text analysis algorithms have mispronunciation rates that are best measured in errors per 10,000 input words and do a reasonable (and improving) job of phrasing and accent assignment. There are a number of factors behind the improvement, but the most important reason is that today's programs simply contain much more information about text than their predecessors. This information may be explicit (e.g., lists of words with their pronunciations) or implicit (statistical rules summarizing the behavior of large bodies of training material).

A similar process has characterized the improvements in speech synthesis proper, the production of sound from a given phonological string. Today's systems are still based on the same general strategy of phonological units sequenced in time. However, the inventory of units is much larger, each unit typically involving two, three, or more phonetic segments, either as distinguishing context for the unit or as part of the unit itself. Often, the internal structure of each unit is much more elaborate, sometimes including an entire stretch of fully specified speech. The timing rules distinguish many more cases, and the procedures for selecting units, combining them, and establishing their time patterns are often quite complex. Between larger tables of units and more complex combination rules, today's systems simply incorporate much more information than Rabiner's system did. Measured in terms of the size in bits of the programs and tables, today's systems are probably two to three orders of magnitude larger.

Although this additional complexity seems essential to improved quality, it is a mixed blessing. It may be argued that most of the recent progress in speech recognition research has been due to two factors:

1. simple architectures that permit program parameters to be optimized with respect to large bodies of actual speech and
2. easily calculated objective evaluation metrics that permit alternative designs to be compared quantitatively.

A similar methodology began to be applied in text analysis more than a decade ago, and it has now become the norm in such work. It is the main reason that text analysis has made such rapid progress, to the point that the real quality bottleneck appears to be in speech synthesis proper, the sound production end of the system.

With some notable exceptions, this methodology has been absent from research in speech synthesis proper until recently. Consider Rabiner's system in light of the two success factors just mentioned. His table of phonemic targets would certainly be amenable to corpus-based optimization; indeed, one can optimize arbitrarily large tables of acoustic targets, as long as enough data are brought to bear. However, Rabiner's method for time-function generation has some properties that would make optimization of its constants somewhat tricky and would hinder optimization of the table of phonemic targets as well. It seems clear that the system was not designed with corpus-based optimization in mind; if it had, Rabiner would no doubt have made certain choices somewhat differently.

Rabiner's 1964 work also does not contain any definition of an evaluation metric that would permit alternative architectures to be compared in a quantitative way. For instance, it is now generally accepted that a single acoustic target for each (surface) phonemic segment is not adequate. Rabiner's 1964 work does not specify a framework in terms of which alternative approaches to the question of subphonemic variation could be compared objectively.

Of course, it is entirely unfair to criticize Rabiner's 1964 work in these terms. His design decisions were not made with these aims in mind. His approach instead seems to be based on a different assumption, which it shared with most other synthesis work of the past 30 years—namely, that success would come from the introduction of a modest amount of fairly high-level scientific knowledge in the form of human-coded programs. From this point of view, the most important goal is not to design a system that can easily be subjected to systematic formal optimization, but rather a system that will permit the introduction of certain scientific models in a convenient and appropriate form.

This was an appropriate point of view in the context of a system as compact and conceptually simple as Rabiner's was. However, the post-Rabiner direction of research in segmental synthesis was (by necessity) toward expanded tables of values, increased algorithmic complexity, and proliferation of special cases in the time-function generation process. These moves made objective optimization even harder to contemplate; at the same time, they brought systems to a level of complexity that taxed the researchers' ability to manage their

development and modification. Researchers like Klatt certainly paid close attention to speech data in setting their parameters, and they often engaged in informal interactive "copy synthesis" as a method for tuning up parameters and algorithms. However, the resulting systems were certainly not designed to facilitate overall optimization of parameters, or objective comparison of alternative algorithms against a large speech database. As the systems grew larger and larger, and their internal interactions grew more and more complex, interactive experimentation by human developers became a less and less viable method for managing the development process.

One might argue that concatenative synthesis methods caught on earlier to the benefits of explicit grounding in large amounts of speech data. Certainly one general lesson of the past decade has been that systems based on minimal manipulation of large bodies of natural speech data often sound better than systems that do deeper and more sophisticated calculations, with a more complex model of how their primitive elements interact. The high quality achieved by some implementations of methods such as PSOLA (pitch-synchronous overlap-add approach) even suggests to some that the apotheosis of superficiality might extend to time domain over frequency domain methods of signal manipulation. However, even very data-intensive concatenative approaches have usually not been quantitatively optimized in the way that speech recognition algorithms routinely are. Instead, someone simply picks an inventory design, a segmentation scheme, and a set of rules for choosing and combining elements and then sets to work building an inventory by manual accumulation of individual elements.

Only within the past few years have we seen a general use of systematic optimization techniques for purposes of inventory design, unit segmentation, unit selection, and unit combination algorithms. The general approach is to define a perceptually reasonable acoustic distortion metric and use it in a global comparison of alternatives (in allophonic clustering, in segmentation points, in unit selection, or whatever). To make this method work effectively, one must usually design the overall system specifically with such a process in view. Psychological tests would be the optimal basis of such an effort, but objective (if psychologically motivated) distortion metrics have the advantage of being quicker and cheaper. Although such objective distortion metrics are far from a perfect image of the human judgments that provide the ultimate evaluation of any synthesis system, they usually provide the only feasible way to perform the massive and systematic comparison of alternatives that is needed. Testing with human subjects can then be used to provide validation at strategically chosen points.

Researchers at NTT and ATR in Japan have been especially prominent in these explorations, and their initial results look very promising. As such methods gain wider application, and especially as we see general availability of the large-scale single-speaker databases that will be required to support them, we can hope to see an increased rate of improvement in segmental speech synthesis quality. Thus, increased investment in speech synthesis research is warranted, both because there is an opportunity created by advances in microelectronics and because there are significant new ideas and new methods waiting to be applied.

As this research goes forward, it faces some pointed questions. What will it take to make synthetic speech that sounds entirely natural, or at least better than word concatenation voice response systems for restricted phrase types such as name and address sequences? Will progress come by a scientific route, through better modeling of human speech production, or by an engineering route, through larger inventories of prerecorded elements with optimal automatic selection and combination methods? How far can we push current ideas about text analysis algorithms? How can we produce more natural-sounding modulation of pitch, amplitude, and timing, and how important are such prosodic improvements relative to segmental improvements?

What will it take to put speech synthesis into true mass market applications? What will those applications be? Will the key development be cheaper hardware, a particular "killer" application, or better-quality synthesis? Will there be a gradual spread of the existing niche markets or a single breakthough?

How should we quantify progress in synthesis quality? What is the proper place for subjective testing relative to objective distortion metrics?

The papers by Carlson and Allen in this volume present a solid foundation of fact for evaluating these questions, and a wide variety of opinions were aired in the symposium discussion, from which an individual point of view has been distilled in this introduction. The next decade will be a lively and interesting time in the field of speech synthesis research, and there is little doubt that the situation will look very different 10 years from now.