The Need for Increased Speech Synthesis Research: Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis

Editors: R. Sproat, M. Ostendorf, and A. Hunt

Contributions from some of the participants in the NSF Speech Synthesis Workshop, listed in Appendix A, with special thanks to M. Macon and J. van Santen.

5 March 1999

Contents

EXECUTIVE SUMMARY

Speech synthesis is the technology that underlies the ability of computers to talk. Talking machines are currently being deployed in a variety of applications ranging from messaging systems that can read one's e-mail over the phone, to aids for the visually impaired, who often depend upon talking machines as their *only* access to the World Wide Web and the wealth of information contained therein. And there is every reason to believe that the use of speech synthesis technology will increase in the coming years, and expand into areas such as language instruction, automated announcement systems, and telephone access to internet services and information for the majority of people without online access. In short, speech synthesis technology is having an increasing impact on how we work and live in our high tech society.

But despite the actual and future potential of speech synthesis, users are not generally happy with the quality of current systems. The human ear is wonderfully sensitive to spoken imperfections and deviations from the norm, and thus speech synthesizers must meet high perceptual standards, but do not yet do so. There are a variety of reasons why current systems are not up to par: both the analysis of the input text (most applications of speech synthesizers start with text as the input) and the different mechanisms associated with generating the speech waveform are imperfect, leading to inappropriate renditions of a sentence and perceptual artifacts that together give synthetic speech an unnatural quality. Solving these problems will involve fundamental research in a variety of areas, including basic understanding of the production and perception of human speech and language, and how to construct computational models of these processes. This report outlines what these areas are, and what kind of research is needed.

The report also outlines a number of sub-topics that the speech synthesis research community must address. The multilingual nature of the world and of our own society dictates that we solve these problems not just once (for English), but many times, for multiple languages. We must understand what distinguishes different speaking voices, styles and moods and be able to reproduce them. We must be able to produce synthetic speech that can be understood in adverse conditions; for example, in noisy environments or when the listener has competing demands for their attention. We need to explore how to effectively combine speech synthesis with other computer interface modalities; for example, in dialog systems with speech recognition, and facial animation synchronized with speech.

Unfortunately, for at least twenty years — since the end of the MITalk research program in synthesis at MIT — essentially all of the research in speech synthesis in the United States has been carried out at large industrial labs or else at small privately held companies whose sole business is building and selling speech synthesis systems. In particular, there has been little government funding available for such work. This is unlike the situation in automatic speech recognition, which has received substantial amounts of government funding and is now moving towards mainstream adoption. It is also unlike the situation in Europe where government funding of a number of multi-year projects has taken place, and where such useful achievements as the open source *Festival* text-to-speech system from the Centre for Speech Technology Research at the University of Edinburgh have been the result.

The work in the United States at the industrial sites *has* advanced the state of the art in speech synthesis, but arguably not at the rate at which it might have advanced if public funding had been available. This is for expected reasons: companies that have invested in speech synthesis technology do so with the desire to get a technological "edge" over their competitors. While such competition drives innovation, it does so at a

cost of wide replication of the same or similar kinds of work at different sites. And while some results from industrial sites are published, these results are often hard to replicate since they frequently depend upon databases that are not public. We therefore believe that future improvement of the technology crucially depends upon the involvement of government funding agencies who are able to support a large-scale public-domain research effort.

The recommendations of this report are as follows:

- **Support annual workshops for reporting progress.**

  A regular process for bringing together speech synthesis researchers will facilitate sharing of ideas, training of new researchers, peer review of research and the opportunity to evaluate and report on advances in this field. Therefore, this report recommends support for workshops every year, which would be devoted in part to reporting results on common speech synthesis tasks.

- **Support a broad portfolio of research.**

  In the twenty years since the research by the MITalk group at MIT, government research funding for speech synthesis in the United States has been extremely limited. The result is a near absence of university research in the area: only a handful of US academics have some interest in the area and perhaps less than a half dozen graduate students are focusing on speech synthesis problems. Therefore, this report recommends support for a broad range of independent research activities to increase the level of long-term research on a host of fundamental speech science issues, to encourage exploration of a diversity of computational models and to develop a greater base of expertise in speech synthesis for ongoing academic and commercial efforts.

- **Build a speech synthesis research infrastructure.**

  Databases of text, recorded human speech and other data are fundamental to speech science. Shared high quality analyses of these databases plus standard tools for manipulating the data are a key component of the proposed research infrastructure. Affordable databases decrease the barriers to entry of new researchers into the field, facilitate scientific comparison of research at different sites, and allow researchers to focus on scientific issues rather than the mechanics of data collection. The collection, labelling and distribution of databases is expensive and time-consuming to an extent that is prohibitive for most institutions — especially academic institutions. Therefore, this report recommends support for producing research databases for speech synthesis, for development of common research tools that support usage of the databases, common evaluation tools, and standards for exchange of text and speech data in common formats.

- **Support multi-disciplinary centers of excellence.**

  In many fields, including the closely related field of speech recognition, the establishment of opportunities for researchers to collaborate for extended periods of time (weeks, months or longer) has driven research. For speech synthesis, the special needs for such collaboration are for the training of new researchers as academic involvement is reestablished and for facilitating multi-disciplinary environments that combine the necessary efforts of researchers from computer science, engineering,

linguistics, psychology and other fields — researchers who typically work in distinct institutions. Therefore, this report contains specific recommendations for funding of centers of excellence.

1 INTRODUCTION: THE NEED FOR IMPROVED SYNTHESIS

On August 6–7, 1998 the Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis was held at the National Science Foundation, Arlington VA. The participants, listed in Appendix A, represented a variety of backgrounds, including industry and university experts on speech synthesis, linguistics, speech recognition, technologists with an interest in applying speech synthesis technology, and government representatives. The stated goals of the workshop were:

- To assess the state-of-the-art in speech synthesis research and the driving application needs;
- To identify promising techniques that should be researched to significantly advance the state-of-the-art; and
- To identify evaluation paradigms and infrastructure that can drive the research agenda.

This report has evolved out of presentations and discussions during and subsequent to that workshop.

After several decades of research into text-to-speech (TTS) conversion, many highly intelligible speech synthesizers are available for a variety of languages. Compared to speech recognition systems (which have also started to become widely available), TTS systems are often computationally cheap, requiring comparatively small amounts of memory and computing power. Fifteen years ago, TTS systems were commonly sold with special purpose hardware: the Digital Equipment Corporation, for example, sold a special DECTalk box that plugged into one's computer. Today practically all commercially available TTS systems are software-only and run on standard PC's.

Concomitant with the availability of higher quality TTS systems, coupled with their decreasing demand on available computational resources, has been an increase in applications involving synthesis. This increase has been partly due to the fact that speech recognition has become more widely available: if one has a system that "understands" speech, one often wants it to be able to talk back. It is partly due also to the growth of the messaging industry: more and more customers are demanding convenient, ubiquitous access to their electronic messages, and many services of this kind – e.g., e-mail reading over the phone – require TTS. But the increase of applications involving synthesis has also been due in part to the belief that synthesis is "ready" for the market in a way that it was not ten years ago. Some current and potential applications of speech synthesis are detailed in Section 2. As we shall argue, the one thing that all of these applications of speech synthesis have in common is that they are clearly beneficial to the users who will interact with them. In many cases, they provide the only means by which someone can get access to information when they want it.

Because of the great potential of synthesis and the fact that it is already being deployed in a number of applications, there is a tendency to think of speech synthesis as a solved problem. However, the reality is that users are not happy with the quality of available synthesizers. Many people find themselves using systems not because they like them, but because they have no other choice. Thus visually impaired users will often use speech output (the only other choice being braille displays) as an alternative display device to help them navigate around a screen, but this does not mean that they are happy with the quality. For instance, most blind senior citizens instantly reject reading machines because of their perception of the voice as too

synthetic. In addition, many commercial application developers intentionally constrain the functionality offered by a system so as to avoid the use of TTS, since the unnatural quality of the output detracts from system quality as perceived by consumers. Clearly synthesis is *not* a solved problem.

What makes synthesizers sound unnatural? The answer is not simple; it is due to a complex set of issues including imperfect linguistic analysis of the input, lack of understanding of the relationship between linguistic structure and the acoustic cues to that structure, inability to appropriately modulate the pitch of the voice, and poor models of the voice itself. We will try to explicate these matters more fully in the sections that follow.

What limits the ability of the speech synthesis research community to better attack these problems is the lack of sufficient resources for research in synthesis; it is this lack of resources that is the major topic of this report. Twenty years ago a landmark project in text-to-speech conversion was wrapping up: the MITalk system, developed in the Research Laboratory of Electronics at MIT during the 1970's was the last major speech synthesis effort (focusing on all aspects of the problem) carried out at a US academic institution. Since then, nearly all research in this area has been carried out at corporations. This is quite unlike the situation in Europe and elsewhere, where funded academic research in synthesis is an active enterprise and has yielded such important accomplishments as the open-source Festival TTS system [Black et al., 1998], discussed later on.

The US companies involved in TTS research have been quite varied, ranging from major telecommunications and computer companies with large in-house research labs, down to small privately held corporations whose sole business is the production of synthesis systems. The major drawback of this commercial basis is obvious: precisely because the concerns involved are commercial, there has been little interest in sharing data, methods or results. Papers are often published, but realistically it would be very difficult for others to replicate the work since it is usually done on proprietary data. There is little doubt that this situation is in the short-term interests of companies who have invested in synthesis research: they are naturally disinclined to give away results that they have paid for. However, it is not in the longterm interest of advancing the state of the art, since such a situation inevitably leads to a state where the same or equivalent techniques are reinvented in parallel. Furthermore, it has led to a situation where very few students are trained in synthesis technology, so industry has difficulty hiring knowledgeable staff. Since so much work is commercial, research-oriented synthesis tools are scarce, constraining academic work in synthesis but also impacting other aspects of speech research that build on this technology, such as research on communications disorders. Therefore, this report recommends a program for comprehensive funding of speech synthesis, as outlined in the Executive Summary and detailed in Section 5.

2 SOCIAL IMPACT OF SPEECH SYNTHESIS

Speech synthesis technology is becoming increasingly important in our society. Speech synthesis systems are being deployed in a wider and wider variety of applications. Furthermore, there is a significant minority – people with various physical disabilities – for whom synthesis technology has been important already for many years. However, synthesis also has a much broader impact, for universal access to computers more generally, as well as on commercial growth and scientific advancement. These issues are explored in the sections to follow.

2.1 Current and Future Applications

To put the matter simply, speech synthesis allows access to information, information which, for various reasons including physical disability or lack of a visual display device, one cannot access by other means. This point is illustrated by the range of applications in which speech synthesis systems are currently employed. Among these are:

- *Messaging systems:* Universal messaging systems provide one point of access to e-mail, paging, fax, voice-mail and other messages over the phone. Speech synthesis is required for e-mail, paging, sometimes fax and other non-audio message formats.

- *Reverse directory systems:* The user enters a telephone number, and the system reads the name(s) and address(es) of the business or individual with that number. Or perhaps the user is interested in all the drug stores within a certain zip code that participate in a particular drug plan: the user enters the zip code using the telephone keypad, and the system reads the names and addresses of the participating businesses.

- *Interactive voice response systems:* For limited domains, such as automated call routing systems, good quality synthesis can be achieved and the resulting automation can have a major impact on the number of calls that can be serviced.

- *Alternative access to information:* For people with certain physical disabilities, speech synthesis can enable access to web pages, news reports and other electronically disseminated information. Indeed, screen readers and voice browsers for the visually impaired constitute one of the best established applications of speech synthesis technology. There is also a growing demand for tools to aid people with reading disabilities more generally, such as dyslexia. Other devices might include a handheld scanner and text-to-speech system that would be used in stores by the visually impaired to read labels and descriptive information.

- *Dual party relay:* This telecommunications service is offered to people who are hearing impaired. The hearing-impaired user types with a special teletype machine, and this typed message is automatically converted to speech for the hearing party at the other end of the line. The response is transcribed by a human operator (current speech recognition technology is not sufficient for this task) and sent back to the hearing-impaired person's teletype. Although a human operator is still needed in the loop, users of such systems prefer having the operator listening to only half of the conversation.

- *Augmentative communication:* Speech synthesis acts as the voice for people who have lost or never had the ability to speak. Probably the most famous individual in this category is the physicist Stephen Hawking.

One can add to this list by considering potential future applications, applications which are not commercially real at present, but where speech synthesis systems can clearly be deployed. In each of these application domains, the key factor that continues to inhibit deployment is not the availability of speech synthesis systems, nor generally their price. Instead, it is that the quality of the existing systems is not acceptable to the majority of potential users.

- *Foreign language instruction and other tutoring systems:* Much language instruction involves drilling on examples. Traditionally this has been handled by native speakers recording many hours of drills onto tape, to be replayed by the language learner. If synthesizers of sufficiently high quality were available, this process could be automated, at a substantial savings. Furthermore, the door would then be open to more flexible language-instruction systems where new materials could be generated and "recorded" on the fly. For example, foreign language instruction could vary the pitch (intonation) to illustrate its effects on meaning of words and phrases. Language instruction is not the only pedagogical context in which speech synthesis systems could be used to advantage. Interactive literacy programs for both children and adults could benefit from high quality synthesis that goes beyond existing pre-recorded phrases. Handheld educational computer games for children could include spoken explanations of mathematical or scientific concepts, since synthesized speech typically requires far less storage than recorded speech. Reading tutors could read aloud stories generated on the fly, inserting information related to each person's reading level and interests, and including the person's name.

- *First language instruction for the hearing impaired.* Aligning facial animation to the synthetic speech provides an additional valuable dimension in language tutoring. Children with hearing-impairment, for example, require guided instruction in speech perception and production. Although there is a long history of using visible cues in speech training for individuals with hearing loss, these cues have usually been abstract or symbolic rather than direct representations of the vocal tract. Some of the distinctions in spoken language cannot be heard with degraded hearing – even when the hearing loss has been compensated by hearing aids or cochlear implants. To overcome this limitation, visible speech can provide speech targets for the child with hearing loss. Furthermore, many of the subtle distinctions among speech segments are not visible on the outside of the face. In addition to providing information about the outside of the face, the skin of a wireframe model can be made transparent in order to show the articulators that are normally hidden. One motivation for developing a hard palate, teeth and tongue [Cohen et al., 1998] was to instruct children with hearing loss by revealing the appropriate articulation via the hard palate, teeth and tongue.

- *Job training simulations:* Simulations using synthesized voice can be used for training pilots as they talk with air traffic controllers, and to provide voice reports of the results of taking certain actions (e.g., movement of a control surface that could not be seen by a maintenance person on an aircraft). These simulations may in time extend to two-way conversations, as in sales skills practice.

- *Technical support for complex tasks:* For example, technicians repairing complex equipment often have their hands and eyes occupied, and would benefit from having their instructions read to them. For large pieces of equipment (for instance, a complex telephone switch) the repair manual may comprise thousands of pages, and it is an expensive and time-consuming proposition to pre-record all of this. A speech synthesis device deployed in a portable or wearable computer, or one accessed by telephone would be able to read relevant pages of the manual to the technician working in the field. This architecture can be applied to a range of applications in which the user is hands-busy, eyes-busy or both, and where conventional display-based computer output is impractical. Examples include:

describing steps and locations for shutting down or repairing a vessel in a chemical plant; driving to a specified location with voice prompts from a global positioning system (GPS); providing guidance while the user performs software functions (voice prompts); and setting up an electronic circuit and taking measurements in a learning lab.

- *Automated announcement systems:* At airports, train stations and other public places, public address systems are used to announce important information. Many of these announcements are highly stylized ("Passengers on UA flight 342 to Cleveland should now proceed directly to gate B43"), and much of this could be handled by speech synthesis systems. Note also that such announcements are often made in more than one language, and the person reading the announcement is frequently not very competent in languages other than their own. In such cases synthesizers have the potential to be more effective than their human counterparts.

- *Information and service access:* Automated announcement systems are a sub-class of the range of information and service provision systems. A large number of interactive voice response systems now deployed on the telephone network could be made more flexible and powerful if responses could be dynamically generated with speech synthesis: for example, systems that use touch tone input to get access to movie timetables, restaurant locations, or product catalogs. Similarly, many of the self-service applications provided on internet web sites could be enhanced with speech synthesis as part of a multimodal interface to simplify the computer interaction.

- *Human-centered computer systems:* Interfaces that allow spoken language communication, both as input and output, will enable a broader segment of society to access computers and will make human-computer interaction more efficient for those who are already computer-literate. Visual speech synthesis for multimodal interaction will add to the appeal of such interfaces for many people. Such systems require research on speech recognition and language understanding for successful speech input, but progress in this area has been rapid and is ongoing.

- *Speech-to-speech translation systems:* Speech translation systems are very appealing for people traveling to countries where their language is not widely spoken, or even for telephone-based transactions. Prototype systems already exist, but are only successful when operating in very constrained domains. Again, the success of this application also depends on advances in speech recognition and language translation.

A speech synthesis research program must have as one of its underlying objectives the aim of making practical most, if not all, of the application types described above. The factor that most often distinguishes between what is a potential application and what is now a deployed and widely used application is the willingness of users to accept the current quality level of speech synthesizers. Where the use of speech synthesis is an absolute requirement to support an application, where users have no alternative means of receiving computer output but through speech synthesis, or where a specific niche exists, speech synthesis is now used. But even with these constraints, it is often reported that user acceptance can be problematic and that obtaining and retaining customers is difficult.

These real and potential applications all serve to underscore the basic point that we made at the beginning of this section: speech synthesis technology is becoming increasingly important in all segments of our society and has tremendous potential to improve access to electronic information and services. The shrinking of computing devices and increasing use of remote computer communication will make speech interfaces a necessity. Furthermore, increasingly global markets will demand multilingual capabilities. The US must increase its support of speech technology research in general, but particularly in synthesis, to maintain its leadership position in information technology.

2.2 Speech Synthesis as an Assistive Technology

Many of the current applications of speech synthesis listed at the start of the previous section are aimed at improving communication for people with disabilities. It is true that a small fraction of people with disabilities depend heavily on speech synthesis technology already, and will continue to do so no matter how other developments in speech synthesis fare. However, far more would make use of speech synthesis if the quality were better. Access to the written word is limited for a significant portion of our society. Disability or physical limitations are a significant segment of this. People with vision impairments and learning disabilities often cannot access printed material, or are much less productive in reading. A large proportion of senior citizens have mild to severe visual impairments. People with physical disabilities or motor impairments may not be able to handle written material to read it. A significant portion of our society is also limited by the inability to speak or to speak clearly. Many deaf people are not able to speak. The lack of sign language interpreters makes this limitation have great significance. People with motor disabilities, or people who have had diseases or injuries that affect their speech abilities, may also not be able to speak at all or speak clearly.

Unfortunately the lack of natural sounding speech synthesis creates barriers: the majority of people with disabilities and the people they interact with are not satisfied with the quality of current systems, which provide degraded communication and access to information relative to natural speech communication. These barriers are worth removing. Speech synthesis technology directly impacts the 20-40 million Americans with disabilities, but also indirectly their friends, families and co-workers. Furthermore, assistive technology is often a stepping stone for products that eventually find broader consumer applications as rapidly growing computer power drives down product costs.

2.3 Speech Synthesis and Universal Access

As a society, information access is critical for education, employment and an informed citizenry. An increasing number of legislative efforts are attempting to require this access as part of civil rights, telecommunications and government procurement activities. The Americans with Disabilities Act and the Telecommunications Act of 1996 both have provisions that provide for improved access to information for people with disabilities. The new federal Section 508 legislation has extensive provisions placing the burden on the federal government to make information accessible to all citizens on an equal basis, with a high standard of effort to make this possible.

Speech synthesis can and should play a part in improving universal access to information. Universal access aims to provide all people with affordable, effective and timely access to the widest possible range

of information and services. Note that universal access is not just an issue of physical disabilities; it also means making the information available on the internet and the power of computers available to citizens who are not computer experts. Our society is heading in a direction of information "haves" and "have nots", where those who are not able to access or use computers – still the majority of the population – will become disenfranchised. The development of spoken language computer interfaces, which include voice output as well as input, can help fight this trend because a much larger percentage of the population has access to a telephone.

A final dimension of access that needs to be considered is where information can be accessed. With audio interfaces on small devices (telephones, mobile phones, personal digital assistants and, in the future, microwaves) speech synthesis will enable information access without the traditional desktop, display-based computer interface. Thus speech synthesis has the potential to enable access to information by more people and in more locations.

In the context of this discussion it is of particular note that there is a new multi-year research focus on Universal Access under the auspices of the Human-Computer Interaction and Knowledge and Cognitive Systems within the Information and Intelligent Systems Division of the National Science Foundation. Speech synthesis is a key component technology of Universal Access, and a research focus on speech synthesis under the auspices of one or more NSF programs would therefore be timely.

2.4 Speech Synthesis and the Scientific Infrastructure

Because speech synthesis research touches on so many disciplines, it also has the potential for broad impact on scientific advancement in several fields. Linguists and speech scientists have frequently used speech synthesis systems as a research tool in experiments aimed at developing a better understanding of the speech communication process. Improved synthesis systems will improve the experimental paradigm and make it possible to pursue deeper questions. As argued in [Cole, 1995],

> Putting effort into synthesis will pay off in terms of better quality synthesis as well as better understanding of spoken language, which will in turn lead to improved models for recognition and understanding.

In addition to automatic speech recognition and understanding, improved models of speech can impact low bit rate coding [Flanagan et al., 1980, Schroeter and Sondhi, 1992], as well as medical applications. An accurate model for human articulation could potentially be used to estimate the shape of the vocal tract using only the acoustic signal – a challenge known as the speech inverse problem (see, for example, [Schroeter and Sondhi, 1994]). A non-invasive, analytical technique such as this would contribute the ability to investigate various communications disorders. Given a better understanding of vocal tract physiology, the effect of certain surgical procedures on the vocal cords or vocal tract could be studied prior to performing the procedure. This would benefit both patients and surgeons alike. Finally, an improved understanding of articulation and the ability to synthesize speech from an articulatory description have applications for speech therapy and language instruction, as described in the previous section on applications.

3 BACKGROUND ON THE TECHNOLOGY

This section reviews, in some technical detail, the history of speech synthesis research, the current status of the technology and assessment methodologies, the commercial systems now available, and areas of active research. (As a reference for the non-specialist, a glossary of technical terms is included in Appendix B.) In reviewing all the areas of speech synthesis the report identifies areas in which increased research will drive improvements in overall quality, which are elaborated on in Section 4. In Section 5, we make specific recommendations on a research program in these areas.

3.1 A Brief History of Speech Synthesis Research

In order to understand the current state of the art as well as the future of speech synthesis technology, it is necessary to have a general sense of the history of the field. To this end, we present in this section a brief overview of work on speech synthesis and text analysis for text-to-speech synthesis. Complete reviews are found elsewhere in published materials, from which much of the current material is drawn, including [Klatt, 1987, Allen, 1992, Olive, 1997].

3.1.1 The Synthesis of Speech Sounds

The history of speech synthesis can be traced to the late 18th century and the first attempts by Kratzenstein and von Kempelen to build mechanical devices to mimic the sounds produced by the human vocal apparatus. These devices consisted of bellows (functioning as "lungs"), and a reed ("vocal chords"), which excited a resonance chamber ("vocal tract"). One could generate different speech-like sounds by manually altering the shape of the resonance chamber, just as the tongue, lips and jaw are used in human speech production to change the shape of the vocal tract.

While Kratzenstein and von Kempelen's devices demonstrated a practical understanding of some of the basic aspects of speech production, it wasn't until the mid 19th century, with studies by von Helmholtz and others that the acoustic basis of speech production began to be understood. A critical result of these studies was the realization that different speech sounds could be produced by careful control of the relative loudness of different regions of the spectrum. Since sounds could be produced electrically as well as mechanically, it followed that one could synthesize speech by electrical rather than mechanical means.

The synthesis of speech by electronic means began in the 1920's with work of J. Q. Stewart on the production of static vowel sounds. The first electronic system which allowed one to synthesize intelligible sentences was the Voder, developed by Homer Dudley at Bell Labs. It consisted of two independent sound (or "excitation") generators, one for periodic sounds (vocal chords during voiced sounds) and the other for noise (turbulence caused by constrictions in the vocal tract). A manually operated filter mimicked the effects of the vocal tract, and an additional control was available to control pitch. Demonstrated at the 1939 World's Fair, the Voder clearly showed the potential of the electronic production of speech. It was also notable in that its fundamental design was based on a separation of the *source* (periodic or noisy excitation) from the *filter* (the vocal tract), and thus anticipated the later *source-filter* model of speech production.

While this early work on speech synthesis clearly demonstrated an understanding of speech production, speech science was handicapped by the lack of a means for visually representing speech in a way that clearly

showed the differences between different classes of sounds. The *sound spectrograph,* developed at Bell Labs by Potter and others in the late 1940's, was the first device to accomplish this. The book describing this work, entitled *Visible Speech* [Potter et al., 1947], Potter *et al.* cataloged a large number of sound combinations in English (and selected other languages), illustrating for the first time precisely how sounds are distinguished from one another by differences in energy at different frequencies. Sound spectrograms (though produced by digital rather than analog means) are still a fundamental tool of speech scientists today. Since spectrographic representations of speech allowed researchers to discover robust acoustic correlates of different movements of the vocal apparatus, it became possible to conceive of automatically generating speech by rule. The sound spectrogram was also important for the development of the *source-filter* model of speech production, mentioned above, and developed most notably by Gunnar Fant at KTH in Stockholm. Simply put, the source-filter theory states that speech can be modeled as one or more periodic or noise sources (the glottis or in the case of fricatives, a constriction), which are then filtered by an acoustic tube, namely the vocal tract (sometimes coupled with the nasal tract) from the point of the source onwards to the lips. Much modern work on speech synthesis depends to a greater or lesser extent upon this model.

Anticipating somewhat the discussion in a later section, modern approaches to synthesis can be classified along two partly-independent dimensions. The first dimension relates to the degree to which the approach has controllable parameters — is *parametric*. The approach may be highly parametric, with many independently controllable parameters: this is the typical situation with articulatory or formant-based systems discussed below. Or the approach may have few or no variable parameters: this is the case in some of the waveform-based approaches to concatenative synthesis that we shall discuss later on. The second dimension relates to how the parameters are derived: the parameters may be derived by rule — *rule-based* systems — or may be trained from speech corpora; the most common instance of the latter are concatenative systems, where the derived parameters take the form of stored and processed segments of real speech.

The two approaches to highly parametric synthesis are *articulatory* and *formant-based* approaches. Articulatory synthesis attempts to produce speech by first understanding how the vocal apparatus changes shape during speech production, then understanding the acoustic problem of how those movements translate into sounds. Formant synthesis does away with the first step and attempts instead to derive rules that directly control the acoustics, given the string of sounds to be synthesized. For *concatenative* systems, where the parameters are in the form of stored segments of speech, one of the main choices (as we shall see later on) is the size of the units that are used in the inventory: in most cases the minimal unit that is practical to use is a transition between two sounds, termed a diphone, and this is the unit that was used in the first concatenative systems.

Articulatory, formant, and concatenative approaches to synthesis have had a roughly equally long history. The earliest articulatory synthesizers include [Nakata and Mitsuoka, 1965, Henke, 1967, Coker, 1968]. and the earliest formant synthesizer was [Kelly and Gerstman, 1961]. Finally, the concept of diphones was discussed in [Peterson et al., 1958], with the first working diphone-based synthesizer being that of [Dixon and Maxey, 1968].

3.1.2 Text Analysis and Prosody

In the previous section we sketched the history of attempts to produce artificial speech given that one knows the sounds that one wants to synthesize: as we have seen, the investigation of this problem dates back about two hundred years. But synthesizing speech is only half of the problem of *text-to-speech synthesis*: a text-to-speech synthesizer must not only be able to synthesize speech from a set of sounds, but it must also be able to figure out an appropriate sequence of sounds – as well as other properties such as the appropriate pitch contour to use – given text. Work on *text analysis* and *prosodic modeling* in synthesis has a much shorter history.

One of the first systems for full text-to-speech conversion for any language — and certainly the best known — was the MITalk American English system, developed at the MIT Research Lab for Electronics during the 1970's [Allen et al., 1987]. MITalk was capable not only of pronouncing (most) ordinary words correctly, but it also dealt sensibly with punctuation, numbers, abbreviations and so forth. These problems were handled with a combination of rules and dictionaries: productive models of word formation based on morphemic analysis were included so that many words that were not explicitly listed in the dictionary could be handled. Part-of-speech analysis and a phrase-level parser provided input for the prosodic rules. (The actual synthesizer in MITalk was a formant synthesizer, a close kin of the well-known Klatt synthesizer, as well as the commercial DECTalk system.) The MITalk project was important not only for being a pioneering full text-to-speech system, but also for the fact that the text analysis components of the system were informed by linguistics — in particular the *generative linguistics* that had been under development for more than a decade at MIT. This helped set the tone for much subsequent work in synthesis, which has generally been well informed about linguistic matters.

The late 1980's saw a blossoming of full text-to-speech systems for numerous languages, as well as much further work on techniques for text analysis. Approaches that are broadly speaking based on hand-constructed rules have remained popular, in particular for the development of so-called *letter-to-sound* rules, which convert from letter strings into appropriate sound (more technically *phoneme*) sequences. However, there has been an increased interest in alternative approaches, including various partially-supervised or fully self-organizing statistical models. Two examples will suffice. First, several groups have investigated the use of neural nets in learning letter-to-sound correspondences given a training corpus of spelled words and their pronunciations. Second, various *corpus-based* approaches to disambiguation have been applied to the problem of deciding, for example, whether the string *1/20* should be read as *one twentieth*, or *January the twentieth.* Such approaches start with a corpus of examples tagged with the correct choice, and learn which kinds of elements in the context (e.g., grammatical properties of words in the immediate context, or nearby words) are the most reliable indicators of each sense.

Of course, it is not enough just to predict what phonemes to produce from a text: one must also determine how to "chunk" sentences into phrases, choose which words to emphasize, compute durations of the phonemes and the appropriate energy and intonation contours associated with the utterance. These different elements of speech are collectively referred to as *prosody*. Aspects of prosody in speech production have been investigated in various phonetic and linguistic traditions for centuries, but their investigation in the context of speech synthesis is, of course, much more recent. For example, according to [Klatt, 1987], the first implemented algorithm for determining a pitch contour was done by Ignatius Mattingly in 1966,

and was incorporated into the Holmes rule-based synthesizer. Intonation patterns were constructed using three basic intonational *tunes* ("falling", "rising", and "fall-rise") which were aligned with the final prominent syllables in clauses. A much later, but quite influential model was the "tone and target" model of [Pierrehumbert, 1980], where pitch contours were derived from combinations of abstract high and low tonal targets. This theoretical model formed the basis for the implemented model that was until recently used in the Bell Labs American English TTS system. Nowadays there are many different implemented models of intonation including a variety of statistically trained models.

One final point relates to multilinguality: not only are there more and more full text-to-speech systems available, but increasing numbers of them have been applied to more than one language. Of these, at least some systems are starting to emerge that are truly *multilingual* in the following restricted sense: a *multilingual text-to-speech system* is one that has been applied to more than one language, and at the same time, for each component of the system, uses the same tools and algorithms for each language. Thus a system that produces speech for English and French, but which has two completely distinct text-analysis modules for the two languages, is not multilingual in this sense. Modern multilingual systems include the Festival system, ETI-Eloquence, Infovox, and the Bell Labs TTS system.

3.2 Current Technology Review: What We Can Do Now?

In this section we describe in some detail the technical challenges involved in synthesizing speech and discuss past and current approaches to meeting these challenges. In addition to enumerating the challenges the point is made that solving the speech synthesis problem requires a multidisciplinary strategy that integrates both theoretical and applied knowledge from the fields of linguistics, computer science, signal processing and more. Following this section's coverage of the research areas, Section 4 recommends a program of research activities in areas that will drive improvements in the quality of speech synthesis. Finally, Section 5 provides the detailed recommendations of the report for a coordinated research program to address the complex, multidisciplinary challenges described here.

Speech synthesis is a problem that is typically broken up into two main sub-problems, as described in Section 3.1. The first phase of speech synthesis is the analysis or generation of input text and its conversion into a representation from which synthesis can then take place. Minimally such a representation will include information on the actual phonemes to be uttered, which words to accent (i.e., emphasize), and how to sensibly chunk up sentences into "information units" (technically termed *prosodic phrases*). Determining the appropriate phoneme sequence and prosodic structure (accents and phrases) are coupled, but they are often treated as separate problems, hence the separation in the discussion to follow. Inevitably even the best algorithms for text analysis and prosodic modeling will fail to produce the desired result in some instances. In such cases, and to facilitate communication with a language generation module, some standardized external markup language is needed to annotate the text in appropriate ways, so we also include here a short section on *markup* schemes. The second phase of speech generation involves synthesizing a waveform given the symbolic representation of the utterance. This stage can be divided into problems of predicting the continuous-valued acoustic parameters associated with the prosodic and segmental context defined by the symbol stream and using signal processing techniques to combine the different elements to obtain the final waveform. Again, the prosody and segmental components will be separated in the discussion, though

they are clearly related. Finally, since multi-modal communication is one important application of speech synthesis, we also include a section on synthesis of visual components of speech (i.e., "talking heads").

3.2.1 Text Analysis and Generation

Text Analysis. As we noted above, the first phase of text-to-speech conversion is the analysis of the input text, and its conversion into a representation from which synthesis can then take place. At the very least, one needs to have a transcription of all the words that occur in the text. Now, deriving the pronunciation of individual words involves much more than simply looking up the words in a dictionary: no dictionary contains all of the words that one might encounter in text, and besides ordinary words, one must also be able to pronounce abbreviations, number strings, and other kinds of material, which one would not generally expect to find in a dictionary. In addition, one must often choose between several alternative pronunciations for a given word.

An example will serve to illustrate the kind of problems faced by a text-analyzer:

```
Dr. Abate lives at 175 Park Dr.  He weighs 175 lb.  His hobbies in-
clude music.  He plays bass in a blues band and he also plays clar-
inet in the local symphony orchestra.  He also likes fishing:  last
week he caught a 20 lb bass.
```

Presumably few competent English readers would have difficulties in reading this short paragraph aloud, yet there are actually a number of problems that must be solved by a speech synthesis system in order for it to perform as a human reader would. Naturally it must pronounce words such as *weighs* or *hobbies* correctly, but these are relatively easy since there is really only one way to pronounce them. But other cases are not so easy. For instance:

- There are two periods in the first sentence, but only one of them marks the end of the sentence, the other one merely serving to mark the abbreviation, *Dr.*, for *doctor*.
- *Dr.* occurs twice in this paragraph. The first instance should be read as *doctor*, the second as *drive.*
- *Abate*, if it were an ordinary English word, would be pronounced as if it were spelled *a-BAIT*. But in this case it is a name of Italian origin: it should be pronounced *a-BAH-tee*.
- *lives*, in other contexts, could be pronounced to rhyme with *jives.* Here it is a verb and rhymes with *gives.*
- *175*, and other numbers, tend to be pronounced differently in some contexts (such as addresses) than they are in others. Many people would say *one seventy five Park Drive.*, but *one hundred and seventy five pounds.*
- *lb.* can be read as both *pound* and *pounds*: roughly speaking, it should be *pounds* if the preceding number is plural, and the phrase *X pounds* is not modifying a following noun (the second instance is *twenty* **pound** *bass*).

- *bass* is either a musical range or instrument (pronounced as *base*), or else a fish (rhyming with *mass*).

As we can see from this short example, text analysis poses numerous problems for a TTS system. The problems it includes can be classified into various subproblems, including, but not limited to: word/sentence segmentation, abbreviation expansion, numeral expansion, word pronunciation, and homograph disambiguation. (Additional problems of accentuation and prosodic phrasing will be discussed in the next section.) All of these problems can be thought of as having two components. The first component is the problem of *enumerating* the legal possibilities for a given input; the second consists of *disambiguation*, or the selection of the contextually appropriate analysis. A couple of examples will illustrate the point. Consider *Dr.*: a text-analysis system for English must know that this abbreviation can be expanded as *doctor* or *drive* (and in other ways too). Thus, it must be able to enumerate the possible expansions of this abbreviation and to disambiguate among them given the context. Text analysis methods for TTS have been reviewed in a number of places, including [Klatt, 1987, Dutoit, 1997, Sproat, 1998]. Here, we will only briefly describe the various approaches that have been taken to the more prominent problems, namely word pronunciation and homograph disambiguation.

The architecturally simplest approach to **word pronunciation** involves letter-to-sound rules. These are rules that map sequences of letters into sequences of sounds, along with other information such as stress placement. This approach is naturally best suited to languages like Spanish or Finnish where there is a relatively simple relation between orthography and phonology. For languages like English, however, it has long been recognized that for accurate word pronunciation one needs a pronouncing dictionary that (at the very least) includes words with irregular pronunciations.

But a simple list is generally not sufficient: words can be morphologically derived from other words (e.g., *giraffishness* can be derived from *giraffe*), and the set of morphological derivatives is open-ended. In various systems, standard techniques for morphological analysis are applied to morphologically complex words, and the pronunciation of the whole is then computed from the pronunciations of the known parts, applying appropriate phonological rules where necessary [Allen et al., 1987].

One class of words that often requires special treatment is personal names. In the United States, for example, many personal names are not of English origin, and special rules of pronunciation apply: a reader who knows that *Dutoit* is a French name, will also generally know that the "oi" is pronounced "wa" and that the final "t" is silent. In some cases morphological analysis can be applied to derive pronunciations for complex names from known morphological pieceparts: e.g., *Phillipson* from *Phillip* plus *son*. However this is not always the best approach, and other approaches have been tried, including reasoning by analogy [Golding, 1991, Dedina and Nusbaum, 1996]. For example, if we have the name *Califano* in our dictionary and know its pronunciation, then we can compute the pronunciation of a hypothetical name *Balifano* by noting that both names share the final substring *alifano*.

Finally, one approach to word pronunciation that has become popular recently is self-organizing methods. A typical approach involves training a neural net, or other statistical learning method, on a list of spelled words paired with their pronunciations. These methods have mostly been tried on English, where relevant resources such as online pronouncing dictionaries abound.

Text analysis problems naturally break down into two parts, namely enumeration and disambiguation. What we have just discussed in word pronunciation constitutes the enumeration part of the problem: the

methods previously described will return a set of pronunciations — though often a set with a single member. In cases where there is more than one viable candidate, one must have some method for choosing between them, a problem referred to as **homograph disambiguation**. One approach is described in [Yarowsky, 1997]. One starts with a training corpus containing tagged examples in context of each pronunciation of a homograph. Significant evidence from both narrow and wide context (e.g., n-gram word patterns containing the homograph in question or words that occur anywhere in the same sentence, which are strongly associated to one of the pronunciations) is used to construct a *decision list*: in the list, each piece of evidence is ordered according to its strength (log likelihood of each pronunciation given the evidence). A new instance of the homograph is then disambiguated by finding the strongest piece of evidence in the context in which the novel instance occurs.

How well do we do? A cursory reading of the literature on text analysis for TTS might convince one that the text analysis problem is almost solved. One finds that in many areas, results in the mid to high 90% range are quoted. But what do these numbers really mean? The first point to observe is that while the percentage of errors for any given subproblem may be small, there are a lot of subproblems, and small errors tend to accumulate. That means that for any reasonable sized text input, the chance that there will be at least one text analysis error is actually quite high. Furthermore the numbers given above are somewhat misleading. In some of these cases, the results are actually better than what the numbers suggest. Word pronunciation results are reported by *type*, meaning that if you go down a random dictionary of words and names, you expect to get around 95% correct. But TTS system developers naturally work hardest on making sure that the most common words are pronounced correctly, with the result that a 95% correct score by type will almost certainly translate into a higher score by token. On the other hand, word pronunciation rates are for "ordinary words": they typically do not include "unusual" strings such as URL's, mixed-case items (*WinNT*), radio station call letters (*KARL*), stock symbols (*AMZN*) and so forth. Often these kinds of strings require special treatment. One might object that these are really problems for a preprocessor and don't fall under the rubric of word pronunciation. But this is a quibble about terminology: the fact is that a TTS system must be able to pronounce such strings, and just because a system handles most ordinary words in a reasonable fashion is no guarantee that it will be able to deal reasonably with these kinds of elements. Finally, evaluations of homograph disambiguation include only the homographs being considered. A system may have models for a few tens of homographs, but one often finds many others that are not treated at all. For example, a system built to read the Wall Street Journal would not work as well for email.

So text analysis for TTS is not nearly as solved a problem as it would at first appear to be: to put it simply, the text analysis problem will only be solved when all strings in a text are pronounced correctly and emphasized correctly, and when all sentences are phrased and intoned correctly. We have not achieved this state yet. Furthermore, text analysis errors contribute significantly to the percepts of *unnaturalness* and *unpleasantness* of existing TTS systems.

Still, one might object that, while the text analysis problem for *general* text is not solved, there are many applications of TTS involving limited domains, where full text analysis capabilities are not required. This is true. However, one must understand that there are many current applications that are *not* limited: among these are e-mail reading (one of the most popular applications of TTS at present), aids for people with disabilities (e.g., page readers for the blind) and general access to information on the web. In none of

these applications can one limit the input to the TTS system. Furthermore, even seemingly limited-domain applications like "reverse-directory" (enter a telephone number, and the system reads the name and address associated with that number) have open-domain problems, since the vocabulary changes more frequently than is practical to update manually. In this particular case one must deal with personal and company names, various kind of standard and non-standard abbreviations, and so on, all of which can present difficulties for a TTS system.

Text Generation. In TTS systems, improved text analysis means that more of the structure, meaning, and function of a string will become apparent. In concept-to-speech (CTS) systems, on the other hand, the meaning and function is given as input, and both the string and its acoustic rendering must be computed. One starts with a data- or meaning-oriented representation of what is to be said, rather than a text string. CTS applications include human-computer dialog, machine translation, explanation generation, training, and report generation. As in TTS, the goal is to produce natural-sounding speech output.

Although the idea of CTS was introduced twenty years ago [Young and Fallside, 1979], there was relatively little activity for many years because of limitations in other aspects of speech and language technology. Recent successes in speech understanding have made it possible to envision human-computer dialogs and hence helped to increase interest in CTS.

CTS conversion combines natural language generation with speech synthesis. At first blush, this looks like a doubly-difficult task because: (1) concept-to-text generation is hard, and (2) text-to-speech synthesis is hard. However, the picture brightens when we consider that a language generator can pass very useful information, above and beyond text string, to a speech synthesizer. For example, concepts in a meaning representation can influence word pronunciation very accurately. A generation system that says "wind" already knows whether it is talking about "*wind* that blows" or "*wind* a clock".

Using concepts to determine pronunciation is but one example of a major theme in CTS research. More generally, research investigates how information produced during the language generation process can improve the quality of synthesized speech. One line of research studies the influence of information structure on prosody. For example, discourse information such as the given/new distinction has been shown to influence intonational prominence [Hirschberg, 1993] while reflection of theme/rheme in syntactic structure can drive the placement of contrastive stress [Steedman, 1996]. Both types of information have been used effectively in CTS systems [Davis and Hirschberg, 1988, Prevost, 1995]. While statistical algorithms have been developed to derive some of this information from text (in particular, given/new [Hirschberg, 1993]), the results are errorful (as is most text analysis); in contrast, this information is naturally produced and used during the course of language generation to determine sentence form and thus, CTS systems have more accurate information available to improve speech quality. Other research looks at differences in accuracy for prosody assignment based on part-of-speech tags, constituency structure, or semantic roles (e.g., agent, patient), when accurately determined by language generation or derived using text analysis [Pan and McKeown, 1998].

While CTS systems to date have exploited information that could conceivably be derived by text analysis techniques though with less accuracy, emerging research explores the potential of information that would be difficult, if not impossible, to derive from text. Semantic and pragmatic information that is routinely used, created, or represented during the process of language generation has the potential for even greater impact

on improvement of prosody because it would more directly correlate intonation with meaning. For example, in planning the content and organization of text, hierarchical discourse plans are typically constructed and rhetorical relations between plan elements labelled (e.g., as *example, contrast*, or *sequence*). While previous work [Hirschberg and Grosz, 1992] has explored the effect of discourse structure on intonation, it seems likely that function of one element in relation to another, such as exemplification, should also have an impact. Information such as the type of each semantic role and speaker intention is available in input to CTS systems and could also play a role in assigning prosodic features. In combination, these could be used to study the impact of affect on intonation. Does a speaker, for example, communicate surprising information (e.g., unexpected abnormal information in medical briefings or unexpected plays in a sports game) differently than the norm? Studies of the effect of a wide variety of information on synthesis is possible within a CTS context because it is already used to determine the language produced.

So far, we have examined the influence of language generation on synthesis. Yet, the fact that speech is the output medium also places demands on and provides opportunities for language generation research [Pan and McKeown, 1996]. Most work in language generation has been done for the written medium. How should syntactic structure and word choice change when speech is the output medium? We know, for example, that human spoken language is more colloquial and less grammatical in most situations than written language, but how much of that do we want to duplicate in computer generated speech? Can an explanation be shorter since the spoken version has prosody to help convey meaning? Finally, spoken language, an auditory medium, is often used in combination with other visual media. In these multimedia contexts, when determining content and form of its output, a spoken language generator must take into account the information that is presented in other media and its form. Again, this places new demands on the language generation process, affecting generation of references, selection of paraphrases, and organization of information in sentences [McKeown et al., 1997].

How well do we do? Research on CTS has only scratched the surface of the problem. The use of discourse information in CTS has demonstrated a marked improvement in the naturalness of synthesized speech and thus, illustrates the potential of this approach. However, use of the rich, accurate information produced by language generation for synthesis has been studied for a very limited set of features. Much more remains to be done.

3.2.2 Prosody Modeling

Prosody, which includes the phrase and accent structure of speech, aids the listener in interpreting utterances by grouping words into larger information units and drawing attention to specific lexical items. That is, prosody helps a listener to more quickly and easily understand the meaning of the incoming speech by making the structural organization of the sequence of words more transparent, where structural organization includes both the syntactic and semantic interpretation of a word sequence as well as the topic structure and speaker intentions. The acoustic cues to prosodic structure include modulation of fundamental frequency (F0), energy, relative timing of phonetic segments and pauses, and phonetic reduction or modification.

Prosody, which includes the phrase and accent structure of speech, is one of the least developed parts of existing speech synthesis systems, according to [Klatt, 1987, Collier, 1991] and many participants at the

workshop. Perhaps the most often cited reason for improving prosody prediction in synthesis is to improve "naturalness", but prosody also affects intelligibility. A domain-specific model of prosody can significantly improve the comprehension of synthesized speech [Silverman, 1993]. Though important, prosody has proved to be difficult to model and to evaluate, in part because the acoustic cues are associated with structures operating at different time scales (from segmental effects to paragraph or dialog structure to speaker emotion) and in part because the same syntactic interpretation of a sentence can be uttered with more than one acceptable prosodic pattern, reflecting a flexibility that may be associated with speaking style, dialect or other factors.

Like the overall synthesis problem, prosody control is often broken into two sub-problems: prediction of symbolic markers to indicate phrasing and emphasis and generation of continuous control parameters for use in speech synthesis. Both of these aspects require strong collaboration between the engineering and linguistics communities. The two sub-problems and current research efforts are described in the following sections.

Prediction of Symbolic Prosodic Markers. Different systems use different inventories of prosodic markers, but there is general consensus that labels are needed for indicating phrase structure and emphasis. Such labels may be used internally and/or as external controls (e.g., breaks and emphasis in the SABLE markup language [Sproat et al., 1998]).

Consider the following example to illustrate the problems of accentuation and phrase prediction:

> *I was talking with Joe last night. He plays bass in a blues band and he also plays clarinet in the local symphony orchestra. Do you play in a band?*

Several factors affect the placement of accents. In the first sentence, *Joe* would be accented if his name is first mentioned in this sentence, but *night* might be accented if the speakers have already been talking about him and the new information is when the speaker last talked to him. If the speaker wanted to especially emphasize the time, s/he might accent both *last* and *night.* Noun phrases such as *blues band* and *symphony orchestra* pose the problem of where to place the main emphasis. Many speakers would say *BLUES band* but *symphony ORCHESTRA*. Words may also be accented to indicate contrast, as in *you* (which contrasts with *Joe*) in the last sentence. Prosodic phrase placement is similarly difficult. The second sentence is long enough that one might expect a good reader to break the sentence into two phrases. The most natural place to do that is between *band* and *and.* However, prosodic phrases are often represented as a hierarchy, with major phrases being composed of a sequence of minor phrases. With a phrase hierarchy, one might put minor phrase boundaries before the prepositions in the second sentence. Finally, there is the issue of the particular type of intonation marker associated with accents and phrase boundaries. For example, one would generally put a falling tone at the phrase boundaries associated with the end of a declarative sentence, a rising tone at the end of a yes-no question, and yet another tone after the first instance of *band* to indicate continuation.

The simplest algorithms for **accentuation** involve making the accenting decision on the basis of broad lexical categories, or part of speech. Content words (nouns, verbs, adjectives, many adverbs) are typically accented, function words (auxiliary verbs, determiners, prepositions) are typically deaccented, and short function words (*to*, *a*, etc.) are deaccented even further or *cliticized.* Methods differ to some extent on how

finely they categorize words into lexical classes, and what exact use is made of that information. In English, and several other languages, accenting not only correlates with lexical category, but also depends upon semantic relations between words and the pragmatic function of words. One important set of constructions in English where part of speech alone cannot suffice are complex noun phrases — a noun preceded by one or more adjectival or nominal modifiers. One finds instances of such constructions where the accent is on the final word (*Park LANE*, *banana PIE*) but seemingly parallel constructions where it is on an earlier word (*PARK Street*, *BANANA bread*). Approaches to this problem typically involve some amount of lexical and crude semantic analysis [Monaghan, 1990, Sproat, 1994]. Accenting in English also appears to be sensitive to prosodic phrase structure in that speakers have a tendency to put accents early and late in a phrase [Shattuck-Hufnagel et al., 1994]. Finally, accenting in English is sensitive to properties of the discourse, including factors like givenness [Hirschberg, 1993] and contrast, as noted above.

As also noted above, another important problem is prediction of **prosodic phrasing**. Punctuation is a fairly reliable cue to a phrase break, assuming a text analysis component that can distinguish abbreviations from sentence boundaries, but there are many cases where phrase breaks are not marked with punctuation and vice versa. Other factors include syntactic structure, which is used to varying degrees by different systems, from simple part-of-speech labels of target and neighboring words [Wang and Hirschberg, 1992] to full syntactic parses [Bachenko and Fitzpatrick, 1990]. Phrase length appears to be a factor [Bachenko and Fitzpatrick, 1990, Ostendorf and Veilleux, 1994], and in addition, it has been hypothesized that focus plays a roll in determining phrase structure (e.g., *last night* forms a minor phrase in the example above when the time is to be strongly emphasized).

For both accent and phrase prediction, all the different factors can be integrated using a variety of models, including rule-based approaches using hand-written or automatically learned rules as well as statistical models. The structure of the model is an important research problem, particularly for automatic training, since there are innumerable syntactic/semantic/discourse contexts to consider as well as variations between speakers, moods and speaking styles. Furthermore, very little prosodically labeled data is currently available from which to learn rules or model parameters. However, text analysis is an equally important limiting factor, particularly with respect to discourse structure, since prosodic parameters are typically predicted from data output from text analysis.

How well do we do? As in text analysis problems, results for accent and phrase prediction of over 90% are reported, but again these results are misleading. First, typical methods for measuring phrasing prediction have a very high baseline: for some corpora, over 90% can be achieved simply by placing a boundary at sentence ends and no boundaries elsewhere. Furthermore, the accuracy is lower for the case where multiple levels of boundaries are predicted, as is needed to account for observed differences in acoustic cues. A problem with accentuation models is that they rarely predict the *strength* of accentuation. A word can be accented (or emphasized) to varying degrees, depending upon many factors including its semantic or pragmatic relation to other terms in the text and the reader's intentions. Variation of strength of accent under the reader's control is part of good reading style and part of what makes a good human reader sound interesting, and lack of such variation is an important reason why even the best synthesis systems sound monotonous. A problem with both accent and phrase prediction results is that they tend to be tuned to a specific style and generalization to different styles has not been demonstrated. As shown in [Ross and Ostendorf, 1996],

an algorithm that gives very good results in one case has performance worse than chance when used on data from a different domain. Finally, as argued in the text analysis case, errors in different components compound through multiple processes in a speech synthesizer, and the current error rates are simply not low enough.

Generation of Acoustic Control Parameters. There are two main acoustic parameters that most systems focus on: duration and fundamental frequency (F0). Additional acoustic cues to prosodic events include various energy measures (total energy, spectral tilt) and glottal source parameters. Since prediction of these control factors are less developed, we will focus the background review on duration and F0, leaving discussion of other factors to Section 4.

In natural speech, the **duration** of a phonetic segment is highly context dependent, and can vary by as much as a factor of ten. For synthetic speech to sound natural, these contextual effects must be understood and computed. This is far from easy, because the number of contextual factors is large (e.g., identities of surrounding phonemes, syllabic stress, word emphasis, effects of phrase boundaries, etc.), and these factors interact. For example, the lengthening effect of a voiced consonant on the duration of the preceding vowel is substantially larger when this consonant is followed by a phrase boundary. Methods used range from manually created knowledge-based rules ("if stressed, add 50 ms"), to applications of general purpose statistical methods (e.g., classification and regression trees [Riley, 1992], to methods that attempt to integrate knowledge in statistical estimation procedures (e.g., [van Santen, 1993]). Although current methods appear to produce perceptually acceptable durations, it is clear that future advances in other synthesis components will require solving several outstanding problems in duration research. These problems include modeling of local variations in natural speaking rate, and understanding non-uniform compression/expansion of phonetic segments.

Duration is relatively well understood compared to intonation, and there are many unresolved questions in **F0 contour generation**. Currently, there is no agreement on how to characterize a fundamental frequency (F0) contour – whether as a sequence of (time, frequency) pairs between which a smooth curve is drawn [O'Shaughnessy, 1979, Pierrehumbert, 1981, Hirst et al., 1991, Aubergé, 1993], as a sequence of quasi-linear movements [Collier, 1991, Taylor and Black, 1994], as a sequence of filtered target functions (e.g., pulses), [Hirose and Fujisaki, 1982, Anderson et al., 1984, van Santen and Möbius, 1997, Ross and Ostendorf, 1999], or a concatenation of local contours defined non-parametrically [Traber, 1992, Malfrére et al., 1998]. There is disagreement as to whether phenomena like the gradual decrease in peak height within a phrase should be represented by adding phrase-level and accent-level contours [Hirose and Fujisaki, 1982, van Santen and Möbius, 1997] vs. predicted based on local context [Pierrehumbert, 1981, Silverman, 1987]. The factors that are important for predicting pitch range and relative prominence of accented words are not well understood, so conservative choices are made that result in only small amounts of variation. Furthermore, there are disagreements about which aspects of intonation are discrete vs. gradient, what is the inventory for the discrete subset, and what is the size of the low-level units, as evidenced by the differences in prosodic control parameters for different speech synthesis systems. Finally, as in all aspects of synthesis, there are differences associated with the possible approaches: rule-based, e.g., [Collier, 1991, Silverman, 1987]; parametric but automatically trained, e.g., [Taylor and Black, 1994, Ross and Ostendorf, 1999]; and data-driven template learning, e.g., [Traber, 1992].

Given these basic disagreements, it is no surprise that our current ability to fit or predict natural fundamental frequency contours is quite limited. Several factors make this task inherently difficult, in essence having to do with our lack of understanding of the acoustic invariance of fundamental frequency contours that carry the same meaning. For example, within a given dialect, a speaker does not have much freedom on how to pronounce a given phoneme, and even duration is fairly strictly controlled; however, there is significant freedom (in English) in what intonational pattern to use. This makes it difficult to conduct well-controlled experiments and to estimate parameters. As a result of these factors, what sounds like the same intonational "meaning" may vary significantly in terms of the measured F0 contour. In summary, the strategy for acoustic intonation modeling remains an open question, at least from the point of view of the specific requirements of speech synthesis.

How well do we do? There are no good quantitative measures to indicate performance of current algorithms. While the mean-squared-error measure is commonly used, it does not take into account the perceptual significance or insignificance of errors in particular contexts, and it requires prohibitively expensive hand-correction to have reliable target values. The best assessment is in perceptual experiments, but these are complicated by the distortion introduced by waveform generation techniques. Without being quantitative, we can say that current methods are capable of generating intonation that is appropriate for a simple reading style and short passages, but the resulting intonation and duration patterns are not reflective of the variability observed in natural speech and longer passages sound monotonous. Prosody modeling is probably the one area that most workshop participants agree is the weakest link in the synthesis chain right now.

3.2.3 Markup

Even with the best text analysis capabilities, we are surely not going to be able to get everything right all the time. If nothing else, the decision that a synthesis system makes on how to render a particular sentence, though reasonable, may not correspond to what the user of the system wants. (Compare the case of an actor reading lines in a play, and being corrected by the director.) This implies the need for synthesis markup schemes that allow the user to control the system's behavior, and such schemes should be *standardized* so that developers and users can switch between synthesis systems and have reasonable confidence that the behavior will be the same.

A number of proposals for standardization are on the table. Among the more serious alternatives are the Java Speech Markup Language, W3C's Cascaded Style Sheets, and SABLE [Sproat et al., 1998]. Each of these standards is potentially useful to the research community because they provide a common way to mark and exchange textual data. Conversely, there is room in speech synthesis markup for research on how to concisely and reliably mark salient features in a way that enables speech synthesizers to produce better quality speech. This research field is only a couple of years old and are relatively immature in comparison to other areas of the speech synthesis problem.

3.2.4 Waveform Generation

Waveform generation is the process of using outputs of the linguistic text analysis and prosody models to generate an intelligible, natural-sounding speech waveform. This is an "ill-posed problem" in the sense that the linguistic feature input specifies only a tiny fraction of the information needed to specify the waveform. The rest must come from basic knowledge and models of a particular talker's speech based on speech production and perception principles and/or examples found in a speech corpus.

Designing a waveform generation system involves specifying a speech model that can faithfully represent any speech sound, and specifying sets of model parameters to synthesize a particular utterance. Traditionally, speech models have been lumped into three broad classes – articulatory models, formant models, and concatenative models – described below. Articulatory and formant synthesis are both considered parametric methods, which have the advantage of a compact representation for a wide range of voices. Concatenative techniques are less flexible, but tend to be less expensive and are a more natural approach for mimicing a particular voice.

Articulatory Methods. In articulatory synthesis, as it is performed today, a geometric model of the vocal tract with its articulators, such as the tongue body, is written into software. The input parameters to an articulatory synthesizer used to create speech include the positions of the model articulators, such as the vertical displacement of the upper lip, and these parameters are specified as trajectories through time. There may be other input parameters that are more aerodynamic in nature, such as subglottal pressure. Physical models of aerodynamic processes and of waveform propagation in a tube are used to convert the input parameters of the synthesizer into sound. Two examples of such synthesizers are the one that was initially designed by [Mermelstein, 1973] and modified at Haskins Laboratories [Rubin et al., 1981] and another designed by [Maeda, 1982].

One of the major impediments to the use of articulatory synthesis in creating natural sounding speech has been a lack of knowledge of the articulatory movement patterns. Further, there is a problem in identifying the articulatory degrees of freedom which are most salient to the production and propagation of sound. It is necessary to know the salient components of articulation because there are simply too many degrees of freedom to hope to make a practical articulatory synthesizer without such knowledge.

Another significant challenge in articulatory synthesis has been modeling of flow-induced sound generation. Presently, wave propagation is computed from the linear wave equation. This ignores the effects of fluid dynamics as expressed in the Navier-Stokes equations. Such aerodynamic effects, particularly at the glottis, can contribute significantly to naturalness. As a result, these effects have been the focus of much recent study (e.g., [Hirschberg, 1992, Pelorson et al., 1980, Pelorson et al., 1994, Shadle et al., 1999]). Similarly, motion of the vocal folds due to fluid flow requires further examination.

Formant Synthesis. The other major parametric synthesis approach has been formant synthesis, some instances of which include the well-known work of John Holmes [Holmes, 1973] and Dennis Klatt [Klatt, 1980]. With formant synthesis there is no longer the problem of translating articulatory trajectories into acoustic parameter trajectories, because the input parameters themselves are acoustic parameters, such as formant frequencies. The formant synthesizer generates a waveform from the formant values and related parameter

values.

Although certain isolated phonetic units can be characterized almost solely by the dynamics of their formant frequencies, the formant trajectories in natural speech are heavily influenced by context. For this reason, the rules in early formant synthesis systems were often fairly complex. However, research into formant timing and the perceptual relevance of specific spectrographic patterns has led to improved models (such as the phone-and-transition model [Hertz, 1991, Hertz and Huffman, 1992]), for formulating rules that predict formant patterns. These models have not only resulted in simpler and more general rules for predicting formant trajectories in a particular language, but also in the expression of a wide range of language-universal patterns.

The rules for formant synthesis are generally derived through an iterative process of rule formulation in accordance with the underlying model, evaluations of the synthetic output through visual and auditory comparisons with natural speech, and refinement of the rules based on the results of these comparisons. This process has been expedited by special interactive rule development tools, such as the Delta System [Hertz, 1990, Hertz and Zsiga, 1995], which allow for straightforward expression and testing of linguistic and phonetic hypotheses. As a result of such improved models and tools, the rules for predicting the relevant acoustic patterns in a given language can now be developed by a trained linguist in a matter of months.

Most rule-based synthesis systems use a Klatt-style formant synthesizer [Klatt, 1980] to generate the waveform from the acoustic parameter values produced by the rules. Copy synthesis, in which synthesizer parameter values are tuned by hand, has shown that the Klatt synthesizer is capable of producing very high-quality speech. However, some researchers have reported difficulty capturing the voice quality of particular speakers. One basic question that needs to be addressed is whether improvements can be made at the level of the synthesizer itself that will result in the ability to better capture the details of a specific voice.

The Klatt synthesizer includes approximately forty input parameters. However, high-quality synthesis can be achieved by controlling fewer than half of these parameters and setting the remainder to default values that remain constant over the course of an utterance, or a specific voice. For any given type of segment, only a handful of parameters needs to be manipulated. Some researchers have also attempted to reduce the number of synthesizer parameters by building on ideas from articulatory synthesis and computing the covariation of acoustic output variables that result from a single articulatory gesture. (For instance, constricting the glottis from a modal voice posture will result in a decrease in both spectral tilt and voice amplitude.) Quasi-articulatory synthesis has been developed to exploit the potential advantages of articulatory synthesis, while avoiding the current deficiencies in our knowledge of articulation and its relation to acoustic output. The quasi-articulatory HLsyn [Stevens and Bickley, 1991] has parameters for upper articulator, nasal, and glottal constrictions so that noise and voice sources as well as nasalization can be modeled in an articulatory manner. However, acoustic parameters such as formant frequencies are included among the input parameters in place of parameters describe the shape of the tongue and lips.

There are a number of areas, then, for further research in formant-based synthesis. These include (1) improvements at the level of the synthesizer itself, (2) determination of the optimal synthesizer parameters to be controlled, and (3) continued development of better methods for predicting the synthesizer parameter values. Also, improved methods for extracting voice quality and noise-related parameters from the speech waveform (see, e.g., [Hanson, 1997]) will facilitate work in these areas. Collectively, the improvements made through further research of this kind should result in increasingly natural-sounding speech output.

Concatenative Synthesis. This method relies on extracting model parameters from speech data (or often just storing the raw waveforms) and concatenating these to create new utterances. Concatenative synthesis is currently the most popular technique in commercial and research TTS systems. This approach holds the promise of representing ill-understood fine detail in the voice and replacing hand-optimization of rules with a data-driven approach to waveform generation. However, even the best concatenative synthesis algorithms still leave much to be desired in terms of speech output quality.

There are two main problems in concatenative synthesis: unit selection and waveform modification. Unit selection involves defining the inventory of units as well as selecting the appropriate unit for a given phonetic (and prosodic) context. Waveform modification is the process of combining the waveform pieces and modifying these to have the desired duration and intonation patterns, and methods differ in terms of the signal model.

Most concatenative algorithms to date have relied on tables of diphones (the transition from the middle of one phone to another) or other such units for solving the **unit selection** problem. Implicit in this approach is the idealized assumption that the units can fully characterize coarticulation in all contexts. This is of course not always true in natural speech, and as a result diphone synthesis is often described as sounding "over-articulated." To compensate for the most significant deviations from this assumption, other context-dependent units must be added to the diphone table by the developer (e.g., to realize aspirated vs. unaspirated /t/ for the same diphone in different syllable positions). An alternative approach is to use sub-phonetic units that define potential splice points and then dynamically search for the sequence of such units that minimizes context mismatch and concatenation costs. In this way, units with size varying from a fraction of a phone to several words can be used together in a dynamic, elegant way. Approaches to this problem draw heavily on the statistical modeling techniques developed in the speech recognition field over the past several years (e.g., HMM model clustering [Nock et al., 1997]). Important aspects of this approach are clustering for defining the unit inventory and definition of the unit selection cost function, and research challenges remain in both areas.

The minimal requirement for **waveform modification models** is independent control of the fundamental frequency (F0) contour and segmental duration of the waveform. In addition to this, other modification controls are desirable: dimensions such as voice quality characteristics (breathiness, creak, etc.), correlates of emotional stress, phonetic reduction, and voice identity could also be altered to expand the range of synthesized effects realizable from a given database. There are several classes of speech models that allow varying degrees of control over the different properties, including: linear predictive coding, based on a source-filter model using a synthetic source [Makhoul, 1975, Rabiner and Schafer, 1978]; pitch-synchronous time-domain models that use copying, shifting and deleting of pitch-period-sized waveform samples (PSOLA [Moulines and Charpentier, 1990], MBROLA [Dutoit and Leich, 1993]); and sinusoidal models that represent the speech waveform as a sum of time-varying sine waves [McAulay and Quatieri, 1986, George and Smith, 1997, Macon, 1996]. Tradeoffs between flexibility, fidelity, and automatic analysis capability exist with each. Although each of these approaches has its own merits, none have yet been proven to simultaneously satisfy all the desired properties of a TTS waveform model.

How well do we do? It is difficult to quantify performance of waveform generators. Copy synthesis techniques can generate speech of extremely high quality for articulatory, quasi-articulatory and formant

synthesis, but this quality cannot yet be achieved using automatic algorithms. Assessments of word intelligibility in isolated sentences gave over 90% accuracy over a decade ago, though performance in noisy conditions degraded substantially, cf. [Klatt, 1987]. However, sometimes those systems scoring highest in standard isolated word or sentence intelligibility tests will be rated lower than competing systems in perceived naturalness. Furthermore, even the highest scoring systems can be improved substantially (in the sense of increased transcription accuracy) with very simple modifications to the prosodic cues in a telephone-based system reading names and addresses [Silverman, 1993]. It is safe to say that there is no text-to-speech system available today that would be mistaken for a human.

3.2.5 Synthesis of Visual Components of Speech

To date, most of the research and development in speech synthesis has been limited to auditory synthesis. It is well-known, however, that information from the talker's face can provide valuable information about the message particularly if the auditory signal is noisy or if the hearing of the listener is limited. A systematic series of research programs have demonstrated that bimodal (auditory/visual) speech is more informative than just auditory speech [Massaro, 1998]. Synthetic visible speech not only can provide additional information about the speech input, it can make the speech sound more natural. When the AT&T synthesis was combined with the Perceptual Science Laboratory's talking head, Baldi [Massaro, 1998], it was played for their vice president. He remarked that they had made great strides in improving the quality of their synthesis. The researchers replied, "Unfortunately not. Close your eyes." He did so and agreed, "Oh, you're right." This example illustrates how visible speech can improve the naturalness of speech synthesis.

Two general classes of facial animation techniques — physically based (PB) and terminal analogue (TA) — have been described [Massaro, 1998]. The PB class involves the use of physical models of the structure and function of the human face. The typical approach here is to use multi-layer tissue models with quasi-muscular controls to change the shape of the face. For the TA class, the goal has been to arrive at the terminal end of the process (a face surface) without resorting to physically based constructs. The typical approach used to achieve this end is to employ a polygon surface with keyframe or parametric controls.

Although there are hundreds of muscles that control the face, only a small subset of these muscles control the speech articulators and thus this technique might be feasible for visible speech synthesis. In fact, one motivation cited for the use of physically based rather than geometric models is that the physiological controls might reduce the number of degrees of freedom required. There is also a potential advantage in the intrinsic simultaneity of visual and auditory synthesis based on the same physical structure.

Under the PB approach, computer animated human faces are made by constructing a computational model for the skin surface, muscle and bone structures of the face [Platt and Badler, 1981, Waters, 1987, Terzoupoulous and Waters, 1990]. At the foundation of this type of model is an approximation of the skull and jaw including the jaw pivot. Simulated muscle tissues and their insertions are placed over the skull. This requires complex elastic models for the compressible tissues. A covering surface layer then changes shape according to the underlying structures. The dynamic information for such a model is defined by a set of contraction-relaxation muscle commands. Many of the PB systems use Ekman and Friesen's [Ekman and Friesen, 1977] "Facial Action Coding System" (FACS) to control the facial model. These codes are based on about 50 facial actions ("action units" or AU's) defined as combinations of muscle

actions. Recently, Pelachaud, Badler, and Steedman [Pelachaud et al., 1996] have extended and updated Platt's [Platt, 1985] seminal PB model, adding manually synchronized natural auditory speech as well as the display of paralinguistic information. A related project [Cassell et al., 1994] incorporated synchronized speech synthesis and automatic gesture production by two automata negotiating a bank transaction.

Some researchers have attempted to derive the muscle commands for driving a PB face from measurements of human muscle actions. For example, Vatikiotis-Bateson, Munhall, Hirayama, Lee, and Terzopolis [Vatikiotis-Bateson et al., 1996] have embarked on a project to drive a PB face utilizing information obtained through neural-net analysis of EMG and facial motion measurements. So far the project has yielded mixed results. A project of more immediate application from the PB school is DECface [Waters and Levergood, 1993]. In order to achieve real-time performance (15 frames per second with texture mapping) on a DEC Alpha workstation, this approach compresses Waters' 3D tissue model to a simple 2D mesh. A set of 55 "viseme" mouth shape key-positions is utilized with a physically motivated nonlinear interpolation algorithm used between these positions. Only this nonlinearity and the history of the workers qualifies the approach as physically based. It might also be viewed as more properly fitting in the TA class, which we now discuss.

In a seminal development [Parke, 1972, Parke, 1974, Parke, 1975, Parke, 1982, Parke, 1991], Parke modeled the facial surface as a polyhedral object composed of about 900 small surfaces arranged in 3D, joined together at the edges and smooth shaded. In his original work [Parke, 1972], key-frames were used to change the shape of the face, but in subsequent work the face was animated by altering the location of various points in the grid under the control of 50 parameters. About 10 of these parameters were used for speech animation, such as jaw rotation, mouth width, lip protrusion, and lower lip "f" tuck. Parke [Parke, 1974] selected and refined the control parameters used for several demonstration sentences by studying his own articulation frame by frame and estimating the control parameter values.

One advantage of the polygon topology strategy is that calculations of the changing surface shapes in the polygon models can be carried out much faster than those for the muscle and tissue simulations. It also may be easier to achieve the desired facial shapes directly rather than in terms of the constituent muscle actions. The animation in the Perceptual Science Laboratory [Cohen and Massaro, 1993, Cohen and Massaro, 1994, Cohen et al., 1996] is a descendant of the Parke software and his particular 3-D talking head. The modifications have included additional and modified control parameters, a tongue (which was lacking in Parke's model), a new visual speech synthesis control strategy with a method for coarticulation, controls for paralinguistic information and affect in the face, text-to-speech synthesis, and bimodal (auditory/visual) synthesis. Most of the current parameters move points on the face by geometric functions as rotation (e.g., jaw rotation) or translation of the points in one or more dimensions (e.g., lower and upper lip height, mouth widening). Other parameters work by interpolating between two alternate faces. Many of the face shape parameters such as cheek shape, neck shape and forehead shape, as well as some affect parameters such as smiling, employ this strategy. Paralleling much of the work in auditory speech synthesis, phonemes are used as the basic unit of animated speech synthesis. Phoneme units are attractive because of their relatively small number. Of course, because of coarticulation, blending a sequence of phonemes requires modifications of each phoneme's realization as a function of its surrounding phonemes.

In contrast to these 3D models, a number of scientists have used 2D vector models for perceptual studies [Montgomery and Soo Hoo, 1982, Brooke, 1989, Brooke, 1992, Brooke and Summerfield, 1983]. Individu-

al key-points define vector-based shapes on the face such as the mouth outline. These controlling key-points move on the basis of articulatory trajectories measured from human speech. Typical of this approach is [Summerfield et al., 1989], which uses faces with and without teeth to examine lipreading performance.

There are several other approaches to the development of a talking head that are actively being pursued by other researchers. Researchers and developers have used morphing as a technique for animation [Hallgren and Lyberg, 1998]. The idea is that video clips of speech segments are stored and chosen to represent some utterance. However, this technique is data intensive and more limited in the amount of control that can be provided. As an example, a parametric synthesis routine can generate emotion independent of speech content but this cannot be easily done with morphing techniques. Another company has developed a system called video rewrite in which morphed images are pasted on to an existing video clip to create new messages [Bregler et al., 1997]. This method would have only more specialized applications in that only the mouth area changes.

How well do we do? A central but somewhat unique aspect of research on visible speech animation is the empirical evaluation of the visible speech synthesis, which is carried out hand-in-hand with its development [Massaro, 1998]. These experiments are aimed at evaluating the realism of their speech synthesis relative to natural speech. Realism of the visible speech is measured in terms of its intelligibility to members of the linguistic community. The goal of this research is to learn how the synthetic visual talker falls short of natural talkers and to modify the synthesis accordingly to bring it more in line with natural visible speech. Successive experiments, data analyses of the confusion matrices, and modifications of the synthetic speech based on these analyses have led to a significant improvement in the quality of the visible speech synthesis. The overall viseme (visually distinct speech segment class) accuracy across three studies improved significantly. The average deficit relative to natural speech was .222, .179, and .106, across the three successive experiments [Massaro, 1998]. We propose that evaluation should be a necessary ingredient of funded research in visible as well as audible speech synthesis.

3.3 Assessment Methodologies

Evaluation must play a prominent role in any research program in speech synthesis: simply put, we need to know how we are doing, and which methods (be they short-term “tweaks” of current methodologies, or fundamentally new approaches) are likely to be most beneficial. As we shall argue, while there have been many proposals for evaluation methodologies, there is little general agreement on which evaluation methods are best. Indeed, appropriate evaluation methodologies are an open research area. This section summarizes the problems in evaluation of speech synthesis systems, and makes the following key points:

1. Speech synthesis evaluation is inherently complicated, more so than evaluation of speech recognition and speech coding systems.

2. Few, if any, research institutes exist that perform extensive multi-system evaluations whose results are publicized.

3. Few, if any, research institutes perform research with as goal the improvement of synthesis evaluation methods.

4. As a result, no information is publicly available about the comparative merits of current synthesis systems.

In Section 4, specific recommendations are made for a program that addresses the limitations of the current evaluation techniques.

Synthesis evaluation serves two groups of people. First, it serves the consumer of synthesis systems. Here, the key issue is identifying the system that best suits the needs of the consumer. Second, it serves the research and technology community. The basic issues here are measuring progress in general, and discovering which approaches to a particular synthesizer component seem most promising. It is important to keep in mind that the evaluation needs of these two groups of people need not be the same.

We propose that the needs of neither group are currently served well, which raises the following questions: What are the underlying causes? How harmful is this situation for scientific progress and commercial success? How do we address the problem?

The Current State in Speech Synthesis Evaluation. There is no lack of evaluation methods. In a recent overview [van Bezooijen and van Heuven, 1997], more than twenty methods were reviewed. To give the reader a flavor of what these methods are about, a few examples are given.

In the Diagnostic Rhyme Test [Voiers et al., 1972] listeners hear a single consonant-vowel-consonant word (e.g., "pen"). Subsequently, they have to decide whether they heard "pen" or "ten". In the Semantically Unpredictable Sentences (SUS) paradigm [Benoit et al., 1996] listeners hear a sentence such as "green ideas eat clouds" and have to transcribe it. Next, in a test by van Bezooijen and Jongenburger, listeners had to rate pleasantness on a 1-10 scale. And, finally, in a test of the grapheme-to-phoneme components of speech synthesizers, phonemic output was manually checked for errors by linguists and lexicographers [Pols, 1992].

Jointly, these tests cover a wide range of evaluation needs. Why is such a wide range needed? First, synthesis systems, in particular TTS systems are multi-component systems, and the relative performance of these components varies across systems due to differences between the responsible groups in terms of resources, priorities, and innovativeness. As a result, different systems excel in different ways. Second, component performance depends critically on text characteristics. For example, including rare words taxes the pronunciation capability of a system, but not prosodic or signal processing capabilities. Likewise, including text whose pronunciation involves rare triphones is primarily a challenge for the synthesis component (if concatenative). Third, the perception of synthetic speech is highly multidimensional. One dimension of perception is the intelligibility vs. naturalness dichotomy. There exist systems that sound quite pleasant but are hard to understand, and vice versa. The naturalness dimension is itself multidimensional; for example, how do we measure a system that produces a click-free yet gravelly voice competing with a system that has a smooth voice but occasional clicks.

The relative importance of the various dimensions of evaluation can be different across application domains or because of other extrinsic factors. In some low-redundancy applications, speech quality is not important so long as intelligibility is sufficient. In other applications, the content of the message is reasonably predictable, which then puts a premium on how pleasant the voice sounds.

Clearly, evaluation of synthesizers is more complicated than evaluation of speech recognition or speech coding systems. (As we shall note in Section 5.1 the classical metric for evaluating speech recognition

systems has been word-error rate, which is easy to measure; and the Diagnostic Rhyme Test is often a quite satisfactory way to evaluate coders.) Yet, for several reasons the needs of customers and scientists are not being served by the currently available tests, the ways they are used, and the ways their results are published. The textual materials are usually atypical, and it is difficult to judge to what degree results obtained with such materials either generalize to real-world applications or provide useful information about a TTS component. For example, the Diagnostic Rhyme Test does not measure performance for vowels, for consonants at the end of syllables, or for unstressed syllables in polysyllabic words. In addition, many test texts are publicly known (e.g., the SUS paradigm), which allows TTS developers to fine-tune their systems.

There is no consensus in the speech synthesis community as to which if any of these tests should be included in an evaluation standard. While many of the tests are methodologically sound, the truth of the matter is that most of them are developed either internally by research institutes, or as part of a small-scale collaborative effort, to serve an ad hoc purpose. Few were developed with standards in mind (e.g., the SUS). And currently there are no research institutes that invest in evaluation research. Few of these methods have been applied to a broad selection of systems with subsequently published results. A key cause of that is that once a method has been researched and developed application of the method to a large set of systems and publishing the results is unexciting, expensive, requires overcoming obstacles (e.g., system installation is often hard), and receives little interest from scientific journal editorial boards. On the other hand, speech technology trade journals regularly perform a disservice by reporting in often glowing terms on vendor-prepared system demonstrations: those of us who work on text-to-speech systems know how easy it is to prepare unrepresentative demonstrations that can be quite compelling.

Harmful Effects of Poor or Absent Evaluation. Currently, the speech synthesizer user has few options that allow making an informed buying decision. The information simply is not there, and conducting in-house studies (such as Nynex performed [Silverman et al., 1990, Basson et al., 1991]) is too expensive for most. One point of light is the TTS Comparison Website at the Linguistic Data Consortium, which allows listeners to automatically select text from large text corpora, and generate speech files from a number of TTS systems for side-by-side comparisons. This alleviates the need for local system installation and text generation. But not all systems participate in the Website yet (although the number is growing). In addition, performing actual assessment using these conveniently generated files still requires resources many individuals or small companies do not have. Furthermore, the Website is more of a service to consumers than to scientists, since it does not provide for the component-level evaluation that is necessary for advancing the state of the art.

Many scientific groups conduct in-house evaluations. Typically, these evaluations are targeted at specific questions of no general interest, and for that reason alone are not published. In fact, they usually do not compare different systems but different variants of one particular component of their own system. However, scientists also need to compare performance of their system with other systems, but because this typically requires more resources than they have, collaborative efforts with published results are needed.

One example of such collaboration is presented by a project involving French language TTS systems in France, Switzerland, Canada, and Belgium, focusing entirely on the pronunciation module. Unfortunately, French language systems in countries other than these four were not invited and the results were kept confidential. Also, the (European) funding for projects of this nature is limited. As a result, currently and in the

past scientists can have little hope for such collaborative efforts to help them in their evaluation needs.

An interesting example of a more open, yet relatively low-cost, collaborative approach is the following. At the Third European Speech Communications Association (ESCA) Workshop on Speech Synthesis at Jenolan Caves, Australia, the first International TTS Evaluation was conducted [van Santen et al., 1998]. This evaluation consisted of three formal test procedures involving SUS, newspaper, and telephone directory listing materials. Participants agreed that they had obtained valuable information about the relative strengths and weaknesses of their system and of various approaches to TTS components. However, because the listeners were the participating scientists themselves, results cannot be considered scientifically valid and hence publishable. Furthermore, by prior agreement, the results of this evaluation must not be published: therefore, while the results are extremely useful to the many speech synthesis professionals who participated in the workshop, they are unfortunately not accessible to others, such as government funding agencies and commercial developers, who have an interest in the current status of TTS systems.

3.4 Commercial Systems

Speech synthesizers have been commercially available since the mid-80s. In the late 90's a range of companies develop and sell speech synthesizers either through direct channels or as OEM (Other Equipment Manufacturer) suppliers. These companies include (but are not limited to): AcuVoice, Inc., Apple Computer, Inc., AT&T, Bellcore, British Telecom, Compaq (Digital), Elan Informatique, Eloquent Technology, Inc., Entropic Inc., E-Speech Corporation, First Byte, IBM Corporation, Infovox, Lernout & Hauspie, Lucent Technologies, Microsoft Corporation, RC Systems, Sensimetrics Corporation, SoftVoice, Inc., and Talktronics, Inc. In addition to those companies, many other companies license and resell speech synthesis products. For instance, some manufacturers of sound cards for personal computers and some speech recognition vendors bundle speech synthesis software with their main products.

The speech synthesis marketplace can be divided by the deployment space: telecommunications, personal computing and embedded systems. The following sections outline the market presence of speech synthesis in each of these domains.

In preparing this report we sought out market data for inclusion but were unsuccessful. From informal communication we infer that the speech synthesis market is substantially smaller than the speech recognition market which is estimated at several hundred million dollars.

3.4.1 Telecommunications

In the telecommunications industry, speech synthesis is currently used for information provision in a selective set of application domains. The dominant forms of speech output in telecommunication applications are pre-recorded speech and *concatenated voice response* systems. In both cases professional actors and speakers are used to record messages for later playback. With pre-recorded speech a recorded utterance is played back exactly as recorded and so is only applicable when the complete set of recorded utterances can be determined in advance.

Concatenated voice response is a technique in which utterances are produced by joining pre-recorded words and phrases to produce new utterances on demand. This technique is very widely used in telecommunication applications. While this technology has some parallels to the concatenative speech synthesis

systems described elsewhere in this report, the technology is much simpler, more restrictive and is not generally thought of as speech synthesis. There is, however, a likelihood that techniques developed for general speech synthesis could be applied to concatenated voice response to increase its flexibility.

In applications in which it is not possible to determine all utterance structures in advance, in which the vocabulary is so large as to prohibit a complete recording, or in which cost is the driving factor, speech synthesis is used. Currently, this is a minority of the telecommunications applications. One instance is the use of speech synthesis for automated reverse directory assistance systems which have intrinsically large vocabulary that includes all the surnames in the US. For example, Bellcore's ORATOR speech synthesizer is deployed for the reverse telephone directory services of Ameritech and Bell Atlantic.

3.4.2 Embedded Computing

A number of speech synthesis vendors have produced compact speech synthesizers for the embedded computing market. However, as with the telecommunications industry, most of the speech produced automatically by small electronic and computing devices is either pre-recorded speech, or large concatenated chunks of speech because of its preferable voice quality. Nevertheless, there are limited circumstances in which speech synthesis is required and is in use. One instance is the provision of spoken directions in car navigation systems. In this application the vocabulary is large (all street and town names in a set of states) and audible instructions are desirable since a driver should not divert his or her attention from looking at the road.

Because of the extreme constraints on CPU power, memory and physical space, the speech synthesizers used in embedded technology have typically been simpler than their desktop counterparts, with a resulting reduction in output quality. The systems are typically formant-based (more compact than concatenative systems), and the text and prosodic processing capabilities are limited or non-existent. As the computing resources in embedded systems continue to increase, more advanced functionality will be integrated into embedded speech synthesizers to improve quality. However, even with a rapid growth of computing power, it will be many years before the more compute-intensive synthesis technologies, such as large-database unit concatenation, can be applied to embedded devices. Therefore, ongoing research on compact technologies, such as formant synthesis, remain important to this application domain.

The key advantage of speech synthesis in this domain is compactness. A speaker or headphone jack can be substantially smaller than using a screen for information output. Thus, improvements in speech synthesis quality will enable a substantial class of personal devices and targeted computing modules that would not be possible with other information output mechanisms. As with the other domains of speech synthesis use, the limited acceptability of speech synthesis quality is the limiting factor.

3.4.3 Personal Computing

Most of the companies producing speech synthesis sell into the personal computer marketplace. Although there are direct sales (shrink-wrap and internet distribution), speech synthesis products are mostly delivered to users as software bundled with sound cards and dictation software.

An important segment of the speech synthesis market for personal computing is assistive technologies. The Trace ResourceBook (1996-1997 Edition) lists over 300 assistive technology products that incorpo-

rate or can utilize synthesized or digitized speech output. Several important product categories use speech synthesis to improve access to information, applications and services for users with a range of physical limitations. Key categories include voice browsers and web browsers with speech output (e.g., Productivity Works pwWebSpeak), screen readers (e.g., JAWS, outSPOKEN), augmentative technology that links with accessibility APIs on computing platforms (e.g., Sun's Java Accessibility API, Microsoft Active Accessibility), and book readers combining scanners with speech synthesis (e.g., Arkenstone).

Despite the potential utility of speech synthesis as an assistive technology, its use is not as widespread as it could be. Of the millions of potential users, only a fraction use speech synthesis. The first inhibiting factor is the price of PCs. The second factor has been the cost of speech synthesizers which, in hardware form, have cost hundreds of dollars. With the substantial drop in PC pricing and the availability of software-based synthesis, the dominant factor is becoming the quality of speech synthesis. Even with a compelling need for speech synthesis, many potential users of assistive technology are deterred by synthetic voice quality. Any improvement in speech synthesis quality that arise from increased research will affect deployment to the accessibility market. This is not, however, a market in which the vendors have a sufficient profit margin to drive the pure research needed to improve speech synthesis.

A quick tour of the history of one product, DECtalk, is instructive on the progression of the speech synthesis market. DECtalk is probably the best known and most recognizable speech synthesizer: e.g., it is the voice of the physicist Stephen Hawking. Developed at Digital's Cambridge Research Lab in Massachusetts with technology from MIT's MITalk project and the Klatt synthesizer (as described in Section 3.1), DECtalk was representative of the first generation of speech synthesizers. It was:

- Hardware-based: a box connected to its host by a serial connection
- Based on formant synthesis
- Expensive at around $4000 per unit.

Along with most commercial speech synthesizers, the evolution of computing technology has had a significant impact upon DECtalk, which is now available as a software product at a much lower cost.

Importantly, the increased computing power of personal computers and the near ubiquitous availability of audio output hardware on those PCs allows speech synthesizers to be deployed as software-only products. The software-only products are significantly cheaper (or virtually free when bundled with other products). For example, for several years, speech synthesizers have been bundled with Apple Macintosh computers and with Creative Labs Sound Blaster cards essentially for free.

3.4.4 Consumer Acceptance

In each of the application domains discussed above the issue of speech synthesis quality emerges. In preparing this report, a search was performed for independent analysis of speech synthesis technology and the market (e.g., through independent consultant reports). Surprisingly little information and analysis was available, especially in comparison to the number of reports and product evaluations for speech recognition technology.

Those reports available point to speech synthesis voice quality as the key issue in the lack of penetration of the technology. The improvements in computing power and price have not been matched by improvements in the voice quality of the products. Thus, while virtually every speech synthesis product is promoted as having "natural-sounding" voices, the perception of most potential customers is not so positive.

For instance, from a report cited by the Gartner Group [Michalski, 1997]

> *"We regret to report that text-to-speech technology has barely advanced over the past decade, probably because it isn't as useful to national security as speech recognition. The "drunken Swede" is alive and well."*

From another report on Telecom Strategies [Forrester, 1998]

> *"Time for the text-to-speech vendors to hire a speech coach. Ever hear how text-to-speech platforms read back words and numbers in a stilted monotone? Whose voice is that? ... since the spoken phrases are made up of a series of words assembled uniquely for each playback, the task of integrating punctuation and inflection has been a bear. Forrester thinks that firms like IBM and Lernout & Hauspie would do well to invest in putting personality into their systems to speed acceptance."*

If we review the list of vendors of speech synthesis, a substantial number are too small to invest the funds required to make a paradigm shift in the techniques and research of speech synthesis. This is compounded by the low margins of the business (because of software bundling) and the relatively small customer base. For the other companies a predominantly commercial atmosphere exists in which research findings are not published to the extent required to foster open scientific collaboration.

We therefore return to the key issue of this report which is that dramatic advances are required to produce the speech synthesis technology required to drive a spoken language interface to computers, that the current research environment will not achieve this objective, and thus a coordinated funding effort is required to revitalize speech synthesis research.

4 FUTURE SCIENCE AND TECHNOLOGY: WHAT WE NEED TO BE ABLE TO DO

This section describes current challenges faced by the synthesis research community. Many of the roadblocks in current-day technology could be overcome more easily with a better fundamental understanding of speech production and perception at a cognitive, psychological, physiological, and acoustic level. Section 4.1 lists a number of basic scientific questions that are in need of investigation.

Application of knowledge must proceed in tandem with basic science. Much can be gained by development of computational models that allow implementation of theories and manipulation of perceptually-relevant dimensions of speech signals. Basic science should provide intuitions for practical models; the shortcomings of practical models should, in turn, pose relevant questions for science. Section 4.2 describes areas of computational implementation that should be explored in tandem with basic scientific study.

Finally, as we have argued in the previous section, speech synthesis research is currently limited by a lack of principled evaluation methodologies. Evaluation is needed for understanding the limitations of current

systems, which is an essential first step towards finding new speech models and synthesis algorithms. While the provision of common resources can help address the problem, there are important research questions associated with evaluation that must be investigated, as described in Section 4.3.

As a preface to the sections that follow, we note that this chapter cannot enumerate all important aspects of speech synthesis research, in part because these will necessarily evolve as research advances. New areas will emerge and others will fold. The issues highlighted here represent areas that many researchers think are important for advancing science and technology, though there is certainly some disagreement on priorities. What should be clear from the listed areas is that the research needed is clearly interdisciplinary, including linguistics, psychology, physics, computer science, and engineering.

4.1 Advances in Basic Science

Speech synthesis is, as we have shown, a highly multidisciplinary field. Not only are the problems diverse, requiring the talents of people from varied backgrounds, but the *integration* of various components into a working system requires that people working on one component be aware of the requirements of other components. So researchers must be specialists, but at the same time must be sensitive to issues outside their own particular areas of expertise. Clearly, then, multidisciplinary approaches should be encouraged, and students should be trained to maintain a holistic view of the entire speech synthesis process (in humans and machines), even while they focus intently on a topic of interest. Similarly, although the research issues discussed in the sections below are lumped into categories, their heavily intertwined, multidisciplinary nature should be kept in mind.

4.1.1 Text Analysis and Generation

Auditory displays of information. Scientific advancements in text analysis must include better understanding of the ways humans encode information in text, and ways in which this information can be rendered via other modalities like speech. We must explore more abstract representations of information that will make it easier to translate among and integrate output modalities. The outcome of better understanding in this area could transform the way we distribute information. In such a framework, user and context could determine whether the "display" is audio, text, graphic or a mixture.

Various domain-specific situations influence human text analysis processes, and these need to be better understood. For example, tables of sports scores might be read quite differently from bus timetables. Rendering two-dimensional layout information from World Wide Web pages is often necessary to understand their content. Understanding the way humans represent hierarchical structure in speech is another important area, especially for prosody prediction functions. This includes study of given/new structure in dialogs, focus, intentional/attentional structure, turn-taking, and topic structure.

Pronunciation. A cataloging of ways in which strings of characters can be read, both within English and across languages should be undertaken. For example: the number sequence "1928" might be read as "one nine two eight" (if it is a telephone number, for instance); or as "nineteen twenty-eight" (if it is a year); or as "one thousand, nine hundred and twenty eight" (if it is interpreted as an ordinary number). While this particular instance may seem trivial — for such number strings there is a handful of ways of saying

them — there are many different classes that need to be covered, and each of these classes needs to be well understood. Thus for each class of text string (digits, capitalized sequences, mixed case words, mixtures of digits and letters, ...), we must better understand this set of options in a general way, for multiple languages. We must also gain a better understanding of the range of kinds of features that affect lexical decision. At present researchers typically use a mixed bag of features (words in the immediate context, words in a wider context, local part of speech sequences, ...) that seem "reasonable". Other features that might influence these decisions should be investigated.

Prosodically-motivated text analysis and generation. We need a better understanding of what factors condition prosodic phrasing and accenting decisions, which are essential to structuring the message for the listener. In the limit both of these problems may, of course, require one to solve the general problem of natural language understanding. However, it is likely that one can fall far short of that general goal, and still make substantial progress on these issues. For example, it may be possible to obtain high quality prosodic phrasing from a coarse clause or noun/verb group bracketing instead of a full syntactic parse. Knowledge of the key syntactic constituents for prosody prediction could inform template-based generation systems that do not make use of detailed syntactic representations. Similarly, a better understanding of what aspects of focus structure and semantic representations are needed for proper accent placement could suggest other simplified text analysis/generation structures. Finally, understanding and then automatically extracting (or generating) the factors that effect *degree* of emphasis is a topic that has barely been addressed at all.

4.1.2 Linguistics and Prosody

Prosody in discourse and dialog. One obvious application of synthesis is in systems that use natural spoken dialog interfaces to information. For these systems to grow beyond simple machine-initiated prompts, synthesizers must be able to convey subtle prosodic cues to the "common ground," or state of mutual understanding between the conversants. The rules for controlling prosody in dialogs must be better understood. Related questions include the following:

- What is the effect of turn taking in discourse on prosody?
- How can we model the common ground?
- How can we deal with the often agrammatical language of dialog?
- How does the given/new attribute (or attentional structure more generally) influence de-accenting in a dialog system?
- How do different factors interact to influence the choice of intonation marker associated with a phrase boundary or accent?

Acoustic and visual correlates of prosody. Traditionally, the result of prosodic analysis in synthesizers has been simply a fundamental frequency contour and segmental durations. This area of speech synthesis still requires considerable research but there are many more variables relevant to human prosody that also need to be modeled.

- How is glottal wave shape and voice quality affected by pitch range, accent placement, and other characteristics? This knowledge should influence waveform generation research.
- How do articulatory constraints in casual or fast speech influence the degree of lenition or coarticulation? What is the relationship between articulatory effort and prosody? For example, how tightly does the speaker make closures for various speech styles?
- How can we move away from duration as a primitive to timing, to sub-phone duration control or timing of gestures? What is the best way to characterize speaking rate variations?
- In a visual speech system, word and phrase boundary information are used to control blinking, eye movements, and head nodding, and pitch is used to control the eyebrows [Massaro, 1998]. An important issue is how best to convey emphasis in the visual synthesis.

Speech styles and emotion. A frequently-cited shortfall of current synthesizers is that they lack the ability to change "style" or emotional characteristics. Laypersons and researchers agree that communication involves much more than simply the linguistic message. Paralinguistic as well as linguistic information is necessary for optimal communication and understanding. However, most people are hard-pressed to define the specific acoustic characteristics of a particular style or emotion, even though they can easily identify the styles themselves. In addressing this area, issues to address should include:

- What findings from current sociolinguistic research on style are relevant to the realm of speech synthesis research?
- What are the determinants of particular styles, the characteristics that make a speaking style coherent and the features that differ with gender, dialect or other factors?
- What are the communicative purposes of various styles?
- Can we find appropriate frameworks to answer these questions in ways that are useful for generalizing knowledge from one variety to another?

Analogous to auditory speech synthesis, research in visible speech synthesis has been directed primarily at the linguistic dimensions of speech but it is important that the paralinguistic ones be developed in parallel. Similar effort should be devoted to the synthesis of emotional displays that occur naturally with speech.

4.1.3 Speech Perception

Further work is needed to understand specific aspects of human speech perception, to aid in the development of advanced algorithms. For example, in concatenative synthesis, the fundamental assumption is made that at every point where waveforms are joined, the acoustics are "similar enough" that humans will not be able to detect a discontinuity. Although recent concatenative systems have utilized measures of "target cost" and "concatenation cost" [Hunt and Black, 1996], few speech perception experiments have been conducted to offer guidance on the appropriateness of these measures.

The literature can provide insights on concepts like "just noticeable differences" and temporal masking windows, but this body of work needs to be augmented by experiments that closer approach the conditions found in speech synthesis algorithms - concatenative or otherwise. The amount of variation considered natural by a listener is heavily dependent on context. Certain unit concatenations or model parameter transitions sound unnatural for the simple reason that listeners have an innate understanding of the patterns and time constants of speech production, and know that certain transitions are impossible. Human listeners integrate a large number of redundant cues, which may not all be necessary for intelligibility but may be important for naturalness. Work leading to better objective measures of perceived discontinuity and the interdependence of redundant cues will allow much more direct optimization of synthesis algorithms. For example, in concatenative synthesis, unit mismatches could be concentrated in segments where they would not be noticed. This is similar to the strategies used in audio compression, where quantization noise is shaped to fit an auditory masking curve.

In addition to perceptual measures of spectral continuity, a better understanding of the perception of prosodic features, style and moods is important in advancing those fields. Also, as a later section explains, research on speech perception is fundamental to development of effective and meaningful assessment paradigms for speech synthesis.

4.1.4 Acoustic Modeling

Fundamental understanding of sound generation in the vocal tract. Because of the difficulties in collecting data and modeling the nonlinear acoustical properties of soft tissue, knowledge of the underlying acoustics of speech production is rather rudimentary. Further work is needed to collect vocal tract geometry data and use these data for high-precision models of flow fields near constrictions. This kind of research underpins research on fully parametric articulatory synthesis, but could also have an impact on the structuring of other kinds of knowledge-based algorithms.

Acoustic manifestations of speaker and speaking style. There is currently great interest in the speech synthesis community for creating synthesizers with desired voice quality, emotion, and dialect. The basic research necessary for this to occur comes from many different areas, including the following:

- The effect of speaking style on the dynamics of the upper vocal tract, including the relative timing of articulatory events, as well as aerodynamics and laryngeal physiology.
- The relation between dialect and intonational characteristics, and the individual speaker differences in this relation.

Variability. In many respects, speech synthesizers have been designed to make "safe" assumptions that avoid gross mistakes. However, this means that they also avoid many aspects of the variability observed in natural speech. Research that characterizes the dimensions of variability in one speaker's voice or across speakers is needed. This research will enable systems to produce speech that is less monotonous for the listener and that incorporates acoustic variations correlated with stress, emotion, and various prosodic cues. A better understanding of cross-speaker differences will lead to methods for transforming speech synthesizer outputs to resemble multiple talkers.

4.1.5 Visible Speech Synthesis

Production Models of Synthesis. An important issue for achieving realistic visible speech is how specific speech segments are combined. Alternative models of coarticulation and different segmental approaches should be evaluated. Simpler models (e.g., one using simpler parametric interpolation) should be tested as a baseline for comparison. Analogous to auditory speech synthesis, a comparison between model-driven parameter synthesis should be evaluated against diphone concatenation models in which coarticulation can go no further than the adjacent phoneme.

Individual Differences and Structural Modeling. Given that talkers differ in both their structure and speech articulation, it is important to model these differences, both for models of speech production and to investigate how human speech perception functions in light of these differences. The use of a variety of talkers is necessary to achieve a true picture of which cues are used. Concentration on too small a sample of natural (or synthetic) faces can lead to lack of generality and ecological validity. Synthetic visual speech should simulate the same variety of talkers that people face in the real world.

Texture and intelligibility. One important issue to be resolved is whether the speech intelligibility of a neutral synthetic face is comparable to that of a texture mapped face. A question to be answered by this comparison is whether having the additional facial information might help performance (e.g., by focusing attention on the critical facial features) or alternatively hurt performance (e.g., by obscuring the pure, abstract facial features).

4.2 Advances in Technology

Many of the scientific challenges outlined in Section 4.1 will require considerable technical innovation, but there are many other technical problems that also need attention. In this section we focus on development of practical computational models based on fundamental scientific results. One component of the goal is still a better understanding of the human speech production process, but the path to get there includes creating practical systems that satisfy the requirements of applications.

4.2.1 Domain-specific Modeling

In the speech recognition community, a useful approach has been to optimize and evaluate system performance on constrained tasks within a particular domain such as air travel reservations, Wall Street Journal financial news reports, etc. It is often argued that progress in speech synthesis has been held back in some ways by trying to tackle too general a problem, that of “unrestricted text input”. Too many compromises must be made to generate “reasonable” output across a range of different domains. In contrast, the lexical, prosodic, and segmental acoustic features of speech constrained within a particular domain are more predictable. Therefore, it may be possible to accelerate progress by studying a portfolio of application domains instead of unrestricted TTS.

Implementation of this approach requires an initiative that lays out compelling applications to engage multiple research groups and user communities. Moreover, applications must be chosen that can drive

research and advance the field for several years to be replaced then by more complex and challenging applications. This approach will have the broadest impact if it is thought about as broadly as possible – that is, if the focus is on developing modular approaches so that technological/research advances can be reapplied to new domains.

4.2.2 Text Analysis and Markup

Language modeling. An important area of future research is the investigation of language modeling techniques in text-to-speech synthesis. The traditional model in speech recognition is the noisy channel model where the speech signal is modeled as a noisy representation of the word sequence. In order to reconstruct what was said, one needs a language model, or in other words a model that can estimate the probability of a given sequence of words. Although some language modeling techniques have been used in text-to-speech synthesis (e.g., in sense disambiguation), there has not been any extensive use of such methods. The common conception is that in text-to-speech conversion one basically knows what the words are, unlike the case of recognition. But this is not really correct: written language is an imperfect representation of the underlying linguistic message intended by the writer. As we have noted elsewhere, strings of characters can often be pronounced in more than one way, and the situation is much worse in some languages than in others. Written language only inconsistently represents intonational phrasing through the use of punctuation, and almost never represents accenting. In order to read a text aloud, one must reconstruct a large amount of information that was not provided by the writer.

Markup languages. A fundamental goal is to endow speech with the properties that make text so useful: this includes composition tools, easy browsing, summarization, etc. Markup languages like those described in Section 3.2.1 solve part of this problem by augmenting text with clues for the system to use in generating speech (e.g., emphasize THIS word). However, none of the current markup languages solve all interface needs. For example, if one wants to specify that a given utterance should have "a final fall starting three syllables from the end and descending at a rate of 30Hz per syllable," there is no way to do so. Quite a few specific synthesis systems have their own idiosyncratic ways of specifying such things but there is no standard.

Thus, there are actually two different levels of markup needed:

- Synthesis markup - the input to control a waveform generator or visual speech synthesizer.
- Language markup - the output of a language generator, or of automatic text-annotation algorithms.

Part of the problem in creating these standards is a lack of general theoretical agreement (especially in the area of intonational modeling and emotional "constituents") on what kinds of things should be under the control of the markup interface and how they should be specified. But part of the issue is that the synthesis and end-user community has not yet systematically thought out the things that one would want to be able control with a markup language. In addition, the natural language generation and speech synthesis communities need to be mutually informed about what NLG can provide and what synthesizers might want to use. Standardizing too early is dangerous because it can constrain the research, but proposing a common

form allows people to debate the merits of specific features. For the special case of synthesizing visual speech, there is also the possibility for graphically defined markup languages. For examples, scientists should be able to graphically draw and display emotional intensities superimposed on the text.

4.2.3 Computational Modeling of Prosody

The failure of many current speech synthesis systems to produce natural prosodic characteristics is a major hindrance to wider application of the technology. Research into better models of prosodic effects is needed, and their implementation in novel applications. There are two classes of "prosodic mappings" to consider. One maps knowledge about the discourse to a description of tone/accent placement and type. The other maps between this (symbolic/categorical) prosodic model and waveform synthesis parameters.

Mapping syntax and discourse states to prosodic models. Computationally efficient methods for using high-level discourse information in prosody prediction should be investigated, including robust learning algorithms for training from sparse data. The idea of developing techniques for specific domains is especially relevant to this problem. New algorithms for uncovering the relationships between syntactic and conceptual structures and prosodic structures should be developed and exploited. A fundamental question concerns how one might build on the mappings discovered/modeled for one discourse style to bootstrap to models for another language variety. In other words, what is the computational framework for style adaptation (or adaptation of dialect, speaker, etc.)? Methods that take advantage of common core knowledge could be automatically trained for new components.

Mapping symbolic prosodic models to waveform synthesis parameters. In this area, continued effort must be made to develop new data driven methods that incorporate, implement, and test experimental results from the linguistic sciences. In general, we need to move away from equating "prosody" with "segmental duration and F0" and explore how prosodic structure affects all aspects of the speech signal. An example is the issue of "duration" vs. "timing": models are needed that go beyond manipulating interval lengths assigned to independent segments in sequence. This is necessary to model variation in speed of transition and transition shape for relevant parameters as well as alignment between the changing parameters. Models should cover spectral parameters (spectral coefficients, spectral tilt etc.), acoustic prosodic features (fundamental frequency, energy etc.) and any other dynamically changing parameters.

4.2.4 Knowledge-based Automatic Learning Techniques

A general area in need of much further research is automatic methods for deriving synthesizer acoustic parameters and "rules" from speech databases. Combining rule-based and automatic learning techniques necessitates the invention of ways to incorporate knowledge, structure, or constraints into automatic learning or optimization. The combination of rule-based and automatic learning approaches is important because they complement one another: knowledge is used to define the space in which the automatic learning procedure works. Knowledge-driven automatic learning has already been explored with success in aspect of prosodic modeling and such work should be encouraged. In addition, more efforts in the area of waveform generation are needed.

Unit selection. In concatenative synthesis, unit selection should be viewed as a general paradigm for models that can "learn" from data, instead of simply a way to paste waveforms together. This calls for knowledge-based models motivated by fundamental science, not brute-force data mining. In particular, more work is needed to find speech signal models that offer as much parametric control freedom as a formant synthesizer, but are able to preserve less well understood fine details of the signals. In addition, the use of temporal feature descriptions broader than the typical 20 ms "frame" of cepstra, etc., should be explored, including models that can take advantage of masking properties of hearing and that incorporate aspects of voice quality and other prosodic dimensions.

Parametric synthesis. In parametric synthesis, phone-to-parameter-trajectory rules can be written so that the trajectories are mathematical functions, such as straight lines, which themselves have the free parameters of slope and intercept. Using analysis-by-synthesis techniques (optimal match of acoustic parameters), these parameters could be adjusted to fit the average speech patterns of a large group of speakers, or the speech pattern of a single speaker. The research into this area would involve determining the perceptually salient acoustic features to use as matching criteria in the analysis-by-synthesis procedure. Research would also need to be conducted on which synthesizer parameter features correspond to the salient acoustic features. Further research into the algorithmic aspects of optimization and automatic learning in this particular application would also be necessary.

4.2.5 Acoustic Parameter Generation and Signal modeling

A set of diverse and innovative techniques should be encouraged to address the problem of generating speech waveforms from a symbolic representation of linguistic information for an utterance. This strategy can serve both the purpose of enhancing our understanding of vocal tract acoustics and in producing usable systems in the near- to medium-term.

Moving from symbolic labels to multi-dimensional trajectories. The transformation from independent symbolic labels to multi-dimensional trajectories can be instantiated as the transformation from phones to articulatory trajectories. This transformation is a vital part of articulatory and quasi-articulatory synthesis. The research in this area has largely been undertaken in relation to health, psychological, and linguistic issues. There has been some interest in this area of research from both the speech synthesis technology and automatic speech recognition communities.

In the United States there has been generally only a small degree of interaction between the linguistic/health related speech physiology researchers and speech technology researchers. A much greater amount of collaboration between the two groups could greatly improve the prospects for parametric articulatory synthesis and formant synthesis, while increasing the knowledge for the implementation of concatenative synthesis.

Voice source quality transformations. The term 'voice quality' usually refers to perceptual features such as breathiness, creakiness, and whisper [Klatt and Klatt, 1990, Childers and Lee, 1991, Childers and Hu, 1994]. Several studies have emphasized that breathiness is important to the synthesis of a realistic female voice

[Klatt and Klatt, 1990, Trittin and de Santos y Lleó, 1995], which is important since in many ways synthesized female voices still talk like men. Voice quality may also be important for implying emotions like sadness or disgust [Murray and Arnott, 1995]. Most applications of voice quality modeling have focused on formant synthesis [Klatt and Klatt, 1990, Childers and Lee, 1991, Childers and Ahn, 1995] or have used synthetic pulse-excited linear prediction [Childers and Hu, 1994, Cummings, 1992, Cummings and Clements, 1993]. For waveform-concatenative synthesis, waveform modification techniques are necessary. While it may be possible to capture some effects as a by-product of unit selection from a large database (e.g., occasional end of utterance creak), the combinatorics of speech features dictate that modification of waveforms will always be needed: it is simply not practical to record and store redundant segments with several voice qualities and other prosodic features. Given the growing popularity of waveform concatenation methods, voice quality modification could have a dramatic impact by offering a new dimension of expressiveness to speech synthesizers. Advancements in waveform modification techniques to achieve a specific voice quality would require techniques for automatic extraction of voice quality parameters and other laryngeal function indicators.

Fundamental frequency modification techniques in waveform synthesizers are another important class of algorithms for advancing the technology. Most current approaches make the assumption that the vocal tract transfer function (i.e., the spectral envelope) and the glottal source (excitation pitch pulses) are independent of each other. This is a useful assumption that gives reasonable results for small modifications, but it ignores important details that are important for more drastic pitch variations. Because of interactions between the glottal source and the vocal tract (due to the time-varying coupling of the trachea to the acoustic system) and reconfigurations of the articulators, the transfer function of the vocal tract changes with pitch and voice source quality [Titze, 1994]. More detailed study of the interdependence of fundamental frequency and the vocal tract transfer function is needed, along with the development of signal processing models that can incorporate this knowledge gracefully.

Voice conversion mappings. The application of synthesis often requires the use of multiple voices. However, the waveform concatenation-based approach requires that a large, annotated corpus of speech be stored on the computer for each voice; formant or articulatory synthesizers require coding of a large set of new model parameters to describe the new voice. Sufficiently accurate labeling of corpora is not yet automatic, and the complications of collecting a good quality corpus are significant. Techniques for rapidly transforming the system's voice characteristics to that of another speaker could potentially allow one to reuse an existing voice to synthesize speech in a another voice based on a small amount of adaptation data.

Most previous approaches to this problem have relied simply on various forms of regression mapping between source-target pairs of spectral features (e.g., LPC or cepstral coefficients)
[Kain and Macon, 1998, Stylianou et al., 1995, Arslan and Talkin, 1997]. Although many acoustic correlates of voice identity are carried in the spectral envelope [Kuwabara and Sagisaka, 1995], a frame-by-frame spectral mapping does not capture systematic cross-speaker variations in timing, voice quality, intonation, or other supra-segmental features. Much further work is needed to allow understanding and modeling these differences to produce more realistic voice conversion results.

4.2.6 Visible Speech Synthesis: Enhancing Realism

A long-term goal of visible speech synthesis would be to simulate, as closely as possible, the physiognomy of a human talker. To enhance the realism and expressiveness of the visual representation of the talking head, it is important to: 1) develop and program a realistic tongue, hard palate, and soft palate with a movable velum; 2) include a fully three dimensional representation of the head including ears, hair, top, side and back of head, and neck; 3) have fine detail of the talker's facial features; such as forehead, eyebrows and neck; and 4) allow independent control of the two sides of the face. A consequence of these improvements will be the increased resolution and display of speech and facial affect and the implementation of facial asymmetries for speech and affect.

4.3 Evaluation of Synthesis Systems

In Section 3.3, we argued that current speech synthesis assessment methodologies are not meeting the needs of either consumers or scientists, and that speech synthesis evaluation is inherently complicated and is an open research question itself. However, it is clear what areas of work are needed:

1. Research on next-generation evaluation tools.
2. Developing system architectures that allow access to internal representations.
3. Agreement on evaluation standards.
4. Performing large-scale evaluations and publishing the results.
5. Making TTS systems accessible via the internet.

Of these steps, the last three are primarily programmatic. Here, we outline some research issues associated with evaluation.

Research on next-generation evaluation tools. Current tests are not sufficiently thought through in terms of their exact goals and are not always well-matched to the different research agendas in speech synthesis. For example, in developing a synthesizer that is customized to a particular task, one might want to assess whether the speech is "appropriate" for the task and not simply whether it is intelligible or natural. Current perceptual tests tend to focus on short words or sentences, whereas synthesizers are often used for longer sections of text and in the future may be used in dialog systems. It is impossible for humans to listen to long stretches of speech from two systems in an A-B comparison, and innovative methods are needed. In this case, advances in synthesis evaluation can also inform research in evaluation of complete dialog systems. Current tests are insufficiently automated (many still use paper-and-pencil), and automation is important for ensuring consistency and sharing resources. In addition, if web-based evaluation sites are to be a goal, as we are proposing, automation is essential. Automating perceptual tests may also inspire new paradigms for evaluation. Quantitative (i.e., not perceptual) tests analogous to word error rate already exist for some components of synthesis, but these could be improved to better account for the allowable variability in speech that is captured in a perceptual test. Finally, new tests are needed that take advantage of text corpora,

where it is possible to guarantee that test text is new for all systems and the statistical and generalizational properties can be controlled.

System architectures that allow access to internal representations. From a scientific perspective, being able to surgically test components is much preferable over having to indirectly infer how well a certain component performs from overall system behavior. A difficulty is that different systems have different internal representations, and these are rarely made available as intermediate control variables. While legislating internal representations will impede research, finding some intermediate representations that can be generally agreed to (e.g., phonemic output) is possible.

5 RECOMMENDATIONS FOR A FUNDED RESEARCH PROGRAM IN SPEECH SYNTHESIS

5.1 Introduction

The preceding sections of this report have supported the view that while speech synthesis is potentially a very important technology for improving access to electronic information and enhancing the human computer interface, it is not yet a "mature" technology and there are many open areas of research. For the reasons outlined in the Introduction, we argue that progress towards mature speech synthesis has been impeded by the lack of public funds for this area of research. In short, only so much progress can be made in industrial labs (the main locus of synthesis research in the US over the last two decades), and progress will be significantly faster if a wider range of researchers can enter the field.

In this final section we detail four key recommendations for how a funded research program in speech synthesis should be structured. These recommendations are based on the view that the role of the government in creating a funded research program in synthesis is to seed a range of research activities that **creates an environment in which real breakthroughs can occur**. For such an environment, it is critical that the timetable be long rather than short, and that multi-year goals be stated which are flexible and open-ended to encourage true innovation. For that reason, the recommendations here are more general and programmatic than prescriptive.

Before we turn to our recommendations (Sections 5.2–5.5), it is useful to consider what we might learn from the most obvious prior model for a funded research program in speech synthesis, namely the DARPA program in speech recognition and understanding carried out over the last two decades. As we shall show next, some aspects of the DARPA program are clearly applicable: indeed, many of the specific recommendations that we will make in Sections 5.2–5.5 are inspired by the DARPA experience. However, we also stress that a direct application of the DARPA evaluation paradigm is difficult to map to synthesis research directly. In part, this difficulty reflects the fact that speech synthesis and recognition are at different levels of maturity, but it is also a difficulty associated with human-computer interaction technologies in general. For that reason, lessons learned from a new program in synthesis could be informative to current and possibly future DARPA programs where the emphasis is more on human-computer dialogs.

Lessons Learned from DARPA Program in Automatic Speech Recognition

Four important hallmarks of the DARPA program in speech and natural language can be identified. These were: (1) the establishment of common benchmarks; (2) an emphasis on demonstrations; (3) the "arranged marriage" of speech recognition and natural language processing researchers; and (4), the organization of annual workshops for program participants. We briefly describe the benefits, the drawbacks and the relevance of each of these features to the speech synthesis research agenda:

1. First, establishing **benchmarks** was important in measuring progress. Before the establishment of common benchmarks, nearly every speech recognition system claimed "99% accuracy." Of course, that number could mean very different things depending on the test set, the training set, the complexity of the task (e.g., vocabulary size), etc. The advantages of common benchmarks are that the results are more interpretable across sites and that they allow for the leverage of shared data and evaluation resources. In addition, they tend to inspire a competitive spirit in which people strive to make progress. Certainly, there was much progress made in lowering word error rate of speech recognizers, the common evaluation metric.

   The main disadvantage was that the evaluations, focused solely on error rate, led to some level of risk aversion. The result was that systems tended to converge technically, relatively few resources were devoted to riskier innovative methods, and progress was incremental, being more evolutionary than revolutionary. Perhaps more importantly, the chosen error-rate-based evaluation metrics drove the research agenda in a rather narrow direction. In particular, the ATIS spoken language system evaluations focused on an utterance-level metric restricted to only the simplest utterances, chosen because the researchers could not agree on a dialog- or task-level evaluation protocol. The consequence was that research in dialog and system initiative was discouraged.

2. The second characteristic of the DARPA program was the emphasis on **demonstrations** of the technology, something that proved very important in getting attention for the field. Demonstrations are compelling; they are a proof of concept and also constitute a type of evaluation. Ideally, they show that things can be done automatically and robustly in a reasonable amount of time. Seeing the technology in action helps inspire funding sources and collaborators, and it helps attract new people to the field.

   On the other hand, significant resources can be diverted from research toward the support and maintenance of demonstrations. On the other hand again, applications and basic research can provide mutual support. The experience of transitioning basic research into applications can lead to unforeseen problems, inspiring new research directions such as work in mixed initiative dialogs and robust parsing.

3. The third important feature of the DARPA program was the so-called **arranged marriage** between speech recognition and natural language technology. In the 1980's, DARPA announced a new program in which funding would not be given out unless the proposed work involved integrating these two areas. The advantages were that the enforced multidisciplinary challenge inspired some new work and caused some people to work together who would not have otherwise done so. New possibilities

were opened up that would not have been possible without such integration. The disadvantages were that the marriage was not exactly "chosen" and to some extent there was a tendency to keep "separate bedrooms." Integration has been slow, and sometimes more token than real.

4. Finally, the DARPA-sponsored **workshops** on speech and language technology were an important driver of technology and of multi-disciplinary work. The long journal publication delays in the speech field at present have increased the importance of workshops in general. The DARPA workshops, in particular, were important because the focus on a common task made it easy to learn from other researchers and because of the inclusion of both speech and language technology. In fact, it was probably these workshops and not the forced marriage that had the most impact on the increasing transfer of methods between speech and language technology. The frequent exchange of information in the context of a common task was important in moving forward the community as a whole. Of course, the cost is that the number of workshops that researchers attend is increasing, and workshop participation must be balanced with time spent on actual research.

The DARPA program in speech recognition and understanding has clearly had many positive effects, many of which can be linked directly to the four key characteristics of the program that we have discussed above. We would therefore do well to emulate the positive aspects of the DARPA program in any funded research program in speech synthesis, and for this reason the recommendations in Sections 5.2–5.5 are inspired to some degree by the above four items.

However, in learning from the positive results of the DARPA program, we must also be willing to learn from the negative results. We wish to avoid a direct application of past DARPA programs in speech recognition to a new program in synthesis, because of the differing natures of the technologies and because of the different levels of maturity of the fields. In particular:

- While all funded research should clearly be evaluable via *appropriate* metrics, we believe that it would be counterproductive to erect a competitive environment in which innovation is stifled because all efforts are aimed towards a one-dimensional of evaluation.

- While we strongly advocate the idea of multidisciplinary interactions, we feel it would not be wise or necessary to enforce a particular set of collaborations. Speech synthesis already has a long history of multidisciplinary research that will surely continue, even without an external mandate.

- Demonstrations and workshop attendance are important, but they should not become so frequent as to be overburdensome.

- Above all, a great deal of thought, and indeed basic research, must be devoted to the question of how to evaluate our progress. In speech synthesis there is no obvious parallel to the simple word-error-rate criterion that was so heavily used in the DARPA program: as we argued in Section 3.3, evaluation in synthesis is complex, and there is no generally agreed method or methods to do it.

Our recommendations borrow heavily from the DARPA program but differ in areas where it is important to establish a program appropriate to speech synthesis research.

5.2 Recommendation: Support annual workshops for reporting progress and exchanging knowledge

We recommend regular workshops to allow for benchmarking and demonstrations (points 1, 2, and 4 above), to facilitate timely exchange of research findings and for training of new speech synthesis researchers. Such workshops, evaluations and demonstrations, if designed correctly, can be beneficial to the field. We recommend that the workshops be held annually for the purposes of exchanging information. Clearly there is a need for regular measurement to chart progress, and there is a need for workshops for reporting on this progress, as well as sharing technical and scientific advances. As we have noted, providing common tasks for program participants makes it easier to interpret the results of others and therefore leads to faster progress for the community as a whole. However, as we have also noted, we must avoid evaluation paradigms that lead to incremental improvements, stifle innovation, and penalize longer-term more speculative research. We have also discussed the lack of a generally agreed upon evaluation metric. With these points in mind, we therefore make the following four sub-recommendations:

Workshops should be held annually in association with benchmark tests, with a goal of the first such workshop being the definition of common evaluation protocols. The goals of the workshops and evaluations would be to cooperatively advance the state of science by facilitating communication across sites and across disciplines. In addition, the workshop could also be a vehicle for demonstrating progress, both within the scientific community and to government organizations funding the program. To that end, such workshops would include system-level results reported on one or more common tasks, site-specific system-level demonstrations, and analyses based on component-level evaluation on common and site-defined tasks. Having some work on common tasks enables researchers to compare results across tasks, but encouraging work on other tasks as well will ensure generalizability of the technology and avoid problems associated with a narrow research focus. Encouraging work on multiple scientific issues will enable advancement in addition to assessment of the current technology.

Resources for common evaluations should support a suite of benchmark tests, designed to encourage research that will support both constrained- and unconstrained-domain tasks. It is clear that there is a broad spectrum of tasks that require speech synthesis and that a research program should not focus on a single dimension. It is not clear whether one approach will serve all needs, and certainly some tasks will be more amenable to short-term approaches than others. For that reason, common evaluations should support at least two tasks — constrained- and unconstrained-domain — though both should be open vocabulary since stored voice response systems can satisfy the closed vocabulary task needs. Since it is important to maintain consistency in the task from year to year in order to chart progress, the chosen tasks should be very difficult.

The program must allow for a subset of the participants to report only on internally-defined evaluation metrics. For those working on whole systems, task-level evaluations are appropriate and a good way to demonstrate progress. However, if the goal is broad involvement and inclusion of academic work, many researchers will not have complete systems and their work will not be amenable to such evaluations. In these cases, the requirement should be to have some site-defined evaluation (discussion of which could

lead to new common evaluation standards) to assess progress on pieces of the common tasks as well as any other site-defined problem. Examples of component-level evaluation metrics might be perceptual experiments assessing copy-synthesis quality for evaluating waveform generation techniques and mean squared error measures of duration difference relative to some target. In many cases, such measures cannot be "common" standards, since different synthesis systems modularize components in different ways. However, all participants must find measures by which they can demonstrate progress.

Resources should be devoted to expanding the web-based evaluation paradigm already in place. From a consumer's perspective, which is to some extent what government funding organizations are, the best way to assess synthesis technology is to evaluate it with text that is representative of their target application domain. For such purposes, a web-based evaluation of synthesis technology is ideal. As described in Section 3.3, there is already such a facility in place and participation is growing. This site can be improved by more active support, but also by allowing for the capability of assessing system tuning capabilities. For instance, techniques are emerging that allow for voice modifications from sample data, and some effort is needed to accommodate this sort of assessment in the web-based evaluation framework.

5.3 Recommendation: Support a broad portfolio of research

Speech synthesis is a highly multidisciplinary undertaking and research in diverse fields is required to advance the quality of speech synthesis. Thus, we recommend that funding include support for a portfolio of research activities intended to address the existing limitations of the technology. In Section 4, we reviewed a range of specific research topics. The following are the general properties that are important for a funded program for speech synthesis to be successful:

Support should be available for work in all areas of basic science underlying speech synthesis, in evaluation methodologies, and in the technology itself. Clearly, if all resources are spent on evaluation or on basic science, there will be little immediate impact on the technology and society will not be well served. On the other hand, investment purely at the technology end is unlikely to engender the kind of paradigm shift needed to significantly advance the state of the art.

Support should be available for a variety of competing approaches to synthesis. As should be clear from the overview of current synthesis technology, there is no one accepted solution to the speech synthesis problem and it is furthermore possible that different solutions will be useful in addressing constrained versus unconstrained domain tasks. "Data driven" and "knowledge driven" approaches are both needed, as well as combinations of the two. Engineering solutions that work well when much data is available may not work as well in the cases of sparse or unreliable data, as one finds when one moves to "higher" linguistic levels (e.g., discourse modeling), or when one moves to new applications. Conversely, data-driven "engineering" solutions can serve to fill in the many gaps where human knowledge is incomplete, particularly with regards to speaker and style variability. Relying on one model is risky in evolution, in finance and in science.

Collaboration should be encouraged — where appropriate. As should be clear from the historical overview, the review of current technology, and the recommendation of areas for future research, there are a large number of disciplines that touch on the problem of speech synthesis, including linguistics, computational linguistics, speech science, psychology, as well as engineering and computer science. Supporting multidisciplinary work is essential, but this means supporting work in multiple disciplines as well as encouraging collaboration. Collaborations between individuals working in multiple technical areas is most feasible in large institutions where there is a wide span of areas of expertise. Certainly such collaborations should be encouraged, but not to the detriment of smaller less interdisciplinary efforts, which nonetheless show promise of advancing the state of the art. We nevertheless recognize the importance of making interdisciplinary collaborations as feasible and as widespread as possible: we therefore will propose (Section 5.5) as a specific recommendation the establishment of interdisciplinary research centers.

Both system-building and smaller-scale component-level research programs should be supported. It is neither efficient nor practical to ask that all research groups build full synthesis systems. In addition to the unnecessary cost of duplicated effort, many research teams will not have the interest or expertise to develop complete state-of-the-art systems. On the other hand, it is frequently the case that the real usefulness of an idea is hard to estimate without implementing the idea as part of a full working system: many of us in the speech synthesis community have seen proposals that seemed reasonable on paper, and even seemed reasonable in demonstrations that are intended to isolate the phenomenon in question; but fail to result in noticeable improvements when implemented as part of a full system. To solve this dilemma we will propose (Section 5.4) a research infrastructure that includes public-domain and/or open-source full speech synthesis systems, that will help research groups working on component-level problems demonstrate their ideas in the context of the problem as a whole.

International collaboration should be encouraged. Finally, one of the goals of synthesis research should be the development of high-quality synthesis technology not only for English, but for many languages. Contrary to popular parlance — one often hears talk of "porting" a particular TTS system to a new language — each language presents its own portfolio of problems. Solutions to a particular problem in English will of course be tremendously useful in attacking similar problems in other languages, but they will not obviate the need for further work, in some cases involving further research, in those languages. The United States is a multicultural society, and we are fortunate to have in the speech technology community a large number of researchers who are native speakers of languages other than English. Inevitably, though, our expertise in some languages cannot match the expertise available in countries where the languages in question are the national languages. This implies a need for international collaboration. Such collaboration should be actively encouraged under the program.

5.4 Recommendation: Build a speech synthesis research infrastructure

Like any large research effort involving multiple sites, a research effort in speech synthesis will require a core set of common public resources. We recommend that funding be made available for the development of the following shared resources that are necessary to support research on speech synthesis.

Common Databases. Public databases of various kinds are necessary, for exploratory data analysis, for system parameter training, and for testing and evaluation. The list of such databases is open-ended, and to a large extent the community will have to decide which ones are most needed, and will have to prioritize based on needs and available resources. A partial list of the kinds of databases needed is the following:

- Labeled text corpora; for example, corpora where all tokens, including number sequences, abbreviations, etc., are labeled with pronunciation, and other linguistic information that could help in training disambiguation models.
- Corpora of text and speech — preferably parallel text and speech — labeled with prosodic information, such as strength of prosodic boundaries and strength of accents.
- Annotated text corpora tailored to the needs of CTS systems, where annotation includes meaning structures as well as prosodic labels.
- Constrained domain (but open vocabulary) and unconstrained domain text and speech data for testing and evaluation.
- Parallel speech and physiological measurements for training articulatory models of waveform generation as well as talking heads.[1]
- Parallel speech and video data for training visual synthesis models.

It is worth reminding the reader that such databases are needed not just for English, but for many other languages as well.

Common Tools. In addition to corpora, various tools are needed. Once again, the community will need to decide what those tools should be, and prioritize accordingly. Again, a partial list would include:

- Tools for corpus annotation.
- Evaluation tools, including graphical interfaces for conducting perceptual experiments, tools for the quantitative analysis of prosodic labeling differences, and so forth.
- Complete public-domain text-to-speech and concept-to-speech systems, so that research groups working on individual components have mechanisms in which to test out their ideas in the context of a full working system.

[1] There are a number of data sources about speech production - both static and dynamic - that can be utilized. These data sources include: 1) ultrasound and electropalatography (EPG) data for the tongue from a single talker [Stone and Lundberg, 1996], 2) X-Ray microbeam [Westbury, 1994] and 3) cineradiographic recordings [Munhall et al., 1995] of the vocal tract during articulation, 4) MRI and EPG data [Alwan et al., 1997], 5) 3D Cyberware laser scans of static facial gestures and visemes, and 6) systematic measurements of visible speech articulation using a computer controlled video motion analyzer.

Standards. Finally the community must agree upon standards for markup of the databases, as well as markup schemes (such as SABLE) to support further technological developments.

Needless to say, some databases, tools, and standards already exist, but none of the existing resources are entirely adequate, and would either need to be extended or else simply replaced with other materials for the purposes of speech synthesis research. For example, the Linguistic Data Consortium (LDC) has large text corpora for many languages. However, only a few of them are annotated, and none of them have the kind of annotations needed for work on text-to-speech synthesis. The transcribed speech corpora available from the LDC are virtually all designed for speech recognition applications, including a small amount of speech from a large number of speakers, which is in contrast to synthesis needs of a large amount of speech from a small number of speakers.

The Festival speech synthesis system [Black et al., 1998], developed by Alan Black, Paul Taylor and others at the University of Edinburgh's Centre for Speech Technology Research, is a widely used freely available public domain system that could serve the community well as a platform for common development of components. Nonetheless, more work will likely be needed to hone the Festival system further to optimize it for use by a wide variety of people working on various components. Furthermore, Festival represents one particular approach to synthesis, and other systems would be useful to support research in alternative paradigms.

For prosodic labeling, the most widely used standard is ToBI. However ToBI does not include conventions for marking strength of accenting, which we believe will be crucial for training models to achieve higher naturalness.

5.5 Recommendation: Support multi-disciplinary centers of excellence

The number of problems needing research attention could consume a large amount of resources. To facilitate sharing of resources and provide a catalyst for multi-disciplinary collaboration, we believe it will be necessary to have funded Centers for speech synthesis research. Two models for such Centers are possible, the first consisting of a loose confederation of multiple laboratories, and the second being funded institutes established at particular (academic) institutions. While such a Center would undoubtably be involved in technical research, it should be primarily supported to serve as a coordinator among various activities and various organizations. The following issues would need to be considered in setting up such a Center:

Facilitate communication and cooperation between commercial and academic institutions. Academic-commercial collaboration is important for technology transfer in both directions. Since so much recent work in synthesis has been done in industry, it is important to have industrial involvement to make sure that wheels are not reinvented and that academic research is relevant. At the same time, the hope of such a program is to push beyond what industrial research can support, in which case there will be a need to transfer technology back to industry. One mechanism for including industry is simply providing funding to support participation in benchmark tests and associated workshops. Another mechanism would be to fund "sabbatical" visiting faculty positions for industrial affiliates. Yet another possibility is for the Center to broker an indus-

trial mentoring program, where individual academic research programs would have industry representatives (from different companies) assigned as "industrial liaisons" or mentors, following the Semiconductor Research Consortium model (see http://www.src.org/mentor/index.dgw), to provide feedback on research goals and help with obtaining resources needed to do state-of-the-art research. Another role might be serving as a resource for setting up students with industrial internships.

Serve a role in developing and distributing common tools and data. Another important function of the Center will be as a repository and distribution center for the public infrastructure materials discussed in Section 5.4. Thus, the Center (or one member of the confederation) would have roughly the function in the synthesis research community as the Linguistic Data Consortium at the University of Pennsylvania has in the speech recognition and language technology community.

Coordinate workshops. A Center should have responsibility for organizing workshops centered around common task evaluation. Apart from organizing the workshops themselves, one of the responsibilities of the Center would be coordinating work on the evaluation methodologies to be used in the task, including both research (as needed) into the evaluation methodologies, and the actual running of the evaluation for the workshops. In addition to these progress evaluation workshops, it will probably be useful to have more extended workshops focusing entirely on scientific and technological progress, as in the Johns Hopkins University summer workshops on speech and language processing or European summer schools on various problems in speech processing hosted by different sites. Such workshops should be organized so as to encourage broad industrial involvement, which means that the limited resources of small companies must be considered and workshops would likely be at least partly "virtual".

Supporting both large and small efforts. In order to have widest possible impact and train the largest number of students, it is important that the program and the Center support projects at different levels of effort from a variety of institutions. One way to ensure broad impact is by simply serving as a tool and data distribution resource. Another idea is to provide funds for students at schools with small programs to spend a semester at schools with big programs taking classes that would not be available to them otherwise.

References

[Allen, 1992] Allen, J. (1992). Overview of text-to-speech systems. In Furui, S. and Sondhi, M. M., editors, *Advances in Speech Signal Processing*, chapter 23. Marcel Dekker, Inc., New York.

[Allen et al., 1987] Allen, J., Hunnicutt, M. S., and Klatt, D. (1987). *From Text to Speech: the MITalk System.* Cambridge University Press, Cambridge.

[Alwan et al., 1997] Alwan, A., Narayanan, S., and Haker, K. (1997). Towards articulatory-acoustic models for liquid consonants based on MRI and EPG data. part II: The rhotics. *Journal of the Acoustical Society of America*, 101(2):1078–1089.

[Anderson et al., 1984] Anderson, M. D., Pierrehumbert, J. B., and Liberman, M. Y. (1984). Synthesis by rule of English intonation patterns. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, pages 2.8.1–2.8.4.

[Arslan and Talkin, 1997] Arslan, L. M. and Talkin, D. (1997). Voice conversion by codebook mapping of line spectral frequencies. In *Proc. Eurospeech*, pages 1347–1350.

[Aubergé, 1993] Aubergé, V. (1993). Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis. In *Proc. ESCA Workshop on Prosody: Working Papers 41*, pages 62–65.

[Bachenko and Fitzpatrick, 1990] Bachenko, J. and Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170.

[Basson et al., 1991] Basson, S., Yashchin, D., Silverman, K. E. A., and Kalyanswamy, A. (1991). Accessing the acceptability of acna: A rigorous comparison of text-to-speech synthesizers. In *Proceedings.* American Voice I/O Systems.

[Benoit et al., 1996] Benoit, C., Grice, M., and Hazan, V. (1996). The SUS text: A method for the assessment of text-to-speech intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18(4):381–392.

[Black et al., 1998] Black, A., Taylor, P., and Caley, R. (1998). The Festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival.html.

[Bregler et al., 1997] Bregler, C., Covell, M., and Slaney, M. (1997). Driving visual speech with audio. In *Computer Graphics Proceedings*, Palo Alto, CA.

[Brooke, 1989] Brooke, N. M. (1989). *Visible speech signals: Investigating their analysis, synthesis and perception.* Elsevier Science Publishers, Amsterdam. Taylor, M. M. and Neel, F. and Bouwhuis, D. G. (Eds.).

[Brooke, 1992] Brooke, N. M. (1992). *Computer graphics synthesis of talking faces.* Elsevier Science Publishers, Amsterdam.

[Brooke and Summerfield, 1983] Brooke, N. M. and Summerfield, A. Q. (1983). Analysis, synthesis, and perception of visible aritculatory movements. *Journal of Phonetics*, 11:63–76.

[Cassell et al., 1994] Cassell, J. C., Pelechaud, C., Badler, N. I., Steedman, M., Achorn, B., Becket, T., Dourville, B., Prevost, S., and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Computer Graphics Annual Conference Series*, pages 413–420.

[Childers and Ahn, 1995] Childers, D. G. and Ahn, C. (1995). Modeling the glottal volume-velocity waveform for three voice types. *Journal of the Acoustical Society of America*, 97:505–519.

[Childers and Hu, 1994] Childers, D. G. and Hu, H. T. (1994). Speech synthesis by glottal linear prediction. *Journal of the Acoustical Society of America*, 96:2026–2036.

[Childers and Lee, 1991] Childers, D. G. and Lee, C.-K. (1991). Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90:2394–2410.

[Cohen et al., 1998] Cohen, M. M., Beskow, J., and Massaro, D. W. (1998). Recent developments in facial animation: An inside view. In *Auditory Visual Speech Perception.* http://mambo.ucsc.edu/psl/avsp98/11.doc.

[Cohen and Massaro, 1993] Cohen, M. M. and Massaro, D. W. (1993). *Modeling coarticulation in synthetic visual speech*. Springer-Verlag, Tokyo.

[Cohen and Massaro, 1994] Cohen, M. M. and Massaro, D. W. (1994). Development and experimentation with synthetic visible speech. *Behavorial Research methods, Instrumentation, and Computer*, 26:260–265.

[Cohen et al., 1996] Cohen, M. M., Walker, R. L., and Massaro, D. W. (1996). *Perception of synthetic visual speech*. Springer-Verlag, New York. Stork, D. G. and Hennecke, M. E. (Eds.).

[Coker, 1968] Coker, C. (1968). Speech synthesis with a parametric articulatory model. In *Speech Symposium*, Kyoto. Reprinted in Flanagan, James and Rabiner, Lawrence (Eds.) *Speech Synthesis*, Dowden, Hutchinson and Ross, Stroudsburg, PA, 135–139, 1973.

[Cole, 1995] Cole, R. e. a. (1995). The challenge of spoken language recognition: Research directions for the 90s. *IEEE Transactions on Speech and Audio Processing*, 3(1):1–21.

[Collier, 1991] Collier, R. (1991). Multi-language intonation synthesis. *Journal of Phonetics*, 19:61–73.

[Cummings, 1992] Cummings, K. E. (1992). *Analysis, Synthesis, and Recognition of Stressed Speech*. PhD thesis, Georgia Institute of Technology, Atlanta, GA.

[Cummings and Clements, 1993] Cummings, K. E. and Clements, M. A. (1993). Application of the analysis of glottal excitation of stressed speech to speaking style modification. *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, II:207–210.

[Davis and Hirschberg, 1988] Davis, J. and Hirschberg, J. (1988). Assigning intonational features in synthesized spoken discourse. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 187–193, Buffalo, New York.

[Dedina and Nusbaum, 1996] Dedina, M. and Nusbaum, H. (1996). PRONOUNCE: a program for pronunciation by analogy. *Computer Speech and Language*, 5:55–64.

[Dixon and Maxey, 1968] Dixon, N. and Maxey, H. (1968). Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Transactions on Audio Electroacoustics*, AU-16:40–50.

[Dutoit, 1997] Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis.* Kluwer, Dordrecht.

[Dutoit and Leich, 1993] Dutoit, T. and Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13:435–440.

[Ekman and Friesen, 1977] Ekman, P. and Friesen, W. V. (1977). *Manual for the Facial Action Coding System.* Consulting Psychologists Press, Palo Alto, CA.

[Flanagan et al., 1980] Flanagan, J. L., Ishizaka, K., and Shiple, K. L. (1980). Signal models for low bit-rate coding of speech. *Journal of the Acoustical Society of America*, 68(3):780–791.

[Forrester, 1998] Forrester (1998). Telecom strategies. Research Report 2, Forrester Research.

[George and Smith, 1997] George, E. B. and Smith, M. J. T. (1997). Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Transactions on Speech and Audio Processing*, 5(5):389–406.

[Golding, 1991] Golding, A. (1991). *Pronouncing Names by a Combination of Rule-Based and Case-Based Reasoning.* PhD thesis, Stanford University.

[Hallgren and Lyberg, 1998] Hallgren, A. and Lyberg, B. (1998). Visual speech synthesis with concatenative speech. In *Auditory Visual Speech Perception.*

[Hanson, 1997] Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, 101:466–481.

[Henke, 1967] Henke, W. L. (1967). Preliminaries to speech synthesis based upon an articulatory model. In *Proceedings of the IEEE Conference on Speech Communication Process*, pages 170–182.

[Hertz, 1990] Hertz, S. R. (1990). *The Delta programming language: an integrated approach to non-linear phonology, phonetics, and speech synthesis.* Cambridge University Press. J. Kingston and M. Beckman (eds.).

[Hertz, 1991] Hertz, S. R. (1991). Streams, phones, and transitions: toward a phonological and phonetic model of formant timing. *Journal of Phonetics*, 19. R. Carlson, Ed.

[Hertz and Huffman, 1992] Hertz, S. R. and Huffman, M. K. (1992). A nucleus-based timing model applied to multi-dialect speech synthesis by rule. In *Proc. of the International Conference on Spoken Language Processing*, volume 2, pages 1171–1174.

[Hertz and Zsiga, 1995] Hertz, S. R. and Zsiga, L. (1995). The delta system with syllt: Increased capabilities for teaching and research in phonetics. In *Proc. of the International Congress on Phonetic Sciences*, volume 2, pages 322–325.

[Hirose and Fujisaki, 1982] Hirose, K. and Fujisaki, H. (1982). Analysis and synthesis of voice fundamental frequency contours of spoken sentences. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, pages 950–953.

[Hirschberg, 1992] Hirschberg, A. (1992). Some fluid dynamic aspects of speech. *Bulletin de la Communication Parlée*, pages 7–30.

[Hirschberg, 1993] Hirschberg, J. (1993). Pitch accent in context: predicting intonational prominence from text. *Artificial Intelligence*, 63(1–2):305–340.

[Hirschberg and Grosz, 1992] Hirschberg, J. and Grosz, B. (1992). Intonational features of local and global discourse structure. In *Proceedings of the Speech and Natural Language Workshop*, pages 441–446, Harriman, New York.

[Hirst et al., 1991] Hirst, D., Nicolas, P., and Espesser, R. (1991). Coding the F0 of a continuous text in French: An experimental approach. In *12eme Congres International des Sciences Phonetiques*, pages 234–237.

[Holmes, 1973] Holmes, J. N. (1973). Influence of the glottal waveform on the naturalness of the speech from a parallel synthesizer. *IEEE Transactions on Audio Electroacoustics*, AU-21:298–305.

[Hunt and Black, 1996] Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 373–376.

[Kain and Macon, 1998] Kain, A. and Macon, M. W. (1998). Spectral voice conversion for text-to-speech synthesis. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 285–288.

[Kelly and Gerstman, 1961] Kelly, J. and Gerstman, L. (1961). An artificial talker driven from phonetic input. *Journal of the Acoustical Society of America*, Supplement 1 33:S35.

[Klatt, 1987] Klatt, D. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82:737–793.

[Klatt, 1980] Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67:971–995.

[Klatt and Klatt, 1990] Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857.

[Kuwabara and Sagisaka, 1995] Kuwabara, H. and Sagisaka, Y. (1995). Acoustic characteristics of speaker individuality: Control and conversion. *Speech Communication*, 16(2):165–174.

[Macon, 1996] Macon, M. W. (1996). *Speech Synthesis Based on Sinusoidal Modeling.* PhD thesis, Georgia Institute of Technology.

[Maeda, 1982] Maeda, S. (1982). A digital simulation method of the vocal tract system. *SpCom*, 1:199–229.

[Makhoul, 1975] Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.

[Malfrére et al., 1998] Malfrére, F., Dutoit, T., and Mertens, P. (1998). Automatic prosody generation using suprasegmental unit selection. In *Proc. of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, pages 323–328.

[Massaro, 1998] Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* MIT Press, Cambridge, MA.

[McAulay and Quatieri, 1986] McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(4):744–754.

[McKeown et al., 1997] McKeown, K., Pan, S., Shaw, J., Jordan, D., and Allen, B. (1997). Language generation for multimedia healthcare briefings. In *Proc. of the Fifth ACL Conf. on ANLP*, pages 277–282.

[Mermelstein, 1973] Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53:1070–1082.

[Michalski, 1997] Michalski, J. (1997). Internet and interactive technologies report: The fridge door. Research report, Edventure Holdings.

[Monaghan, 1990] Monaghan, A. (1990). Rhythm and stress-shift in speech synthesis. *Computer Speech and Language*, 4:71–78.

[Montgomery and Soo Hoo, 1982] Montgomery, A. A. and Soo Hoo, G. (1982). Animat: A set of programs to generate, edit and display sequences of vector-based images. *Behavioral Research Methods and Instrumentation*, 14:39–40.

[Moulines and Charpentier, 1990] Moulines, E. and Charpentier, F. (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453–467.

[Munhall et al., 1995] Munhall, K. G., Vatikiotis-Bateson, E., and Tohkura, Y. (1995). X-ray film database for speech research. *Journal of the Acoustical Society of America*, 98:1222–1224. http://pavlov.psyc.queensu.ca/faculty/munhall/x-ray/.

[Murray and Arnott, 1995] Murray, I. R. and Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16(4):369–390.

[Nakata and Mitsuoka, 1965] Nakata, K. and Mitsuoka, T. (1965). Phonemic transformation and control aspects of synthesis of connected speech. *Journal of Radio Research Laboratories*, 12:171–186.

[Nock et al., 1997] Nock, H. J., Gales, M. J. F., and Young, S. J. (1997). A comparative study of methods for phonetic decision-tree state clustering. In *Proc. Eurospeech*, pages 111–114.

[Olive, 1997] Olive, J. (1997). The talking computer: Text-to-speech synthesis. In Stork, D., editor, *Hal's Legacy: 2001's Computer as Dream and Reality*. MIT Press, Cambridge, MA.

[O'Shaughnessy, 1979] O'Shaughnessy, D. (1979). Linguistic features in fundamental frequency patterns. *Journal of Phonetics*, 7:119–145.

[Ostendorf and Veilleux, 1994] Ostendorf, M. and Veilleux, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, 20(1):27–54.

[Pan and McKeown, 1996] Pan, S. and McKeown, K. (1996). Spoken language generation in a multimedia system. In *Proceedings of ICSLP*, volume 1, Philadelphia.

[Pan and McKeown, 1998] Pan, S. and McKeown, K. (1998). Learning intonation rules for concept to speech generation. In *Proceedings of COLING/ACL'98*, Montreal, Canada.

[Parke, 1972] Parke, F. I. (1972). Computer generated animation of faces. In *Proc. ACM National Conference*, volume 1, pages 451–457.

[Parke, 1974] Parke, F. I. (1974). A parametric model for human faces. Technical Report UTEC-CSc-75-047, University of Utah, Salt Lake City, Utah.

[Parke, 1975] Parke, F. I. (1975). A model for human faces that allows speech synchronizated animation. *Computers and Graphics Journal*, 1:1–4.

[Parke, 1982] Parke, F. I. (1982). Parameterized models for facial animation. *IEEE Computer Graphics*, 2(9):61–68.

[Parke, 1991] Parke, F. I. (1991). *Control parameterization for facial animation.* Springer-Verlag, Tokyo.

[Pelachaud et al., 1996] Pelachaud, C., Badler, N. I., and Steedman, M. (1996). Generatin faciall expressions for speech. *Cognitive Science*, 20:1–46.

[Pelorson et al., 1980] Pelorson, X., Hirschberg, A., van Hassel, R. R., Wijnands, A. P. J., , and Auregan, Y. (1980). Signal models for low bit-rate coding of speech. *Journal of the Acoustical Society of America*, 68(3):780–791.

[Pelorson et al., 1994] Pelorson, X., Hirschberg, A., van Hassel, R. R., Wijnands, A. P. J., and Auregan, Y. (1994). Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. application to a modified two-mass model. *Journal of the Acoustical Society of America*, 96(6):3416–3431.

[Peterson et al., 1958] Peterson, G., Wang, W., and Sivertsen, E. (1958). Segmentation techniques in speech synthesis. *Journal of the Acoustical Society of America*, 30:739–742. Supplement 1 33.

[Pierrehumbert, 1980] Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA. Distributed by the Indiana University Linguistics Club.

[Pierrehumbert, 1981] Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America*, 70(4):985–995.

[Platt, 1985] Platt, S. M. (1985). *A Structural Model of the Human Face.* PhD thesis, University of Pennsylvania.

[Platt and Badler, 1981] Platt, S. M. and Badler, N. I. (1981). Animating facial expressions. *IEEE Computer Graphics*, 15:245–252.

[Pols, 1992] Pols, L. C. W. (1992). Quality assessment of text-to-speech synthesis by rule. In Furui, S. and Sondhi, M. M., editors, *Advances in Speech Signal Processing*, pages 387–416. Marcel Dekker, Inc., New York.

[Potter et al., 1947] Potter, R., Kopp, G., and Green, H. (1947). *Visible Speech.* van Nostrand, New York.

[Prevost, 1995] Prevost, S. (1995). *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation.* PhD thesis, University of Pennsylvania.

[Rabiner and Schafer, 1978] Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals.* Prentice-Hall, Englewood Cliffs, New Jersey.

[Riley, 1992] Riley, M. (1992). Tree-based modeling for speech synthesis. In Bailly, G. and Benoit, C., editors, *Talking Machines: Theories, Models, and Designs*, pages 265–273. Elsevier, Amsterdam.

[Ross and Ostendorf, 1996] Ross, K. and Ostendorf, M. (1996). Prediction of abstract prosodic labels for speech synthesis. *Computer, Speech and Language*, 10(3):155–185.

[Ross and Ostendorf, 1999] Ross, K. and Ostendorf, M. (1999). A dynamical system model for generating fundamental frequency for speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 7(3). to appear.

[Rubin et al., 1981] Rubin, P., Baer, T., and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70:321–328.

[Schroeter and Sondhi, 1992] Schroeter, J. and Sondhi, M. M. (1992). Speech coding based on physiological models of speech production. In Furui, S. and Sondhi, M. M., editors, *Advances in Speech Signal Processing*, pages 231–267. Marcel Dekker, Inc., New York.

[Schroeter and Sondhi, 1994] Schroeter, J. and Sondhi, M. M. (1994). Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1):133–150.

[Shadle et al., 1999] Shadle, C. H., Barney, A., and Davies, P. O. A. L. (1999). Fluid flow in a dynamic mechanical model of the vocal folds and tract. ii. implications for speech production studies. *Journal of the Acoustical Society of America*, 105(1):456–466.

[Shattuck-Hufnagel et al., 1994] Shattuck-Hufnagel, S., Ostendorf, M., and Ross, K. (1994). Stress shift and early pitch accent placement in lexical items in American English. *Journal of Phonetics*, 22:357–388.

[Silverman, 1987] Silverman, K. (1987). *The Structure and Processing of Fundamental Frequency Contours*. PhD thesis, Cambridge University.

[Silverman, 1993] Silverman, K. (1993). Customizing prosody in speech synthesis: names and addresses as a case in point. In *Proc. DARPA Workshop on Human Language Technology*, pages 317–322.

[Silverman et al., 1990] Silverman, K. E. A., Basson, S., and Levas, S. (1990). Evaluating synthesizer performance: Is segmental intelligibility enough? In *Proc. of the International Conference on Spoken Language Processing*, Kobe, Japan.

[Sproat, 1994] Sproat, R. (1994). English noun-phrase accent prediction for text-to-speech. *Computer Speech and Language*, 8:79–94.

[Sproat et al., 1998] Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K., and Edgington, M. (1998). Sable: A standard for tts markup. In *Proc. of the International Conference on Spoken Language Processing*, volume 5, pages 1719–1722.

[Sproat, 1998] Sproat, R. W., editor (1998). *Multilingual Text-to-Speech Synthesis: the Bell Labs Approach*. Kluwer, Dordrecht, the Netherlands.

[Steedman, 1996] Steedman, M. (1996). Representing discourse information for spoken dialogue generation. In *Proceedings of the International Symposium on Spoken Dialogue*, pages 89–92, Philadelphia.

[Stevens and Bickley, 1991] Stevens, K. N. and Bickley, C. A. (1991). Constraints among parameters simplify control of Klatt formant synthesizer. *Journal of Phonetics*, 19:161–174.

[Stone and Lundberg, 1996] Stone, M. and Lundberg, L. (1996). Three-dimensional tongue surface shapes of english consonants and vowels. *Journal of the Acoustical Society of America*, 99:1–10.

[Stylianou et al., 1995] Stylianou, Y., Cappé, O., and Moulines, E. (1995). Statistical methods for voice quality transformation. *Proc. Eurospeech*, pages 447–450.

[Summerfield et al., 1989] Summerfield, A. Q., MacLoed, A., McGrath, M., and Brooke, M. (1989). *Lips, teeth, and the benefits of lipreading.* Elsevier Science Publishers, Amsterdam.

[Taylor and Black, 1994] Taylor, P. and Black, A. (1994). Synthesizing conversational intonation from a linguistically rich input. In *Proc. of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 175–178.

[Terzoupoulous and Waters, 1990] Terzoupoulous, D. and Waters, K. (1990). Muscle parameter estimation from image sequences. *SIGGRAPH Facial Animation Course Notes*, pages 146–155.

[Titze, 1994] Titze, I. R. (1994). *Principles of Voice Production.* Prentice-Hall.

[Traber, 1992] Traber, C. (1992). F_0 generation with a data base of natural F_0 patterns and with a neural network. In Bailly, G. and Benoit, C., editors, *Talking Machines: Theories, Models, and Designs*, pages 287–304. Elsevier, Amsterdam.

[Trittin and de Santos y Lleó, 1995] Trittin, P. J. and de Santos y Lleó, A. (1995). Voice quality analysis of male and female Spanish speakers. *Speech Communication*, 16(4):359–368.

[van Bezooijen and van Heuven, 1997] van Bezooijen, R. and van Heuven, V. (1997). Assessment of Synthesis Systems. In Gibbon, D., Moore, R., and Winsky, R., editors, *Handbook of standards and resources for spoken language systems*, chapter 12, pages 481–563. Walter de Gruyter & Co, Berlin.

[van Santen, 1993] van Santen, J. (1993). Quantitative modeling of segmental duration. In *Proceedings of the Human Language Technology Workshop*, pages 323–328, Princeton, NJ.

[van Santen et al., 1998] van Santen, J., Pols, L., Abe, M., Kahn, D., Keller, E., and Vonwiller, J. (1998). Report on the third ESCA TTS workshop evaluation procedure. In *Proc. of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, pages 329–331.

[van Santen and Möbius, 1997] van Santen, J. P. H. and Möbius, B. (1997). Modeling pitch accent curves. In *Proceedings ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications.*

[Vatikiotis-Bateson et al., 1996] Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., and Terzopoulous, D. (1996). *The dynamics of audiovisual behavior in speech.* Springer-Verlag, New York. Stork, D. G. and Hennecke, M. E. (Eds.).

[Voiers et al., 1972] Voiers, W. D., Sharpley, A. D., and Hehmsoth, C. J. (1972). Research on diagnostic evaluation of speech intelligibility. Research Report AFCRL-72-0694, Cambridge Research Laboratories.

[Wang and Hirschberg, 1992] Wang, M. and Hirschberg, J. (1992). Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196.

[Waters, 1987] Waters, K. (1987). A muscle model for animating three-dimensional facial expression. *IEEE Computer Graphics*, 21:17–24.

[Waters and Levergood, 1993] Waters, K. and Levergood, T. M. (1993). Decface: An automatic lip-synchronization algorithm for synthetic faces. Technical Report CRL 93/4, Digital Equipment Corporation, Cambridge Research Laboratory.

[Westbury, 1994] Westbury, J. R. (1994). *X-Ray Microbeam Speech Production Database User's Handbook*. Waisman Center on Mental Retardation and Human Development, University of Wisconsin, Madison, WI.

[Yarowsky, 1997] Yarowsky, D. (1997). Homograph disambiguation in text-to-speech synthesis. In SSOH, editor, *Progress in Speech Synthesis*, pages 157–172. Springer, New York, NY.

[Young and Fallside, 1979] Young, S. and Fallside, F. (1979). Speech synthesis from concept: A method for speech output from information systems. *Journal of the Acoustical Society of America*, 66(3):685–695.

A THE AUGUST 1998 NSF WORKSHOP ATTENDEES

- Alex Acero, Microsoft Corporation
- Mary Beckman, Ohio State University
- Alan Black, Edinburgh University
- Olin Campbell, Vanderbilt University
- Susan Chipman, Office of Naval Research
- Chris Cieri, University of Pennsylvania, Linguistic Data Consortium
- Jim Flanagan, Rutgers University
- Jim Fruchterman, Arkenstone
- Francisco Galanes, Entropic Research Laboratories
- Helen Gigley, Office of Naval Research
- Jim Glass, MIT
- Harry Hersh, Fidelity Investments
- Sue Hertz, Eloquent Technology, Inc., and Cornell University
- Andrew Hunt, Sun Microsystems Laboratories
- Frederick Jelinek, Johns Hopkins University
- Kevin Lenzo, Carnegie Mellon University
- Kevin Knight, Information Sciences Institute
- Michael Macon, Oregon Graduate Institute
- Marian Macchi, E-Speech Corporation
- Dominic Massaro, University of California, Santa Cruz
- Richard McGowan, Sensimetrics Corporation
- Kathy McKeown, Columbia University
- Joe Olive, Lucent Technologies
- Mari Ostendorf, Boston University
- Dave Pallett, NIST

- Louis Pols, Institute of Phonetic Sciences
- Patti Price, SRI International
- T. V. Raman, Adobe Systems
- Juergen Schroeter, AT&T
- Kim Silverman, Apple Computer, Inc.
- Allen Sears, DARPA
- Dan Sinder, Rutgers University
- Richard Sproat, Lucent Technologies
- Gary Strong, NSF
- Ann Syrdal, AT&T
- David Talkin, General Magic
- Jan van Santen, Lucent Technologies
- Jon Yi, MIT

B GLOSSARY

Accent. The "prominence" or degree of "emphasis" associated with spoken words. Roughly speaking, a word is said to be accented if it is fairly prominent, and deaccented if it lacks any prominence.

Articulator. One of several organs — e.g., the lips, jaw, tongue tip, tongue blade and tongue body — that move during speech, and alter the shape of the vocal tract.

Articulatory Synthesis. A *parametric* approach to synthesis that is based on modeling the movement of articulators, and the acoustic consequences of those movements.

Coarticulation. The influence of one speech sound upon an adjacent or nearby speech sound.

Concatenated Voice Response. A speech output system which constructs utterances by splicing together raw speech segments, usually corresponding to whole words or phrases. Unlike concatenative synthesis systems, concatenated voice response systems usually do not attempt to ensure good fits between the spliced segments.

Concatenative Synthesis. An approach to speech synthesis that involves splicing processed segments of real speech together, with signal processing and unit selection methods being used to optimize the fits between the spliced segments.

Concept-to-Speech. The generation of speech from "semantic" representations (usually the output of a language generation system) rather than from text.

Copy Synthesis. Synthesis of speech from parameters that are derived (and often tuned by hand) from an original utterance that is being "copied".

Deaccented. See *Accent.*

Dialog System. A speech or natural-language system that interacts with a user in a cooperative manner to achieve a certain goal or task.

Diphone. A unit in a concatenative synthesis system that consists of the transition between two adjacent sounds.

Discourse Structure. A formal model of the organization of utterances into a conversation or monolog, including facets such as topic structure and intent.

Excitation. A periodic or noise source that excites the acoustic resonances of, e.g., the vocal tract. The term may refer to the physical source or a simple mathematical model as used in computer waveform generation.

F_0. See *Fundamental Frequency.*

Formant. One of the resonances of the vocal tract.

Formant Synthesis. A *parametric* approach to synthesis that starts with a model that generates formants and other acoustic features of the voice.

Fricative A "noisy" speech sound such as an "s", an "f" or a "v", that involves a partial constriction in the vocal tract.

Fundamental Frequency. The frequency, usually measured in Hertz, of a speaker's voice. The *fundamental frequency contour* is the primary physical manifestation of intonation.

Glottis. The vocal cords, or vocal folds, a pair of muscles in the larynx that vibrate to produce voiced sounds, and remain open to produce voiceless or breathy sounds.

Homographs. A set of words that are spelled the same, but have different meanings and/or different pronunciations.

Interactive Voice Response Systems. An automatic information access or call routing system that uses speech synthesis or concatenated voice response and automatic speech recognition technology.

Intonation. A level of linguistic representation which determines, among other things, the fundamental frequency contour. In popular parlance it is often termed "inflection".

Language Modeling. A model that determines the likelihood of various sequences of words. A good language model will give a high likelihood for sentences that are both syntactically well-formed and semantically and pragmatically plausible, and a low likelihood for others.

Laryngeal. Pertaining to the larynx or *glottis.*

Letter-to-Sound Rules. A set of rules that converts from spellings into pronunciations.

Markup. A scheme for annotating a text with information useful for a particular task. For instance, HTML is a markup language for web documents, where the particular task is display by a browser.

Morphology. The study of the formal properties, in particular the structure of words.

Parametric Models. Models that are based on a (usually small) set of controllable parameters.

Periodic Sounds. Speech sounds that have a regular periodic voiced source (or *excitation*). Vowels, for instance, are periodic sounds, whereas *fricatives* are not.

Phoneme. A basic contrastive sound of a language. By "contrastive" we mean that it is possible to find two words that differ in exactly one sound: in that case the two sounds involved would be phonemes. Thus "t" (*tack*) is a phoneme of English, as is "b" (*back*). However, the "t" in "top" and the "t" in American English "butter", which are quite different, are nonetheless not different phonemes since they are not contrastive in the defined sense.

Phone. A basic sound of a language. The "t" in "top" and the "t" in American English "butter" would be different phones; cf. *phoneme.*

Phrasing. See *Prosodic Phrase.*

Pragmatics. The study of the meanings of linguistic utterances in context.

Prosody. A level of linguistic representation that includes *intonation*, *accent*, and *prosodic phrasing*. The acoustic correlates associated with prosody include fundamental frequency, segmental durations (or timing), and energy (or other manifestations of vocal effort).

Prosodic Phrase. A (usually) meaningful chunk of speech that is delimited by various prosodic cues. A prosodic phrase is often marked by such effects as a final pause, final segmental lengthening, a final drop or rise in the fundamental frequency, inter alia.

Rule-Based Systems. Typically used to refer to a system that is based upon hand-written rules, rather than on trainable parameters.

Segment. A basic sound unit of speech that corresponds roughly to a single phone or phoneme.

Source-Filter Model. A model of speech synthesis that assumes a more-or-less clean separation between the glottal or other source excitation and the vocal tract.

Spectral Tilt. If one computes a spectrum of a short window of speech, there is a constant energy roll-off as one looks at higher and higher frequencies in the spectrum. This roll-off is spectral tilt.

Symbolic Representation. A representation involving discrete symbols, or discrete-valued parameters, as opposed to continuous-valued functions.

Text Analysis. The analysis of the input text into a symbolic representation from which waveform generation can then proceed.

Text-to-Speech. The generation of speech from unrestricted input text.

TTS. See *Text-to-Speech*.

Vocal Tract. The acoustic tube formed by the oral and nasal passages.

Vocal Tract Transfer Function. A mathematical function that describes the acoustic properties of the vocal tract in a given state. The vocal tract transfer function changes as one speaks and thereby changes the shape of the vocal tract.

Voice Browser. An aural analog of a web browser such as Netscape Communicator.

Voice Conversion. The automatic manipulation of one voice to sound like another voice.

Voice Quality. An impressionistic description of the voice in terms of such properties as its “harshness”, “shrillness”, “muffledness”, “breathiness”, etc. In speech synthesis evaluation, it is often used in contradistinction to other evaluable aspects of the synthesis such as the naturalness of the prosody or the accuracy of the text analysis.

Waveform Generation. The generation of a listenable speech waveform from a symbolic representation of the message to be spoken.