Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9932–9937, October 1995
Colloquium Paper

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

Models of speech synthesis

ROLF CARLSON

Department of Speech Communication and Music Acoustics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

ABSTRACT **The term "speech synthesis" has been used for diverse technical approaches. In this paper, some of the approaches used to generate synthetic speech in a text-to-speech system are reviewed, and some of the basic motivations for choosing one method over another are discussed. It is important to keep in mind, however, that speech synthesis models are needed not just for speech generation but to help us understand how speech is created, or even how articulation can explain language structure. General issues such as the synthesis of different voices, accents, and multiple languages are discussed as special challenges facing the speech synthesis community.**

The term "speech synthesis" has been used for diverse technical approaches. Unfortunately, any speech output from computers has been claimed to be speech synthesis, perhaps with the exception of playback of recorded speech.* Some of the approaches used to generate true synthetic speech as well as high-quality waveform concatenation methods are presented below.

Knowledge About Natural Speech

Synthesis development can be grouped into three main categories: acoustic models, articulatory models, and models based on the coding of natural speech. The last group includes both predictive coding and concatenative synthesis using speech waveforms. Acoustic and articulatory models have had a long history of development, while natural speech models represent a somewhat newer field. The first commercial systems were based on the acoustic terminal analog synthesizer. However, at that time, the voice quality was not good enough for general use, and approaches based on coding attracted increased interest. Articulatory models have been under continuous development, but so far this field has not been exposed to commercial applications due to incomplete models and high processing costs.

We can position the different synthesis methods along a "knowledge about speech" scale. Obviously, articulatory synthesis needs considerable understanding of the speech act itself, while models based on coding use such knowledge only to a limited extent. All synthesis methods have to model something that is partly unknown. Unfortunately, artificial obstacles due to simplifications or lack of coverage will also be introduced. A trend in current speech technology, both in speech understanding and speech production, is to avoid explicit formulation of knowledge and to use automatic methods to aid the development of the system. Since such analysis methods lack the human ability to generalize, the generalization has to be present in the data itself. Thus, these methods need large amounts of speech data. Models working close to the waveform are now typically making use of increased unit sizes while still modeling prosody by rule. In the middle of the scale, "formant synthesis" is moving toward the articulatory models by looking for "higher-level parameters" or to larger prestored units. Articulatory synthesis, hampered by lack of data, still has some way to go but is yielding improved quality, due mostly to advanced analysis-synthesis techniques.

Flexibility and Technical Dimensions

The synthesis field can be viewed from many different angles. We can group the models along a "flexibility" scale. Multilingual systems demand flexibility. Individual voices, speaking styles, and accents also need a flexible system in which explicit transformations can be modeled. Most of these variations are continuous rather than discrete. The importance of separating the modeling of speech knowledge from acoustic realization must be emphasized in this context.

In the overview by Furui (6), synthesis techniques are divided into three main classes: waveform coding, analysis-synthesis, and synthesis by rule. The analysis-synthesis method is defined as a method in which human speech is transformed into parameter sequences, which are stored. The output is created by a synthesis based on concatenation of the prestored parameters. In a synthesis-by-rule system the output is generated with the help of transformation rules that control the synthesis model such as a vocal tract model, a terminal analog, or some kind of coding.

It is not an easy task to place different synthesis methods into unique classes. Some of the common "labels" are often used to characterize a complete system rather than the model it stands for. A rule-based system using waveform coding is a perfectly possible combination, as is speech coding using a terminal analog or a rule-based diphone system using an articulatory model. In the following pages, synthesis models will be described from two different perspectives: the sound-generating part and the control part of the system.

THE SOUND-GENERATING PART

The sound-generating part of the synthesis system can be divided into two subclasses, depending on the dimensions in which the model is controlled. A vocal tract model can be controlled by spectral parameters such as frequency and bandwidth or shape parameters such as size and length. The source model that excites the vocal tract usually has parameters to control the shape of the source waveform. The combination of time-based and frequency-based controls is powerful in the sense that each part of the system is expressed in its most

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

*The foundations for speech synthesis based on acoustical or articulatory modeling can be found in Fant (1), Holmes *et al.* (2), Flanagan (3), Klatt (4), and Allen *et al* (5). The paper by Klatt (73) gives an extensive review of the developments in speech synthesis technology.

explanatory dimensions. A drawback of the combined approach can be that it makes interaction between the source and the filter difficult. However, the merits seem to outweigh the drawbacks.

Simple Waveform Concatenation

The most radical solution to the synthesizer problem is simply to have a set of prerecorded messages stored for reproduction. Simple coding of the speech wave might be performed in order to reduce the amount of memory needed. The quality is high, but the usage is limited to applications with few messages. If units smaller than sentences are used, the quality degenerates because of the problem of connecting the pieces without distortion and overcoming prosodic inconsistencies. One important and often forgotten aspect in this context is that a vocabulary change can be an expensive and time-consuming process, since the same speaker and recording facility have to be used as with the original material. The whole system might have to be completely rebuilt in order to maintain equal quality of the speech segments.

Analysis-Synthesis Systems

Synthesis systems based on coding have as long a history as the vocoder. The underlying philosophy is that natural speech is analyzed and stored in such a way that it can be assembled into new utterances. Synthesizers such as the systems from AT&T Bell Laboratories (7–9), Nippon Telephone & Telegraph (NTT) (10, 11), and ATR Interpreting Telephone Research Laboratories (ATR) (12, 13) are based on the source-filter technique where the filter is represented in terms of linear predictive coding (LPC) or equivalent parameters. This filter is excited by a source model that can be of the same kind as the one used in terminal analog systems. The source must be able to handle all types of sounds: voiced and unvoiced vowels and consonants.

Considerable success has been achieved by systems that base sound generation on concatenation of natural speech units (14). Sophisticated techniques have been developed to manipulate these units, especially with respect to duration and fundamental frequency. The most important aspects of prosody can be imposed on synthetic speech without considerable loss of quality. The pitch-synchronous overlap-add approach (PSOLA) (15) methods are based on concatenation of waveform pieces. The frequency domain approach (FD-PSOLA) is used to modify the spectral characteristics of the signal; the time domain approach (TD-PSOLA) provides efficient solutions for real-time implementation of synthesis systems. Earlier systems like SOLA (16) and systems for divers' speech restoration also did direct processing of the waveform (17).

Fig. 1 shows the basic function of a PSOLA-type system. A data base of carefully selected utterances is recorded, and each pitch period is marked. The speech signal is split into a sequence of windowed samples of the speech wave. At resynthesis time the waveforms are added according to the desired pitch, amplitude, and duration.

Source Models

The traditional source model for the voiced segments has been a simple or double impulse. This is one reason why text-to-speech systems from the 1980s have had serious problems, especially when different voices are modeled. While the male voice sometimes has been regarded to be generally acceptable, an improved glottal source will open the way to more realistic synthesis of child and female voices and also to more naturalness and variation in male voices.

Most source models work in the time domain with different controls to manipulate the pulse shape (19–23). One version

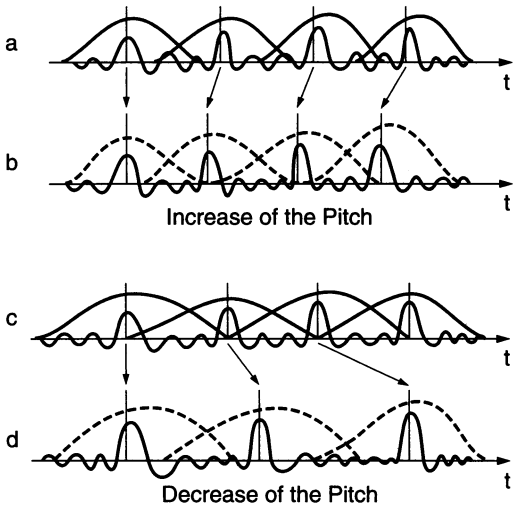

FIG. 1. Example of the PSOLA method (18).

of such a voice source is the LF-model (24). It has a truncated exponential sinusoid followed by a variable cut-off 6 dB/octave low-pass filter modeling the effect of the return phase, that is, the time from maximum excitation of the vocal tract to complete closure of the vocal folds. Fig. 2 explains the function of the control parameters. In addition to the amplitude and fundamental frequency control, two parameters influence the amplitudes of the two to three lowest harmonics, and one parameter influences the high-frequency content of the spectrum. Another vocal source parameter is the diplophonia

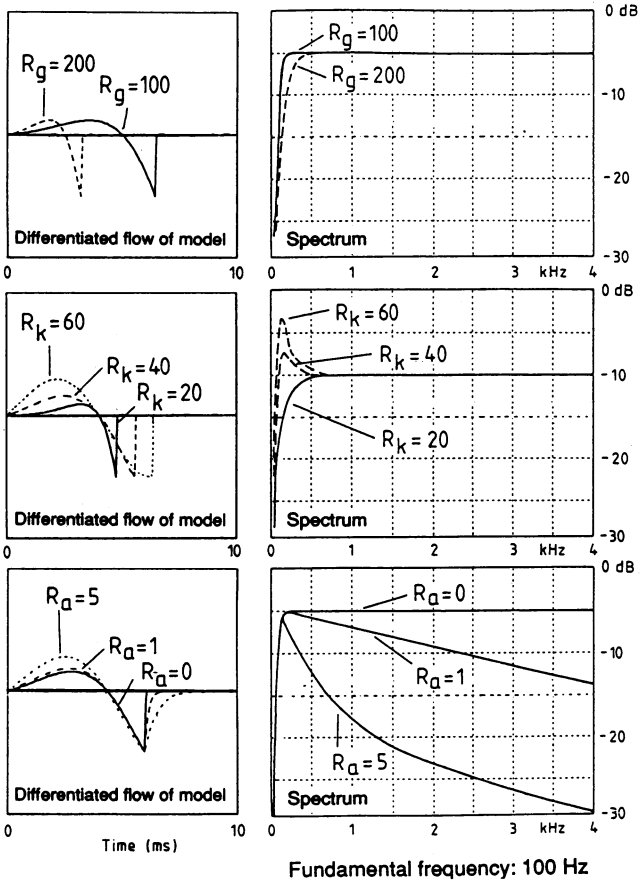

FIG. 2. Influence of the parameters R^g, R^k, and R^a on the differentiated glottal flow pulse shape and spectrum (25). The spectra are preemphasized by 6 dB/octave.

parameter (22) with which creak, laryngalization, or diplophonia can be simulated. This parameter influences the function of the voiced source in such a way that every second pulse is lowered in amplitude and shifted in time.

The next generation of source models has to include adequate modeling of noise excitation in order to synthesize a natural change between voiced and unvoiced segments. The work of Rothenberg (26) can serve as a guide for future implementations. In some earlier work at the Royal Institute of Technology (KTH), we were able to use a model that included a noise source (27). High-quality synthesis of extralinguistic sounds such as laughter could be produced with this model in addition to reasonable voiced-unvoiced transitions.

The acoustic interactions between the glottal source and the vocal tract also must be considered (28). One of the major factors in this respect is the varying bandwidth of the formants. This is especially true for the first formant, which can be heavily damped during the open phase of the glottal source. However, it is not clear that such a variation can be perceived by a listener (29). Listeners tend to be rather insensitive to bandwidth variation (3). In more complex models the output is glottal opening rather than glottal flow. The subglottal cavities can then be included in an articulatory model.

Noise sources have attracted much less research effort than the voiced source. However, some aspects have been discussed by Stevens (30), Shadle (31), and Badin and Fant (32). Today, simple white noise typically is filtered by resonances that are stationary within each parameter frame. The new synthesizers do have some interaction between the voice source and the noise source, but the interaction is rather primitive. Transient sounds and aspiration dependent on vocal cord opening are still under development.

Formant-Based Terminal Analog

The traditional text-to-speech systems use a terminal analog based on formant filters. The vocal tract is simulated by a sequence of second-order filters in cascade while a parallel structure is used mostly for the synthesis of consonants. One important advantage of a cascade synthesizer is the automatic setting of formant amplitudes. The disadvantage is that it sometimes can be hard to do detailed spectral matching between natural and synthesized spectra because of the simplified model. Parallel synthesizers such as that of Holmes (33) do not have this limitation.

The Klatt model is widely used in research for both general synthesis purposes and perceptual experiments. A simplified version of this system is used in all commercial products that stem from synthesis work at the Massachusetts Institute of Technology (MIT): MITalk (5), DECtalk, and the system at Speech Technology Laboratory (34). An improved version of the system has been commercialized as a research vehicle by Sensimetrics Corporation (35). Similar configurations were used in the ESPRIT/Polyglot project (36).

A formant terminal analog GLOVE (37), based on the OVE synthesizer (38), has been developed at KTH and is used in current text-to-speech modeling (39, 40). The main difference between these and the Klatt model is the manner in which consonants are modeled. In the OVE a fricative is filtered by a zero-pole-pole configuration rather than by a parallel system. The same is true for the nasal branch of the synthesizer.

New parameters have been added to the terminal analog model so that it is now possible to simulate most human voices and to replicate an utterance without noticeable quality reduction. However, it is interesting to note that some voices are easier to model than others. Despite the progress, speech quality is not natural enough in all applications of text to speech. The main reasons for the limited success in formant-based synthesis can be explained by incomplete phonetic knowledge. It should be noted that the transfer of knowledge from phonetics to speech technology has not been an easy process. Another reason is that the efforts using formant synthesis have not explored control methods other than the explicit rule-based description.

Higher-Level Parameters

Since the control of a formant synthesizer can be a very complex task, some efforts have been made to help the developer. The "higher-level parameters" (35, 41) explore an intermediate level that is more understandable from the developer's point of view compared to the detailed synthesizer specifications. The goal of this approach is to find a synthesis framework to simplify the process and to incorporate the constraints that are known to exist within the process. A formant frequency should not have to be adjusted specifically by the rule developer depending on nasality or glottal opening. This type of adjustment might be better handled automatically according to a well-specified model. The same process should occur with other parameters such as bandwidths and glottal settings. The approach requires a detailed understanding of the relationship between acoustic and articulatory phonetics.

Articulatory Models

An articulatory model will ultimately be the most interesting and flexible solution for the sound-generating part of text-to-speech systems. Development is also advancing in this area, but the lack of reliable articulatory data and appropriate control strategies still presents challenges. One possible solution that has attracted interest is to automatically train neural networks to control such a synthesizer. Rahim *et al.* (42) and Bailly *et al.* (43) have explored such methods.

Articulatory models, now under improvement, stem from basic work carried out at such laboratories as AT&T Bell Laboratories, MIT, and KTH. At each time interval, an approximation of the vocal tract is used either to calculate the corresponding transfer function or to directly filter a source waveform. Different vocal tract models have been used based on varying assumptions and simplifications. The models by Flanagan *et al.* (44), Coker (45), and Mermelstein (46) have been studied by many researchers in their development of current articulatory synthesis.

The term "articulatory modeling" is often used rather loosely. Only part of the synthesis model is usually described in physical terms, while the remaining part is described in a simplified manner. Compare, for example, the difference between a tube model that models a static shape of the vocal tract with a dynamic physical model that actually describes how the articulators move. Thus, a complete articulatory model for speech synthesis has to include several transformations. The relationship between an articulatory gesture and a sequence of vocal tract shapes must be modeled. Each shape must be transformed into some kind of tube model with its acoustic characteristics. The acoustics of the vocal tract can then be modeled in terms of an electronic network. At this point, the developer can choose to use the network as such to filter the source signal. Alternatively, the acoustics of the network can be expressed in terms of resonances that can control a formant-based synthesizer. The main difference is the domain, time, or frequency in which the acoustics is simulated.

The developer has to choose at which level the controlling part of the synthesis system should connect to the synthesis model. All levels are possible, and many have been used. One of the pioneering efforts using articulatory synthesis as part of a text-to-speech system was done by AT&T Bell Laboratories (45). Lip, jaw, and tongue positions were controlled by rule. The final synthesis step was done by a formant-based terminal analog. Current efforts at KTH by Lin and Fant (47) use a parallel synthesizer with parameters derived from an articu-

latory model. In the development of articulatory modeling for text to speech, we can take advantage of parallel work on speech coding based on articulatory modeling (48). This work focuses not only on synthesizing speech but also on how to extract appropriate vocal tract configurations. Thus, it will also help us to get articulatory data through an analysis-synthesis procedure. This section has not dealt with the important work carried out to describe speech production in terms of physical models. The inclusion of such models still lies in the future, beyond the next generation of text to speech systems, but the results of these experiments will improve the current articulatory and terminal analog models.

THE CONTROL PART

Models of segmental coarticulation and other phonetic factors are an important part of a text-to-speech system. The control part of a synthesis system calculates the parameter values at each time frame. Two main types of approaches can be distinguished: rule-based methods that use an explicit formulation of existing knowledge and library-based methods that replace rules by a collection of segment combinations. Clearly, each approach has its advantages. If the data are coded in terms of targets and slopes, we need methods to calculate the parameter tracks. The efforts of Holmes *et al.* (2) and the filtered square wave approach by Liljencrants (49) provide some classical examples in this context.

To illustrate the problem, I have chosen some recent work by Slater and Hawkins (50). The work was motivated by the need to improve the rule system in a text-to-speech system for British English. Data for the second formant frequency at the onset of a vowel after a velar stop and at the midpoint in the vowel were analyzed, and, as expected, a clear correlation between the frequencies at these positions could be noted. The data could be described by one, two, or three regression lines, depending on the need for accuracy. This could then be modeled by a set of rules. As an alternative, all data points can be listed. Unfortunately, the regression lines change their coefficients depending on a number of factors such as position and stress. To increase the coverage, we need to expand the analysis window and include more dimensions or increase the number of units. Eventually, we will reach a point where the rules become too complex or the data collection becomes too huge. This is the point where new dimensions such as articulatory parameters might be the ultimate solution.

Concatenation of Units

One of the major problems in concatenative synthesis is to make the best selection of units and describe how to combine them. Two major factors create problems: distortion because of spectral discontinuity at the connecting points and distortion because of the limited size of the unit set. Systems using elements of different lengths depending on the target phoneme and its function have been explored by several research groups. In a paper by Olive (8), a new method for concatenating "acoustic inventory elements" of different sizes is described. The system, developed at ATR, is also based on nonuniform units (13).

Special methods to generate a unit inventory have been proposed by the research group at NTT in Japan (10, 11). The synthesis allophones are selected with the help of the context-oriented clustering (COC) method. The COC searches for the phoneme sequences of different sizes that best describe the phoneme realization.

The context-oriented clustering approach is a good illustration of a current trend in speech synthesis: automatic methods based on data bases. The studies are concerned with much wider phonetic contexts than before. (It might be appropriate to remind the reader of similar trends in speech recognition.) One cannot take into account all possible coarticulation effects by simply increasing the number of units. At some point, the total number might be too high or some units might be based on very few observations. In this case a normalization of data might be a good solution before the actual unit is chosen. The system will become a rule-based system. However, the rules can be automatically trained from data in the same way as speech recognition (51).

Rules and Notations

Development tools for text-to-speech systems have attracted considerable efforts. The publication of *The Sound Pattern of English* by Chomsky and Halle (52) impelled a new kind of synthesis system based on rewrite rules. Their ideas inspired researchers to create special rule compilers for text-to-speech developments in the early 1970s. New software is still being developed according to this basic principle, but the implementations vary depending on the developer's tastes. It is important to note that crucial decisions often are hidden in the systems. The rules might operate rule by rule or segment by segment. Other important decisions are based on the following questions: How is the backtrack organized? Can nonlinear phonology be used (53), as in the systems described by Hertz (54, 55) and the Institute for Perception Research (56, 57)? Are the default values in the phoneme library primarily referred to by labels or features? These questions might seem trivial, but we see many examples of how the explicit design of a system influences the thinking of the researcher.

Automatic Learning

Synthesis has traditionally been based on very labor-intensive optimization work. Until recently, the notion of analysis by synthesis had been explored mainly by manual comparisons between hand-tuned spectral slices and a reference spectrum. The work of Holmes and Pearce (58) is a good example of how to speed up this process. With the help of a synthesis model, spectra are automatically matched against analyzed speech. Automatic techniques, such as this, will probably also play an important role in making speaker-dependent adjustments. One advantage of these methods is that the optimization is done in the same framework as that to be used in the production. The synthesizer constraints are thus already imposed in the initial state.

Methods for pitch-synchronous analysis will be of major importance in this context. Experiments such as the one presented by Talkin and Rowley (59) will lead to better estimates of pitch and vocal tract shape. These automatic procedures will, in the future, make it possible to gather a large amount of data. Lack of glottal source data currently is a major obstacle for the development of speech synthesis with improved naturalness.

Given that we have a collection of parameter data from analyzed speech corpora, we are in a good position to look for coarticulation rules and context-dependent variations. The collection of speech corpora also facilitates the possibilities of testing duration and intonation models (60–63).

SPEAKING CHARACTERISTICS AND SPEAKING STYLES

Currently available text-to-speech systems are not characterized by a great amount of flexibility, especially not when it comes to variations in voice or speaking style. On the contrary, the emphasis has been on a neutral way of reading, modeled after the reading of nonrelated sentences. There is, however, a very practical need for different speaking styles in text-to-speech systems. Such systems are now used in a variety of applications, and many more are projected as the quality is

improved. The range of applications demands a variation close to that found in human speakers. General use in reading stock quotations, weather reports, electronic mail, or warning messages are examples in which humans would choose rather different ways of reading. Apart from these practical needs in text-to-speech systems, there is the scientific interest in formulating our understanding of human speech variability in explicit models.

The current ambition in speech synthesis research is to model natural speech at a global level, allowing for changes of speaker characteristics and speaking style. One obvious reason is the limited success in enhancing the general speech quality by only improving the segmental models. The speaker-specific aspects are regarded as playing a very important role in the acceptability of synthetic speech. This is especially true when the systems are used to signal semantic and pragmatic knowledge.

One interesting effort to include speaker characteristics in a complex system has been reported by the ATR group in Japan. The basic concept is to preserve speaker characteristics in interpreting systems (64). The proposed voice conversion technique consists of two steps: mapping code book generation of LPC parameters and a conversion synthesis using the mapping code book. The effort has stimulated much discussion, especially considering the application as such. The method has been extended from a frame-by-frame transformation to a segment-by-segment transformation (65).

One concern with this type of effort is that the speaker characteristics are specified through training without a specific higher-level model of the speaker. It would be helpful if the speaker characteristics could be modeled by a limited number of parameters. Only a small number of sentences might in this case be needed to adjust the synthesis to one specific speaker. The needs in both speech synthesis and speech recognition are very similar in this respect.

A voice conversion system that combines the PSOLA technique for modifying prosody with a source-filter decomposition that enables spectral transformations has been proposed (66).

Duration-dependent vowel reduction has been another topic of research in this area. It seems that vowel reduction as a function of speech tempo is a speaker-dependent factor (67). Duration and intonation structures and pause insertion strategies reflecting variability in the dynamic speaking style are other important speaker-dependent factors. Parameters such as consonant-vowel ratio and source dynamics are typical parameters that must be considered in addition to basic physiological variations.

The differences between male and female speech have been studied by a few researchers (22, 68). A few systems, such as that of Syrdal (69), use a female voice as a reference speaker. The male voice differs from the female voice in many respects in addition to the physiological aspects. To a great extent, speaking habits are formed by the social environment, dialect region, sex, education, and by a communicative situation that may require formal or informal speech. A speaker's characteristics must be viewed as a complete description of the speaker in which all aspects are linked to each other in a unique framework (70, 71).

The ultimate test of our descriptions is our ability to successfully synthesize not only different voices and accents but also different speaking styles (72). Appropriate modeling of these factors will increase both the naturalness and intelligibility of synthetic speech.

Multilingual Synthesis

Many societies in the world are increasingly multilingual. The situation in Europe is an especially striking example of this. Most of the population is in touch with more than one language. This is natural in multilingual societies such as Switzerland and Belgium. Most schools in Europe have foreign languages on their mandatory curriculum. With the opening of the borders in Europe, more and more people will be in direct contact with several languages on an almost daily basis. For this reason, text-to-speech devices, whether they are used professionally or not, ought to have a multilingual capability.

Based on this understanding, many synthesis efforts are multilingual in nature. The Polyglot project, supported by the European ESPRIT program, was a joint effort by several laboratories in several countries. The common software in this project was, to a great extent, language independent, and the language-specific features were specified by rules, lexica, and definitions rather than by the software itself. This is also the key to the multilingual effort at KTH. About one-third of the systems delivered by the company INFOVOX are multilingual. The synthesis work pursued at companies such as ATR, CNET, DEC, and AT&T Bell Laboratories is also multilingual. It is interesting to see that the world's research community is rather small. Several of the efforts are joint ventures such as the CNET and CSTR British synthesis and the cooperation between Japanese (ATR) and U.S. partners. The Japanese company Matsushita even has a U.S. branch (STL) for its English effort, originally based on MITalk.

Speech Quality

The ultimate goal for synthesis research, with few exceptions, is to produce the highest speech quality possible. The quality and the intelligibility of speech are usually very difficult to measure. No single test is able to pinpoint where the problems lie. The Department of Psychology at the University of Indiana started a new wave of innovation in evaluation of synthesis systems to which a number of groups have made subsequent substantial contributions. But we are still looking for a simple way to measure progress quickly and reliably as we continue development of speech synthesis systems. The recent work that has been done in the ESPRIT/SAM projects, the COCOSDA group, and special workshops will set new standards for the future.

CONCLUDING REMARKS

In this paper a number of different synthesis methods and research goals to improve current text-to-speech systems have been touched on. It might be appropriate to remind the reader that nearly all methods are based on historical developments, where new knowledge has been added piece by piece to old knowledge rather than by a sudden change of approach. Perhaps the most dramatic change is in the field of synthesis tools rather than in the understanding of the "speech code." However, considerable progress can be seen in terms of improved speech synthesis quality. Today, speech synthesis is an appreciated facility even outside the research world, especially as applied to speaking aids for persons with disabilities. New synthesis techniques under development in speech research laboratories will play a key role in future man-machine interaction.

I would like to thank Björn Granström for valuable discussions during the preparation of this paper. This work has been supported by grants from the Swedish National Board for Technical Development.

1. Fant, G. (1960) *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands).
2. Holmes, J., Mattingly, I. G. & Shearme, J. N. (1964) *Lang. Speech* **7**, 127–143.
3. Flanagan, J. L. (1972) *Speech Analysis, Synthesis and Perception* (Springer, Berlin).
4. Klatt, D. K. (1976) *IEEE Trans.* **ASSP-24.**

5. Allen, J., Hunnicutt, M. S. & Klatt, D. (1987) *The MITalk System* (Cambridge Univ. Press, Cambridge, U.K.).
6. Furui, S. (1989) *Digital Speech Processing, Synthesis, and Recognition* (Dekker, New York).
7. Olive, J. P. (1977) *Proc. ICASSP-77*, 568–570.
8. Olive, J. P. (1990) in *Proceedings of the ESCA Workshop on Speech Synthesis* (Autrans, France).
9. Olive, J. P. & Liberman, M. Y. (1985) *J. Acoust. Soc. Am.* **78,** Suppl. 1, S6.
10. Hakoda, K., Nakajima, S., Hirokawa, T. & Mizuno, H. (1990) in *Proceedings of the International Conferences on Spoken Language Processing* (Kobe, Japan).
11. Nakajima, S. & Hamada, H. (1988) *Proc. ICASSP-88.*
12. Sagisaka, Y. (1988) *Proc. ICASSP-88.*
13. Sagisaka, Y., Kaikin, N., Iwahashi, N. & Mimura, K. (1992) in *Proceedings of the International Conference on Spoken Language Processing* (Banff, Canada).
14. Moulines, E., *et al.* (1990) *Proc. ICASSP-90.*
15. Charpentier, F. & Moulines, E. (1990) *Speech Commun.* **9,** 453–467.
16. Roucos, S. & Wilgus, A. (1985) *Proc. ICASSP-85*, 493–496.
17. Liljencrants, J. (1974) Swedish Patent Number 362975.
18. Sagisaka Y. (1990) *IEEE Commun. Mag.*
19. Ananthapadmanabha, T. V. (1984) *STL-QPSR 2-3/1984*, 1–24.
20. Hedelin, P. (1984) *Proc. IEEE*, 1.6.1–1.6.4.
21. Holmes, J. N. (1973) *IEEE Trans. Audio Electroacoust.* **21,** 298–305
22. Klatt, D. & Klatt, L. (1990) *J. Acoust. Soc. Am.* **87,** 820–857.
23. Rosenberg, A. E. (1971) *J. Acoust. Soc. Am,* **53,** 1632–1645.
24. Fant, G., Liljencrants, J. & Lin, Q. (1985) *Speech Transm. Lab. Q. Status Rep.* **4.**
25. Gobl, C. & Karlsson, I. (1991) in *Proceedings of the Vocal Fold Physiology Conference* eds. Gauffin & Hammerberg (Singular Publishing, San Diego).
26. Rothenberg, M. (1981) in *Vocal Fold Physiology*, eds. Stevens, K. M. & Hirano, M. (Univ. of Tokyo Press, Tokyo), pp. 303–323.
27. Rothenberg, M., Carlson, R., Granström, B. & Lindqvist-Gauffin, J. (1975) in *Speech Communication* (Almqvist & Wiksell, Stockholm), Vol. 2, pp. 235–243.
28. Bickley, C. & Stevens, K. (1986) *J. Phon.* **14,** 373–382.
29. Ananthapadmanabha, T. V., Nord, L. & Fant, G. (1982) in *Proceedings of the Representation of Speech in the Peripheral Auditory System* (Elsevier, Amsterdam), pp. 217–222.
30. Stevens, K. N. (1971) *J. Acoust. Soc. Am.* **50,** 1180–1192.
31. Shadle, C. H. (1985) Ph.D. thesis (Massachusetts Institute of Technology, Cambridge).
32. Badin, P. & Fant, G. (1989) in *Proceedings of the European Conference on Speech Technology.*
33. Holmes, J. (1983) *Speech Commun.* **2,** 251–273.
34. Javkin., H., *et al.* (1989) *Proc. ICASSP-89.*
35. Williams, D., Bickley, C. & Stevens, K. (1992) in *Proceedings of the International Conference on Spoken Lanugage Processing* (Banff, Canada), pp. 571–574.
36. Boves, L. (1991) *J. Phon.* **19.**
37. Carlson, R., Granström, B. & Karlsson, I. (1991) *Speech Commun.* **10,** 481–489.
38. Liljencrants, J. (1968) *IEEE Trans. Audio Electroacoust.* **16,** 137–140.
39. Carlson, R., Granström, B. & Hunnicutt, S. (1982) *Proc. ICASSP-82* **3,** 1604–1607.
40. Carlson, R., Granström, B. & Hunnicutt, S. (1991) in *Advances in Speech, Hearing, and Language Processing*, ed. Ainsworth, A. W. (JAI Press, London).
41. Stevens, K. & Bickley, C. (1991) *J. Phon.* **19.**
42. Rahim, M., Coodyear, C., Kleijn, B., Schroeter, J. & Sondi, M. (1993) *J. Acoust. Soc. Am.* **93,** 1109–1121.
43. Bailly, G., Laboissihre, R. & Schwartz, J. L. (1991) *J. Phon.* **19.**
44. Flanagan, J. L., Ishizaka, K. & Shipley, K. L. (1975) *Bell Syst. Tech. J.* **54,** 485–506.
45. Coker, C. H. (1976) *Proc. IEEE* **64,** 452–460.
46. Mermelstein, P. (1973) *J. Acoust. Soc. Am.* **53,** 1070–1082.
47. Lin, Q. & Fant, G. (1992) *Proc. ICASSP-92.*
48. Sondhi, M. M. & Schroeter, J. (1987) *IEEE Trans.* **ASSP-35.**
49. Liljencrants, J. (1969) *STL-QPSR 4/1969*, 43–50.
50. Slater, A. & Hawkins, S. (1992) in *Proceedings of the International Conference on Spoken Language Processing* (Banff, Canada).
51. Philips, M., Glass, J. & Zue, V. (1991) in *Proceedings of the European Conference on Speech Communication and Technology.*
52. Chomsky, N. & Halle, M. (1968) *The Sound Pattern of English* (Harper & Row, New York).
53. Pierrehumbert, J. B. (1987) *The Phonetics of English Intonation* (Indiana Univ. Press, Bloomington).
54. Hertz, S. R. (1991) *J. Phon.* **19.**
55. Hertz, S. R., Kadin, J. & Karplus, K. J. (1985) *Proc. IEEE* **73,** 11.
56. Van Leeuwen, H. C. & te Lindert, E. (1991) *Proc. ICASSP-91.*
57. Van Leeuwen, H. C. & te Lindert, E. (1993) *Comput. Speech Lang.* **7,** 149–168.
58. Holmes, W. J. & Pearce, D. J. B. (1990) in *Proceedings of the ESCA Workshop on Speech Synthesis* (Autrans, France).
59. Talkin, D. & Rowley, M. (1990) in *Proceedings of the ESCA Workshop on Speech Synthesis* (Autrans, France).
60. Carlson, R. & Granström, B. (1986) *Phonetica* **43,** 140–154.
61. Kaiki, N., Takeda, K. & Sagisaka, Y. (1990) in *Proceedings of the International Conference on Spoken Language Processing* (Kobe, Japan).
62. Riley, M. (1990) in *Proceedings of the ESCA Workshop on Speech Synthesis* (Autrans, France).
63. Van Santen, J. & Olive, J. P. (1990) *Comput. Speech Lang.* **4.**
64. Abe, M., Shikano, K. & Kuwabara, H. (1990) in *Proceedings of Speaker Characterisation in Speech Technology* (Edinburgh, U.K.).
65. Abe, M. (1991) *Proc. ICASSP-91.*
66. Valbret, H., Moulines, E. & Tubach, J. P. (1992) *Proc. ICASSP-92*, I-145–I-148.
67. Van Son, R. J. J. H. & Pols, L. (1989) in *Proceedings of the European Conference on Speech Communication and Technology.*
68. Karlsson, I. (1992) *Speech Commun.* **11,** 491–497.
69. Syrdal, A. K. (1992) *J. Acoust. Soc. Am.* **92.**
70. Cohen, H. M. (1989) Ph.D. thesis (Univ. of California, Berkeley).
71. Eskénazi, M. & Lacheret-Dujour, A. (1991) *Speech Commun.* **10,** 249–264.
72. Bladon, A., Carlson, R., Granström, B., Hunnicutt, S. & Karlsson, I. (1987) in *Proceedings of the European Conference on Speech Technology*, eds. Laver, J. & Jack, M. A. (Edinburgh, U.K.).
73. Klatt, D. K. (1987) *J. Acoust. Soc. Am.* **82,** 737–793.