Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Models of speech synthesis

Carlson, R.

journal: STL-QPSR
volume: 34
number: 1
year: 1993
pages: 001-014

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

MODELS OF SPEECH SYNTHESIS*

Rolf Carlson

Abstract

We will in this paper review some of the approaches used to generate synthetic speech and discuss some of the basic motivations for choosing one method over another. Primarily, we will discuss different methods of generating synthetic speech in a text-to-speech system. In the last part of the paper, general issues such as different voices, accents, and multiple languages are discussed.

INTRODUCTION

The term speech synthesis has been used for diverse technical approaches. Basically, any speech output from computers has been claimed to be speech synthesis, perhaps with the exception of play-back of recorded speech. We will in this paper review some of the approaches used to generate synthetic speech and discuss some of the basic motivations for choosing one method over another. In the publications by Fant (1960); Holmes, Mattingly, & Shearme (1964); Flanagan (1972); Klatt (1976); and Allen, Hunnicutt, & Klatt (1987), the foundations for speech synthesis based on acoustical or articulatory modelling can be found. The paper by Klatt (1987) gives an extensive review of the developments the speech synthesis technique.

Primarily, this paper will discuss different methods of generating synthetic speech in a text-to-speech system. However, there are other reasons for developing synthesis models. For example, the model might be used to understand how speech is in fact created, or how articulation can explain language structure. In the last part of the paper general issues such as different voices, accents and multiple languages are discussed.

Knowledge about natural speech

Synthesis development can be grouped into a few main categories: acoustic or articulatory modelling, and models based on coding of natural speech. In the last group, both predictive coding and concatenative synthesis using speech waveforms are included. The first two groups have had a long history of development while the last is a slightly younger 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 cost.

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 only use such knowledge to a

*This is a draft version of a paper presented at the "Colloquium on Human-Machine Communication by Voice", Irvine, California, February 8-9, 1993, organized by the National Academy of Sciences, USA.

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 the 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 modelling prosody by rule. The middle way, 'formant synthesis," is reaching towards the articulatory models 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 better quality, much due 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 ask for flexibility. Different voices, speaking styles, and accents also demand a flexible system in which explicit transformations can be modelled. Most of these variations are continuous rather than discrete. The importance of separating the modelling of speech knowledge from acoustic realization must be emphasized in this context.

In the overview by Furui (1989), synthesis techniques are divided according to 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 which 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. We will in the following describe synthesis models from two different perspectives: the sound generating part and the control part.

THE SOUND-GENERATING PART

The sound-generating part of the synthesis system can be divided into two subclasses depending upon in which dimensions 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 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 be dominating.

Simple waveform concatenation

The most radical solution to the synthesizer problem is simply to have a set of pre-recorded 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 to connect the pieces without distortion and to overcome 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. We will not further discuss these methods in this contribution.

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 (Olive, 1977; 1990; Olive & Liberman, 1985), NTT (Hakoda, Nakajima, Hirokawa, & Mizuno, 1990; Nakajima & Hamada, 1988), and ATR (Sagisaka, 1988; Sagisaka, Kaiki, Iwahashi, & Mimura, 1992) are based on the source-filter technique where the filter is represented in terms of 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, aspirative, and fricative.

Considerable success has been achieved by systems that base sound generation on concatenation of natural speech units (Mouline & al., 1990). 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 PSOLA (Carpentier & Moulines, 1989) methods are based on a pitch-synchronous overlap-add approach for concatenating 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 (Roucos & Wilgus, 1985), and systems for diver's speech restoration also did direct processing of the waveform (Liljencrants, 1974).

The basic function of a PSOLA type system is fairly simple. A database 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 decided pitch and amplitude.

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 last decade have serious problems especially when different voices are modelled. 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 variations in male voices.

Most source models work in the time domain with different controls to manipulate the pulse shape (Ananthapadmanabha, 1984; Hedelin, 1984; Holmes, 1973; Klatt & Klatt, 1990; Rosenberg, 1971). One version of such a voice source is the LF-model (Fant, Liljencrants, & Lin, 1985). It has a truncated exponential sinusoid followed by a variable cut-off -6 dB/octave low-pass filter modelling the effect of the return phase, i.e., the time from maximum excitation of the vocal tract to complete closure of the vocal folds. In addition to the amplitude and fundamental frequency control, two parameters influence the amplitudes of the two to three lowest harmonics, and one parameter, the high-frequency content of the spectrum. Another vocal source parameter is the diplophonia parameter with which creak, laryngalization, or diplophonia can be simulated (Klatt & Klatt, 1990). 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 modelling of noise excitation in order to synthesize a natural change between voiced and unvoiced segments. The work by Rothenberg (1981) can be guiding for future implementations. In some earlier work at KTH, we were able to use this model that included a noise source (Rothenberg, Carlson, Granström, & Lindqvist-Gauffin, 1975). High quality synthesis of extra-linguistic sounds, such as laughter, could be produced with this model in addition to reasonable voice-unvoiced transitions.

The acoustic interactions between the glottal source and the vocal tract also have to be considered (Bickley & Stevens, 1986). 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 (Ananthapadmanabha, Nord, & Fant, 1982). Listeners tend to be rather insensitive to bandwidth variation (Flanagan, 1972). When more complex models should be included, the output from the model has to change from a glottal flow model to a model of the glottal opening. The subglottal cavities can then be included in an articulatory model.

Noise sources have attracted much less research effort compared to the voiced source. However, some aspects have been discussed by Stevens (1971), Shadle (1985), and Badin & Fant (1989). Typically today simple white noise is filtered by resonances which are stationary in-between 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 mostly used for the synthesis of consonants. The classical configuration by Klatt (1980) has been used by many researchers. 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 the one by Holmes (1983), do not have this limitation.

The Klatt model is widely used in research both for general synthesis purposes and for perceptual experiments. A simplified version of this system is used in all

commercial products that stem from MIT: MITalk (Allen, Hunnicutt, & Klatt, 1987), DECtalk, and the system at Speech Technology Laboratory (Javkin & al., 1989). An improved version of the system has been commercialised as a research vehicle by Sensimetrics Corporation (Williams, Bickley, & Stevens, 1992). Similar configurations were used in the ESPRIT/Polyglot project (Boves, 1991).

A formant terminal analog, GLOVE (Carlson, Granström, & Karlsson, 1991), based on the OVE synthesizer (Liljencrants, 1968) has been developed at KTH and is used in current text-to-speech modelling (Carlson, Granström, & Hunnicutt, 1982; 1991). The main difference between the two traditions can be found in how the consonants are modelled. In the OVE case, a fricative is filtered by a zero-pole-pole configuration rather than 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, the speech quality is not good 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 alternative control methods to 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" (Stevens & Bickley, 1991; Williams, Bickley, & Stevens, 1992) explore an intermediate level that is more understandable from the developer's point of view compared to the detailed synthesizer specifications. The goal with 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 detailed understanding of the relation between acoustic and articulatory phonetics.

Articulatory models

Ultimately, an articulatory model will be the most interesting solution for the sound-generating part of text-to-speech systems, when the total flexibility of such a system is appreciated. Development is also going forward in this area, but the lack of reliable articulatory data and appropriate control strategies is still some of the bottlenecks. One possible solution that has attracted interest is to automatically train neural networks to control such a synthesizer. The works by Rahm, Kleijn, & Schroeter (1991) and Bailly, Laboissière, & Schwartz (1991) explore such methods.

Articulatory models, which now are under improvement, stem from the basic work carried out at laboratories such as Bell, MIT, and KTH. At each time interval, an approximation of the vocal tract is used either to calculate the corresponding transfer filter or to directly filter a source waveform. Different vocal tract models have been

used based on varying assumptions and simplifications. The models by Flanagan, Ishizaka, & Shipley (1975); Coker (1967); and Mermelstein (1973) have been studied by many researchers in the development of current articulatory synthesis.

The term "articulatory modelling" is often used in a rather loose way. The distinction between static and dynamic models has to be kept in mind when a synthesis approach is discussed. A complete model has to include several transformations from the control signal to the actual speech output. The relation between an articulatory gesture and a sequence of vocal tract shapes has to be modelled. Each shape should be transformed into some kind of tube model which has its acoustic characteristics. The vocal tract is then modelled 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 which 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 Bell Labs (Coker, 1967). 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 & Fant (1992) use a parallel synthesizer with parameters derived from an articulatory model.

In the development of articulatory modelling for text-to-speech, we can take advantage of parallel work on speech coding based on articulatory modelling (Sondhi & Schroeter, 1987). This work not only focuses 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.

In this section, we have not dealt with the important work carried out to model speech production in terms of volumes, masses, and airflow. 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

Modelling segmental coarticulation and other phonetic factors is 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 the library-based methods that replace rules by a collection of segment combinations of different unit lengths. Clearly both approaches have their advantages. If the data is coded in terms of targets and slopes, we need methods to calculate the parameter tracks. The efforts by Holmes, Mattingly, & Shearme (1964) and the filtered square wave approach by Liljencrants (1969) are some classical examples in this context.

To illustrate the problem, we have chosen some recent work by Slater & Hawkins (1992). The work was motivated by the need to improve the rule system in a text-to-speech system for British English. Data for the onset of the second formant in vowels after a velar stop and the midpoint in the vowel were analyzed and, as expected, a clear correlation between the 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 modelled 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. At some point, we will reach a maximum where the rules become too complex or the data collection 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 to 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 (1990), a new method was described to concatenate "acoustic inventory elements" of different sizes. The system developed at ATR is also based on non-uniform units (Sagisaka, Kaiki, Iwahashi, & Mimura, 1992).

Special methods to generate a unit inventory have been proposed by the research group at NTT in Japan (Hakoda & al., 1990; Nakajima & Hamada, 1988). The synthesis allophones are selected with the help of the context-oriented clustering method, COC. 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 databases. 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.) It is not possible to 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 a very few observations. In this case, a normalization of data might be a good solution before the actual unit is chosen. The system will be changed to a rule-based system. However, the rules can be automatically trained from data the same way as in speech recognition (Philips, Glass, & Zue, 1991).

Rules and notations

Development tools for text-to-speech systems have attracted considerable efforts. The publication of *The Sound Pattern of English* by Chomsky & Halle (1968) started 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 seventies. New software is still being developed according to this basic principle, but the implementations vary depending on the developer's taste. 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 organised? Can non-linear phonology be used (Pierrehumbert, 1987), as in the systems described by Hertz (1991) and Hertz, Kadin, & Karplus, 1985) and IPO (Leeuwen & Lindert, 1991)? Are the default values in the phoneme library primarily referred to by labels or by 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 labour-intensive optimization work. The notion "analysis by synthesis" has not been explored except by manual comparisons between hand-tuned spectral slices and reference spectra. The work by Holmes & Pearce (1990) is a good example of how to speed up this process. Automatic techniques, such as this, will probably also play an important role in making speaker-dependent adjustments. One advantage with 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 & Rowley (1990) 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 is currently 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 possibilities to test duration and intonation models (Carlson & Granström, 1986; Kaiki, Takeda, & Sagisaka, 1990; Riley, 1990; van Santen & Olive, 1990).

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 varying of voice or speaking style. On the contrary, the emphasis has been on a neutral way of reading, modelled after reading of non-related 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 developed. The range of applications asks for 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 on a global level, allowing 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 (Abe, Shikano, & Kuwabara, 1990). 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 (Abe, 1991).

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 modelled 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 has been proposed that combines the PSOLA technique for modifying prosody with a source-filter decomposition which enables spectral transformations (Valbret, Moulines, & Tubach, 1992).

Duration-dependent vowel reduction has been one topic of research in this context. It seems that vowel reduction as a function of speech tempo is a speaker-dependent factor (Gopal, Manzella, & Carey, 1991; van Son & Pols, 1989). 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 have to be considered in addition to basic physiological variation.

The difference between male/female speech has been studied by a few researchers (Klatt & Klatt, 1990; Karlsson, 1990; 1992a; 1992b). A few systems, such as Syrdal (1992), use a female voice as reference speaker. The male voice differs from the female in many respects, not only based on physiological aspects. To a great extent, speaking habits are formed by the social environment, dialect region, sex, education and also by a communicative situation which may require formal or informal speech. The speaker characteristic aspects have to be viewed as a complete description of the speaker in which all aspects are linked to each other into a unique framework (Cohen, 1989; Eskénazi, 1991; Eskénazi & Lacheret-Dujour, 1992).

The ultimate test of our descriptions is our ability to successfully synthesize not only different voices and accents but also different speaking styles (Bladon, Carlson, Granström, Hunnicutt, & Karlsson, 1987). Appropriate modelling of these factors will increase both 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 in the software itself. This is also the key to the multilingual effort at KTH. About one-third of the delivered systems by the INFOVOX company is multilingual. The synthesis developments pursued at companies such as ATR, CNET, DEC, and AT&T are all examples of multilingual projects. It is interesting to see that the research community in the world is rather

small. Several of the efforts are joint ventures such as the CNET British together with CSTR in Edinburgh, and the co-operation between Japanese (ATR) and U.S. partners. The Japanese company, Matsushita, even has a US branch (STL) for its English effort, originally based on MITalk.

Speech quality

The ultimate goal for all synthesis research with few exceptions is to produce as high speech quality as possible. The quality and the intelligibility of speech are usually very difficult tasks to measure. No single measure is able to pinpoint where the problems are. The research in Pisoni´s group (University of Indiana) has pushed the evaluation methods further, but we are still looking for the simple way to measure progress as a fast and reliable station on the synthesis development path. Thus, research tends to be heavily influenced by fast and subjective judgements by the developer in front of the computer screen. 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 review we have touched upon a number of different synthesis methods and research goals to improve current text-to-speech systems. It might be in place to remind the reader that nearly all methods are based on a historic development, where new knowledge has been added piece by piece to old knowledge rather than by a dramatic change of approach. Perhaps the most dramatic change is in the field of 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 a common facility even outside the research world, especially as 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.

ACKNOWLEDGEMENTS

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.

REFERENCES

Abe, M. (1991): "A segment-based approach to voice conversion," *Proc. ICASSP-91.*

Abe, M., Shikano, K., & Kuwabara, H. (1990): "Voice conversion for an interpreting telephone," *Proc. Speaker Characterisation in Speech Technology,* Edinburgh, UK.

Allen, J., Hunnicutt, M.S., & Klatt, D. (1987): *From Text to Speech. The MITalk System,* Cambridge University Press, Cambridge, England.

Ananthapadmanabha, T.V. (1984): "Acoustic analysis of voice source dynamics," *STL-QPSR* No. 2-3/1984, pp. 1-24.

Ananthapadmanabha, T.V., Nord, L., & Fant, G. (1982): "Perceptual discriminability of non-exponential/exponential damping of the first formant of vowel sounds," pp. 217-222 in *Proc. of the Representation of Speech in the Peripheral Auditory System,* Elsevier Biomedical Press, Amsterdam.

Badin, P. & Fant, G. (1989), "Fricative modeling: some essentials," *Proc. European Conference on Speech Technology,* Paris.

Bailly, G., Laboissi, R., & Schwartz, J.L. (1991): "Formant trajectories as audible gestures: an alternative for speech synthesis," *J. Phonetics* **19,** No 1.

Bickley, C. & Stevens, K. (1986): "Effects of the vocal tract constriction on the glottal source: Experimental and modelling studies," *J.Phonetics* **14,** pp 373-382.

Bladon, A., Carlson, R., Granström, B., Hunnicutt, S., & Karlsson, I. (1987): "A text-to-speech system for British English, and issues of dialect and style," pp. 55-58 in (J. Laver & M.A. Jack, eds.) *Proc. European Conference on Speech Technology, Vol. 1,* Edinburgh.

Boves, L. (1991): "Considerations in the design of a multi-lingual text-to-speech system," *J.Phonetics* **19,** No 1.

Carlson, R. & Granström, B. (1986): "A search for durational rules in a real-speech data base," *Phonetica* **43,** pp. 140-154.

Carlson, R. Granström, B., & Karlsson, I. (1991): "Experiments with voice modelling in speech synthesis," *Speech Communic.* **10,** pp 481-489.

Carlson, R., Granström, B., & Hunnicutt, S. (1982): "A multi-language text-to-speech module," pp. 1604-1607 in *Proc. ICASSP 82, Vol. 3,* Paris.

Carlson, R., Granström, B., & Hunnicutt, S. (1991): "Multilingual text-to-speech development and applications," in (A.W. Ainsworth, ed.), *Advances in Speech, Hearing and Language Processing,* JAI Press, London, UK.

Carpentier, F. & Moulines, E. (1990): "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communic.* **9**:5-6, pp. 453-467.

Chomsky, N. & Halle, M. (1968): *Sound Pattern of English,* Harper and Row, New York.

Cohen, H.M. (1989): *Phonological Structures for Speech Recognition,* Ph.D. Thesis, Computer Science Division, Univ. California, Berkeley, U.S.A.

Coker, C.H. (1967): "Synthesis by rule from articulatory parameters," *Proc. 1967 IEEE Boston Speech Conference.*

Eskénazi, M. (1992): "Changing speech styles: strategies in read speech and causal and careful spontaneous speech," *Proc. ICSLP92,* Banff, Canada.

Eskénazi, M. & Lacheret-Dujour, A. (1991): "Exploration of individual strategies in continuous speech," *Speech Communic.* **10,** pp. 249-264.

Fant, G. (1960): *Acoustic Theory of Speech Production,* Mouton, the Hague.

Fant, G., Liljencrants, J., & Lin, Q. (1985): "A four parameter model of glottal flow," *STL-QPSR,* No 4, pp. 1-13.

Flanagan, J.L. (1972): *Speech Analysis, Synthesis and Perception*, Springer Verlag, Berlin.

Flanagan, J.L., Ishizaka, K., & Shipley, K.L. (1975): "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell Syst.Techn.J.* **54**, pp. 485-506.

Furui, S. (1989): *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, New York.

Gopal, H.S., Manzella, J., & Carey,, C. (1991): "Factors influencing the spectral representation of front-back vowels in American English," *J.Acoust.Soc.Am.* **89**:4, 4SP10.

Hakoda, K.S., Nakajima, T., Hirokawa, & Mizuno, H. (1990): "A new Japanese text-to-speech synthesizer based on COC synthesis method," *Proc. ICSLP90*, Kobe, Japan.

Hedelin, P. (1984): "A glottal LPC-vocoder," *Proc. IEEE*, San Diego, pp. 1.6.1-1.6.4.

Hertz, S.R. (1991): "Streams, phones, and transitions: toward a new phonological and phonetic model of formant timing," *J.Phonetics* **19**, No 1.

Hertz, S.R., Kadin, J., & Karplus, K.J. (1985): "The Delta rule development system for speech synthesis from text," *Proc. IEEE*, **73**,

Holmberg, E.B., Hillman, R.E., & Perkell, J.S. (1988): "Glottal air flow and pressure measurements for loudness variation by male and female speakers", *J.Acoust.Soc.Am.* 84, pp. 511-529.

Holmes, J. (1983): "Formant synthesizers, cascade or parallel," *Speech Communic.* **2**, pp. 251-273.

Holmes, J.N. (1973): "Influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio and Electroacoustics*, **AU-21**, pp. 298-305.

Holmes, J., Mattingly, I.G., & Shearme, J.N. (1964): "Speech synthesis by rule," *Lang. & Speech* **7**, pp. 127-143.

Holmes, W.J. & Pearce, D.J.B. (1990): "Automatic derivation of segment models for synthesis-by-rule," *Proc. ESCA Workshop on Speech Synthesis*, Autrans, France.

Javkin., H. et al. (1989): "A multi-lingual text-to-speech system," *Proc. ICASSP-89.*

Kaiki, N. Takeda, K., & Sagisaka, Y. (1990): "Statistical analysis for segmental duration rules in Japanese speech synthesis," *Proc. ICSLP*, Kobe, Japan.

Karlsson, I. (1991): "Female voices in speech synthesis," *J.Phonetics* **19**, No 1.

Karlsson, I. (1992a): "Modelling speaking styles in female speech synthesis," *Speech Communic.* **11**, pp. 491-497.

Karlsson, I. (1992b): *Analysis and Synthesis of Different Voices with Emphasis on Female Speech*, DSc thesis, Dept. of Speech Communication & Music Acoustics, KTH.

Klatt, D.K. (1980): "Software for a cascade/parallel formant synthesizer," *J.Acoust.Soc.Am.* **67**, pp. 971-955.

Klatt, D.K. & Klatt, L. (1990): "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J.Acoust.Soc.Am.* **87**, pp. 820-857.

Klatt, D.K. (1976): "Structure of a phonological rule component for a synthesis-by-rule program," *IEEE Trans. ASSP-24.*

Klatt, D.K. (1987): "Review of text-to-speech conversion for English," *J.Acoust.Soc.Am.* **82**:3, pp. 737-793.

van Leeuwen, H.C. & te Lindert, E. (1991): "Speechmaker, text-to-speech synthesis based on a multilevel, synchronized data structure," *Proc. ICASSP-91.*

Liljencrants, J. (1968): "The OVE III Speech synthesizer," *IEEE Trans on Audio and Electroacoustics*, **Au-16**, No. 1, pp 137-140.

Liljencrants, J. (1969): "Speech synthesizer control by smoothed step functions," *STL-QPSR* No. 4, pp 43-50.

Liljencrants, J. (1974): "Metoder för proportionell frekvenstransponering av en signal," Swedish patent number 362975.

Lin, Q. & Fant, G. (1992): "An articulatory speech synthesizer based on a frequency domain simulation of the vocal tract," *Proc. ICASSP-92.*

Mouline., E. et al. (1990): "A real-time French text-to-speech system generating high quality synthetic speech," *Proc. ICASSP-90.*

Mermelstein, P. (1973): "Articulatory model for the study of speech production," *J.Acoust.Soc.Am.* **53**, pp. 1070-1082.

Nakajima, S. &. Hamada, H. (1988): "Automatic generation of synthesis units based on context oriented clustering," *Proc. ICASSP-88.*

Olive, J.P. (1977): "Rule synthesis of speech from dyadic units," *Proc. ICASSP-77*, pp 568-570.

Olive, J.P. (1990) "A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds," *Proc. ESCA Workshop on Speech Synthesis*, Autrans, France.

Olive, J.P. & Liberman, M.Y. (1985): "Text-to-speech -- an overview." *J.Acoust.Soc.Am.* **78**, Suppl. 1, S6.

Philips, M., Glass, J., & Zue, V. (1991): "Automatic learning of lexical representations for subword unit based speech recognition systems," *Proc. European Conf. on Speech Communication and Technology.*

Pierrehumbert, J.B. (1987): *The Phonetics of English Intonation*, Bloomington, IULC.

Rahm, M., Kleijn, B., & Schroeter J. (1991): "Acoustic to articulatory parameter mapping using an assembly of neural networks," *Proc. ICASSP-91.*

Riley, M. (1990): "Tree-based modeling for speech synthesis," *Proc. ESCA Workshop on Speech Synthesis*, Autrans, France.

Rosenberg, A.E. (1971): "Effect of glottal pulse shape on the quality of natural vowels," *J.Acoust.Soc.Am.* **53**, pp. 1632-1645.

Rothenberg, M. (1981): "Acoustic interactions between the glottal source and the vocal-tract," pp. 303-323 in (K.N. Stevens & M. Hirano, eds.) *Vocal Fold Physiology*, University of Tokyo Press.

Rothenberg, M., Carlson, R., Granström, B., & Lindqvist-Gauffin, J. (1975): "A three-parameter voice source for speech synthesis", pp. 235-243 in (G. Fant, ed.) *Speech Communication, Vol.* 2, Almqvist & Wiksell, Stockholm.

Roucos, S. & Wilgus, A. (1985): " High quality time-scale modification for speech," *ICASSP-8,5* pp. 493-496.

Sagisaka, Y. (1988): "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," *Proc. ICASSP-88.*

Sagisaka, Y., Kaiki, N., Iwahashi, N., & Mimura, K. (1992): "ATR v-TALK speech synthesis system," *Proc. ICSLP92*, Banff, Canada.

van Santen, J. & Olive, J.P. (1990): "The analysis of segmental effect on segmental duration," *Computer Speech and Lang.* No 4.

Shadle, C.H. (1985): *The Acoustics of Fricative Consonants*, Ph.D. Thesis, MIT, Cambridge, MA.

Slater, A. & Hawkins, S. (1992): "Effects of stress and vowel context on velar stops in British English," *Proc. ICSLP92*, Banff, Canada.

van Son, R.J.J.H. & Pols, L. (1989): "Comparing formant movements in fast and normal rate speech," *Proc. European Conf. on Speech Communication and Technology.*

Sondhi, M.M. & Schroeter, J. (1987): "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. ASSP* **35,** No. 7.

Stevens, K.N. & Bickley, C. (1991): "Constraints among parameters simplify control of Klatt formant synthesizer," *J.Phonetics* **19,** No 1.

Stevens, K.N. (1971): "Airflow and turbulence noise for fricative and stop consonants: Static considerations," *J.Acoust.Soc.Am.* **50:**4, pp. 1180-1192.

Syrdal, A.K. (1992): "Development of a female voice for a concatenative synthesis text-to-speech system," *J.Acoust.Soc.Am.* 92, 5pSP12.

Talkin, D. & Rowley, M. (1990): "Pitch-synchronous analysis and synthesis for TTS systems," *Proc. ESCA Workshop on Speech Synthesis,* Autrans, France.

Valbret, H., Moulines, E., & Tubach, J.P. (1992): "Voice transformation using PSOLA technique," pp I-145 - I-148 in *Proc. ICASSP-92,* San Francisco, U.S.A.

Williams, D., Bickley, C., & Stevens, K.N. (1992): "Inventory of phonetic contrasts generated by high-level control of a formant synthesizer," pp. 571-574 in *Proc. ICSLP92,* Banff, Canada.