SYNTHESIS: MODELING VARIABILITY AND CONSTRAINTS

Rolf Carlson

This is a modified version of a keynote paper given at the "European Conference on Speech Communication and Technology 91," September 24-26, Genova, Italy. The paper was prepared while the author was a guest reseacher at the Spoken Language Systems Group, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA.

Abstract

This paper discusses some important topics in current speech synthesis research. Modeling of speaker characteristics and emotions are used as a examples of new trends in the speech synthesis field. The relation to speech recognition research is emphasized. New methods such as automatic learning and the use of new analysis techniques are also discussed.

Introduction

The title of this paper might at first glance seem to be a mistake. Concepts such as variability and constraints are traditionally more related to speech recognition than speech synthesis. Variability is something that creates problems in speech recognition and many different methods have been developed to process speech in such a way that the variability to some extent can be handled. Similarly, constraint is an often-used term in speech recognition. Speech synthesis research has only recently started to deal with these two concepts.

It is a basic goal for speech research to understand when variability is allowed and when constraints are applied. It is also an important task for speech synthesis development to model the cause of variation: Is it a free variation or is it the result of a specific circumstance? Constraints have many different shapes. The freedom and the limitation of the vocal tract shape has been studied for many years. The constraints can in this case be expressed by size and mass limitations. Other useful constraints can be in the form of possible control parameter combinations in a formant type synthesizer. The move to explore higher level parameters, Stevens and Bickley (1990), is an example of how constraints are introduced into the control structure itself rather than by explicitly formulated rules. The description of prosody in terms of synchronized and unsynchronized turning points is another example of how constraints described in autosegmental phonology terminology has had a positive influence on the speech synthesis development, Boves (1990), Hertz (1990), Leeuwen and Lindert (1991) and Pierrehumbert (1987).

Modeling of variability is a new trend in speech synthesis. Speaker characteristics are beginning to play a more important role in the specification of a text-to-speech system. Similarly inter-speaker variation is put into focus as a way of improving the naturalness of the synthesis. Emphasis, focus and emotions are starting to be important concepts in the speech synthesis community. Better understanding of these areas will have an impact on several applications in speech technology in terms of improved quality. A systematic account of speech variability helps in creating speaker adaptable speech recognition systems and more flexible synthesis schemes.

In this paper we will give some examples from the speech synthesis area where this type of thinking has been productive. It is clear that speech synthesis research has changed during the last ten years. After rather slow progress we now have a very productive phase with many new directions. The special workshops in Autrans[1] and Edinburgh[2] gave many good examples of this new trend. The special issue of Journal of Phonetics[3] is another manifestation of the current interest in speech synthesis research.

Synthesizers and control parameters

We currently have a number of different classes of synthesizers in our systems. The long term goal of having articulatory synthesis is also attracting considerable research effort. It is generally agreed upon that the ultimate goal, to use an articulatory-based synthesizer, will be the best solution. However we are still far from including these models into our text-to-speech systems. One problem for this development is still the lack of articulatory data. Despite new analysis methods, the data collection is one of many bottlenecks. The efforts to use neural networks to go directly from the speech waveform to articulatory parameters are thus of considerable interest, Bailly and Laboissi (1990) and Rahm, Kleijn and Schroeter(1991). It is clear that we will see many papers of this nature in the future.

In the other end of the synthesizer continua we have the PSOLA type of method, Carpentier and Moulines (1989). The algorithms are based on a pitch-synchronous overlap-add approach for modifying the speech prosody and concatenating diphone waveforms. The frequency domain approach is used to modify the spectral characteristics of the signal while the time domain approach provides efficient solutions for real time implementation of synthesis systems. The PSOLA method has been very successfully applied in high quality text-to-speech synthesis systems, Mouline (1990). However, there are limitations in these approaches. Speaker transformation and unit selection can cause serious problems.

---

[1] The ESCA workshop on Speech Synthesis in Autrans (France) September 1990.

[2] The ESCA workshop on Speaker Characterization in Speech Technology in Edinburgh (UK) June 1990.

[3] Journal of Phonetics Special Issue: Speech Synthesis and Phonetics, Volume 19 No 1 January 1991.

Despite the difficulties in controlling formant-based synthesizers they are still used by many researchers. For example such synthesizers are used in the ESPRIT-project polyglot, Boves (1990), and in other multilingual efforts, Carlson, Granström and Hunnicutt (1991) and Javkin (1989). The current formant-based synthesizer systems are slowly incorporating some of the regularities found in true articulation. This is especially the case with the glottal source models Carlson et al. (1989), Fant, Liljencrants and Lin (1985), Klatt and Klatt (1990) and Stevens (1991).

Since the control of a formant synthesizer can be a very complex task, some efforts have been made to help the users. The introduction of "higher level parameters" should be mentioned in this context, Stevens and Bickley (1990). These parameters can be used at an intermediate level that is more understandable from the user's point of view compared to the detailed synthesizer specifications. Thus, the first goal is to find a framework to simplify the process and to incorporate within the synthesis process the constraints that are known to exist. 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 second goal for the introduction of higher level parameters is more basic in terms of understanding of the relation between the two levels of controls. This requires detailed understanding of the relation between acoustic and articulatory phonetics.

As a small test of this type of articulatory-based thinking, test stimuli along different feature dimensions have been synthesized, Bickley, Stevens and Carlson (1991). One intention was to illustrate the power of higher level parameters in speech synthesis. It was hypothesized that the higher level parameters explored more natural dimensions than the lower level controls. Thus, the phoneme identification of intermediate stimuli should be easier for the subjects in these experiments compared to similar experiments carried out before. It was concluded that the transition between two phoneme identities along the continuum was very abrupt, supporting this view.

Modeling of the Glottal Source

During the last decade we have seen a strong effort to study the glottal source. In addition to the Vocal Fold Symposiums, special sections dealing with this subject have been arranged at ASA meetings and also at the Spoken Language Processing meeting in Kobe, Japan. It has been felt that understanding of the glottal source is one of the most important goals in speech synthesis work. The unnatural quality of the synthetic speech has to a large extent been blamed on a simplified glottal source. The work to synthesize female voices has supported this view Carlson, Granström and Karlsson(1990) and Karlsson, (1990, 1991). Several glottal models have been proposed Fant, Liljencrants and Lin (1985) and Klatt, and Klatt (1990). The improvement of speech quality by including an elaborate glottal model has in some cases been very impressive. It is clear that the simple models used up to now have been a major obstacle and that source modeling will be a critical aspect in the next generation of synthesis systems.

However, despite the current emphasis on source modeling, it should be noted that other aspects have equal importance. For example, improved models of the higher vocal tract resonances or the fricative spectrum in formant-type synthesis have a very strong impact on the speech quality.

It should be emphasized that quality improvement can be made in many different ways. Correct intonation can in some cases lead to high acceptability of the synthetic speech despite segmental problems. On the other hand, inferior quality in synthesis with many unnatural discontinuities and missing cues can not be "hidden" by a good prosodic model.

Speaker characteristics

Synthesis research has, to some extent, changed direction during recent years. The emphasis on CV syllables has been reduced and general aspects such as speaker characteristics, prosodic models and linguistic analysis have been given higher priority. The reasons for this change are many. 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 and Kuwabara (1990). The proposed voice conversion technique consists of two steps: mapping codebook 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 specified through training without any specific underlying 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. Thus, it is a challenge for the future to find the best way to classify a speaker. The needs in both speech synthesis and speech recognition are very similar in this respect.

Several studies have recently been published concerning how a speaker adjusts to the listener and to the environment. A speaker is expected to vary the speech along a continuum of hypo- and hyperspeech, Lindblom (1990). It is argued that one important research task is to study the sufficient discriminability needed for communication rather than the notion of phonetic invariance. Duration-dependent vowel reduction has been one topic of research in this context. However, it seems that vowel reduction as a function of speech tempo is a speaker-dependent factor, Gopal, Manzella and Carey (1991) and van Son and Pols (1989).

Duration and intonation structures and pause insertion strategies reflecting variability in the dynamic speaking style are other important speaker dependent factors, Fant, Kruckenberg and Nord (1990), Sagisaka and Kaiki (1991) and Sorin, Larreur and Llorca (1987). Parameters like consonant-vowel ratio and source dynamics are typical parameters that have to be considered in addition to basic physiological variation. The ultimate test of our descriptions is our ability to successfully synthesize not only different voices but also different styles, Bladon et. al. (1987). Appropriate modeling of these factors will increase both naturalness and intelligibility of a synthetic speech.

Synthesis of emotions

In acoustic-phonetic research most studies deal with function and realization of linguistic elements. With a few exceptions, (e. g. , Scherer, 1989 and Williams and Stevens, 1972) the acoustics of emotions have not been extensively studied. Most studies have dealt with the task of identifying extralinguistic dimensions qualitatively. Sometimes these studies have also included efforts to quantify these dimensions, by using scaling methods for example. Spontaneous speech has been used as well as read speech with simulated emotional expressions in these experiments.

An interesting alternative is to ask subjects to adjust test stimuli to some internal reference, such as joy, anger etc. This is typically done by using synthetic speech. The speech should not be of too poor quality if emotions should be conveyed. Recent experiments using DECtalk have been reported by Cahn (1990). The special "affect-editor" was developed to control the synthesizer. Its success in generating recognizable affects was confirmed in an experiment in which the affect intended was perceived as such for the majority of the presentations.

Similar efforts have been reported by Murray et.al (1988,1991). The system HAMLET was developed for use in speech prostheses for the nonvocal, and was designed for incorporation into communication systems. The system uses DECtalk as an output device just as in the experiments by Cahn. Any of six emotions can be selected from a menu. The corresponding rules then operate on the phonemes and the voice quality settings, which are sent to the text-to-speech system.

The amount of interaction between the emotive speech and the linguistic content of a sentence is difficult to ascertain, but has to be taken into account. The voice does not always give away the speaker's attitude. It is often observed that misinterpretation of emotions occurs if the listener is perceiving the speech signal without reference to visual cues. Depending on contextual references, it is thus easy to confuse anger with joy, fright with sorrow, etc.

Systematic variation in speech synthesis has been used as a tool to explore possible style and speaker dimensions, Granström and Nord (1991). Preliminary listening experiments were carried out with the aim of describing different synthesis samples according to different attitudinal and emotional dimensions. It was shown that such a method can be extremely valuable in exploring extralinguistic types of variations.

Automatic learning

We have recently noticed very interesting efforts to collect segmental data for synthesis with the help of automatic procedures. Formant-type synthesis has traditionally been based on very labor-intensive optimization work. The notion "analysis by synthesis" has not been explored except by manual comparisons between hand-tuned spectral slices and a reference spectra. The work by Holmes and Pearce (1990) is a good example of how to speed up this process. With the help of a synthesis model, the spectra is automatically matched against analyzed speech. The matching is done on a linear power scale to emphasize the importance of spectral peaks. The ambition is to make a broad collection of such analyzed segments and to use a clustering technique to reduce the size of the collection. 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-synchronus analysis will be of major importance in this context. Experiments such as the one presented by Talkin and 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 an analyzed speech corpora, we are in a good position to look for coarticulation rules and context-dependent variations. Detailed analysis work such as the study of vowels by Huang (1990) can be complemented with automatic procedures. Rule extraction algorithms such as the one described by Bosch (1990) can be applied to these types of data.

The collection of huge speech corpora has also facilitated a new possibility to test duration and intonation models on a grand scale, Carlson (1991), Kaiki, Takeda and Sagisaka (1990), Riley (1990) and van Santen and Olive (1990). Some of the old "knowledge" has been revised in this context. The new type of methods can easily create large amounts of analysis results. It will be the task for the speech synthesis researcher to summarize these in understandable models that can be used in the next generation of synthesizers, Campbell and Isard (1990), Collier (1990) and Kohler (1990).

Unit size

A special method to generate an allophone inventory has been proposed by the research group at NTT in Japan, Hakoda et. al. (1990) and Nakajima and 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 most affect the phoneme realization. The system developed using these synthesis units was regarded to have superior speech quality compared to an earlier synthesis system based on diphones.

The context-oriented clustering approach is a good illustration of a new trend in speech synthesis. Our studies are concerned with much wider 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 and Zue (1991).

Systems using elements of different lengths depending on the target phoneme and its function are 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 (1988). These units have been statistically chosen to cover a specific domain.

Speech synthesis in speech recognition

One purpose of this paper has been to show how "variability" and "constraints" are relevant aspects to consider in speech synthesis just as in speech recognition. The borders between synthesis and recognition are slowly disappearing. Speaker-independent recognition and speaker-dependent synthesis have many similar problems to handle. It is, for example, possible to limit the number of input dimensions in a recognition model by using a speaker-dependent synthesis model. Constraints such as vocal tract length, source variation or segment durations can be applied in such a model.

The text-to-speech project at the Royal Institute of Technology has recently focused on modeling different speaker characteristics and speaking styles. Methods in the speech recognition project have been influenced by this work. This has made our speech recognition efforts slightly different from the general trend. The research program, "Nebula" Blomberg, et al. (1988), includes prediction models based on speech synthesis. A description of speech on a level closer to articulation, rather than the acoustic base that is used in present-day speech recognition will make generalization of different speakers easier.

Intra-speaker voice source variation can cause severe spectral distortion and contributes to recognition errors in current speaker-independent as well as in speaker-dependent recognition systems. The prosodic information carried by the voice source is important and should not be discarded. This information is lost in many of the current techniques using parameter estimation methods intended to be insensitive to voice source behavior. Since the voice characteristics are changing during an utterance, the speaker adaptation should be part of the recognition process itself. Modeling the source of variation rather than the effect on the speech acoustics potentially makes adaptation more efficient. The production component in the form of a speech synthesis system will ideally make the collection of training data unnecessary. During the last year, special projects studying speaker-independent recognition based on stored phoneme prototypes have been undertaken Blomberg (1989, 1990). In these experiments, the references are synthesized

during the recognition process itself. The synthetic references can be modified to match the voice of the current speaker. The experiments have shown promising results.

Speech synthesis research presented in public

Speech synthesis research has had a long tradition as a subject for papers and reports. Most work has been presented at meetings such as the ASA or ICASSP. However we can see a tendency towards reduced focus on synthesis papers in these meetings. The European Conference on Speech Technology meetings have shown more variety in synthesis papers. The same is true for the first ICSLP 90 meeting. The successful meeting in Autrans was devoted totally to speech synthesis and showed the breadth of this exciting research area.

One possible reason for lack of publication of detailed work is the reluctance to discuss a particular subject such as "How I improved the synthesis of /s/", "How much f0 movement do I need to turn a statement into a question" or "How I synthesized different degrees of emphasis with a global parameter change". The list just like the research area is endless. Somehow this type of paper is not as acceptable as it used to be. Because of this, some work also stops just before it attains scientific value. At best the result gets hidden in a laboratory system or in some cases in a company product. The message is that the tuning of systems or testing of new solutions must be treated as good research, not as an uninteresting optimization. If we can change this attitude we will get an exciting selection of presentations that will push speech research quality and synthesis quality forward.

Conclusion

In this review we have focused on a few exciting research areas which are just beginning to demonstrate their potential. "Speaker variation" and "speaker variability" are key-words in future synthesis research. However we need to go further than just understanding the problem. We need to build new models that can capture the basic parameters along these new dimensions. We need to set up synthesis systems that can incorporate this variation without having to rewrite our software from the beginning. We need to create new synthesizers that can model all the needed control parameters. And we need to structure these parameters in such a way that the users can handle them without too much effort.

To reach these goals we need to make use of methods outside the current speech synthesis domain. Automatic procedures have to be developed to adjust our models to specific speakers and to gather huge amounts of speaker-dependent data. However, we should not get lost in this data. It needs to be structured in new models. Thus the long history of using speech synthesis to evaluate gained knowledge should be continued and expanded, Fant (1990) and Stevens (1991). Speech synthesis will also be an important research tool in the future.

Acknowledgements

I would like to thank Björn Granström for valuable discussions during the preparation of this paper. I would also like to thank Victor Zue for the opportunity of working this past year with the Spoken Language Systems Group, Laboratory for Computer Science, MIT, USA during which time this paper was prepared.

References

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

M. Abe, K. Shikano and H. Kuwabara (1990), "Voice conversion for an interpreting telephone," Proc. Speaker characterization in speech technology, Edinburgh, UK.

G. Bailly and R. Laboissi (1990), "Formant trajectories as audible gestures: an alternative for speech synthesis," J. Phonetics 19, No 1.

C. Bickley, K. Stevens, and R. Carlson (1991), "Synthesis of manner and voicing continua based on speech production models," J. Acoust. Soc. Am. , Vol 89. No4 3SP7.

A. Bladon, R. Carlson, B. Granström, S. Hunnicutt and I. Karlsson (1987), "Text-to-speech system for British English, and issues of dialect and style," Proc. European Conference on Speech Communication and Technology, Edinburgh, UK.

M. Blomberg (1989), "Synthetic phoneme prototypes in a connected-word speech recognition system," Proc. ICASSP-89

M. Blomberg (1990), "Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references," Proc. ESCA Workshop on Speaker Characterization in Speech Technology, Edinburgh, UK.

M. Blomberg, R. Carlson, K. Elenius, B. Granström and S. Hunnicutt (1988), "Word recognition using synthesized reference templates. ," Proc. Second Symposium on Advanced Man-Machine Interface Through Spoken Language, Hawaii, USA, also in STL-QPSR 2-3/1988.

L. Bosch. (1990), "Rule extraction for allophone synthesis," Proc. ESCA Workshop on Speech Synthesis, Autrans, France.

L. Boves (1990), "Considerations in the design of a multi-lingual text-to-speech system," J. Phonetics 19, No 1.

J. E. Cahn (1990), "The generation of affect in synthesized speech," Journal of the American Voice I/O Society, vol. 8.

N. Campbell and S. D. Isard (1990), "Segment durations in a syllable frame," J. Phonetics 19, No 1.

R. Carlson (1991), "Duration models in use," Proc. XIIth ICPhS, Aix en Provence, France.

R. Carlson, G. Fant, C. Gobl, B. Granström, I. Karlsson, and Q. Lin (1989), "Voice source rules for text-to-speech synthesis," Proc. ICASSP-89.

R. Carlson, B. Granström and I. Karlsson(1990), "Experiments with voice modeling in speech synthesis," Proc. ESCA workshop on Speaker Characterization in Speech Technology, Edinburgh, UK.

R. Carlson, B. Granström and S. Hunnicutt (1991), "Multilingual text-to-speech development and applications," A. W. Ainsworth (Ed.), Advances in speech, hearing and language processing, JAI Press, London, UK.

F. Carpentier and E. Moulines (1989), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Proc. European Conference on Speech Communication and Technology 89.

R. Collier (1990), "Multi-language intonation synthesis," J. Phonetics 19, No 1.

G. Fant (1990), "Basic research as a support for speech synthesis," J. Phonetics 19, No 1.

G. Fant, A. Kruckenberg and L. Nord (1990), "Prosodic and segmental speaker variations," Proc. ESCA Workshop on Speaker Characterization in Speech Technology, Edinburgh, UK.

G. Fant, J. Liljencrants and Q. Lin (1985), "A four parameter model of glottal flow," Speech Transmission Laboratory Quarterly and Status Report STL-QPSR No 4.

H. S. Gopal, J. Manzella and C. Carey (1991), "Factors influencing the spectral representation of front-back vowels in American English," J. Acoust. Soc. Am. 4SP10, Vol 89, No4.

B. Granström and L. Nord (1991), "Ways of exploring speaker characteristics and speaking styles," Proc. XIIth ICPhS, Aix en Provence, France.

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

S. Hertz (1990), "Streams, phones, and transitions: toward a new phonological and phonetic model of formant timing," J. Phonetics 19, No 1.

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

C. Huang (1990), "Effects on context, stress, and, speech style on American vowels," Proc. ICSLP90, Kobe, Japan.

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

N. Kaiki, K. Takeda and Y. Sagisaka (1990), "Statistical analysis for segmental duration rules in Japanese speech synthesis," Proc. Int. Conf. on Spoken Language Processing, Kobe, Japan.

I. Karlsson (1990), "Female voices in speech synthesis," J. Phonetics 19, No 1.

I. Karlsson (1991), "Dynamic voice quality variations in female speech," Proc. XIIth, International Congress of Phonetic Sciences, Aix-en-Provance, France.

D. Klatt and L. Klatt (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. , vol. 87(2)

K. Kohler (1990), "Prosody in speech synthesis: The interplay between basic research and TTS application," J. Phonetics 19, No 1.

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

B. Lindblom (1990), "Explaining phonetic variation: A sketch of the H&H theory," in Speech production modeling Hardcastle and Marchal (eds.) Kluwer Academic Publishers, Netherlands.

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

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

I.R. Murray, J.L. Arnott and A.F. Newell (1988), " HAMLET - simulating emotion in synthetic speech," Proc. Speech '88, 7th FASE Symposium.

I.R. Murray, J.L. Arnott, N. Alm and A.F. Newell (1991), " A communiaction system for the disabled with emotional synthetic speech produced by rule," Proc. European Conference on Speech Communication and Technology 91.

J. P. Olive (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.

M. Philips, J. Glass and V. Zue (1991), "Automatic learning of lexical representations for sub-word unit based speech recognition systems," Proc. European Conference on Speech Communication and Technology 91.

J. B. Pierrehumbert (1987), "The phonetics of English intonation,"Bloomington: IULC.

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

M. Riley (1990), "Tree-based modeling for speech synthesis," Autrans, France.

Y. Sagisaka and N. Kaiki (1991), "Prosody control for spontaneous speech synthesis," XIIth, International Congress of Phonetic Sciences, Aix-en-Provance, France.

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

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

K. Scherer (1989), "Vocal correlates of emotion," in (eds. Wagner, H. and Manstead, T.), Handbook of Psychophysiology: Emotion and Social Behavior, Chichester: Wiley, UK.

R. J. J. H. van Son. and L. Pols (1989), "Comparing formant movements in fast and normal rate speech," Proc. European Conference on Speech Communication and Technology 89.

C. Sorin, D. Larreur and R. Llorca (1987), "A rythm-based prosodic parser for text-to-speech systems in French," Proc. XIth ICphS, Tallin,Estonia.

K. Stevens and C. Bickley (1990), "Constraints among parameter simplify control of Klatt formant synthesizer," J. Phonetics 19, No 1.

K. Stevens (1991), "The contribution of speech synthesis to phonetics: Dennis Klatt's legacy," Proc. XIIth, International Congress of Phonetic Sciences, Aix-en-Provance, France.

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

C. E. Williams and K. Stevens (1972), "Emotions and speech: some acoustical correlates," J. Acoust. Soc. Am. Vol. 52.