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

Perception and synthesis of speech

Carlson, R. and Granström, B.

journal: STL-QPSR
volume: 18
number: 1
year: 1977
pages: 001-016

KTH Computer Science and Communication

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

I. SPEECH PERCEPTION AND SPEECH SYNTHESIS

A. PERCEPTION AND SYNTHESIS OF SPEECH

R. Carlson and B. Granström*

Abstract

Descriptive models for speech production have been critically analyzed from a perceptual point of view. In these studies we have taken advantage of the speech synthesis technique, which is the second topic of this thesis.

Vowels are normally described by the first two or three formants. In perceptual experiments we have found a better technique to describe the vowel space in two dimensions than to just leave the third and higher formants out. The influence of f_0 and the discriminability of spectral slope have also been studied.

Consonantal cues realized as rapid changes and short durations in the acoustic waveform have been found to carry a large part of the segmental information in the speech communication process.

The perception of prosodic aspects like segment duration seems to be highly dependent on segment type, position, and stress. This has consequences for the understanding of descriptive models for prosody.

In the second part of our thesis we present a powerful tool, a programming language, for formalizations of dynamic models like speech synthesis systems. This tool has been used to construct a text-to-speech system that transforms a written text to synthesized speech. The system has been evaluated as a reading machine for the blind with promising results.

The importance of using speech synthesis techniques in perceptual research and vice versa has hopefully been demonstrated.

---

* This is a summary of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Science to be officially defended on May 17, 1977 at KTH.

Introduction

This summary relies on work performed during the last eight years and is intended only as an introduction to the eleven publications which form the thesis.

The combination of speech synthesis and perception research has been a tradition since the fifties, especially at Haskins Laboratories, where a considerable research effort has been invested in clarifying the relations between parameters of the speech wave and its perceived identity. Some of these early results could be disputed in the light of our experiments, but what should not be neglected is the profound influence this line of research has had on our thinking. We have also been inspired by the work of the research group of Chistovich and Kozhevnikov with their efforts in building models on diverse evidence, especially for the auditory-perceptual system.

Besides the basic theme of perceptual studies we have spent a considerable effort in developing speech synthesis techniques. In this part of our research we rely on influences and interactions with several groups involved in speech synthesis, such as MIT, Haskins Laboratories, and Bell Laboratories. The most important influence in this field has, however, come from the unique and pioneering research in acoustic modeling and speech synthesis in the Dept. of Speech Communication, KTH in Stockholm.

In this laboratory there is a long tradition of combining basic phonetic research with mathematical modeling and technical development. This has resulted in not only a better understanding of the speech communication process per se but has also led to useful applications in fields like communication technology and aids for people with different kinds of communication handicaps. This tradition we have tried to continue to the best of our abilities.

In an attempt to put our effort in its appropriate place we will mention some instances where we believe that we have made original contributions.

For vowels we have constructed an auditory two-dimensional mapping and also proposed a functional model relating this perceptual representation to the physical vowel stimulus. The information content in consonants has been shown to have a different distribution than assumed in the literature. Brief stop explosions (10-20 msec) could in certain contexts overrule the traditional articulation place information of formant transitions. Manner cues normally believed to reside primarily in the more steady state central position in consonants are for laterals and

nasals found in the first glottal pulses of the adjacent vowel. The importance of durational variation in continuous speech can, according to our results, not be predicted from psychoacoustic data. Vowel duration as such, rather than intervals between vowel onsets, appears most important, contrary to current theories.

In speech synthesis a starting point has been on the results from synthesis based on analysis of natural speech and the phoneme based square wave synthesis technique developed in our laboratory. Our contribution in this field has primarily been in developing a programming language for rule synthesis and implementing a total and general system for transformation of ordinary Swedish text to control parameters for the synthesizer.

Rather than summarizing every paper in turn we will deal with groups of papers under what we consider relevant headings. Our goal has been to concentrate on ideas and major or typical results referring readers to the original papers for details. For the same reason we have left out literature references that would otherwise substantiate our positions.

Behind the title "Perception and synthesis of speech" there is a wish to demonstrate the fruitful interaction between these two areas of speech research.

Stimulus generating techniques for speech perception research

The work presented in this summary relates to different disciplines like linguistics, experimental psychology, acoustics, and engineering. The tendency to break down the boundaries between these areas has proved to be very productive in speech research. Since the following sections will contain discussions and reviews dealing with speech rather than with technical problems, some general remarks should be made about the technical methods used. The development of these research tools and procedures has been a non-negligible part of our work.

In perceptual experiments there are a variety of methods for stimulus production that in principle are possible. Each has advantages and drawbacks. The work on vowels needed a flexible method in which both the number of formants and their frequencies and levels could easily be changed. A simulated parallel synthesis configuration was a good solution giving the subjects the possibility of interactively varying one or more parameters by turning knobs. A drawback to this method, however, is that unwanted antiresonances that are hard to control are introduced into the spectrum. This is especially unacceptable in discrimina-

tion tests, where a detailed specification is necessary. We have for this purpose preferred to describe the vowel as a sum of partials. These can be produced by well-controlled simulated frequency generators and added before play-back. This method, however, gives a constant pitch and because of that a reduced naturalness and applicability.

In speech perception research there is sometimes a need to use human speech, but in a well-defined way. We may want to prolong, amplify, or replace a certain segment or insert synthesized samples in a natural context. A computer system has been programmed in which digitized signals can be manipulated according to the wish of the researcher. This tool has been used among other things in the study of rapid consonant cues and is referred to as a speech editing technique.

Special effort has been made to construct a software system that by rule can transform characters, features, and variables. Dynamic processes are easily described within this system without risking the researcher becoming lost in computer programming. The system and its application to speech synthesis is described in more detail below.

Perception of vowels
(mainly referring to papers (1)(4)(6)(7)(10))

Understanding how vowels are perceived is basic to any model of speech perception. Vowels are the smallest units in speech production in that they function as syllables with or without associated consonants and can hence constitute a whole message. The production of vowels is normally modeled as a simple source-filter structure. The sound source at the glottis can be regarded as essentially independent of the filtering effect of the vocal tract. This filter is for most purposes conveniently described by a number of resonances or formants.

Our primary interest has been in how variations in the filter are perceived, since these variations are the main determinants of vowel quality. We have also studied how source characteristics, such as spectral slope and glottal frequency, are perceived and how these parameters interact with the perception of the "filter".

Vowels are often described by a few (two or three) dimensions such as the front-back and high-low position of the tongue and rounding-spreading of the lips. In multidimensional scaling techniques the most significant dimensions have been interpreted

in terms of acoustic parameters. High correlation to the first and second formant has been found but the third dimension has been given no unique interpretation. This might be due to the fact that no obvious physical parameter is used in the subjects' classification. We have tried to approach the problem with a different technique by attempting a reduction to two independent variables in the vowel stimulus itself.

The nature of this reduction was studied in matching experiments. Four-formant vowels were synthesized with a software parallel synthesizer but with formant levels calculated according to the serial model. These reference stimuli were matched to stimuli consisting of two formants only, whose levels, bandwidths, and resonance frequencies could be controlled by the subjects. Both types of stimuli were given the same natural sounding envelopes and fundamental frequency inflexions. The result can be summarized in three points. Levels and bandwidths were of little or no importance to phonetic quality over a wide range of values; the first resonance of the two-formant stimulus coincided well with F1 of the reference stimulus; the second resonance of the two-formant vowel ($F2'$) coincided with F2 for back vowels but had considerably higher values for front vowels, unrounded in particular.

The nature of these matches seemed to rely partially on the vowel identity given to the reference. Various methods were investigated to see to what extent the matching results could be predicted from the physical characteristics of the reference. The most successful method for parameter reduction, giving a good agreement to the matching results, was to make a histogram of the zero-cross frequency of the output in different filters of a filter bank. This filter bank was simulated to conform to Flanagan's formalization of von Békésy's measurements of basilar membrane motion. The interpretation of the parameter reduction as taking place in the cochlea, must, however, be rejected, as four-formant vowels - split between ears in different ways - were found to evoke the same vowel identification as when all formants were presented to one ear.

In addition to this psychoacoustically oriented method an empirical formula has been constructed to predict $F2'$ from a given set of formant frequencies. To test the validity of the matching results a two-formant identification test was performed. The resulting $F1\text{-}F2'$ plane has the property that the Swedish vowels, especially when plotted on perceptually relevant mel scales, are approximately equally spaced. This result may clarify some of the perceptual restrictions on vowel systems.

Vowel identification can also be affected by different conditioning factors including the phonetic context, dialect, speaker, idiosyncrasies, etc. One such factor is the fundamental frequency that is, on the average, inversely related to vocal tract size.

Considering the width of a critical frequency band it is clear that individual harmonics could be resolved by the auditory system for low-frequency formants. It has been argued that the most audible or phonetically significant harmonic will be the basis of vowel identity decision. In an experiment with systematically varied F0 and F1 we concluded that the phoneme boundary between [i] and [e] changes monotonically with fundamental frequency (conditioning effect) and that the estimate of the lowest resonance (F1) was based, if not on envelope detection, at least on a weighted mean of two loudest harmonics in the F1 region (parameter extraction).

In a study concerning the ambiguity between [ø] and [e] we also noticed the interaction between conditioning factors and spectral analysis. By considering the entire upper part of the spectrum the ambiguity could be resolved. Most of our studies have been concerned with perception of variations in the vocal tract. Besides the fundamental frequency, some source variations affecting the spectrum have been studied. The possible effect of these variations is not primarily in the phonetic identity but rather to signal speaker identity and mood, as well as stress and emphasis, and will hence be dealt with under the heading perception of prosody.

Perception of consonants
(mainly referring to paper (2)(9))

The perception of consonants has been claimed to be quite dif ferent than the perception of vowels. This can partly be due to the fact that vowels are studied both in isolation as well as in more complex context. Consonants normally don't form syllables without a vowel as kernel and thus experiments with isolated consonants are unnatural and uncommon.

We have in our paper (2) taken a special interest in the perception of consonantal acoustic cues realized in the speech wave as rapid changes, e. g. the consonant-vowel boundaries, and events of short duration, such as stop explosions. This is of particular interest since perceptual research concerning consonants mostly has been focused on place cues signaled by formant transitions or manner cues realized during comparatively stationary consonant segments. Exceptions are, however, the experiments, carried

out at Haskins Laboratories, in which stop explosions have been studied with highly simplified synthetic speech. In these studies the optimal specification, especially for velars, was shown to be strongly dependent on vowel context with considerable overlap between [p] and [k] explosions. This result could be expected on the basis of analysis data since velars in particular are strongly coarticulated.

We have, however, studied the distinctive importance of explosions in natural voiceless stops using a speech-editing technique. The explosion of a [p]V syllable was replaced by a [k] or [t] explosion of 10-20 msec duration. The result showed that the acoustic cues in the stop explosion could govern the consonant identification despite contradicting formant transition cues.

This leads us to the conclusion that it is of great importance both for the understanding of the perception as well as the synthesis of speech that we get a far more detailed description of the release. By selectively filtering the explosions we found, for example, that the [t] explosion must contain energy above 4 kHz to be perceptually efficient, while energy below 4 kHz is of little importance. The [k] explosion in front of [a] also contains important energy above 2 kHz and around .5 kHz apart from the characteristic and well-known resonance around 1.7 kHz.

The second subject dealt with in paper (2) concerns the consonant-vowel boundary. The relatively close similarity between [n] and [l] is obvious in the analysis of speech. In a speech-editing experiment the oral occlusion plus 0-3 glottal pulses from the following vowel were interchanged between the two categories. We found that the first glottal pulse contained enough information to change the result from 50 %-100 % [l] responses. On the other hand, we needed several pulses to increase the [n] responses. We conclude that the abrupt change of intensity and spectral shape at the phoneme boundary is of great importance for the [l], while the nasalization of the vowel must be realized during several glottal pulses for the [n]. In other words: the listener has to decide whether the speaker makes a fast change of the mouth cavity in the [l] or a slow opening or closing of the nasal tract for the [n].

Thus, events of short duration with fast changes, corresponding to transitional states of the speech apparatus, carry relatively greater importance for consonant perception than the slowly varying or steady state portions.

Segmental duration per se is, however, of considerable perceptual interest as will be demonstrated in the next section.

Perception of prosody
(mainly referring to papers (8)(10))

The prosodic structure of a sentence has been studied on different levels and the results have been formulated in models that predict prosodic factors such as duration, intonation, and vocal effort. One of these models is used in our text-to-speech system, paper (3), taking into account the position of a segment within the syllable, word, and sentence, as well as the overall stress pattern. The data behind these models, however, show astonishingly low variance in certain positions but marked variation in others. The question remains as to whether these variations can be perceived or not and in which positions a correct duration or spectral slope is critical.

Spectral slope has been claimed to mediate linguistic information like stress, emphasis, and, possibly, position within a phrase. We have in our paper (10) asked: Does this description have a close relationship to perceptual factors? The answer has consequences for the use of the complex glottal source, described in paper (4), in speech synthesis.

In paper (8) we have dealt with three questions concerning duration. a) How accurately can a listener perceive a segment duration? b) Is this accuracy influenced by segment type as well as the duration of neighboring segments? c) If a segment duration is changed, what compensatory effects does a listener expect?

Let us, before we discuss our data, rewrite the questions above into one which according to our opinion has been too much neglected in speech research: How accurate does a descriptive model have to be if perception sets the limits?

We have tried to estimate the detectability of differences in intensity and spectral slope for various vowels, paper (10). The vowel stimuli were produced by adding sinusoids from simulated frequency generators and presented in ABX tests. A strong vowel dependence was apparent and the estimated DLs for spectral slope changed from about .75 dB/oct for [i] to nearly 3 dB/oct for [u]. In reorganizing our data taking the F3 level difference as parameter we found a nearly vowel independent description of the result. The rationale behind this choice of parameter is that for all vowels the level difference is great in this region and the audibility is comparatively high, except for [u], where masking effects occur. We conclude that level differences in prominent spectral regions seem more likely as a perceptual decision parameter than the spectral slope. This result cannot, however,

be generalized to natural fluent speech without caution before additional experiments have been carried out.

In a series of discrimination and rating experiments, paper (8), we measured the difference limen for duration in different kinds of segments using a speech-editing technique. In these experiments we also studied compensatory effects in perception. This was done by moving the segment boundary between two segments, keeping their total duration constant. Vowels, voiceless fricatives, and nasals were used as test segments in the word context.

In addition to these experiments, production tests were carried out to study compensatory effects. The duration of a segment in a synthesized sentence was changed randomly and the subject had to compensate this variation by manually changing a test segment which was different from the randomized segment. The subject could, by this method, optimize the quality of the distorted utterance. The relation between the randomized segment and the segment controlled by the subject could then be measured.

We conclude from these experiments that psychoacoustic results regarding DL for the duration of tone, noise, and silence cannot explain the discriminability of duration in speech. Vowel duration proved to be most critical in the discrimination and rating experiments, indicating that vowel length per se is more important than the time between vowels or vowel onsets. This is in disagreement with current theories, such as the vowel onset theory. Our results support an approach that takes vowel length as a basic unit of the descriptive framework. Compensation within a consonant cluster has been shown both to decrease the discriminability and to increase the acceptability. In homorganic clusters a higher degree of compensation might be preferred possibly depending on less pronounced articulatory constraints. However, the demand for compensation between consonants and vowels is either very weak or non-existent irrespective of the position relative to the vowel or shared voicing cue. This points to the fact that vowel length (rather than just onset time) plays a primary role in perception. The positive correlation between vowel lengths within a word is of major importance, far more so than the normalizing effect of the sentence or the intrasyllable balance.

A rule synthesis facility
(mainly referring to papers (4)(5)(9))

Ideally, a speech synthesis program should be able to absorb any knowledge that has been gained concerning speech production, but it should also be easily modified to include new hypotheses about the speech production process, that could be falsified or verified by, e.g. listening tests. There have been at least two obstacles to accomplishing this. One is that the knowledge or hypothesis is in a non-explicit and perhaps even non-testable form; the other is that it is quite difficult to formulate linguistic facts in ordinary computer languages. The first problem is left with the individual speech researcher but the second we have tried to at least partially solve by implementing a new computer language. The basic structure of this language is close to the notation used in generative phonology.

This notation has the virtue of being familiar to most speech scientists and it has proved to be efficient in formulating phonological processes. The reason for this is that rules can be formulated sensitive to alternative contexts in which, e.g., optional elements can be specified. Also, the rules can operate on distinctive features rather than segments. This is, however, not sufficient when rules for phonetic realization are to be formulated. We therefore extended the notation to include continuous variables and arithmetic capabilities. Definitions of the symbols to be used and their associated features and variables are given by the user.

The program, i.e. the rule system, presupposes some input string and rules are applied in order, gradually transforming the string.

Special monitoring and control capabilities were also developed to increase the system's utility as a research tool. For example, the result of the transformational process can be followed on the terminal screen, intermediate results can be stored for later inspection, and statistics on rule application can be collected to establish the productivity of each rule. Moreover, in the synthesis process the user can alter the function of a rule by changing variables in the rule by means of a joy stick facilitating perceptual production experiments, as described in paper (10).

The ultimate output of the program is in most applications, though not all, a set of control functions for a synthesizer. In these cases a software interface is necessary, which includes display programs for parameter tracks. At present such an adaptation is made for both music and speech synthesis.

The speech synthesizer used is a serial terminal analog (OVE IIIb) with some important additions and modifications. The simple pulse source of the synthesizer is replaced by a functional glottal source model with physiologically related control paramters as described in paper (4). This addition has increased the possibilities of including knowledge concerning glottal articulation and hence produce better specified synthetic speech.

We have also included the possibility of mixing digitally stored signals with the output of the synthesizer under program control. This is useful for certain types of experiments where, e.g., segments from human speech can be included in the synthesis. Short transient sounds such as stop explosions are examples of sounds that are difficult to generate in the present synthesizer but could be of great importance, as demonstrated in paper (2).

A text-to-speech system
(mainly referring to papers (3)(4)(9))

The development of a text-to-speech system can be motivated by at least two different goals. First, to meet the need for talking machines in different kinds of applications as, for example, reading machines for the blind. Even for sighted persons it is sometimes inconvenient or even impossible to use the visual system for communication as in the use of an ordinary telephone or at some meter's distance from a graphical display. In several cases the visual system has to be focused on a specific point in order to handle a certain problem. Auditory presentation of information is in this situation advantageous.

The second, but not less important goal, is to answer the scientific question whether it is in fact possible to make such a system. This leads us to the ultimate test of our present knowledge of speech. Simple solutions can be used for selective applications but a general rule-based system producing high quality unrestricted speech is dependent on phonetic and linguistic research, since it can be regarded as a functional model of the speaking act. Therefore the total system may guide the general work on speech to important subjects and force researchers to formulate knowledge gained about speech into well-specified rules.

Our present work on a text-to-speech system has been guided by some views on the reading process. We know that a speaker normally pronounces a written word correctly, even if it is unknown to him. This process could be formulated as general pronunciation rules that exist in most languages. By using the rule language described in the preceding section we have been able to specify some of these rules together with rules for prosody and phonetic realization. We have with this method proved that such a system is possible for Swedish. The deficiencies in speech quality, however, reflect the restrictions put on the system and our present lack of knowledge about speech.

The text-to-speech system presented in papers (3)(4) and (9) consists of a sequence of rules that transform the text in successive steps. The first rules change the graphic form to a phonetic transcription taking both accent and phonetic quantity into account. In this process we make use both of consonant clusters and vowel quality to mark stressed and unstressed syllables. Endings are stripped away from the root morphemes and compounds broken down to their constituents.

Rules of this kind could be replaced by a huge lexicon containing diverse types of linguistic information. However, a lexicon of this type has to be of the order of 100.000 words to cover a corpus of about one million words.

In future applications of the system it is of course possible to include the use of a lexicon. In a practical text-to-speech system it could, for example, be advantageous to store a few thousand of the most common words in order to increase processing speed and accuracy. We have, however, tried the general and theoretically more interesting lexicon free approach. In doing so some faults in the phonetic transcription will be produced. The philosophy behind the system is, however, that all mistakes that are impossible to correct by some general rule should be according to Swedish pronunciation rules and thus understandable by the listeners despite the wrong pronunciation.

The prosodic part of the system is a variety of context-sensitive rules. According to the model described in paper (3) segment duration and intonation are dependent on different linguistic factors like stress, syllable, word and sentence structure. Clustering of unstressed words has in this process appeared to be of great importance. A limited dictionary of unstressed words has hence been introduced in the system. The special glottal source, paper (4) is used to signal stressed and unstressed syllables by spectral changes.

The phonetic part of the system takes care of the coarticulation and reduction effects. The notation used is highly efficient in formulating rules pertaining to degree of aspiration, extent of nasalization, etc. Since the terminal analog OVE IIIb and the glottal source are not articulatory models the rules have to operate on an acoustic descriptive level. This could be regarded as a drawback but, on the other hand, it is much easier to study speech production on an acoustic than on, e.g., a muscular level.

In this section we have not intended to describe the text-to-speech system in detail. The rules are continuously being changed and the quality increased. However, the goal has been to show that a system based on rules is a possible solution, giving a general talking machine for an unrestricted text. In the next section an evaluation of this system is presented.

Evaluation
(mainly referring to paper (11))

The text-to-speech system described in the preceding section is continuously developed in an iterative manner. By judging the output of the rule systems on different levels with perceptual techniques and comparison to human speech production, we obtain a solid base for changing the system constructively. The quality of the output will improve as more knowledge of the speech process is accumulated in the rule system.

In January 1976, however, we felt the need to make a more general evaluation of the total system. Since we wanted the system to include the use as a reading machine for the blind, an evaluation was performed in cooperation with the Swedish Association of the Blind. Different aspects were studied covering both intelligibility, acceptability, and learning effects as well as possible applications.

Test lists containing 25 sentences each were presented to eight visually handicapped subjects who were asked to repeat the sentences. No feedback concerning the correct response was given. Two test sessions of about half an hour each were arranged with a one-week interval. This method allowed a study of both the learning effects during one test session as well as a possible decrement in performance after the interval. A longer passage of speech was also presented and opinions about synthetic speech as an aid for the visually handicapped were collected.

We found rapid learning which remained unaffected after the one-week interval. The subjects could be divided into two groups, one claiming that the test sessions were tiring and one of the opposite opinion. However, in the end of the experiment the two groups had about the same test result.

All subjects thought that speech synthesis will become a useful tool in the near future. They were also of the opinion that synthesis should at first be used for textbooks, peripheral literature, and newspaper material and to shorten or eliminate the time between printing publications and recordings of talking books.

The result of this experiment is encouraging in the sense that it has demonstrated the feasibility and usefulness of a reading machine for the blind developed according to our concept of speech synthesis.

Final remarks

We have, in our thesis, demonstrated the importance of putting hypotheses within the frame of testable models. Since the topic is speech, there has been a need to develop methods of generating and manipulating speech in different ways, the choice of technique being dependent upon the actual question. Models of speech production can be tested on different levels, the most obvious being the acoustic level. Parameters on this level can however, vary greatly without affecting the message. Since in speech communication there is both a speaker and a listener involved we think that there are good reasons to include what we know about perception in evaluating speech production models and also to use easily manipulable speech in speech perception research.

Apart from the theoretical advantages of putting facts into models, there is in the case of speech some interesting applications of such models. We have already mentioned the text-to-speech system that could be used as a reading machine for the visually handicapped and as a talking aid for people with speech defects, let alone applications in those situations, where a spoken presentation is preferred also for non-handicapped. Work within this framework is also under way in studying imperfect rule systems typical for deaf speakers or interactions from different rule systems found in the speech of foreigners studying Swedish. These

new systems of rules could be used not only to demonstrate certain abnormalities in speech, but more important they could be used to evaluate different pedagogical efforts. The optimal learning strategy will most certainly be shown to be different for people with different language backgrounds.

The work presented in this thesis is a contribution to the never exhausted field of speech research, an endeavour that gradually increases our understanding of the human being as a communicator and that will make possible new methods and products in, e.g., communication, teaching, and rehabilitation.

Acknowledgments

Our sincere thanks are due to Prof. Gunnar Fant, Dept. of Speech Communication, Royal Institute of Technology (KTH) for his support, encouragement, and cooperation. A thesis work like ours is not an isolated manifestation done in order to get a degree, a title, or a hat. It is just a part of the general work within our laboratory to understand more about the speech communication act and to use this knowledge in applications. We thank all friends at the Dept. of Speech Communication for letting us take part in this process.

We thank the members of the Linguistic Department, University of Stockholm for profitable cooperation.

We thank all friends at home and abroad for collaboration support, encouragement, and most important, friendship.

Our thanks are also due to the Swedish Board for Technical Development (STU) and the Bank of Sweden Tercentenary Foundation (RJ), that have supported this work financially.

References

The thesis consists of the present summary and the following papers:

1. R. Carlson, B. Granström, and G. Fant: "Some studies concerning perception of isolated vowels", STL-QPSR 2-3/1970, pp. 19-35.
2. R. Carlson, B. Granström, and S. Pauli: "Perceptive evaluation of segmental cues", STL-QPSR 1/1972, pp. 18-24, also pp. 206-209 in Conf. Record 1972 Conference on Speech Communication and Processing, April 24-26, 1972, Boston, Mass.
3. R. Carlson and B. Granström: "Word accent, emphatic stress, and syntax in a synthesis by rule scheme for Swedish", STL-QPSR 2-3/1973, pp. 31-36.
4. M. Rothenberg, R. Carlson, B. Granström, and J. Lindqvist-Gauffin: "A three-parameter voice source for speech synthesis", pp. 235-243 in Speech Communication, Vol. 2 (ed. G. Fant), Almqvist & Wiksell, Stockholm 1975.
5. R. Carlson and B. Granström: "A phonetically oriented programming language for rule description of speech", pp. 245-253 in Speech Communication, Vol. 2 (ed. G. Fant), Almqvist & Wiksell, Stockholm 1975.
6. G. Fant, R. Carlson, and B. Granström: "The [e]-[ø] ambiguity", pp. 117-121 in Speech Communication, Vol. 3 (ed. G. Fant), Almqvist & Wiksell, Stockholm 1975.
7. R. Carlson, G. Fant, and B. Granström: "Two-formant models, pitch and vowel perception", pp. 55-82 in Auditory Analysis and Perception of Speech (eds. G. Fant and M.A.A. Tatham), Academic Press, London 1975.
8. R. Carlson and B. Granström: "Perception of segmental duration", pp. 90-104 in Structure and Process in Speech Perception (eds. A. Cohen and S. G. Nooteboom), Springer-Verlag, Berlin 1975.
9. R. Carlson and B. Granström: "A text-to-speech system based entirely on rules", pp. 686-688 in Conf. Record 1976 IEEE International Conf. on Acoustics, Speech, and Signal Processing, April 12-14, 1976, Philadelphia, Pa.
10. R. Carlson and B. Granström: "Detectability of changes of level and spectral slope in vowels", STL-QPSR 2-3/1976, pp. 1-4.
11. R. Carlson, B. Granström, and K. Larsson: "Evaluation of a text-to-speech system as a reading machine for the blind", STL-QPSR 2-3/1976, pp. 9-13.