MODELLING DIFFERENT WAYS OF SPEAKING IN A TEXT-TO-SPEECH SYSTEM.

Rolf Carlson och Björn Granström
Department of speech communication and music acoustics
KTH, Stockholm
Phone 468 - 790 7568

Introduction

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. Most of the testing of these systems have been carried out with isolated words, frequently monosyllabic nonsense words. Varying the voice of the synthesis, if at all possible, has frequently been modelled as a simple transformation in the synthesis parameter domain. Likewise basic variation of speaking rate has been thought of as a linear change in the time domain. Vowels, consonants and pauses are in this case affected by the same factor, irrespective of stress, essentially by changing the time interval between synthesis parameter updates. We have started to look into some of these shortcomings in the context of our text-to-speech system.

The demand for different speaking styles

There is 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 ask for a variation close to the one found in human speakers. General use in reading stock quotations, weather reports, electronic mail or warning messages are examples where humans would choose rather different ways of reading. The most common application today is in aids for the handicapped.

Visually impaired persons have very different needs (Carlson & Granström, 1986). On one extreme end of the style continuum they want hypercorrect speech that gives them maximally exact information how the text is written. Spelling is in this case a possibility, but it is too slow and gives hardly any understanding of the text content. The other extreme is very fast speech for information scanning. Using the device as a speech prosthesis implies the general need of variation that any human feel. There is, however, a problem to make the variation available to the handicapped person in this case, since communication speed is the primary concern. Much of the variation found in natural speech is not controlled deliberately, at least not at a very conscious level. The possibility has been discussed to connect the degree of emphasis to some myoelectric signal rather than controling it with special symbols in the text . To our knowledge this has not yet been successfully tried.

Apart from these practical needs in text-to-speech systems there is the scientific interest to formulate our understanding of human speech variability in explicit models

Current possibilities

In our current system there are several possibilities to vary the speech. On the global level there are possibilities to vary speech tempo, level and voice parameters like amount of aspiration, mean pitch and pitch dynamics, vocal tract length, general speed of articulation etc. Some of these parameters have been combined to define different voices similar also to female or child voices. The simulations of such voices are, however, still not very convincing. One aspect that we currently are working on is to include a more sophisticated voice source in the model. This new model will make it easier to simulate different ways of speaking.

The text-to-speech system consists of a combination of rules and lexica. It is always possible to enter a phonetic transcription, thereby getting full control of allophone selection. In the lexica, information on basic pronunciation is given but also on the dichotomy: function word/content word. So called strong and weak forms of words can both be given and selected by later rule governed processes. Lexical items can also be marked with extra information such as parts of speech or tags for phonological reduction processes.

In the rule components all sorts of effects could be modelled i.e. as long as they are rule governed. The conditioning factors for these rules could either be given as analysis by the system, such as syntax or some measure of predictability of words, as commands to the system or it could be given as extra information in the input text. One example of the latter is the emphasis control in the present system. By adding a number before each word in the phonetic string we can over-rule such things as default sentence stress assignment and function word reduction and also force different degrees of emphasis.

Experiments with speaking styles for British English

Although the concept of a multi-lingual text-to-speech system is a familiar one, rather little attention has been given to the question of the variety of each language that is synthesized. Language variation is currently being incorporated into our system both as concerns dialect and style (Bladon et al.,1987).

Our British text-to-speech implementation has been extended to provide a "style variable", a user-set range of ten values. This device can be used, for example, to propagate more affrication with a "lower" style number. The area of the system in which we first explored this style variable was in fact that of the forty or so function words ('can, have, for, them' etc.) of British English whose pronunciation, though not their spelling, varies considerably with sentence context and style. As an example, the word 'can' in a phrase 'I can go' may have a large number of realizations, some of which may be just acoustically specifiable subtleties, but some at least of which can be rendered transcriptionally: [kæn , kᵊn , kəŋ , kŋ̍ ,ʔŋ̍]
It is probably reasonable to rank these forms from left to right as graded from most formal to casual. They can therefore be synthesized with style variable values of say 9, 7, 5, 3, and 1, respectively. There are doubts whether style is possible to model on a one dimensional variable, especially for an entire utterance. Style may also change within the utterance and there may well be style variants that are lexically, and individually, determind. An illustration of the simple approach can be seen in figure 1. The three utterances are controlled by the same phonetic string, and interestingly there is no rule specifically concerned with speech rate. The faster speech of the "casual" style is due primarily to the greater reductions allowed.

To undertake this style ranking more widely through English phonetics is, in the present state of knowledge, rather an uncertain exercise. The normative data have hardly been collected at all. At the same time, there are two particularly good motives for pressing ahead. One is that, at present, the text-to-speech developer is faced with some uncomfortable decisions of simplification when specifying such a highly variable word as "can". Another reason is a research issue.

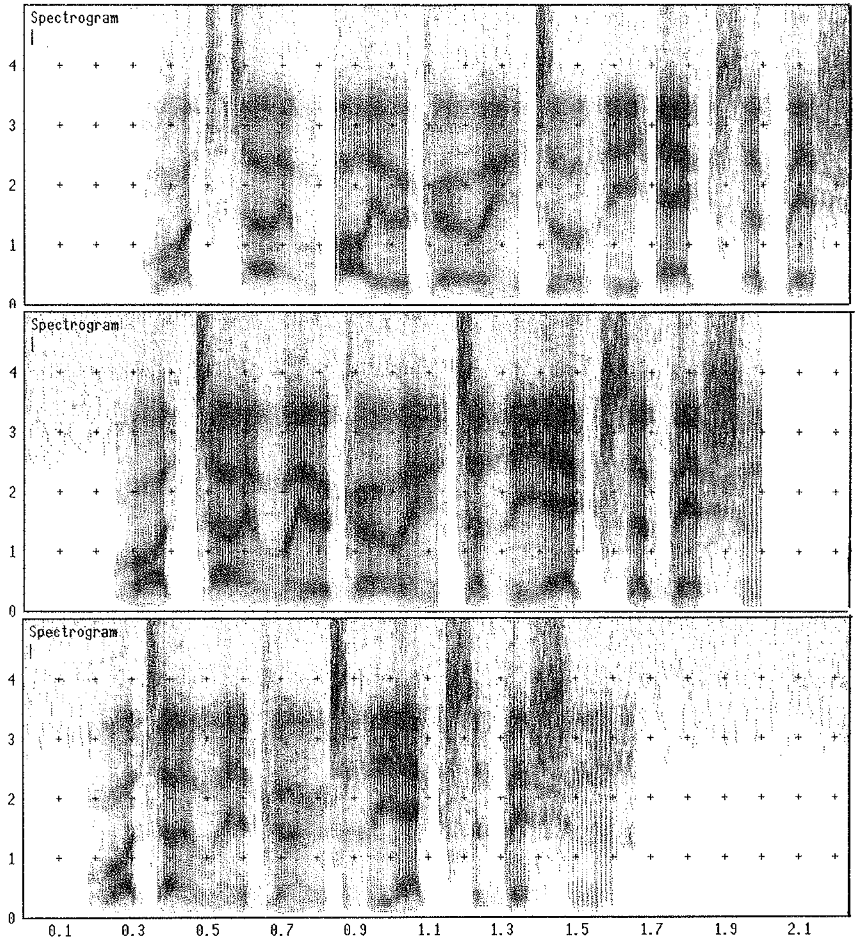

Figure 1.The synthezised sentence "What time are you going to the exhibition" spoken in "formal"(top), "normal"(middle) and "casual"(bottom) speaking style.

Fast speech for the visually impaired

Even if there are means of varying the speaking rate in the normal text-to-speech system they are not appropriate for the the extremely high speaking rates demanded by the visually impaired. Normal speaking rate is often estimated to be around 150 words per minute (wpm). This measure will be language and text dependent and it will also increase if pauses are not included. The demand from the blind is to obtain speaking rates of around 500 wpm to approximate fast silent reading by sighted persons. One way to obtain these "super human" rates would be to ignore progressively more of the low information content of the text. This would mean some kind of keyword reading without any language stucture. After preliminary tests we abandoned this idea. One problem with this solution is how to predict where the keywords are. It also seemed to us that the lack of linguistic structure was very confusing. If at all possible it is our conviction that we should model the speech on human performance. At these high speeds, however, there is no good human templet. We don't want to extrapolate from the point where human speech production breaks down.

The rationale in the present attempt to increase speed further is that the phonetic component could be changed for a reduced and faster component at runtime appropriate for very high speaking rates. Especially prosodic rules are simplified or taken away. The differences between stress and unstressed syllables are still marked by duration and fundamental frequency, a new set of inherent durations are established and mean pitch also needed to be increased. To most listeners the speech is close to unintelligible at this speed though it is claimed to be useful by experienced blind listeners. In figure 2 two reading speeds are shown.

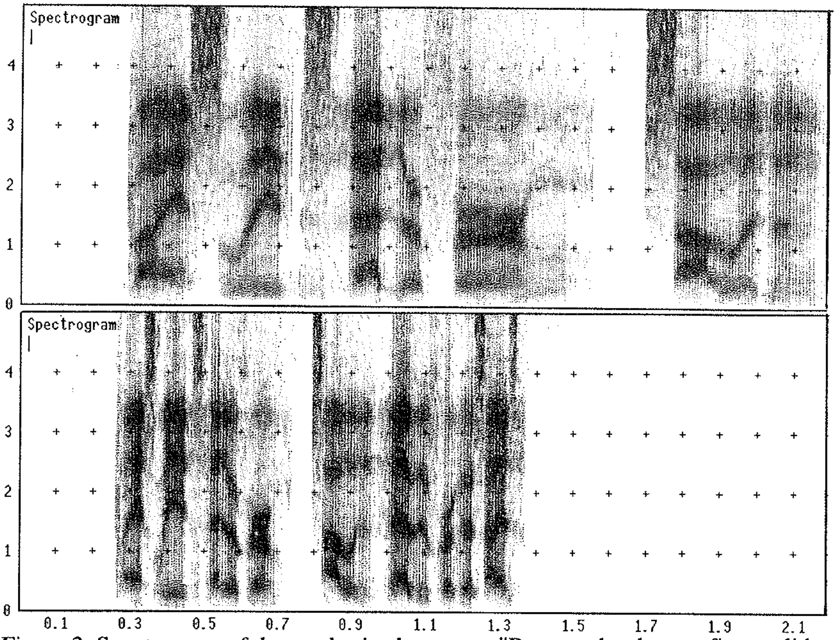

Figure 2. Spectrogram of the synthesized sentence "Buy swedes that are firm, solid and heavy for their size" at normal speaking rate(150 wpm, truncated) and at very high rate (500 wpm)

References

Bladon A., Carlson R., Granström B, Hunnicutt S. & Karlsson I.(1987): "Text-to-speech system for British English, and issues of dialect and style",European Conference on Speech Technology, vol. 1, Edinburgh, Scotland.

Carlson R. & Granström B. (1986): "Applications of a multi-lingual text-to-speech system for the visually impaired", pp. 87-96 in (P.L. Emiliani, Ed.): Development of Electronic Aids for the Visually Impaired, Martinus Nijhoff/Dr. W. Junk Publ., Dordrecht.