RULSYS - THE SWEDISH MULTILINGUAL TEXT-TO-SPEECH APPROACH.

Rolf Carlson, Björn Granstrom and Sheri Hunnicutt*

Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology, Stockholm, Sweden

*names in alphabetical order

ABSTRACT

Speech synthesis has been a major field of research at our department for several decades. The projects contain everything from basic research on speech production models to applications of speech technology, e.g., for handicapped persons. In this contribution we will concentrate on the development strategies, describe the development environment and discuss some recent results. The synthesis is based on a combination of modules including lexica and rule components. Even if the number of components are about the same for different languages, the emphasis on the different parts varies considerably due to language structure. Rule development is done in the generative phonology tradition. The development system, originally written for a different computer, has now been moved to our network of Apollo workstations and integrated with speech analysis and resynthesis software. Expanded use of morphological and syntactical analysis has proved useful in several languages. Recent experiments with an expanded synthesis model including a more realistic voice source, the LF-model, has given new possibilities to vary both speaker type and speaking style.

THE CASE FOR MULTI-LINGUAL TEXT-TO-SPEECH

Many societies in the world are increasingly multi-lingual. The situation in Europe is an especially striking example of this. Most of the population are in touch with more than one language. This is natural in multi-lingual societies like 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 get in direct contact with several languages on an almost daily basis. Text-to-speech devices, whether they are used professionally or not, ought to have a multi-lingual capability. Already today, Infovox, the only producer of a truly multi-lingual system, estimates that about 25% of their sales are delivered with more than one language. Multi-lingual text-to-speech then obviously fills a very practical need, but it also opens ways of scientifically studying and comparing languages.

THE RULSYS APPROACH

The text-to-speech system developed in our department was originally designed to be multi-lingual. The language-specific parts are mostly formulated in a notation close to the one commonly used in the linguistic tradition referred to as generative phonology. The language independent parts of the system are, however, efficiently coded, partially in assembler language for the microprocessor and the signal processor involved. Ideally, this needs to be done only once. The task of improving the system for a particular language or adding a new language to the system will be a continued effort. This work is then done in a development environment that is easy and familiar to many speech and language researchers. There is no need for conventional skills in computer programming in this process. This means that the expert directly can test and implement ideas for the text-to-speech system without the problem of communicating them to a computer programmer with the potential risk of misunderstandings or loss of information. This working situation is the key to the many languages presently available.

Versions of the text-to-speech program are now commercially available in British and American English, German, French, Italian, Spanish, Norwegian, Danish and Swedish (Bladon et. al., 1987; Kohler, 1988; Barber et. al., 1988; Granström & Gustafson, 1986; Granström et. al., 1987). Some other languages have also been studied in this context, but not brought to a state where they are ready for public use. The generality of the system has been exploited in many other research applications such as in a music synthesis project and in modelling the speech of deaf persons.

STRUCTURE OF A TYPICAL TEXT-TO-SPEECH SYSTEM

The general structure of the language dependent parts of the system can be seen in figure 1, and is described in detail elsewhere (Carlson et. al., 1982).

The DIG box transforms numeric expressions, including, for example, monetary amounts, to pronunciations/phonetic strings. The LEX component is the only component not formulated as rules. It contains both whole words and roots that for some reasons are not handled by rule. The SUF component strips the endings from words to be looked up in the LEX. The ROT component merges roots and endings on the phonetic level. The BLISS is a component that is used for syntactic analysis. Finally the FON component produces parameters for the language-independent speech production model, the speech synthesizer.

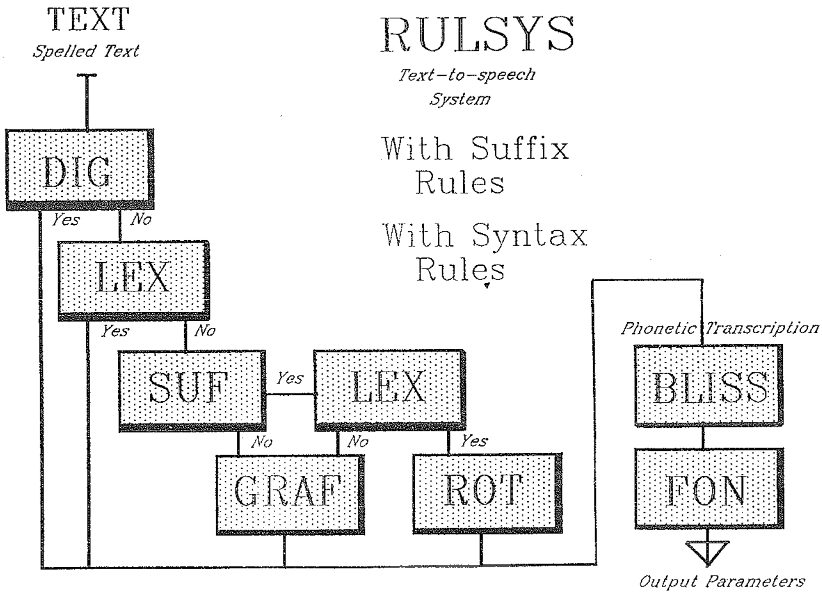

Figure 1. The general structure of the text-to-speech system.

ADDING A NEW LANGUAGE

The driving force for developing a new language for the system has varied. Some of the languages were originally developed under Swedish research contracts. Being part of a very small (9 million persons) language community, Swedes frequently have to work in other languages. For further development, cooperation has been established with other research teams throughout Europe. Some of this development for the major European languages has been supported by Infovox, anticipating commercial benefits.Other projects like Danish and Norwegian have been financed as handicap research to make the technology and applications available to the handicapped community.

It is hard to give an estimate to the effort and cost needed to create a version of the text-to-speech system for a new language. The quality criterion for a system is strongly tied to the motivation of the user and the specific application. The quality will only gradually, after many years of hard work, get close to the quality of human speech. A workable system, however, has in some instances been running in less than a man-year. Languages differ very much in the complexity needed for the different components of Figure 1. The background knowledge available for different languages is also quite different. In some countries, with a strong tradition in acoustic phonetics and linguistics, it will be much easier to find the right experts. Much of the knowledge is already available, even if it is not expected to be in a form appropriate for the text-to-speech system. One main factor for deciding the complexity of the task is how obvious and rule governed the relation between the orthography and the phonetic transcription is. Another important factor is how similar the phonetic structure is to other languages already developed.

DEVELOPMENT TOOLS

The central part of the development facility is the rule development environment, where rules can be formulated, compiled and tried out in separate rule components. Trace facilities of different kinds exist, from simple printout of the result to detailed traces of individual rule applications. Statistics of rule productivity for example can also be gathered. On the parameter level, graphs of the parameter tracks can be displayed. Support programs for maintaining frequency ordered reference word corpora have been developed. The corpora exist in orthographic and phonetic form and also in some cases with additional information such as parts of speech tags and morphological structure. These corpora are used to generate the appropriate lexical components for the text-to-speech system. The development environment is implemented on several computer systems including IBM/PC compatibles and most recently on our network of Apollo workstations. The Apollo system is also used for general speech analysis, and it has been possible to combine speech analysis and synthesis in an environment efficient for developing the phonetic component of the text-to-speech system. One important component is the rule controlled data bank search (Carlson & Granström, 1986a). The system accesses label files in our recorded speech data bank. These label files have the phonetic transcriptions stored together with the location in time of all segment boundaries. A search rule system, written in the RULSYS notation, uses these files as input. Depending on the need, different phonemes in specific positions can be marked and stored for examination. We can, for example, formulate a rule that picks all final tense /e:/ vowels. These vowels can be analyzed and presented in a contour histogram representation. The bottom graph in Figure 2 is an illustration of this method.

An important complement to the rule generated synthesis is manually controlled synthesis, so called HI-FI synthesis or "analysis by synthesis". We have developed an extension of RULSYS, called HISYS, which can use rule generated parameter data as input. This will give the user a fast start and also create a link between the two approaches. These data can then be edited according to the user's wishes. Figure 2 gives an example of this type of work. On the top is a spectrogram of the Swedish word "enighet," that should be synthesized. The phonetic transcription of this word is entered in RULSYS and the rule-controlled parameter tracks are sent to the HISYS component. The spectrogram is put under these tracks on the screen and the duration for each segment is adjusted. The adjusted parameter tracks can be seen in the figure.

With the help of a cursor, each parameter can be adjusted, deleted or new points inserted. During the whole process, the synthesis can be listened to in an interactive way. The sound can also be compared to earlier versions. At a certain point, the user might want to study the spectral shape at a specific time frame in the synthesis. The point is marked and a software simulation of the synthesizer is used to create the corresponding spectrum. This can, of course, be compared to the original recording, but as an alternative, the result of a data bank search can be used.

A NEW GLOTTAL SOURCE MODEL

For many years research on an improved glottal source model has beencarried out in our department. A review of the general properties of our four-parameter "LF"-model is given in Fant et al (l985).

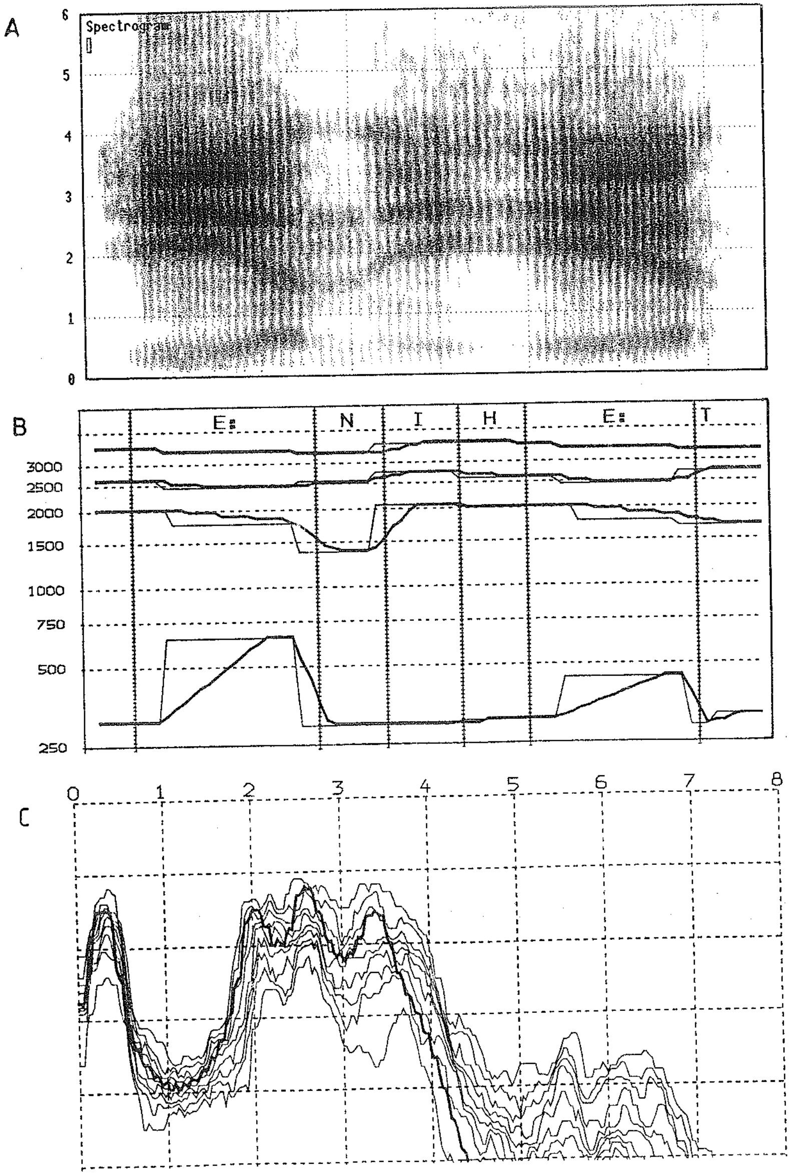

Figure 2. Spectrogram, parameter tracks and spectral sections used in synthesis development.

Accumulated experience of voice source variation in continuous speech has also been structured into

synthesis rules for female and child's speech (Fant et. al., 1987; Gobl, 1988; Karlsson, 1988). The new voice source has been integrated into our text-to speech system and we attempt to organize the rules into physiologically interpretable parameter variations operating within a linguistic frame, covering contextual variations, segmental assimilations and prosodic categories. Interactions between the voice source parameters and parameters of the vocal tract such as formant bandwidths are also handled by rule. It is the dynamic variations rather than the stationary properties of the model which contribute to naturalness.

LINGUISTIC ANALYSIS

Syntactic analysis

It is well known that different aspects of the syntactic structure have an important influence on the way speech is produced. This is especially true of prosodic realizations, e.g., pause insertions and phrase-final lengthening. A full syntactic parsing normally relies on an extensive dictionary with parts-of-speech information. In our text-to-speech system, this is at present not feasible. However, complete parses are not always required. Local clause and phrase information is expected to contribute essentially to the quality of the speech output. An alternative approach is to use only a limited dictionary and to predict parts-of-speech from surface word structure. To this end we have created such a system based on a small rule component that has been evaluated on the 10,000 word vocabulary (Carlson & Granström, 1986b) It was first subjected to a parts-of-speech labelling which was done in a semi-automatic fashion. A short rule system was developed that made parts-of-speech assignment based on orthographic surface structure criteria. The results demonstrate among other things the effect of word frequency. For the 200 most common words, we get a total prediction error of 90%. This is obviously due to the high proportion of function words in this frequency class. With decreasing word frequency, the total error approaches 20%. We obtain the smallest error for nouns, which is also the predominant category. We regard these results as very promising. The most frequent one thousand words cover about 63% of a typical Swedish text. A small dictionary with these words combined with the rules gives around 90% correct parts-of-speech assignment on a newspaper text. This is expected to form a good basis for a parsing system without an extensive dictionary as long as complete parses are not required.

The lexical component

Even though the main emphasis in our approach is put on the rule components, the lexical component is of vital importance for the total performance of the system. The size needed varies widely between languages and depends on such factors as the complexity of the letter-to-sound correspondence and the amount of syntactic analysis we want to include. French, Spanish and Finnish represent languages that show very few exceptions to the letter-to-sound rules. English and Swedish have quite a number of exceptions even to rather extensive rule sets.

Our reference word corpora contain frequency-sorted full words as they appeared in the sample texts. There could be several problems in using these corpora directly to create the exceptions dictionary. One is that the vocabulary is atypical for the application of the text-to-speech system. Using corpora of the kind generally available, i.e., based on newspaper text, gives low frequencies of everyday words, e.g., second person pronouns. In languages with a rich inflectional structure, many forms of common words will appear with low frequencies in the corpora, e.g., verb tenses or number and person inflections for nouns and adjectives. Another problem, especially in German and the Scandinavian languages, is that orthographic compounding makes a word dictionary less productive.

This calls for a more complex structure of the lexical component. Fig. 1 shows how it is handled in our present system. The LEX component contains both whole words and fragments and is searched twice. First the whole word is checked. If it is not found, it is passed through a word decomposition rule component. At present only word endings are removed, being the far most productive word modifications. Some rewriting is carried out at this stage (e.g. for English relied - rely+ed). The "root" is checked for occurrence in the dictionary and, if found, the pronunciation is combined with the ending in the word composition component. This procedure is rather simple compared to the recent effort at Bell Labs or the full morph decomposition in MITalk , but still it reduces the number of entries in the dictionary by about 30% for English and Swedish on the 10,000 word reference corpus. The size of the dictionary decreases even more and most importantly the productivity of the combined system increases.

A morphologically based orthographic normalizer for Danish

A Nordic cooperative project has been started to develop a text-to-speech device for the Nordic languages. The development is based on the system originally created in Stockholm. Danish poses some special problems for a text-to-speech system because the relation between the standard orthography and pronunciation is rather complicated. To tackle this, we have included a unique component in the system that transforms words into an idealized normalized orthography. This is accomplished through a morphological analysis based on a set of moderately large morph lexica. This component is written in C and precedes the ordinary "graph" rules. With a limited set of rules, the result is transformed to a phonetic transcription, including stress. The normally unstressed function words are identified through the ordinary lexical component. The combined system is designed on a frequency-ordered list of Danish words and guarantees correct pronunciation of the 14,000 most common words. For the future, the morphologically based approach shows good promise as a base for parts-of-speech analysis.

REFERENCES

Barber, S., Granström, B.,& Touati, P. (1988): "French prosody in a rule-based text-to-speech system", Proceedings of Speech '88, 7th FASE symposium, Edinburgh, Scotland.Edinburgh, Scotland.

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

Carlson, R. & Granström, B. (1986b): "Linguistic processing in the KTH multi-lingual text-to-speech system", pp. 2403-2406 in Proc. ICASSP 86, Vol. 4, Tokyo.

Carlson, R., Granström, B. & Hunnicutt, S. (1982): "A multi-language text-to-speech module", Conf Rec IEEE-ICASSP, 1982, Paris.

Fant, G. Liljencrants, J & Lin, Q. (1985): "A four-parameter model of glottal flow", STL-QPSR 4/1985.

Fant, G., Gobl, C., & Karlsson, I. (1987): "The female voice - Experiments and overviews", J.Acoust.Soc.Am. 82, S92(A).

Gobl, C.(1988): "Voice source dynamics in connected speech" STL-QPSR 1/1988.

Granström, B., Gustafson, K.(1986): "Toneme 1 1/2 in a Norwegian text-to-speech system",Nordisk Prosodi IV, Odense.

Granström, B., Molbaek Hansen, P., & Gronnum Thorsen, N. (1987): "A Danish text-to-speech system using a text normalizer based on morph analysis", European Conference on Speech Technology, vol. 1 , Edinburgh, Scotland.

Karlsson, I. (1988) "Glottal waveform parameters for different speaker types", Proc. of 7th FASE symp. SPEECH, Edinburgh, 1988.

Kohler, K., (1988): "An intonation model for a German text-to-speech system" Proceedings of Speech '88, 7th FASE symposium, Edinburgh, Scotland.