TEXT TO SPEECH SYNTHESIS BASED ON GRAPHEME AND PHONEME CLUSTERS

K. Fellbaum*, J. Rook*

ABSTRACT

We are developing a text to speech synthesis system for the German language, which has the same type of speech elements in both, the linguistic-phonetic transcription and the phonemization level, namely 'clusters'. They are defined as sequences of graphemes or phonemes of the same type (vowel or consonant clusters). Compared to other speech elements (e.g. diphones or demisyllables), the number of clusters is remarkably lower. Another advantage of clusters is the fact that cluster borders are usually correlated with stress borders, which reduces concatenating problems and improves the naturalness of synthesized speech. This paper describes our transcription and phonemization techniques and some hardware aspects.

INTRODUCTION

Speech synthesis is necessary in all cases where an unlimited vocabulary has to be pronounced (an example is the 'reading machine' for blind people). Even if the vocabulary is restricted, a synthesis sometimes has advantages compared to 'voice store and forward' techniques, above all, if in certain applications the vocabulary has to be changed or extended in a flexible manner. On the other hand, all synthesis techniques, which have been developed so far, suffer from a 'robotic' sound, which might lead to acceptance problems. There are some reasons for unnaturalness which will be discussed later.

A speech synthesizer consists mainly of three modules (fig.1). The first one transforms the ordinary orthographic text into phonetic symbols which describe the pronunciation much more precisely than orthographic text.

Fig. 1: Principle of speech synthesis

* K. Fellbaum and Jörg Rook, Technical University of Berlin, West Germany
We call this process 'linguistic-phonetic transcription'. In the second module ('phonemization') the symbols are assigned to the related phonetic elements. Finally, the third module ('signal reconstruction') concatenates the phonetic elements to continuous speech.

LINGUISTIC-PHONETIC TRANSCRIPTION

Starting point of our transcription are sentences of orthographic text in order to analyze sentence-prosodic parameters. At the moment we have only a very simple evaluation of sentence information (mainly punctuation marks). Additionally, a special list of abbreviations (e.g., a.s.o., e.t.c.) and algorithms which interpret numeric information (digits, dates etc.) have been implemented. The transcription works properly on the word level, more precisely on the morpheme level. For the morpheme recognition we have to determine the morpheme borders, which is in fact a difficult problem. It cannot be solved by comparing letter sequences with stored morphemes since the number of these morphemes is too high (in the order of 11,000).

A very useful approach was published by KÄSTNER (ref.1). He defined the unit 'cluster' as a sequence of consonants (consonant cluster) or vowels (vowel cluster). According to the position of the cluster within the word he distinguished 'initial', 'medial' and 'final' clusters. As an example, the German word 'ANGSTSCHREI' (shout of fear) has the initial cluster 'A', the medial cluster 'NGSTSCHR' and the final cluster 'EI'. Medial clusters often are further divisible. If not (or not further), we call them 'basis clusters'. In our example, the medial cluster 'NGSTSCHR' can be divided further into the basis cluster 'NGST' and 'SCHR'. Initial and final cluster are usually basis clusters.

It has been shown that in most cases the borders of final clusters are morpheme borders as well. Thus our search of morphemes (or its borders) is led back to a decomposition into clusters. A similar approach was published some years ago by RÜHL (ref. 2) and BREIDBACH (ref. 3). Our investigations have shown that a German text can be represented by about 400 basis clusters (which include some important foreign word stems). In addition to the clusters we try to find affixes which determine morpheme borders as well.

In order to get more reliable results we disintegrate each word forward and backward, which very often yields different results. Each of both disintegrations is then weighted and selected by certain decision rules which are based on probability measures.

PHONEMIZATION AND SIGNAL GENERATION

In principle transcription and phonemization are independent of each other. In our first synthesizer realization we used a commercially available IC (VOTRAX SC02) both for the phonemization and the signal generation. VOTRAX makes use of phonemes as elements. Since phonemes and concatenating procedures are implemented on the IC, we have a very simple, inexpensive and one-board solution for a complete synthesizer. On the other hand, the control parameters are largely fixed and the speech quality is rather modest.

We have therefore realized a new phonemization and signal generation module which is based on the LPC technique. Phonemization elements are phoneme clusters which widely correspond with the grapheme clusters.
The LPC technique has three advantages: firstly the fundamental frequency (pitch) and the signal energy can be modified and corrected if necessary. Secondly, prosodic information (stress, dynamic, sound durations) can be realized by well defined variations of the LPC parameters. Finally, it has been found out that the LPC filter concatenates the phonetic elements in a manner similar to the human vocal tract which makes preprocessing procedures easier and results in a good speech quality. Nevertheless, selecting clusters from natural speech is a very troublesome and time consuming process to be done by hand and with permanent hearing control. For this reason we have developed an interactive system with a graphic display and a graphic tablet which allows to mark the beginning and end of a signal section. This section can then be magnified and pronounced. If the section was wrong the marks are changed until a satisfying speech quality has been reached. In a second step the procedure has to be repeated with LPC analyzed and reconstructed speech since it might be different from the original speech. Finally certain corrections of the pitch and the level have to be made in order to get a better adaptation of cluster sequences and to generate a kind of micro melody (intrinsic pitch).

Fig. 2 shows an example of the selection process, where the clusters /tr/, /ai/ and /c/) are extracted from the words 'Strafe', 'mein' and 'dich' and then concatenated to the new word 'Streich'.

![Fig. 2: Selection and concatenation process for the clusters](image)

**HARDWARE CONSIDERATIONS AND CONCLUSIONS**

As mentioned above, we have realized an one-board synthesizer. The transcription is carried on a microprocessor MOTOROLA 68 000, which controls the VOTRAX SC02 as well. In a second version of our synthesizer we use the same transcription (including the M 68 000). Additionally we need a second board to perform the LPC phonemization and signal reconstruction by a signal processor TMS 32010 (fig. 3). A future version of the system will have mini-sized IC's which allow to place the whole synthesizer on one board.
Our first evaluation tests have indicated an acceptable speech quality, remarkably better than that of the VOTRAX version. Nevertheless, the sentence prosody still leaves something to desire. Thus in our future work we will investigate aspects of syntax and semantic analysis in more detail.

Fig. 3: Signal processor-based phonemization and reconstruction module

ACKNOWLEDGEMENTS

We have to thank Thomas Blume, Shahla Sehhati, Martin Rusch and Lutz Fliegner for valuable contributions. The synthesis project was partly supported by the PHILIPS company and the German Ministry of Research and Technology.

REFERENCES