TEXT-TO-SPEECH CONVERSION FOR GERMAN USING A CASCADE/PARALLEL FORMANT SYNTHESIZER

GERHARD RIGOLL**

Fraunhofer-Institute (IAO)
Dept. of Advanced Information and Communication Technologies
Holzgartenstr. 17, 7000 Stuttgart 1, West Germany

ABSTRACT

The paper describes some aspects of the use of the cascade/parallel formant synthesizer for German text-to-speech synthesis. Since the algorithms used for speech synthesis are relatively similar for the most languages, the paper emphasizes some novel approaches and special phonetic problems, such as the use of a mathematical model for the cascade/parallel formant synthesizer for the determination of the synthesizer control parameters or the synthesis of phonemes with special articulation, rather than to describe more generally the development of the entire system.

INTRODUCTION

In 1983, the work on the development of German text-to-speech converters, based on the cascade/parallel formant synthesizer developed by D.H. Klatt, has started. In general, the development of a text-to-speech system can be described under different aspects, e.g. emphasizing more the common technical problems, or the letter-to-sound conversion, or the fact that the system might have been modified for a different language, which usually requires a complicated modification procedure that was also performed for the system described here. In this paper, the phonetic aspects of the use of the cascade/parallel formant synthesizer for the German language are especially considered. The accurate determination of the main control parameters for the cascade/parallel formant synthesizer is probably the most important step in order to achieve high voice quality of the final system, although many different steps, e.g. letter-to-sound conversion or prosodics, which are not considered in this paper, are also responsible for the overall quality of the system.

THE CASCADE/PARALLEL FORMANT SYNTHESIZER

The formant synthesizer which was used for German synthesis is a modified version of the synthesizer described in /2/ which was improved by D.H. Klatt during the last years. The current version is now very flexible and capable of synthesizing different voices, including women and children voices. The most important control parameters are still the first three formants and bandwidths and the fundamental frequency. Additionally, it is possible to control special parameters for the voicing source and for prosodics, which is mainly used for the generation of different speaker characteristics. The German vowels and sonorants are synthesized using the cascade branch, while voiceless fricatives and plosives are generated by the parallel branch. Only for voiced obstruents, both branches of the synthesizer are excited.

MATHEMATICAL MODELLING OF THE CASCADE/ PARALLEL FORMANT SYNTHESIZER

There are basically three possibilities to obtain the values for the control parameters of the synthesizer. The fastest and simplest method is a perceptually based method, where the parameters for every phoneme are tuned as long until the synthesis of the phoneme sounds very similar to the original utterance of the phoneme. Although it is possible to find relatively fast a parameter constellation which is leading to an acceptable result for every single phoneme, the overall quality of such a system is mostly poor and many phonemes which used to sound natural when they where tested in isolation, sound unnatural in connected speech. The second method is based on the calculation of the parameters with the use of speech analysis tools, e.g. formant calculation based on an LPC analysis and the generation of the phonemes with the parameters obtained from this analysis for each phoneme. But also this method leads to problems because usually the generation of a sound, using the formant values which were obtained from an analysis of an utterance of this sound, does not lead to a synthetic sound with exactly the same acoustic and spectral properties of the natural utterance. The third method is a spectral tuning of the synthesizer parameters to the spectral properties of one single speaker. This is a very time consuming iterative process, where at the first step the initial values of the synthesizer parameters are derived from an analysis of a natural utterance, as in method 2. In the following steps, the parameters are tuned as long until the spectral analysis of the synthetic utterance is similar enough to the spectral analysis of the natural utterance. In this way, the system is forced to have almost the same spectral features as the single test speaker, which can lead to a high amount of naturalness of the synthetic voice. If such a procedure is used, it is obvious that the success is also dependent on the tools and algorithms which are used for the analysis and comparison of natural and synthetic speech. Beside the more traditional analysis methods like spectrograms and spectra, a novel approach was tested during the development of the German text-to-speech system. This approach is based on a mathematical model of the cascade/parallel formant synthesizer. Based on the fact, that both branches of the synthesizer are composed of digital resonators with the transfer function

$$G(z) = \frac{1 + c_i + d_i}{1 + c_i z^{-1} + d_i z^{-2}} \qquad (1)$$

where the resonator coefficients c_i and d_i contain the according formant F_i and bandwidth B_i as nonlinear functions:

$$c_i = -2e^{-\pi B_i T} \cdot \cos(2\pi F_i T) \qquad (2)$$

$$d_i = e^{-2\pi B_i T} \qquad (3)$$

** The author is currently with the Continuous Speech Recognition Group, Dept. of Computer Sciences, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

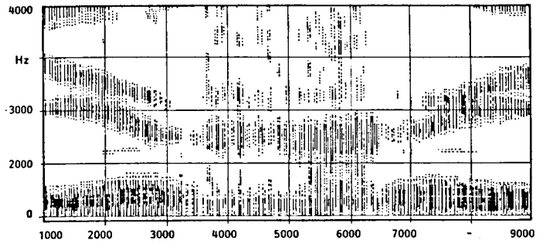

Fig. 2: Spectrogram of the phoneme sequence /iRi/

This periodical change can be demonstrated even better by looking at the formant tracks of this sequence in Fig. 3, obtained from the earlier described nonlinear parameter estimation procedure.

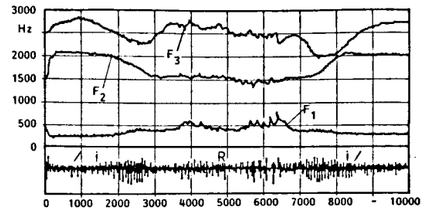

Fig. 3: Formant tracks of the phoneme sequence /iRi/ calculated with the use of a nonlinear parameter estimation algorithm

The synthesis of the uvular /R/ can be performed either by a modulation of the gain AV of the voicing source or by a modulation of the formants. It was decided to use the latter method in the current version, where only the second formant was modulated by a certain percentage of his stationary value, which is shown in Fig. 4. Simultaneously to this modulation, frication noise is added via the parallel branch by setting the gain A_2 of the second parallel resonator to a value different from zero. In this way, the uvular /R/ is handled similar to a voiced fricative.

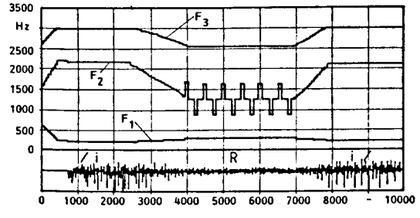

Fig. 4: Formant tracks given to the cascade synthesizer branch for the production of the phoneme sequence /iRi/

COARTICULATION

A very important module of a text-to-speech system are the rules for coarticulation. In the current version, coarticulation is performed in several ways

- the coarticulation of the spectrum of the stationary part of vowels and sonorants with the preceeding and the following phonemes is calculated by a general coarticulation formula which is applied to the formant values and reflects the percentage of the influence of the formant values of the neighbour phonemes to the formant value of the current phoneme
- the variations of boundary values in the transitions of consonants to different vowels are automatically taken into account by the application of the locus theory
- the variations of the spectral properties of consonants in the environment of different phoneme classes is taken into account by modification of the gains of the parallel resonators according to the current environment
- special extended rules are applied to some sonorants which show strong coarticulation. In the German language these are especially the phonemes /l/ and the uvular /R/. Again the /R/ is the most difficult phoneme to handle scince it requires complex rules for the formant values as well as for the control of the frication by the parameter A_2 if it appears in different environments. The coarticulation of these sounds can sometimes be effected only by the left or sometimes only by the right neighbour phoneme and often also by both neighbour phonemes, depending on these phonemes

Since the recording and tuning of phonemes is never carried out with phonemes spoken in isolation, special attention has to be paid to incorporate the coarticulation rules into the process of the spectral tuning of the various sounds because these rules will modify the originally entered parameter values according to the current phonetic environment, in which a specific phoneme is recorded, analyzed and tuned. This is a serious problem scince at the begin of the tuning task, the coarticulation rules are usually not yet known, but they are theoretically required in order to obtain optimal tuning results.

CONCLUSION

Some important steps of the synthesis of German with the cascade/parallel formant synthesizer have been described as well as some new approaches for the analysis of speech to obtain the values for the synthesizer control parameters. The description of the development of the entire text-to-speech system is beyond the scope of this paper. The experiences which were gained during this development have shown that the careful and time consuming tuning of every single phoneme and the consideration of many special cases and exceptions is the key to obtain a synthesizer with a high voice quality.

REFERENCES

/1/ G. Rigoll: The DECtalk System for German: A Study of the Modification of a Text-to-Speech Converter for a Foreign Language. Proc. IEEE-ICASSP, Dallas, 1987.

/2/ D.H. Klatt: Software for a cascade/parallel formant synthesizer. J.A.S.A., Vol. 67, No. 3, 1980.

/3/ G. Rigoll: A New Algorithm for Estimation of Formant Trajectories Directly from the Speech Signal Based on an Extended Kalman-Filter. Proc. IEEE-ICASSP, Tokyo, 1986.

/4/ A. Gelb: Applied Optimal Estimation. M.I.T. Press, Cambridge 1974.