

HIGH QUALITY AND REDUCED MEMORY TEXT-TO-SPEECH SYNTHESIS OF THE GREEK LANGUAGE

N. Yiourgalis*, G.Kokkinakis*.

ABSTRACT

A software, unlimited vocabulary text-to-speech synthesis system of the Greek language is presented. The system uses a cascade/parallel formant synthesizer, 125 speech segments and several new or modified techniques for segment coding, concatenation, intonation, etc., which provide a high speech quality along with a reduced memory for the control parameters.

INTRODUCTION

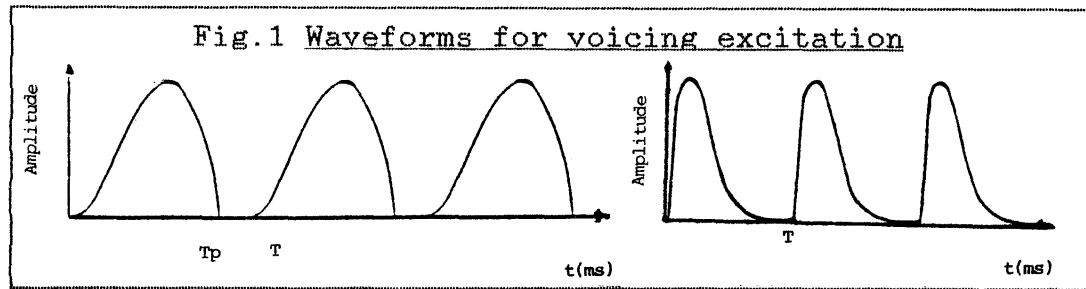
Unlimited vocabulary text-to-speech synthesis of the Greek language has been achieved at the Patras University, initially with a time-domain software synthesizer using 134 LP-coded speech segments and a lattice filter. The speech segments of the type V,C, CV, CCV etc, were extracted from natural speech and were standardized (ref 1). Work on this synthesizer is continued in order to improve the speech quality by adding prosodic features. Parallel to that, experimentation with a formant synthesizer using the same 134 character units was decided in hope that a better speech quality could be achieved. The D. Klatt synthesizer (ref 2) was chosen and an analysis-synthesis procedure was started in order to isolate appropriate speech segments from natural speech and to extract the formant parameters etc. During this procedure a reduction of the segments' number to 125 was decided.

As soon as a sufficient number of segments was analysed, experimenting with the synthesis of words and phrases was started. In this phase, several modifications of the synthesizer and new techniques were introduced, which substantially improved the quality of the synthetic speech and reduced the necessary memory for the control parameters. These are presented below along with comments on the results and the further work which has to be done.

SPEECH QUALITY IMPROVEMENT

a. Source excitation : For voiced sounds the waveform given in fig.1a was used, instead of the waveform given by Klatt (fig.1b). This waveform, which comes closer to natural source excitation, is described by a polynomial relating time t to periods T_p and T_n the glottis is open and closed respectively (ref 3). Several acoustic tests have proved its superiority.

* Wire Communications Lab., University of Patras, Greece.



b. Influence of the glottis condition on formants : The frequency and bandwidth of the formants depends on the open or closed condition of the glottis. The dependance is greater on the first formant F_1 , whose frequency during the open time of the glottis rises to a value larger than that during the closed time. To consider this change, two values of F_1 were introduced for voiced sounds: F_{1op} and F_{1cl} corresponding to the time the glottis is open and closed respectively. Depending on the time t of the appearance of an excitation source sample ($t < T_p$ or $t > T_p$), this sample excites the F_{1op} - or F_{1cl} - resonator.

c. Aspiration sounds : Aspiration sounds like /h/ are produced using the cascaded configuration of the synthesizer excited by the noise source. Aspiration noise tends to excite all but the first formant. To simulate this in the synthesis program, the first resonator is bypassed whenever aspiration is used to excite the vocal tract filter. A better /h/ was synthesized in this way.

d. Speech segment duration and V.O.T. : For the speech segments, a normalized length of 200 ms and a normalized V.O.T, if it exists, is provided. These values are then modified according to the phonemic environment of the segment. E.g. a vowel or sonorant is significantly shortened if it is followed by a voiceless plosive. Also, the V.O.T of a segment is shortened by 9 to 13 ms, if the segment before a prestressed / π , τ , κ / is a nasal. Several rules used for the modification of a segment's duration and V.O.T. are used which have significantly improved speech intelligibility.

e. Concatenation of segments : Generally the various segments are abutted directly. However when the end of a segment and the beginning of the next one are both voiced, concatenation is applied. The concatenation algorithm used in our case is a modified version of the well known method proposed by L.R. Rabiner et al (ref 4) for the concatenation of formant coded words. The algorithm works upon the last six frames of the first segment and the first six frames of the second segment. For each frame which has a length of 10 ms, the values of the formant frequencies are taken through contour-interpolation and stored in a matrix. The twelve values of each parameter are then merged in such a way that six new frames with smooth transition values from one frame to the next are produced. This procedure shortens the two segments by three frames each. In order to recover the

original length, the point in each segment with the smallest spectral derivative of the interpolated values is located and three frames with values equal to the value of the previous frame are inserted in each segment at that point.

The described method gives good results since no objectionable transients are created.

f. Intonation contours : The intonation algorithm creates an envelope for the four amplitude control parameters for the excitation sources AV, AVS, AH, AF and the pitch parameter FO, according to the breath group level or the punctuation mark at the end of a sentence. There are four punctuation marks in Greek: The comma, the full-stop, the question mark and the exclamation mark. Each of the above parameters has been given a contour for each punctuation mark, with a standard length of 1800 ms. Therefore, all cases are described by 20 contours. The contours were defined on the basis of statistical data obtained from pitch analysis of real speech.

The intonation algorithm provides that the contours chosen are expanded or compressed to fit the length of the text. It also modifies the pitch contour in voiced and unvoiced consonant places as well as in stressed segments.

MEMORY REDUCTION

Memory reduction results from a reduction of the data which must be stored for coding each of the 125 segments and from the way these data are stored and processed. In detail:

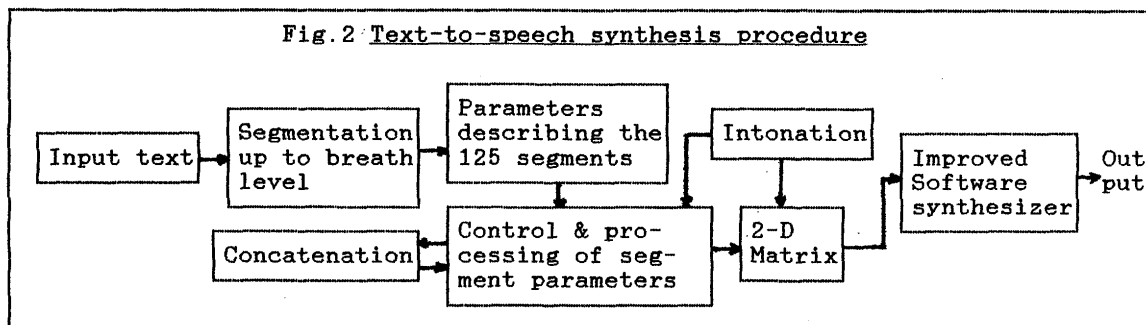
The 13 constant parameters of the synthesizer, which have the same value for all speech segments are kept in file No-1 used for the synthesis of all segments. The 18 variable parameters have different values for each segment and include the "actual parameters", which get different values every 10 ms (the frame update rate). They are stored in file No-2 consisting of four sections for each segment, in which the 18 parameter values, the code numbers of the actual parameters, the values of the contours and the amplitude control parameters (AV, AVS, AH, AF) for the excitation sources are kept respectively. Each contour is defined only by its pronounced points. The intermediate values are calculated by means of a raised cosine interpolation algorithm.

The above arrangement of the parameters in memory allows the synthesis program to become structured and fast. All the necessary data for decoding a segment are stored together in a compact way. The parameters are processed consequently from up to down, so that all the intermediate results needed in a point for some calculation have already been calculated in previous stages. Thus, processing time and memory are saved.

TEXT-TO-SPEECH SYNTHESIS SYSTEM

The speech synthesis algorithm has been implemented on a HP-9845B desk-top computer, which plays the role of both a controller and a synthesizer.

The algorithm uses a different subroutine for each complete process, e.g. concatenation, intonation, etc. The whole synthesis procedure is given in fig.2 as a block diagram.



The above system has been used as a research tool to examine the effect of different synthesis strategies, different contours etc. upon the intelligibility and quality of the produced speech. Both have been checked and improved during experimentation. Finally a speech of high intelligibility and quality has been achieved with the segments analysed up to now. Work is continued to complete the segment analysis and to perform systematic intelligibility and quality assessment tests. In parallel to this a hardware implementation of the system is under way.

CONCLUSION

A software text-to-speech synthesis system for the Greek language has been described, using formant synthesis, 125 speech segments and several new or modified synthesis techniques. The intelligibility and quality of the produced speech up to now with a limited amount of segments are very good while the memory needed for the parameter storage is substantially reduced in comparison to synthesis with LP-coded segments.

REFERENCES

1. P Stathopoulou, G. Kokkinakis, A.G. Mian, Inter Conf Info Sciences and Systems, Patras Univ, pp. 506-515 (1979).
2. D H Klatt, JASA, Vol 67, No 3, (1980).
3. N Yiourgalis, G. Kokkinakis, Digital Tech in Simulation, Communication and Control, IMACS (1985).
4. L R Rabiner et all, Bell Systems tech jour, pp. 1541-1558 (1971).