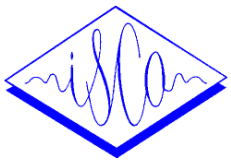


The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

QUALITY AND INTELLIGIBILITY IMPROVEMENTS IN A GREEK TEXT-TO-SPEECH SYSTEM

N. Yiourgalis, G. Kokkinakis

Laboratory of Wire Communications, School of Engineering, University of Patras, Patras, Greece
Tel: +30/61/991722, Fax: +30/061/991855, Tlx: UNPA GR 312 447

ABSTRACT

The quality and intelligibility of speech produced by a text-to-speech system for Greek using parametric synthesis by rules and 131 formant coded speech segments, have been substantially improved by controlling the V.O.T. and duration of each segment. This paper presents the technique devised for controlling these parameters along with a short description of the coding scheme and the concatenation algorithm.

1. INTRODUCTION

A modular software text-to-speech system for Greek using parametric synthesis by rules and Formant coded speech segments has already been presented [1, 2, 3, 4]. This system includes a sophisticated text pre-processor, a number handling module and an intonation routine working on a breath length level. Its main characteristics are the concatenation of a small number of speech segments instead of phonemes, the very short memory required for the storage of the segment parameters and the possibility to easily extend the number of exception words, abbreviations etc. that can be treated.

The quality and intelligibility of speech produced by the above system have been substantially improved by controlling the Voice Onset Time (V.O.T.) and the duration of each speech segment according to its phonemic environment. In the following the technique used for this control is described after a short presentation of the segment coding scheme and the concatenation algorithm.

2. CODING AND CONCATENATION OF THE SPEECH SEGMENTS

A block diagram of the Text-to-Speech system is shown in fig. (2.1).

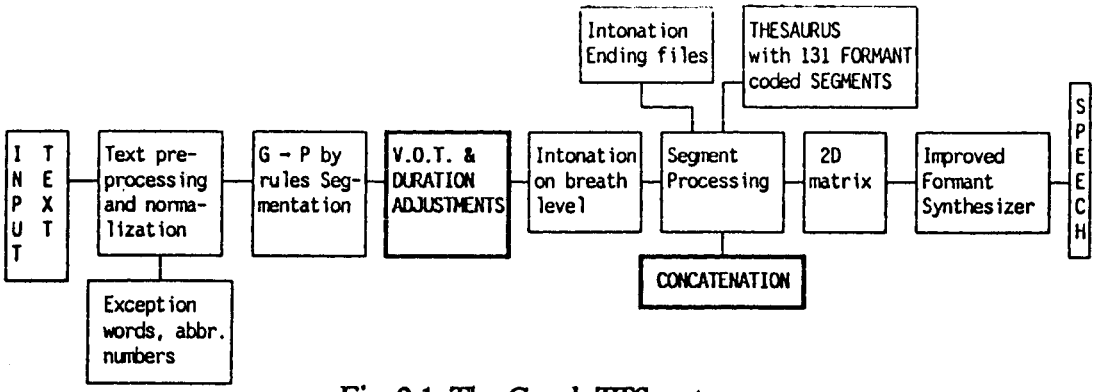


Fig. 2.1 The Greek TTS system

The system is based on an improved version of the Klatt synthesizer and uses a thesaurus of 131 segments of the Consonant (C), Vowel (V), Consonant-Vowel (CV) and Consonant-Consonant-Vowel (CCV) type. The decision to use such speech segments was based on the observation that any Greek word can be easily segmented immediately after each vowel. The CV and CCV segments represent the part of speech from the beginning of a consonant to the steady state of the vowel including in this way the coarticulation characteristics across consonant-vowel boundaries. The segments are formant-coded in the lexicon using a reduced-memory scheme [3]. Details for the coding and storage can be found in [1, 3]. Briefly, there are 9 constant parameters with the same values for all speech segments. The parameters describing each segment consist of 3 groups containing 18 variable parameters, the contour they follow, if any, and the activation/deactivation data for the four excitation sources, respectively. This coding allows the program to be structured and fast in execution. All necessary data for decoding a segment are stored together and data are processed top to bottom, so intermediate results needed for some calculations have been already calculated in previous stages.

The concatenation algorithm processes the last 10 frames of the first segment together with the first 10 frames of the second one, for each of the F1op, F1cl, F2, F3 ,FNP, FNZ parameters of the synthesizer in such a way, that smooth trajectories around their concatenation point are obtained. The formula used is given below:

$$F(j,k) = \frac{F(j,k)*k^{1.5}*Aveder(j,1)+F(j,k+10)*(k+10)*Aveder(j,2)}{k^{1.5}*Aveder(j,1)+(k+10)*Aveder(j,2)} \qquad (2.1)$$

$$Aveder(j,1) = \sum_{i=end-10}^{end} (F(j,i) - F(j,i-1)) \,/10 + 0.6 \qquad (2.2)$$

$$Aveder(j,2) = \sum_{i=end}^{end+10} (F(j,i) - F(j,i-1)) \,/10 \qquad (2.3)$$

where k = frame number
F(j,k) = value of parameter J at frame k.
Aveder(j,1) = mean spectral derivative for the parameter j of the first segment
j represents one of the concatenating trajectories
The range for k is from end-10 to end, where "end" is the length of the first segment

Formula (2.1) calculates a new frame value for the jth parameter, using two weighted values one from each segment. The frame number is raised to a power so that the new value departs from the first trajectory slowly, to reach the second one. It is obvious that concatenation shortens each segment by 5 frames. In order to compensate for this shortening, the smallest formant variation in the trajectory of each segment is located and 5 frames are inserted there.

The concatenation algorithm works only in voiced boundaries. The vowel-consonant boundaries are amplitude matched by allowing the energy of the vowel to decay during its last 10 ms, in order to match the low consonant's energy.

3. VOT and DURATION ADJUSTMENTS

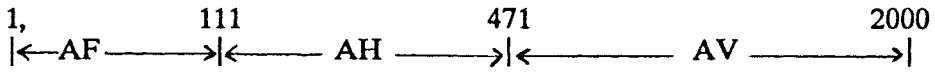
V.O.T. depends on articulation and on the identity of the following vowel or sonorant consonant [5,6]. It is a very important factor for the synthesis of complex segments. As a complex segment we define one for the production of which the voicing source does not participate

from its beginning but it becomes activated after some specific duration of noise excitation. For example, in the case of /pa/ the first source activated is AF (noise) for a duration of 5ms, the second for the next 40ms being AH source (aspiration) and finally participation of the AV source starts. In this case, V.O.T. is 45ms, that is the time elapsed before the glottis starts to oscillate. In a case where the V.O.T has not the correct value, a different sound is produced such as /ta/, or /pa/ with inferior quality. V.O.T values for all complex segments are extracted from natural voice by means of mainly perceptual tools.They are normalized and sorted in the lexicon (thesaurus) where all other segment coding parameters are stored.

In a synthesis scheme using concatenation of segments, V.O.T and duration adjustments are vital if speech of good quality and intelligibility is to be produced. This is due mainly because both VOT and duration vary depending heavily on the immediate phonemic environment of the segment and on its position in the word. For example, the distinction between /ba/ and /pa/ in the beginning of an utterance before a stressed vowel is manifested predominantly as a difference in V.O.T. The V.O.T value is altered in different positions of the segment in the utterance.

In the same way, modification of each segment's duration according to the phonemic environment is necessary, as it certainly deviates from the 200 ms normalized value that is used in the lexicon. For example, vowels or semivowels are of shorter duration if they are followed by voiced plosives.

To make these modifications a simple task, special attention has been paid to the coding scheme used to store the segments in the lexicon. V.O.T and duration modifications to an accuracy of one sample has been possible. Four parameters, one for each excitation source (AV, AH, AF, AVS) have been included. Each parameter takes two values which activate or deactivate the corresponding source. As an example, the timing (in sample numbers) for the different sources participating in the production of segment /pa/ are given below:



These are coded in our model as:

The two values for each source parameter AV, AH, AF and AVS, show the starting sample for the activation of that source and its duration in number of samples. Hence, in the above example: For samples 1 to 111, AF is the only source exciting the vocal tract, for samples 111 to 470, AH is the exciting source and for samples 471 to 2000, AV is the exciting source. The source AVS is deactivated.

Notice that the duration of this segment has been normalized to 2000 samples.Depending on the position of the segment in an utterance, the algorithm first calculates an offset sample value at which the first source becomes ON. Next, rules are applied to modify the VOT and duration depending on the corresponding segment environment. In this example for segment /pa/ the V.O.T is shown to be 470 samples as the AV source starts to oscillate after 470 samples. This value can easily be modified by rules according to the segment position and it can further be adjusted according to its environment.

Table (3.1) shows some of the rules implemented to adjust the duration and V.O.T of some segments. The rules have been extracted from experiments and they are going to be expanded in number to cover most of the cases met in the Greek language.

Table 3.1
V.O.T and Duration of some segments in different phonemic environments

SEGMENT	VOT	DURATION
-new segment added	no change in initial segment	vowel in initial segment is shortened by 66 %
-vowel+unvoiced plosive		shortened by 30ms
-semivowel+unvoiced plosive		shortened by 15ms
-nasal+stressed complex segment with /p,t,k/	shortened by 10ms	
-/p,t,k/ between vowels or /p,t,k/+ semivowel	15ms shorter if initial and 35ms shorter in any other case	
-unstressed segment between vowels	shortened	
-initial consonant		increased by 40ms
-final consonant		shortened by 40ms
-s+unvoiced complex segment	shortened by 20ms	
-s+unvoiced plosive +semivowel	shortened by 30ms	
-voiced plosive+liquid		liquid shortened by 15ms
-initial unvoiced plosive +liquid	increases by 13ms	liquid increased by 25ms
-vowel+stressed complex segments	increased	
-voiced fricatives+vowel		Vowel increased

3. CONCLUSION

The quality and intelligibility of a TTS system for Greek based on segment concatenation have been substantially improved by controlling the VOT and duration of each segment. Work on completing the VOT and duration adjustment rules is continued. Further improvents of quality and intelligibility are expected from the application of syntactic analysis which is currently under investigation

REFERENCES

[1] N. Yiourgalis, G. Kokkinakis: "High quality and reduced memory TTS of the Greek language", European conference on Speech Technology, pp. 187-190, Edinburgh, 1987.

[2] N. Yiourgalis, G. Kokkinakis: "Text processing for unrestricted Text-to-speech Synthesis of Greek", 7th FASE Symposium, pp. 247-254, Edinburgh, 1988.

[3] N. Yiourgalis: "A unrestricted input TTS for Greek using parametric synthesis by rules and Formant coded segments ", PhD dissertation, Patras University, 1989.

[4] N. Yiourgalis, G. Kokkinakis: "Text normalization and intonation in text-to-speech synthesis of Greek", VERBA 90, (ALCATEL FACE), Italy.

[5] D. Klatt: "VOT, Frication and Aspiration in word-Initial Consonant Clusters", Journal of Speech and hearing research, pp. 686-705, 1975.

[6] C. Darwin et al: "What tells us when voicing has started?", Speech Communication 1, pp. 29-44, 1982.