SOME IMPORTANT CUES ON IMPROVING THE QUALITY OF A TTS SYSTEM

N. Yiourgalis, G. Epitropakis, G. Kokkinakis

Wire Communications Laboratory
University of Patras

ABSTRACT - This paper presents the experience acquired by improving the quality of a rule-based parametric formant Greek TTS system developed in our laboratory. Improvements are achieved by :
i. the addition of rules which control the duration, concatenation and coarticulation behaviour of the appropriate speech segments that are to be abutted. The intended definition/firing of these rules was ensured by the use of a specially designed graphical environment,
ii. the introduction of pitch synchronous excitation which removed the appearing spikes in the speech output due to filter instabilities and
iii. the application of an intonation scheme based on the syntactic analysis results of the input text, in order to resemble the intonational boundaries and the important prosodic features of natural speech.

INTRODUCTION.

The style of speech produced by modern Text-To-Speech systems is far from natural. Although several commercial systems exist, much of the interesting work in the field, is still in the research phase. In this paper, the effort experienced in improving the speech quality of a rule-based parametric formant Greek synthesizer, is presented (Yiourgalis, 1991). The improvements were introduced gradually into two phases :
First, by eliminating the spurious unwanted clicks presented at the output. To this end, pitch synchronous updating of the frame parameters was chosen instead of the fixed time (5ms or 10ms) commonly used, ensuring that the glottal filters are always stable. Also the speech segments corresponding to the text, are concatenated smoothly under all circumstances. Finally, coarticulation phenomena are treated by rules, that take care of all the peculiarities in segment boundaries which exhibit a rapid spectral movement under specific phonemic context.
Second, by resembling the rhythm of the synthetic voice as naturally as possible. For this purpose, the duration of each segment was controlled by context sensitive rules extracted from analysis of natural speech. Last but not least, a new intonation scheme that is completely based on syntactic analysis of the text has been integrated in the TTS system.

All improvements were practically integrated by means of a graphical environment tool built for this purpose. The user of this tool can watch the duration of each segment and the formant , pitch, intensity and energy contours, all aligned with a labelled time domain presentation of the synthetic speech. This tool has recently revealed to us, that the Rosenberg excitation source that is currently used by our TTS system has to be replaced by a more flexible parametric source, in order to achieve the high quality synthetic output which we aim at.

PITCH SYNCHRONOUS EXCITATION (PSE)

In parametric synthesizers the vocal tract is usually simulated by a number of cascaded/parallel resonators (Klatt, 1980). It is these second order filters that cause clicks, due to the possible

instability situations that they may fall into when excited frame synchronously (Boves, 1987).

The resonator case.

The equation of a 2nd order resonator is $y_n = ax_n + by_{n-1} + cy_{n-2}$ where a,b,c are skilled coefficients that are nonlinearly related to the centre frequency (formant) and bandwidth of the filter, x_n is the system's input, y_{n-1}, y_{n-2} is its memory and y_n is the output voice sample. Updating b and c when y_{n-1} and y_{n-2} have a large value, will cause instability, as in the case of fixed frame synchronous parameters updating. Large changes in b and c coefficients co-occurring with high y_{n-1} and y_{n-2} values do not matter, **only** if the latter are low, as in the case of pitch synchronous parameters updating, where the filter experiences its decaying phase. PSE also allows large movements in the formant contours, in the case of the second formant of segment /mi/ for example, without exhibiting any disturbing click in the output.

The anti-resonator case.

The equation of a 2nd order anti-resonator is $y_n = a_1xn + b_1x_{n-1} + c_1x_{n-2}$. In this case inconsistency between a_1, b_1, c_1 coefficients and the output cannot exist, since there is not memory in the system. In case of instability the filter stabilizes after at most two samples. Thus, fast changes in the coefficients of an anti-resonator are permitted (nasal+vowel transition).

Based on the above observations, the implementation of the pitch synchronous update scheme was integrated into our Greek TTS. Improvement in the overall quality was realised immediately not only by the elimination of the clicks, but also by the individual improvements experienced in some segments that needed a more abruptly changing formant contour in their coding representation. Some important points that were spotted during implementation of the pitch synchronous updating feature, are discussed below :

- The excitation pulse must be completed at the end of a segment, even if its duration has been completed and does not coincide with the end of the pulse (Fig. 1). However the extra samples are counted in the duration of the next segment so that the oveall duration of the text is not exceeded.
- In case of a Voiced-Unvoiced concatenation, there is a time of zero excitation at the end of the first segment, when the final pitch exceeds its duration (Fig. 2).

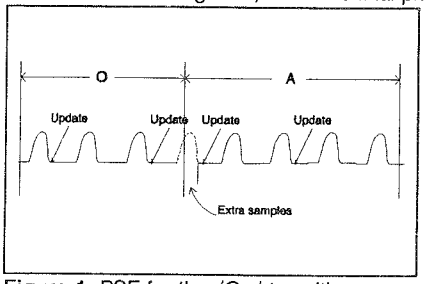

Figure 1. PSE for the /Oa/ transition.

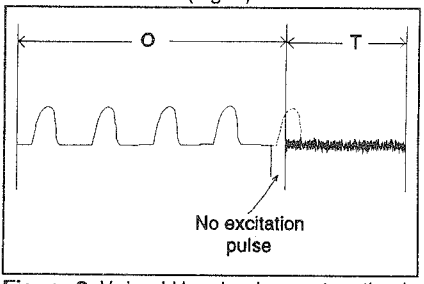

Figure 2. Voiced-Unvoiced concatenation in PSE environment

CONCATENATION SCHEME.

All efforts to string together chunks of speech, coded or not, no matter their size, have failed because of the well known coarticulatory effects between adjacent phonemes. These effects cause substantial changes to the acoustic manifestations of the phoneme, depending on the context. Since coarticulatory influences tend to be minimal at the acoustic centre of a phoneme, we have captured the Consonant-Vowel transition in our synthesis strategy, in the definition of the "segment" unit. A thesaurus of 134 formant coded speech segments of the types Consonant(C), Vowel(V), CV and CCV has been built (Yiourgalis, 1987) . This leaves the remaining VC transitions to be spectrally smoothed during synthesis by means of a sophisticated "concatenation" algorithm. Due to the segment approach of our synthesis, there are seven cases that define VC or VV boundaries. Namely they are :
V-V, V-Semivowel, V-CV, CV-V, CV-Semivowel, Semivowel-CV and the CV-CV boundary.
Cases 1 and 2. (Fig. 3).

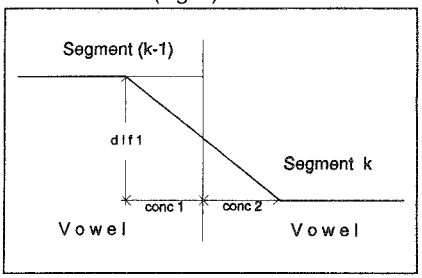

Figure 3. Vowel-Vowel transition.

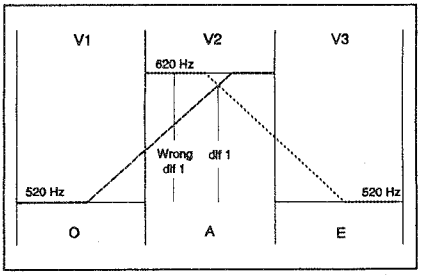

Figure 4. Special case in V-V concatenation.

The two contours are smoothed by linear interpolation over the range conc1+conc2 that is dynamically changed according to dif1, the contours difference at the boundary.

Special Case.

In cases where there is a $V_1V_2V_3$ sequence such as in the word "ο αετός" (/O aet'Os/ the eagle), V_2V_3 needs concatenation. But V_2 has been already changed due to the V_1V_2 smoothing procedure that has been preceded. In this case dif1 is defined as in Figure 4.

Cases 3, 6, and 7 (Fig. 5)

The algorithm in this situation is two fold, since only linear interpolation would destroy the Consonant contour that is very important for the quality of the CV segment. Thus linear interpolation is followed from F_a to F_c, followed by a complex filter (equation 1) smooths the contour between F_c and F_b over the conc2 range.

$$F_{fi} = \frac{F_1[l] * phonw1 * aveder[0] + F_{fi}[l] * phonw2 * aveder[1]}{phonw1 * aveder[0] + phonw2 * aveder[1]} \qquad (1)$$

where : $F_{fi}[l]$ is the final formant value, for the l sample ($0 < l < conc2$),
phonw1 = conc1-l and phonw2 = $(l+1)^{1.5}$ are the corresponding formant weighting factors,
aveder[0] is the spectral derivative of the first segment and

aveder[1] that of the second.

Cases 4 and 5. (Fig 6)

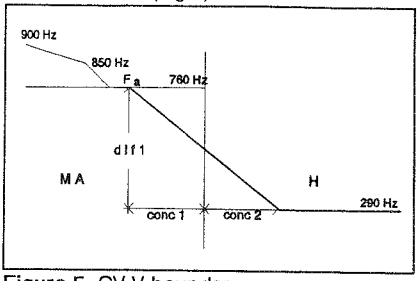

Figure 5. CV-V boundary.

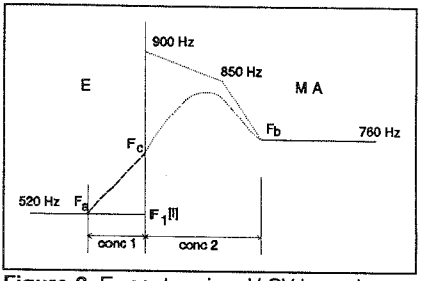

Figure 6. F_1 contour in a V-CV boundary.

Linear interpolation over the conc1 + conc2 range that is also dynamically adapted according to the dif1 value, is followed. F_a is estimated at the first zero of the spectral derivative of the first segment.

COARTICULATION RULES.

Even though the PSE technique and the spectral smoothing algorithm were implemented in our TTS system, there were still transitions that produced disturbing spectral discontinuities in the synthetic output. Such cases were confined to situations involving very large differences in the target values stored in the thesaurus for the first, second, third formant, the nasal pole and the nasal zero of the segments under concatenation. These differences could not be smoothed by the concatenation algorithm since they were very large. For example, the second formant transition in "ιου" (/iu/) starts from 2260 Hz and ends at 890 Hz with a difference of 1370 Hz. Fifteen rules have been written to deal with these context dependent coarticulation effects. The rules of the type : **segm1 (before/after) segm2 ——> change (parameter, + value, segment)** when fired, change the target value of the specific parameter of the involved segment.

DURATION RULES.

Segmental duration is one of the cues that in many cases determine features of the neighbouring segments (Klatt, 1976). It also influences the average speaking rate and determines the emphasis or contrastive stress locations. In our TTS, we have developed about fifty context sensitive rules, that adjust the normalised (average) durations of every segment in such a way that the sentence has the rhythm of the naturally spoken. Our model assumes that each segment has an inherent duration obtained from the thesaurus. This duration is the average one measured from the analysis of 200 recorded words resembling good phonemic coverage of the Greek language. Every fired rule, tries to effect an increase or decrease on this duration.

The rules were created by a time-consuming method that is described below : The 200 recorded words were manually segmented and labelled using the SAMBA software. The average duration of every segment in the thesaurus together with the duration under any context were coded into a data base. As contexts we have decided on the following categories : Stressed Vowels,

Unstressed Vowels, Nasals, Liquids, Glides, Voiced Fricatives, Unvoiced Fricatives, Voiced Stops and Unvoiced Stops. Finally, a menu driven tool was built that assisted the developer to retrieve the informations from the data base in a graphical presentation (Fig. 7, 8). After careful inspection the graphical informations were turned into the rules that were integrated into the system.

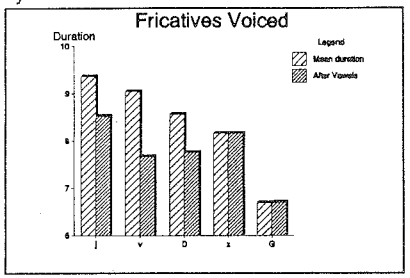

Figure 8. Voiced Fricatives before Vowels.

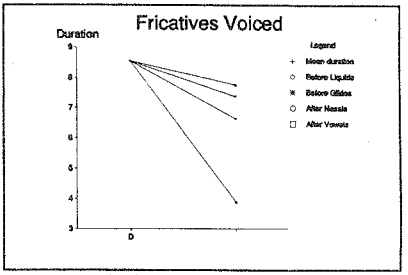

Figure 7. /D/ under specific context.

INTONATION SCHEME

The relationship between syntactic and intonational structure is well documented in the linguistic literature. Allen (1985) points out that conventional parsing programs do not readily accommodate the needs of TTS systems for fast analysis into structure constituents. One the one hand, traditional parsers do not operate in real time; on the other, much of their effort ie concentrated in deciding between alternative hierarchical relationships between phrase-level units, to yield interpretative analyses which are of limited usefulness to current TTS systems. In the long term, such information will enhance the naturalness of audible output, when the relationship between the syntactic and prosodic levels is fully understood.

Following the above observations, a syntactic parser based on a fast accessed morph-type structured lexicon and a word class assignment algorithm, groups the sentence into phrases, estimating in this way the place and the duration of the important intonational boundaries (pauses), that have to be "spoken" during synthesis. The parser deals with unrestricted input at real time. A syntax-to-Pitch algorithm (Epitropakis, 1992) realises a pitch contour that greatly resembles that of natural speech.

GRAPHICAL ENVIRONMENT

The correct definition and firing of all experiences discussed above, was ensured by the use of a specially designed graphical environment tool (Fig. 10). The user of this tool can watch the absolute duration of each segment by means of cutting a pasting through a moving pointer (Fig.10d). Pitch synchronous updating (Fig. 10a) the first three formant contours (Fig. 10b) and the intonation contours (Fig. 10c), all aligned with a labelled time domain representation of the synthetic speech can be watched any time after synthesis. It is this tool that recently revealed to as problems referring to our voiced excitation source. Its closing time should be about 10 samples, otherwise the high pitched portions of speech (stressed points) presented unwanted clicks in the output.

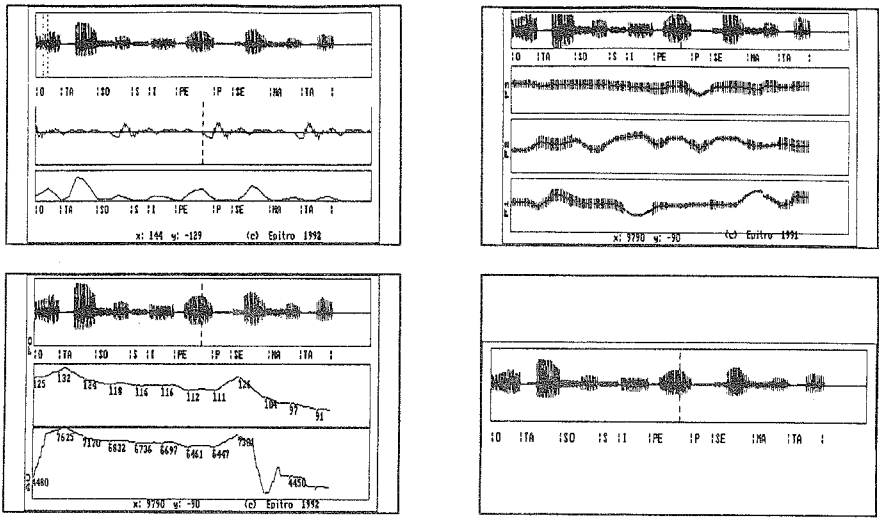

Figure 10. a. Pitch synchronous b. Concatenation and coarticulation rules
c. Intonation d. Duration

CONCLUSIONS

Improvements in the quality of a rule-based formant TTS for the Greek language have been achieved through a sequence of changes and addition of rules, that were ensured by means of a graphical environment tool. A better understanding of semantic and pragmatic factors would improve our ability to model intonational phenomena. Until it becomes feasible to derive such information from the input text, as human readers appear to do, progress in quality of the synthetic speech will be limited.

REFERENCES

Allen. J. (1985), *Speech Synthesis from text*, Computer Speech processing, Prentice Hall.

Boves, L. & Kerkohoff, J. & Loman, H. (1987), *A new synthesis model for an allophone-based TTS system,* Proceedings of 1st European Speech Technology Conference, 2, 385-388

Epitropakis, G. et all (1992), *Prosody assignment to TTS systems,* SST'92

Klatt, D. H. (1976), *Linguistic uses of segmental duration in English : acoustic and perceptual evidence,* J. Acoust. Soc. Am. 59, 1208-1221

Klatt, D. H. (1980), *Software of Cascade/Parallel Synthesizer,* J. Acoust. Soc. Am. 67, 971-995

Yiourgalis, N. et all (1987), *High quality and reduced memory TTS synthesizer for the Greek language*, Proceedings of 1st European Speech Technology Conference, 1, 187-190

Yiourgalis, N. & Kokkinakis, G. (1991) *A TTS system for the Greek,* ICASSP'91, 525-529