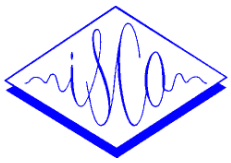


The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

CONTRIBUTION OF THE ANALYSIS OF PUNCTUATION TO IMPROVING THE PROSODY OF SPEECH SYNTHESIS

Isabelle GUAITELLA & Serge SANTI
Institut de Phonétique d'Aix-en-Provence, Université de Provence 1
29, Av. R. Schuman, 13621 Aix-en-Provence, FRANCE

ABSTRACT
Generating speech synthesis that would sound more natural is necessary. An original perception test, in which subjects listened to read and spontaneous speech and were asked to state how each should be ponctuated, leaded us to built up prosodic rules based on the punctuation of the to-be-synthesized text. The advantages of such a system are examined. These basic rules and algorithms that have been applied to speech synthesis are described.

1- INTRODUCTION
Descriptions of most text-to-speech synthesis systems do not deal much with the analysis of punctuation. As far as we know, except from the procedure developed by Choppy and Liénard (1977,1979), only its role in syntactic segmentation is considered, where it serves as the basis for generating the prosody of the to-be-synthesized text. This approach appears to be sufficient to give an account of the prosodic style of read speech, where punctuational and syntactic segmentation is relatively fixed, however, it is not sufficient for generating speech that is closer to spontaneous style (for applications such as remote inquiry of vocal databases, conversational systems, etc.). Nevertheless, Choppy and Liénard have shown that it is not necessary to go through a syntactic analysis to obtain a coherent result.

Such a synthesizer should be able to give a "reading" or a "natural" version of a single written text according to the user's decision, the supposed differences between them appearing at the segmental and suprasegmental levels. The prosodic output of the two versions should be generated by rules, based on the signs and places of punctuation. The main advantages of such a system is to avoid to go through an upper level analysis (syntactic, semantic,...) and to do without extra graphic markers in the text. From a theorical point of view, segmental and supra-segmental levels should not be dissociated, however, we thought it useful to start applying the prosodic rules obtained from our punctuation test.

2- A PUNCTUATION TEST
2-1- Hypothesis
We consider that punctuation is the conventional process used to note intonation and rhythm. Our hypothesis is that if we ask listeners to write down the punctuation coresponding to spontaneous or read oral texts, the discordances between reading and spontaneous prosodic systems will appear through the use of the punctuation code. According to us, the advantage of such an experiment is to apprehend the "direct impression" of the listeners with regard to possible prosodic differences, without any phonetic or acoustic analysis.

2-2- Main results
The main results in our experience only will be mentioned here. Readers willing to have further information about our methodology and more detailled analysis of our results can refer to GUAITELLA, SANTI, CAVE 1990 and GUAITELLA, SANTI 1990.

For a given textual content, the location of punctuation signs differs from reading to spontaneous speech.This would confirm that listeners have not punctuated the texts according to syntax but according to the specificities of each kind of speech.

Silent pauses and the amplitude of the melodic contour (> 30 Hz for a single syllable) have been considered to be the determining factors in the punctuation setting.

The following figures show the distribution (in percentage) of the places of punctuation (grey) and the number of puctuation signs (black) for each combination of "pause" and "amplitude" parameters for both read (Figure 1) and spontaneous speech (Figure 2).

The analysis of the number of punctuation signs of each place shows the low importance of the amplitude parameter alone and the marginality of places with no pause nor amplitude. Exept cases where pause and amplitude are associated, in spontaneous speech, the amplitude determines the setting of punctuation signs.This phenomenon can be interpreted by asserting that, in spontaneous speech, the silent pause plays a role close to the one played by hesitation, that is to say not perceived as an intentional act of text structuration. In an other hand, in reading, pause is pre-programated and decoded as something voluntary and significant. In spontaneous speech, the absence of this structuration cue may be made up by the amplitude of the F0 variation. As a

consequence, subjects should adapt themselves to the pragmatic constraints of each kind of speech situation and interpret the prosodic phenomena with regard to these constraints.

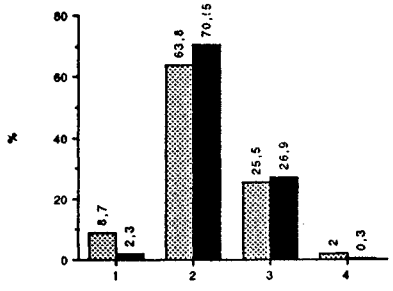


figure 1

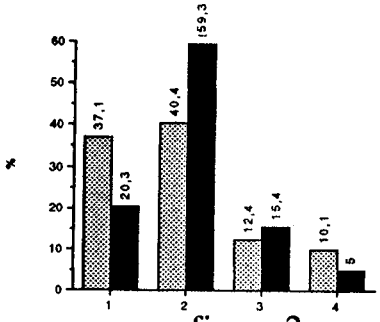


figure 2

(1 - amplitude, 2 - pause + amplitude, 3 - pause, 4 - no pause nor amplitude)

This analysis is used as a background for the elaboration of our prosodic rules which are supposed to generate an expressive reading or a more natural speech style. Our task is now to start from punctuation to obtain the right prosodic configurations of the two kinds of speech.

3- AN APPLICATION TO SYNTHESIS

3-1- Méthodology

3-1-1- corpus

Our corpus is constituted by two short extracts of texts, chosen according to the following criteria:

One of these texts comes from a graphic transcription of spontaneous speech and has been punctuated in a normative way. The other one has been taken from a written text and has also been ponctuated according to standard punctuation. Unsimilar origins have been chosen in order to show that any type of text (syntacticaly coherent) can be synthesized according to the following styles of speech, "reading" and "natural".

3-1-2- synthesizer

Our modelisation of prosody has to be adaptable on any synthesizer able to modify the F0 as a function of time. The synthesizer used in these first tries is a natural speech coder , the PSOLA system (ESPESSER 1987, ROUCOS & WILGUS 1985), and a F0 modelisation program MOMEL elaborated in Aix-en-Provence by D. HIRST and R. ESPESSER. This tool is able to modelize any kind of melodic curve in placing, with the help of a graphic input, a series of points (time -> X-axis and F0 ->Y-axis) automaticaly linked together by a curve (spline function).

The main advantage of such a system is its ability to modify the temporal evolution of the F0 at our convenience without modifying the segmental level and avoiding, a priori, problems of intelligibility. Pauses have been included manualy with the help of the signal editor used previously for the punctuation test. All of our programs, speech acquisition and treatments were running on a MASSCOMP 5400 mini-computer.

3-2- Prosodic rules

Our rules use two kinds of parameters : the presence/absence of a silent pause and the temporal evolution of the fundamental frequency

The application of the rules and of the conditions has to be preceeded by a syllabic segmentation. In a text to speech synthesizer, this segmentation will have to be automatic, the difficulty of this task being directely dependant on the type of synthesizer and on the chosen synthesis method.

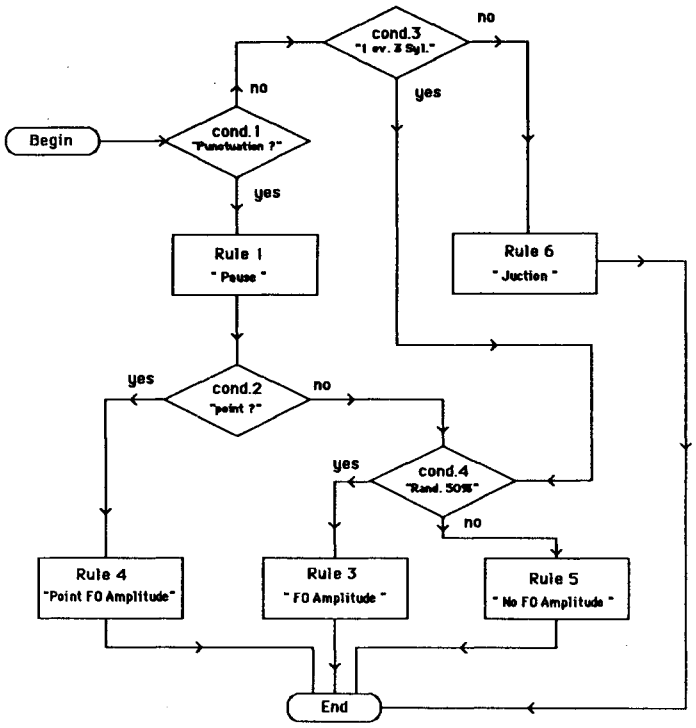
The rules and conditions described below are applied according to algorithm 1 (for the "read style") and algorithm 2 (for the "natural style").

RULE 1: Apply a silent pause (P) with duration (x) for the following punctuation signs (symbol "•" is used for " end of paragraph"): [,] --->P(x1) with x1 = 300 ms; [...], [;], [(], [)] --->P(x2) with x2 = 400 ms; [.], [!], [?] --->P(x3) with x3 = 500 ms; [•] --->P(x4) with x4 = 600 ms

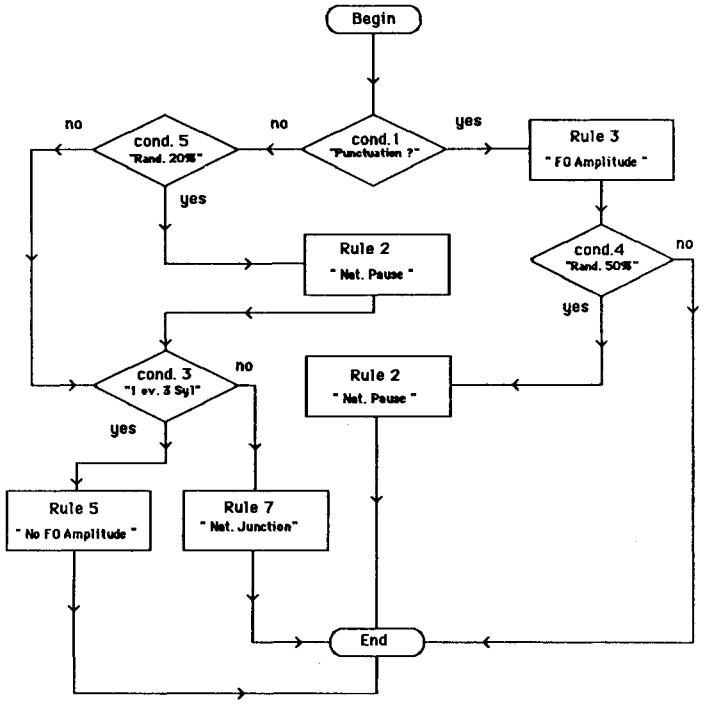
RULE 2: Apply a silent pause (P) with duration (x); P(x1), P(x2), P(x3), P(x4) being determined at random

RULE 3: Apply a melodic contour (A) with amplitude (y) with y1 = 30 Hz, y2 = 50 Hz, y3 = 70 Hz, y4 = 90 Hz; A(y1), A(y2), A(y3), A(y4) being determined at random; and, apply a falling (F) or rising (R) contour, F or R being determined at random

RULE 4:Apply a melodic contour (A) with amplitude (y) with y1 = 30 Hz, y2 = 50 Hz, y3 = 70 Hz, y4 = 90 Hz; A(y1), A(y2), A(y3), A(y4) being determined at random; and, apply a falling contour (F)



Algorithm 1: "reading version" algorithm, for each segmented syllable



Algorithm 2: "natural version" algorithm, for each segmented syllable

RULE 5: Apply a melodic contour (N) with amplitude (z) with z1 = 10 Hz, z2 = 15 Hz, z3 = 20 Hz, z4 = 25 Hz; N(z1), N(z2), N(z3), N(z4) being determined at random; and, apply a falling (F) or rising (R) contour, F or R being determined at random

RULE 6 and **RULE 7**: Used to determine the F0 trajectories between the contours generated by the rule 3, 4 and 5, these rules are totaly independant on punctuation and will be described later.

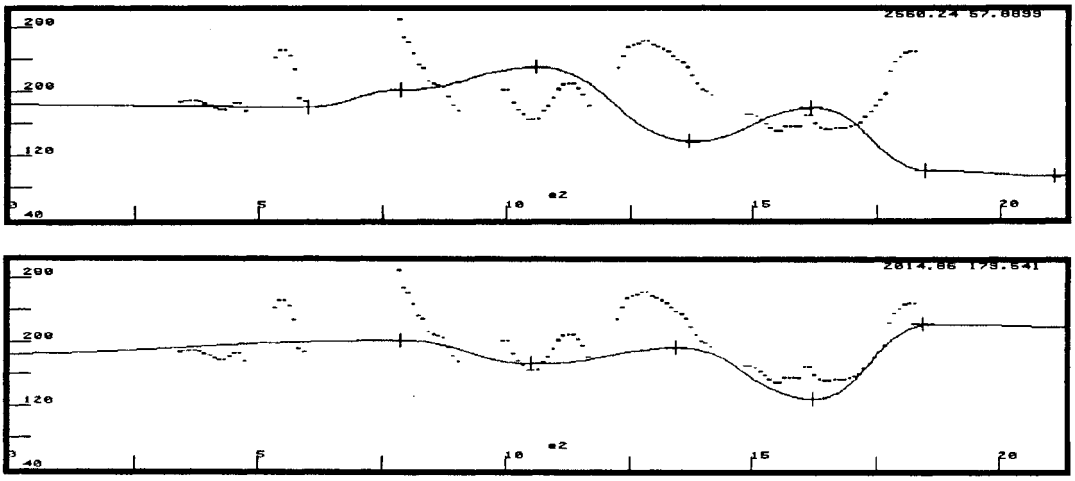
COND 1: Is this syllable followed by a punctuation sign ?

COND 2: Is this syllable followed by a point ?

COND 3: Is this syllable the third syllable since the begining of the text or since the last syllable that has satisfied to this condition ?

COND 4: Random choice with 50% of chances for the answer to be "yes"

COND 5: Random choice with 20% of chances for the answer to be "yes"



Example of modelisation : "reading" (up) and "natural" (down) versions of the segment "L'objectif de ces deuxièmes journées".(dotted line: F0 of natural input segment)

4- CONCLUSION

Prosodic rules alone are not supposed to account for the overall variations existing between read and spontaneous speech. Other phenomenons as the repetition of certain sequences, hesitations, rate variations, segmental duration modifications and some coarticulation phenomena will have to be taken into account for delivering a natural sounding synthetic speech closest to spontaneous speech. We have tried, above all, to show the validity of such a process, making a more natural vocal synthesis is possible and necessary. This work is a first step in reaching this goal and these first results seem to be encouraging. Moreover, such an approach, taking into account the specificities of spontaneous speech, may help us to better understand the linguistic behavior of the user in man-machine communication.

REFERENCES

CHOPPY C., 1979, "La ponctuation, indicateur prosodique pour la synthèse à partir du texte, étude de la virgule", Actes des 10e J.E.P., Grenoble, 185-191

CHOPPY C., LIENARD J. S., 1977, "Prosodie automatique pour la synthèse par diphonèmes", Actes des 8e J.E.P. , Aix-en-Provence, 211-217

ESPESSER R., 1987, "De la précision d'une méthode de variation du débit de parole", Actes des 16e J.E.P., Hammamet, 22-24

FONAGY I., FONAGY J., 1983, "L'intonation et l'organisation du discours", B.S.L. Paris, tome 78, fasc. 1, Kliencksieck, 161-209

GUAITELLA I., 1990, "Propositions pour une méthode d'analyse de l'intonation en parole spontanée", Actes du premier congrès d'acoustique", Lyon, Les Editions de Physique, 515-518

GUAITELLA I., SANTI S., 1990, "Ponctuation et organisation rythmique de l'oral", Proceedings of LP'90 Conference, Prague (à paraître)

GUAITELLA I., SANTI S., CAVE C., 1990, "Relations ponctuation/prosodie en lecture et en parole spontanée, Actes des 18e J.E.P., Montréal

MARTIN P., 1982, "Phonetic realisations of prosodic contours in french, Speech Communication 1, North-Holland Publishing Company, 283-294

ROUCOS S., WILGUS A. M., 1985, "High quality time-scale modification for speech", Proceedings of ICASSP, 493-496