Generation and evaluation of rhythmic patterns for text-to-speech synthesis

P. Barbosa & G. Bailly
Institut de la Communication Parlée, U.R.A. CNRS n° 368
ENSERG/INPG -Université Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex 1, France

ABSTRACT

This paper presents a characterization of durational contours based on phasing relations between noticeable acoustic events and an internal clock. We generate segmental durations in two stages : the duration of a rhythmic programming unit is computed according to a reference clock and then is distributed among its segmental constituents. A perception experiment evaluates the necessity of the rhythmic patterns found in analysis. A method for mastering speaker's speech rate is described and is analysed to propose some guidelines for the integration of the pause phenomenon into automatic generation.

INTRODUCTION

"(...)intonation manages *to do what it does* by continuing *to be what it is* (...)" (Bolinger, 89). That is our point of view about prosody: it performs a linguistic task under biological constraints. Perception is guided by motor schemes: listeners use information from kinematic patterns (Viviani & Stucchi, 92).

We assume the existence of an underlying internal clock, a timekeeping function, used for synchronization of impulses transmitted to the muscles (Turvey *et al.*, 90). We show that the regularity of this clock is maintained through pauses (cf. section 3).

We generate segmental duration by a two-stage model (Campbell, 91) but our approach is different from Campbell's because duration is obtained by a control signal emitted at each Perceptual-center (PC).

1. THE INTER-PERCEPTUAL-CENTER GROUP

A two-rate, 88-sentence corpus was explored in order to study the rhythmic patterns of read sentences. It was designed for answering at: (1) are continuously increasing patterns (cf. fig. 1) needed to the perception of accentuation or are they just an artefact of production constraints ? (2) Are this typical configuration needed to the perception of any kind of isochrony in French connected speech ?

Pompino-Marschall's experiments have tried to estimate an absolute localization for PC using *syllable/beat* and *beat/syllable* sequences. Despite the diversity of the consonants the PC seems to be at the neighbourhood of the vocalic onset. The perception of momentary tempo is better characterized by inter-PC intervals.

Thus the PC location in our work is fixed at the vocalic onset. The importance of this event is largely developed in the literature (Dogil & Braun, 88; Stevens & Blumstein, 78; Fant & Kruckenberg, 89). The term PC will be maintained because of the allusion on perception and our hypothesis of an internal clock guiding the production of the programming rhythmic units. The lenghtening of IPCGs is characterized by a single factor k, by computing $\sum \exp(\mu_i + k.\sigma_i) =$ IPCG duration, where μ_i and σ_i are the mean and standard deviation of the log-transformed durations (in milliseconds) of the realizations of the phoneme i from a corpus of logatoms at comfortable rate. Rhythmic patterns of the corpora are characterized by k averaged for each IPCG and segmental durations are in turn computed for synthesis by the exponential expression above.

The analysis of the corpus evidences a rhythmic pattern by concatenation of elementary movements (cf fig. 1). These movements are monotonously increasing, mark clearly prosodic boundaries, start with a reset of k at 0 and exhibit a more or less exponential increase.

2. THE PERCEPTION EXPERIMENT

Method

Ten pairs of sentences were used for this experiment. They were listened in binoral presentation by eleven subjects working in the laboratory but not in synthesis domain. The duration of the test was between 10 and 15 minutes. Each pair contains a reference durational pattern (A) and a pattern to be tested (B). The two sequences (AB and BA) were listened in random order within a session. Listeners were asked to answer what sentence was the more natural by pressing on the keyboard "1" to the first one, "2" to the second one or "?" if a doubt persists. Listening may be repeated twice. An example of a sentence pair is showed below.

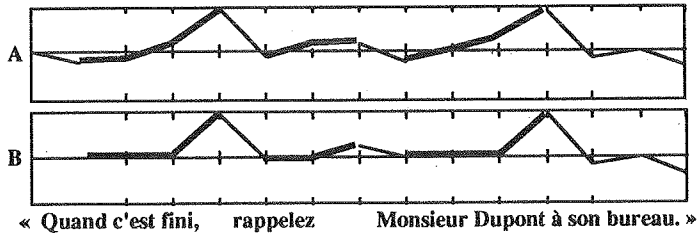

Figure 1. *Sentence 56 : Example of A and B configurations. Discrete values of k are connected by lines for sake of visibility. Prosodic groups are represented by thicker lines. The last IPCG of the utterance is not presented here.*

The A configuration was obtained by calculating the successive k-factors of all IPCGs in a sentence. In the B configuration k-factors preceding the accent were set to 0 and the ones associated with the accent were not modified. Segmental durations in the two configurations were computed by applying the formula mentioned above. Segmental durations in A patterns are different from the ones in the natural sentences (k is averaged in each IPCG) but since the IPCG durations are the same, the VO timing is identical in both ones. Silent pause durations and all other parameters were unchanged. A high quality speech analysis/resynthesis system was used to obtain the above parameters (Moulines, 92). We are testing the perceptual prominence of a gradual versus an abrupt pattern of accent realization.

Results

Considering all subjects 77% of A answers are obtained. Taking also the question-mark answers the result is 65% of A, 20% of B and 15% of question-mark answers.

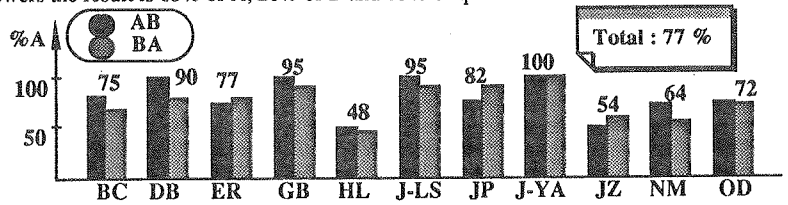

Figure 2. *Mean scores of A answers by subject and mean over all subjects*

There were no significant effect of presentation order. All listeners agree on the difficulty of the task : "During the experiment I thought I have changed the criterion of utterance choice. At the beginning, I have chosen the ones were more *constant* concerning

the rhythm" (JLS), "I have chosen utterances that were the most *constant* at rhythm level." (OD), "Are you sure they are not identical ?" (JYA) !!

Results show subjects' preference for the gradual accent realization, despite the finesse of the task. The term *constant* was used to oppose the two stimuli.

Discussion

The clear general preference for the gradual pattern lead us to think that this configuration is necessary to the accent perception. Too small differences between A and B configurations explain the poor score of one of the utterances.

Lack of lengthening of previous IPCGs sounds abrupt. The internal clock hypothesis may explain the subjects' perceptual behaviour : shorter IPCIs are cues of an unexpected local acceleration. Gradual lenghtening contributes to the perception of isosyllabicity (Duez, 87 ; Lehiste, 77). But there is no implicit conclusion that human beings use k coefficients to produce the accent pattern.

3. TOWARD A MODEL INCLUDING PAUSE EMERGENCE

About the clock beats

Grosjean's performance structures are built using pause duration as a cue of the strength of corresponding prosodic juncture (markers here). Grosjean's approach do not take into account an underlying rhythmic activity which constrains pause durations to be realized by an integer number of clock units (Fant & Kruckenberg, 89): this is illustrated by the relatively low 86% correlation between cues obtained at two different speaking rates in their experiment (Monnin & Grosjean, in press).

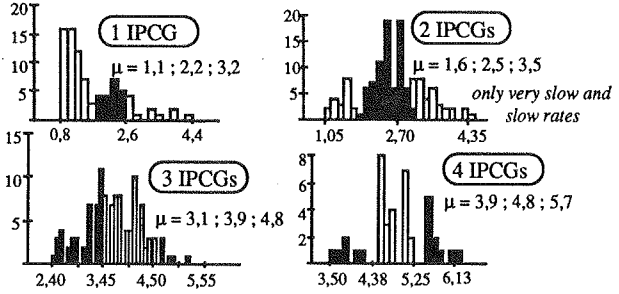

Figure 3. *Histograms and clusters by number of IPCGs in the PG*

A five-rate, 20-sentence corpus was recorded, pronounced by our speaker in order to evidence the influence of pause insertion on the rhythmic structure. This experiment aims at developing an automatic generation of duration including the pause phenomenon. The speaker was asked to answer to interrogative synthetic utterances with predefined sentences. These utterances were obtained using our text-to-speech system by multiplying μ_i and.σ_i by a phonation factor (Wightman *et al.*, 92) .

The analysis of this corpus confirms for all rates the general trend of rhythmic patterns: 80% of them are monotonously increasing. To study the pause emergence a characterization of total duration of PGs have been developed: the ratio between the total duration of each PG and the internal clock duration is computed.

The choice of an internal clock duration is particularly delicate, but necessary. How to choose an unaccented unit if lengthening is gradual over the PG ? Taking into account that: (1) the last IPCG in the prosodic group is clearly the main lengthened unit; (2) if there is a pause, we cannot separate silent and sound intervals : they are elements of the same phenomenon (Duez, 87); (3) the first IPCG of a PG is often shortened, the internal clock durations are computed for each utterance as the mean among the non-accentuated IPCGs. We assume that there is a reseting of the internal clock after the accent realization.

These ratios were submited to a cluster analysis for each PG length (cf. fig.3). Clusters are similar between the rates. The analysis was made over the five rates (in order to have statistically significant data).

Results

Mean values of the clusters are closely associated with integer number of clocks. The standard deviations represent between 10 to 20% of the respective means.

We then observed each cluster for each individual rate : (1) the additional clock units are associated with the presence of silent pauses in slow rates and only lengthening in the fastest rate (except for the strongest markers); (2) when there is a silent pause, lengthening of the preceding group is represented by an integer number of clock units; (3) the strongest markers are associated with the greatest values of the ratios.

4. COMMENTS AND PERSPECTIVES

Some deviations in the cluster sets may be due to wrong segmentations (several unvoiced plosives after silent pauses) and the choice of the clock unit.

Automatic generation of duration will depend on the strength of the prosodic marker associated with the juncture : strongest markers will receive more clock units. The main differences between the speech rates are : (1) the frequency of the internal clock clearly differenciates the rates (except very slow and slow rates) ; (2) in slow rates the subject seems to prefer lengthening plus silent pause to realize the accent whereas in fast rates, only lengthening ; (3) markers can be removed in fast rates to form a PG that will contain more IPCGs (marker deletion).

In this perspective, rhythmic patterns are monotonous decelerations which modulate in frequency a carrier clock.

REFERENCES

Barbosa, P. & Bailly, G. (1992) "Generating segmental duration by P-centers", *4th Workshop on Rhythm Perception and Production*, Bourges, France, June, 163-168.

Bolinger, D.(1989) *Intonation and its uses*, (Edward Arnold).

Campbell, W. N. & Isard, S. D. (1991) "Segment durations in a syllable frame", *Journal of Phonetics*, **19**, 37-47.

Dogil, G. & Braun, G. (1988) *The PIVOT model of speech parsing* , (Verlag, Wien).

Duez, D. (1987) "Contribution à l'étude de la structuration temporelle de la parole en français", *Thèse d'état.*

Fant, G. & Kruckenberg, A. (1989) "Preliminaries to the study of Swedish prose reading and reading style", *STL-QPSR*, **2**, 1-80.

Lehiste, I. (1977) "Isochrony reconsidered", *Journal of Phonetics*, **5**, 253-263.

Monnin, P. & Grosjean, F. (in press) "Les structures de performance en français : caractérisation et prédiction", *Année Psychologique.*

Moulines, E. (1992) "Synthesis models : a discussion". In : *Talking machines : theories, models and designs* (Bailly, G. & Benoît, C., Eds), 7-12.

Pompino-Marschall, B. (1992) "The P-center and the perception of rhythm in connected speech", *4th Workshop on Rhythm Perception and Production*, Bourges, France, June, 157-162.

Stevens, K. & Blumstein, S. (1978) "Invariant cues for place of articulation in stop consonants", *J. Acoust. Soc. Am.*, **64(5)**, 1358-1368.

Turvey, M.T., Schmidt, R.C. & Rosenblum, L. (1990) "Clock and motor components in absolute coordination of rhythmic movements", *Haskins Laboratories Status Report on Speech Research*, 231-242.

Viviani, P. & Stucchi, N. (1992) "Biological movements look uniform : evidence of motor-perceptual interactions", *Journal of Experimental Psychology : Human Perception and Performance*, **18(3)**, 603-623.

Wightman, C. W., Shattuck-Hufnagel, S. Ostendorf, M. & Price, P. J. (1992) "Segmental durations in the vicinity of prosodic boundaries", *J. Acoust. Soc. Am.*, **91(3)**, 1707-1717.