Esta pasta
pertence ao
Proc. - 21 P - 18417 - 94
de Plínio Almeida
Barbosa

Prédiction de la durée segmentale : le paradigme des groupes inter-«P-Centers»

Barbosa, P. and Bailly, G.

Institut de la Communication Parlée
INPG/ENSERG - Université Stendhal, URA CNRS n°368,
46, av. Félix Viallet, 38031 Grenoble CEDEX FRANCE

RÉSUMÉ

À partir du modèle de génération de la durée segmentale à deux étapes de N. Campbell, nous avons pu montrer précédemment, au moyen d'une étude des corrélations entre les facteurs de déformation (k) entre segments adjacents, la pertinence du groupe inter «P-Center» (GIPC) comme unité de programmation rythmique.

Dans cette présente étude, nous mettons en évidence que les facteurs de déformation associés au GIPC démarquent mieux les frontières entre les mots prosodiques et font apparaître pour ces mots un patron accentuel monotoniquement ascendant.

Un étiquetage de ces patrons nous a permis d'entamer une étude sur la généralité de cette caractéristique et d'envisager une méthode moins dépendante du contenu segmental pour la génération automatique de la durée.

1. INTRODUCTION

Pour générer les paramètres prosodiques classiques (la fréquence fondamentale, la durée et l'intensité), on doit passer d'une représentation généralement phonologique à une réalisation possible au moyen d'un modèle. Ces modèles, statistiques ou stochastiques [Bartkova & Sorin, 87 ; van Santen & Olive, 90 ; Traber, 92 ; Sagisaka, 90], permettent d'associer à chaque phonème un nombre constant de valeurs de paramètres prosodiques et ceci en fonction de leur contexte et de la nature de l'unité phonologique à laquelle ils appartiennent.

En ce qui concerne la fréquence fondamentale, la conception de modèles stochastiques en vue de sa prédiction est facilitée par la relation de la courbe mélodique aux constituants phonématiques : chaque unité est caractérisée par un nombre constant de cibles (1 à 3, typiquement) dont les valeurs sont considérées indépendantes de la nature des constituants (la microprosodie étant modélisée séparément). Pour la durée segmentale on ne peut pas, en principe, concevoir la même démarche car la durée est étroitement dépendante de la nature intrinsèque du segment et de son contexte [Di Cristo, 85]. Notre travail propose un modèle de génération de la durée segmentale qui bénéficie des avantages classiques d'interpolation et de lissage des systèmes à base d'apprentissage. Ceci nous permettra, en outre, de rapprocher les méthodes de génération de la fréquence fondamentale et de la durée de façon à rendre aisée l'obtention intégrée de ces deux paramètres [Todd, 89].

Pour réaliser ce projet, le modèle de Nick Campbell nous a paru intéressant. Il propose la génération de la durée segmentale en deux étapes : la durée de chaque syllabe du message est obtenue à la sortie d'un réseau connexionniste et la durée segmentale, au moyen d'un algorithme de répartition.[Campbell, 92b] Cet algorithme suppose que la durée de la syllabe est la somme de la durée de chacun de ses constituants et que chacun d'entre eux est soumis à une même déformation *k*. La durée de la réalisation phonémique peut être obtenue alors par la formule suivante :

$$\text{durée} = \exp(\mu + k.\sigma)$$

où μ est la moyenne des réalisations du phonème concerné, obtenue à partir d'un corpus de logatomes et σ est leur écart-type.

2. DE LA NATURE DE L'UNITÉ DE PROGRAMMATION RYTHMIQUE

Le *P-center* [Marcus, 76] ou Centre de Perception de l'isochronisme est un point de repère psychoacoustique de la perception de l'isochronisme en parole. Les expériences de Pompino-Marschall [89] semblent montrer qu'au dépit de la variété des consonnes et des voyelles ce *P-center* est toujours au voisinage de l'établissement du noyau vocalique. Cependant il a montré qu'il existe une grande difficulté pour le repérer en parole continue [Pompino-Marschall, 92]. Dans notre travail, le *P-center* coïncide avec l'établissement vocalique.

Dans des articles précédents [Barbosa & Bailly, 92a et 92b], nous avons mis en évidence que le groupe inter *P-center* (GIPC), c'est à dire, l'ensemble des unités phonémiques entre deux *P-centers* adjacents semble être l'unité de programmation rythmique.

Cette conclusion a été possible grâce à une étude de corrélation entre les facteurs de déformation adjacents associés à chaque segment pour le corpus et pour deux débits. Chaque segment, dont la durée réelle était connue, a été étiqueté en *établissement, noyau* et *coda*. Au moyen de la formule précédente et des valeurs des moyennes et écarts-type du corpus de logatomes, on a calculé les facteurs *k*.

Les résultats ainsi obtenus sont montrés ci-dessous, Figure 1.

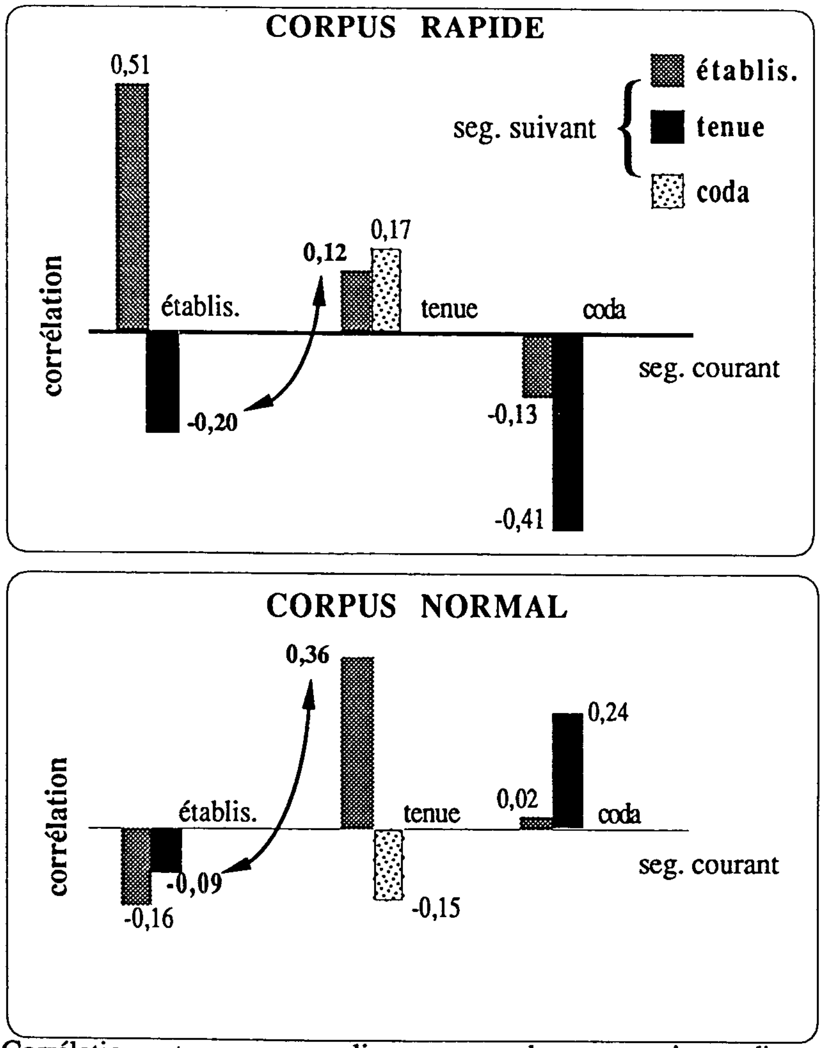


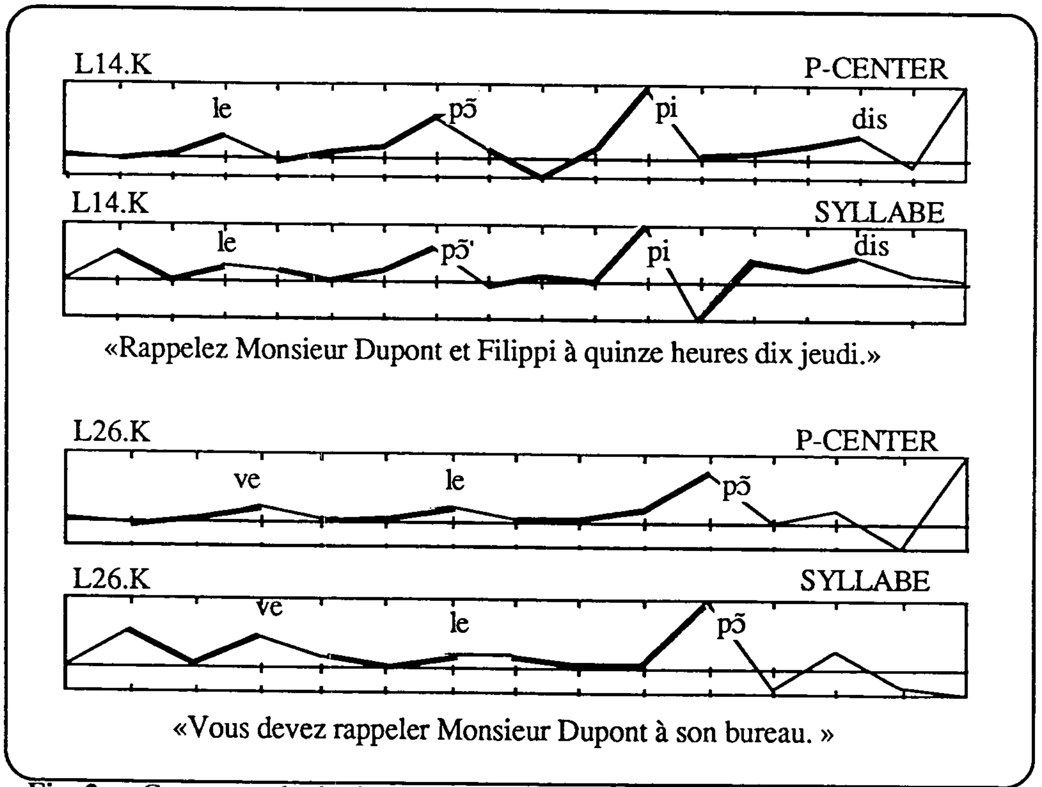
Fig. 1: Corrélation entre segments adjacents pour chaque constituant d'une syllabe.

Les numéros reliés par les flèches sont statistiquement signifiants (p << 0.001) et montrent la complémentarité entre la corrélation positive des tenues courantes et les établissements qui les suivent (bien sûr, dans la syllabe suivante) alors que la corrélation des établissements et des tenues au sein de la même syllabe est négative.

3. LA GÉNÉRATION DE LA DURÉE

Nous générons la durée segmentale en deux étapes, selon la méthode de Campbell. À la différence de celui-ci, c'est la durée du GIPC que l'on présente à la sortie du réseau (à la différence de Campbell et à l'instar de Traber, nous avons adopté un réseau séquentiel qui introduit une contrainte de lissage supplémentaire et une meilleure modélisation des dynamiques gestuelles [Bailly et al, 91]) en phase d'apprentissage. La connaissance de la durée de chaque groupe nous permet de calculer par l'algorithme de répartition le facteur de déformation *k* qui est associé à ses constituants.

En prenant les vraies durées de ces GIPC, on a calculé les facteurs de déformation pour tous les groupes dans toutes les phrases de notre corpus. On l'a fait pour le corpus à débit normal. Deux phrases sont présentées comme exemple, Figure 2.



«Rappelez Monsieur Dupont et Filippi à quinze heures dix jeudi.»

«Vous devez rappeler Monsieur Dupont à son bureau. »

Fig. 2 : «Contours» de durée pour les phrases 14 et 26 du corpus à débit normal.

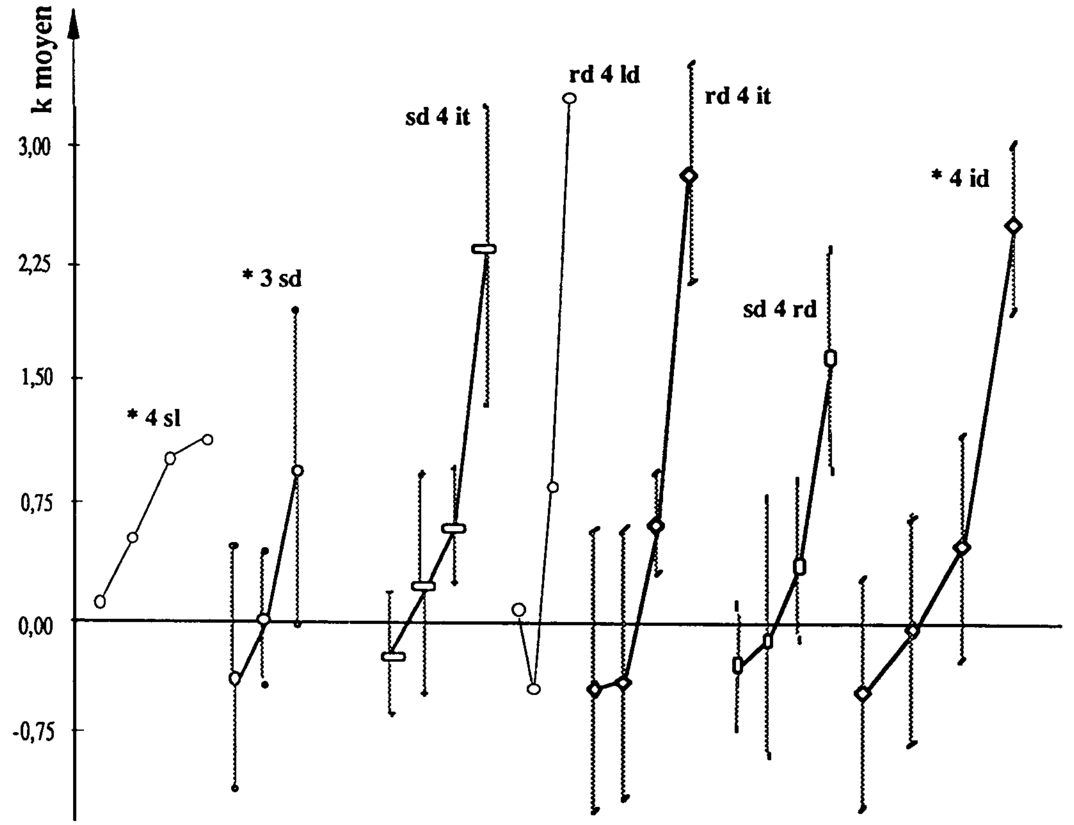
Les valeurs *k* au-dessus de la ligne horizontale sont positives et celles au-dessous, négatives. Ces valeurs discrètes ont été reliées par des lignes pleines. L'ensemble des valeurs rattachées à un mot prosodique sera appelé d'ores et déjà un *contour.*

Pour l'approche PC on voit (1) des contours généralement monotones et ascendants et (2) un marquage clair des frontières prosodiques. Ces phénomènes ne sont pas vérifiés quand la syllabe est l'unité concernée. Les phrases où les syllabes semblent être la bonne unité sont rares.

3.1. Une Base de Données contours

À partir de critères perceptifs (de l'accent), on a découpé manuellement les phrases du corpus à débit normal en mots prosodiques. À une exception près (oubli d'une marque prosodique entre "dès que vous aurez" et "fini."), toute frontière de mot prosodique était associée à une marque, ces marques ayant été positionnées semi-automatiquement sur le corpus dès le départ.

Deux mots prosodiques sont considérés comme ayant le même contour si et seulement si les marques au début et à la fin du groupe sont les mêmes et s'il contient le même nombre de groupes GIPC. Cet étiquetage nous a permis d'obtenir 65 contours-type différents pour les 400 mots prosodiques du corpus. Parmi ces contours-type, une dizaine est statistiquement significative. Nous en présentons quelques exemples dans la Figure 3.



*** 4 sl : Monsieur Dupont** vous rappelera à quinze heures dix.

*** 3 sd : Contactez** le président à son bureau à quinze heures quarante-cinq.

sd 4 it : Rappelez **Monsieur Dupont** jeudi à quinze heures dix.

rd 4 ld : À quinze heures **quarante-cinq** rappelez Monsieur Dupont à son bureau.

rd 4 it : Contactez le président **européen** à son bureau à quinze heures quarante-cinq.

sd 4 rd : Contactez **le Président**-Directeur-Général européen à son bureau
à quinze heures quarante-cinq.

*** 4 id : Monsieur Tapie,** vous devez rappeler Monsieur Dupont à son bureau.

Fig. 3 : Quelques contours moyens du corpus à débit normal. En bas, sont présentés en gras des exemples d'occurrences de ces contours dans les phrases du corpus.

Les configurations monotones et ascendantes de certains de ces contours (43 %) confortent le fait qu'ils peuvent être des indices de marquage prosodique. Pour les autres cas, on peut avancer l'hypothèse que l'accent aurait été effacé. D'autres problèmes sont présentés dans la section suivante.

4. COMMENTAIRES

Les contours de durée présentés nous permettent de lancer l'hypothèse que les durées croissantes sont un indice du «geste» temporel nécessaire pour la réalisation de l'accent [Pasdeloup, 92]. La valeur moyenne du *k* du dernier GIPC (cible de ces patrons croissants) est d'autant plus forte que la marque cible est plus forte. Dans la Figure 4, on peut voir ce comportement. On peut comparer ces histogrammes avec la courbe expérimentale de quantification des durées d'accent proposée par Wightman et al [92].

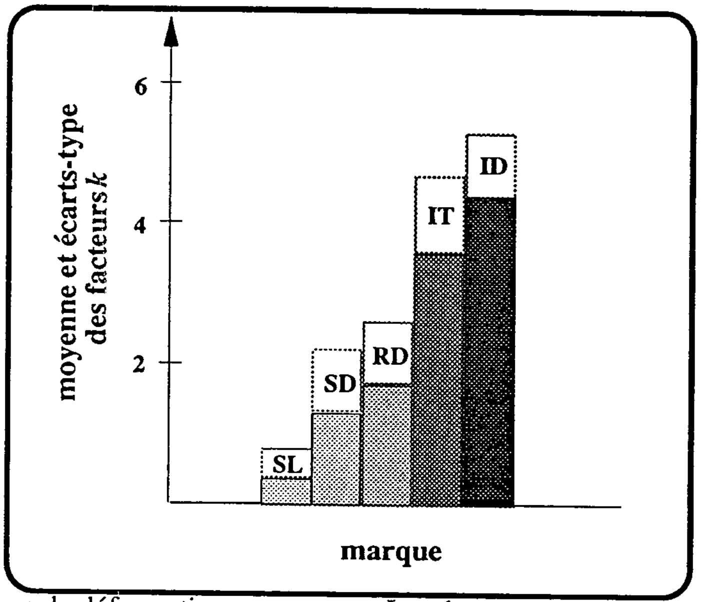


Fig. 4 : Facteurs de déformation par marque. Les écarts-type sont figurés en blanc et superposés aux moyennes.

La force croissante de ces marques est la même que celle des pauses entraînées par ces marques et présente la même hiérarchie des marqueurs proposés par Bailly [1989].

Quelques questions cependant peuvent se poser. Ce patron monotone et croissant est-il nécessaire pour la perception de l'accent ou est-il un artefact des contraintes du système de production? Cette configuration typique est-elle nécessaire pour la perception de n'importe quel type d'isochronisme en Français? Des expériences doivent être mises au point pour tester ces hypothèses.

5. PERSPECTIVES

La relative indépendance des patrons de durée vis à vis de leur contenu segmental suggère la mise en oeuvre d'une génération de la durée par lexiques dynamiques ou par réseaux récurrents appris sur une base de données de contours en *k*. À la place de la durée du GIPC, le générateur de contours délivrerait, à la sortie, le facteur de déformation *k* associé au GIPC courant selon une entrée qui comprendrait les marques et le nombre d'éléments du mot prosodique courant: le prédicteur serait donc désormais indépendant de la nature des constituants.

Il faut, bien sûr, résoudre le problème du débit. En principe, les résultats déjà obtenus nous conduisent à adopter une base de données de contours et un corpus de moyennes et écarts-type (basés sur des logatomes) par débit.

Une première proposition est présentée Figure 5.

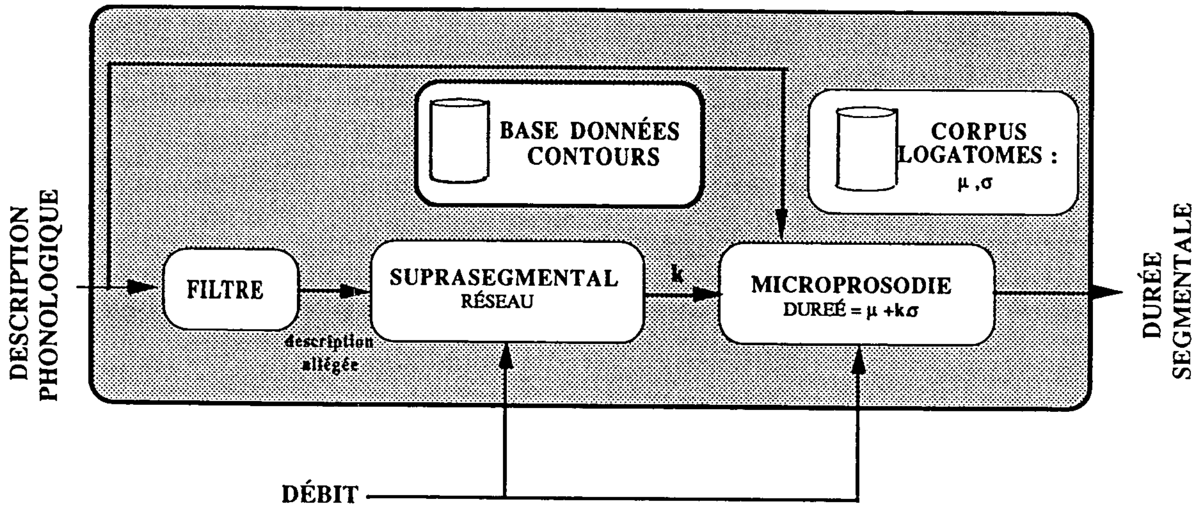


Fig. 5 : Génération de la durée segmentale en deux étapes : modèle de prédiction du facteur d'allongement et modèle de répartition tenant compte des propriétés intrinsèques des sons.

La durée segmentale serait obtenue à partir de la description phonologique en deux étapes. Le premier bloc filtre tout ce qui est de la nature des phonèmes dans la chaîne phonologique.

Pour la génération, on aura pourtant quelques problèmes à résoudre. Puisque il s'agit du GIPC, comment faire pour générer les segments avant la première voyelle de la phrase? Et ceux qui suivent la dernière voyelle (en incluant celle-ci)? On peut avancer que l'exclusion de la dernière voyelle de la structure rythmique, produite par notre stratégie, peut être mise en relation avec la grande variabilité observée du facteur d'allongement final.

Pour ce qui est du débit, l'étude de l'émergence de la pause mérite plus d'attention. Elle doit «apparaître» naturellement selon le débit, la force de la marque correspondante et le nombre d'unités concernées: la marque prosodique, de part sa force, détermine la hiérarchie des pauses mais c'est le débit qui décide de leur éventuelle réalisation physique (pause silencieuse ou subjective). Leur durée doit aussi émerger du modèle. Une étude complète de l'influence du débit dans le système est absolument nécessaire.

RÉFÉRENCES

Allen, G.D. (1975) "Speech rhythm: its relation to performance universals and articulatory timing", *Journal of Phonetics*, 3, 75-86.

Aubergé, V. (1992) "Developing a structured lexicon for synthesis of prosody", in *Talking machines : theories, models and designs* (Bailly,G. & Benoît, C., Eds.), 307-321.

Bailly, G. (1986) "Un modèle de congruence relationnel pour la synthèse de la prosdie du français", *Actes des 15es Journées d'Étude sur la Parole,* Aix-en-Provence, 75-78.

Bailly, G. (1989) "Integration of rhythmic and syntactic constraints in a model of generation of French prosody", *Speech Communication*, 8, 137-146.

Bailly, G. ,Laboissière, R. & Schwartz, J-L. (1991) "Formant trajectories as audible gestures: an alternative for speech synthesis", *Journal of Phonetics*, 19, 9-23.

Barbosa, P. & Bailly, G. (1992a) "Generating segmental duration by P-centers", *4th Workshop on Rhythm Perception and Production,* Bourges, 8-12 juin.

Barbosa, P. & Bailly, G. (1992b) "Génération automatique des *P-centers*", *Actes des 19es Journées d'Étude sur la Parole*, Bruxelles, 19-22 mai, 357-361.

Bartkova, K. (1991) "Speaking rate modelization in French : application to speech synthesis", *Proceedings of the XII th ICPS,* Aix-en-Provence, France, August 19-24, **3**, 482-485.

Bartkova, K. & Sorin, C. (1987) "A model of segmental duration for speech synthesis in French", *Speech Communication,* **6**, 245-260.

Campbell, W.N. (1992a) "Segmental elasticity and timing in Japanese speech". In : *Speech perception, production and linguistic structure* (Tohkura, Y., Vatikiotis-Bateson, C. and Sagisaka, Y., Eds.), 403-418.

Campbell, W.N. (1992b) "Syllable-based segmental duration". In : *Talking machines : theories, models and designs* (Bailly,G. & Benoît, C., Eds.), 211-224.

Di Cristo, A. (1985) *De la microprosodie à l'intonosyntaxe*, Université de Provence.

Emerard, F. (1977) *Synthèse par diphones et traitement de la prosodie, Thèse 3e cycle,* Grenoble, France.

Fraisse, P. (1974) *La psychologie du rythme*, PUF, Paris.

Jordan, M. (1986) "Serial order: a parallel distributed processing approach", *Technical Report ICS-8604.* La Jolla: University of California - Institute for Cognitive Science.

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

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

Marcus, S. M. (1976) *Perceptual centres, Thèse de doctorat non publiée,* Cambridge University.

Monnin, P. & Grosjean, F. (à paraître) "Les structures de performance en français : caracterisation et prédiction", (submitted to *l'Année Psychologiqu* e).

O'Shaughnessy, D. (1981) "A study of French vowel and consonant durations", *Journal of Phonetics*, 9, 385-406.

Pasdeloup, V.(1992) "Durée intersyllabique dans le groupe accentuel en français", *Actes des 19es Journées d'Étude sur la Parole*, Bruxelles, 19-22 mai, 531-536.

Pike, K.L. (1945) *The intonation of American English.* Ann Arbor: University of Michigan Press.

Pompino-Marschall, B. (1992) "The P-center and the perception of rhythm in connected speech", présenté au *Fourth Rhythm Workshop : Rhythm Perception and Production,* Bourges, France, 8-12 juin.

Pompino-Marschall, B. (1989) "On the psychoacoustic nature of the P-center phenomenon", *Journal of Phonetics* , 17, 175-192.

Sagisaka, Y. (1990) "On the prediction of global Fo shapes for Japanese text-to-speech", *Int. Conf. on Acoust. Speech & Sig. Proc.*, **1**, 325-328.

Semjen, A., Schulze, H-H. & Vorberg, D. (1992) "Temporal control in the coordination1 between repetitive tapping and periodic external stimuli", présenté au *Fourth Rhythm Workshop : Rhythm Perception and Production*, Bourges, France, 8-12 juin.

Todd, P.(1989) "A connectionist approach to algorithmic composition", *Computer Music Journal*, 13, 4,27-43.

Traber, C. (1992) "Fo generation with a database of natural Fo patterns and with a neural network", in *Talking machines : theories, models and designs* (Bailly,G. & Benoît, C., Eds.), 287-304.

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.

van Santen, J.P.H. & Olive, J.P. (1990) "The analysis of contextual effects on segmental durations", *Computer, Speech & Language*, 4, 359-390.

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

GENERATING SEGMENTAL DURATION BY P-CENTERS

P. Barbosa & G. Bailly

Institut de la Communication Parlée
INPG/ENSERG - Université Stendhal, URA CNRS n°368,
46, av. Félix Viallet, 38031 Grenoble CEDEX FRANCE

Abstract

k= z-score

A model for the generation of segmental duration has been proposed by N. Campbell. It is based on a prediction of the duration of a high-level unit followed by a distribution of the duration of this interval between the constituents of the group by a statistical model. The purpose of our study is to show that the inter "P-Center" interval (IPCI) is merely a better suited high-level unit for such a prediction. We show by a study of the correlations between the deformation factors (k), obtained by the statistical model, for adjacent segments that P-center groups (IPCG) are the most appropriate high-level unit. The sequences of factors k obtained for each IPCG were plotted for each sentence of our corpus. They delimite clearly the prosodic boundaries better than the syllable and exhibit an increasing monotonic accentual pattern along the prosodic group.

1. Introduction

The first models of automatic prediction of prosody for text-to-speech generation were based on a phonological description of the prosodic continuum. Inductive methods may be then used to adapt such a qualitative framework to the real production of a speaker in a specific situation [Emerard, 77 ; Traber, 92 ; Aubergé, 92]. These systems associate each phoneme with a constant number of values of prosodic parameters according to its phonetic context and the phonological unit it belongs to.

As segmental duration is closely dependent on the intrinsic nature and the context of the segment [Di Cristo, 85 ; Lehiste, 77], statistical approach is used systematically in this research field. Most generative models of segmental duration [O'Shaughnessy, 81; Bartkova & Sorin, 87 ; van Santen & Olive, 90] thus use the multiplicative model, first introduced by Klatt [76], which combine complex co-intrinsic factors (for vowels followed by voiced fricatives, unvoiced plosives or liquids, according to the vowel aperture...) and suprasegmental factors (prominence of an accent...).

By merging segmental and suprasegmental factors in such a statistical model, the latter approach do not explicitely use a higher-level unit such as the syllable as a rhythmic unit. This non-hierarchical model needs then a large amount of natural data to learn all necessary configurations and *ad hoc* adaptation to take into account variable speaking rate [Bartkova, 91]. The modeling approach [Todd, 89] has the advantage of generalization and smoothing that permit one to overcome the inadequacies of the learning corpora and to solve problems like the interpolation of contours in the synthesis domain. This approach also allows one to constrain learning by the predictor structure and not only by the data structure.

Our multiparametric predictor of the French prosody is based on the notion of gestural control [Bailly *et al.*, 91] : phonological units describe the task and the controler executes it according to rhythmic constraints specific to the language. The predictor consists of a sequential network [Jordan, 86] which generates timing of *Perceptual-Centers* (PC) locations. The time interval between successive locations of PC is then shared between the constituants according to a simple statistical model [Campbell, 92a].

We describe here a data analysis which justify the choice of the IPCI as the rhythmic control unit.

2. About the nature of the rhythmic programming unit

The quasi-periodic movement of the jaw has earlier given prominence to the role of the syllable in the rhythmic organization of the speech both in perception and production [Fraisse, 74].The work developed by Marcus [76] about PC shows that listeners are capable to adjust consistantly the delay between two syllabic recurrences (V versus CV) to perceive such an alternance as isochronous.

This tendancy to isochrony seems to be one of the components of the rhythmic control system as attested by the discussion about isochrony and isosyllabism [Pike, 45 ; Lehiste, 77]. In a experiment described by Pompino-Marschall [89] the PC position is considered to be in the middle of two beat clicks in a sequence *syllable/beat* or *beat/syllable* after adjusting the beat sequence to perceive the alternation as isochronous. Despite the diversity of the consonants the PC seems to be at the neighbourhood of the vocalic onset. His results also show that the distance between two consecutive PC determines probably the perception of momentary tempo. In our work we call this distance the inter PC interval (IPCI) and the units it contains form the inter PC group (IPCG).

In next section we envisage the candidatures of the syllable and inter PC interval as rhythmic programming units.

2.1 The prediction of segmental duration proposed by Campbell

Taking the syllable as the rhythmic programming unit, Campbell presents a two-step generative model of segmental duration for English [Campbell, 92b]: (1) generation of syllabic durations which depends on accentual typology, number and nature of constituents of the syllable ; (2) distribution of this duration between these constituents by a statistical model.

The small dimensionnality of the input (six parameters per syllable) is adequate, in the first stage, to an automatic learning by multilayers network.

Secondly, each segmental duration Dur is calculated by a deformation factor k applied to the mean μ and standard deviation σ in log-transformed milliseconds give by the formula : Dur = exp (μ + k.σ), where k is obtained by the relation $\sum$Dur = syllable duration.

The analysis of the mean values of the k-factors for four types of segments (*onset, peak, coda* and *medial*) shows a good homogeneity of behavior despite the considerable difference of the temporal characteristics of vowels and consonants. It is important to note that in Campbell's approach the coda of long syllables and the onset of the final ones have a different coefficient compared to the other constituents.

We have found similar results in French but we think that differences in mean values cannot be interpreted significantly and we developed an analysis in terms of correlation which permits to clarify what is happening locally.

2.2 The distribution model applied to French

2.2.1 The database

Two corpora were used : (1) a corpus of 2,012 logatoms in normal rate. Statistics of log-transformed durations of phonemic realizations were performed on it and the values are shown in the appendix ; (2) a corpus to predict the durations with 88 sentences in normal and fast rate. The oral reading style was the delivery of information (type : “Rappelez Monsieur Dupont à son bureau.”). The sentences have between 4 and 34 syllables.

2.2.2 Correlations between deformation factors of adjacent segments

We tried to test by a study of correlation what was the validity of the Campbell's hypothesis, i.e., the uniform behaviour of the deformation for all segments in the syllable. For this, the constituents of the syllable were classified in *onset* (any consonant preceding the vowel), *peak* (the vowel), *coda* (any consonant following the vowel). Final syllables were not considered in the analysis as suggested by Campbell.

The correlations concern a current segment and the following one : the results will permit to decide for a syllabic approach or to propose another unit for the rhythmic gesture. The results are showed in Figure 1. There is no *coda/coda* pairs in our corpora.

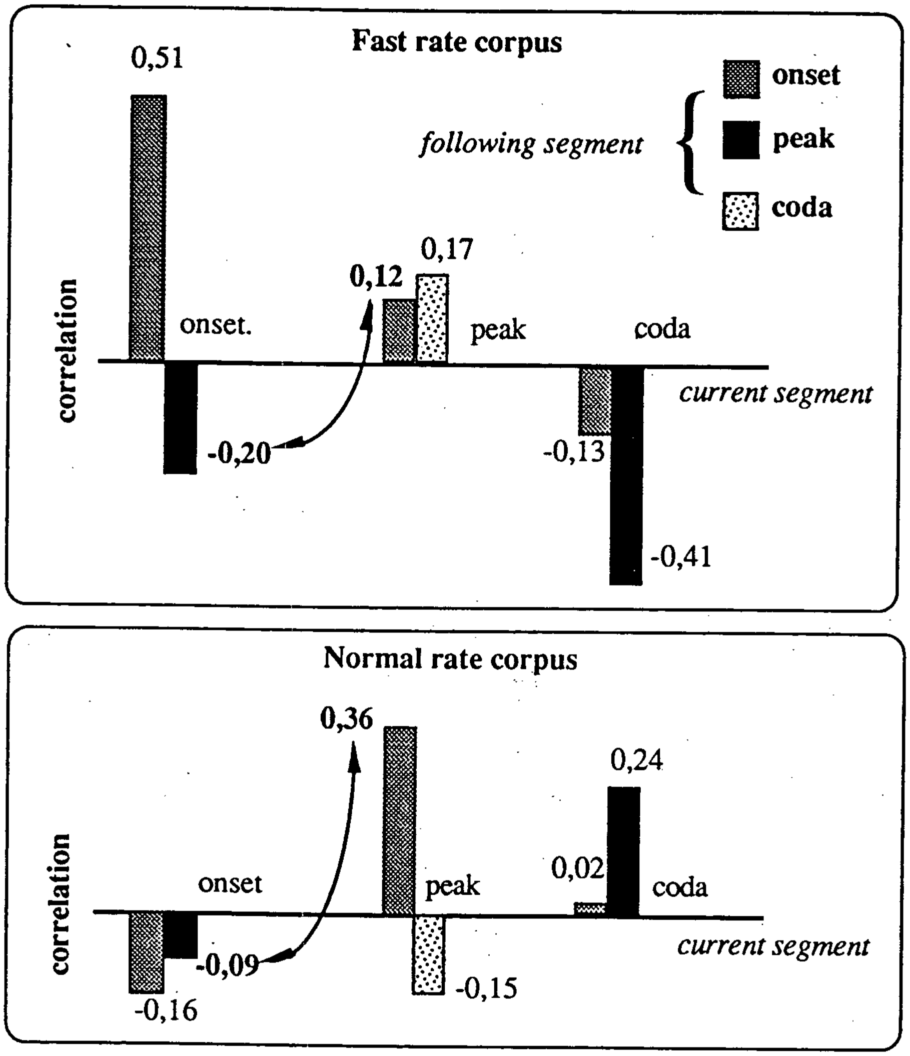


Fig. 1 : Correlation between adjacent segments. The arrows evidence the significant correlation between the vocalic peak and onset of the following syllable which favors P-Centers.

The numbers related by the arrows are statistically significant (p << 0.001) and they show a positive correlation between the current nucleus and the onset that follows it (of course, in the following classic syllable). Conversely, a negative correlation was obtained between adjacent onset and nucleus in the same syllable.

If we consider all the pairs for the syllable (onset/nucleus, onset/onset and nucleus/coda) and PC (nucleus/onset, onset/onset, nucleus/coda, coda/onset and coda/nucleus) approaches we find a correlation respectively of -0.00 (p<<0.0001) and of 0.24 (p<<0.0001), for the fast rate and a correlation of -0.02 (p<<0.0001) and of 0.31 (p<<0.0001) for the normal one.

These results allow us to envisage a group as the PC between two consecutive vocalic onsets, so called IPCG.

3. Generation of segmental durations

The main hypotheses of our model are : (1) the several inter PC durations are obtained by a deformation from an internal clock [Allen, 75]. This biological clock is the point of reference for the isochrony and the release mechanism of the motor actions [Turvey *et al.*, 90]. Results of studies on coordination between repetitive tapping and periodic external stimuli tend to reforce this assumption [Semjen *et al.*, *in press*]. This internal clock would be a rhythmic attractor : the accentual movement would consist in indicating that "something has happened" as pointed out by the phase difference between the two clocks. They would fall in phase after the realization of the accent ; (2) the segmental durations may be obtained statistically from the distributions of each class of sound if we consider that the correlation values in the last section are not negligible.

3.1 IPCI generation by a sequential network

Due to the difficult to delimit the factors influencing the duration of the inter PC group we develop a sequential connectionist architecture [Jordan, 86].

The input of the network contains the phonological description, characterized by the current and following prosodic marker, the nature of the current and following vowel, the number of consonants between two consecutive PC. Negative slopes generated at accentual and phrasal group levels indicate the position of the PC group in the current unit. This description is updated at each period of the internal clock. The total number of periods per phrase is the number of inter PC groups.

The output unit is the duration of the current inter PC group in terms of percentage of the total duration of the phrase. The distribution of the syllables coded like this fits the normal one (99.6%). Half of the corpus has been used for the learning phase and the other for test. The total error of network prediction for the test phase was less than 1.3 that of learning corpus.

3.2 Distribution of the IPCI duration between the constituents

The network output is the input of an algorithm which distributes the duration of the group between its constituents.

For instance let is take the /as/ and /aʀ/ groups (to simplify, means and standard deviations are presented in ms). μ_a= 114 ms, σ_a = 40 ms ; μ_S = 166 ms, σ_S = 48 ms ; μ_R = 94 ms, σ_R = 39 ms. For a total duration of 200 ms we obtain for /as/, Dur_a = 78 ms, Dur_S = 122 ms (k = -0.9) and Dur_a = 110 ms, Dur_R = 90 ms (k = -0.1) for the /aʀ/ group. The /a/ that preceeds the /ʀ/ is longer than the one that preceeds the /s/ by 35 %, coefficient similar to the ones for the correction of the co-intrinsic durations in some studies on duration in French [Di Cristo, 85 ; Bartkova & Sorin, 87].

For the fast rate the errors were of 18 ms per segment if we take the true inter PC durations (20 % of the mean of segmental duration in the corpus) and 25 ms per segment (28 % of the mean) for the durations predicted by the network (to both learning and test corpus).

Although connectionist models allow to know the command parameters pertinents to the task, the slow convergency of the learning phase may be a cue of redundancy in the phonological description, of the lack of pertinence of certain factors to determine the rhythmic structure of the phrase and, of course, the description of the input may not be sufficiently precise.

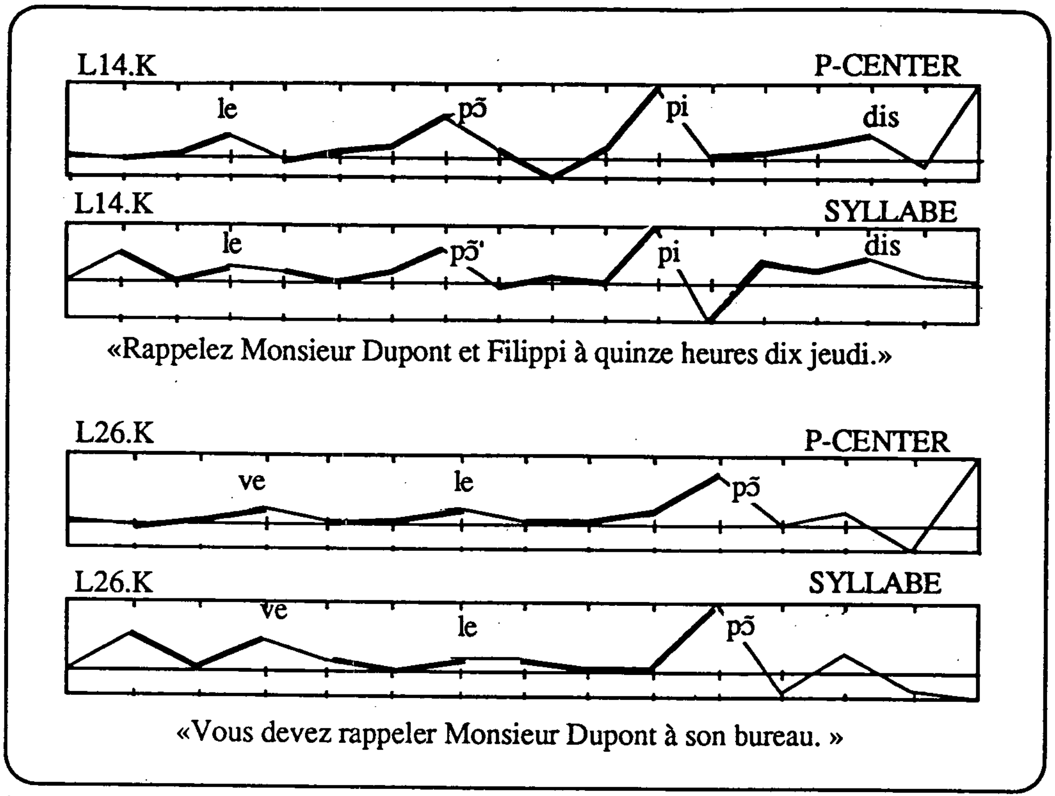
4. The durational contours

The k-factors calculated in the second step were plotted to each phrase in two configurations : in the first one, the syllable is the rhythmic unit and in the other one, the IPCG fills this function. The results for two phrases are given as example.

The k-values above the horizontal line in the curves are positive and the other ones are negative. These discrete values are tighted by filled lines.

We can see to the PC approach (1) monotonic and increasing durational "contours" by prosodic group and (2) clear marking of prosodic boundaries. These phenomena are not verified when the syllable is the unit concerned. The sentences where the syllable seems to be a good unit are rare.

We have also studied the relation between the k-deformations and the prosodic markers we use in our research [Bailly, 89]. The results shown in Table 1 permit to stablish a degree of hierarchy that is the same of the degree of strength for the markers proposed by Bailly.



«Rappelez Monsieur Dupont et Filippi à quinze heures dix jeudi.»

«Vous devez rappeler Monsieur Dupont à son bureau. »

Fig. 2 : Durational contours for phrases 14 and 26 from the normal rate corpus.

Table 1 : Deformations and pauses by marker type

marker type	pause(nb & perc)	k-mean	k-s.-dev.
ID	26/27 (96%)	4.4	0.9
IT	106/115 (92%)	3.6	1.1
RD	23/81 (28%)	1.7	0.9
SD	17/91 (19%)	1.3	0.9
SL	-	0.4	0.4

5. Discussion

The durational contours we have shown confort the assumption that increasing durations may be an important cue of the temporal gesture necessary to the realization of the accent. This increasing pattern is stronger when stronger is the marker between two adjacent words.

Are this continuous increasing pattern necessary to the perception of the accent or it is just an artefact of the constraints of the production system ? Are this typical configuration necessary to the perception of any kind of isochrony in French continuous speech ? Experiments need to be carried out to bring these questions to light.

The localizations of the PC in connected speech specially when we have consonant clusters or unstable sounds like the French *schwa* are important answers to direct our research.

6. Perspectives

The similarity of durational patterns for different natures of the units in the prosodic groups suggest the conception of a general method to generating the duration contours. We would remove the nature of the vowel at the network input level and would present for learning the k-values at the output level. In the second step the duration of the segments would be obtained from a fixed set of means and standard deviations (maybe a set per typical rate) by the formula given in section 2.1.

The pause attached to the vocalic segment preceding it (Monnin & Grosjean, *in press)* should emerge progressively from the input information (speed rate, number of units in the prosodic group) as its presence as well as its duration. No simple relation being found between statistics in the two corpora (normal and fast rates), a complete study of the influence of the rate variation is absolutely necessary.

References

Allen, G.D. (1975) "Speech rhythm: its relation to performance universals and articulatory timing", *Journal of Phonetics*, **3**, 75-86.

Aubergé, V. (1992) "Developing a structured lexicon for synthesis of prosody", in *Talking machines; theories, models and designs* (Bailly,G. & Benoît, C., Eds.), 307-321.

Bailly, G. (1989) "Integration of rhythmic and syntactic constraints in a model of generation of French prosody", *Speech Communication*, **8**, 137-146.

Bailly, G. ,Laboissière, R. & Schwartz, J-L. (1991) "Formant trajectories as audible gestures: an alternative for speech synthesis", *Journal of Phonetics*, **19**, 9-23.

Bartkova, K. (1991) "Speaking rate modelization in French : application to speech synthesis", *Proceedings of the XII th ICPS*, Aix-en-Provence, France, August 19-24, **3**, 482-485.

Bartkova, K. & Sorin, C. (1987) "A model of segmental duration for speech synthesis in French", *Speech Communication*, **6**, 245-260.

Campbell, W.N. (1992a) "Segmental elasticity and timing in Japanese speech". In : *Speech perception, production and linguistic structure* (Tohkura, Y., Vatikiotis-Bateson, C. and Sagisaka, Y., Eds.), 403-418.

Campbell, W.N. (1992b) "Syllable-based segmental duration". In : *Talking machines : theories, models and designs* (Bailly,G. & Benoît, C., Eds.), 211-224.

Di Cristo, A. (1985) *De la microprosodie à l'intonosyntaxe*, Université de Provence.

Emerard, F. (1977) *Synthèse par diphones et traitement de la prosodie, Thèse 3e cycle,* Grenoble, France.

Fraisse, P. (1974) *La psychologie du rythme*, Presses Universitaires de France, Paris.

Jordan, M. (1986) "Serial order: a parallel distributed processing approach", *Technical Report ICS-8604*. La Jolla: University of California - Institute for Cognitive Science.

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

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

Marcus, S. M. (1976) *Perceptual centres, Thèse de doctorat non publiée,* Cambridge University.

Monnin, P. & Grosjean, F. (in press) "Les structures de performance en français : caracterisation et prédiction", (submitted to *l'Année Psychologiqu* e).

O'Shaughnessy, D. (1981) "A study of French vowel and consonant durations", *Journal of Phonetics*, **9**, 385-406.

Pike, K.L. (1945) *The intonation of American English.* Ann Arbor: University of Michigan Press.

Pompino-Marschall, B. (1989) "On the psychoacoustic nature of the P-center phenomenon", *Journal of Phonetics* , **17**, 175-192.

Semjen, A., Schulze, H-H. & Vorberg, D. (in press) "Temporal control in the coordination between repetitive tapping and periodic external stimuli", presented at the *Fourth Rhythm Workshop : Rhythm Perception and Production*, Bourges, France

Todd, P.(1989) "A connectionist approach to algorithmic composition", *Computer Music Journal*, **13-4**,27-43.

Traber, C. (1992) "Fo generation with a database of natural Fo patterns and with a neural network", in *Talking machines : theories, models and designs* (Bailly,G. & Benoît, C., Eds.), 287-304.

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.

van Santen, J.P.H. & Olive, J.P. (1990) "The analysis of contextual effects on segmental durations", *Computer, Speech & Language*, **4**, 359-390.

AT LEAST TWO MACRORHYTHMIC UNITS ARE NECESSARY FOR MODELING BRAZILIAN PORTUGUESE DURATION

Plínio Almeida Barbosa
LAFAPE/IEL/Unicamp, CP 6045, 13081-970 - Campinas, Brazil (plinio@iel.unicamp.br)

Abstract

By characterizing Brazilian Portuguese acoustic duration, this work presents two arguments in favor of macrorhythmic units. First, the emergence of distinct durational patterns for lexical and phrasal accents. Second, the homogeneous lengthening (shortening) effect of segments correlating syllables at lexical stress and IPCGs at phrasal accent. A two-stage model of segmental duration generation was derived.

Résumé

La caractérisation de la durée en portugais brésilien (PB) permet de faire émerger une typologie accentuelle signalant la présence de deux unités macrorythmiques. Les maxima des *z-scores* de la syllabe coïncident systématiquement avec la position de l'accent lexical tandis que les maxima des *z-scores* du GIPC démarquent les frontières prosodiques de l'énoncé. Un modèle à deux étapes permettant la génération simplifiée de la durée segmentale du PB en est dérivé.

Introduction

Is it really necessary to consider the existence of (macro)rhythmic programming units (RPU) in order to generate segmental duration? Certainly not, as van Santen's work has convincingly shown so far [12]. In early work [1] we have shown that the durational structure associated with a set of read sentences reveals a certain kind of organization over the segment level. An approach for duration generation that takes into account the macrorhythmic organization of speech was obtained that drastically simplifies the mechanism of duration assignment.

As has already been demonstrated for French, normalizing the acoustic duration of consecutive segments points to an organization into higher units whose boundaries are two consecutive vowel onsets [1]. This unit was named inter-perceptual-center group (IPCG) by reference to research on p-centers [8][11], whose findings suggest that their optimal location is the vicinity of the vowel onset[6].

Normalized durations are obtained through Campbell's *z-score* model [3]. The z value of each segment *s* is computed by writing: $\mathrm{Dur}_s = \exp\,(\mu_s + z.\sigma_s)$ (1), where Dur_s is segment duration and (μ_s, σ_s) stands for the average and the standard-deviation of the log-transformed durations of all *s* realizations in an *ad hoc corpus*. The strong elasticity hypothesis in Campbell's model says that all segments in a syllable frame have the same z-score: a single value of *z* per syllable can be computed by writing: $\mathrm{Dur}\ (\mathrm{syllable}) = \Sigma_s \exp\,(\mu_s + z.\sigma_s)$ (2).

French data allow us to propose a weaker elasticity hypothesis where the rhythmic unit is the IPCG, not the syllable. Brazilian Portuguese (BP) data, on the other hand, indicate that at least two macrorhythmic units are necessary to model the durational structure of read sentences. BP rhythm can shed light on the segmental/suprasegmental controversy thanks to the greater complexity of its accentual typology.

BP accentual typology

In BP, lexical accent can be assigned to the final, penultimate or antepenultimate syllable. Stressed syllables can be enhanced as they are uttered by carrying phrasal accent. Only lexically stressable syllables can bear phrasal accent.

The acoustical correlates of accent are often the greater duration of the unit bearing the accent and the decrease of intensity in the post-stressed syllables (if any)[9][1]. Our data reveal that the accent is carried out by units over the segmental level.

The following results are based on the analysis of two *corpora*.

Speaker's duration statistics

Segmental durations were determine from a 1195-nonsense word *corpus*, containing all BP phonemes and allophones of a native 30-year speaker (from the Paulista dialect). Statistical analyses were then performed. The results on Table 1 confirm current knowledge on duration in BP (which is in agreement with universal trends [7]): (a) for front and back vowels, the higher the vowel, the shorter its average duration; (b) post-stress vowels (/ɐ, ɪ, ʊ/) are shorter than their stressed counter-parts (/a, i, u/); (c) nasal vowels are longer than their oral counterparts; (d) voiceless consonants are longer than their voiced counterparts.

Table 1: Mean duration (and standard-deviation) in ms of the BP phones for our speaker.

i	145 (37)	ɐ	111 (45)	ĩ	209 (25)	tʃ	149 (20)	f	138 (14)	n	76 (15)
e	170 (36)	ɪ	98 (44)	õ	229 (26)	k	121 (21)	s	143 (26)	ɲ	103 (24)
ɛ	175 (32)	ʊ	77 (19)	ũ	215 (29)	b	86 (17)	ʃ	143 (16)	ɾ	47 (16)
a	165 (28)	j	92 (10)	j̃	136 (14)	d	71 (17)	v	78 (16)	r	81 (12)
ɔ	183 (29)	w	97 (25)	w̃	139 (23)	dʒ	109 (18)	z	87 (21)	ʁ	62 (15)
o	168 (35)	ɐ̃	174 (46)	p	120 (20)	g	67 (16)	ʒ	89 (12)	l	73 (16)
u	134 (42)	ẽ	210 (44)	t	113 (20)			m	90 (12)	ʎ	77 (14)

The log-transformed versions of these data were used in formula (2) (where *syllable* is either a phonological syllable or an IPCG) to compute the z-scores of syllables and IPCGs of 100 sentences read by the same speaker. This corpus was manually segmented and carefully labeled by the author. Sentence length varies between one and 84 syllables. Syntactic boundaries were also marked using a set of eight hierarchical labels [1].

Extracting durational contour from z-score evolution

Segments were grouped into two kinds of RPU's: syllable and IPCG. By using the raw duration of each group in formula (2) above, the *z-scores* were computed for all RPU's in each sentence. An example is shown in Figure 1.

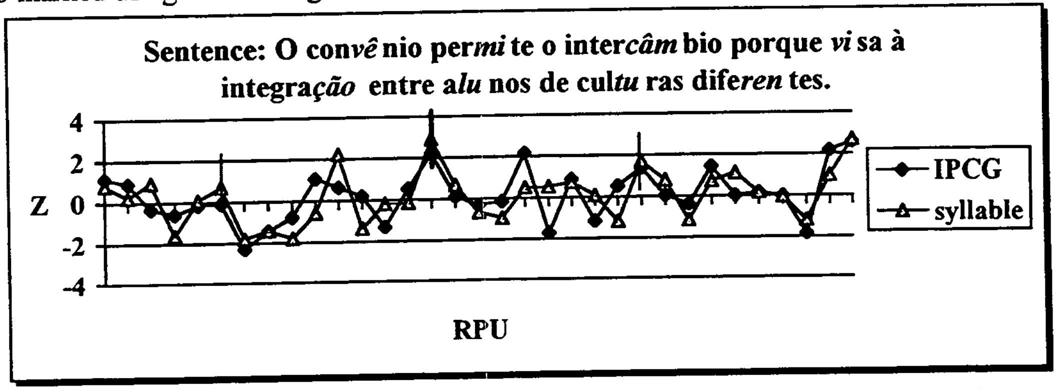


Figure 1: Z-scores (vertical axis) for IPCG's and syllables in the sentence "O convênio permite o intercâmbio porque visa à integração entre alunos de culturas diferentes." ("The agreement allows for interchange since it aims at the integration among students from different cultures."). Lexically stressed syllables are italicized. Tick marks on the horizontal axis represent the vowels along the sentence (signaling both syllables and IPCG's; NB: orthographic *-io* – in *convênio* and *intercâmbio* – is a semivowel/vowel cluster).

In all 100 rhythmic contours, syllable z-scores indicate the (lexically) stressed syllables of the utterance: the highest z-score within each word coincides with the lexically stressed syllable. On the other hand, if the highest IPCG z-scores within non-clitic words *at lexical stress* are taken as a criterion for boundary placement as well as a measure of boundary strength, coherent prosodic groups are obtained for all sentences in the *corpus*. In Figure 1, the strongest boundary (IPCG z-score of *vi* in *visa*) splits the sentence into two chunks of 16 IPCG's each. The property of eurhythmicity is very clear and the coherence between this result and those of Grosjean's performance trees [10] is notorious.

IPCG z-scores delimit accentual groups (prosodic words) where the rhythmic pattern is characterized by frequent alternation of z-score values at the beginning of the accentual group followed by a duration *crescendo* (starting at least on the penultimate syllable) towards the last stressed syllable in the group.

Statistical confirmation of the results

Statistical analyses confirm that actual segmental z-scores (by using raw segmental duration into formula 1) are strongly correlated in RPU frames. For these analyses, phrasal boundaries were marked by choosing positions in the utterance corresponding to IPCG z-score maxima (coinciding with lexically stressed RPU) and by carefully hearing the utterances in order to confirm these candidates to prominence.

The results in table 2 show that, in phrasal accent position, onset and nucleus segments are negatively correlated (-31%) whereas *nucleus/coda* segment sequences in the rhyme are positively correlated (76%). On the other hand, *onset/nucleus* and *nucleus/coda* segment sequences confirm that lexically stressed syllables (not bearing phrasal accent) form a homogeneous unit. In non prominent positions, VC and V#C sequences seem to suggest that the IPCG is a homogeneous unit whose rhyme component is enhanced at phrasal accent position.

These results corroborate Campbell's predictions [4] only in part: since he adopts the syllable as the RPU, onset segments in syllables at prosodic boundaries are overpredicted. The observed final lengthening affects the rhyme, not the onset. This is produced by lesser articulator stiffness associated with closure movement [5].

Table 2: Correlation (in percentage) between consecutive segmental *z-scores* according to accentuation degree[2].

	lexical accent	*phrasal accent*	*other positions*
onset/ nucleus	63	-31	4
nucleus/ onset	ns	26	56
nucleus/ coda	48	76	63

Segmental duration generation

RPU z-scores can be a means of deriving the segmental durations of a particular sentence (if formula 1 is applied by setting $z = z_{syllabe}$ or $z = z_{IPCG}$). The above durational contours can be easily generated by neural networks, as implemented by the author for French [1]. In this work segmental durations were computed in a second stage by sequentially applying formula 2 and 1 (named repartition algorithm) with the IPCG duration delivered by the network output.

For French, a sequential, recurrent network was used to learn to associate a phonological, prosodic description of the sentence to the respective rhythmic pattern expressed by the evolution of IPCG duration over the sentence. This method allows for preservation of macrorhythmic unit durations maintaining the rhythmicity of the original sentence. Although the vowel quality of the syllable nucleus is sequentially described at the network input, only the number of the intervening consonants between two vowels and not their nature are represented. This implicitly means that consonant nature is not important to derive IPCG duration (although microrhythmic differences in segmental duration can be captured in the second stage by the repartition algorithm).

For BP, a simpler network was used by simulating a multilayered perceptron. In the perceptron input, a phonological, prosodic description of each sentence is used to infer the network output, the IPCG and syllable z-score evolution over the sentence. Thanks to a greater coherence between accentual typology and z-score patterning, the learning of the network was, in fact, faster. Furthermore, z-score patterns are smoother than RPU duration one. But original RPU durations are no longer preserved. What is preserved here is the rhythmic structure as represented by the z-score patterning.

Our model of segmental duration generation was applied to the learning and also to the test *corpus* subsets. Results respectively show 32-ms and 36-ms standard-deviation error scores per phoneme. The model was capable to generalize even when lexical stress position was manipulated.

Conclusions

By claiming that RPU duration and boundary are cued by linguistic factors such as prosodic prominence and lexical stress we corroborate the idea that explicit models of speech production need a rhythmic input as suggested by the rhythmic tier in Articulatory Phonology [2]. The existence of at least two macrorhythmic units suggests however a modification on the AP framework since rhythmic nodes should dominate different suprasegmental units.

It is also clear that speech production models should include a cognitive, phonologically oriented input that would account for distinct representations for lexical and phrasal accent.

The BP results correlating lexical stress with the syllable frame, on one hand, and phrasal accent with the IPCG frame, on the other, constitute a strong argument for macrorhythmic units in prosodic modeling.

Acknowledgments

The author is supported by a grant (n° 94/3358-0) from the FAPESP. We are grateful to CPqD-Telebrás and to PUC-SP for allowing us to use their *corpora*. We are in debt to Sandra Madureira and Eleonora Albano for their helpful suggestions and comments.

References

[1]Barbosa, P.A. & Bailly, G. (1994) *Characterisation of rhythmic patterns for text-to-speech synthesis,* Speech Communication, 15 (1-2), 127-137.

[2]Browman, C.P. & Goldstein, L. (1990) *Tiers in articulatory phonology with some implication for casual speech.* Kingston, J. & Beckman, M.E. (Eds.). Papers in Laboratory Phonology 1. Cambridge University Press, 341-376.

[3]Campbell, N.W. (1992) Syllable-based segmental duration. In: Talking Machines: theories, models, and designs (Bailly, G. & Benoît, C. Eds.) 211-224.

[4]Campbell, N.W. (1993) *Automatic detection of prosodic boundaries in speech*, Speech Communication, 13, 343-354.

[5]Edwards, J., Beckman, M.E. & Fletcher, J. (1991*) The articulatory kinematics of final lengthening*. J. Acoust. Soc. Am. 89 (1), 369-382.

[6]Janker, P.M. (1995) *On the influence of the internal structure of a syllable on the p-center perception.* XIII ICPhS, August 13-19, Stockholm, Sweden, 2, 510-513.

[7]Lehiste, I. (1970) *Suprasegmentals.* Cambridge, Massachusetts: MIT Press.

[8]Marcus, S.M. (1981) *Acoustic determinants of Perceptual-center (p-center) location*, Perception and Psychophysics, 30(3), 247-256.

[9]Massini, G. (1991) *A Duração no estudo do acento e do ritmo em português,* Tese de Mestrado, Unicamp.

[10]Monnin & Grosjean (1993) *Les Structures de performance en français : caractérisation et prédiction*, L'Année Psychologique 93, 9-30.

[11]Pompino-Marschall, B. (1991) *The syllable as a prosodic unit and the so-called P-centre effect.* FIPSK der Universität München, 29, 65-123.

[12]van Santen, J.P.H. (1994) *Assignment of segmental duration in text-to-speech synthesis*. Computer, Speech and Language 8, 95-128.

[1] Fundamental frequency is not an acoustic correlate of lexical accent.

[2] Segments were categorized for each phonological syllable with three labels: *onset,* for each segment in onset position, *nucleus*, for the vowel nucleus, and *coda*, for each segment in coda position. Sequences as *nucleus-onset* span necessarily over syllable boundaries. Lexical accent category refers to lexical stressed RPU not bearing phrasal accent. Only the significant values are reproduced here.

TELEMO '96

UM CONVERSOR TEXTO-FALA PARA O PORTUGUÊS BRASILEIRO COM PROCESSAMENTO LINGÜÍSTICO DE ALTA QUALIDADE*

Fábio Violaro[†], Plínio A. Barbosa[‡], Eleonora C. Albano[‡] e Edson Françozo[‡]
[†]FEE/DECOM, CP 6101 e [‡]IEL/LAFAPE, CP 6045, UNICAMP, 13083-970 - Campinas, SP
e-mail: fabio@decom.fee.unicamp.br

*Este trabalho tem os apoios da FAPESP, através do projeto temático n° 93/0565-2 e da bolsa de pós-doutorado n° 94/3358-0 para o segundo autor, e do CNPq, através do projeto temático n° 523555/94-6 (NV).

Resumo

Este trabalho descreve um conversor texto-fala em desenvolvimento na UNICAMP. O sistema utiliza um método de síntese concatenativo aliado a técnicas PSOLA ou Híbrida, e está em implementação no DECOM/FEE. O sistema busca a alta qualidade pela integração de processamentos lingüísticos oriundos do conhecimento em português brasileiro adquirido no LAFAPE/IEL.

Abstract

This paper describes a TTS (text-to-speech) system under development at UNICAMP. The system uses concatenation of poliphones, employs both the PSOLA or the Hybrid synthesis techniques, and incorporates a high degree of linguistic processing for the Portuguese language spoken in Brazil.

1- Introdução

A Fig.1 abaixo apresenta um diagrama de blocos pormenorizado do conversor texto-fala atualmente em desenvolvimento. Ele emprega o método de síntese concatenativo, em que um texto genérico irrestrito é convertido em sinal de fala através da concatenação de unidades básicas denominadas polifones. Esses polifones incorporam as transições entre fones, de modo que a concatenação é feita de maneira relativamente suave, pela junção de regiões do sinal com conteúdo espectral semelhante. Quanto à técnica empregada para a obtenção do sinal acústico, ele pode operar tanto com a técnica tradicional PSOLA (*Pitch Synchronous OverLap and Add*) [1] quanto com a técnica Híbrida [2,3], que tolera maiores variações prosódicas.

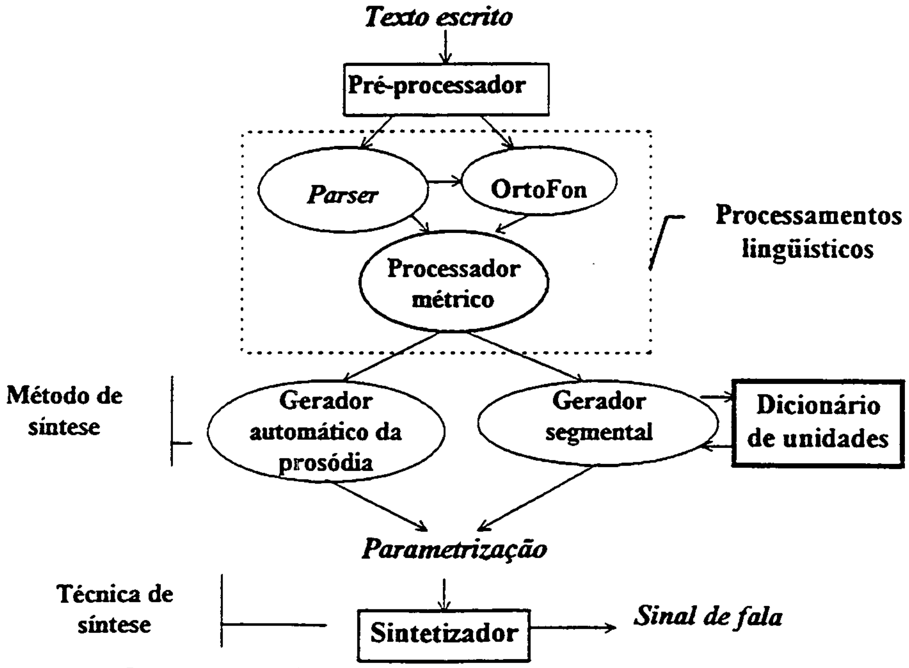


Fig. 1: Conversor texto-fala de alta qualidade em desenvolvimento na Unicamp

O conversor texto-fala incorpora um nível elevado de processamento lingüístico: concepção de um dicionário de unidades de modo a minimizar as descontinuidades na concatenação entre polifones, conversor ortográfico-fônico (OrtoFon) integrando noções modernas da teoria fonológica visando à realização da transdução letra/fonema de maneira versátil (transcrição larga seguida de transcrição estreita, esta última realizada por um processador métrico), analisador gramatical ou *parser*, para resolver os problemas de ambigüidade na pronúncia de palavras homógrafas-heterofônicas e marcar fronteiras sintáticas, gerador automático da prosódia, para controle automático de duração e *pitch* da fala sintetizada. No caso da síntese concatenativa, o gerador segmental apenas realiza o endereçamento dos polifones pré-armazenados no dicionário referido acima. Estes processamentos lingüísticos são imprescindíveis para assegurar à fala sintetizada uma aparência de fala natural, em contraposição à fala tida como "de robô" de sistemas mais simples, que empregam processamentos lingüísticos (muito) mais simplificados.

2- Processamento Lingüístico

A desvantagem, já apontada por Klatt [4], do método concatenativo, a saber, a possibilidade de concatenação de unidades que apresentariam descontinuidades espectrais responsáveis pela sensação de "voz metálica" será minimizada pela seleção rigorosa, quando da elaboração do dicionário de unidades, dos polifones e do locutor para a gravação. Estes polifones serão selecionados a partir do conhecimento em Fonética Acústica do LAFAPE, tornando possível a escolha de estruturas aptas a conservar as características de coprodução [5] dos segmentos e aptas a dar conta da variabilidade da combinatória de segmentos presentes na fala.

Há vários trabalhos em preparação no LAFAPE cobrindo a descrição de sons altamente coarticuláveis. Esteş trabalhos revelam que, no português brasileiro, a coarticulação de vogal a vogal [6] estende-se, em muitos casos, até o meio da vogal [7]. Assim, o segundo formante é relativamente alto no [ɐ] final de "pipa" e relativamente baixo no [ɐ] final de "pupa". Um sistema de síntese concatenativa deveria, portanto, ao menos em princípio, tratar /pa/ do contexto /i-/ e /pa/ do contexto /u-/ como unidades distintas, empregando, na realidade os trifones /ipa/ e /upa/. Levando em conta, por outro lado, o grande número de seqüências da forma VCV (Vogal-Consoante-Vogal) na língua, parece sensato ignorar esse tipo de coarticulação, apesar do ganho em naturalidade que a sua incorporação ao sistema idealmente poderia trazer. No entanto, seqüências VCV em que a consoante é altamente coproduzida com as vogais adjacentes devem ser mantidas, tais como para [ɾ], [ɲ] ou [ʎ]. Estes tipos de decisão permitirão reduzir um número inicial de 5 000 para algo pouco maior do que 2 000 polifones. Estes polifones constituem o dicionário de unidades.

Quanto à conversão texto-fala (vide Fig. 1), o texto escrito dá entrada em um pré-processador que reescreve por extenso números, abreviaturas e siglas. Em seguida, operam os módulos definidos pelo *parser* (de onde o texto retorna com uma análise morfossintática para uso do OrtoFon), e o OrtoFon. Este é composto de um transcritor propriamente dito e um dicionário de exceções. O transcritor aplica regras para reescrever fonicamente os caracteres ortográficos e marcar outras informações relevantes, tais como acentuação lexical e fronteiras silábicas. O dicionário de exceções busca e transcreve exceções às regras de transcrição.

O atual dicionário de exceções é constituído pelos 1383 verbetes do Minidicionário Aurélio [8] nos quais as regras do transcritor persistiram produzindo erros após sucessivos refinamentos (a taxa de acerto das regras é de 96%). O transcritor tem atualmente um processador de vogais médias em oposições gramaticais, o qual, na verdade, inclui um *mini-parser*. Ele reconhece paroxítonos com *e*, *o* ortográficos na penúltima sílaba e procura a palavra numa lista de homógrafos verbo-substantivo e verbo-adjetivo. No caso de encontrá-la, um conjunto de regras simples, de aplicação provável mas falível, decide se ela pertence a uma ou outra classe. Assim, por exemplo, a forma ortográfica 'acordo' é localizada na lista e testada quanto à possibilidade de tratar-se do substantivo (pronunciado com [o]) ou do verbo (pronunciado com [ɔ]). Esse procedimento dá conta de frases de estrutura sintático-semântica simples tais como "Eu acordo às seis", "O governo fechou o acordo", mas falha em casos mais complicados tais como "Eu acordo não faço não". A precariedade dessa análise deverá ser superada quando o verdadeiro *parser* estiver em operação, mas isso não garantirá um acerto de 100%, visto que certas ambiguidades só podem se dirimir via análise semântico-pragmática discursiva. Por exemplo, nas frases "O medo subsiste no Irã", "Esvaziou-se a forma do escultor", "A sede da torcida é grande", as vogais tônicas de "medo", "forma" e "sede" podem tanto ser abertas quanto fechadas.

O OrtoFon dá conta assim da fonologia lexical (ligada à formação da palavra) e alimenta o módulo de refinamento da transcrição ou processador métrico que realiza os cálculos métricos necessários para dar conta da fonologia pós-lexical (ligada ao que se passa quando a palavra está inserida em uma frase). A transcrição fônico-métrica resultante recebe então duas traduções. Uma, feita pelo gerador prosódico, calcula a variação de tom, duração e intensidade de cada enunciado de acordo com as cesuras e demais distinções "suprassegmentais" estabelecidas. A outra, feita pelo gerador segmental, recodifica os segmentos nos termos do sistema de síntese escolhido. Na síntese concatenativa, são explicitadas as unidades a concatenar. Embora já operante, o OrtoFon está sendo melhorado quanto ao funcionamento interno e quanto à comunicação com o resto do sistema de síntese.

Um *parser*, composto de um léxico e analisador morfológico-sintático, é um módulo indispensável à qualidade de um sistema de síntese na medida em que, na fala lida, a prosódia é uma interpretação rítmico-melódica da sintaxe e da semântica. Um grupo de pesquisadores do LAFAPE está elaborando um *parser* a partir de um léxico elaborado pelo grupo que se ocupa do OrtoFon. Esse léxico constitui não só o embrião do *parser* mas também uma ferramenta importante de teste do OrtoFon. Ele foi criado manualmente através da digitação dos 27.078 verbetes do Minidicionário Aurélio. Um conjunto de regras de reescrita acrescentou automaticamente as flexões dos verbos regulares. As flexões dos verbos irregulares foram acrescidas à mão.

O modelo de geração automática da prosódia permite a atualização do sinal de fala sintetizado em relação à nova informação, em termos de parâmetros prosódicos, veiculada pelo texto escrito. Juntamente com a escolha dos polifones, o gerador de prosódia garante a alta qualidade de um sistema de síntese. O modelo de geração da duração segmental [9,10] foi concebido em duas etapas:

1- na primeira, um *perceptron* multicamadas gera um padrão de formas rítmicas associado à frase a sintetizar, a partir de uma descrição fonológico-prosódica fornecida pelo processador métrico. Tais formas são expressas por uma medida de duração normalizada (*z-score* [9,11]) de unidades de programação rítmica mínimas: sílaba e GIPC. O GIPC (grupo inter-*perceptual-center*) é delimitado por dois *onsets* acústicos consecutivos da vogal, conforme trabalho realizado para o francês [9] e o português brasileiro [11]. Em português, a sílaba é uma unidade coesa em posição de acento lexical (não marcado frasalmente) enquanto que o GIPC é uma unidade coesa nas demais posições, em particular em posição de acento frasal, onde é notório o alongamento restrito à rima da sílaba nesta posição. Além disso, os pontos de máximo do *z-score* da sílaba para cada palavra da frase coincidem com a posição de acento lexical, enquanto que os pontos de máximo do *z-score* do GIPC demarcam a frase em grupos acentuais que revelam a estrutura prosódica ou de *performance* das sentenças.

2- na segunda etapa, um algoritmo de repartição distribui as durações normalizadas geradas na primeira etapa, em termos de durações absolutas dos segmentos que compõem as respectivas unidades rítmicas. O valor do *z-score* das sílabas em posição de acento lexical (não marcado frasalmente) e o valor do *z-score* dos GIPCs nas demais posições são usados para a obtenção da duração absoluta dos segmentos que formam as respectivas unidades rítmicas (para maiores detalhes ver [11,12]).

A informação de duração dos segmentos pode passar por um processo de parametrização antes de entrar no sintetizador propriamente dito. As duas etapas de geração se utilizam de informação contida em *corpora* de fala: o *perceptron* aprende a gerar formas rítmicas a partir do padrão presente em um subconjunto das frases de um *corpus* de frases isoladas lidas e o algoritmo de repartição se baseia em análise estatística do *corpus* de polifones e do *corpus* de frases isoladas.

Em um primeiro teste do sistema, as formas entoacionais serão obtidas por um modelo similar ao usado para a duração. Neste modelo, as formas entoacionais serão obtidas via treinamento de um *perceptron* multicamadas possuindo, em sua entrada, a mesma descrição fonológico-prosódica usada para o modelo de duração e, em sua saída, uma representação da variação da curva de freqüência fundamental ao longo da frase.

3-Sintetizador propriamente dito: técnicas disponíveis

O sistema pode operar com duas técnicas alternativas de síntese: a técnica tradicional PSOLA [1] e a técnica Híbrida [2,3,14]. Ambas as técnicas operam de maneira síncrona com o período de *pitch*, daí a necessidade de se submeter os logatomas de onde serão extraídos os polifones a uma marcação de *pitch*. Inicialmente são determinadas cadeias de máximos e mínimos locais calculadas segundo a referência [13]. Essas cadeias sofrem em seguida uma série de processamentos, tais como eliminações, defasagens, interpolações, extrapolações, etc., fornecendo uma seqüência de marcas espaçadas pelo período de *pitch* e posicionadas nos picos do sinal nas porções sonoras, e regularmente espaçadas de 10 ms nas porções não sonoras.

Na técnica PSOLA os polifones são armazenados sob forma temporal e as variações prosódicas quando da síntese são realizadas através da manipulação de segmentos janelados do sinal, segmentos estes centrados nas marcas de *pitch*. As variações de duração são obtidas através da eliminação/repetição de segmentos janelados. As variações de freqüência de *pitch*, por sua vez, são obtidas através da maior/menor superposição entre os segmentos janelados. Apesar de amplamente utilizada em sistemas comerciais de conversão texto-fala, a técnica PSOLA apresenta alguns inconvenientes que surgem quando da realização de alterações prosódicas (duração e *pitch*):

1- as modificações de duração só podem ser implementadas de maneira quantizada (..., 1/2, 2/3, 3/4, ..., 4/3, 3/2, 2/1, ...);

2- as modificações de *pitch* introduzem uma alteração de duração, causada pela maior ou menor superposição dos segmentos janelados, variação esta que deve ser compensada apropriadamente;

3- durante o aumento de duração efetuado em porções não sonoras do sinal de fala, a repetição de segmentos introduz uma periodicidade que é responsável por uma aparência "metálica" na fala sintetizada;

4- variações elevadas de *pitch* causam distorções sensíveis, uma vez que a envoltória resultante da superposição das janelas se afasta consideravelmente de um nível constante.

Para contornar as deficiências acima, foi desenvolvido uma técnica alternativa de síntese, denominada técnica Híbrida, e que será explicada sucintamente a seguir.

Na técnica Híbrida o sinal de fala correspondente aos polifones é separado em duas componentes: componente harmônica e componente de ruído. A separação da componente de ruído é feita através de uma filtragem passa-altas com freqüência de corte variável e que é implementada via um procedimento baseado em DFT (*Discrete Fourier Transform*). Seja *pm[i]* a posição das marcas de *pitch*. Para cada valor de *i*, é calculada uma DFT para o segmento de fala se estendendo de *pm[i-1]* a *pm[i+1]*. Em seguida são tomados os termos da DFT até uma freqüência máxima *fm[i]*, calculada a DFT inversa e o sinal resultante é janelado através de uma janela de Hanning assimétrica centrada em *pm[i]* e se estendendo de *pm[i-1]* a *pm[i+1]*. A realização do procedimento acima para todos valores de *i* e a somatória dos segmentos janelados gera um sinal que, após subtração do sinal de fala original, corresponde à componente de ruído.

Devido aos diferentes processamentos a que serão submetidas as duas componentes da técnica Híbrida, é importante que a componente harmônica seja calculada até uma freqüência máxima *fm[i]* que varia a cada segmento de fala. Assim, uma determinação automática de *fm[i]* é efetuada através de uma análise de autocorrelação da componente de ruído calculada no intervalo de *pm[i-1]* a *pm[i+1]*. Essa freqüência máxima pode assumir um entre os seguintes valores: 2, 3, 4 e 5 kHz. Se a autocorrelação da componente de ruído se apresenta elevada para um atraso em torno do valor do período de *pitch*, *fm[i]* é incrementada de 1 kHz, uma nova componente de ruído é calculada e assim sucessivamente.

A componente de ruído sofre um modelamento AR (excitação de ruído branco aplicada a um filtro LPC de 15 pólos para um sinal limitado em 8 kHz e amostrado a 16 kHz). A cada marca de *pitch* é associado um conjunto de coeficientes correspondentes ao modelamento AR: 15 coeficientes do filtro e coeficientes de ganho, um a cada 2 ms, calculados através de uma filtragem passa-baixas das amplitudes quadradas do ruído.

Nas porções sonoras do sinal, ambas as componentes do modelo podem coexistir, ao passo que nas porções não sonoras, apenas a componente de ruído prevalece.

A componente harmônica calculada acima, via DFT, não se presta para a síntese na presença de variações prosódicas devido à sua peridiocidade a cada 2 períodos de pitch $(\omega_0[i] = 2.\pi/(pm[i+1]-pm[i-1]))$. Nesse caso, de posse do valor *fm[i]*. a componente harmônica do modelo é calculada através da análise do segmento de fala que se estende de *pm[i-1]* a *pm[i]+(pm[i]-pm[i-1])*. e empregando uma freqüência fundamental $\omega_0[i]=2.\pi/(pm[i]-pm[i-1])$, onde $pm[i]-pm[i-1]$ é o período de *pitch* associado à marca $pm[i]$. Adicionalmente deve ser imposta uma ponderação de erro, o que exige que o cálculo das componente harmônicas nas freqüências múltiplas de $\omega_0[i]$ seja feito através da resolução de um sistema sobredeterminado de equações. Após este cálculo, a cada marca de *pitch* é associado o valor de $\omega_0[i]$ e de $(fm[i].\pi)/(\omega_0[i].8kHz)$ componentes harmônicas (amplitude e fase).

O uso de um modelamento paramétrico para a componente de ruído e de um modelo de forma de onda para a componente harmônica é que justifica o termo Híbrida. A síntese da componente harmônica é feita através do modelo de síntese senoidal [15], que providencia uma interpolação gradual da amplitude, freqüência e fase das harmônicas entre duas marcas de *pitch* adjacentes.

Na presença de variações prosódicas, a posição das marcas $pm[i]$ se altera e por isso a descrição passa agora a ser feita em termos de marcas $m[i]$, com os parâmetros AR e harmônicos associados a cada marca $m[i]$. Adicionalmente, sejam $\alpha[i]$ e $\beta[i]$ os coeficientes de controle de duração e freqüência de *pitch* respectivamente. Assim, na presença de uma variação de duração, tem-se $m[i]-m[i-1] = \alpha[i].(pm[i]-pm[i-1])$.

A síntese LPC é empregada para a componente de ruído. Uma interpolação dos coeficientes de reflexão associados a $m[i-1]$ e $m[i]$ e uma atualização de ganho é providenciada a cada $2.\alpha[i]$ ms. Já a variação de duração e *pitch* da componente harmônica é realizada segundo o procedimento descrito em [16] e que se encontra ilustrada na Fig. 2. No caso de uma variação da freqüência de *pitch*, tem-se uma nova freqüência fundamental $\beta[i].\omega_0[i]$, e um novo conjunto de coeficientes harmônicos deve ser calculado. Isso é feito através de uma reamostragem da envoltória do espectro, que pode ser obtida pela simples interpolação linear entre as partes real e imaginária dos coeficientes harmônicos, seguida de uma correção de amplitude para manter a energia do sinal. Como, após uma variação de duração ou de *pitch*, um número não inteiro de períodos de *pitch* pode estar presente entre as marcas $m[i-1]$ e $m[i]$, um atraso $t_0[i]$ deve ser introduzido para efetuar uma correção de fase nas harmônicas associadas à marca $m[i]$. O cálculo dessa correção para a *késima* harmônica, na presença de variações tanto de duração quanto de *pitch*, é dado pelas equações a seguir (onde $L[i]$ é um inteiro calculado para assegurar o mínimo valor para $|t_0[i]|$):

$$\theta'_k[i] = \theta_k[i]-k.\beta[i].\omega_0[i].t_0[i];\ t_0[i] = t_0[i-1]+L[i].pitch'[i]-\alpha[i].pitch[i]$$
$$pitch[i] = pm[i]-pm[i-1];\ pitch'[i] = pitch[i]/\beta[i]$$

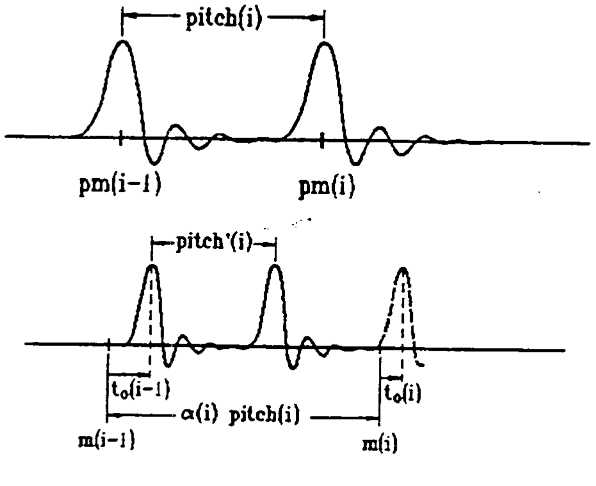


Fig. 2: Variação de duração e *pitch* na técnica Híbrida

Apesar da maior complexidade computacional requerida, a síntese Híbrida apresenta uma série de vantagens sobre a síntese PSOLA, daí sua inclusão no sistema em desenvolvimento: as variações de duração e *pitch* se processam de maneira independente e contínua, a aparência de som metálico é evitada quando do aumento da duração dos segmentos não sonoros, e a fala pode suportar modificações elevadas de duração e *pitch* sem apresentar distorções significativas. Devido a essas vantagens, a síntese Híbrida já está sendo utilizada para o teste e avaliação do gerador automático de prosódia.

4- Estágio Atual e Conclusões

Vários dos módulos do conversor texto-fala estão completamente testados e aguardando uma integração no sistema como um todo. Com referência à Fig. 1, já estão concluídos e testados o Pré-processador, o OrtoFon, o módulo de duração do gerador automático de prosódia e os sintetizadores baseados nas técnicas PSOLA e Híbrida. O dicionário de unidades está em fase final de definição. O *parser* e o módulo de entonação do gerador automático de prosódia se encontram em elaboração.

Ao associar processamento de sinal sofisticado a processamento lingüístico de alta qualidade estamos concebendo um sistema de síntese da fala que prima pela modularidade (simplificando a modificação de módulos específicos para diversos fins e simplificando a modernização do sistema) e pelo automatismo. Estas características permitirão o teste de teorias lingüísticas ou o aprimoramento de técnicas de processamento do sinal. Esta versatilidade torna nosso sistema uma ferramenta indispensável para lingüistas e engenheiros.

5- Referências Bibliográficas

[1] Moulines, E., Charpentier, F. (1990) *Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones*. Speech Communication, 9 (5/6), 453-467.

[2] Böeffard, O., Violaro, F. (1994) *Improving the Robustness of Text-to-Speech Synthesizers for Large Prosodic Variations*. Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, USA, 111-114.

[3] Böeffard, O., Violaro, F. (1994) *Using a Hybrid Model in a Text-to-Speech System to Enlarge Prosodic Modifications*. International Conference on Spoken Language Processing (ICSLP '94), Yokohama, Japan, 727-730.

[4] Klatt, D.H (1987) *Review of text-to-speech conversion for English*, J. Acoust. Soc. Am. 82, 737-793.

[5] Fowler, C. (1981) *Production and perception of coarticulation among stressed and unstressed vowels*, J.S.H.R., 47, 127-139.

[6] Öhman, S. (1966) *Coarticulation in VCV utterances*, J. Acoust. Soc. Am., 39, 151-168.

[7] Albano, E.C., Rossi, A.J.G. & Silva, A.H.P. (em preparação) Coarticulação de vogal a vogal em português.

[8] Ferreira, A. (1977) *Minidicionário Aurélio*. Rio de Janeiro: Nova Fronteira.

[9] Barbosa, P.A. & Bailly, G. (1994a) *Characterisation of rhythmic patterns for text-to-speech synthesis*, Speech Communication. 15 (1-2), 127-137.

[10] Campbell, N.W. (1992) *Syllable-based segmental duration*. In: Talking Machines: theories, models, and designs (Bailly, G. & Benoît, C. Eds.) 211-224.

[11] Barbosa, P. A. (1996) *At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration*. First ESCA-TRW on Speech Production Modeling, 20-25 de Maio, Autrans, França.

[12] Barbosa, P. A. (1994) *Caractérisation et génération automatique de la structuration rythmique du français*. Thèse de troisième cycle. INP de Grenoble, França.

[13] Schäfer-Vincent, K. (1983) *Pitch Period Detection and Chaining: Method and Evaluation*. Phonetica, 40, 177-202.

[14] Laroche, J., Stylianou, Y., and Moulines, E. (1993) *HNS: Speech Modification Based on a Harmonic + Noise Model*. Proc. IEEE ICASSP, Vol. II, 550-553.

[15] Marques, J.S., and Almeida, L.B. (1989) *Sinusoidal Modelling of Voiced and Unvoiced Speech*. Proc. Eurospeech, Vol. II, 203-206.

[16] Quatieri, T.F., and McAulay, R.J. (1992) *Shape Invariant Time-Scale and Pitch Modification of Speech*. IEEE Trans. ASSP, Vol. ASSP-40, (3), 497-510.

Generation of pauses within the z-score model

Plínio Barbosa[†] & Gérard Bailly
Institut de la Communication Parlée, U.R.A.CNRS n° 368, INPG/Univ. Stendhal
46, av.Félix Viallet, 38031 - Grenoble CEDEX 1, France

Introduction

Many models of generation of segmental duration are based on statistical characterization of *ad hoc* corpora. A large set of coefficients accounting for rate, context, mode and place of articulation effects is obtained from this kind of description [5, 3, 10].

In these models pause is generally considered separately. For example, in Klatt's model pause durations are given *a priori* and not integrated in the same mechanism for generating segmental durations: "*Rule 1:* Insert a brief pause before each sentence-internal main clause and at other boundaries delimited by an orthographic comma." [5, p. 761] The models proposed by Bartkova & Sorin for French and by O'Shaughnessy for Canadian French modify also an inherent (or intrinsic) segment duration following modifications induced by phonemic environment. Their models also take a-priori durations for pauses that are placed according to phonotactic and syntactic criteria. In this kind of generation, corpus dependent results may be easily obtained due to exhaustive computation of specific context effects. The main problem with these models is the choice of the phoneme as the basic unit for duration generation. The choice of a higher level unit as the foot or the syllable (see [4] for a review) avoids the tuning of a large set of coefficients (about 63 in Bartkova-Sorin's model!) and an easier integration of duration in the prosodic structure.

A high-level basic unit for duration is presented here: the inter-Perceptual-center group (IPCG). This unit has allowed us to elaborate a model integrating the automatic placing and calculation of pauses. This model is guided by rhythmic constraints in a coherent and homogeneous framework.

The Perceptual-center

Research developed by Marcus and his colleagues [9, 7] on Perceptual-center (PC) show that listeners perceive regularity in syllable sequences despite important differences among absolute durations: "we were forced to ask ourselves what it was that was regular in a rhythmic list. To simplify our discussions we defined this as the P-center of each item."[9, p. 405] The PC does not seem to correspond to any simple articulatory or acoustic correlate but there is a clear interaction between speaker production and perception systems in order to produce speech sequences that the listener perceives as isochronous. This tendency to isochrony seems to be a major characteristic of the rhythmic control system as attested by the discussion on isochrony [6].

Experiments carried out by Marcus [8], Pompino-Marschall [11] and Scott [12] have shown that despite the nature of experimental conditions (perception or production experiments) and the nature of phonemes, the PC seems to be located at the vicinity of vocalic onset. Although their experiments do not concern connected speech, we have set the location of the PC at the vocalic onset, when there is no silence before the syllable. If there is a silence before it, the PC is usually placed early in the syllable[1]:

[†] Supported by CNPq-Brasilia/Brasil.

[1] Scott has implemented a model for PC location based on the first derivative of the intensity function of the acoustic signal. According to her model, significant increases of intensity in the band frequency from 195 to 1638 Hz strongly determines the PC location.

We have taken the beginning of the leftmost voiced consonant of the syllable onset as the PC. Two consecutive PCs define the boundaries of the inter-Perceptual-center interval (IPCI). The segments it contains form the IPCG.

The Campbell model

The elasticity principle [4] in its strongest version says that all segmental durations in a syllable frame are obtained by a same and single factor k —the so-called z-score or normalized duration [13]— as follows: $Dur_i = \exp(\mu_i + k\sigma_i)$ (1), where μ_i and σ_i are the mean and standard deviation of the log-transformed durations (in ms) of the realizations of the phoneme i. These k-factors are computed over the syllable by: $\sum_i Dur_i = syllable\,duration$ (2). The successive use of formulas 2 and 1 is referenced as the repartition model. Campbell's model proceeds in two steps: (a) prediction of the syllable duration from phonotactical and phonological information by a statistical model (a multilayer perceptron); (b) application of the repartition model [4].

We have already demonstrated the optimal consistency of the rhythmic behaviour of the segmental units inside an IPCG [1]: the k-factors of the syllable onset are positively correlated with the preceding rime whereas they are negatively correlated with the following nucleus. If the k-factors of all IPCGs of a sentence are represented graphically, temporal organization of these units exhibits monotonous ascending movements towards each phrase accent with resetting just after accent realization. These patterns can be easily captured by dynamical nonlinear predictors synchronously with F0 movements.

The Barbosa-Bailly model

A five-rate[2], 20-sentence corpus was recorded to study the influence of pause emergence on the overall rhythmic structure. In order to simplify the problem of locating PCs, all sentences of the corpus are CV sequences (all pauses were preceded by a vowel).

As in the Campbell model, our duration predictor proceeds in two stages. Nevertheless, significant differences have to be noticed: (a) a sequential network constrained by an internal clock (a measure of each utterance rate represented by the mean of the non accentuated IPCIs of the utterance [2]) generates timing of PC locations; (b) the IPCIs are then shared out among the IPCG constituents according to the repartition model which presently includes the emergence of pauses.

Incorporating pause phenomenon to the repartition model

In order to know the amount of silence duration that must be assigned to each IPCI in the generation stage, we have studied the relation between the actual k-factors of vowels: $k_s = (\ln(Dur_{vowel\,V}) - \mu_V)/\sigma_V$ and the virtual k-factors of the same vowel added to an eventual adjacent silence duration: $k_p = (\ln(Dur_{vowel\,V} + Dur_{adj.\,silence}) - \mu_V)/\sigma_V$. The critical k is the point where the regression line, computed for $k_s \neq k_p$[3], crosses the straight line $k_s = k_p$ (see figure 1, where *critical* $k = 0.83$).

A minimum silence duration was set for each speech rate as a result of analysis carried out on the corpus (see Table 1).

The repartition algorithm proceeds as follows: (a) computation of the k-factor for a given IPCG (k_g); (b) if k_g is greater than the critical k, the corresponding k_s (sound part) is obtained by the regression formula (by

[2] Effective separated rates were obtained by controlling the speaker's utterance with five sets of synthetic questions obtained roughly using our text-to-speech system. We thank especially Frédéric Bimbot who offered his voice.

[3] Equation: $(k_s + 5) = (k_p + 5)^{0.59}\exp(0.72)$

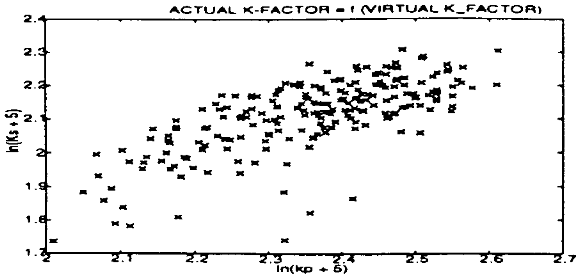


Figure 1: Scatterplot of log-transformed actual k-factors versus log-transformed virtual k-factors for all speech rates.

Table 1: Number of errors obtained by the repartition algorithm in placing the silences. The number of actual silences by speech rate in the natural utterances is also given (there are 285 IPCGs by rate in the corpus).

rate	clock (ms)	min. pause (ms)	loc. errors	actual silences
v. slow	360	75	15	99
slow	325	67	13	64
normal	270	57	11	32
fast	210	51	-	26
v. fast	165	59	1	5

setting $k_p = k_g$); (c) the segmental durations are computed with the repartition model and added up. The difference between this result and the IPCG original duration is assigned to the silence; (d) if the silence duration is greater than the minimum the procedure is over; (e) if not, no silence is inserted and the k-factor to be used to compute the corresponding segmental durations is k_g; (e) If k_g is lesser than the critical k, the segmental durations are computed using k_g as the k-factor and no duration is assigned to the silence.

It is important to note that no constraints on location —like accentuable position— are imposed to the silences: they are placed as a result of the repartition algorithm described above.

Using original IPCG durations, only few silences were placed in such positions not actually chosen by the speaker (see table 1) but all positions were assigned to a latent location for accent realization (a prosodic marker).

As described above, a sequential network was trained in order to generate the successive IPCIs for each sentence of the corpus. Fifty sentences (ten sentences in five rates) from the corpus (learning set) were chosen for the learning phase. The network generalizes with more or less success the corresponding IPCI sequences for all sentences in the corpus (compare standard deviations of errors between learning and test sentences in table 2). Synthetic segmental and silent durations were obtained by applying the repartition model described above.

The errors between the original segmental durations and the ones obtained by our model were computed for each speech rate and two types of unit: realizations of phonemes and silences (see table 2). The histograms of these errors for all learning-set sentences were strongly correlated with the normal distribution (minimum 95 %).

Table 2: Mean (and standard deviations) in milliseconds from the histograms of errors between the segmented and generated durations for learning-set and test-set sentences.

	learning set		*test set*	
rate	sil.	seg.	sil.	seg.
v.sl.	-41(137)	-1(49)	3(207)	4(56)
sl.	70(116)	-2(40)	-82(147)	6(46)
nm.	67(122)	0(37)	-105(113)	5(43)
ft.	-189(84)	3(30)	64(144)	0(28)
v.ft	-4(170)	2(25)	-49(193)	-1(23)

We conducted an experiment in order to test the perceptual relevance of IPCGs organization.The stimuli were obtained by an analysis-resynthesis technique where only the segmental durations were modified. Two versions of each utterance were compared: the first one (tagged as *model*) is paced by our model and the second (tagged as *random*) is obtained by modifying the original segmental durations according to a random amount from a normal distribution with the same mean and standard-deviation of the associated histogram of errors (see table 2).

Perception test

Fifteen subjects[4] working at *ICP* participated in this perception experiment. Each session lasted between 7 and 10 minutes. In a test ABBA the subjects were asked to select the most natural utterance. Listening may be repeated once. A question mark could be used if both utterances seem similar to the subject. All subjects considered, 89% of *model* utterances were chosen as being the most natural. In 15% of the answers there was a doubt between the utterances in the pair. Individual scores can be seen in Fig. 2.

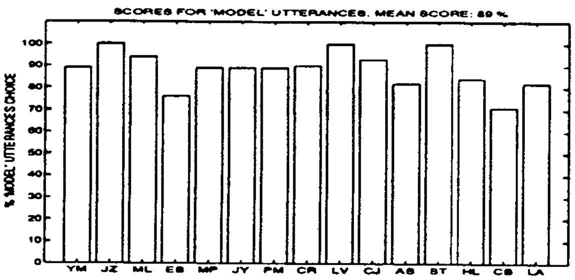


Figure 2: Individual scores for the choice of "model" utterances. Please note that inter-subject results are similar.

Some comments: "When the utterances continuous without interruption they seemed natural to me" (HL), "Utterances with rhythm change at the middle of a word did not seem natural to me!" (LV).

Discussion

Subjects' preference for an utterance in the pair involves judgements of global aspects of rhythm (see two comments above). In both stimuli the nature of the segments does not seem unnatural to them.

Results seem to show that the clear preference for the utterances obtained with our paradigm claim for a model that maintains the basic rhythmic structure of natural speech by using a new higher-level rhythmic programming unit: the IPCG.

References

[1] Barbosa, P. and Bailly, G. Generating segmental duration by p-centers. In Auxiette, C., Drake, C., and Gérard, C., editors, *4th Rhythm Workshop: Rhythm Perception and Production*, pages 163–168. Bourges, France, 1992.

[2] Barbosa, P. and Bailly, G. Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, in press.

[3] Bartkova, K. and Sorin, C. A model of segmental duration for speech synthesis in French. *Speech Communication*, 6:245–260, 1987.

[4] Campbell, W.N. *Multi-level Timing in Speech.* PhD thesis, University of Sussex, 1992.

[5] Klatt, D.H. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.*, 82:737–793, 1987.

[6] Lehiste, I. Isochrony reconsidered. *Journal of Phonetics*, 5:253–263, 1977.

[7] Marcus, S.M. Perceptual centres. Unpublished PhD thesis, Cambridge University, 1976.

[8] Marcus, S.M. Acoustic determinants of Perceptual center (p-center) location. *Perception and Psychophysics*, 30(3):247–256, 1981.

[9] Morton, J., Marcus, S., and Frankish, C. Perceptual centers (p-centers). *Psychological Revue*, 83(5):405–408, 1976.

[10] O'Shaughnessy, D. A study of French vowel and consonant durations. *Journal of Phonetics*, 9:385–406, 1981.

[11] Pompino-Marschall, B. On the psychoacoustic nature of the p-center phenomenon. *Journal of Phonetics*, 17:175–192, 1989.

[12] Scott, S.K. *Perceptual Centres in Speech - An Acoustic Analysis.* PhD thesis, University College London, 1993.

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

[4] We thank our listeners who offered their ears to the perception task.

Characterisation of rhythmic patterns for text-to-speech synthesis

Plínio BARBOSA and Gérard BAILLY
Institut de la Communication Parlée
U.R.A. CNRS n° 368 INPG/ENSERG - Université Stendhal
46, av. Félix Viallet - 38031 Grenoble CEDEX 1 France

Abstract. This article proposes an alternative rhythmic unit to the syllable: the inter-perceptual-center group (IPCG). This group is delimited by events which can be detected using only acoustic correlates (Pompino-Marschall 1989). The rhythmic patterns for French are described using this characterisation: we show that realisation of accents is gradual over the trailed accentual group and that this gradual lengthening is needed for perception. A model of repartition of the IPCG duration among its segmental constituents incorporating automatic generation of pauses (emergence and duration) according to speech rate is then described.

Zusammenfassung. In diesem Artikel wird eine Alternative für die Silbe als rhythmische Einheit vorgeschliegen: Die inter-perceptual-Center Gruppe (IPCG). Die Begrenzung der Gruppe ist durch Ereignisse gegeben, die ausschließlich durch akustische Korrelate entdeckt werden (Pompino-Marschall 1989). Die Rhythmusmuster auf Französisch werden unter Verwendung dieser Charakterisierung beschrieben: Wir zeigen, daß die Umsetzung von Akzenten entlang der akzentualen Gruppe langsam verläuft und eine ständige Verlängerung der Gruppe für die Wahrnehmung nötig ist. Es wird ein Modell beschrieben, worin die Dauer der IPCG der Dauer der Realisationen einzelner Phoneme zugewiesen wird, welche die Gruppe bilden. Dieses Modell schließt die Erzeugung der Pausen (Auftreten und Dauer) entsprechend der Sprachrate ein.

Résumé. On propose dans cet article une unité rythmique alternative à la syllabe: le groupe inter-perceptual-center (GIPC). Ce groupe est délimité par des événements qui peuvent être caractérisés par des corrélats purement acoustiques parallèlement au décodage acoustico-phonétique (Pompino-Marschall 1989). Les patrons rythmiques du francais sont ainsi décrits: la réalisation de l'accent est graduelle au long du groupe accentuel et cet allongement progressif de la durée est nécessaire à la perception. On décrit un modèle de répartition de la durée du GIPC entre les réalisations des phonèmes qui le composent. Ce modèle incorpore la génération de la pause (son émergence et sa durée) selon le débit souhaité.

1 Introduction

Concatenative speech synthesis offers the most efficient solution to the problem of dealing with large coarticulation effects which are extremely difficult to describe using synthesis-by-rule systems. The search for high intelligibility and naturalness of the resulting synthetic speech leads researchers to use larger and larger synthesis units (Takeda, Abe & Sagisaka 1992; Hauptmann 1993). This trend is not strongly reflected in the models of prediction of prosody: the synthesis-by-rule approach is widely accepted for generating melodic trajectories and segmental durations. The most common melodic models describe the melodic curve as a discrete sequence of targets connected by interpolating functions (t' Hart & Collier 1973; Pierrehumbert 1981; Hirst & di Cristo forthcoming) which are easily related to a phonological surface structure (see for example the HL model in Pierrehumbert 1980). Inductive methods may be used to learn situation- and speaker-specific strategies. Except some attempts cited below, the statistical methods used take the syllable as the basic unit: f0 "movements" are generated by appropriate phonotactical information given as input, e.g. the position of the syllable in the accentual group (Scordilis & Gowdy 1989; Sagisaka 1990). The models of generation of segmental durations are even more local and generally use the phoneme as a basic unit: complex rules eventually trained on large corpora using statistical methods (Bartkova & Sorin 1987; van Santen & Olive 1990) combine co-intrinsic and linguistic factors into additive (O'Shaughnessy 1984) or multiplicative models (Klatt 1976).

Few studies have tried to generate prosody using globa, methods. Most of them have been dedicated to the generation of melody using retrieval of melodic "contours" in a hierarchical lexicon indexed by linguistic keys (Aubergé 1992) or using dynamical non-linear predictors (Ljolje & Fallside 1986; Traber 1992) with powerful learning abilities.

This study provides evidence of the possibility of automatic generation of segmental durations using a new rhythmic programming unit, i.e. the inter-P-center group. When these units are represented by a normalised lengthening factor, they generally exhibit a monotonic increase with resetting instants that can be easily generated by dynamical non-linear predictors synchronously with f0 movements.

2 About the nature of the rhythmic programming unit

The most widely accepted prosodic unit is the syllable. The quasi-periodic movement of the jaw has earlier given prominence to the role of the syllable in the rhythmic organisation of the speech both in perception and production (Fraisse 1974). As mentioned by Daniel Hirst (1993, p. 34): "Most of the arguments in favour of the syllable as a unit are really arguments in favour of syllabicity rather than syllabic constituents (...)". In this section we examine an alternative candidate for the rhythmic programming unit.

2.1 The Perceptual-center

The work developed by Marcus (1976) on Perceptual-centers (PC) shows that listeners are capable of adjusting consistently the delay between two syllabic occurrences (V versus CV) to perceive such an alternation as isochronous. This tendency to isochrony seems to be one of the components of the rhythmic control system as attested by the discussion about isochrony and isosyllabism (Pike 1945; Lehiste 1977; Wenk & Wioland 1982).

Lehiste gives a brief overview of the search for absolute isochrony in English speech production: Shen & Peterson (1962) reject this notion; O'Connor (1965) expressed doubt on the existence of perfect isochrony. This strong version of the notion is definitely abandoned by Lea (1974): "the average time intervals appeared to increase almost linearly with the number of intervening syllables." (Lehiste 1977, p. 254) These results are corroborated more recently, by Nooteboom (1991) or Fant, Kruckenberg & Nord (1991). Nevertheless, according to Lehiste, isochrony exists, but only under favourable circumstances (she quotes Classe, 1939) as similar phonetic structure of the component syllables, grammatical structure and grammatical connection between the inter-stress groups. Although acoustic phenomena argue for a tendency towards isochrony, such as the trend to the preservation of word durations, she is in favour of a perceptual quest for isochrony: listeners seem to impose a rhythmic structure on sequential stimuli and then "we may hear sequences of only approximately equal time intervals as more equal than they really are." (1977, p. 258) This approach is supported in a more recent contribution by Fant and his colleagues: "Stress timing is not a matter of physical isochrony of inter-stress intervals but a perceptual dominance of heavy syllables, the succession of which is sensed as quasi-periodical." (Fant, Kruckenberg & Nord 1991, p. 363) Marcus proposes a perception model which accounts for this quasi-periodicity by defining the PC.

Pompino-Marschall (1989) describes an experiment designed to compute the absolute position of the PC. His stimuli consisted of five synthetic syllable continua (S) presented to the subjects via headphones alternating with a click (5 ms, 1 kHz tone burst: C) with an overall tempo of 120 signals per minute: C-S-C-S-C or S-C-S-C-S. The subjects had to adjust the temporal alignment of the click sequence relative to the test syllable sequence to perceive isochrony by turning a potentiometer knob. The time instant bisecting the duration between two successive clicks measured relative to the acoustical syllable onset was taking as an indicator of the PC location. Despite the diversity of the consonants the PC location seems to be at the neighbourhood of the vocalic onset in either production or perception experiments. His results also show that the distance between two consecutive PCs probably determines the perception of momentary tempo.

Other recent work (Howell 1988), however, places the PC at the location of a significant change in the derivative of the intensity function. Thus the PC would shift left of vocalic onset if a silence precedes the CV syllable. That is why in our work the PC location is fixed at the vocalic onset if there is no silent pause before the the current syllable in the connected speech and at the syllable onset when there is a silence preceding the current syllable. The distance between two adjacent PCs has been called the inter-PC interval (IPCI) and the segmental units it contains form the inter-PC group (IPCG).

In the following we will demonstrate the optimal consistency of the rhythmic behaviour of the segmental units inside a IPCG.

2.2 Campbell's model

The elasticity principle (Campbell & Isard 1991) in its strongest version says that all segmental durations in a syllable frame are obtained using a single factor k in the formula

$$Dur_i = \exp(\mu_i + k.\sigma_i) \tag{1}$$

In this equation, μ_i and σ_i are the mean and standard deviation of the log-transformed durations (in milliseconds) of the realisations of the phoneme i. In the second step of generation, a simple algorithm permits the calculation of segmental durations by computing the k-factor from:

$$\sum_i Dur_i = syllable\,duration \tag{2}$$

This sum is over all segments in the syllable. This model of repartition of the syllable duration among its segmental constituents simplifies the automatic generation of segmental duration which then proceeds in two steps (Campbell 1992): (a) prediction of the syllable duration from phonotactical and phonological information by a statistical model (a multilayer perceptron); (b) application of the repartition model.

It is worth noting that Campbell proposes further modification of the algorithm (Campbell 1992, p. 218): "This is currently implemented with a use of a decay variable that reduces the effect of any lengthening on segments earlier in the syllable (...)". This means that the onset receives a weaker lengthening factor k than the rime. This is confirmed experimentally by our own work (Barbosa & Bailly 1992a) showing that the k-factors of the onset are positively correlated with the preceding rime and they are negatively correlated with their own rime.

2.3 Computing the elasticity characteristics for our reference speaker

The elasticity characteristics of the above model μ_i and σ_i are computed using a large corpus of natural speech (200 phonetically balanced sentences in Campbell 1992). As our model of generation of prosody will be part of a text-to-speech system using concatenative synthesis (Bailly, Barbe & Wang 1992), we used the database of 2012 labelled logatoms designed to gather our "polysounds". The resulting elasticity characteristics for our reference speaker are listed in table 1. Notice that this table captures speaker's articulatory habits such as the unusual lengthening of the realisations of the /s/.

a	115(38)	ɛ	109(32)	e	118(41)	i	101(33)	œ	107(36)	ø	116(29)
y	105(30)	ɔ	110(32)	o	125(34)	u	108(32)	j	94(27)	w	95(25)
ɥ	92(20)	ɑ̃	136(42)	ɔ̃	138(42)	ɛ̃	134(39)				
p	125(38)	t	128(39)	k	114(35)	b	92(22)	d	94(25)	g	87(23)
f	143(32)	s	163(36)	ʃ	144(26)	v	106(30)	z	116(25)	ʒ	104(26)
m	104(24)	n	102(24)	ɲ	110(26)	R	94(30)	l	93(27)		

Table 1: Means (and standard deviations) of the durations (in ms) of the French phonemes for our speaker.

3 Characterisation of rhythmic patterns

3.1 The corpus

We recorded a corpus of 88 sentences pronounced at two speech rates (slow and normal[1]) by our reference speaker. Segmental durations were determined semi-automatically using the *Temporal Decomposition* algorithm described in (Bailly, Barbe & Wang 1992): first, centres of phonemes are marked by hand and then, segmental boundaries are set automatically. In order to study the relation between rhythmic and linguistic structure, this corpus was labelled using a simple hierarchical model of prosodic marking (Bailly 1986).

3.2 The rhythmic patterns

The sequences of k-factors computed for each IPCG can be easily segmented into prosodic contours characterised by:

[1] The speaker was initially instructed to utter at three different rates i.e. slow, normal and fast. But no significant variation in tempo was found between normal and fast! This justifies a posteriori the use of synthetic stimuli in Section 4.

(a) monotonically increasing durational patterns marking prosodic boundaries: the end of the patterns always coincide with a prosodic marker and the amount of target lengthening correlates with the supposed strength of the marker[2] (see Table 2). Please note that the hierarchy of markers induced by mean target lengthening also holds for the pause insertion rate;

(b) resetting near to zero after accent realisation;

(c) these contours often exhibit a quasi-exponential increase[3]: these rhythmic "movements" are generated with a positive acceleration of k-factors towards the target lengthening.

	Marker				
	ID	IT	RD	SD	LD
k-values	4.4 (0.9)	3.6 (1.1)	1.7 (0.9)	1.3 (0.9)	0.4(0.4)
pauses (%)	96 (26/27)	92 (106/115)	28 (23/81)	19 (17/91)	-

Table 2: Mean values (and standard deviations) of k-factors versus the prosodic marker (associated with a prosodic boundary) for the slow rate. The percentage of boundaries followed by a pause is also given (with the number of actual occurrence).

The second and third properties mentioned above may be associated with both syllable and IPCG, but the first is an essential property of the IPCG and not of the syllable (Fig. 1).

Two questions were then asked: (1) are continuously increasing patterns needed for the perception of accentuation or they are just an artifact of production constraints? (2) Is this typical configuration needed to the perception of any kind of isochrony in French connected speech? A perception experiment was carried out to answer these questions.

4 The perceptual relevance of rhythmic patterns

4.1 Method

A preference test was designed to show the perceptual relevance of this production phenomenon (Viviani & Stucchi 1991). Ten pairs of sentences were generated using a high-quality analysis-resynthesis system (Moulines 1992; Bailly, Barbe & Wang 1992): the A configuration was obtained by calculating the successive k-factors of all IPCGs in a sentence. In the B configuration only the k-factors ending an accentual group (cued by a marker) are kept to their original values whereas all others were set to 0. Segmental

[2] The markers were defined by Bailly (1989) as follows: *Independence* (**ID**) for strong phrase boundaries, e.g. *M. Tapie, (ID) rappelez M. Dupont jeudi.* (*Mr. Tapie, call Mr. Dupont on Thursday.*). *Interdependence* (**IT**) as in enumerations, e.g. *Pierre, (IT) Jean (IT)et Marie boivent du lait.* (Pierre, Jean and Marie drink milk.). *Right dependence* (**RD**) for right dependence, e.g. *Le chat (RD) noir boit du lait.* (*The black cat drink milk.*). *Strong right dependence* (**SD**) as between verbe and complement, e.g. *Le chat noir boit (SD) du lait.. Left dependence* (**LD**) for left dependence, e.g. *Le petit (LD) chat joue.* (*The little cat plays*). *Strong left dependence* (**SL**) as between subject and verbe, e.g. *Papa (SL)mange.* (*Dad eats.*

[3] The tendency of the gradual lengthening over the accentual group in French has been already mentioned in (Vaissiere 1980; Touati 1987; Pasdeloup 1992)

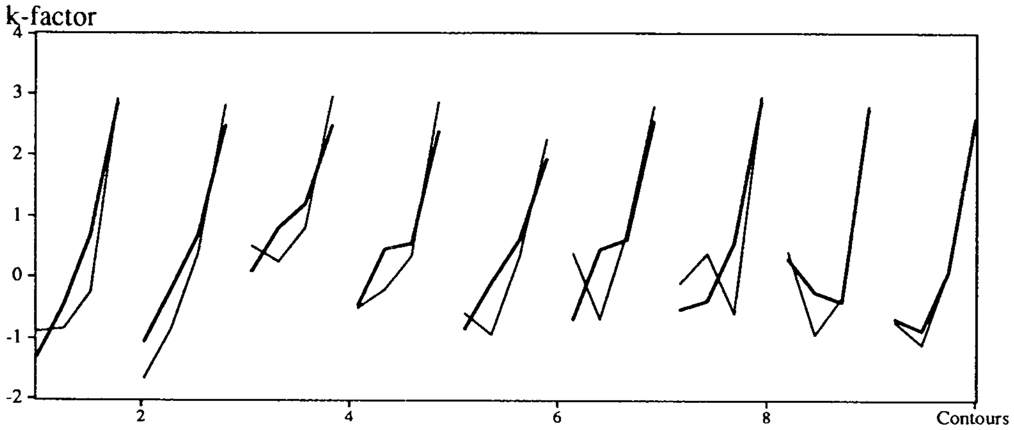


Figure 1: Superposition of k-factor evolution for two kinds of group: the syllable (thin lines) and the IPCG (thick lines). The k-factors are discrete values but they were attached to each other for ease of visibility. Note that the third, the fifth, the sixth and the seventh thick contours (IPCG) are monotonic and their syllable counterparts are not. In the other contours the evolution of the k-factor for the IPCG is less abrupt. All these selected contours have 4 syllables and end with the ID mark.

durations in the two configurations A and B were then computed by applying the repartition algorithm. Note that in A (a) segmental durations are different from the ones in the natural sentences since k is averaged in each IPCG and (b) on the other hand, the timing of the successive vocalic onsets is identical in both cases (natural sentences and A) since IPCG durations are the same. Silent pause durations and all other parameters were unchanged. An example of computed k-factors for the sequence of IPCGs (A configuration) and the sequence of k-factors to be tested (B configuration) of an utterance is shown in Fig. 2.

Eleven subjects working in the *Institut de la Communication Parlée* but not in the synthesis domain participated in this perception experiment. Each session lasted between 10 and 15 minutes. In an ABBA test the subjects were asked to select the most natural utterance. They were allowed to listen a maximum of two times to the stimuli. At the end of the two possible repetitions they could use a question mark to indicate that they were not able to decide.

4.2 Results

Considering all the subjects, the scores are 65% for A, 20% for B and 15% undecided. If only the A and B choices are taking into account, 77% of the preference is for A (65 × 100/(65 + 20) = 77%). Individual scores can be seen in Fig. 3.

There were no significant effect of presentation order. All listeners agree on the difficulty of the task: "During the experiment I thought I changed the criterion of utterance choice. In the beginning, I chose the ones that were more constant concerning the rhythm" (JL), "I chose utterances that were the most constant at rhythm level." (OD), "Are you sure they are not identical?" (JY) !!

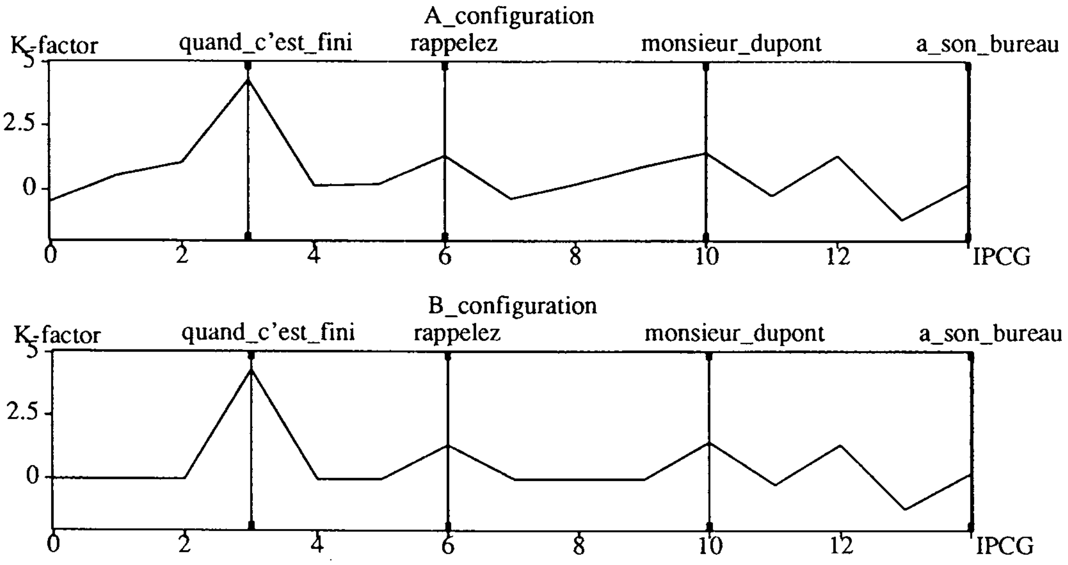


Figure 2: Example of A and B configurations for the sentence: "Quand c'est fini, rappelez Monsieur Dupont à son bureau." Discrete values of k are connected by lines for the sake of visibility. Prosodic groups are indicated by vertical bars. The k of the last IPCG of the utterance is taken equal to the k-factor of the vowel.

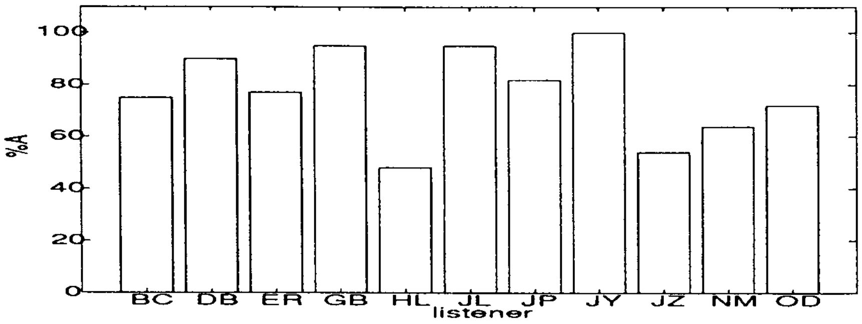


Figure 3: Mean scores (across presentation order) of A selection by subject.

4.3 Discussion

The clear general preference for the gradual rhythmic pattern leads us to think that this configuration is necessary for the perception of an accent: a lack of lengthening in previous IPCGs sounds abrupt. The *internal clock* (see below) hypothesis may explain the subjects' perceptual behaviour : shorter IPCIs are cues to an unexpected local acceleration of speech. Gradual lengthening contributes to the perception of isosyllabicity (Duez 1987; Lehiste 1977). But there is no implicit conclusion that human beings use k-factors to control rhythmic patterns.

5 Incorporating the automatic generation of pauses

5.1 The corpus

A five-rate, 20-sentence corpus was recorded in order to evidence the influence of pause insertion on the rhythmic structure. In order to study the PC phenomenon all sentences of the corpus are CV sequences (all pauses were preceded by a vowel). The analysis of this second corpus confirms for all rates the general trend of rhythmic patterns: 80% of them are monotonically increasing sequences of k-factors.

This experiment aims at developing an automatic generation of duration including the pause phenomenon. To avoid overlapping of speech rates, as noted in section 3.1 the speaker was asked to answer synthetic interrogative utterances with predefined sentences. These synthetic utterances were obtained roughly using our text-to-speech system by multiplying both μ_i and σ_i by a phonation factor (Wightman, Shattuck-Hufnagel, Ostendorf & Price 1992).

5.2 Characterising speech rate by an internal clock

Our basic assumption of how PC beats are generated and perceived is that the IPCIs are obtained by a deformation from an *internal clock* (Allen 1975). This biological clock is the reference for isochrony and the release mechanism of the motor actions (Turvey, Schmidt & Rosenblum 1990). Results of studies on coordination between repetitive tapping and periodic external stimuli tend to reinforce this assumption (Allen 1975; Semjen, Schulze & Vorberg 1992). This internal clock would be a rhythmic attractor: the accentual movement would consist in indicating that "something has happened" as pointed out by the phase difference between the internal clock and the *PC clock*, constituted by the sequence of PC beats (the motor realisation cycle). Pause phenomena will ease the resynchronisation of these two clocks (Fant & Kruckenberg 1989).

5.3 Analysing pause insertion

Active control of the clock duration is an attractive way for characterising speech rate. But the choice of an internal clock duration is particularly delicate. How can an unaccented unit be chosen if lengthening is gradual over the prosodic group? Taking into account that: (a) the last IPCG in the prosodic group is clearly the main lengthened unit; (b) if there is a silent pause, we cannot separate silence and sound intervals : they are elements of the same phenomenon (Duez 1987); (c) the first IPCG of a prosodic group is often shortened, the internal clock durations are computed for each utterance as the mean among the nonterminal (excluded prepausal) IPCGs.

This characterisation of speech rate aims to analyse the emergence and durations of actual pauses according to prosodic structure and speech rate. Although most phonological models would assume that the prosodic structure remains the same despite rate changes (Monnin & Grosjean 1993) there are no clear proposals for modelling pause emergence. To our knowledge all text-to-speech systems generate pauses at a phonological level according to ad-hoc phonological and phonotactical constraints. We think that pause emergence has to be solved at the performance level and thus is partly speaker-specific as the parameters of the elasticity model.

In order to know the amount of silence duration that must be assigned to each IPCI in the generation stage, we have studied the relation between the k-factors of vowels (k_s) and the k-factors of the same vowel added to the adjacent silence duration (k_p)[4]. We have plotted the values $k_s = f(k_p)$: there is a effective silence after the vowel (see Fig. 4) where k_s is different from k_p. The critical k is the point where the regression line, computed only for $k_s \neq k_p$, crosses the straight line $k_s = k_p$ (here *critical* $k = 0.83$). The regression line for $k_s \neq k_p$ of the log-transformed k-factors has the following equation:

$$(k_s + 5) = (k_p + 5)^{0.59} \exp(0.72) \qquad (3)$$

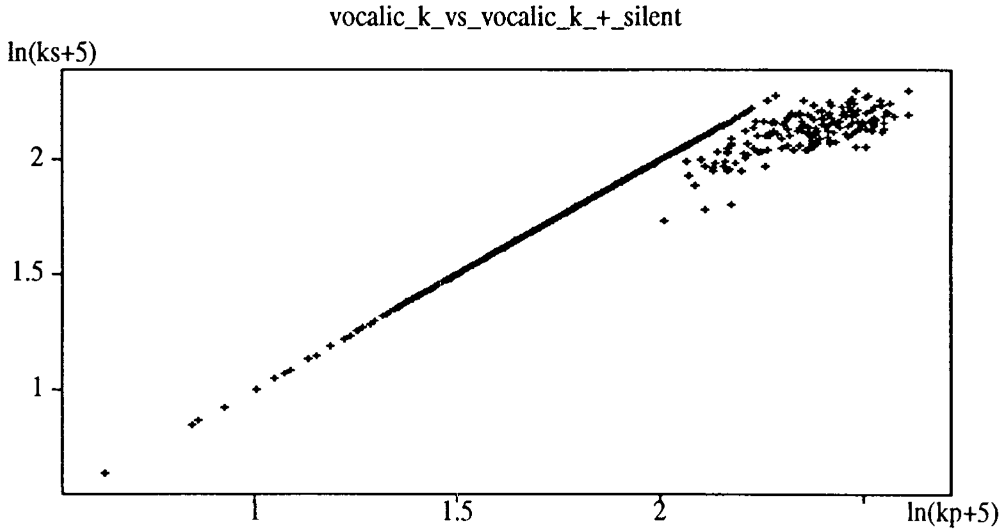


Figure 4: Scatterplot of log-transformed k-factors of vowel versus vowel+silence for all speech rates.

5.4 Generating sound and silence durations

A minimum silence duration was set for each speech rate as a result of analysis on the corpus (see Table 3).

The repartition algorithm proceeds as follows: (a) computation of the k-factor for a given IPCG (k_g); (b) if k_g is greater than the critical k, the corresponding k_s (sound part) is obtained by the regression formula 3 (by setting $k_p = k_g$); (c) the segmental durations are computed by the formula 1 and added up. The difference between this result and the

[4] If there is no silence, we have set *silence duration* $= 0$.

IPCG original duration is assigned to the silence; (d) if the silence duration is greater than the minimum the procedure is over; (e) if not, no duration is assigned to the silence and the k-factor to be used to compute the corresponding segmental durations is k_g; (e) If k_g is less than the critical k, the segmental durations are computed using k_g as the k-factor and no duration is assigned to the silence.

rate	clk.	*Vowels*			*Consonants*			min.	*Silent pauses*		
	(ms)	SY	VO	PC	SY	VO	PC	(ms)	SY	VO	PC
v. slow	360	34(32)	35(32)	34(29)	43(46)	42(43)	37(46)	75	67(58)	76(62)	55(53)
slow	325	38(40)	42(41)	36(32)	32(32)	34(31)	28(26)	67	73(52)	73(52)	73(56)
normal	270	26(26)	30(38)	29(23)	27(26)	28(26)	25(22)	57	54(37)	64(36)	56(35)
fast	210	17(19)	18(19)	19(20)	21(25)	21(27)	19(21)	51	75(77)	68(99)	49(78)
v. fast	165	12(12)	11(13)	12(14)	13(13)	12(17)	11(12)	59	59(33)	93(28)	44(20)

Table 3: Means (and standard deviations) of prediction errors of three sound classes for different rhythmic units using the modified Campbell's model. Average clock durations (clk.) computed with all IPCIs in non accentuable position show that rates are effectively different (tests of significancy have done $p < 10^{-4}$). The units are: the syllable (SY), the inter-Vocalic Onset group (VO) and the inter-Perceptual-center group (PC). The thresholds are used in the determination of silent pauses.

Results of segmental duration generation were computed for three different rhythmic units (see Table 3). Prediction errors are similar between the rhythmic units for vowels but for the prediction of consonants and silence durations the PC unit seems to give better results: smaller mean and coefficient of variation ($100 \times standarddeviation/mean$)values than for SY and VO groups (see differences in very slow, fast and very fast speech rates for example).

5.5 Evaluation

It is important to note that no constraints on location — like accentuable position — are imposed on the silences: they are placed as a result of the repartition algorithm described above.

Only a few silences were placed in such positions not chosen by the speaker(see Table 4) but all positions were assigned to a latent location for accent realisation (a prosodic marker). The only silence inside a word (in the slow rate) was after the syllable /sɔ/ in /sɔmɛ/ ("sommet"). After having verified that the /m/ duration was very long (291 ms) we realized that the speaker had in fact pronounced a geminated /m/ (which is possible in emphatic French). We then resegmented "sommet" as /sɔmmɛ/ and the algorithm did not place a silence in that position anymore.

6 Perspectives

As in Campbell's model, our segmental duration predictor (Barbosa & Bailly 1992b), proceeds in two stages. Significant differences have to be noticed: (a) a sequential network (Jordan 1990) constrained by an internal clock generates yhe timing of PC locations; (b) then the IPCIs are shared between the IPCG constituents according to the repartition model which now includes the emergence of pauses. If the *internal clock* is effectively a rhythmical attractor phase resetting has to occur frequently. Our first measurements do

rate	clock (ms)	loc. errors	actual silences
v. slow	360	15	99
slow	325	13	64
normal	270	11	32
fast	210	-	26
v. fast	165	1	5

Table 4: Number of errors obtained by the repartition algorithm in placing the silences. The algorithm does not assign a non-silence where there is a corresponding silence in the natural utterance. All errors are due to extra silences. The number of actual silences by speech rate in the natural utterances is also given (there are 285 IPCGs by rate in the corpus).

not show clear quantisation of the durations of accentual groups (phonation+eventual following silent pause) by the clock as evidenced by Fant & Kruckenberg (1989). We are currently working on a more robust estimation of the clock duration.

The integration of generation of melody in our predictor is now being planned since both melody and duration will be described by global patterns in our framework. An alternative to this *control* approach could be to use a lexicon of multiparametric contours as proposed by Aubergé (1992). Although the latter approach is more compatible with the concept of prosodic prototypes accessed by linguistic and paralinguistic keys, the control approach has be proven to be more efficient (Traber 1992).

7 Acknowledgements

The first author was supported by CNPq-Brasilia/Brasil. We thank especially Frédéric Bimbot who offered his voice, the kind reviewers who read this paper carefully, Jürgen Hess for the German version of the abstract and our listeners who offered their ears for such a difficult perception task.

References

G. Allen (1975). Speech rhythm: its relation to performance universals and articulatory timing. *Journal of Phonetics 3*, 75–86.

V. Aubergé (1992). Developing a structured lexicon for synthesis of prosody. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*, pp. 307–321. Elsevier B.V.

G. Bailly (1986). Un modèle de congruence relationnel pour la synthèse de la prosodie du francais. In *15es Journées d'Étude sur la Parole*, Aix-en-Provence, France, pp. 75–78. Organisée par le GALF.

G. Bailly, T. Barbe & H. Wang (1992). Automatic labelling of large prosodic databases: tools, methodology and links with a text-to-speech system. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*, pp. 323–333. Elsevier B.V.

P. Barbosa & G. Bailly (1992a). Generating segmental duration by p-centers. In C. Auxiette, C. Drake & C. Gérard (Eds.), *4^{th} Rhythm Workshop: Rhythm Perception and Production*, pp. 163–168. Bourges, France.

P. Barbosa & G. Bailly (1992b). Génération automatique des p-centers. *XIXe Journées d'Etudes sur la Parole*, 357–361.

K. Bartkova & C. Sorin (1987). A model of segmental duration for speech synthesis in french. *Speech Communication 6*, 245–260.

W. Campbell (1992). Syllable-based segmental duration. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*, pp. 211–224. Elsevier B.V.

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

A. Classe (1939). *The Rhythm of English Prose.* Oxford: Blackwell.

D. Duez (1987). *Contribution à l'étude de la structuration temporelle de la parole en francais.* Ph. D. thesis, Université de Provence.

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

G. Fant, A. Kruckenberg & L. Nord (1991). Durational correlates of stress in Swedish, French and English. *Journal of Phonetics 19*, 351-361.

P. Fraisse (1974). *La psychologie du rythme.* Paris: Presses Universitaires de France.

A. G. Hauptmann (1993). Speakez: a first experiment in concatenation synthesis from a large corpus. Volume 3, Berlin, pp. 1701–1704.

D. Hirst (1993). Peak, boundary and cohesion characteristics of prosodic grouping. In D. House & P. Touati (Eds.), *Proc. ESCA Workshop on Prosody*, Lund, Sweden, 27-29 September, pp. 32–37.

D. J. Hirst & A. di Cristo (forthcoming). *Intonation systems: a survey of twenty languages.* Cambridge: Cambridge University Press.

P. Howell (1988). Prediction of p-center location from the distribution of energy in the amplitude envelope. *Perception and Psychophysics 43*, 90–93.

M. I. Jordan (1990). Motor learning and the degrees of freedom problem. In M. Jeannerod (Ed.), *Attention and Performance*, Volume XIII. Hillsdale, NJ: Lawrence Erlbaum.

D. H. Klatt (1976). Linguistic uses of segmental duration in english : acoustic and perception evidence. *Journal of the Acoustical Society of America 59*, 1208–1221.

W. A. Lea (1974). Prosodic aids to speech recognition: IV. A general strategy for prosodically-guided speech understanding. Univac Report PX10791, Sperry Univac, DSD, St. Paul, Minnesota, USA.

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

A. Ljolje & F. Fallside (1986). Synthesis of natural sounding pitch contours in isolated utterances using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing 34*, 1074–1080.

S. Marcus (1976). *Perceptual centres.* Ph. D. thesis, Cambridge University.

P. Monnin & F. Grosjean (1993). Les structures de performance en français : caractérisation et prédiction. *L'Année Psychologique 93*, 9–30.

E. Moulines (1992). Synthesis models: a discussion. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*, pp. 7–12. Elsevier B.V.

S. G. Nooteboom (1991, August). Some observations on the temporal organisation and rhythm of speech. In *Proceedings of the XIIth International Conference on Phonetic Sciences*, Volume 1, Aix-en-Provence, France, pp. 228–237.

J. D. O'Connor (1965). The perception of time intervals. Progress Report 2, Phonetics Laboratory, University College, London, UK.

D. O'Shaughnessy (1984). A multispeaker analysis of durations in read French paragraphs. *Journal of the Acoustical Society of America 76*(6), 1664–1672.

J. Pierrehumbert (1980). *The phonetics and phonology of English intonation.* Ph. D. thesis, MIT.

J. Pierrehumbert (1981). Synthetizing intonation. *Journal of the Acoustical Society of America 70*(4), 985–995.

K. Pike (1945). *The intonation of American English.* Ann Arbor: University of Michigan Press.

B. Pompino-Marschall (1989). On the psychoacoustic nature of the p-center phenomenon. *Journal of Phonetics 17*, 175–192.

Y. Sagisaka (1990). On the prediction of global fo shapes for japanese text-to-speech. *IEEE International Conference on Acoustics, Speech and Signal Processing 1*, 325–328.

M. Scordilis & J. Gowdy (1989). Neural network based generation of fundamental frequency contours. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 219–222.

A. Semjen, H.-H. Schulze & D. Vorberg (1992). Temporal control in the coordination between repetitive tapping and periodic external stimuli. In C. Auxiette, C. Drake & C. Gérard (Eds.), *Fourth Rhythm Workshop : Rhythm Perception and Production*, Bourges, pp. 73–78.

Y. Shen & G. G. Peterson (1962). Isochronism in English. *Studies in Linguistics, Occasional Papers 9*, 1–36.

J. t' Hart & R. Collier (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics 1*, 309–327.

K. Takeda, K. Abe & Y. Sagisaka (1992). On the basic scheme and algorithms in non-uniform unit speech synthesis. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*, pp. 93–105. Elsevier B.V.

P. Touati (1987). *Structures prosodiques du suédois et du français : profils temporels et configurations tonales.* Lund University Press.

C. Traber (1992). Fo generation with a database of natural fo patterns and with a neural network. In G. Bailly & C. Benoît (Eds.), *Talking Machines: Theories, Models and Designs*, pp. 287–304. Elsevier B.V.

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

J. Vaissière (1980). La structuration acoustique de la phrase francaise. In *Annali della Scuola Normale Superiore di Pisa X*, Volume 2, pp. 529–560.

J. van Santen & J. Olive (1990). The analysis of contextual effects on segmental durations. *Computer, Speech and Language 4*, 359–390.

P. Viviani & N. Stucchi (1991). Motor-perceptual interactions. In J. Requin & G. Stelmach (Eds.), *Tutorials in Motor Behavior II.* The Netherlands: Elsevier.

B. J. Wenk & F. Wioland (1982). Is French really syllable-timed? *Journal of Phonetics 10*, 193–216.

C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf & P. Price (1992). Segmental durations in the vicinity of prosodic boundaries. *Journal of the Acoustical Society of America 91*(3), 1707-1717.

Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia de fala

Plínio Almeida Barbosa, LAFAPE/IEL/Unicamp

> *Un petit coup au carreau, comme si quelque chose l'avait heurté, suivi d'une ample chute légère comme de grains de sable qu'on eût laissés tomber d'une fenêtre au-dessus, puis la chute s'étendant, se réglant, adoptant un rythme, devenant fluide, sonore, musicale, innombrable, universelle : c'était la pluie.*
> *MARCEL PROUST, Du côté de chez Swann*

DA NATUREZA DO RITMO

Ninguém melhor que os bons escritores para despertar nossa sensibilidade ao aspecto rítmico da língua. Como não perceber neste parágrafo proustiano antológico a descrição quase tangível de uma chuva repentina? O espaçamento entre as vírgulas guia a leitura e as amplas proposições do início do texto vão aos poucos sendo substituídas por outras, mais curtas, formadas por palavras que caracterizam a chuva: fluidez, sonoridade, musicalidade. Ou, como diz o próprio autor, as palavras vão adotando um ritmo. Mesmo para aquele que não compreende o francês é possível perceber esta ritmicidade pois, ao ritmo oriundo da leitura, corresponde um ritmo visual em que os objetos percebidos são delimitados pelos sinais de pontuação.

Poder-se-ia argumentar que não há nada de espantoso neste efeito visto que a atividade rítmica se encontra em todo lugar da experiência quotidiana e é uma propriedade fundamental da natureza viva (Fraisse, 1974). De fato, não se pode negar a universalidade da fenomenologia do ritmo e seu papel na regulação das atividades primárias dos seres vivos. A adequação entre os ritmos biológicos internos e os ciclos naturais é condição *sine qua non* de sobrevivência para estes seres, entre os quais nos incluímos.

Sendo de apreensão evidente, seria natural supor que a noção de ritmo seja antiga. No que concerne à etimologia, constata-se que a palavra *ritmo* provém do latim *rhythmus*, palavra que, por sua vez, é oriunda do grego *ῥυθμός*. Segundo Benveniste (1951), o termo *ῥυθμός* teve na filosofia iônica (sobretudo de Demócrito e Leucipo) o significado de *forma*, evoluindo para *forma ligada aos movimentos humanos* com Platão (*Leis, 665a*): "É a ordem no movimento". Assim, ainda segundo Fraisse, "le concept de rythme ne viendrait pas de quelque expérience de la nature mais bien de l'organisation du mouvement humain."

No que diz respeito à atividade fonatória, o Dicionário de Lingüística e Fonética de Crystal (1985) refere-se ao ritmo como "regularidade percebida de unidades proeminentes na fala". Entretanto, mesmo que, ao se falar de ritmo, se tenha em mente que os fatores temporais sejam de importância principal (Fraisse, 1974) para a percepção desta regularidade, esta só é percebida pelo concurso de outros parâmetros prosódicos além da duração. Ao se falar de prosódia, é preciso distinguir seu aspecto de produção (identificado pelos três parâmetros clássicos: a duração – representada pela diferença de tempo entre dois eventos –, a freqüência fundamental e a intensidade), de seu aspecto de percepção (identificado pelas noções de duração percebida, altura e volume).

Para que se possa experimentar a sensação perceptiva da duração é preciso que dois eventos acústicos singulares ocorram no tempo e que estes sejam associados em nossa memória de curto termo afinal, "sans une mémoire élémentaire qui relie les deux instants l'un à l'autre, il n'y aura que l'un ou l'autre des deux, un instant unique par conséquent, pas d'avant et d'après, pas de succession, pas de temps." (Bergson, 1968). Além do tempo, a freqüência fundamental e a intensidade também concorrem para a percepção da duração[1] (Fraisse, 1974). A sensação de duração percebida é obtida portanto, pelo concurso dos parâmetros prosódicos como um todo, e não apenas pela duração mensurável por instrumentos de medida de tempo (que estamos chamando de **duração observada** ou simplesmente **duração**).

Estas considerações nos permitem sugerir a definição de ritmo como *a variação a longo termo da duração percebida*. Visto acreditarmos em uma visão teleológica do ritmo (é produzido *para* ser percebido), talvez ele deva ser melhor definido a partir da experiência perceptiva.

A variação da duração percebida a longo termo condiciona a percepção não somente de regularidade, como também de estruturação (Woodrow, 1951, p. 1232. Grifo nosso):

> *By rhythm in the psychological sense, is meant the perception of a series of stimuli as a series of groups. The successive groups are ordinarily of similar pattern and experienced as repetitive.* Each group is perceived as a whole *and therefore has a length lying within the psychological present.*

O aspecto estrutural do ritmo permite relacionar o grupo rítmico à noção de forma introduzida pela *Gestalttheorie* no sentido de que os estímulos não são percebidos de maneira independente, mas interagem entre si favorecendo a percepção de uma globalidade, a *Fugengestalt*. Fraisse (1974, p. 74) explica que, sob certas condições de sucessão, os estímulos são percebidos como agrupados e que a repetição destes grupos dá nascimento à percepção do ritmo.

No que diz respeito à fala, o ritmo é estudado empiricamente pela observação sistemática dos mecanismos de produção e percepção. Testemunho disto é a sensação observada precocemente (desde o século XVIII, segundo Abercrombie, 1967) de que o inglês tem a tendência a produzir sílabas acentuadas a intervalos regulares de tempo (*isochronous stressed syllabes*). Para uma discussão

[1] Uma sílaba é percebida como sendo mais curta se ela possui um tom estático (Lehiste, 1978).

detalhada sobre o isocronismo[2] ou isossilabismo na fala, ver Lehiste (1977) e Barbosa (1994, p.60-74).

A tarefa de fazer emergir a estrutura rítmica a partir do estudo dos fenômenos de *parole* é, no entanto, árdua. Tradicionalmente se procura analisar o ritmo pelo estudo da duração observada, deixando-se de lado os papéis desempenhados pela freqüência fundamental e pela intensidade em sua percepção. A tarefa não deixa de ser menos complexa, visto que é preciso escolher um par de eventos representativo do fenômeno observado dentro de uma constelação de eventos detectáveis no sinal de fala.

O estudo exaustivo de Abry e colegas (1985) apresenta uma dezena de eventos do sinal acústico susceptíveis de funcionar como fronteiras para a delimitação da duração. Alguns deles são o início e o fim do vozeamento, o início e o fim do vozeamento vocálico, o início e o fim da fricção consonantal. O estudo da duração implicará na escolha de eventos pertinentes. Contudo, a detecção destes eventos no sinal acústico não é evidente, nem a partir da forma de onda do sinal, nem a partir de informação espectral (presente em um espectrograma, por exemplo): uma dose de incerteza sempre existe devido aos fenômenos de coarticulação – a influência de um segmento sobre o outro (Fowler, 1981). Ao processo de marcação dos eventos que foram escolhidos para a caracterização da duração, dá-se o nome de segmentação. Tendo-se escolhido estes eventos, mesmo uma cuidadosa segmentação manual introduz erros devido a decisões altamente subjetivas e à fadiga inerente a uma tarefa que consome muito tempo (Leung & Zue, 1984).

É justamente a partir da segmentação manual de um *corpus* de cem frases lidas por um locutor profissional que iniciamos um estudo sistemático da estruturação rítmica do português brasileiro-PB. A duração observada é o único parâmetro prosódico que será considerado, tendo em vista o objetivo visado por este estudo, a saber, a geração automática da duração segmental para um sistema de síntese da fala.

Veremos mais adiante que a duração segmental (duração de um segmento de fala delimitado por eventos acústicos singulares) é obtida a partir de unidades de nível superior ao segmento, que garantem a ritmicidade da frase a ser sintetizada e confirmam hipóteses em produção e percepção de fala (cf. discussão). Antes de se mostrar como a duração pode ser gerada automaticamente, convém introduzir um breve histórico da síntese da fala, nascida do antigo desejo do homem de reproduzir sua voz por máquinas falantes.

[2] A isocronismo dá-se o nome ao fato de que, em línguas ditas de ritmo acentual, o acento frasal tende a ocorrer a instantes iguais de tempo (cf. Lehiste, 1977 para uma revisão). Pode-se aplicar o termo de isossilabismo ao fato de que, em línguas ditas de ritmo silábico, a sílaba tende a ocorrer a instantes iguais de tempo. Dentro desta tipologia, o francês é normalmente considerado de ritmo silábico (mas Wenk & Wioland, 1982 para uma crítica) e o português do Brasil, de ritmo acentual (Major, 1981).

DO DESAFIO DAS MÁQUINAS FALANTES

Toda história de síntese da fala remonta aos gregos e suas estátuas falantes, utilizadas por sacerdotes que desejavam impressionar seus fiéis (Flanagan & Rabiner, 1973). Mas a verdadeira tentativa de se compreender e reproduzir os sons da linguagem articulada viria bem mais tarde, com o barão von Kempelen, em 1791. A máquina que construiu era composta de um fole, de um bocal cuja variação de volume era efetuada pela mão esquerda (para a produção de vogais), de narinas e apitos acionados por alavancas controladas pela mão direita (para a produção das consoantes). Esta máquina, em relação à qual von Kempelen era virtuose, podia emitir uns vinte sons diferentes (Calliope, 1989).

Outras máquinas foram construídas no século XIX, deixando entrever dois métodos para a reprodução da fala, como se depreende do testemunho de du Moncel (1880; *apud* Köster, 1973, p. 148) a respeito da máquina do professor de matemática Joseph Faber, apresentada pela primeira vez em 1835:

> *On s'est étonné que la machine parlante qui nous est venue, il y a quelques années d'Amérique (Barnum hatte die Maschine nach einer Amerikatournee 1875 nach Paris gebracht), et qui a été exhibée au Grand-Hôtel fût d'une extrême complication, alors que le phonographe résolvait le problème d'une manière simple : c'est que l'une de ces machines ne faisait que reproduire la parole, tandis que l'autre l'émettait, et l'inventeur de cette dernière machine avait dû, dans son mécanisme, mettre à contribution tous les organes, qui dans notre organisme, concourent à la production de la parole.*

A possibilidade de realizar a síntese da fala a partir do texto que, como o próprio nome sugere, significa emitir os sons da fala a partir de uma representação textual da mensagem, desde cedo suscitou duas linhas de pesquisa. A primeira linha busca reproduzir da melhor forma possível um sinal acústico que *pareça* com o sinal da fala (chamaremos de abordagem *fazer-parecido*). A segunda linha procura obter sinal acústico a partir das causas que o propiciaram, reproduzindo o mecanismo fonatório da forma *como* ele funciona no ser humano (chamaremos de abordagem *fazer-como-se-fosse*[3]).

O *fazer-como-se-fosse* é realizado pela síntese articulatória e representa o estado-da-arte da pesquisa internacional (Coker, 1968; Bailly, Laboissière & Schwartz, 1991). Tem por meta científica obter mensagem sonora que, não apenas pareça aquela oriunda de um aparelho fonador, mas também reproduza esta mesma mensagem *como* o aparelho fonador o faz. Neste sentido é que se diz que a síntese articulatória se encontra no extremo antropomorfizante entre os sistemas de síntese da fala.

[3] As duas abordagens aqui referidas são uma adaptação dos termos *faire-semblant* e *faire-comme* propostos pelo *Institut de la Communication Parlée* (ICP, 1994).

Este desafio científico está sendo alcançado pelo estudo da dinâmica dos articuladores envolvidos com a fonação, das fontes sonoras (controle dos movimentos das cordas vocais, ruídos de fricção, efeitos de turbulência), do papel da percepção na seleção dos gestos articulatórios que podem ser produzidos e pela conseqüente simulação computacional destes fenômenos.

Por outro lado, a abordagem *fazer-parecido* é ainda muito presente no cenário internacional. A inexistência – até o presente momento – de um sistema de síntese articulatória eficaz e operacional garante a necessidade de sistemas de síntese menos custosos computacionalmente em que se possa testar várias hipóteses oriundas de estudos articulatório-perceptivos. Ao *fazer-parecido*, pode-se distinguir entre métodos e técnicas de síntese de fala a partir do texto.

Métodos de síntese

Excetuando-se a síntese articulatória, em que método e técnica de síntese fazem um todo orgânico, os métodos de síntese de fala a vocabulário ilimitado[4] se restringem à síntese concatenativa e à síntese por regras (Klatt, 1987).

O primeiro deles se propõe a gerar sinal de fala pela concatenação de porções de sinal pré-armazenadas e organizadas em um dicionário. Estas porções de sinal são recuperadas por um gerador segmental que as alinha, constituindo então o sinal concatenado. As porções de sinal pré-armazenadas possuem tamanhos diversos e são delimitadas por dois pontos de quase-estacionaridade (relativamente estáveis do ponto de vista da variação a curto prazo da evolução das formantes) do sinal de fala. São constituídas por difones, contendo apenas uma transição de segmento a segmento ou, de maneira geral, por polifones, contendo transições mais complexas, como no caso da seqüência /-a.ɾʊ/ em *caro*.

O segundo método, a síntese por regras, parte de uma descrição detalhada das regras que regem os movimentos dos formantes (sobretudo durante as transições entre segmentos) presentes no sinal de fala que se deseja gerar, caracterizando acusticamente a dinâmica da fonação. O sinal de fala é gerado *a posteriori* através de um sintetizador de formantes.

[4] Falaremos aqui apenas da síntese de fala a vocabulário ilimitado por favorecer a geração de som a partir de um texto qualquer. A síntese a vocabulário limitado, cuja unidade utilizada para reprodução do som é a palavra, não permite obter sinal acústico a partir de um texto genérico, não constituindo para nós fonte de interesse científico: desde que novas frases se façam necessárias, a explosão do vocabulário obriga a adoção de métodos cuja unidade manipulada é inferior à palavra.

Nos dois métodos, o conhecimento lingüístico extraído do texto é integrado em etapas anteriores do processamento e usado para a atualização da mensagem (ver figura 1). Esta atualização é efetivada por um gerador automático de prosódia, que fornece a informação relativa à variação dos parâmetros prosódicos clássicos ao sintetizador, segundo o conteúdo presente no texto escrito.

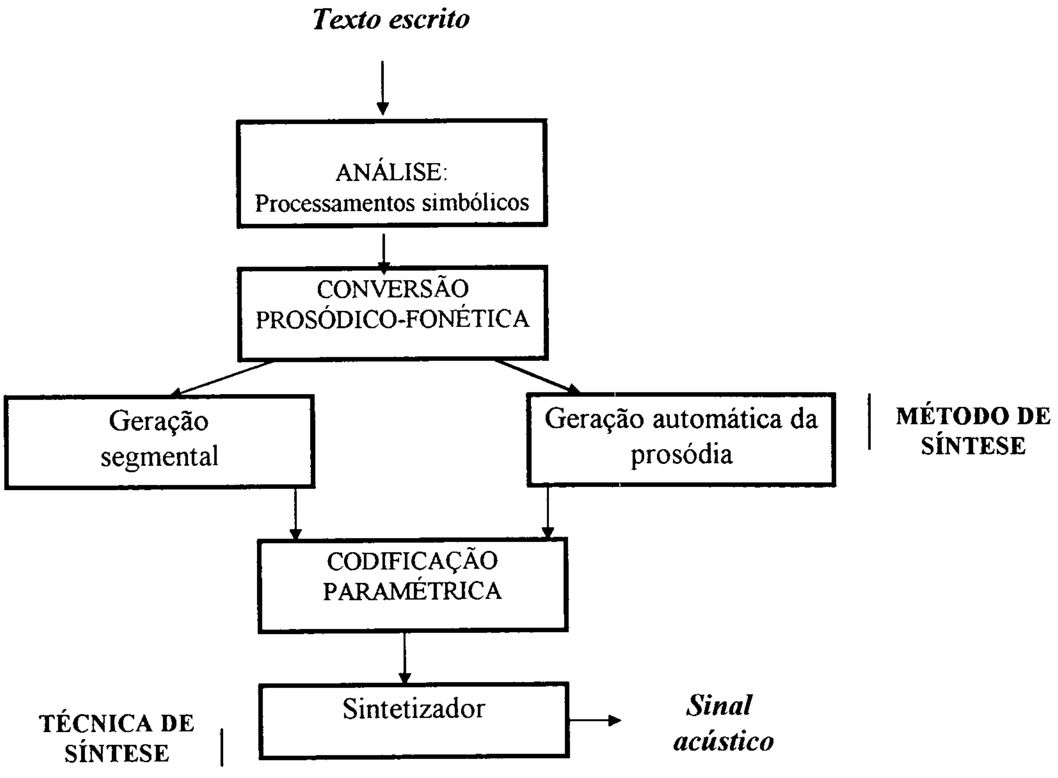


Figura 1: Esquema geral de um sistema de síntese da fala

Técnicas de síntese

Além de um método, utiliza-se uma *técnica de síntese* para a obtenção do sinal acústico. Ela constitui o sintetizador propriamente dito, a etapa final de um sistema de síntese. As técnicas usadas comumente são a PSOLA (*Pitch Synchronous OverLap and Add*), a LPC *(Linear Predictive Coding*) e aquela representada por um sintetizador de formantes.

A técnica PSOLA opera diretamente sobre o sinal de fala, modificando os valores de duração, freqüência fundamental e intensidade do sinal concatenado via multiplicação, redução, compressão ou expansão de períodos glotais. A técnica LPC pressupõe que o sinal de fala seja produzido por uma fonte sonora (pulsos glotais ou ruído de fricção) aplicada a um filtro (trato vocal). O filtro é implementado por um conjunto fixo e pré-definido de parâmetros (dentre os quais, *n* coeficientes LPC) que permitem a obtenção da amostra atual do sinal de fala a partir de *n* amostras anteriores. O sintetizador de formantes é usualmente empregado com um método de síntese por regras e gera o sinal acústico a partir da informação, a uma taxa de amostragem pré-definida, dos três primeiros formantes, de suas larguras de banda e de suas amplitudes, além da informação prosódica dada pela freqüência fundamental, pela duração e pela intensidade.

Em um sistema de síntese concatenativo, o gerador segmental se ocupa da concatenação efetiva dos polifones. Mesmo com polifones não é possível embutir na porção de sinal pré-armazenada fenômenos coarticulatórios mais extensos (a combinatória causaria um aumento contra-producente do número de unidades concatenantes). A coarticulação é planificada de antemão e desempenha papel importante para a eficácia da comunicação (Fowler, 1981; Whalen, 1990). Sendo assim, a naturalidade do sinal acústico em um sistema de síntese concatenativo não é garantida apenas com a geração segmental, o gerador automático de prosódia desempenha uma função crucial: assegurar que o sinal acústico concatenado a partir da informação textual reproduza as variações rítmica, entoacional e de energia que seriam obtidas com a leitura do trecho escrito por um ser humano. No que concerne à ritmicidade, garantida primordialmente pelas modulações de duração ao longo da frase, faz-se necessária a implementação de um modelo de geração automática da duração segmental.

MODELOS DE GERAÇÃO DA DURAÇÃO SEGMENTAL

Dois modelos de geração da duração serão apresentados: o modelo de Dennis Klatt, que procura dar conta da influência do contexto prosódico-fonético sobre a duração do segmento, e o modelo de Nick Campbell, que gera a duração do segmento a partir da duração de uma unidade de nível superior, a sílaba.

O Modelo de Klatt

O sistema de predição de duração de Klatt (1987) para o inglês serviu e serve de referência para um grande número de modelos desenvolvidos por outros pesquisadores (cf. van Santen, 1994, p. 102). Todos eles tomam o segmento como paradigma para a obtenção da duração segmental. Esta duração é obtida após a aplicação sucessiva de um certo número de regras.

Os princípios fundamentais do modelo klattiano são: (a) a cada segmento se associa uma duração intrínseca específica, representando uma de suas propriedades distintivas; (b) cada regra procura introduzir uma certa porcentagem de modificação à duração de cada segmento e (c) os segmentos não podem ser comprimidos aquém de uma duração mínima.

A duração segmental pode então ser expressa por um modelo aditivo-multiplicativo da forma:

$$\mathrm{Dur} = \mathrm{DurMin} + \frac{(\mathrm{DurInt} - \mathrm{DurMin}).\ \mathrm{PRNCT}}{100} \qquad (1)$$

Onde *Dur* é a duração do segmento, *DurInt* é a duração intrínseca, *DurMin* é a duração mínima, calculada em função de *DurInt* (em geral *DurMin = 0,45 .DurInt*) para cada segmento não acentuado. O valor *PRNCT* corresponde à porcentagem de encolhimento, determinada de maneira cíclica e cumulativa pela aplicação das regras (uma regra introduz em geral um fator multiplicativo que é reaplicado ao valor atual de *PRNCT*, fornecido por uma regra anterior). Os fatores que condicionam o valor final de *PRNCT* são o contexto fonético imediato e o ambiente sintático-prosódico do segmento. Pausas silenciosas são atribuídas desde o início, de maneira não integrada ao mecanismo geral das regras (na verdade funcionam como fator de influência para a duração segmental).

Klatt afirma que influências rítmicas e semânticas podem ser introduzidas *a posteriori*, embora não as incorpore explicitamente em seu modelo (*ibidem*, p. 761). É justamente para garantir a influência de níveis superiores do processamento lingüístico que Campbell propõe um modelo que tem uma unidade de programação rítmica como paradigma para a derivação da duração segmental.

Quando falamos de unidade de programação rítmica nos referimos a uma unidade rítmica mínima (UPRM) que seja operacional tanto em produção quanto em percepção de fala. O termo de programação refere-se ao fato de que tal unidade é planejada com antecedência e participa na organização do ritmo a diversos níveis de sua estruturação. Sem justificar empiricamente o porquê[5], a sílaba é adotada como UPRM por Campbell.

Esta programação diz respeito à organização temporal de gestos vocálicos e consonantais, organização esta que se manifesta articulatoriamente através de duas estratégias distintas (Edwards *et al.*, 91): (1) a rigidez própria aos gestos de abertura e de fechamento, de ordem intragestual, ligada à precisão do gesto; (2) a organização temporal intergestual, entre dois gestos consecutivos, ligada à duração do gesto. A primeira estratégia desempenha um papel preponderante na variação da taxa de elocução (*speech rate*) e, a segunda, no mecanismo acentual. Estas estratégias se relacionam a unidades de programação acima do segmento (a UPRM) e caracterizam o macrorritmo (Barbosa, 1994) ou ritmo propriamente dito.

A UPRM é uma unidade que age como elemento estruturante a níveis superiores de organização rítmica ao mesmo tempo em que fornece um *frame* no qual o *timing* dos gestos vocálicos e consonantais são computados, a nível microrrítmico. Muitos tomam a UPRM como sendo a sílaba (Mehler e colaboradores, por exemplo). Mas, como disse Hirst (1993): "A maior parte dos argumentos em favor da sílaba como unidade são de fato argumentos em favor de silabicidade."

O Modelo de Campbell

Campbell (1992) define seu modelo de predição da duração segmental para o inglês britânico como um modelo combinado: ele separa o controle do tempo ao nível da sílaba (procurando assim descrever a estruturação rítmica da fala) do cálculo da duração do segmento (a um nível inferior), cálculo este que é efetuado a partir do paradigma temporal fornecido pela sílaba. Na literatura fonética, Kozhevnikov & Chistovich (1965) e Collier (1992, p. 206), entre outros, sugerem a existência de unidades de programação rítmica superiores ao segmento.

O modelo concebido por Campbell opera em duas etapas. Na primeira, a duração silábica é obtida por aprendizado automático pelo uso de uma rede conexionista do tipo perceptron multicamadas (cf. apêndice para explicação dos termos). Em uma segunda etapa, esta duração é distribuída entre os segmentos que formam a sílaba pelo uso de um modelo estatístico que chamamos de modelo de repartição (Barbosa, 1994).

A rede conexionista é treinada para aprender a associar uma descrição fonológica da sílaba e de seu contexto frasal (no domínio simbólico) à duração real desta mesma sílaba (no domínio físico). A rede

[5] Uma justificativa mais rigorosa empiricamente seria necessária devido à hipótese forte relacionando duração da sílaba e duração dos segmentos que Campbell pressupõe em seu modelo. Esta hipótese é apresentada adiante. Uma tentativa de justificar mais rigorosamente a adoção de uma UPRM culminou em nossa proposta de uma unidade delimitada por dois onsets vocálicos consecutivos para o francês (Barbosa, 1994) e, para o português brasileiro (apresentado aqui), a proposta de duas unidades de programação.

realiza uma passagem complexa entre o código simbólico e uma realização. A descrição fonológica da sílaba utilizada por Campbell à entrada do perceptron é composta pelos itens seguintes: (a) número de fonemas; (b) natureza do núcleo (vogal reduzida, vogal *lax* ou *tense*, consoante silábica, ditongo ou tritongo) ; (c) posição no grupo tonal; (d) tipo de pé; (e) natureza acentual (*stressed* ou *unstressed*) e (f) classe da palavra contendo a sílaba (clítica ou não clítica). A rede conexionista se ocupa portanto do componente macrorrítmico da fala.

O modelo de repartição é baseado em um princípio de elasticidade (Campbell & Isard, 1991) que, em sua versão mais forte, estabelece que todos os fonemas de uma determinada sílaba possuem um único fator de alongamento *z* (de agora em diante *z-score*) que impõe que a duração da mesma é dada por:

$$\text{Duração (sílaba)} = \sum_{i=1}^{n} \exp(\mu_i + z.\sigma_i) \qquad (2)$$

Onde a duração de cada segmento *i* é obtida pelas parcelas $\exp(\mu_i+z.\sigma_i)$. O par estatístico (μ_i,σ_i) representa a média e o desvio-padrão associados à distribuição formada pelas durações das realizações do fonema *i*. Esta distribuição é obtida pela análise de um *corpus ad hoc* de frases lidas. A função exponencial *exp()* é necessária porque se usa o logaritmo das durações segmentais: a distribuição assim obtida se aproxima mais da distribuição gaussiana do que aquela que seria obtida com a duração expressa em milissegundos, por exemplo (Barbosa, 1994; Campbell, 1992).

É importante notar na fórmula 2 acima que um único valor *z* é utilizado para todos os segmentos que compõem a sílaba. É a este fato que Campbell se refere quando fala de hipótese (forte) de elasticidade uniforme para a sílaba: todos os segmentos que a compõem estão sujeitos ao mesmo alongamento (ou compressão). O *z-score* enunciado aqui é uma medida da distância (em unidades da soma dos desvios-padrão dos segmentos) da duração da UPRM em relação à soma das durações médias dos segmentos que a formam. O *z-score* pode ser chamado de duração normalizada (cuja norma são os pares estatísticos calculados sobre um *corpus ad hoc*), na medida em que procura fornecer o alongamento (ou compressão) da UPRM, independentemente da duração intrínseca de seus segmentos. Os valores de *z-score* possibilitam a obtenção da duração segmental e, portanto, do componente microrrítmico da fala.

A elaboração de um modelo de geração da duração segmental, que também procede em duas etapas (obtenção da duração de uma UPRM e distribuição da duração da unidade entre os segmentos que a compõem), foi efetuada inicialmente para o francês (Barbosa, 1994) através do teste da hipótese forte de elasticidade enunciada por Campbell.

O modelo de geração da estruturação rítmica do francês proposto a partir das análises das durações dos segmentos presentes em *corpora* de fala mostrou que a sílaba não era a melhor UPRM para esta língua. Uma outra unidade, delimitada por dois *onsets* (acusticamente definidos) vocálicos consecutivos mostrou uma coerência maior entre seus elementos constitutivos, em termos de alongamento homogêneo. Devido ao fato do *onset* vocálico ser o *point d'ancrage* por excelência para a percepção ou a produção da ritmicidade segundo os estudos em torno das noções de isocronismo e *perceptual-center* (Marcus, 1981; Morton et al., 1976) – que representa o evento acústico singular que seria usado pelos auditores para alinhar estímulos sonoros e perceber o isocronismo da fala –, essa unidade foi denominada grupo *inter-perceptual-center* ou GIPC. O GIPC é então composto pela rima de uma sílaba e o ataque da sílaba seguinte, quando este é presente.

PARA O MODELO DE BARBOSA-BAILLY EM PB: ANÁLISE DE *CORPORA*

O modelo desenvolvido para o francês (Barbosa, 1994; Barbosa & Bailly, 1997) foi adaptado para o PB (Barbosa, 1997a, 1997b). Ele permite a obtenção automática da duração segmental em duas etapas: aprendizado de formas rítmicas por uma rede conexionista e distribuição da duração das UPRM entre seus segmentos constituintes. Para que isto fosse possível uma análise detalhada da fenomenologia duracional presente em dois *corpora* de fala lidos por um locutor profissional paulista (cerca de 30 anos, da região de Campinas) foi empreendida. Os *corpora* foram segmentados manualmente (mais de 6000 fronteiras para os segmentos foram introduzidas).

O *Corpus* de logatomas e a distribuição das durações segmentais

Um *corpus* contendo 1195 polifones foi gravado para a constituição de um dicionário de unidades para um sistema de síntese da fala concatenativo. Este *corpus* foi usado para a obtenção dos pares estatísticos (μ,σ) associados aos fonemas (e alguns alofones) do PB, calculados a partir da distribuição das durações dos segmentos, expressos em logaritmo natural. Para uma melhor clareza, a tabela que se segue apresenta os resultados expressos em unidade de tempo. Princípios fonológicos e articulatórios bem conhecidos corroboram estes resultados.

Tabela 1: Duração média (e desvio-padrão) dos fones do PB (em ms) para o locutor[6].

i	145 (37)	ĩ	209 (25)	f	138 (14)
e	170 (36)	õ	229 (26)	s	143 (26)
ɛ	175 (32)	ũ	215 (29)	ʃ	143 (16)
a	165 (28)	j̃	136 (14)	v	78 (16)
u	134 (42)	w̃	139 (23)	z	87 (21)
o	168 (35)	p	120 (20)	ʒ	89 (12)
ɔ	183 (29)	t	113 (20)	m	90 (12)
ɐ	111 (45)	tʃ	149 (20)	n	76 (15)
ɪ	98 (44)	k	121 (21)	ɲ	103 (24)
ʊ	77 (19)	b	86 (17)	ɾ	47 (16)
j	92 (10)	d	71 (17)	r	81 (12)
w	97 (25)	dʒ	109 (18)	ʁ	62 (15)
ɐ̃	174 (46)	g	67 (16)	l	73 (16)
ẽ	210 (44)			ʎ	77 (14)

Pode-se verificar pela tabela acima a coerência, em termos acústico-articulatórios, das médias e desvios-padrão: à abertura vocálica corresponde um *crescendo* de duração[7], as consoantes surdas são mais longas que as sonoras correspondentes, as vogais nasais são mais longas que as orais correspondentes (Sousa, 1994), as vogais pós-tônicas são mais curtas e variam mais do que as tônicas correspondentes.

Um outro *corpus* permitiu testar a sílaba e o GIPC como unidades de programação para o PB.

O *Corpus* de frases lidas e o papel das UPRM

Um *corpus* de cem frases lidas foi gravado e segmentado. De posse das durações segmentais, é possível calcular o *z-score* associado a cada segmento, pois a duração dos mesmos é obtida pelas parcelas da fórmula 2:

$$\text{Duração (segmento)} = \exp(\mu_{\text{segmento}} + z_{\text{segmento}}.\sigma_{\text{segmento}}) \quad (3)$$

Admitindo a hipótese forte de elasticidade seja para a sílaba, seja para o GIPC, também é possível calcular os *z-scores* para as sílabas e os GIPCs das frases do *corpus*. Os valores são obtidos pelo cálculo recorrente usando a fórmula 2 acima para a sílaba e para o GIPC. Se a hipótese de elasticidade é correta o valor único do *z-score* da UPRM deveria ser o mesmo que os valores individuais dos *z-scores* de cada segmento (calculados pela fórmula 3 acima). Na prática este não é o caso: o *z-score* da UPRM é uma média ponderada dos *z-scores* dos segmentos que a compõem. Independentemente da

[6] Para os fins da síntese concatenativa, interessa aqui a realização fonética dos fonemas. Assim, mesmo que /t/ e /tʃ/ não representem fonemas distintos, suas durações são expressas aqui por se tratar de realizações físicas envolvendo modos de articulação distintos com conseqüência na duração. O mesmo vale para as vogais e semi-vogais nasais. Três formas de "r" aparecem aqui. A vibrante múltipla /r/ ocorreu em final de sílaba. A versão forte de "r" (de *carro* ou *rosa*) foi realizada por uma fricativa. Todas as suas ocorrências foram categorizadas pela fricativa uvular.

[7] Os segmentos /ɐ/ e /ʊ/ parecem ir de encontro a esta e à última constatação que fazemos a partir da tabela. Além de idiossincrasia do locutor, somente fatores ligados a características inerentes à escolha dos logatomas do *corpus* explicam estes fatos.

validade da hipótese de Campbell, pode-se calcular os valores dos *z-scores* da sílaba e do GIPC e analisar suas evoluções ao longo da frase.

A vantagem de uma medida de duração normalizada como o *z-score* da UPRM em relação à duração observada é a de evitar durações mais longas para a unidade de programação simplesmente por conter maior número de fonemas (observar na figura 2, para a sílaba, o menor valor de duração da sílaba /o/ - décima posição na abscissa -, na palavra "ou" (pronunciado /o/), em relação à sílaba /bo/ - segunda posição na abscissa -, na palavra "bolsa" (pronunciado /'bo.sɐ/). O mesmo não ocorre com o GIPC: a unidade /os/, entre as palavras "ou" e "sofrerá" (pronunciado /so.fre.'ra/), tem duração superior a /os/ da palavra "bolsa"). O valor do *z-score* é uma indicação do alongamento sofrido pela unidade de programação, independentemente do número de seus elementos constituintes.

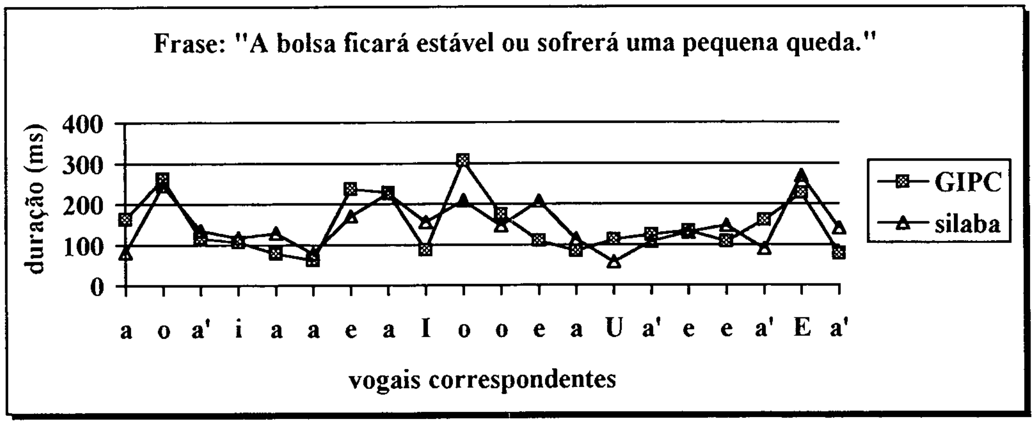


Figura 2: Evolução da duração da sílaba (dsílaba) e do GIPC (dgipc) ao longo da frase "A bolsa ficará estável ou sofrerá uma pequena queda." As vogais correspondentes são indicadas no eixo horizontal (a' representa o alofone /ɐ/).

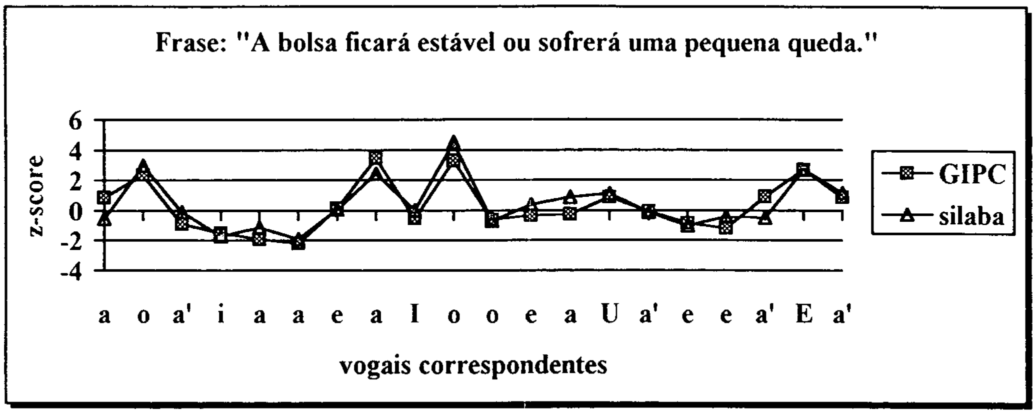


Figura 3: Evolução do *z-score* da sílaba (zsílaba) e do GIPC (zgipc) ao longo da frase "A bolsa ficará estável ou sofrerá uma pequena queda." As vogais correspondentes são indicadas no eixo horizontal (a' representa o alofone /ɐ/). O último valor de z-score (GIPC) mostrado aqui é na verdade o z-score da última rima da frase (/ɐ/). Em final de frase isolada a última rima nunca forma um GIPC, por não haver fronteira à direita (*onset* vocálico) que o delimite.

Diferenças entre a duração intrínseca das UPRM e o valor de seus *z-scores* ainda são bastante visíveis ao se comparar as figuras 2 e 3. Para o GIPC, na figura 2, a palavra "estável" tem dois picos (pois os GIPCs são "est" e "av": três e dois fonemas, respectivamente). Os respectivos *z-scores* dos GIPCs dessa palavra na figura 3 têm apenas um pico na unidade "av". Para a sílaba, o pico da sílaba "fre" ("sofrerá") da duração intrínseca (figura 2), é transferido à sílaba "rá", na figura 3.

Ao se utilizar os valores dos *z-scores* da sílaba ($z_{\text{sílaba}}$) e do GIPC (z_{GIPC}), dois pontos se esclarecem. Primeiro, o $z_{\text{sílaba}}$ assinala sistematicamente o acento lexical (excetuando-se os casos – pouco freqüentes neste *corpus* – de desacentuação. Um exemplo é o da frase acima, na palavra "ficará"): os máximos de $z_{\text{sílaba}}$ para cada palavra coincidem com os acentos lexicais das mesmas. Segundo, o z_{GIPC} assinala sistematicamente a fronteira frasal: os máximos de z_{GIPC} que correspondem a uma posição de acento lexical demarcam a frase em grupos acentuais (ou palavras prosódicas). O valor do z_{GIPC} que termina cada grupo acentual representa a força da ligação entre este grupo e o seguinte.

A hierarquia introduzida pela força das fronteiras rítmicas[8] não é a mesma das introduzidas pelas fronteiras sintáticas[9]. De fato, ao observarmos a figura 4, a mais forte fronteira frasal, representada

[8] Pode-se conceber uma hierarquia rítmica (ou árvore de *performance*, na terminologia de Grosjean) definida pelos diferentes graus de força (maior duração) que as pausas subjetivas fazem emergir ao longo da frase. A palavra ligada à pausa mais forte divide a frase em dois blocos e assim por diante, de acordo com a força que as demais pausas representam.

[9] As fronteiras sintáticas foram demarcadas manualmente, mas por um mecanismo automatizável. A partir de uma gramática de dependência (Tesnière, 1965; Martin, 1981), um conjunto de nove marcas distintas (versão modificada das marcas de Bailly, 1986) são obtidas pela projeção da árvore de superfície (cuja cabeça é

pelo z_{GIPC} de *vi* em *visa*, divide a frase em dois blocos de 16 GIPCs cada. Um divisão baseada no conteúdo sintático colocaria a fronteira mais forte entre a oração principal e sua subordinada, ou seja, o acento recairia sobre a sílaba -câm- em "intercâmbio", separando "O convênio permite o intercâmbio" de "porque visa à integração entre alunos de culturas diferentes." Os grupos acentuais delimitados pelas posições de máximo do *z-score* do GIPC (coincidente com uma posição de acento lexical) permitem inferir uma regra para a determinação das fronteiras frasais que alia informação sintática a princípios fonotáticos. Para este locutor, nesta taxa de elocução e para esta situação de leitura de frases isoladas, a regra poderia ser enunciada da seguinte maneira.

1. inserir acento frasal nas posições correspondentes às marcas mais fortes (IF e TF: cf. nota de rodapé 8);

2. dividir os blocos resultantes em grupos de dois ou três sub-blocos, segundo as marcas seguintes, em termos de hierarquia de força (ID, DF ou GF), procurando obter sub-blocos de tamanho comparável;

3. se os sub-blocos contêm mais do que um número pré-definido *maxsil* de sílabas (que no caso deste locutor, para a taxa de elocução que usou é de 10 sílabas. Este valor corresponde ao maior grupo acentual, em termos de número de sílabas, observado no *corpus*), subdividi-los segundo as marcas mais fortes restantes;

4. se dois sub-blocos juntos contêm um número menor do que *maxsil* sílabas, reuni-los em um só bloco;

5. verificar a eurritmia da disposição resultante e reorganizar novamente, se necessário, a partir do segundo item. A eurritmia estabelece que os sub-blocos sucessivos (respeitando-se, é claro, aquelas posições impostas pelo item primeiro) devam possuir um número próximo de sílabas e que a estrutura rítmica resultante deva ser hierarquicamente aceitável.

A coerência da regra enunciada aqui com as árvores de *performance* de Grosjean e colegas (1983) é evidente. Este autor teoriza que, em condição de enunciação, o locutor transforma a árvore de

representada pelo verbo) sobre o eixo sintagmático. A força entre nós adjacentes sobre este eixo é indicada pela relação de dependência entre os mesmos. As marcas são: IF (quando os nós pertencem a árvores distintas, o que corresponde, por exemplo, a posições de sinal de pontuação fortes ou conjunções coordenativas); TF (quando os dois nós dependem do mesmo nó, representado pelo verbo); DF (quando o dominado está à direita do dominante, representado pelo verbo. Exemplo: entre verbo e complemento); GF (quando o dominado está à esquerda do dominante, representado pelo verbo. Exemplo: entre sujeito diretamente seguido do verbo); ID (quando os nós não estão diretamente relacionados, mas estão na mesma árvore); DD (quando o dominado está à direita do dominante, que é normalmente um substantivo. Exemplo: entre substantivo e adjetivo posposto); DG (quando o dominado está à esquerda do dominante, que é normalmente um substantivo. Exemplo: entre substantivo e adjetivo anteposto); IT (quando os dois nós dependem do mesmo nó, que não é verbo. Exemplo: entre adjetivos qualificando um mesmo substantivo); FF (final de frase). Os exemplos que seguem ilustram o processo de marcação. "O gatinho <GF> bebeu <DF> leite <TF> numa tigela <DD> verde <FF>." e "Ontem, <IF> o calmo <DG> gatinho <DD> preto <ID> bebeu <DF> leite <TF> numa tigela <DD> verde <IT> e rosa <FF>."

competência (de natureza sintática) em uma árvore de *performance* (de natureza prosódica). A árvore de *performance* é obtida pela hierarquia gerada pela força das fronteiras frasais, indicada pela pausa precedendo a fronteira (quando falamos pausa nos referimos a um alongamento da rima seguido ou não de pausa silenciosa e percebido subjetivamente como uma desaceleração da enunciação. Cf. Duez, 1987 e Barbosa & Bailly, 1997).

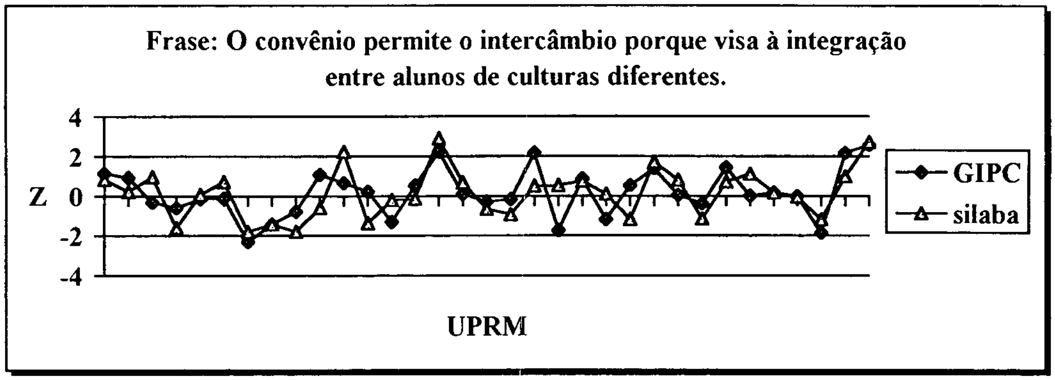


Figura 4: Z-scores (eixo vertical) para sílabas e GIPCs na frase "O convênio permite o intercâmbio porque visa à integração entre alunos de culturas diferentes." Os traços no eixo horizontal representam a posição da vogal da sentença (assinalando tanto a sílaba como o GIPC, visto que as duas unidades têm a vogal em comum). O último valor de z-score (GIPC) mostrado aqui é na verdade o *z-score* da última rima da frase (/ɪs/).

Uma análise de correlação entre os *z-scores* de segmentos adjacentes (obtidos com a fórmula 3), também foi realizada para se verificar a coerência entre os segmentos que formam a sílaba ou o GIPC nas posições de acento lexical e acento frasal , segundo a hipótese de elasticidade homogênea enunciada por Campbell. Para realizar tal correlação, todos os segmentos foram categorizados com três etiquetas: *onset*, *núcleo*, *coda*. Cada consoante integrante do ataque silábico recebeu a etiqueta *onset*, a vogal do núcleo recebeu a etiqueta *núcleo* e as consoantes e vogais assilábicas da coda receberam a etiqueta *coda*. Percebe-se então que, quando se fala de correlação entre segmentos adjacentes e em seqüência, correlações do tipo *onset/núcleo* são diferentes das do tipo *núcleo/onset*. A primeira representa uma seqüência necessariamente na mesma sílaba, a segunda, uma seqüência contendo uma fronteira silábica entre os segmentos. Correlações do tipo onset/onset são possíveis, como no caso de grupos consonantais formados com as líquidas. Os resultados são apresentados na tabela abaixo.

Tabela 2: Correlações (em porcentagem) entre *z-scores* de segmentos adjacentes segundo acentuabilidade. Apenas os valores significativos foram reproduzidos.

	acento lexical	*acento frasal*	*outras posições*
onset/núcleo	63	-31	4
núcleo/onset	ns	26	56
núcleo/coda	48	76	63

Da tabela acima se depreende uma forte coesão da sílaba (sobretudo para a seqüência CV onde C é consoante de *onset* e V é a vogal do núcleo) em posição de acento lexical. A coesão entre *núcleo* e *coda* (correspondendo na maior parte das vezes neste *corpus* à rima da sílaba) é grande em posição de acento frasal, ao mesmo tempo em que se observa uma decorrelação entre o *onset* e o *núcleo* na mesma posição. Nas outras posições, o GIPC parece ser a unidade mais coesa pois as seqüências *núcleo/onset* e *núcleo/coda* estão bem correlacionadas mas a seqüência *onset/núcleo* (entre onset e núcleo de uma mesma sílaba) tem uma correlação baixa. Também se depreende da tabela 2 que a hipótese forte de elasticidade não é válida para o GIPC ou a sílaba. Se fosse o caso, a correlação deveria ser de 100%[10]. Somente uma hipótese fraca de elasticidade uniforme da sílaba em posição de acento lexical e do GIPC, nas demais posições é possível a partir desses resultados[11].

Os resultados acima confirmam que, para o PB, ao menos duas UPRM são necessárias para a caracterização de sua estrutura rítmica: a sílaba e o GIPC. A acentuabilidade lexical é carregada pela sílaba como um todo (Massini, 1991) enquanto que a frasal, pelo GIPC como um todo. Em termos de produção, os resultados obtidos parecem indicar que os gestos de abertura e fechamento da mandíbula (associados à sílaba) são acentuados (hiperarticulados, segundo a tipologia hipo- e hiperarticulação de Lindblom, 1990) em posição de acento lexical enquanto que os gestos de fechamento da mandíbula (associados ao GIPC) formam um todo homogêneo que é hiperarticulado em posição de acento frasal (acento lexical carregando informação prosódica adicional). Uma das conseqüências desta tipologia acentual é que há segmentos que nunca são hiperarticulados, podendo ser melhor caracterizados como segmentos fracos (cf. Albano *et al.*, 1997).

A coerência entre as estruturas sintático-fonológicas (no domínio da competência) e a realização do contornos duracionais (no domínio da *performance*) representada pela evolução dos *z-scores* das UPRM abre a possibilidade de geração automática. Para o aprendizado com redes conexionistas, por exemplo, os contornos duracionais expressos pelo z-score são mais homogêneos que aqueles que seriam obtidos com contornos expressos pela duração observada (como na figura 1) ou mesmo pela duração observada expressa em porcentagem (da média das durações dos GIPCs da frase ou da duração total da

[10] Além de uma inclinação de 1 para a reta correspondente à regressão linear.

[11] Para o francês, a hipótese fraca também foi verificada para os segmentos do GIPC (Barbosa, 1994).

frase, por exemplo). Além disso, por ser um multiplicador do desvio-padrão, medindo em média um terço da duração média (ver tabela 1), um erro cometido no aprendizado da rede com as curvas *z-score* é menos conseqüente do que a mesma porcentagem de erro cometida com as curvas de duração observada. Estes aspectos motivaram o aperfeiçoamento do modelo de geração da duração segmental (desenvolvido originalmente com o francês) para o PB.

A GERAÇÃO AUTOMÁTICA DA DURAÇÃO SEGMENTAL

As curvas rítmicas exemplificadas acima podem ser aprendidas por uma rede conexionista que descreva em sua entrada a informação sintático-fonológica pertinente para a caracterização do *z-score* das UPRM. No caso do PB utilizou-se um perceptron multicamadas com 17 neurônios para a entrada, 2 neurônios para a saída e 25 neurônios na camada escondida. Para a entrada os parâmetros usados foram: linha de declinação, relógio interno (estimativa do período médio do GIPC expresso em segundos), acentuabilidade da unidade (tônica, pré-tônica e pós-tônica) corrente e das três anteriores e das três posteriores, marca sintática dominando o GIPC corrente e marca seguinte, vogal corrente, três anteriores e vogal seguinte, número de consoantes do GIPC e da rima e função dente de serra cujo período é o número de GIPCs de cada grupo acentual. Para a saída, os valores dos *z-scores* da sílaba e do GIPC.

A rede foi treinada com 39 frases do *corpus*. O grau de convergência é bastante satisfatório. Uma vez terminado o aprendizado, testar-se-á seu grau de generalização para as outras 50 frases do *corpus*. Generalizar significa ser capaz de prever os padrões duracionais das frases que não faziam parte do conjunto usado na fase de aprendizado.

Os valores de *z-score* fornecidos pela rede são usados pelo modelo de repartição (fórmula 3) para a obtenção das durações segmentais. Os erros cometidos na atribuição da duração são fruto dos erros cometidos pela rede conexionista e pelo modelo de repartição (pois a hipótese de elasticidade foi enfraquecida).

Pode-se testar o desempenho do modelo de repartição (operacional), com os valores teóricos do *z-score* das UPRM (o que corresponderia a um aprendizado perfeito da rede conexionista e, portanto, de caráter ideal). Os erros cometidos (em relação às durações observadas obtidas pela segmentação manual) por tal modelo aprensentam média nula (esperada, visto que conserva-se em nosso modelo, a duração de uma unidade de nível superior ao fonema: um erro aumentando o valor da duração de um segmento diminui necessariamente na mesma proporção os valores das durações dos segmentos fazendo parte da mesma UPRM) e desvio-padrão de 20 ms. Considerando que a distribuição dos erros é quase-gaussiana, isto significa que 68% dos erros cometidos são inferiores ou iguais a 20 ms e que 97% dos erros cometidos são inferiores ou iguais a 40 ms. Independentemente de se saber se tais erros estão acima ou abaixo do limiar de percepção, é importante lembrar um resultado obtido com um teste de percepção realizado para o PB (Barbosa, 1997a, 1997b).

Neste teste gerou-se frases cujas durações segmentais naturais foram modificadas de acordo com dois modelos, ambos com a mesma distribuição de erros. A diferença consistia no fato de que um dos modelos (o nosso) procurava preservar a posição dos *onsets* vocálicos da frase natural[12] na frase modificada. O outro distribua os erros aleatoriamente (com distribuicão gaussiana) na frase modificada. Convém notar que este teste não foi realizado com fala sintética, pois frases assim obtidas seriam de difícil avaliação quanto à naturalidade, dada a precariedade da concatenação ao nível segmental (cf. seção *Métodos de síntese*).

Ao serem solicitados, em um teste do tipo ABBA, para designarem a frase que lhes parecia mais natural, os 15 sujeitos preferiram em 67% dos casos, a frase gerada pelo nosso modelo de geração. Este resultado é um indício de que um modelo que busca conservar a duração de unidades macrofonêmicas assegura, ao menos em parte, a ritmicidade da frase[13]. Pesquisas tanto em produção quanto em percepção de fala mostram que a existência de tais unidades parece estar alicerçada na realidade cognitiva e psicoacústica.

DISCUSSÃO: AS REFERÊNCIAS COGNITIVA E PSICOACÚSTICA DO RITMO

Tendo estudado a coordenação absoluta dos movimentos rítmicos do corpo humano, Turvey e colegas (1990) propõem um modelo representativo do ato de locomoção baseado em duas funções: uma função de manutenção de temporalidade (*timekeeping function*), executada por células centrais ou por populações de células centrais que produzem um sinal periódico (relógio interno) e uma função motora (*motor function*) que, realizada por populações de células centrais que se servem da referência representada pelo sinal periódico, transmite impulsos aos músculos. Este sinal periódico parece ser a referência para o controle do tempo do ato de enunciação como sugerido por Allen (1973).

Resultados de estudos de coordenação entre o *tapping* (batimento regular do dedo sobre uma superfície plana) e um estímulo periódico externo (uma seqüência de sílabas) tendem a reforçar a hipótese do relógio interno (Fraisse, 1974; Allen 1975). A velocidade de correção dos sujeitos após perturbações introduzidas na seqüência sonora (de forma a manter o sincronismo entre as seqüências de *tapping* e sonora cujo ponto de contato se situa em torno do *onset* acústico da vogal) levam Semjen (1992) a supor a existência de um relógio interno único. O relógio interno funciona como um verdadeiro marca-passo, constituindo uma referência temporal cognitiva para as atividades rítmicas. Esta linha de pensamento encontra respaldo em uma concepção filosófica cognitivista do tempo, em que

[12] Para o francês, o GIPC é suficiente para descrever os padrões duracionais. É importante lembrar que o francês é uma língua sem acento lexical.

[13] Ritmicidade também revelada pelas modulações de freqüência fundamental e energia (cf. Madureira, 1997, 1994 e Aubergé, 1990).

"a estrutura do tempo é dada *a priori* para um agente cognitivo, que dela faz uso para representar os processos temporais com os quais se depara em sua experiência." (Pereira Jr., 1995).

Este relógio interno deve funcionar como referência não apenas para o indivíduo em atividade enunciativa, como também para o seu provável auditor. É o que foi buscado pelos defensores de uma noção de isocronismo em percepção de fala (Allen, 1975; Lehiste, 1977), ou seja, a tendência a perceber pontos de referência presentes no sinal acústico como ocorrendo a intervalos regulares de tempo. Este ponto de referência foi chamado de *perceptual-center* por Marcus (1976). As experiências de Pompino-Marschall (1989) demonstraram que ele se encontra na vizinhança do *onset* acústico da vogal (e não em algum *onset* articulatório inferido a partir do sinal, como sugerido por Fowler, 1979).

A predominância de pontos de referência temporal em torno da vogal sugere que as primeiras unidades de percepção seriam suprafonêmicas e do tamanho da sílaba. De fato, os trabalhos de Studdert-Kennedy e Mehler e colegas buscam mostrar que "phones are not directly perceived, but must rather be derived from a running analysis of the signal over stretches of at least syllable length." (Liberman & Studdert-Kennedy, 1978, p. 153, *apud* Studdert-Kennedy, 1981). Fonemas são derivados a posteriori pelo auditor porque este aprendeu a falar e sabe como estas unidades funcionam em seu sistema fonológico (Studdert-Kennedy, 1981).

Os resultados obtidos por nós para o francês, em produção e percepção de fala, e para o PB, em produção, aliados aos desta seção, para o inglês, reforçam a necessidade de se conceber a existência de unidades de nível superior ao segmento, ao menos no que diz respeito à substância da expressão.

Mesmo que os dados aqui discutidos só tenham sido validados para o inglês, o francês e o português, tendo em vista que os dados cognitivos e psicoacústicos são próprios do homem, abre-se a possibilidade da universalidade do fenômeno. Porém, mais dados empíricos (sobretudo de outras línguas) são necessários para se confirmar tal asserção.

Conclusão

A proposta de um modelo de geração automática da duração segmental que preserva unidades de programação rítmica (do tamanho da sílaba) operantes em PB possibilita utilizar conhecimento científico de ponta para garantir um melhor desempenho das máquinas falantes. A automaticidade do modelo apresentado fornece a capacitação tecnológica para que o gerador da estruturação rítmica do PB seja inserido em um sistema de síntese de fala de alta qualidade.

Agradecimentos

Algumas idéias que permeiam este artigo se encontram desenvolvidas com mais vagar em nossa tese de doutorado (Barbosa, 1994) financiada pelo CNPq. O trabalho aqui apresentado com o PB foi realizado a nível de pós-doutoramento com bolsa da Fapesp (processo nº 94/3358-0 vinculado ao projeto temático "Processamento de texto e sinal acústico em português: uma interface lingüística-engenharia para a ciência e tecnologia de fala", proc. nº 93/0565-2 cujo outorgado é Eleonora Cavalcante Albano}. Agradeço aos locutores por nos terem emprestado suas vozes, ao CPQD-Telebrás pela gentil cessão de seus *corpora* e particularmente a Eleonora C. Albano e a Sandra Madureira pelos comentários e sugestões.

REFERÊNCIAS BIBLIOGRÁFICAS

Abercrombie, D. (1967) *Elements of General Phonetics*, Edinburgh University Press.

Abry, C. *et al.* (1985) *Un modèle de congruence relationnel pour la synthèse de la prosodie du français.* Actes des 14^{es} Journées d' Étude du Groupe Communication Parlée,Paris, 156-163.

Albano. E.C., Silva, A.P., Moreira, A.A., Aquino, P.A & Kakinohana, R.K. (1997) *Um conversor ortográfico-fônico e uma notação prosódica mínima para síntese de fala em língua portuguesa.* Neste volume.

Allen, G. D. (1975) *Speech rhythm: its relation to performance universals and articulatory timing.* Journal of Phonetics 3, 75-86.

Allen, G. D. (1973) *Segmental timing control in speech production.* Journal of Phonetics 1, 219-237.

Aubergé, V. (1990) *Semi-automatic constitution of a prosodic contour lexicons for the text-to-speech synthesis*, Proc. of the ESCA Workshop on Speech Synthesis, Autrans, 215-218.

Bailly, G. (1986) *Un modèle de congruence relationnel pour la synthèse de la prosodie du français.* Actes des 15^{es} Journées d' Étude sur la Parole, Aix-en-Provence, 75-78.

Bailly, G., Laboissière, R & Schwartz, J.-L. (1991) *Formant trajectories as audible gestures: an alternative for speech synthesis.* Journal of Phonetics 19(1), 9-23.

Barbosa, P. A. (1997a) A model of segment (and pause) duration generation for Brazilian Portuguese text-to-speech synthesis. A aparecer em Proceedings of the EUROSPEECH'97, 22 a 25 de setembro, Rodes, Grécia.

Barbosa, P. A. (1997b) *At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration: emphasis on segmental duration generation.* A aparecer em Caderno de Estudos Lingüísticos, Unicamp.

Barbosa, P. A. (1995) *Conexionismo: estruturas neuronais, mentais e aplicações em prosódia.* Resenha de Conferência ministrada na Escola Paulista de Medicina para a Sociedade Brasileira de Neuropsicologia-SBNp, 27 de abril.

Barbosa, P. A. (1994) *Caractérisation et génération automatique de la structuration rythmique du français*, Thèse de doctorat de troisième cycle. ICP/INP de Grenoble, França.

Barbosa, P. A. & Bailly, G. (1997) *Generation of pauses within the z-score model.* In: *Progress in Speech Synthesis.* van Santen, J.P.H., Sproat, R.W., Olive, J.P. & Hirshberg, J. (Eds.), New York: Springer-Verlag, 365-381.

Benoît, C. (1990) An *intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity.* Speech Communication 9, 293-303.

Benveniste, E. (1951) *La notion de « rythme » dans son expression linguistique.* J. Psychol. Norm. Path. 44, 401-411.

Bergson, H. (1968) *Durée et simultanéité.* Paris: Presses Universitaires de France.

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

Calliope (1989) *La Parole et son traitement automatique.* Paris: Masson.

Campbell, N.W. (1992) *Syllable-based segmental duration.* In: Talking Machines: theories, models, and designs (Bailly, G. & Benoît, C. Eds.) 211-224. Elsevier B.V.

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

Coker, C.H. (1968) *Speech synthesis with a parametric articulatory model.* In: Flanagan, J.L. & Rabiner, L.R. (Eds.) Speech Symposium, reprinted in Speech Synthesis, 135-139. Dowden, Hutchinson and Ross, Stroudsburg, PA.

Collier, R. (1992) *A comment on the prediction of prosody.* In: Talking Machines: theories, models, and designs (Bailly, G. & Benoît, C. Eds.) 205-208. Elsevier B.V.

Crystal. D. (Ed.) (1985) *A dictionnary of Linguistics and Phonetics.* Basil Blacwell in association with André Deutsch.

Dauer, R. M. (1983) *Stress-timing and syllable-timing re-analysed*, Journal of Phonetics, 11, 51-62.

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

Edwards, J., Beckman, M.E. & Fletcher, J. (1991) *The articulatory kinematics of final lengthening*. Journal of the Acoustical Society of America 89 (1), 369-382.

Flanagan, J.-L. & Rabiner, L.-R. (1973) *Speech synthesis*. Benchmark papers in acoustics. Dowden, Hutchinson & Roos inc., Stroudsburg, Pennsylvania.

Fowler, C. (1981) *Production and perception of coarticulation among stressed and unstressed vowels*, J.S.H.R., 47, 127-139.

Fowler, C. (1979) *"Perceptual-centres" in speech production and perception*, Perception and Psychophysics, 25, 375-388.

Fraisse, P. (1974) *La psychologie du rythme*, Paris: Presses Universitaires de France.

Grosjean, F. & Dommergues, J-Y (1983) *Les Structures de performance en psycholinguistique*, L'Année Psychologique, 83, 513-536.

Hirst, D.J. (1993) *Peak, boundary and cohesion characteristics of prosodic grouping*. Proc. ESCA Workshop on Prosody, Lund, Suécia, 27 a 29 de setembro, 32-37.

ICP (1994) *Du faire-semblant au faire-comme*. Équipe Synthèse. Rapport d'Activité de l'Institut de la Communication Parlée. 49-50.

Klatt, D.H (1987) *Review of text-to-speech conversion for English*, J. Acoust. Soc. Am. 82, 737-793.

Köster, J.-P. (1973) *Historische Entwicklung von Syntheseapparaten*. Hamburg: Helmut Buske Verlag Hamburg.

Kozhevnikov, V.A. & Chistovich, L.A. (1965) *Speech articulation and perception*. In: Joint Publications Research Service, 543.

Lehiste, I. (1978) *Temporal organization and prosody. Perceptual aspects*. In Joint Meeting of A.S.A. and A.S.J. 1, Honolulu, 1-17.

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

Lehiste, I. (1970) *Suprasegmentals*. Cambridge, Massachussets: MIT Press.

Leung, H.C. & Zue, V. W. (1984) *A procedure for automatic alignment of phonetic transcriptions with continuous speech*. Proceedings of the IEEE ICASSP, 1, San Diego, 2.7.1-2.7.4.

Lindblom, B. (1990) *Explaining phonetic variation: a sketch of the H & H theory*. In: Hardcastle, H.J. & Marchal, A. (Eds.) *Speech Production and Speech Modelling*, 403-440, Dordrecht: Kluwer.

Lippmann, R. (1987) *An introduction to computing with neural nets*. IEEE on Acoustics, Speech and Signal Processing Magazine, 4-22.

Madureira, S. (1997) *Entoação e síntese de fala: modelos e parâmetros*. Neste volume.

Madureira, S. (1994) *Pitch patterns in Brazilian Portuguese: an acoustic phonetic analysis*. Vth Australian International Conference on Speech Science and Technology, 5 a 9 de Dezembro, Perth, Austrália.

Major, R. C.(1981) *Stress-timing in Brazilian Portuguese*, Journal of Phonetics, 9, 343-351.

Marcus, S.M. (1981) *Acoustic determinants of Perceptual-center (p-center) location*, Perception and Psychophysics, 30(3), 247-256.

Marcus, S.M (1976) *Perceptual-centres*. Unpublished PhD Thesis, Cambridge University.

Martin, P. (1981) *L'Intonation est-elle une structure congruente à la syntaxe ?* In: *L'Intonation : de l'acoustique à la sémantique*, 234-271. Paris: Klincksieck.

Massini, G. (1991) *A Duração no estudo do acento e do ritmo em português*, Tese de Mestrado, Unicamp.

Morton, Marcus & Frankish (1976) *Perceptual-centers (P-centers)*, Psychological Revue, 83 (5), 405-408.

Pereira Jr., A. (1995) *Tempo e irreversibilidade física: algumas distinções conceituais*. Manuscrito, XVII (1), 97-152.

Pompino-Marschall, B. (1989) *On the psychoacoustic nature of the P-center phenomenon*, Journal of Phonetics, 17, 175-192.

Rosenblatt, R. (1959) *Principles of neurodynamics*. New York: Spartan Books.

Sousa, E. M. G. (1995*) Towards an Acoustic Description of Brazilian Portuguese Nasal Vowels*. XIII International Congress of Phonetic Sciences, 13 a 19 de Agosto, Estocolmo, Suécia.

Semjen, A., Schulze, H.-H. & Vorberg, D. (1992) *Temporal control in the coordination between repetitive tapping and periodic external stimuli*. Fourth Rhythm Workshop: Rhythm Perception and Production, 73-78. Bourges, França, Junho.

Studdert-Kennedy, M. (1981) *Perceiving phonetic segments*. In: The Cognitive Representation of Speech. Myers, T., Laver, J. & Anderson, J. (Eds.) The Netherlands: Elsevier Science Publishers B.V.

Tesnière, L. (1965) *Éléments de syntaxe structurale*. Paris: Klincksieck.

Turvey, M.T., Schmidt, R.C. & Rosenblum, L. (1990) *Clock and motor components in absolute coordination of rhythmic movements*. Status Report on Speech Research SR-101/102, Haskins Labs.

van Santen, J.P.H. (1994) *Assignment of segmental duration in text-to-speech synthesis*. Computer, Speech and Language 8, 95-128.

Wenk, B.J. & Wioland, F. (1982) *Is French really syllable-timed?* Journal of Phonetics, 10, 193-216.

Whalen, D.(1990) *Coarticulation is largely planned.* Journal of Phonetics 18 (1), 3-35.

Woodrow, H. (1951) *Time perception.* In: Stevens, S. (Ed.) Handbook of Experimental psychology. 1224-1236. New York: Wiley.

APÊNDICE: AS REDES CONEXIONISTAS

Uma rede conexionista ou rede de neurônios formal consiste fisicamente em um conjunto de processadores elementares interconectados chamados nós ou neurônios formais. Alguns destes neurônios constituem a entrada da rede, outros constituem a saída. A maneira pela qual os neurônios estão ligados definem a arquitetura da rede (ver Barbosa, 1995 para resumo histórico).

Cada neurônio recebe em sua entrada um conjunto de conexões provindas de um sub-conjunto de neurônios da rede. Um peso é associado a cada conexão, representando a influência do neurônio aferente sobre o neurônio eferente. Define-se um valor de ativação para o neurônio eferente por um cálculo que, na maioria das vezes, é obtido em duas etapas: (a) combinação linear dos valores de ativação dos neurônios aferentes ponderados pelos pesos das respectivas conexões e (b) aplicação de uma função diferenciável e limitada – normalmente a função sigmóide $f(x) = 1/(1 + \exp(-x))$.

Os valores de ativação dos neurônios de saída são os valores de saída da rede.

A finalidade de uma rede conexionista é simular uma processo físico ou uma função quaisquer sem a necessidade de se conhecer sua descrição analítica. O que a torna de utilidade comprovada, visto que tal descrição é muitas vezes desconhecida. Assim, uma rede com aprendizado supervisionado (Lippmann, 1987), como a usada em nosso trabalho, deve aprender a associar um espaço de entrada a um espaço de saída por meio da apresentação sucessiva de pares entrada/saída representativos do fenômeno estudado. Isto é realizado por meio de uma regra de aprendizado que, pela modificação dos pesos das conexões, busca aproximar a saída desejada (dos pares entrada/saída apresentados) à saída atual da rede (obtida aplicando-se os processos calculatórios definidos para os neurônios).

Ao terminar o aprendizado, a rede está pronta para operar. Quanto mais representativos do fenômeno subjacente forem os exemplos entrada/saída apresentados à rede na fase de aprendizado, maior será a possibilidade da generalizar: a uma nova entrada (não apresentada anteriormente), ela fornecerá uma saída que é uma boa aproximação da resposta (à mesma entrada) do sistema simulado.

O *perceptron* foi a primeira rede conexionista (Rosenblatt, 1959). Ele é organizado em camadas de neurônios. A primeira forma a entrada, a última, a saída e as demais, as camadas escondidas. Todas as conexões vão do sentido da entrada para a saída (os neurônios na mesma camada não estão ligados entre si) e cada neurônio de uma camada recebe normalmente (diz-se que o perceptron está densamente conectado) conexões de todos os neurônios da camada anterior.

A INTERFACE FONÉTICA-FONOLOGIA E A INTERAÇÃO PROSÓDIA-SEGMENTOS

Eleonora ALBANO (Coordenadora - UNICAMP), Plínio BARBOSA (UNICAMP), Aglael GAMA-ROSSI (UNICAMP - PG), Sandra MADUREIRA (PUC - SP), Adelaide SILVA (UNICAMP - PG)

ABSTRACT: *The phonetics-phonology interface is discussed in light of acoustic phonetic data on the interaction between prosody and segments. While agreeing with the spirit of current trends such as Articulatory Phonology in countering the traditional component sequence "phonology-linguistic phonetics-universal phonetics", the data speak against effacing the phonetics-phonology distinction.*

KEY WORDS: *phonetics, phonology, interface, prosody, segments.*

0. Introdução

Este grupo de trabalho discutiu a interface Fonética-Fonologia a partir de dados fonético-acústicos sobre a interação prosódia-segmentos. A variação dos segmentos sob diferentes condições prosódicas interessa ao debate sobre as relações entre a Fonética e a Fonologia porque tende a envolver diferenças fonéticas que parecem não ser binárias, mas gradientes. Gradientes passíveis de redução a escalas discretas podem ser acomodados na Fonologia. Gradientes idênticos a contínuos físicos só fazem sentido dentro da Fonética.

1. Histórico

Embora tendo surgido muito depois, o termo Fonologia foi usado como sinônimo de Fonética até o advento da Lingüística contemporânea, que inaugurou a atual distinção. Para o fundador da Fonologia, Trubetzkoy, a Fonética e a Fonologia eram tão distantes como a Numismática e a Economia (1964, p. 12).

A literatura estruturalista pouco disse sobre as relações entre as duas disciplinas além de afirmar a sua separação. Coube à Fonologia Gerativa fazer a primeira proposta explícita a esse respeito. Ela está no capítulo 7 de SPE (Chomsky e Halle 1968, pp. 293-329), onde se elaboram as diferenças entre sistemas lingüísticos e sistemas físicos.

Para isolar as fontes de variabilidade lingüística das demais, Chomsky e Halle propõem derivar o sinal de fala em três etapas. Na primeira, uma forma básica da palavra ou morfema é transformada numa forma superficial pela manipulação de traços binários, num componente da gramática denominado fonológico, onde se introduzem todas as variações devidas ao contexto morfológico ou fonológico. Na segunda etapa, essa forma superficial é especificada conforme a língua, em outro componente da gramática, denominado fonético. Esse introduz o chamado "detalhe fonético lingüístico" pela manipulação de traços gradientes, que se distinguem, contudo, de verdadeiros contínuos físicos por exibirem gradações discretas. Finalmente, a forma fonética resultante dá entrada no sistema de produção do falante, que é constituído por características individuais subordinadas a restrições fonéticas universais. Surge então o objeto físico que é o sinal de fala.

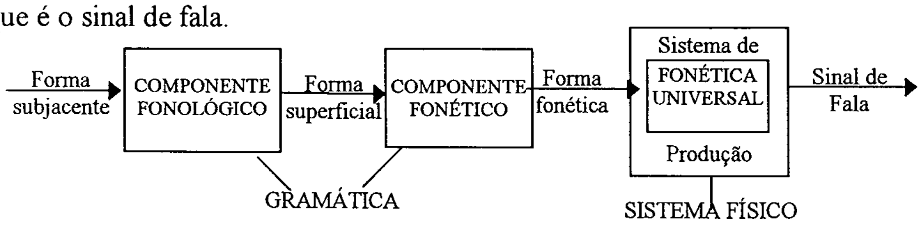


Figura 1 - Relações entre a Fonética e a Fonologia segundo Chomsky e Halle.

Nessa concepção, os componentes da gramática distinguem-se dos sistemas físicos por manipularem símbolos (binários ou escalares) e não, como esses, números. Um exemplo de regra fonológica é a transformação de um segmento em outro, tal como acontece com o /t/ e o /d/ do inglês americano, que podem ambos se realizar como *flaps* em posição intervocálica. Um exemplo de regra de detalhe fonético é o ligeiro alongamento da vogal diante de consoante sonora, que ocorre em inglês bem como em outras línguas. Segundo Chomsky e Halle, essa regra, embora muito comum, não chega a ser um automatismo universal, pois se manifesta diferentemente de língua para língua. Automática seria apenas a interpretação das escalas do componente fonético em termos de restrições articulatórias e auditivas universais e/ou individuais.

A literatura fonética dos últimos vinte anos vem mostrando que a idealização da Figura 1 não corresponde à realidade. Fox e Terbeek (1977) demonstraram que a regra de *flapping* do inglês americano não neutraliza de fato a distinção /t,d/, pois a vogal que precede o *flap* proveniente de /d/ é mais longa que aquela que precede o *flap* proveniente de /t/. Ora, nos termos da Figura 1, esse achado é extremamente problemático: o *flapping* é uma regra fonológica, enquanto o alongamento da vogal antes de consoante sonora é uma regra de detalhe fonético. A segunda não deveria, portanto, aplicar-se depois da primeira, pois o *flap* é sempre sonoro. Ou se postula *ad hoc* um *flap* "com memória surda", ou se revêem os critérios para a separação entre os componentes.

Três tipos de argumentos se acumulam na literatura para questionar essa separação. Há, como no exemplo acima, paradoxos de ordenação de regras, sugerindo que a oposição binário/gradiente não define componentes (p. ex., Port e O' Dell 1985). Há também evidências de que certas distinções tradicionalmente vistas como binárias são, na verdade, gradientes (p.ex., Zue e Shattuck-Hufnagel 1980). Há, finalmente, o leque de variação interlingüística dos parâmetros ditos gradientes, tão grande que faz suspeitar que se trata, de fato, de contínuos (p. ex., Keating 1985).

Problemas como esses têm levado alguns autores a questionar não só as escalas fonéticas, que tenderiam a se expandir indefinidamente para acomodar diferenças interlingüísticas, como também as próprias unidades fonológicas, tais como o segmento ou o traço distintivo. Browman e Goldstein (1992), no que se refere aos segmentos e traços, e Port, Cummins e Gasser (1995), no que se refere ao ritmo, propõem a fusão da Fonética e da Fonologia num componente único que representaria diretamente a dinâmica da produção da fala:

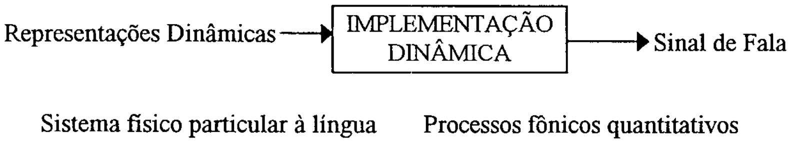

Representações Dinâmicas
IMPLEMENTAÇÃO DINÂMICA
Sinal de Fala
Sistema físico particular à língua
Processos fônicos quantitativos

Figura 2 - Fusão da Fonética com a Fonologia segundo propostas recentes.

Assim, ao invés de segmentos, traços, sílabas e acentos estáticos, haveria gestos articulatórios dotados de tempo intrínseco - i.e., pulsações que, agindo como osciladores ou molas, "empurrariam" os articuladores. Essas entidades dinâmicas não traduziriam símbolos estáticos, mas constituiriam as próprias unidades lingüísticas representadas no léxico - obliterando a distinção entre gramáticas e sistemas físicos. A variação lingüística da pronúncia, antes captada por processos simbólicos de reescrita, seria agora captada exclusivamente por processos - necessariamente quantitativos e particulares a línguas - de implementação física dessas representações dinâmicas.

O discurso revolucionário desses autores nem sempre justifica claramente a afirmação radical de que existem sistemas físicos vinculados a línguas particulares. Para tanto, seria preciso demonstrar (a) que há muitos gradientes fonéticos lingüisticamente pertinentes; (b) que processos fônicos tradicionalmente vistos como categóricos são, na verdade, gradientes; e (c) que os gradientes fonéticos têm uma lógica semelhante à dos contínuos físicos. Foi esse o objetivo das exposições resumidas a seguir.

2. Ritmo do português adulto

Plínio Barbosa mostrou que o ritmo do português tem aspectos quantitativos captáveis pela noção de contorno duracional. Na palavra, a duração das sílabas forma contornos que culminam no acento lexical, lembrando os graus acentuais de Mattoso Câmara (Câmara Jr. 1969). Na frase, uma outra unidade do tamanho da sílaba, compreendida entre duas vogais - o grupo *inter perceptual center* (GIPC) - é o lugar da culminação. Assim, os acentos lexicais alongam não só a vogal, mas também a consoante da sílaba acentuada, enquanto os acentos frasais alongam também a consoante seguinte à vogal acentuada.

A variabilidade interlingüística dos contornos duracionais é atestada pelo próprio trabalho anterior de Barbosa (1994), que mostra que a unidade da culminação em francês é o GIPC. O caráter quantitativo - e não meramente escalar - desses contornos é, por sua vez, corroborado pelo grande número de gradações de cada frase e pela sua sensibilidade à taxa de elocução. A Figura 3 mostra os contornos da frase "Ele guarda a sela do cavalo numa prateleira de uma antiga cela", emitida nas taxas lenta, normal e rápida. Notar que a última tem picos mais abruptos e agrega elementos lexicalmente distintos, tais como "ele" e "guarda".

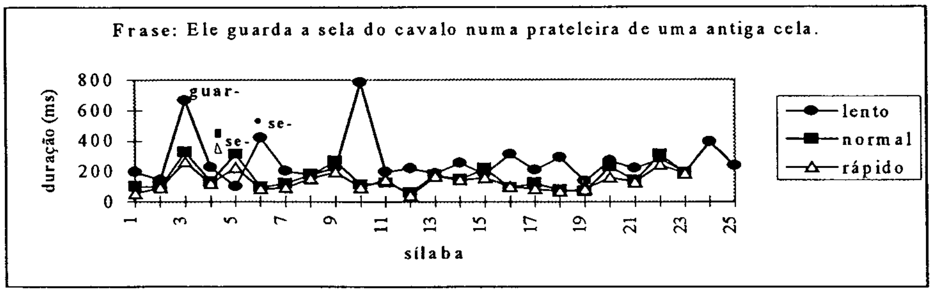


Figura 3 - Variação de contornos duracionais conforme taxa de elocução.

O papel dos contornos duracionais em caracterizar a pronúncia brasileira pode ser estudado pela sua implementação computacional para a fala sintética, trabalho que Barbosa vem desenvolvendo atualmente.

3. Ritmo do português infantil

Uma outra maneira de inquirir a pertinência lingüística dos contornos duracionais é estudar a sua aquisição. Aglael Gama-Rossi abordou essa questão com um estudo das diferenças fonético-acústicas entre crianças e adultos.

Numa tarefa de repetição de frases com apoio de estímulos visuais, as durações de segmentos e sílabas de crianças na faixa dos 4-5 anos foram comparadas às do modelo adulto. Um resultado típico é o da Figura 4, onde se vê que, embora não sendo idênticas, as durações das sílabas da criança e do adulto se aproximam o suficiente para dizermos que se trata do mesmo contorno.

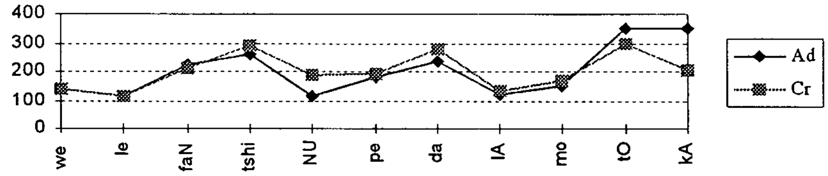


Figura 4 - Coerência entre contornos duracionais de uma criança e de um adulto-modelo.

Parece, pois, que a criança não está propriamente imitando as durações da fala adulta, mas reproduzindo, com os seus próprios meios, relações duracionais ritmicamente significativas. O grande número de gradações reforça o argumento discutido a propósito da exposição de Barbosa.

4. Qualidade de voz e redução de vogal em diferentes fronteiras prosódicas

Outro parâmetro prosódico cujas gradações são lingüisticamente pertinentes é a qualidade de voz. Sandra Madureira expôs dados em que a vogal final da mesma palavra apresenta graus distintos de laringealização conforme ocorra numa fronteira prosódica forte ou fraca. Observem-se as formas de onda da Figura 5, que correspondem às duas ocorrências de "manga" na frase "Ele sujou a *manga* da camisa ao comer uma suculenta *manga*" (fronteiras, respectivamente, de palavra fonológica e de sintagma entoacional).

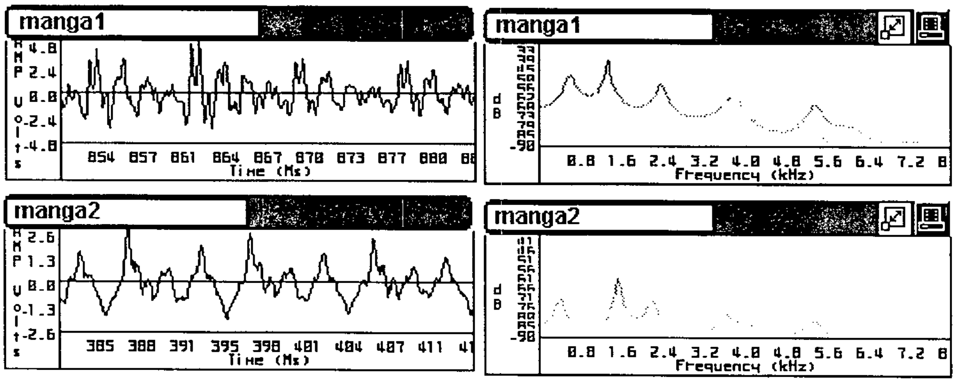


Figura 5 - Diferentes graus de laringealização e de redução de /a/ em fronteiras prosódicas.

Madureira demonstrou ainda que essas mesmas fronteiras afetam também um processo segmental: a redução vocálica. Assim, o /a/ final do segundo "manga" tem um F1 mais baixo e um F2 mais alto do que os do primeiro, apesar de serem ambos reduzidos.

Esses achados engrossam o rol dos gradientes fonéticos lingüisticamente pertinentes e conduzem à segunda linha de argumentação acima proposta: a redução de vogais, processo geralmente visto como fonológico, parece, com esse comportamento gradiente, entender-se melhor como fonético.

5. Espirantização da vibrante apical

Outras evidências de que processos tradicionalmente vistos como categóricos podem ser, de fato, gradientes foram apresentadas por Adelaide Silva.

Leituras repetidas de um texto por um informante cujo /r/ forte é em geral uma vibrante apical foram examinadas com vistas ao processo de espirantização. Como se vê na Figura 6, além da vibrante canônica constituída de curtos segmentos vocálicos separados por breves oclusões e da sua variante fricativa, onde não se discernem esses eventos, surgiu uma pronúncia intermediária, onde o ruído fricativo se superpõe a uma vibrante suavizada:

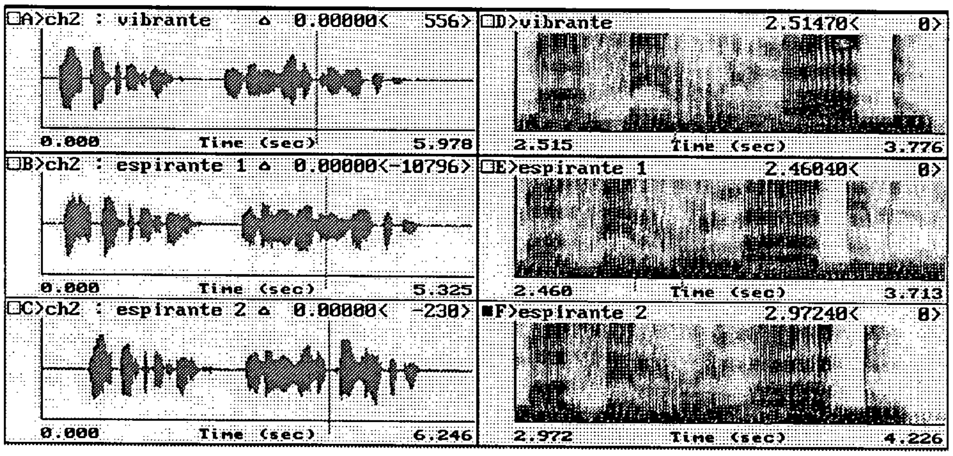


Figura 6 - Diferentes graus de espirantização da vibrante apical.

É interessante observar que, embora se trate de leituras repetidas do mesmo texto, esses enunciados têm estruturas prosódicas diferentes, como sugere o padrão de pausas indicado pelo cursor nas formas de onda. Aqui, como no caso acima, a variabilidade potencial das condições prosódicas sob as quais emerge o gradiente faz duvidar que ele se possa exprimir em termos discretos.

6. Elisão parcial de vogal em juntura com vogal

Finalmente um processo cuja gradiência desafia as descrições fonológicas tradicionais foi apresentado por Eleonora Albano. Trata-se da elisão parcial de /a/ átono em juntura com outra vogal átona.

Num experimento de leitura repetida com aumento progressivo da taxa de elocução, foram contrastados pares de frases tais como: "Não se faz isso com um cara idoso" e "Não se faz isso com um caridoso". As pronúncias observadas para a primeira frase variaram desde a lenta, com /a/ e /i/ em hiato, até a rápida, com elisão de /a/, homófona à segunda frase. Ocorreu, porém, também uma leitura intermediária, ambígua entre "cara idoso" e "caridoso", onde, nos 20 primeiros ms após a liberação do /r/, há evidência de um resquício de /a/, manifesto por um F1 superior e um F2 inferior aos de /i/. Superpondo-se, como na Figura 7, os espectros tomados nesse mesmo ponto para as três leituras, tem-se, de fato, a impressão de um contínuo entre o hiato e a elisão da vogal.

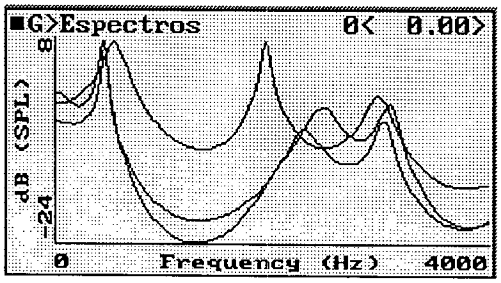


Figura 7 - Diferentes graus de elisão de vogal em fronteira com vogal.

A força desses dados está em que uma escala discreta que varie entre zero e a duração plena da vogal não faz sentido. A duração é um parâmetro sujeito a outros condicionamentos particulares a línguas, tais como o ritmo e as influências dos segmentos adjacentes. Uma escala duracional que acomodasse todas essas possibilidades de gradação tenderia a crescer indefinidamente.

A visão escalar torna-se ainda mais implausível quando se considera a capacidade dos modelos dinâmicos de dar conta de fenômenos como esse. Na verdade, a presença maior ou menor do /a/ na fronteira vocabular se explica facilmente através do mecanismo de superposição de gestos articulatórios proposto por Browman e Goldstein (*op. cit.*): o gesto do /a/ reduz-se em amplitude e partilha o seu tempo progressivamente com os dos segmentos adjacentes, até ocultar-se sob eles.

7. Conclusões

O exposto cumpriu o objetivo de demonstrar a pertinência lingüística dos gradientes fonéticos e a dificuldade de reduzi-los a escalas discretas. Isso corrobora a visão dos autores supracitados a respeito do dinamismo da fala: muitos processos comumente entendidos como fonológicos devem ser vistos como fonéticos, i.e., resultantes da variação contínua de relações dinâmicas entre os articuladores. Cabe perguntar, então, se a proposta de fundir a Fonética com a Fonologia também se sustenta.

A existência de processos fônicos contínuos não abala apenas a histórica teoria de Chomsky e Halle, mas toda a sua descendência. A Fonologia Não-Linear contemporânea, apesar de inegáveis avanços, continua a tratar fenômenos como o enfraquecimento, a assimilação, o apagamento e a inserção de segmentos como reescrita de símbolos, ignorando os problemas levantados pela descoberta das suas versões gradientes. Radicalizar a tese da continuidade negando toda e qualquer representação simbólica dos fenômenos fônicos é rejeitar em bloco toda a Fonologia desde SPE.

O perigo de fazê-lo reside em esvaziar a interface entre a Fonética e a Gramática - agora reduzida aos componentes morfológico e sintático. Sem a Fonologia enquanto gramática dos sons de fala, surge um abismo incomensurável entre os aspectos simbólicos e físicos dos sistemas lingüísticos, visto que a metalinguagem dos foneticistas pouco tem em comum com a dos lingüistas.

Os dados examinados sugerem que uma revolução dessa monta seria prematura. Em primeiro lugar, casos como a redução e elisão de vogais e a espirantização da vibrante recolocam o velho problema da representação lexical: como escolher uma forma vocabular básica ao longo de um contínuo? Em segundo lugar, o caráter distintivo do acento em línguas como o português reitera a questão no domínio da prosódia: como escolher um contorno acentual básico ao longo de um contínuo? Em ambos os casos, arbitrar um valor numérico é ainda mais complicado do que arbitrar uma variante simbólica. Em terceiro lugar, os ambientes disparadores de alguns dos processos discutidos envolvem fronteiras entre constituintes prosódicos tais como a palavra e o sintagma fonológico. É difícil imaginar uma fronteira que não seja abstrata e, portanto, não-fonética. A literatura fonética mostra, aliás, que os índices articulatórios e acústicos das fronteiras não são unívocos e distribuem-se por vários segmentos. Integrá-los numa única representação resulta em postular algo muito semelhante a uma parentetização ou a um pulso métrico mudo.

Finalmente, cabe lembrar que uma parte da fonologia é, em línguas como o português, lexical ou morfofonológica. Sziga (1995) demonstrou que as palatalizações pós-lexicais do inglês são gradientes, enquanto as lexicais parecem ser categóricas, visto que as palatais derivadas são tão estáveis quanto as subjacentes. Isso acresce ao acima exposto para sugerir cautela em abrir mão de representações simbólicas. A Figura 8 mostra uma visão das relações entre a Fonética e a Fonologia igualmente atual, porém mais prudente, que a da Figura 2.

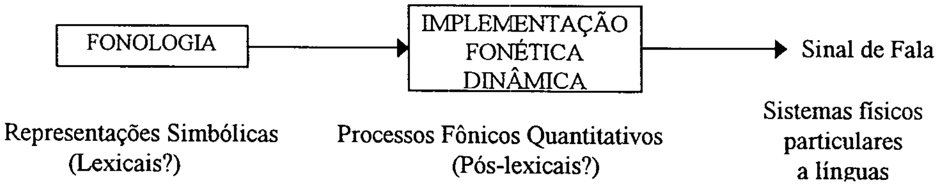

FONOLOGIA
IMPLEMENTAÇÃO FONÉTICA DINÂMICA
Sinal de Fala
Representações Simbólicas (Lexicais?)
Processos Fônicos Quantitativos (Pós-lexicais?)
Sistemas físicos particulares a línguas

Figura 8 - Relações entre a Fonética e a Fonologia segundo a visão do grupo.

De acordo com essa visão, é preciso rediscutir a Fonologia, buscando uma metalinguagem que permita constituir uma interface ótima entre os números da Fonética e os símbolos da Gramática. Sendo o tema extenso e o espaço curto, fica agendada a discussão para o próximo Seminário.

RESUMO: *Discute-se a interface fonética-fonologia à luz de dados fonético-acústicos sobre a interação prosódia-segmentos. Embora corroborando o espírito de visões como a Fonologia Articulatória em rejeitar a organização componencial "fonologia - fonética lingüística - fonética universal", os dados desaconselham a fusão entre a Fonética e a Fonologia.*

PALAVRAS-CHAVE: *Fonética, Fonologia, interface, prosódia, segmentos*

REFERÊNCIAS BIBLIOGRÁFICAS

BARBOSA, P. (1994) Caractérisation et génération automatique de la structuration rythmique du français. Tese de doutorado inédita. ICP-Grenoble.

BROWMAN, C. e GOLDSTEIN, L. (1992) Articulatory phonology: an overview. *Phonetica* **49**: 155-180.

CÂMARA JR., J. M. (1969) *Problemas de lingüística descritiva*. Petrópolis: Vozes.

CHOMSKY, N. e HALLE, M.(1968) *The sound pattern of English*. Nova Iorque: Harper e Row.

FOX, R. e TERBEEK, D. (1977) Dental flaps, vowel duration and rule ordering in American English. *Journal of Phonetics* **5**: 27-34.

KEATING, P. (1985) Universal phonetics and the organization of grammars. In: FROMKIN, V.(org.) *Phonetic Linguistics: essays in honor of Peter Ladefoged*. Nova Iorque: Academic Press, pp.115-132.

PORT, R. e O' DELL, M.(1985) Neutralization of syllable-final voicing in German. *Journal of Phonetics* **13**: 455-471.

_____, CUMMINS, F. e GASSER, M.(1995) A dynamic approach to rhythm in language: toward a temporal phonology. Manuscrito inédito, Indiana University.

SZIGA, E. (1995) An acoustic and eletropalatographic study of lexical and postlexical palatalization in American English. In: CONNELL, B. e ARVANITI, A.(orgs.) *Phonology and phonetic evidence: papers in laboratory phonology IV*. Cambridge: Cambridge University Press, pp. 282-302.

TRUBETZKOY, N. (1964[1939]) *Principes de phonologie*. Paris: Klincksieck.

ZUE, V. e SHATTUCK-HUFNAGEL ,S. (1980) Palatalization of /s/ in American English: when is an /ʃ/ not an /ʃ/? *Journal of the Acoustical Society of America* **67**, S27.

A MODEL OF SEGMENT (AND PAUSE) DURATION GENERATION FOR BRAZILIAN PORTUGUESE TEXT-TO-SPEECH SYNTHESIS

Plínio A. Barbosa
Laboratório de Fonética Acústica e Psicolingüística Experimental-LAFAPE
Instituto de Estudos da Linguagem
Universidade Estadual de Campinas
CP 6045 - 13081-970 Campinas-SP, Brazil
Tel: +55 19 2397784, FAX: +55 19 2391501, E-mail: plinio@iel.unicamp.br

ABSTRACT

This work presents and evaluates a model of segmental duration generation for Brazilian Portuguese where the notion of macrorhythmic unit is the starting point to drastically simplify duration assignment and to allow pause insertion as an integrated procedure of generation. This model is preferred to random assignment with the same error distribution.

Some aspects of rhythm phonetics and phonology are also discussed that constitute a first step to the understanding of the prosodic component of the language under study.

1. INTRODUCTION

An earlier model for French automatic segmental duration generation was proposed that has represented an important advance in rhythmic structure generators for TTS systems ([1]). That model has allowed to straightforwardly assign segment and pause duration as parts of the same mechanism of generation.

This ability is possible thanks to the notion of a normalized duration , the syllable z-score, as outlined by Campbell ([2]). This notion was extended to IPCG (inter-p-center group) and used within "a neural net plus repartition algorithm" model for generating segment and pause durations. Z-scores' computation for IPCGs including possible silence (as a part of final lengthening in French) has enabled duration assignment and automatic insertion of silent pause. It is important to note that the use of macrorhythmic units has drastically simplified the segmental duration assignment task.

An implicit assumption in our framework is to conceive rhythm production as a by-product of the subject's choice of one among many strategies to perform an underlying metrical structure whose building blocks are syllable-size units called rhythmic programming units (RPU). In French, as the subject speaks, syllables (typically CV sequences) at the first (macro)rhythmic level are performed as IPCGs (typically VC units). In Brazilian Portuguese (henceforth BP), on the other hand, it has been shown ([3]) that at least two macrorhythmic units are necessary to model duration: syllables and IPCGs. In that language, lexical stress can be assigned to the last, penultimate (the most frequent) or antepenultimate syllable (oxyton, paroxytons and proparoxytons, respectively) and the acoustic correlates of stress are often the greater duration of the stressed syllable and the decrease of intensity in the post-stressed ones. Stressed syllables can be enhanced as they are uttered by carrying phrasal accent. Only lexically stressable syllables can bear phrasal accent, where duration and pitch play the major roles in building prominence.

The need for considering two RPUs and the existence of (weak) post-stressed syllables constitute challenges for building an automatic BP segmental duration generator. As in French, the BP duration generator also computes segmental durations in two stages. But in this case, the neural net is conceived in such a way as to sequentially map a phonological, prosodic description for each sentence at the input to the corresponding syllable and IPCG z-scores at the output. The repartition algorithm in the second stage distributes a given duration among IPCG segments bearing phrasal accent and among syllable segments lexically stressed (not phrasally prominent in the sentence). As IPCG consonants include next syllable onsets, the assignment of duration to segments is not a simple task.

2. OVERVIEW OF THE Z-SCORE MODEL

Normalized durations are obtained through Campbell's *z-score* model. The z value of each segment *s* is computed by writing:

$$\mathrm{Dur}_s = \exp(\mu_s + z.\sigma_s) \ (1),$$

where Dur_s is segment duration and (μ_s, σ_s) stands for the average and the standard-deviation of the log-transformed durations of all *s* realizations in an *ad hoc corpus*. A strong elasticity hypothesis says that all segments in a RPU frame have the same z-score: a single value of *z* per RPU can then be obtained recursively by writing:

$$\mathrm{Dur}\ (\mathrm{RPU}) = \textstyle\sum_s \exp(\mu_s + z.\sigma_s) \ (2)$$

The log-transformed durations were determined from a 1195-nonsense word *corpus*, containing all BP phonemes (from the São Paulo State dialect). Average and standard-deviation per phone have confirmed (see table 1 in [3]) current knowledge on duration in BP which is in agreement with universal trends.

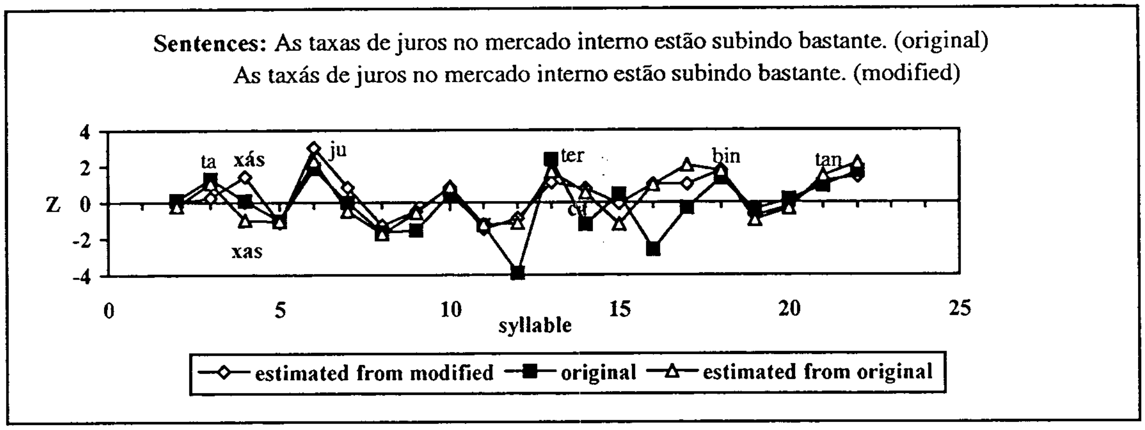


Figure 1: Comparison between the rhythmic patterns for the syllables in the sentence "As taxas de juros no mercado interno estão subindo bastante." (original) and "As taxás de juros no mercado interno estão subindo bastante." (modified). Notice that the accentual peak (learned by the network) on "ta-" (2nd position) migrates to the 3rd position (corresponding to "xás") and that the remaining z-scores for estimated from original and from modified sentences are close to each other.

Another *corpus* was used in order to study rhythmic patterns that emerge from BP sentences. Syllable and IPCG z-scores were computed for 100 sentences read by the same speaker. This *corpus* was manually segmented and carefully labeled by the author. Sentence length varies between one and 84 syllables. Syntactic boundaries were also marked using a set of eight hierarchical labels (see [1] for details).

3. THE RHYTHMIC STRUCTURE GENERATOR

In all 100 rhythmic patterns (represented as durational contours of syllable and IPCG z-scores), syllable z-scores indicate the (lexically) stressed syllables of the utterance: the highest z-score within each word coincides with the lexically stressed syllable. On the other hand, if the highest IPCG z-scores within non-clitic words *at lexical stressed position* are taken as a criterion for boundary placement and as a measure of boundary strength, coherent prosodic groups are obtained for all sentences in the *corpus*. IPCG z-scores delimitate accentual groups (prosodic words) where rhythmic patterns are characterized by frequent alternation of z-score values at the beginning of the accentual group followed by a duration *crescendo* (starting at least on the penultimate syllable) towards the last stressed syllable in the group.

Statistical analyses carried on adjacent segmental z-scores (from formula 1) have confirmed that segment durations are strongly correlated within syllable (at lexical stress) and within IPCG (at phrasal accent) frames.

The observed regularities were used for training a multilayered perceptron. In the network input, a phonological, prosodic description of each sentence is used to infer the network output, the IPCG and syllable z-score evolution over the sentence.

Thanks to a greater coherence between accentual typology and z-score patterning, the network learning was, in fact, faster than that of the sequential network used for French (where durations expressed in units of a clock were used instead of z-scores). But original RPU durations are no longer preserved in the BP case. What is preserved is the rhythmic structure as represented by the z-score patterning.

Our model of segmental duration generation was applied to learning and also to test *corpus* subsets. The model was capable to generalize even when lexical stress position was manipulated, as can be observed in figure 1, where the pseudo-word "taxás" has an estimated duration contour coherent with the oxyton pattern.

The example above shows clearly that the network was capable of associating the crucial linguistic information concerning lexical stress to a specific durational patterning.

In the model's next stage, segmental z-scores are computed simply by taking the z-scores previously obtained at the network output and using them in formula 1 for each segment in the sentence. It is important to say that, in BP, two kinds of z-scores must be carefully manipulated. By default, each IPCG z-score is used for computing the durations of the segments that constitute this rhythmic unit. But at lexical stress not enhanced phrasally, the syllable z-score must be used to compute the durations of the segments in the syllable frame. By doing this, the consonants in the syllable onset that are supposed to be part of the previous IPCG would have their durations computed again. To correct this, nucleus and coda segment durations of the previous IPCG are computed by using an averaged IPCG z-score (mean between previous and next IPCG z-scores) and onset segment durations at lexical stress are computed once by using the syllable z-score. Onset segments in the next syllable are not taken into account with this kind of computation. Their duration can be obtained by using the same z-score of the IPCG sharing the same vowel with the corresponding syllable (a kind of reset value).

With this algorithm, the error means between original and estimated duration were -1 ms for the learning *subcorpus* and 2 ms for the test *subcorpus* (both are not statistically different from 0). The standard-deviations are 32 ms, for the learning *subcorpus* and 36 ms, for the test one.

4. EVALUATION OF THE MODEL

The results presented here show clearly that it is possible to obtain a segmental duration generator integrating two important characteristics for a speech synthesis system: automation and correct reproduction of the BP natural rhythm. The capacity of the model to generalize to new sentences was also shown.

A perception test was also performed in order to evaluate the model of segmental duration generation. An ABBA test allowed us to evaluate 10 utterances whose segmental durations were modified by analysis-resynthesis with the Hybrid Model ([4]). Segmental durations were assigned to utterances presented in pairs according to two models: our rhythmic model of segmental duration generation (utterances of type *model*) and a model with the same error distribution as our generator but having durations assigned according to a Gaussian number generator (utterances of type *random).* What is being evaluated when these two models are compared is the tendency for our rhythmic generator to preserve homogeneous lengthening of syllables at lexical stress and of IPCGs at phrasal accent. This tendency is not taken into account by the random model.

The ten pairs of utterances were presented to fifteen listeners. Each pair consists of a model utterance and a random utterance ramdomly ordered. Utterance pairs are also randomly organized in a sequence for each listener. During the session, each pair is heard twice by the listener via headphones (in this case, in the same order). After listening, the subject must decide which utterance seemed less unnatural by writing down on a specific sheet.

The results shows a preference of about 67% (significantly different from chance) for the utterance modified by our segmental duration generator. All subjects said that the utterances sound quite artificial. (This aspect is inherent to the Hybrid Model, which is still being improved). Three analysis-resynthesis-generated utterances whose durations were obtained by our model can be heard here [sound A0467S01.WAV, A0467S02.WAV, A0467S03.WAV].

This weak - but stable - preference for our model can be explained by the type of test prepared. Both models have a 27-ms standard-deviation for segmental durations. If the perception thresholds for durations (30 ms for vowels and 40 ms for consonants) proposed by some authors like Goedemans & van Heuven ([5]) are taken as an approximation, an important amount of the utterances' duration errors for each utterance would be very close or under the threshold. If this assumption is true, it is very hard for the listeners to perceive any difference between the two versions of the original utterance. A certain amount of *don't care* responses reinforces this hypothesis.

5. AUTOMATIC PAUSE INSERTION

A new *corpus,* pronounced by another speaker (40-year-old State of Pernambuco dialect), was recorded to study BP rhythm and, particularly, the problem of silent pause insertion. This *corpus* was also manually labelled and segmented and presents subsets of sentences and nonsense words embedded in carrier sentences uttered at three speech rates (self-chosen normal rate, and metronome-controlled slow and fast rates).

Final lengthening and pause phenomena are challenging in BP. In French, pausing is said to be part of final lengthening. The same can be said for oxytons in BP. But in paroxytons and proparoxytons, prepausal stressed syllables (phrasal accent position, in our terminology) are lengthened and are (often) followed by non-lengthened post-accented ones.The first analyses of the durational contours for this speaker confirmed our previous finding of two macrorhythmic units in BP and seem to indicate that pre-stressed and post-stressed syllable-size units function as a reference clock. At phrasal accent, post-accented syllables are likely to be a kind of filler of a quantized beat interval ([6]). In this position, the IPCG z-score can be computed as follows.

The duration of the entire unit, from current to next vowel onset (including eventual silence) is taken as the starting point. From this duration we extract *n* times the duration of a previously computed reference clock period, where *n* is the number of post-accented IPCGs in the unit. The z-score is computed using (2) with the segments of the accented IPCG.

At the generation stage, if the estimated z-score for that position is greater than a previously assumed value, an amount of "sound" z-score will be computed, as it was done for French. This "sound" z-score is used to compute segment durations for the accented IPCG. Post-accented IPCGs will have their segment durations computed in two steps: z-score computation using a given reference clock period as the left member in (2) and segment duration computation using (1). The silent interval is the amount of duration necessary to complete the next integer number of reference clock periods for the entire unit.

The analyses of the durational contours also show that if duration contours are compared across the three speech rates, some differences between the phonology and the

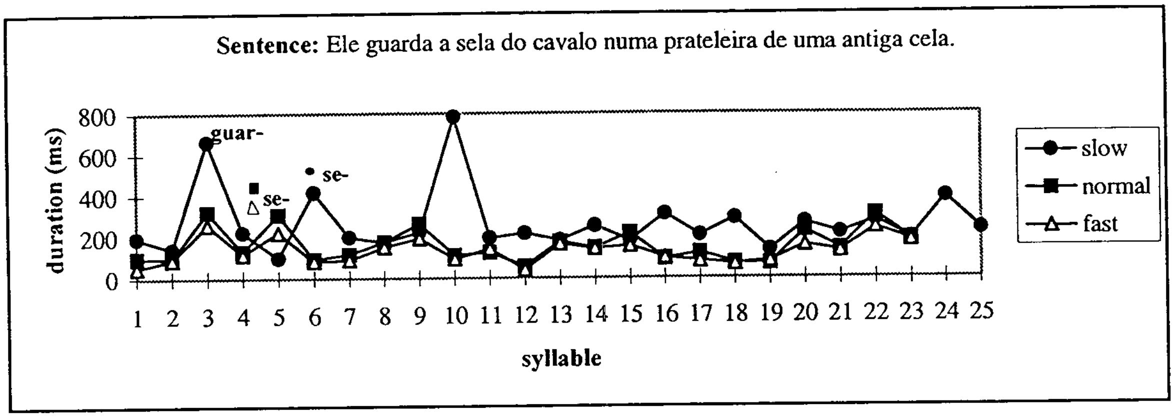


Figure 2: Durational contours for the sentence "Ele guarda a sela do cavalo numa prateleira de uma antiga cela." ("He keeps the horse saddle on a shelf in an ancient cell."). Vertical axis stands for syllable duration in milliseconds and horizontal one, the position of the syllable in the sentence. Please note the evolution of the first four z-scores in the sentence (corresponding to "Ele guarda"). Contour lines are showed only for the sake of visibility.

phonetics of rhythm can be observed that contitute an argument in favor of the initial assumption.

6. CONCLUSION

Clearly stated, this assumption says that rhythm is performed with a continuous phonetic variation subject to physical laws as inertia but can still be understood as the pragmatic-conditioned subjective interpretation of a single metrical representation.

Suppose that the following metrical grid for the first four positions in the sentence in figure 2 is assumed (where *guar-* is phrasally prominent):

```
      x
  x   x
  x   x
  x x  x  x
  Ele guarda
```

This grid fits with the slow rate phonetic pattern. As speech rate increases, the need to accelerate the accent realization on *guar-* forces the durational contour to be a *crescendo* towards the strongest position in the grid. This behaviour resembles the entrainment model proposed by Port and colleagues ([7]) if we take the demands of an accentual prominence clock as stronger in a hierarchy than the syllable one (this one would be synchronized to the vowel onset succession). Syllable durations are shown here because they are traditional units, but the pattern shown above is even more *crescendo*-like for IPCG durations and z-scores.

Our rhythm model is part of a project for building a concatenative text-to-speech synthesis system for BP that aims at incorporating well-grounded knowledge of linguistic structure during all research stages (from unit inventory choice [8] to acoustic signal generation). We advocate that high-quality concatenative synthesis is possible once an accurate understanding of the phonetics-to-semantics aspects of a language is assumed.

The careful understanding of phonetic and phonological aspects of rhythm as tentatively presented here is a crucial step in this direction.

ACKNOWLEDGEMENT

This work has been supported by grants (n^{os}. 95/9708-6 and 96/7832-4) from the *State of São Paulo Foundation for the Promotion of Science* (FAPESP).

REFERENCES

[1]Barbosa, P.A. & Bailly, G. (1997) *Generating pauses within the z-score model.* In: Progress in Speech Synthesis. van Santen, J.P.H.,Sproat, R.W., Olive, J.P. & Hirschberg, J. (Eds.), New York: Springer-Verlag. 365-381.

[2]Campbell, N.W. (1992) *Syllable-based segmental duration.* In: Talking Machines: theories, models, and designs (Bailly, G. & Benoît, C. Eds.), 211-224.

[3]Barbosa, P.A. (1996) *At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration.* Proceedings of the 1st ESCA TRW on Speech Production Modeling, Autrans, France, 85-88.

[4]Böeffard, O. & Violaro, F. (1994) *Using a hybrid model in a Text-to-Speech system to enlarge prosodic modifications.* Proceedings of the ICSLP'94, Yokohama, Japan, 727-730.

[5]Fant, G. & Kruckenberg, A. (1996) *On the quantal nature of speech timing.* Proceedings of the ICSLP'96, pp. 2044-2047.

[6]Goedemans, R. & van Heuven, V.J. (1995) *Duration perception in subsyllabic constituents.* Proceedings of the EUROSPEECH'95, Madrid, Spain, 1315-1318.

[7]Port, R., Cummins, F. & Gasser, M. (1996) *A dynamic approach to rhythm in language: Toward a temporal phonology.* In B. Luka and B. Need (eds) CLS-31: Proceedings of the Chicago Linguistics Society, pp. 375-397.

[8]Albano, E.C & Aquino, P.A. (1997) *Linguistic criteria for building and recording units for concatenative speech synthesis in Brazilian Portuguese.* In these Proceedings.

30

Generation of Pauses Within the *z*-score Model

Plínio Almeida Barbosa
Gérard Bailly

ABSTRACT We have previously proposed [BB94] a model for the generation of segmental durations that proceeds in two steps: (1) prediction of the timing of a salient acoustic event per syllable according to phonotactic and syntactic information, and (2) application of a repartition model that determines the duration of each individual segment between these events. This chapter focusses on the repartition model and describes how the initial model has been enriched to account for the emergence of pauses as speech rate is decreased. It describes a perceptual evaluation of the whole model. This evaluation shows that, for the same distribution of prediction errors, a precise timing of these events is perceptually more relevant than a segment-based method aiming at predicting precisely each individual segmental duration.

30.1 Introduction

Many generation models for segmental duration are based on statistical analysis of ad hoc corpora. In these models, a large set of coefficients accounts for different factors influencing the duration of segments. Several types of factors are used: intrinsic factors such as mode and place of articulation of the phoneme, contextual factors such as identities of the surrounding segments, phonotactic constraints such as the position of the segment within the syllable, phonological factors such as the word prominence, or linguistic factors such as the word class [Kla82, BS87, OSh81, van94]. These models can be referred to as phoneme-based models of prediction. In the most recent models, parameters are tuned automatically on large corpora in order to get the least squares minimum of segmental prediction errors. All segments have a priori the same weight in the optimization problem.

Other models use intermediate higher-level units, such as the syllable [MG93, Cam92] or the foot [Wit77, KOH86]. These models focus on macroscopic rhythms and isochronies. The choice of a *rhythmic unit* is expected to reveal prosodic regularities or units [GJM94]. Compared to phoneme-based models, these models are expected to be more robust to variation in speech rate [MG93] and segmental components [Cam92]. These models rely on the fact that rhythm can be defined as a perceptual structure in which units are grouped according to recurrent patterns. The tentative definition of rhythm proposed by Woodrow [Woo51, p. 1232]:

By rhythm, in the psychological sense, is meant the perception of a series of stimuli as a series of groups. The successive groups are ordinarily of similar pattern and experienced as repetitive. Each group is perceived as a whole and therefore has a length lying within the psychological present

is confirmed by Fraisse [Fra74, p. 75]:

Sans perception globale, il n'y a pas de structure rythmique[1]

Regularity of these patterns is thus considered as the prime object of the perception of rhythm. The exact durations of the segmental constituents of each unit are thus less important than the functions assumed by the rhythmical patterns of the groups. The nature of these units and the language-specific constraints acting on the regularity of the patterns have been largely discussed (see the debate on isochrony/isosyllabism in [Lea74, Leh77, Dau83, Noo91, Wh94]).

Our generation model is based on an original rhythmic unit called the interperceptual center group (IPCG). This IPCG is delimited by *salient acoustic events*. That is, the *perceptual centers* (see section 30.2) and the succession of IPCGs evidence rhythmical groups that are characterized for French as a gradual lengthening of IPCGs. The perceptual experiment described in section 30.5 demonstrates that listeners are quite sensitive to the distribution of "prediction errors" along the utterance: The relative timing of salient acoustic events such as the vocalic onset is more relevant for the perception of momentary tempo than the distribution of segmental durations between these events.

Relative timing of these events is conditioned by both linguistic and phonotactic factors and is sensitive to variation in speech rate. Our model is based on the assumption that momentary tempo is encoded by the speaker and decoded by the listener according to a reference clock.[2] Such a hypothesis is clearly stated by other authors [All75, Fra74]. This is also corroborated by data on speech acquisition [Kon91, Smi78], which show that babbling departs from an initial isochrony near 16 months of age to incorporate language-specific rythmic structure. Systematic deviations of IPCG durations from this reference clock can be thus interpreted as rhythmical patterns assuming various linguistic or paralinguistic functions. We use this paradigm to control speech rate: The duration of each IPCG is expressed in terms of clock units.[3] Then the repartition model divides the duration of each IPCG among its segmental constituents. We show that this repartition model can easily be in charge with pause emergence: durations of pauses are thus intimately tied with rhythmical patterns because pauses as other segments contribute to "fill" each IPCG when the stretch of segments is too large. This is a partial answer to Fant's rhythmical coherence of pauses and tempo observed for English, French, and Swedish [Fan91, p. 248]:

the average inter-stress interval within a short time memory span of about 4 seconds preceding a pause . . . synchronises an internal beat generating clock which sets a preferred pause duration.

This tendency to restart phonation at a time in relation with prior tempo—eventually with intervening "silent" beats—has been evidenced also by Couper-Kuhlen [Cou91] for turn-taking.

It is important to notice that in all duration models, pause is generally considered separately. In Klatt's model, for example, pause durations are given a priori and not integrated in the same mechanism responsible for generating segmental durations [Kla82, p. 761]

Rule 1: Insert a brief pause before each sentence-internal main clause and at other boundaries delimited by an orthographic comma.

The models proposed by Bartkova and Sorin for French and by O'Shaughnessy for Canadian French modify also an inherent (or intrinsic) segment duration as a function of the phonemic environment. Their models also take a priori durations for pauses that are placed according to phonotactic and syntactic criteria. In this kind of generation, corpus-dependent results may be easily obtained due to exhaustive computation of specific context effects.

The choice of a higher-level unit such as the syllable or the foot avoids the tuning of a large set of coefficients (about 80 in Bartkova-Sorin's model!). It also should enable an easier integration of duration in the prosodic structure via the characterization of the durational contours: these contours can be predicted in parallel with fundamental frequency contours by automatic learning procedures recently developed using dynamical predictors [Tra92] or stored prototypes [Aub92, AS92]. Nevertheless, none of the models using rhythmic units integrates the pause phenomenon as a by-product of *performance*: In Campbell's model, pause is not considered for generation. In Kohler's model ([Koh86], chapter 37 this volume), pause is also attributed at the beginning of the application of rules. Monnin and Grosjean [MG93] predict the duration of the last vowel of each word and the optional following pause from an index of linguistic complexity. An average correlation of 0.86 between the normal and slow rates is found. Although the concept of performance structures was introduced using very slow rates [GG83] in order to obtain a pause following each word, the performance structure is now built using the normal rate. No quantitative model is proposed to describe how the duration of each final group is shared by the vowel and the optional pause.

We next describe a model that integrates the automatic placing and calculation of pauses (silent—if present—and lengthened segments) as a continuous function of speech rate. Our model is guided by rhythmic constraints in a coherent and homogeneous framework.

[1] Without global perception there is no rhythmical structure.

[2] Existence of such internal clocks has been clearly hypothesized by neurophysiologists to coordinate biological movements [TSR90, Lli89]. Coordination of movements is controlled by two cooperative structures: a timekeeper, which delivers periodic signals, and a motor function, which uses these landmarks to deliver muscular activations.

[3] The actual value of the clock is computed for each utterance using the average duration of unstressed IPCGs.

30.2 Rhythm and the Perceptual Center

30.2.1 Arguments for a Reference Point per Syllable

Strong arguments for the characterization of momentary tempo by only one event per syllable come mainly from synchronization of speech with other acoustic stimuli or gestural activities. Research developed by Marcus and his colleagues [MMF76, Mar76] on perceptual centers (PCs) show that listeners perceive regularity in syllable sequences despite important differences among absolute durations [MMF76, p. 405]:

> *we were forced to ask ourselves what it was that was regular in a rhythmic list. To simplify our discussions we defined this as the P-centre of each item.*

The PC does not seem to correspond to any simple articulatory or acoustic correlate, but there is a clear interaction between speaker production and perception systems in order to produce speech sequences that the listener perceives as isochronous [Al175, Leh77]. Subjects must be capable of relating to some points of reference located in the speech flow in order to perceive such sequences as isochronous. This point of reference (or beat) was coined by Marcus [Mar75] as the perceptual center of the word (for monosyllabic ones). Similar experiments involving alternation or synchronization of speech and tapping or musical performance [Fra74, GA92] shows the consistency of this paradigm.

30.2.2 The Cyclic Attractor

These experiments do not concern connected speech and force the subjects to focus on a precise synchronization. However, more ecological experiments involving speech and tapping [Al175, BAL91] report converging results. Allen shows that subjects place their down-beats near the onsets of stressed syllables. For French, Berthier et al. show that there is a temporal relationship between the percussion of the knife on the whistle and the vocal onset that follows (from 0 to 100 ms) but never with the end of the vowels preceding it. One of the interesting questions to answer here is *what* is synchronized with *what*. By hypothesis, the performance of such tasks requires an internal time-measuring device or clock that provides target time points at which the tapping movements or syllabic gestures must produce their effect. The concept of an internal clock put forward by neurophysiologists [HM87, Sha82, SSV92] is central to biological movements. The natural frequency of the opening/closing of the vocal tract driven by the jaw (equal to 6 Hz ± 1 [SGE80]) could be easily adjusted to the frequency of regular beats of other gestural activies such as the leg/foot or arm/hand movements, which can reach 4-6 Hz. However, in most experiments the beat is not strictly synchronous with the metronome (clicks, taps), but rather precedes it. What could be the explanation of this? A partial answer is given by Fraisse [Fra80]: The subjects were asked to perform the beat with their hands or their feet. What is basically observed is that in the foot condition the lead of the beat over the click was larger than in the hand condition. The difference is roughly equivalent to the difference in nerve conduction time due to the foot/brain versus the hand/brain distance. This suggests that synchrony is maintained at the level of the two perceived events and that the motor commands can be issued well ahead to secure synchrony of the two percepts [Pri92]. We have thus to keep in mind that momentary tempo is a perceptual construct [Al175, Leh77].

30.2.3 Acoustic Correlate of P-Centers

Perception experiments carried out by various authors [Mar81, Pom89, Sco93] confirm that despite the nature of experimental conditions (perception or production experiments) and the nature of phonemes, the PC seems to be located at the vicinity of vocalic onset. Although their experiments do not concern connected speech, we have set the location of the PC at the vocalic onset when the syllable is not preceded by a silence. If there is a silence, the PC is usually placed earlier in the syllable:[4] in this case, only we have taken the beginning of the left-most voiced consonant of the syllable onset as the PC, that is, at the major increase in low-frequency energy within the syllable.[5]

Two consecutive PCs define the boundaries of the inter-perceptual-center interval (IPCI). The segments it contains form the IPCG. In the next sections, two prediction models for duration that use a rhythmic programming unit are confronted: Campbell's model, which uses the syllable, and our model, which uses the IPCG.

30.2.4 Importance of P-Centers as Sensory Regulators

Our experimental and theoretical work is strongly based on the fact that speech is produced by biological movements that have to be coordinated with other activities as important as breathing. Section 30.2.2 showed that such a sensorimotor synchronization can be explained by the coupling between internal and such "external" clocks. Such claim is not beyond the scope of speech synthesis: Speech is multimodal, and multimedia applications will need to synchronize speech with other movements (artificial or synthetic). Our two-stage model of control of segmental durations, which first computes the timing of one event per syllable, satisfies these requirements. We next demonstrate that this approach offers an efficient way to control speech rate and the emergence of pauses that are not constrained to appear in ad hoc and predefined positions.

[4] Scott has implemented a model for PC location based on the first derivative of the intensity function of the acoustic signal. According to her model, significant increases of intensity in the band frequency from 195 to 1638 Hz strongly determine the PC location.

[5] The exact PC location could be easily associated with each synthesis unit in concatenative synthesis systems.

TABLE 30.1. Parameters of the z-score model for two speakers, EV and FB, computed over logatoms. Means ± standard deviations of log-transformed segmental durations are given.

phon	EV	FB	phon	EV	FB	phon	EV	FB
a	4.58 ± .30	4.74 ± .34						
ϵ	4.43 ± .32	4.68 ± .27	œ	4.61 ± .34	4.68 ± .34	ɔ	4.62 ± .38	4.69 ± .30
e	4.49 ± .33	4.73 ± .33	Φ	4.65 ± .31	4.75 ± .27	o	4.67 ± .32	4.81 ± .27
i	4.49 ± .38	4.55 ± .33	y	4.59 ± .40	4.61 ± .29	u	4.64 ± .40	4.69 ± .31
˜ɑ	4.80 ± .30	4.89 ± .29	˜ɔ	4.81 ± .29	4.89 ± .31	˜ɛ	4.75 ± .28	4.93 ± .31
w	4.36 ± .21	4.51 ± .25	j	4.28 ± .23	4.47 ± .29	ɥ	4.33 ± .26	4.48 ± .23
l	4.16 ± .19	4.48 ± .26	∇	4.25 ± .19	4.48 ± .32	ə	4.16 ± .28	4.86 ± .43
p	4.60 ± .19	4.58 ± .53	t	4.51 ± .26	4.54 ± .55	k	4.60 ± .23	4.56 ± .43
b	4.34 ± .19	4.48 ± .24	d	4.28 ± .21	4.48 ± .23	g	4.32 ± .18	4.42 ± .24
v	4.34 ± .24	4.59 ± .28	z	4.39 ± .23	4.71 ± .20	ʒ	4.43 ± .23	4.62 ± .21
f	4.73 ± .25	5.04 ± .38	s	4.87 ± .24	5.26 ± .37	ʃ	4.81 ± .23	4.99 ± .25
m	4.46 ± .24	4.61 ± .23	n	4.35 ± .26	4.59 ± .24	ŋ	4.75 ± .24	4.68 ± .23

30.3 The Campbell Model

The elasticity principle [Cam92] in its strongest version says that all segmental durations in a syllable frame are obtained by a same and single factor z—the so-called z-score or normalized duration [WSOP92]—as follows:

$$Dur_i = \exp(\mu_i + z\sigma_i), \qquad (30.1)$$

where μ_i and σ_i are the mean and standard deviation of the log-transformed durations (in ms) of the realizations of the phoneme i. These z-scores are computed over the syllable by:

$$\sum_i Dur_i = syllable\ duration. \qquad (30.2)$$

The successive use of equations (30.1) and (30.2) is referenced as the repartition algorithm. Campbell's model proceeds in two steps:

- prediction of the syllable duration from phonotactic and phonological information by a statistical model (a multilayer perceptron)
- use of the repartition algorithm [Cam92]

As input to the perceptron, Campbell describes the syllable by six factors: (1) number of phonemes; (2) nature of the vowel (reduced, lax, and tense vowels or syllabic consonants or diphthongs); (3) position in intonational group; (4) type of foot (used to describe the rhythmic context of the syllable); (5) stress; and (6) word class to which the syllable belongs (lexical or functional words). In the output, the syllable duration is expressed in natural logarithmic form.

We have already demonstrated the optimal consistency of the rhythmic behavior of the segmental units inside an IPCG [BB92]: the z-scores of the syllable onset are positively correlated with the preceding rime whereas they are negatively correlated with the following nucleus. If the z-scores of all IPCGs of a sentence are represented graphically, temporal organization of these units exhibits monotonous ascending movements toward each phrase accent with a reset just after accent realization. For French, these patterns delimit rhythmic units characterized by a gradual lengthening toward the accented syllable [Pas92]. Such elementary macrorhythmic "contours" can be easily captured by dynamical nonlinear predictors.

30.4 The Barbosa-Bailly Model

30.4.1 Prediction of IPCG Durations

As in Campbell's model, our duration predictor proceeds in two stages. Nevertheless, significant differences have to be noticed:

- A sequential network [Jor89] constrained by an internal clock (a measure of the speaking rate for each utterance computed as the mean of the IPCIs in the utterance that do not correspond to a prosodic marker [BB94]) generates the timing of PC locations in terms of internal clock units (see figure 30.1).
- The IPCIs are then distributed among the IPCG constituents according to the repartition model, which presently includes the emergence of pauses.

In the network input, prosodic and phonological information that seems to be relevant for the prediction of the duration of the current IPCI is sequentially de-

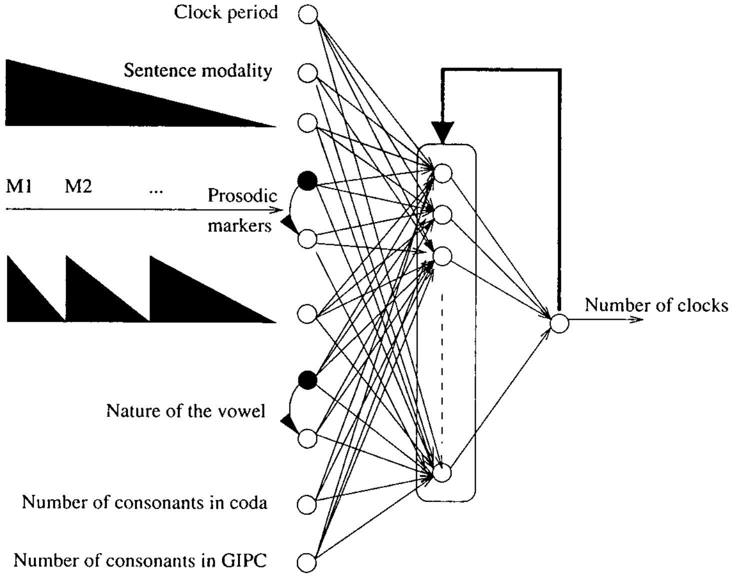


FIGURE 30.1. The sequential network for generating timing of PC locations in terms of internal clock units. The network transforms simple ramps indicating the length and the function of each linguistic unit of the utterance into rhythmic contours according to speech rate and phonotactic constraints. The input cells shown with filled circles store the values of connected cells with delay −1, i.e., following prosodic marker and next nucleus.

scribed. Two cells remain constant: (1) the frequency of the internal clock and (2) the sentence modality. Two other cells depict possible basic rhythmic patterns. Ramps indicate the extent of two basic prosodic units: (3) the sentence and (4) the prosodic group.[6] These linear ramps are set to a value equal to the number of IPCGs of the unit and reach zero value at the end of the unit. The six remaining input cells are fed with: (5) current prosodic marker, (6) next prosodic marker, (7) nature of current vowel, (8) nature of next vowel, (9) number of consonants in the IPCG, and (10) the coda (to be able to correct the PC estimate in case of pause emergence).

The parameters of the repartition model for French are given in table 30.1 for two speakers of our text-to-speech system. These parameters have been obtained by statistical analysis of the segmental durations of logatoms produced in isolation (speaker FB) or with a carrier sentence (speaker EV). Only data of the di- and trisyllabic logatoms are used. Although FB is globally faster than EV, well-known characteristics of intrinsic phonemic durations are captured: nasal vowels longer than oral ones, voiced consonants shorter than unvoiced ones, and so forth.

The repartition algorithm was modified to include pause emergence. For this, an original procedure for recording versions of the same sentence at different speech rates was tested.

30.4.2 *The Corpus*

A corpus with 20 sentences uttered at five different speaking rates was recorded to study the influence of pause emergence on overall rhythmic structure. In order to simplify the problem of locating PCs, all sentences of the corpus are CV sequences (all pauses were thus preceded by a vowel). Effective separated rates were obtained by controlling the speaker's speaking rate with synthetic questions. The speaker was instructed to answer these questions with sentences presented on a screen at the same speaking rate as that of the questions. The speaking rate of the synthetic questions was controlled by multiplying the coefficients μ_i and σ_i of each phoneme by a scaling factor, as suggested by Wightman and colleagues [WSOP92]. This procedure has been particularly effective in differentiating five speaking rates from very slow to very fast in a rather continuous way. The average values of the five speech rates are given in table 30. 3.

[6] The prosodic group is defined on a linguistic basis: It consists of each content word and any depending function words. The prosodic groups are linked by prosodic markers [Bai89]. These markers are indexed by the degree of cohesion between the adjacent prosodic groups.

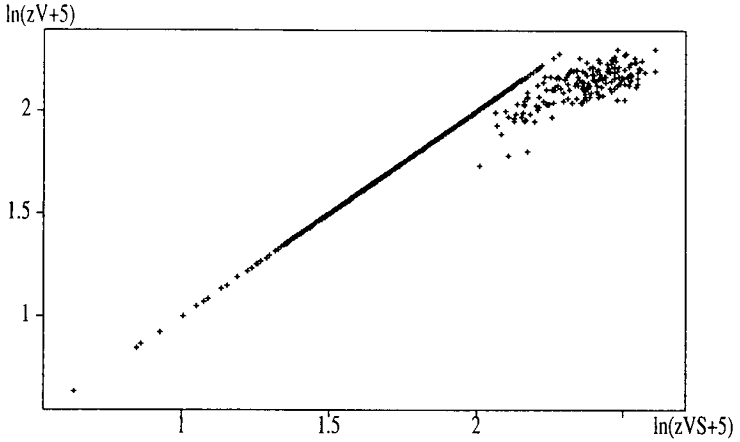


FIGURE 30.2. Scatterplot of log-transformed actual z-scores z_V of each vowel of the corpus versus log-transformed virtual z-scores z_{VS} for all speech rates.

30.4.3 *Incorporating Pause Phenomenon to the Repartition Algorithm*

In order to compute the silence duration that must be assigned to each IPCI in the generation stage, we have studied the relation between the actual z-scores of vowels $z_V = (\ln(Dur_V) - \mu_V)/\sigma_V$ and virtual z-scores of the same vowels. These virtual z-scores are computed by adding to the actual duration of the vowel Dur_V the duration of the optional adjacent silence as $z_{VS} = (\ln(Dur_{\text{vowel}} + Dur_{\text{silence}}) - \mu_V)/\sigma_V$.

As presented in figure 30.2, $z_V = z_{VS}$ up to the critical point $z_{VS} \simeq 2.4$. Then, between $z_{VS} = 2.4$ and $z_{VS} \simeq 4.5$, the speaker has two strategies:

(1) The speaker carries on lengthening the vowel with no pause, that is, a subjective pause is produced; this corresponds to the end of the segment $z_V = z_{VS}$ already mentioned.

(2) The speaker introduces a silent pause that corresponds to the scatterplot, which becomes more and more dense as z_{VS} increases.

Finally, for large z_{VS}, only the strategy (2) is retained, and a pause is always inserted.

Figure 30.3 presents a zoom on the scatterplot presented in figure 30.2. Figure 30.3 shows that despite the emergence of a pause, the speaker still lengthens the vocalic part of the IPCG. Our model of pause emergence uses the regression line (see equation (30.3) below) for $z_V \neq z_{VS}$ as the computational model of pause duration.

$$(z_V + 5) = (z_{VS} + 5)^{0.59} \exp(0.72) \qquad (30.3)$$

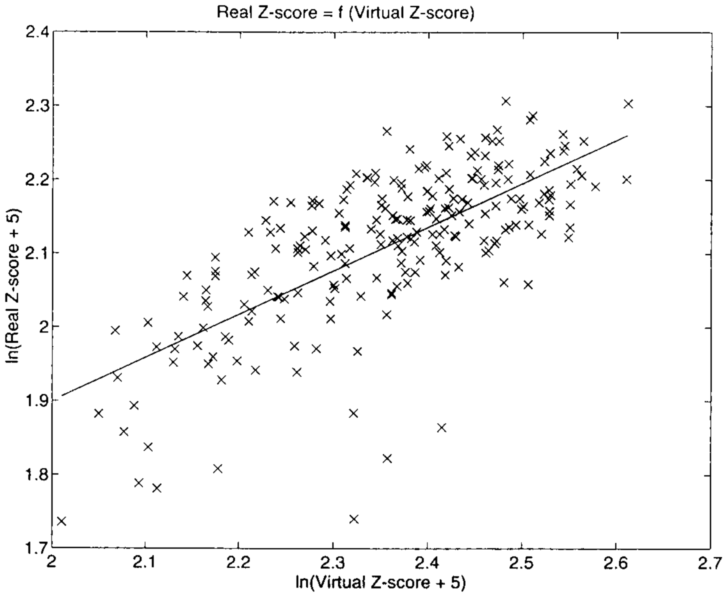


FIGURE 30.3. Scatterplot of log-transformed actual z-scores versus log-transformed virtual z-scores for $z_{VS} \neq z_V$.

A pause can thus theoretically "emerge" when this regression line crosses the straight line $z_V = z_{VS}$. This critical point is thus $z_{VSc} = 0.79$. No silence per se is generated between z_{VSc} and the minimal $z_{VS} \simeq 2.4$ because a minimum silence duration is required for relaxing speech articulators. A minimum silence duration was set for each speech rate as a result of analysis carried out on the corpus (see table 30.2). Surprisingly, this minimum duration does not vary very much as a function of speech rate and remains around $Dur_{\text{silence}} \simeq 60$ ms.

The modified repartition algorithm proceeds as follows:

- Computation of the z-score for a given IPCG (z_g).
- If z_g is smaller or equal to the critical z_{VSc}, the procedure is over: segmental durations are obtained by using equation (30.1).
- If z_g is greater than z_{VSc}, the z-score of the vowel z_V is obtained by regression equation (30.1), by setting $z_{VS} = z_g$.
- The segmental durations are computed with the repartition algorithm with $z_g = z_V$ and added up. The difference between this result and the original IPCG duration gives the duration of the silence.
- If the silence duration is greater than the minimum the procedure is over.
- If not, no silence is inserted and the z-score of the IPCG is kept equal to z_g.

TABLE 30.2. Averages ± standard deviations of absolute prediction errors for three sound classes using the modified repartition algorithm on original data. The distribution of errors is compared for three different rhythmic units: the syllable (SY), the inter-vocalic onset group (VO), and the IPCG (PC).

rate	*Vowels*			*Consonants*			*Silences*		
	SY	VO	PC	SY	VO	PC	SY	VO	PC
v. slow	34 ± 32	35 ± 32	34 ± 29	43 ± 46	42 ± 43	37 ± 46	67 ± 58	76 ± 62	55 ± 53
slow	38 ± 40	42 ± 41	36 ± 32	32 ± 32	34 ± 31	28 ± 26	73 ± 52	73 ± 52	73 ± 56
normal	26 ± 26	30 ± 38	29 ± 23	27 ± 26	28 ± 26	25 ± 22	54 ± 37	64 ± 36	56 ± 35
fast	17 ± 19	18 ± 19	19 ± 20	21 ± 25	21 ± 27	19 ± 21	75 ± 77	68 ± 99	49 ± 78
v. fast	12 ± 12	11 ± 13	12 ± 14	13 ± 13	12 ± 17	11 ± 12	59 ± 33	93 ± 28	44 ± 20

TABLE 30.3. Number of errors obtained by the repartition algorithm in placing the silences. The number of actual silences by speech rate in the natural utterances is also given (there are 285 IPCGs by rate in the corpus).

rate	clock (ms)	minimum pause (ms)	location errors	actual silences
very slow	360	75	15	99
slow	325	67	13	64
normal	270	57	11	32
fast	210	51	—	26
very fast	165	59	1	5

It is important to note that no constraint on location—such as accentable position—is imposed on the silences; they are placed as a result of the modified repartition algorithm described above (see comments on the last audio example, see section 30.6). We applied this modified repartition algorithm to the original data for the five speech rates; the distribution of errors is given in table 30.2. This table shows that the IPCG predicts the durations of consonants and pauses with more accuracy than the syllable or the inter-vocalic onset group.

Using original IPCG durations, only a few silences were placed in positions not actually chosen by the speaker (see table 30.3); but all positions were assigned to a latent location for accent realization (a prosodic marker).

30.4.4 *Automatic Learning*

As described above, a sequential network was trained in order to generate the successive IPCIs for each sentence of the corpus. Fifty sentences (ten sentences in five rates) from the corpus (learning set) were chosen for the learning phase. The network generalizes quite successfully the corresponding IPCI sequences for all sentences in the corpus (compare standard deviations of errors between learning and test sentences in table 30.4). Synthetic segmental and silent durations were obtained by applying the modified repartition algorithm described above.

The errors between the original segmental durations and the ones obtained by our model were computed for each speech rate and for two types of unit: segments

TABLE 30.4. Means (and standard deviations) in milliseconds from the histograms of errors between the segmented and generated durations for learning-set and test-set sentences.

	learning set		*test set*	
rate	silences	segments	silences	segments
very slow	−41(137)	−1(49)	3(207)	4(56)
slow	70(116)	−2(40)	−82(147)	6(46)
normal	67(122)	0(37)	−105(113)	5(43)
fast	−189(84)	3(30)	64(144)	0(28)
very fast	−4(170)	2(25)	−49(193)	−1(23)

and silences (see table 30.4). The histograms of these errors for all learning-set sentences were strongly correlated with the normal distribution (minimum 95%).

We conducted an experiment to test the perceptual relevance of IPCG's organization. The stimuli were obtained by a PSOLA [CM90, BBW92] analysis-resynthesis technique, in which only the segmental durations were modified.[7] Two versions of each utterance were compared: the first one (tagged as *model*) is paced by our model and the second (tagged as *random*) is obtained by adding a Gaussian noise to the original segmental durations. The noise distribution is equal to the distribution of prediction errors produced by our model. Three distinct distributions are used for vowels, consonants, and silences (see table 30.2). The two versions differ only in the time structure of errors: The *model* version is expected to predict with more accuracy the sequence of PCs.

30.5 Perception Test

Fifteen subjects studying at ICP participated in this perception experiment. Each session lasted between 7 and 10 minutes. In an ABBA test the subjects were asked to select the utterance with the most adequate prosody. Listening may be repeated once. A question mark could be used if both utterances seemed similar to the subject. All subjects considered, 89% of *model* utterances were chosen as being the most natural. In 15% of the answers there was doubt between the utterances. Individual scores can be seen in figure 30.4.

Comments are quite instructive to evaluate what perceptual cues were used by listeners. Some of them are listed below:

- "When the utterances were continuous without interruption they seemed natural to me" (HL).

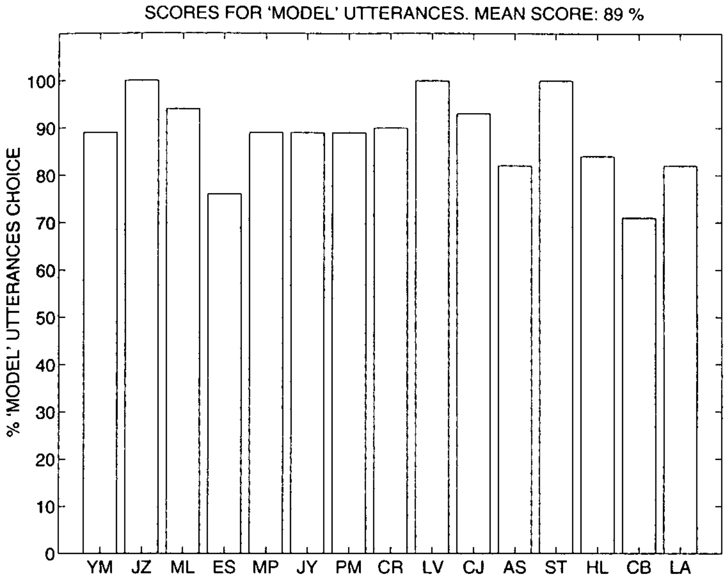


FIGURE 30.4. Individual scores for the choice of *model* utterances. Please note that intersubject results are similar.

- "My decisions were guided by word sequencing: if the sequence was interrupted or rhythm was not continuous, these utterance did not seem natural to me." (ML).
- "Utterances with a rhythmic change at the middle of a word did not seem natural to me!" (LV).

Subjects' preference for one utterance in the pair involves judgments of global aspects of rhythm (see last two comments above). In both stimuli the nature of the segments does not seem unnatural to them. Results are very similar among subjects. These results thus indicate the reliability of rhythmic judgments.

Results show a clear preference for the utterances obtained with our generation procedure. They support a model that maintains the basic rhythmic structure of natural speech by using a new higher-level programming unit: the IPCG.

30.6 Conclusions

We have presented a complete model for the rhythmic control of synthetic utterances. This model is based on the assumption that the rhythmic structure of speech is perceived by extracting salient acoustic events from the speech signal. This extraction is thus parallel to the acoustic-to-phonetic decoding of utterances and can explain why prosody and segments can be processed separately [CN88, PMK93].

[7] The insertion/deletion of short-term signals is ruled by emergence functions computed by temporal decomposition [BMA89]. Fundamental frequency values are recoded with only three values by segment: beginning, middle, and final, and energy is recorded by an average value.

As the synthetic rhythm is controlled by anchoring each syllable to its perceptual center—syllables expand or retract according to the repartition algorithm to fill in the IPCGs—the rhythmic trajectory produced can be further synchronized with other parametric trajectories such as melody or loudness to characterize in a unified way the prosodic structure of the synthetic utterances. We are currently working on the prediction of melody using a global approach [MAB95].

Acknowledgements

The first author was supported by CNPq/Brazil when he was working at ICP. Now he works as a post-doctoral student at the Instituto de Estudos da Linguagem, Unicamp, Campinas, Brazil. We thank Frédéric Bimbot for his voice and our listeners who offered their ears to the perception task.

References

[All75] G. Allen. Speech rhythm: its relation to performance universals and articulatory timing. *J. Phonetics* 3:75–86, 1975.

[AS92] M. Abe and H. Sato. Two-stage F_0 control model using syllable based F_0 units. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 53–56, 1992.

[Aub92] V. Aubergé. Developing a structured lexicon for synthesis of prosody. In *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoit, eds. Elsevier B.V., North-Holland, Amsterdam, 307–321, 1992.

[Bai89] G. Bailly, Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Comm.* 8:137–146, 1989.

[BAL91] V. Berthier, C. Abry, and T. Lallouache. Coordination du geste et de la parole dans la production d'un instrument traditionnel. In *Proceedings, Twelfth XIIe International Congress of Phonetic Sciences*, vol. 4, Aix-en-Provence, France, 34–37, 1991.

[BB92] P. Barbosa and G. Bailly. Generating segmental duration by p-centres. In *Fourth Rhythm Workshop: Rhythm Perception and Production*, C. Auxiette, C. Drake, and C. Gérard, eds. Ville de Bourges, Bourges - France, 163–168, 1992.

[BB94] P. Barbosa and G. Bailly. Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Comm.* 15:127–137, 1994.

[BBW92] G. Bailly, T. Barbe, and H. Wang. Automatic labelling of large prosodic databases: tools, methodology and links with a text-to-speech system. In *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoit, eds. Elsevier B.V., North-Holland, Amsterdam, 323–333, 1992.

[BMA89] G. Bailly, P. F. Marteau, and C. Abry. A new algorithm for temporal decomposition of speech. application to a numerical model of coarticulation. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, 508–511, 1989.

[BS87] K. Bartkova and C. Sorin. A model of segmental duration for speech synthesis in French. *Speech Comm.* 6:245–260, 1987.

[Cam92] W. Campbell. *Multi-level Timing in Speech.* Ph.D. thesis, University of Sussex, Sussex, UK, 1992.

[CM90] F. Charpentier and E. Moulines. Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Comm.* 9(5-6):453–467, 1990.

[CN88] A. Cutler and D. Norris. The role of strong syllables in segmentation for lexical access. *J. Experimental Psychology: Human Perception and Performance* 14:113–121, 1988.

[Cou91] E. Couper-Kuhlen. A rhythm-based metric for turn-taking. In *Proceedings, Twelfth International Congress of Phonetic Sciences*, vol. 1, Aix-en-Provence, France, 275–278, 1991.

[Dau83] R. M. Dauer. Stress-timing and syllable-timing re-analyzed. *J. Phonetics* 11:51–62, 1983.

[Fan91] G. Fant. Units of temporal organization. Stress groups versus syllables and words. In *Proceedings, Twelfth International Congress of Phonetic Sciences*, vol. 1, Aix-en-Provence, 247–250, France, 1991.

[Fra74] P. Fraisse. *La psychologie du rythme.* Presses Universitaires de France, Paris, 1974.

[Fra80] P. Fraisse. Des synchronisations sensori-motrices aux rythmes. In *Anticipation et Comportement*, J. Requin, ed. Editions du CNRS, Paris, 233–257, 1980.

[GA94] C. Gérard and C. Auxiette. The processing of musical prosody by musical and nonmusical children. *Music Perception* 9:471–503, 1992.

[GG83] J. P. Gee and F. Grosjean. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology* 15:418–458, 1983.

[GJM94] L. A. Gerken, P. W. Jusczyk, and D. R. Mandel. When prosody fails to cue syntactic structure: 9-months-olds' sensitivity to phonological versus syntactic phrases. *Cognition*, 20:237–265, 1994.

[HM87] D. Hary and G. P. Moore. Synchronizing human movement with an external clock source. *Biological Cybernetics* 56:305–311, 1987.

[Jor89] M. I. Jordan. Serial order: A parallel, distributed processing approach. In *Advances in Connectionist Theory: Speech*, J. L. Elman and D. E. Rumelhart, eds. Lawrence Erlbaum, Hillsdale, NJ, 1989.

[Kla82] D. H. Klatt. The KLATTalk text-to-speech conversion system. In *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, France, 1589–1592, 1982.

[Koh86] K. J. Kohler. Invariability and variability in speech timing: from utterance to segment in German. In *Invariance and Variability in Speech Processes*, J. Perkell and D. H. Klatt, eds. Lawrence Erlbaum, Hillsdale, NJ, 268–298, 1986.

[Kon91] G. Konopczynski. Acquisition de la proéminence dans le langage émergent. In *Proceedings, International Congress of Phonetic Sciences*, vol. 1, Aix-en-Provence, France, 333–337, 1991.

[Lea74] W. A. Lea. *Prosodic Aids to Speech Recognition: IV. A General Strategy for Prosodically-guided Speech Understanding.* Univac Report PX10791, Sperry Univac, DSD, St. Paul, MN, 1974.

[Leh77] I. Lehiste. Isochrony reconsidered. *J. Phonetics* 5:253–263, 1977.

[Lli89] R. R. Llinás. The role of the intrinsic electrophysiological properties of central neurons in oscillation and resonance. In *Cell to Cell Signalling: From Experiments to Theoretical Models*, A. Goldbeter, ed. Academic Press, New York, 3–16, 1989.

[MAB95] Y. Morlec, V. Aubergé, and G. Bailly. Evaluation of automatic generation of prosody with a superposition model. *International Congress of Phonetic Sciences*, Stockholm, Sweden, 1995.

[Mar75] S. M. Marcus. Perceptual centres. Unpublished fellowship dissertation, King's College, Cambridge, UK, 1975.

[Mar76] S. M. Marcus. *Perceptual centres.* PhD thesis, Cambridge University, Cambridge, 1976.

[Mar81] S. M. Marcus. Acoustic determinants of Perceptual center (p-center) location. *Perception and Psychophysics* 30(3):247–256, 1981.

[MG93] P. Monnin and F. Grosjean. Les structures de performance en français: Caractérisation et prédiction. *L'Année Psychologique* 93:9–30, 1993.

[MMF76] J. Morton, S. Marcus, and C. Frankish. Perceptual centers (p-centers). *Psychological Revue* 83(5):405–408, 1976.

[Noo91] S. G. Nooteboom. Some observations on the temporal organisation and rhythm of speech. In *Proceedings, International Congress of Phonetic Sciences*, vol. 1, Aix-en-Provence, France, 228–237, 1991.

[OSh81] D. O'Shaughnessy. A study of French vowel and consonant durations. *J. Phonetics* 9:385–406, 1981.

[Pas92] V. Pasdeloup. Durée inter-syllabique dans le groupe accentuel en français. In *XIXe Journées d'Etudes sur la Parole*, 531–536, 1992.

[PMK93] V. Pasdeloup, J. Morais, and R. Kolinsky. Are stress and phonemic string processed separately? Evidence from speech illusions. In *Proceedings of the European Conference on Speech Communication and Technology*, vol. 2, Berlin, 775–778, 1993.

[Pom89] B. Pompino-Marschall. On the psychoacoustic nature of the p-center phenomenon. *J. Phonetics* 17:175–192, 1989.

[Pri92] W. Prinz. Distal focussing in action control. In *Fourth Rhythm Workshop: Rhythm Perception and Production*, C. Auxiette, C. Drake, and C. Gérard, eds. Bourges, 65–71, 1992.

[Sco93] S. Scott. *Perceptual Centres in Speech—An Acoustic Analysis.* Ph.D. thesis, University College, London, 1993.

[SGE80] V. N. Sorokin, T. Gay, and W. Ewan. Some biomechanical correlates of the jaw movements. *J. Acoust. Soc. Amer.* 68:S32, 1980.

[Sha82] L. H. Shaffer. Rhythm and timing in skill. *Psychological Review* 89:109–122, 1982.

[Smi78] B. L. Smith. Temporal aspects of English speech production: A developmental perspective. *J. Phonetics* 6:37–67, 1978.

[SSV92] A. Semjen, H. H. Schulze, and D. Vorberg. Temporal control in the coordination between repetitive tapping and periodic external stimuli. In *Fourth Rhythm Workshop : Rhythm Perception and Production*, C. Auxiette, C. Drake, and C. Gérard, eds. Bourges, 73–78, 1992.

[Tra92] C. Traber. F_0 generation with a database of natural F_0 patterns and with a neural network. In *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoit, eds. Elsevier B.V., North-Holland, Amsterdam, 287–304, 1992.

[TSR90] M. Turvey, R. Schmidt, and L. Rosenblum. Clock and motor components in absolute coordination of rhythmic movements. *Haskins Laboratories Status Report on Speech Research*, New Haven, CT, 231–242, 1990.

[van94] J. P. H. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer, Speech and Language* 8:95–128, 1994.

[WH94] B. Williams and S. M. Hiller. The question of randomness in English foot timing: A control experiment. *J. Phonetics* 22:423–439, 1994.

[Wit77] I. H. Witten. A flexible scheme for assigning timing and pitch to synthetic speech. *Language and Speech* 20:240–260, 1977.

[Woo51] H. Woodrow. Time perception. In *Handbook of Experimental Psychology*, S. Stevens, ed. Wiley, New York, 1224–1236, 1951.

[WSOP92] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price. Segmental durations in the vicinity of prosodic boundaries. *J. Acoust. Soc. Amer.* 91(3):1707–1717, 1992.

Audio Demos

Original and synthetic signals are given in the CD-ROM to assess the quality of our generation. Three versions of three sentences of the corpus at different speech rates are given: (a) the original signal, (b) the *random* version, and (c) the *model* version. These last two synthetic versions are obtained by TD-PSOLA resynthesis applied to (a). The three sentences are:

1: Very fast: "Tu peins les sommets du Mont Karimo!"

2: Normal: "Tôt dans la matinée, Thibaud peint les sommets du Mont Karimo!"

3: Very slow: "Ta main, posée sous le manteau, visait les sommets du Mont Karimo, tôt dans la matinée!"

Then three purely synthetic versions produced by our current text-to-speech system driven by the original prosodic contours of the sentences are given.

Finally sentence 2 is synthesized using the original F_0 and energy contours of the *very slow* version and the durations computed with a clock at 450 ms. Six pauses have been inserted, including one after the unaccented word "dans."

Cad.Est.Ling., Campinas, (31) :33-53, Jul./Dez. 1996

AT LEAST TWO MACRORHYTHMIC UNITS ARE NECESSARY FOR MODELING BRAZILIAN PORTUGUESE DURATION: EMPHASIS ON AUTOMATIC SEGMENTAL DURATION GENERATION [*]

PLÍNIO ALMEIDA BARBOSA
(UNICAMP)

> *Et le vent, la vague, l'étoile, l'oiseau, l'horloge, vous répondront, il est l'heure de s'enivrer ;* pour ne pas être les esclaves martyrisés du temps, *enivrez-vous, enivrez-vous sans cesse de vin, de poésie, de vertu, à votre guise.*
> Les Petits poèmes en prose, *Charles Baudelaire*

ABSTRACT: By modeling Brazilian Portuguese acoustic duration, this work presents two arguments in favor of macrorhythmic units. First, the emergence of distinct durational patterns for lexical and phrasal accents. Second, the homogeneous lengthening (shortening) effect of segments correlating syllables at lexical stress and IPCGs at phrasal accent. A two-stage model of segmental duration generation is derived.

RÉSUMÉ: La caractérisation de la durée en portugais du Brésil (PB) permet de faire émerger une typologie accentuelle signalant la présence de deux unités macrorythmiques. Les maxima des *z-scores* de la syllabe coïncident avec la position de l'accent lexical tandis que les maxima des *z-scores* du GIPC démarquent les frontières prosodiques de l'énoncé. Un modèle à deux étapes permettant la génération simplifiée de la durée segmentale du PB en est dérivé.

RESUMO: O modelamento da duração acústica no português do Brasil (PB) tornou possível a emergência de uma tipologia acentual que revela a existência de ao menos duas unidades de programação macrorrítmicas: a sílaba e o GIPC. Os pontos de

[*] A shorter version of this paper is published on the *Proceedings of the 1st ESCA Tutorial and Research Workshop on Speech Production Modeling & 4th Speech Production Seminar*, Autrans, France, May 20 to 24th, 1996, under the title: *At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration.*

máximo dos *z-scores* da sílaba coincidem com a posição do acento lexical enquanto os pontos de máximo dos *z-scores* do GIPC (coincidindo com posição de acento lexical) demarcam as fronteiras prosódicas do enunciado. Um modelo rítmico possibilitando a geração automática e simplificada da duração segmental é proposto para ser integrado em um sistema de síntese da fala em PB.

INTRODUCTION

Recently, van Santen's work (1994) convincingly showed that it is not necessary to consider the existence of (macro)rhythmic programming units (RPU) in order to generate segmental duration. His phoneme-based approach involves, however, a huge computational cost. In early work, on other hand, Barbosa and Bailly (1997) have shown that the durational structure (expressed by a syllable size z-score model) associated with a set of read sentences reveals a certain kind of organization over the segment level. We were able to propose an approach for duration generation that takes into account the macrorhythmic organization of speech and drastically simplifies the mechanism of duration assignment.

As has already been demonstrated for French, normalizing the acoustic duration of consecutive segments points to an organization into higher order units whose boundaries are two consecutive vowel onsets (Barbosa & Bailly 1994). This unit was named inter-perceptual-center group (IPCG) by reference to research on p-centers (Marcus 1981; Pompino-Marschall 1991), whose findings suggest that their optimal location is the vicinity of the vowel onset, as can be observed by careful examination of Pompino-Marshall's figures and can be confirmed in a more ecological work (Janker 1995).

Normalized durations are obtained through Campbell's *z-score* model (1992). The z value of each segment *s* is computed by writing:

$$\text{Dur}_s = \exp\,(\mu_s + z.\sigma_s) \qquad (1)$$

where Dur_s is the segment duration and (μ_s, σ_s) stands for the average and the standard-deviation of the log-transformed durations of all *s* realizations in an *ad hoc corpus*. The strong elasticity hypothesis in Campbell's model says that all segments in a syllable frame have the same z-score: a single value of *z* per syllable can then be computed by writing:

$$\text{Dur (syllable)} = \Sigma_s \exp\,(\mu_s + z.\sigma_s) \qquad (2)$$

French data allow us to propose a weaker elasticity hypothesis where the rhythmic unit is the IPCG, not the syllable. Brazilian Portuguese (BP) data, on the other hand, indicate that at least two macrorhythmic units are necessary to model the durational structure of read sentences. BP rhythm can then shed light on the segmental/suprasegmental controversy thanks to the greater complexity of its accentual typology.

In BP, lexical stress can be assigned to the final, penultimate or antepenultimate syllable. Stressed syllables can be enhanced as they are uttered by carrying phrasal accent. Only lexically stressable syllables can bear phrasal accent.

The acoustical correlates of stress are often the greater duration of the stressed unit and the decrease of intensity in the post-stressed syllables (if any) (Massini 1991)[1]. Massini also says that stress is carried by the syllable as a whole – and not by the vowel alone– but our data reveal a slightly more complex situation.

The results shown below are based on the analysis of two *corpora*. A nonsense word *corpus*, meant to capture the speaker durational characteristics, and a read-sentence corpus, meant to analyze utterances' rhythmic structure.

SPEAKER'S DURATION STATISTICS

Segmental durations were determined from a 1195-nonsense word *corpus*, containing all BP phonemes (and also some of the BP allophones) of a native 30-year-old professional speaker (from the Paulista dialect). Statistical analyses were then performed. The results on Table 1 confirm current knowledge on duration in BP (which is in agreement with universal trends: Lehiste 1970): (a) for front and back vowels, the higher the vowel, the shorter its average duration; (b) post-stressed vowels (/ɐ, ɪ, ʊ/) are shorter than their stressed counterparts (/a, i, u/); (c) nasal vowels are longer than their oral counterparts; (d) voiceless consonants are longer than their voiced counterparts.

Table 1: Mean duration (and standard-deviation) of the BP phones (in ms) for our speaker[2].

i	145 (37)	ɐ	111 (45)	ĩ	209 (25)	tʃ	149 (20)	f	138 (14)	n	76 (15)
e	170 (36)	ɪ	98 (44)	õ	229 (26)	k	121 (21)	s	143 (26)	ɲ	103 (24)
ɛ	175 (32)	ʊ	77 (19)	ũ	215 (29)	b	86 (17)	ʃ	143 (16)	ɾ	47 (16)
a	165 (28)	j	92 (10)	j̃	136 (14)	d	71 (17)	v	78 (16)	r	81 (12)
u	134 (42)	w	97 (25)	w̃	139 (23)	dʒ	109 (18)	z	87 (21)	ʁ	62 (15)
o	168 (35)	ɐ̃	174 (46)	p	120 (20)	g	67 (16)	ʒ	89 (12)	l	73 (16)
ɔ	183 (29)	ẽ	210 (44)	t	113 (20)			rr	90 (12)	ʎ	77 (14)

The log-transformed versions of these data were used in formula (2) (where *syllable* is either phonological syllable or IPCG) to compute the z-scores of syllables and IPCGs of 100 sentences read by the same speaker. This corpus was manually segmented and carefully labeled by the author (a total of 2,055 syllables). Sentence length varies between one and 84 syllables. Syntactic boundaries were also marked using a set of

[1] Fundamental frequency is not an acoustic correlate of lexical stress but is a cue of phrasal prominence.

[2] The /r/ phoneme was realized as a trill ([r]) in syllable-final position and as a fricative (transcribed as the uvular [ʁ]) elsewhere (e.g. in *carro* or *rosa*).

eight hierarchical labels (obtained by the projection of a surface tree – from a dependence grammar where the root is the verb – over the paradigmatic axis. See Barbosa & Bailly 1994 for a more detailed description of these labels).

EXTRACTING DURATIONAL CONTOURS FROM Z-SCORE EVOLUTION

Segments were grouped into two kinds of RPU's: syllable and IPCG. By using the raw duration of each group in formula (2) above, the *z-scores* were computed for all RPU's in each sentence. An example is shown in Figure 1.

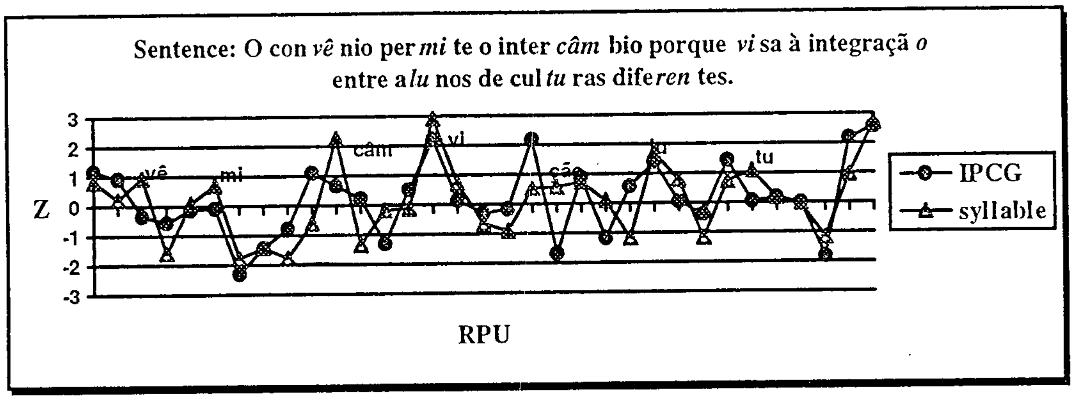


Figure 1: Z-scores (vertical axis) for IPCG's and syllables in the sentence "O convênio permite o intercâmbio porque visa à integração entre alunos de culturas diferentes." ("The agreement allows for interchange since it aims at the integration among students from different cultures."). Lexically stressed syllables are italicized. Tick marks on the horizontal axis represent the vowels along the sentence (signaling both syllables and IPCG's; NB: orthographic *-io* – in *convênio* and *intercâmbio* – is a semivowel/vowel cluster). Note that the greatest z-scores for the syllable correspond to lexically stressed vowels.

In all 100 rhythmic contours, syllable z-scores indicate the (lexically) stressed syllables of the utterance: the highest z-score within each word coincides with the lexically stressed syllable. On the other hand, if the highest IPCG z-scores within non-clitic words *at lexical stressed position* are taken as a criterion for boundary placement and as a measure of boundary strength, coherent prosodic groups are obtained for all sentences in the *corpus*. In Figure 1, the strongest boundary (IPCG z-score of *vi* in *visa*) splits the sentence into two chunks of 16 IPCG's each. The property of eurhythmicity is very clear and the coherence between this result and those of Grosjean's performance trees (1993) is notorious.

IPCG z-scores delimitate accentual groups (prosodic words) where rhythmic pattern is characterized by frequent alternation of z-score values at the beginning of the accentual group followed by a duration *crescendo* (starting at least on the penultimate syllable) towards the last stressed syllable in the group.

STATISTICAL CONFIRMATION OF THE RESULTS

Statistical analyses confirm that actual segmental z-scores (by using raw segmental duration into formula 1) are strongly correlated in RPU frames. For these analyses, phrasal boundaries were marked by choosing positions in the utterance corresponding to IPCG z-score maxima (coinciding with lexically stressed RPU) and by carefully hearing the utterances in order to confirm these candidates to prominence.

The results in table 2 show that, in phrasal accent position, *onset/nucleus* segment sequences (in the syllable) are negatively correlated (-31%) whereas *nucleus/coda* segment sequences (necessarily in the rhyme) are positively correlated (76%). On the other hand, *onset/ nucleus* and *nucleus /coda* segment sequences confirm that lexically stressed syllables not bearing phrasal accent form a homogeneous unit. In non prominent positions, VC and V#C sequences seem to suggest that the IPCG is a homogeneous unit whose rhyme component is enhanced at phrasal accent position.

Table 2: Correlation (in percentage) between consecutive segmental *z-scores* according to accentuation degree. Segments were categorized for each phonological syllable with three labels: *onset*, for each segment in onset position, *nucleus*, for the vowel nucleus, and *coda*, for each segment in coda position. Sequences as *nucleus-onset* span necessarily over syllable boundaries. The lexical stress category refers to lexical stressed RPU not bearing phrasal accent. Only the significant values are reproduced here.

	lexical stress	*phrasal accent*	*other positions*
onset/nucleus	63	-31	4
nucleus /onset	ns	26	56
nucleus /coda	48	76	63

These results corroborate Campbell's predictions (1993) only in part: since he adopts the syllable as the RPU, onset segments in syllables at prosodic boundaries are overpredicted. The observed final lengthening affects the rhyme, not the onset. This is produced by lesser articulator stiffness associated with closure movement (Edwards, Beckman, & Fletcher 1991).

NEURAL NETWORK ARCHITECTURE, TRAINING AND TEST

RPU z-scores can be a means of deriving the segmental durations of a particular sentence (if formula 1 is applied by setting $z = z_{syllabe}$ or $z = z_{IPCG}$). The above durational contours can be easily generated by neural networks, as implemented by the author for French (Barbosa & Bailly, *op.cit.*). In this work segmental durations were computed in a second stage by sequentially applying formula 2 and 1 (these two steps constitute the so-called *repartition algorithm*) with the IPCG duration delivered by the network output (this two-stage model is proposed in Campbell 1992).

For French, a sequential, recurrent network was used to learn to associate a phonological, prosodic description of the sentence to the respective rhythmic pattern expressed by the evolution of IPCG duration over the sentence. This method allows for preservation of macrorhythmic unit durations maintaining the rhythmicity of the original sentence. Although the vowel quality of the syllable nucleus is sequentially described at the network input, only the number of the intervening consonants between two vowels and not their nature are represented. This implicitly means that consonant nature is not important to derive IPCG duration (although microrhythmic differences in segmental duration can be captured in the second stage by the repartition algorithm).

For BP, a simpler network was used by implementing a multilayered perceptron (Rohani, Chen & Manry 1992). At the perceptron input, a phonological, prosodic description of each sentence is used to infer the network output, the IPCG and syllable z-score evolution over the sentence.

Thanks to a greater coherence between accentual typology and z-score patterning, the network learning was, in fact, faster. Furthermore, z-score patterns are smoother than that of RPU duration. But original RPU durations are no longer preserved. What is preserved in this framework is the rhythmic structure as represented by the z-score patterning. In a typical consonant contrast as *ti/bi*, for instance, different consonants (/t/ and /b/) induce different RPU durations, since the same RPU z-score delivered by the network allows to obtain different consonant durations in the repartition algorithm but the same vowel duration for /i/. In the early French version an identical duration for the pair *ti/bi* would have been imposed and different durations for /t/, /b/ and /i/ would have been obtained.

Our model of segmental duration generation was applied to the learning and also to the test *corpus* subsets. The model was capable to generalize even when lexical stress position was manipulated.

The 17 formal neurons used at the perceptron input stand for a phonological, syntactic description of each sentence. They are coded with real numbers between 0 and 1 and are briefly described below.

1. Internal clock. This is a measure of the utterance speech rate. This value is computed per sentence by averaging the IPCG durations not associated with a syntactic marker[3];

2. Declination line. (Decreasing) number of IPCGs between the current position and the final in the sentence. Normalization factor: 100;

3. Lexical stress (pre-stressed, stressed and post-stressed) 3 IPCGs before the current position;

4. Lexical stress (pre-stressed, stressed and post-stressed) 2 IPCGs before the current position;

5. Lexical stress (pre-stressed, stressed and post-stressed) immediately before the current position;

6. Lexical stress (pre-stressed, stressed and post-stressed) at the current position;

7. "Declination line" associated with the current phrase. (Decreasing) number of IPCGs between the current position and the next syntactic marker;

8. Syntactic marker that dominates the current phrase;

9. Syntactic marker that dominates the next phrase;

10. Vowel nature 3 GIPCS before the one at the current position;

11. Vowel nature 2 GIPCS before the one at the current position;

12. Vowel nature immediately before the one at the current position;

13. Vowel nature at the current position;

14. Number of consonants in the IPCG immediately before the current one;

15. Number of consonants in the current IPCG;

16. Number of consonants in the coda of the syllable immediately before the current one;

17. Number of consonants in the coda of the current syllable.

[3] The syntactic boundaries were marked manually (an automatic assignment can be obtained with a parser). A set of nine distinct markers were extracted from a dependence grammar analysis where the head is the verb (Tesnière 1965, Martin 1981). This set is a modified version of Bailly's markers (Barbosa & Bailly 1994). These markers are obtained by projecting the surface tree nodes over the syntagmatic axis. The strength between adjacent nodes is indicated by the dependence relation between them. The markers are: IF (the two adjacent nodes come from distinct trees which is common with strong ponctuation signs and coordinated clauses); TF (the two adjacent nodes are dominated by the verb); DF (the dominated node is to the right of the dominant verb. Example: between verb and complement); GF (the dominated node is to the left of the dominant verb. Example: between subject and adjacent verb); ID (the two adjacent nodes are not directly related but are in the same tree); DD (the dominated node is to the right of the dominant one, which is normally a noun. Example: between noun and postposed adjective); DG (the dominated node is to the left of the dominant one, which is normally a noun. Example: between anteposed adjective and noun); IT (the two adjacent nodes are dominated by a same node which is not the verb); FF (sentence ending). Here two examples: "O gatinho <GF> bebeu <DF> leite <TF> numa tigela <DD> verde <FF>." e "Ontem, <IF> o calmo <DG> gatinho <DD> preto <ID> bebeu <DF> leite <TF> numa tigela <DD> verde <IT> e rosa <FF>."

The number of (formal) neurons in the hidden layer (25) was estimated by Manry's software (for a global learning error of 0.5. This error is computed between the desired IPCG and syllable z-scores - previously calculated using formula 1 - and the IPCG and syllable z-scores obtained at the network output). The 2 nodes at the output stand for the IPCG and syllable z-score corresponding to a current position in the sentence (current syllable and IPCG sharing the same vowel).

During the learning phase, 39 sentences were included in the training (these sentences constitute the learning *subcorpus*. The remaining ones, the test *subcorpus*). They contain 663 RPUs presented iteratively to the network by examples. An example is a pair formed by the linguistic description of the current RPU (performed by the 17 nodes described above) at the input and the corresponding IPCG and syllable z-scores at the output.

The network computes at each iteration the error between the original and the estimated z-scores (for IPCGs and syllables). This error allows to modify the connections' weights in the direction of a better approximation between estimated and original outputs.

The degree of convergence for this process is satisfactory, as can be seen by the gradation of estimated rhythmic patterns modifications in the sequence of figures 3, 4 and 5 (for IPCGs only). The sentence in these examples is: "As taxas de juros no mercado interno estão subindo bastante."

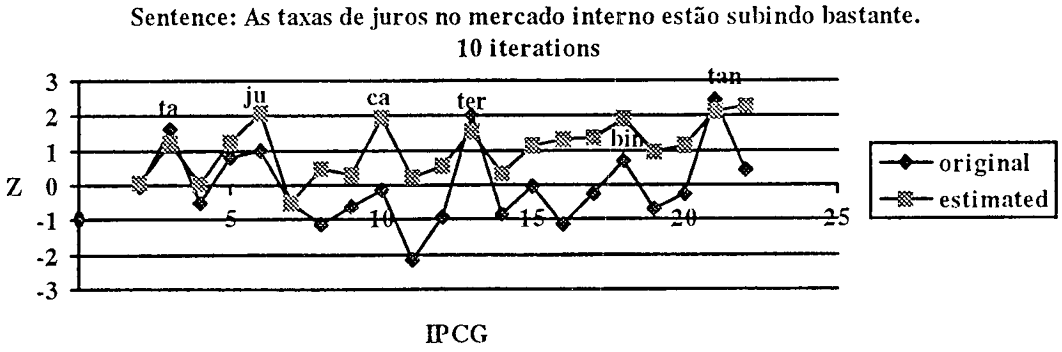


Figure 3: Comparison between original IPCG z-scores (presented to the network as the output element of this example) and IPCG z-scores estimated by the network after 10 iterations. Notice the distance between original and estimated rhythmic patterns. After 10 iterations the phrasal accent at "-ter" (from "interno") is already reproduced. Syllable (instead of IPCG) orthography is indicated at corresponding positions for ease of reading.

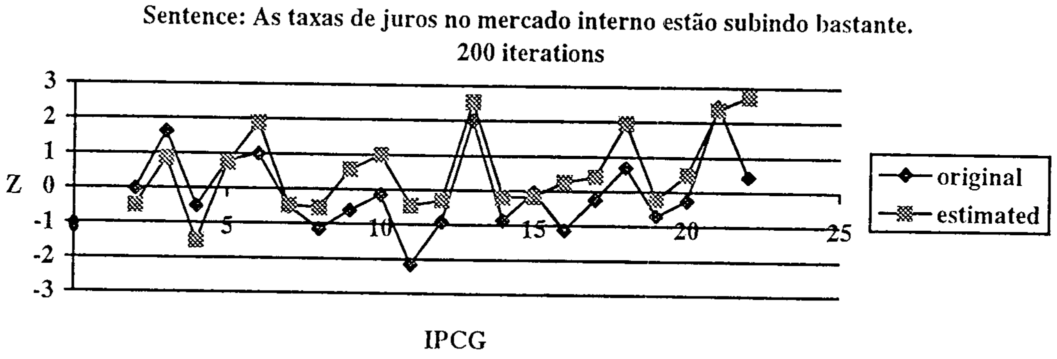


Figure 4: Comparison between original IPCG z-scores (presented to the network as the output element of this example) and IPCG z-scores estimated by the network after 200 iterations. Notice that the pattern corresponding to "interno" was improved compared to the one in the previous figure. In 200 iterations the sentence rhythmic pattern is well reproduced.

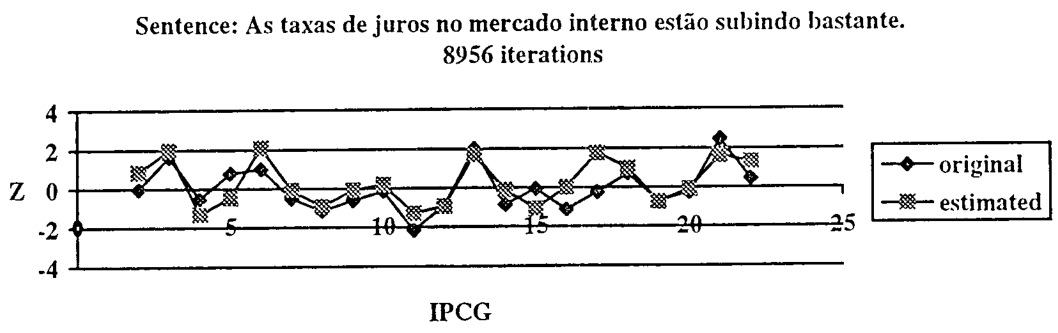


Figure 5: Comparison between original IPCG z-scores (presented to the network as the output element of this example) and IPCG z-scores estimated by the network after 8956 iterations. Although the pattern is not perfectly well reproduced, the network captured the saliences of "ju-" (from "juros", 5th position) and "subindo" (16th, 17th and 18th positions).

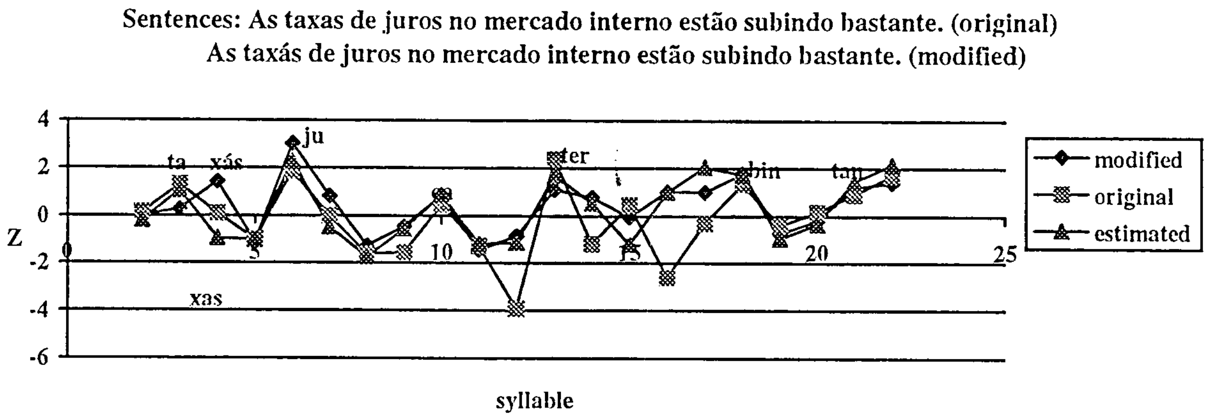


Figure 6: Comparison between the rhythmic patterns for the syllables in the sentence "As taxas de juros no mercado interno estão subindo bastante." (original) and "As taxás de juros no mercado interno estão subindo bastante." (modified). Notice that the accentual peak (learned by the network) on "ta-" (2th position) migrates to the 3th position (corresponding to "xás"). The next z-scores are maintained close to each other (compare remaining positions for estimated and modified z-scores).

The evolution of the syllable z-scores has the same degree of coherence and can be seen in figure 6 (compare original and estimated z-scores). As it is shown here, the network is capable of learning the sentence rhythmic patterns (actual *gestalten*) by estimating from the input what is more important to determine the desired output. It can be said that the neural net is a data-driven system because it learns from the empirical domain represented by the speech *corpora*.

Generalizing still more means to be able to capture the underlying mechanism allowing to associate input and output, that is, to learn to pacemake from a symbolic representation. In order to test this capacity, a modification from the original sentence "As taxas de juros no mercado interno estão subindo bastante." is introduced here. The lexical stress on "ta-" (from "taxas") is placed at the next syllable ("-xas"), creating the pseudo word "taxás". The results are shown below.

Figure 6 shows that the neural net is able to mimic a rhythmic pattern very well. Nevertheless, the non exact matching between original and estimated patterns introduces additional errors to the segmental duration generation model when the network works together with the repartition algorithm.

SEGMENTAL DURATION GENERATION

A program in C language was developed to integrate the two stages of duration prediction (perceptron and repartition algorithm).

All segmental durations in the learning and test *subcorpora* were predicted by the model. It is also important to note that the estimated z-scores can generate a segmental duration (by using formula 2) smaller than the minimum allowed by the phonatory system. In order to prevent this serious error a minimum duration associated with each segment is computed from the corpora and used as a threshold. No segmental duration is generated for a particular phone below that minimum.

The error means between original and estimated duration were -1 ms for the learning subcorpus and 2 ms for the test subcorpus (both are not statistically different from 0). The standard-deviations are 32 ms, for the learning *subcorpus* and 36 ms, for the test one. As was said before, the standard-deviations are greater than that obtained at the output of the repartition algorithm alone. The closeness between standard-deviations for the two *subcorpora* is a sign of a good performance in generalizing.

Segmental durations predicted by the model are presented below for two sentences (in table 3, a learning *subcorpus* utterance and, in table 4, a test *subcorpus* one).

Table 3: Natural and estimated segmental durations (in ms). Estimation performed by the Barbosa-Bailly model with the sentence "As taxas de juros no mercado interno estão subindo bastante.", one of the sentences in the learning *subcorpus*.

Segment	a	s	t	a	ʃ	ɐ	z	dʒ	ɪ	ʒ	u	ɾ
Original duration	75	104	69	137	84	70	47	41	50	85	118	25
Estimated duration	108	97	78	108	71	40	38	53	46	92	147	59

Segment	ʊ	z	n	ʊ	m	e	r	k	a	d	ʊ	
Original duration	58	44	28	39	44	64	45	83	99	39	28	
Estimated duration	46	51	47	38	49	88	47	82	107	31	33	

Segment	ĩ	t	ɛ	r	n	ʊ	ɪ	s	t	ɐ̃	w̃	s
Original duration	75	86	175	54	35	38	68	86	46	51	49	106
Estimated duration	105	57	133	51	63	46	53	81	79	124	93	114

Segment	u	b	ĩ	d	ʊ	b	a	s	t	ɐ̃	tʃ	ɪ
Original duration	75	86	175	54	35	38	68	86	46	51	49	106
Estimated duration	105	57	133	51	63	46	53	81	79	124	93	114

Table 4: Natural and estimated segmental durations (in ms). Estimation performed by the Barbosa-Bailly model with the sentence "As operações de crédito continuam.", one of the sentences in the learning *subcorpus*.

Segment	a	s	o	p	e	ɾ	a	s	õ	j̃	z	dʒ	ɪ		
Original duration	64	62	70	81	71	16	78	140	71	77	35	56	24		
Estimated duration	75	38	79	60	69	17	59	59	105	66	34	85	118		
Segment	k	r	ɛ	dʒ	ɪ	t	ʊ	k	o	tʃ	ɪ	n	u	ẽ	w̃
Original duration	126	52	175	70	44	78	38	78	99	85	24	44	163	100	34
Estimated duration	99	54	117	74	74	76	45	69	103	66	55	47	126	102	83

EVALUATION OF THE RHYTHM GENERATION MODEL

The results presented here show clearly how it is possible to obtain a segmental duration generator integrating two important characteristics for a speech synthesis system: automation and correct reproduction of the BP natural rhythm. The capacity of the model to generalize to new sentences was also shown.

The network learning can be continued. A typical computation lasts 30 minutes for about 1,000 iterations. Nevertheless, for this generator, the performance of the entire model will never be better than the performance of the repartition algorithm alone (standard-deviation of 24 ms). But the results are very satisfactory. We cannot forget that a experimented phonetician makes mistakes with about 10-ms standard-deviations when segmenting the speech signal (Leung & Zue, 1984).

A perception test was also performed in order to evaluate the model of segmental duration generation. An ABBA test allowed us to evaluate 10 utterances whose segmental durations were modified[4] by analysis-resynthesis with the Hybrid Model (Böeffard & Violaro 1994). Segmental durations were assigned to utterances presented in pairs according to two models: our rhythmic model of segmental duration generation (utterances of type *model*) and a model with the same error distribution[5] as our generator but having durations assigned according to a Gaussian number generator (utterances of type *random)*. What is being evaluated when these two models are compared is the tendency for our rhythmic generator to preserve homogeneous lengthening of syllables at lexical stress and of IPCGs at phrasal accent. This tendency is not taken into account by the random model.

The ten pairs of utterances were presented to fifteen listeners. Each pair has a model utterance and a random utterance ramdomly ordered. Utterance pairs are also randomly organized in a sequence for each listener. During the session, each pair is heard twice by the listener via headphones (in this case, in the same order). After listening, the subject must decide which utterance seemed less unnatural (the first or the second one) by writing down on a specific sheet. A *don't-care* option could be used if the two utterances seemed identical. Some listeners' comments were taken at the session end.

The results point to a preference of about 67% (significantly different from chance) for the utterance modified by our segmental duration generator. All subjects said that the utterances sound quite artificial. (This aspect is inherent to the Hybrid Model, which is in phase of improvement.)

This weak - but stable - preference for our model can be explained by the type of test prepared. Both models have a 27-ms standard-deviation for segmental durations. If the perception thresholds for durations (30 ms for vowels and 40 ms for consonants) proposed by some authors like Goedemans & van Heuven (1995) are taken as true, an important amount of the utterances' duration errors would be very close or under the

[4] Fundamental frequency and intensity were unchanged.

[5] Our generator presented a gaussian-like distribution of errors when comparing original and predicted durations. The two models have same means and standard-deviations.

threshold. If this assumption is true, it is very hard for the listeners to perceive any difference between the two versions of the original utterance. A great amount of *don't care* options reinforces this hypothesis.

CONCLUSIONS

The systematic findings correlating lexical stress in syllable frames and phrasal accent in IPCG frames constitute arguments in favor of the existence of rhythmic units above the segment level.

These distinct macrorhythmic units are also an evidence for a lexical and a post-lexical rhythm component, as suggested by Keating (1995), and point to the need for a rhythm tier where rhythmic nodes dominate different linguistic units in explicit models of speech production (as outlined by Articulatory Phonology in Browman & Goldstein 1990). (This is equivalent to saying that models of speech production should receive explicit linguistic input. And, from this point of view, this work insists on the need for cognitive theories of phonetics.)

The persistence of rhythmic homogeneity of lexically stressed syllables at the utterance level may be explained by a strong constraint on phasing of articulatory gestures (Browman & Goldstein 1986; 1990; 1992) that operates on lexical units (a microrhythmic adjustment in order to enhance the stressed syllable of lexical entries, as suggested by Perkell 1980) and broadly maintains the former phase relations at the utterance level.

Post-lexical metrical (macrorhythmic) readjustments as the Rhythm rule and microrhythmic readjustments as external sandhi rules (very common in BP) manipulating the gesture constellation would intervene during the elaboration of the connected-speech plan for each utterance.

Finally, teleological production (and perception)-guided readjustments may also intervene. Their function has a connection with the need for monitoring utterance production and for ensuring that articulatory movements are produced comfortably (low in articulatory cost). (A connection with the need for ensuring listeners comfortable decoding would also have an important communicative function.) Research insisting on the existence of a internal[6] clock enabling and monitoring rhythmic productions such as speech may favor the former assumption (Turvey, Schmidt & Rosenblum 1990; Pöppel 1989; Semjen, Schulze & Vorberg 1992; Leiner, Leiner & Dow 1991). Research as the one carried out by Edwards, Beckman & Fletcher, 1991, showing that increased duration at phrasal final position is realized by stiffness decreasing of VC units confirms the latter assumption and the IPCG homogeneity at the phrasal accent level.

The comfortable speech production would be realized by ballistic closure (jaw) movements contained in the VC frame (the jaw O-C movement is the articulatory correlate of macrorhythmic unit succession). In this sequencing, vowels are the most important segments as it is pointed out by B&G (1990, p. 352): "The X-ray data we have analyzed (...) have consistently supported the contention that consonant articulations are superimposed on continuous vowel articulations, which themselves minimally overlap." (In our work this is confirmed by the closeness of the vowel z-score contours with the IPCG ones. Syllables z-score contours are never correlates of phrasal accentuation.)

Under these assumptions, IPCGs and syllables have distinct status: the former are ease-of-production-and-monitoring units (or teleological units) at the phrasal level and the latter, gestural units (in the sense of the lexical orientation of Articulatory Phonology) at the word level. In the utterance framework, accentual groups are delimited by IPCG duration increasing followed by a durational contour reset (a typical alternation-*crescendo* contour. Whether this duration *crescendo* is related to a stiffness *descendo*, is a matter of investigation. But see Edwards, Beckman & Fletcher, 1991, note 2). Accentual group (AG) boundaries should also delimit regions where phenomena as sandhi rules or V-to-V coarticulation would be possible (but not across the AG boundaries).

[6] The word *internal* is used here in cognitive terms and means "in the brain".

ACKNOWLEDGMENTS

The author is currently supported by a grant (96/7832-4 and 95/9532-6) from the Fundação de Amparo à Pesquisa do Estado de São Paulo-FAPESP. The speaker and the listeners are strongly acknowledged. I am also in debt with Eleonora Albano e Sandra Madureira for their helpful suggestions in previous versions of this paper.

REFERENCES

BARBOSA, P.A. & Bailly, G. *Generating pauses within the z-score model.* In: *Progress in Speech Synthesis*. van Santen, J.P.H., Sproat, R.W., Olive, J.P. & Hirschberg, J. (Eds.), New York: Springer-Verlag, 1997:365-381.

BARBOSA, P.A. & Bailly, G. *Characterisation of rhythmic patterns for text-to-speech synthesis*, Speech Communication, 15 (1-2), 1994:127-137.

BÖEFFARD, O. & Violaro, F. *Using a hybrid model in a Text-to Speech system to enlarge prosodic modifications.* International Conference on Spoken Language Processing (ICSLP '94), Yokohama, Japan, 1994:727-730.

BROWMAN, C.P. & Goldstein, L. *Articulatory Phonology: an overview.* Phonetica, 49, 1992:155-180.

BROWMAN, C.P. & Goldstein, L. *Tiers in articulatory phonology with some implication for casual speech.* Kingston, J. & Beckman, M.E. (Eds.). Papers in Laboratory Phonology 1. Cambridge University Press, 1990:341-376.

BROWMAN, C.P. & Goldstein, L. *Towards an articulatory phonology.* Phonology Yearbook, 3, 1986:219-252.

CAMPBELL, N.W. Syllable-based segmental duration. In: Talking Machines: theories, models, and designs (Bailly, G. & Benoît, C. Eds.), 1992:211-224.

CAMPBELL, N.W. *Automatic detection of prosodic boundaries in speech*, Speech Communication, 13, 1993:343-354.

EDWARDS, J., Beckman, M.E. & Fletcher, J. *The articulatory kinematics of final lengthening*. J. Acoust. Soc. Am. 89 (1), 1991:369-382.

GOEDEMANS, R. & van Heuven, V.J. *Duration perception in subsyllabic constituents*. Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH), September, 18-21ª, Madrid, Spain, 2, 1995:1315-1318.

JANKER, P.M. *On the influence of the internal structure of a syllable on the p-center perception*. XIII International Congress of Phonetic Sciences, August 13-19, Stockolm, Sweden, 2, 1995:510-513.

KEATING, P.A. *Segmental Phonology and Non-segmental Phonetics*. XIII International Congiess of Phonetic Sciences, August 13-19, Stockolm, Sweden, 3, 1995:26-32.

LEHISTE, I. *Suprasegmentals*. Cambridge, Massachussets: MIT Press, 1970.

LEINER, H.C., Leiner, A.L. & Dow, R.-S. *The human cerebro-cerebellar system: its computing, cognitive, and language skills*. Behavioural Brain Research, 44, 1991:113-128.

LEUNG, H.C. & Zue, V. W. *A procedure for automatic alignment of phonetic transcriptions with continuous speech*. Proceedings of the IEEE ICASSP, 1, San Diego, 1984:2.7.1-2.7.4.

MARCUS, S.M. *Acoustic determinants of Perceptual-center (p-center) location*, Perception and Psychophysics, 30(3), 1981:247-256.

MARTIN, P. *L'Intonation est-elle une structure congruente à la syntaxe ?* In: *L'Intonation : de l'acoustique à la sémantique*, Paris: Klincksieck, 1981:234-271.

MASSINI, G. *A Duração no estudo do acento e do ritmo em português,* Master's thesis, Unicamp, 1991.

MONNIN & Grosjean, *Les Structures de performance en français : caractérisation et prédiction*, L'Année Psychologique 93, 1993:9-30.

PERKELL, J.S. *Phonetic Features and the Physiology of Speech Production*. In: *Language Production: Speech and Talk*. Butterworth, B. (Ed.). Academic Press: London. Vol 1, 1980: pp 33-372.

POMPINO-MARSCHALL, B. *The syllable as a prosodic unit and the so-called P-centre effect*. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, 29, 1991:65-123.

PÖPPEL, E. *The Measurement of Music and the Cerebral Clock: a new theory*. LEONARDO, 22(1), 1989:83-89.

ROHANI, K., Chen, M.S. & Manry, M.T. *Neural subnet design by direct polynomial mapping*. IEEE Transactions on Neural Nets, 3 (6), 1992:1024-1026.

SEMJEN, A., Schulze, H.-H. & Vorberg, D. *Temporal control in the coordination between repetitive tapping and periodic external stimuli*. Fourth Rhythm Workshop: Rhythm Perception and Production, Bourges, June, France, 1992:73-78.

SOUSA, E.M.G. *Towards an Acoustic Description of Brazilian Portuguese Nasal Vowels*. XIII International Congress of Phonetic Sciences, August 13-19, Stockolm, Sweden, 1995.

TESNIÈRE, L. *Éléments de syntaxe structurale*. Paris: Klincksieck, 1965.

TURVEY, M.T., Schmidt, R.C. & Rosenblum, L. *Clock and motor components in absolute coordination of rhythmic movements*. Status Report on Speech Research SR-101/102, Haskins Labs,1990.

VAN SANTEN, J.P.H. *Assignment of segmental duration in text-to-speech synthesis*. Computer, Speech and Language 8, 1994:95-128.

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.

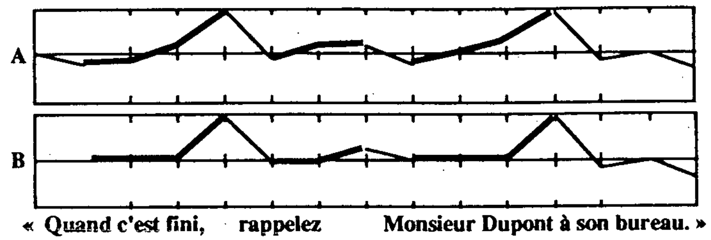


Figure 1. *Sentence 56 : Example of A and B configurations. Discrete values of k are connected by lines for sake of visibility. Prosodic groups are represented by thicker lines. The last IPCG of the utterance is not presented here.*

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.

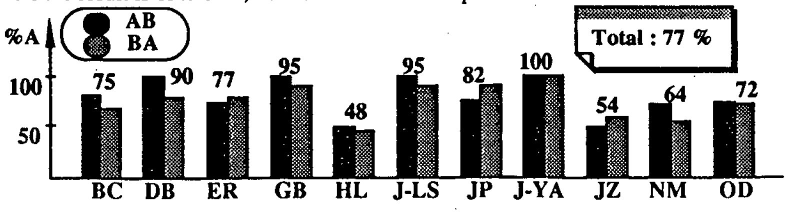


Figure 2. *Mean scores of A answers by subject and mean over all subjects*

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).

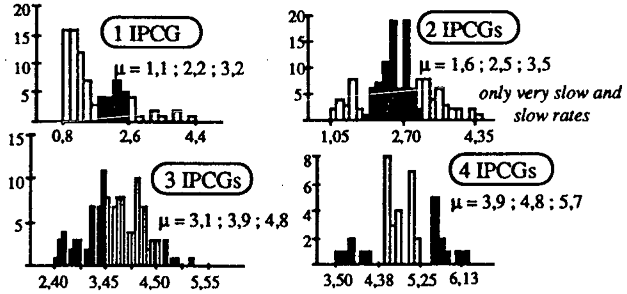


Figure 3. *Histograms and clusters by number of IPCGs in the PG*

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.