

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

Plínio Almeida Barbosa, 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 cotidiana e é uma propriedade fundamental da natureza viva (Fraisse, 1974). De fato, não se pode negar o universalismo 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 Linguística 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 frequência 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 frequência fundamental e a intensidade também concorrem para a percepção da duração¹ (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**).

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 syllables*). Para uma discussão detalhada sobre o isocronismo² 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 frequência 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,

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

²A 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).

convém introduzir um breve histórico da síntese da fala, nascida do antigo desejo do homem de reproduzir sua voz por máquinas falantes.

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. 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 ilimitado⁴ 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-

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).

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.

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.ru/ 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.

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.

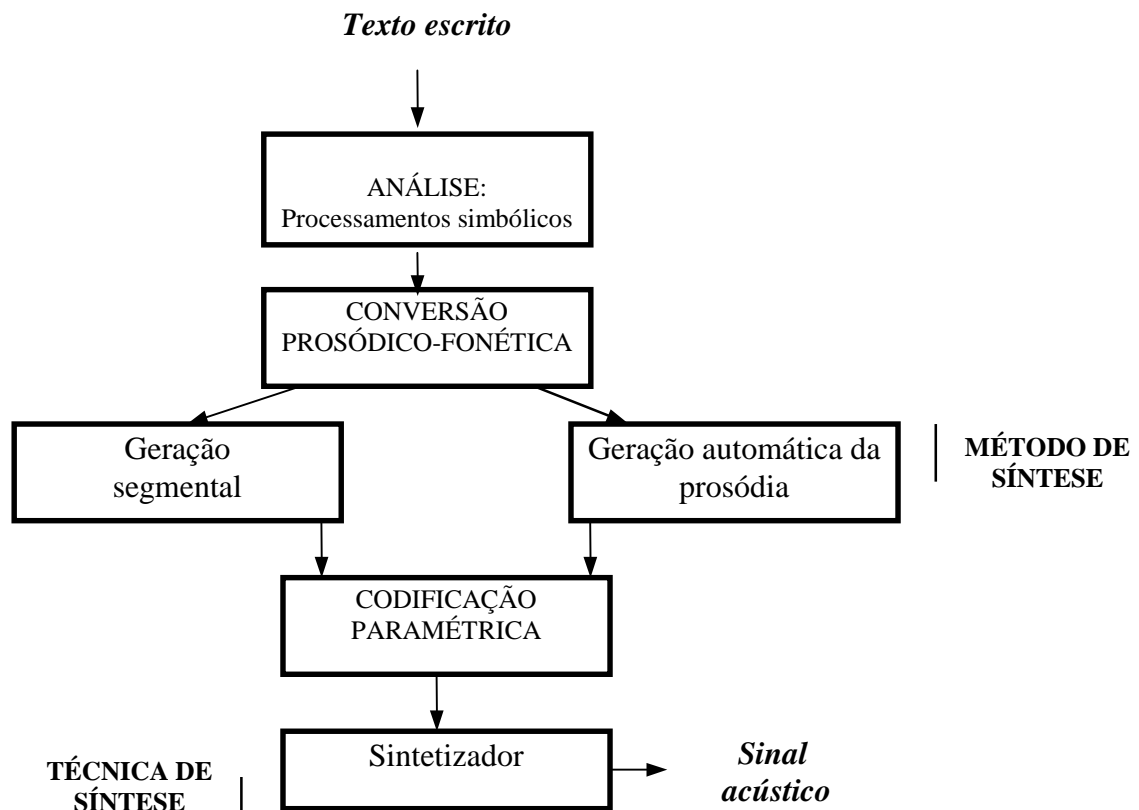


Figura 1: Esquema geral de um sistema de síntese da fala

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, frequência 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 frequência 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:

$$Dur = \frac{DurMin + (DurInt - DurMin) \cdot PRNCT}{100} \quad (1)$$

100

Onde *Dur* é a duração do segmento, *DurInt* é a duração intrínseca, *DurMin* é a duração mínima, calculada em função de *DurInt* (em geral $DurMin = 0,45 \cdot DurInt$) 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 porquê⁵, 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

⁵Uma 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

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 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:

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

Onde a duração de cada segmento i é obtida pelas parcelas $\exp(\mu_i + z \cdot \sigma_i)$. 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.

de uma unidade delimitada por dois onsets vocálicos consecutivos para o francês (Barbosa & Bailly, 1994) e,

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, no prelo) está sendo adaptado para o PB. Ele permitirá 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 seja 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

para o português brasileiro (apresentado aqui), a proposta de duas unidades de programação.

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 locutor⁶.

i	145 (37)	ĩ	209 (25)	f	138 (14)
e	170 (36)	õ	229 (26)	s	143 (26)
ε	175 (32)	ũ	215 (29)	ʃ	143 (16)
a	165 (28)	ĵ	136 (14)	v	78 (16)
u	134 (42)	ũ̃	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)	r	47 (16)
j	92 (10)	d	71 (17)	ʀ	81 (12)
w	97 (25)	dʒ	109 (18)	ʁ	62 (15)
ẽ	174 (46)	g	67 (16)	l	73 (16)
ẽ̃	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 duração⁷, 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:

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

⁶Para 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 idiosincrasia do locutor, somente fatores ligados a características inerentes à escolha dos logotomas do *corpus* explicam estes fatos.

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 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 /^lbo.sɐ/). O mesmo não ocorre com o GIPC: a unidade /os/, entre as palavras “ou” e “sofrerá” (pronunciado /so.fre.^lra/), 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.

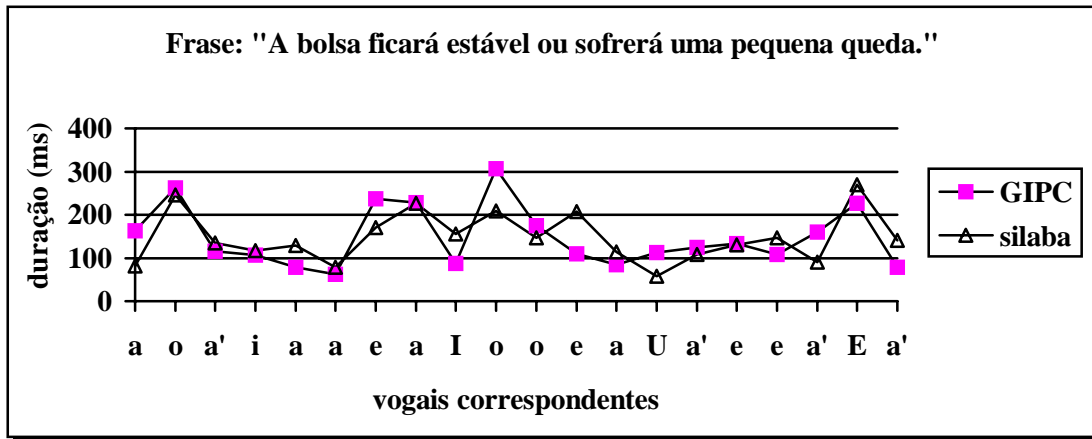


Figura 2: Evolução da duração da sílaba (*dsílaba*) e do GIPC (*dgipc*) ao longo da frase “A bolsa ficará estável ou sofrerá uma pequena queda.” As vogais correspondentes são indicadas no eixo horizontal (a' representa o alofone /ɐ/).

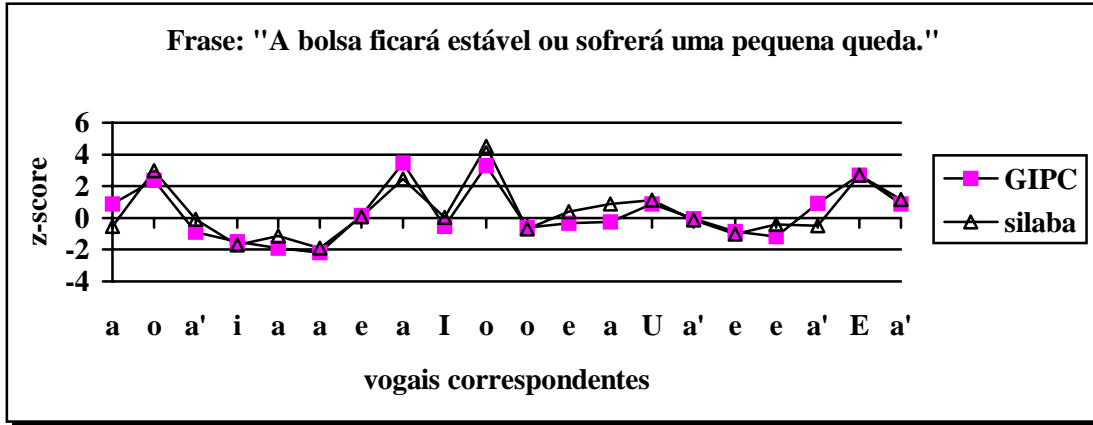


Figura 3: Evolução do z -score da sílaba ($z_{\text{sílaba}}$) e do GIPC (z_{gipc}) ao longo da frase “A bolsa ficará estável ou sofrerá uma pequena queda.” As vogais correspondentes são indicadas no eixo horizontal (a' representa o alofone /ə/). O último valor de z -score (GIPC) mostrado aqui é na verdade o z -score da última rima da frase (/ə/). 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 frequentes 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 rítmicas⁸ não é a mesma das introduzidas pelas fronteiras sintáticas⁹. De fato, ao observarmos a figura 4, a mais forte fronteira frasal,

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

representada 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;

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 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>.”

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 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, no prelo).

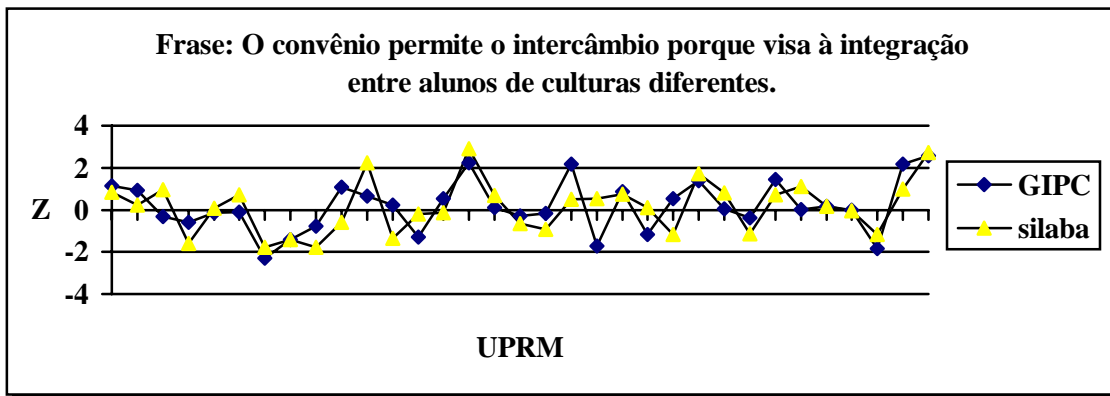


Figura 4: Z-scores (eixo vertical) para sílabas e GIPCs na frase “O convênio permite o intercâmbio porque visa à integração entre alunos de culturas diferentes.” Os traços no eixo horizontal representam a posição da vogal da sentença (assinalando tanto a sílaba como o GIPC, visto que as duas unidades 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 (/Is/).

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.

	<i>acento lexical</i>	<i>acento frasal</i>	<i>outras posições</i>
<i>onset/núcleo</i>	63	-31	4
<i>núcleo/onset</i>	ns	26	56
<i>núcleo/coda</i>	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 resultados¹¹.

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

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).

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.*, neste volume).

A coerência entre as estruturas sintático-fonológicas (no domínio da competência) e a realização dos 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 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 6 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 se encontra em fase de aprendizado para 50 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 apresentam 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 francês (Barbosa, 1994).

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 natural¹² na frase modificada. O outro distribuía os erros aleatoriamente (com distribuição 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 sujeitos preferiram em 89% dos casos, a frase gerada pelo nosso modelo de repartição. 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 frase¹³. Pesquisas tanto em produção quanto em percepção de fala mostram que a existência de tais unidades parece estar alicerçada na realidade cognitiva e psicoacústica.

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

¹³Ritmicidade também revelada pelas modulações de frequência fundamental e energia (cf. Madureira, 1994 e Aubergé, 1990).

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 “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 buscaram 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 em desenvolvimento para o PB é 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. (neste volume) *Um conversor ortográfico-fonético flexível para o português*.
- 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. (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.
- Barbosa, P.A. & Bailly, G. (1994) *Characterisation of rhythmic patterns for text-to-speech synthesis*, Speech Communication, 15 (1-2), 127-137.
- 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 dictionary of Linguistics and Phonetics*. Basil Blackwell 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*. Equipe 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, Massachusetts: 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. (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 conexcionista 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 providas 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 um 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 de 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.