MÁQUINAS FALANTES COMO INSTRUMENTOS LINGÜÍSTICOS: POR UM HUMANISMO *ÉCLAIRÉ*

Plínio Almeida Barbosa
LAFAPE/IEL - Unicamp

Toute langue est Beauté,
bijou, rosée, écho, souvenir :
un jeu. Un foisonnement de structures
accord folâtre de sonorités
âme passé présent mystère avenir
ANTOINE D'ARBOISE

RESUMO: *Este trabalho apresenta aspectos-chave da história da Síntese de Fala e de sua utilização como instrumento de pesquisa lingüística. Acompanhando o fascínio pela voz humana desde o tempo das tentativas de reprodução de cabeças e andróides falantes, passando pela realização da máquina falante do barão von Kempelen e pelos primeiros sintetizadores elétricos e primeiros sistemas de síntese digitais, queremos defender a tese de que a aventura pluri-inter-transdisciplinar de construir uma máquina falante se insere na mais absoluta tradição humanista e exige um movimento de tolerância das ciências humanas em relação às ciências naturais e vice-versa.*

ABSTRACT: *This work presents some landmarks in the history of speech synthesis and in the usage of speech synthesis for linguistic research. It traces the fascination for the human voice back to the first attempts to reproduce human-like talking figures, taking into consideration the century of the automata that culminates with Kempelen's speaking machine, and presents the first speech synthesizers and speech synthesis systems. By showing the pluridisciplinary character of this unique research adventure, this paper aims at defending at least one single thesis: the building of a speaking machine is an integral part of the humanistic tradition, and whatever the original affiliation of its builder may be, his/her work must be characterized by a tolerant attitude vis-à-vis the disciplines concerned.*

0. Introdução

EM UM SÉCULO em que se discutem as sérias questões bioéticas, filosóficas, políticas, econômicas e religiosas relacionadas à clonagem

de seres humanos, século herdeiro de um sem número de revoluções sociais, políticas, econômicas, culturais e científicas (em termos kuhnianos), a menção a *máquinas falantes* pode soar como mais uma ameaça ao Humanismo e à nossa humanidade.

Longe de representar uma ameaça, procuraremos mostrar neste artigo que a construção de uma máquina falante, e as reflexões científicas e tecnológicas em torno da mesma, convidam e levam justamente a um conhecimento aprofundado daquilo que é humano em nós, daquilo que mais caracteriza nossa humanidade: a capacidade de falar pelo uso de uma dupla articulação (morfemática e fonemática). Todos os trabalhos realizados em primatologia para avaliar a capacidade dos antropóides em se servir de línguas de sinais ou versões simplificadas das mesmas, têm o mesmo fim: delimitar o que nos é próprio.[1] Como estruturamos nossa mensagem pré-verbal, como utilizamos o conhecimento semântico-sintático e prosódico para organizar frases e como utilizamos o conhecimento fônico para pronunciá-las são algumas das questões que a construção de uma máquina falante pode ajudar a responder.[2]

Para se engajar em sua construção veremos que um conhecimento pluri-inter-transdisciplinar é necessário, mas não é suficiente. É ainda preciso que seu construtor ou cada um de seus construtores se distancie dos pressupostos teóricos e metodológicos de sua disciplina para questionar a natureza da relação entre os simbolismos discretizantes (afeitos à Lógica e às Linguagens Formais) e os simbolismos continuístas (afeitos ao Cálculo e às Leis da Física). Esses questionamentos são parte do dia-a-dia de uma área de pesquisa que ainda tateia ao procurar construir sua epistemе: as Ciências da Fala (vide apêndice).

Inserida plena e genuinamente nas Ciências da Fala, a Síntese de Fala, área de pesquisa que visa à construção de máquinas falantes, é um fascínio que tem suas raízes na mais remota história da humanidade. Na Antiguidade, as estátuas "falantes" de deuses e heróis míticos gregos, entre elas uma estátua parcialmente oca do Oráculo de Orfeu, na ilha de Lesbos (Dudley, Riesz & Watkins 1939, Flanagan 1972), eram usadas para simular vozes divinas. Mas se a voz se prestou a fins que hoje questionaríamos, o mesmo fascínio tomará de assalto diversos personagens a quem se atribui a construção de cabeças falantes, entre eles Gerbert d'Aurillac, o futuro papa Silvestre II, no século X, e Roger Bacon e Alberto, o Grande, no século XIII (Beaune 1980), para então ganhar direito a discussão filosófica no século XVIII. A discussão acalorada do Século das Luzes se origina a

partir de uma interpretação exclusivista do mecanicismo cartesiano e sua noção de *animal-máquina*, das discussões levantadas pelo empirismo dos britânicos Locke e Hume, sem excluir as inúmeras discussões suscitadas pelos autômatos, destacando-se aqueles de Jacques de Vaucanson, hábil construtor apaixonado pela Filosofia e pelas ciências da Anatomia e da Física.

Na primeira seção apresentamos *en vol d'oiseau* o interesse por máquinas simulando os movimentos humanos (incluindo os dos articuladores da fala), começando pelos autômatos e terminando no primeiro sintetizador articulatório, a máquina de von Kempelen. Na segunda seção, apresentamos os primeiros sintetizadores elétricos, obtidos a partir da analogia entre as leis físicas governando o movimento dos corpos e aquelas governando o funcionamento dos circuitos elétricos. A terceira seção é dedicada aos primeiros (verdadeiros) sistemas de síntese de fala integrando a manipulação de informação lingüística abstrata. A quarta seção conclui o artigo defendendo uma visão humanista esclarecida e tolerante com respeito à relação entre as disciplinas da Lingüística e entre essas e a Engenharia de Telecomunicações. Apresenta também os rumos da pesquisa em Síntese de Fala para a delimitação do que é especificamente humano em nós.

Outras revisões e histórias da Síntese de Fala encontram-se em Dudley (1955), Flanagan (1965, 1972) e Mattingly (1974), além da excelente página na Internet de Rubin e Vatikiotis-Bateson (s.d.). O livro de Flanagan & Rabiner (1973) é uma coletânea de artigos pioneiros na área. O livro de Köster (1973) apresenta uma visão detalhada das máquinas falantes e o artigo de Dennis Klatt (1987), sobre a história da síntese do inglês e o desenvolvimento do sintetizador que veio a receber seu nome, tornou-se um clássico.[3]

1. Dos autônomos como instrumentos filosóficos à máquina de Von Kempelen

A passagem a seguir, extraída do *Discours de la méthode*, apresenta claramente a noção do animal como máquina (trecho 1): para Descartes, os animais não são dotados de razão ou psiquismo (trecho 6), mas agem de acordo com a disposição atual de seus órgãos. Se hoje concebemos um psiquismo animal a ponto de falarmos de uma zoopsicologia, não podemos deixar de notar que o pensamento cartesiano tem um forte sabor de modernidade no que diz respeito ao humano em nós. Insinuando o seu

caráter simbolizante, o filósofo francês apresenta uma visão não preconceituosa do que parece ser a referência a uma proto-língua de sinais[4] (trecho 5). Além disso, procura assinalar dois pontos fundamentais de contraste entre a fala das máquinas falantes (como os papagaios, as pegas e um provável autômato falante) e a fala dos homens. A fala sintética é produto de uma ação que não parte do conhecimento de algo (trecho 4), como também não é produto de uma interação comunicativa (trechos 2 e 3). Esses dois pontos serão discutidos na seção 3 à luz da pesquisa recente em Síntese de Fala e Sistemas Automáticos de Diálogo.

"Ce qui ne semblera nullement étrange à ceux qui, sachant combien de divers *automates*, ou machines mouvantes, l'industrie des hommes peut faire, sans y employer que fort peu de pièces, à comparaison de la grande multitude des os, des muscles, des nerfs, des artères, des veines, et de toutes les autres parties qui sont dans le corps de chaque animal, considéreront ce corps comme une machine, qui, ayant été faite des mains de Dieu, est incomparablement mieux ordonnée et a en soi des mouvements plus admirables qu'aucune de celles qui peuvent être inventées par les hommes. **[1] Et je m'étois ici particulièrement arrêté à faire voir que s'il y avoit de telles machines qui eussent les organes et la figure extérieure d'un singe ou de quelque autre animal sans raison, nous n'aurions aucun moyen pour reconnoître qu'elles ne seroient pas en tout de même nature que ces animaux**; au lieu que s'il y en avoit qui eussent la ressemblance de nos corps, et imitassent autant nos actions que moralement il seroit possible, nous aurions toujours deux moyens très certains pour reconnoître qu'elles ne seroient point pour cela de vrais hommes : dont le premier est que jamais elles ne pourroient user de paroles ni d'autres signes en les composant, comme nous faisons pour déclarer aux autres nos pensées : **[2] car on peut bien concevoir qu'une machine soit tellement faite qu'elle profère des paroles, et même qu'elle en profère quelques unes à propos des actions corporelles qui causeront quelque changement en ses organes**, comme, si on la touche en quelque endroit, qu'elle demande ce qu'on lui veut dire; si en un autre, qu'elle crie qu'on lui fait mal, et choses semblables; **[3] mais non pas qu'elle les arrange diversement pour répondre au sens de tout ce qui se dira en sa présence, ainsi que les hommes les plus hébétés peuvent faire.** Et le second est que, **[4] bien qu'elles fissent plusieurs choses aussi bien ou peut-être mieux qu'aucun de nous, elles manqueroient infailliblement en quelques autres, par lesquelles on découvriroit**

qu'elles n'agiroient pas par connoissance, mais seulement par la disposition de leurs organes: car, au lieu que la raison est un instrument universel qui peut servir en toutes sortes de rencontres, ces organes ont besoin de quelque particulière disposition pour chaque action particulière; d'où vient qu'il est moralement impossible qu'il y en ait assez de divers en une machine pour la faire agir en toutes les occurrences de la vie de même façon que notre raison nous fait agir. Or, par ces deux mêmes moyens, on peut aussi connoître la différence qui est entre les hommes et les bêtes. Car c'est une chose bien remarquable qu'il n'y a point d'hommes si hébétés et si stupides, sans en excepter même les insensés, qu'ils ne soient capables d'arranger ensemble diverses paroles, et d'en composer un discours par lequel ils fassent entendre leurs pensées; et qu'au contraire il n'y a point d'autre animal, tant parfait et tant heureusement né qu'il puisse être, qui fasse le semblable. Ce qui n'arrive pas de ce qu'ils ont faute d'organes : car on voit que les pies et les perroquets peuvent proférer des paroles ainsi que nous, et toutefois ne peuvent parler ainsi que nous, c'est-à-dire en témoignant qu'ils pensent ce qu'ils lisent; **[5] au lieu que les hommes qui étant nés sourds et muets sont privés des organes qui servent aux autres pour parler,- autant ou plus que les bêtes, ont coutume d'inventer d'eux-mêmes quelques signes, par lesquels ils se font entendre à ceux qui étant ordinairement avec eux ont loisir d'apprendre leur langue. [6] Et ceci ne témoigne pas seulement que les bêtes ont moins de raison que les hommes, mais qu'elles n'en ont point du tout** : car on voit qu'il n'en faut que fort peu pour savoir parler; et d'autant qu'on remarque de l'inégalité entre les animaux d'une même espèce, aussi bien qu'entre les hommes, et que les uns sont plus aisés à dresser que les autres, il n'est pas croyable qu'un singe ou un perroquet qui seroit des plus parfait de son espèce n'égalât en cela un enfant des plus stupides, ou du moins un enfant qui auroit le cerveau troublé, si leur âme n'étoit d'une nature toute différente de la nôtre. Et on ne doit pas confondre les paroles avec les mouvements naturels, qui témoignent les passions, et peuvent être imités par des machines aussi bien que par les animaux; ni penser, comme quelques anciens, que les bêtes parlent, bien que nous n'entendions pas leur langage. Car s'il étoit vrai, puisqu'elles ont plusieurs organes qui se rapportent aux nôtres, elles pourroient aussi bien se faire entendre à nous qu'à leurs semblables. C'est aussi une chose fort remarquable que, bien qu'il y ait plusieurs animaux qui témoignent plus d'industrie que nous en quelques unes de leurs actions, on voit

toutefois que les mêmes n'en témoignent point du tout en beaucoup d'autres: de façon que ce qu'ils font mieux que nous ne prouve pas qu'ils ont de l'esprit, car à ce compte ils en auroient plus qu'aucun de nous et feroient mieux en toute autre chose; mais plutôt qu'ils n'en ont point, et que c'est la nature qui agit en eux selon la disposition de leurs organes : ainsi qu'on voit qu'un horloge, qui n'est composé que de roues et de ressorts, peut compter les heures et mesurer le temps plus justement que nous avec toute notre prudence."[5] (Descartes 1824, V, p. 185-189) [negritos e colchetes meus, itálico de Descartes].

O discurso cartesiano suscitou nos homens de seu século a vontade de construir máquinas que imitassem perfeitamente os movimentos naturais, incluindo-se os movimentos humanos. Opondo-se ao racionalismo cartesiano, as idéias empiristas de Locke foram traduzidas para o francês desde o início do século XVIII e continuadas pelo empirismo de Hume. O empirismo afirma que a origem do conhecimento deve ser buscada exteriormente ao homem, no mundo das sensações (ver também a obra de Condillac) ou na reflexão interna feita pelo homem a partir dessas mesmas sensações (Auroux 1990). Se a origem do conhecimento é exterior ao homem, entende-se a curiosidade do século da *Encyclopédie* pela observação, experimentação e imitação da natureza.

A imitação da natureza chegará à perfeição técnica em 1738 e 1739, quando da apresentação à Academia de Ciências de Paris dos autômatos do grenoblês Jacques de Vaucanson:[6] *Le Joueur de flûte traversière, Le Canard digérateur* e *Le Joueur de galoubet.*

O andróide-flautista, de elevada perfeição na coordenação dos movimentos dos lábios e da mandíbula, tocava doze árias e recebeu a atenção entusiasmada de todos os que ouviram suas performances musicais, como confirma o relato do abade Desfontaines (cf. Doyon & Liaigre 1967, p. 49-51):[7]

"C'est un faune assis sur un rocher qui joue de la flûte traversière et qui exécute, avec autant de force et d'élégance que de justesse et de précision, plusieurs airs de symphonie, dont quelques-uns sont assez difficiles tels que le Rossignol [ton ramage tendre] de Blavet dont ce faune a été le disciple.

C'est surtout dans les airs en do, la, ré qu'il brille parce que ce sont les plus favorables pour la flûte. Coups de langue marqués et précis sans enflés, et diminués avec goût, tenues gracieuses, ports de voix, pinces, coulés,

tremblements vifs, cadences perlées, échos mêmes ; aucun agrément n'est inconnu au flûteur inanimé.

Il joue des airs lents et rapides, de tendresse et de mouvement. Ici nulle supercherie : le vent qui sort par la bouche de l'automate se brisant au trou de l'embouchure forme les vibrations modifiées par ses doigts. Ce sont ses doigts posés différemment sur les trous de la flûte qui varient les tons, qui les pincent, qui les flattent, qui les cadencent. En un mot, l'art fait ici tout ce que fait la nature dans ceux qui jouent bien de la flûte. C'est ce qui se voit et s'entend, sans qu'il soit permis d'en douter."

O flautista também alimentou a imaginação de La Mettrie, já banido da França e da Holanda por suas idéias materialistas: "si il a fallu plus d'art à Vaucanson pour faire son flûteur que pour son canard, il eût dû employer encore davantage pour faire un parleur, machine qui ne peut être regardée comme impossible." (La Mettrie 1751 apud Boë 1997, p. 18). Esse entusiasmo do médico e filósofo se explica em meio a um clima de exílio, provocado pelo radicalismo que imprimiu às idéias cartesianas de animal-máquina estendidas ao homem: seu *Homme-Machine* (1748) defendia a tese de que o pensamento humano e a sua própria "alma" nada mais são do que uma propriedade da matéria orgânica.

Não é necessário fazer deduções como as de La Mettrie para nos admirarmos da possibilidade de um autômato falante, como prometera realizar Vaucanson, com o apoio real de Luís XV. As necessidades prementes da indústria da seda orientaram entretanto o genial construtor para outras preocupações (Doyon & Liaigre 1985, Balpe 1997). A tarefa de realizar um autômato falante não seria portanto cumprida. No entanto, no mesmo século, a máquina falante operada manualmente pelo barão húngaro von Kempelen é a mais próxima e mais impressionante realização técnica nessa direção (Fig. 1).[8]

Wolfgang Ritter von Kempelen (de seu verdadeiro nome Kempelen Farkas), advogado, engenheiro, construtor de autômatos,[9] artista e dramaturgo, construiu para a corte de Viena diversos dispositivos para facilitar a vida dos habitantes das cidades das atuais Áustria, Romênia, Hungria, República Checa e Eslováquia, incluindo uma máquina de imprimir para a senhorita Maria Theresia Paradis, escritora e musicista cega (Pompino-Marschall 1991, Ondrejovic 1996). Kempelen iniciou a construção de sua máquina em 1769, mas só a terminou cerca de vinte anos depois, em 1791 (apresentou, contudo, versões parciais ou exploratórias da mesma em uma turnê pela Europa, entre 1783 e 1785[10]).

Ao que tudo indica, o que norteou a construção da máquina foi a preocupação de Kempelen com a educação dos surdos-mudos. Ele descreveu a máquina e as etapas de sua construção de forma minuciosa na quinta parte de seu livro (Kempelen 1791), em que discorre ao longo de 456 páginas sobre linguagem, origem das línguas e aspectos diversos de produção e percepção da fala. O lançamento dos 195 exemplares de seu livro (122 em alemão e o restante em francês) foi um grande evento na época, com a presença de diversas personalidades, e é considerado um marco na história da Fonética Experimental e da Síntese de Fala (Ondrejovic 1996, Dudley & Tarnoczy 1950, Pompino-Marschall 1991). Ernst von Brücke, no século XIX, considerava-o "one of the best physiological books" e recomendava que todos os lingüistas que "want to make themselves throughly acquainted with the purely mechanical parts of sound theory" deveriam lê-lo (apud Tillmann 1994, p. 3083).

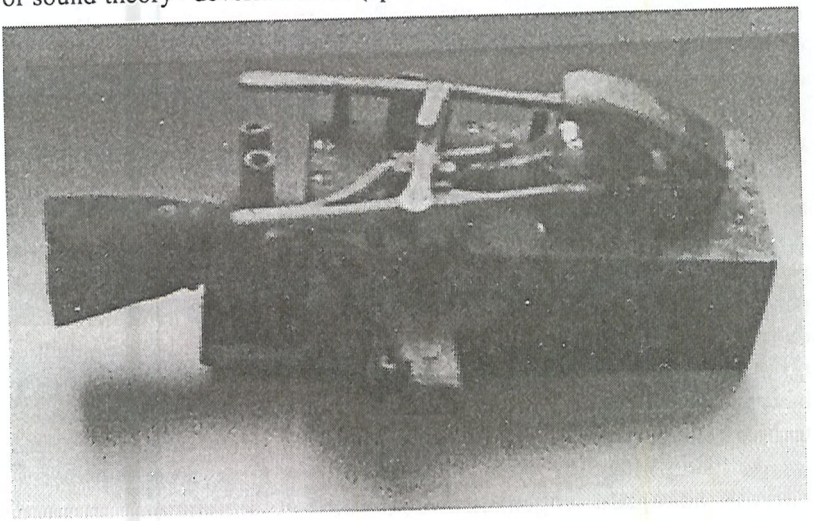

Figura 1: Foto da máquina de von Kempelen (do Deutsches Museum de Munique), reproduzida do site <http://www.ling.su.se/staff/hartmut/kemplne.htm> com a permissão de Hartmut Traunmüller.

É impressionante como Kempelen sintetiza de forma magistral aquilo que é mais relevante no funcionamento do aparelho fonador: "Zu einer Sprechmaschine braucht man also weiter nichts, dacht ich, als eine Lunge, eine Stimmritze, und einen Mund"[11] (Kempelen 1791, p. 398). Reconhece assim o que hoje denominamos os três subsistemas de produção da fala: o

respiratório (o *pulmão* é o órgão principal), o laríngeo (as pregas vocais, que definem a *glote*, são as principais estruturas) e o supralaríngeo (a língua que está na base da *boca* e o palato que *a* limita superiormente, são os órgãos mais importantes).

A máquina de von Kempelen simula, pela ação manual de suas estruturas, a etapa final do mecanismo de produção de fala: a articulação. Sua operação se dá através de treinamento de um operador da maneira que se segue (referir-se à figura 1). Pela ação cadenciada do cotovelo direito, o operador aciona um fole que gera a fonte de ar (o fole, colocado no extremo direito da máquina vista na foto, volta a sua posição de repouso pela ação de um contrapeso). Os dedos da mão direita controlam as alavancas e fecham (sons orais) ou abrem (sons nasais) os orifícios visíveis na figura1. Com essas ações manuais, simula o som de várias fricativas, oclusivas e nasais, num total de dezenove. As vogais apresentavam problemas de inteligibilidade, mas seu contraste é obtido deformando-se o ressoador de couro do extremo esquerdo da foto com a mão esquerda (Dudley & Tarnoczy 1950, Tillmann 1994). Os vinte anos da construção da máquina foram recompensados por uma grande perfeição técnica: a máquina da foto é operacional até hoje (Traunmüller 2000 testemunha que a operou em 1997 e se surpreendeu por seu bom estado de conservação e pela voz feminina ou quase infantil que sai de sua boca de couro[12]). Um testemunho anônimo entusiasmado, tirado do *Journal de Sçavans*, de 1783 (p. 629-630), atesta:

"Mais on voit chez lui [Kempelen] une autre machine qui n'a pas moins de mérite : c'est une machine qui parle & qui articule assez distinctement : 'Maman, aimez moi, allons à Paris', &c. Nous avons rendu compte du Mémoire qui a remporté le Prix de l'Académie de Pétersbourg en 1780,1 par M. 'Kratzenstein', sur la manière d'exprimer les sons des voyelles par des tuyaux d'orgues : mais on n'éttoit pas encore parvenue à imiter l'articulation des consonnes, & cette enterprise de M. de Kempelen annonce un talent également singulier ; il est à désirer qu'il publie bientôt les moyens" (apud Pompino-Marschall 1991, p. 199).

A máquina falante de Kempelen impressionou seus ouvintes por quase dois séculos. No século XIX, várias foram suas reproduções, destacando-se a do físico britânico Sir Charles Wheatstone, a de Alexander Graham Bell e a do matemático suíço Joseph Faber (Dudley & Tarnoczy 1950, Traunmüller 2000).

A máquina de Joseph Faber, construída em 1835 e demonstrada em Londres em 1846, denominava-se *Euphonia* e apresentava duas novidades: a operação via teclado e o controle do tom laríngeo via pedal (Fig. 2) (Flanagan 1972).

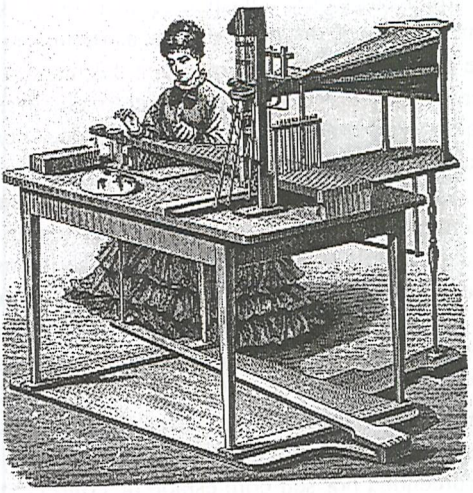

Figura 2: A Euphonia, reproduzida do site <http://mambo.ucsc.edu/psl/smus/smus.html> com a permissão de D. Massaro

Figura 3: A sra. Harper demonstrando o Voder. Figura reproduzida do site <http://mambo.ucsc.edu/psl/smus/smus.html> com a permissão de D. Massaro

É dela o testemunho entusiasmado de du Moncel:

"On s'est étonné que la machine parlante qui nous est venue, il y a quelques années d'Amérique, et qui a été exihibée au Grand-Hôtel fût d'une extrême complication, alors que le phonographe résolvait le problème d'une manière si 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" (Moncel 1880 apud Köster 1973).

Identificam-se na passagem acima dois métodos de reprodução sonora: o primeiro, o do fonógrafo, manipula o próprio sinal acústico, sem referência a como foi produzido. O segundo, o da máquina falante de Faber, simula a movimentação dos articuladores do aparelho fonador. Reconhecem-se nesse testemunho os dois métodos modernos de Síntese de Fala: a síntese acústica e a síntese articulatória, respectivamente (vide seção 2 para mais detalhes).

No século XX, com raras exceções (como Riesz 1937 apud Cater 1983 e Rubin & Vatikiotis-Bateson s.d.), as tentativas de construção de máquinas falantes não mais procuravam dispor e organizar engrenagens e outros dispositivos mecânicos, mas simular a produção de som através de equivalentes elétricos constituídos por circuitos dotados de resistores, capacitores e indutores, como se verá na próxima seção.

Todas essas máquinas, sejam elas mecânicas ou elétricas, apesar de representarem uma contribuição sem precedentes à Fonética, revelam tão somente um único aspecto da produção de fala: a produção de som a partir da disposição e movimento dos articuladores da fala. Essas máquinas constituem o domínio da Síntese Articulatória (que tem a máquina de Kempelen como ancestral).

Por sua vez, o uso intenso da espectrografia e mesmo o uso de fitas magnéticas vão naturalmente conduzir às primeiras experiências com a percepção da fala, por meio das primeiras sílabas e palavras sintéticas. A partir desses experimentos, um conhecimento fonético-acústico se constitui e permite o aparecimento da Síntese Acústica.

2. Os sintetizadores elétricos: simulação de aspectos mecânicos da produção da fala e primeiros experimentos em percepção da fala

Para realizar a síntese da fala é preciso reunir e organizar, segundo critérios lingüísticos e de produção de fala, além de decisões tecnológicas, diversos elementos, blocos e subestruturas lingüísticas e sonoras previamente analisadas. Toda síntese da fala, pressupõe, portanto, uma etapa de análise. Também a pré-história da Síntese de Fala, através dos sintetizadores mecânicos acima referidos e dos sintetizadores elétricos que serão apresentados nesta seção, testemunha várias décadas de experimentos com os sons da fala, experimentos estes orientados por teorias acústicas que procuravam mostrar como o movimento dos articuladores do aparelho fonador, excitados pela fonte de ar dos pulmões, produz som.

No que diz respeito às vogais, há mais de dois séculos que tanto sua percepção quanto sua produção estão associadas ao conhecimento de que o som é mais intenso em determinadas regiões no domínio da freqüência (Russel 1928 apud Dunn 1950). Em tais regiões se encontra o que chamamos de *formantes*, que são freqüências de ressonância, freqüências de ondas estacionárias que se formam nos tubos constituídos na boca, na faringe e eventualmente no trato nasal (se o véu palatino estiver abaixado) pela posição atual dos articuladores da fala. O fenômeno da ressonância pela formação de ondas estacionárias é o mesmo que se dá quando tocamos as cordas de um violão ou de um piano, quando tocamos uma flauta ou qualquer outro instrumento de sopro, quando empurramos alguém em um balanço ou mesmo quando um cantor lírico quebra um cristal.

Os formantes de uma vogal podem ser facilmente identificados em um espectrograma como o da figura 4, obtido a partir de um espectrógrafo digital (do CSL 4300B, da Kay Elemetrics): são os trechos mais escuros que evoluem temporalmente da esquerda para a direita e que apresentam três ou quatro faixas mais escuras por vogal (no sentido vertical). Um conjunto de formantes caracteriza uma vogal específica (mas a caracterização precisa varia segundo o contexto fônico segmental e prosódico, a situação, o estado emocional, o momento da elocução e a pessoa que enuncia). As consoantes têm uma descrição mais complexa e são responsáveis pela modificação da pronúncia de uma vogal, introduzindo claras alterações nas margens vocálicas que se estendem freqüentemente até a região central de tais vogais.

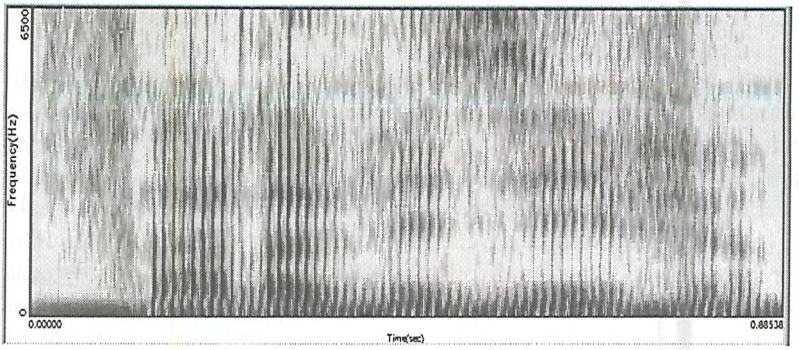

Figura 4: Espectrograma do enunciado "Fala visível", produzido pelo autor

2.1. Sintetizadores articulatórios

A compreensão da natureza ressoante do som foi possível graças ao trabalho de figuras como a de Helmholtz (1877). Baseado em noções helmholtzeanas, como a de que cada vogal possuiria até dois formantes, o análogo elétrico de trato vocal de Stewart (1922) é o primeiro sintetizador totalmente elétrico.[13]

O desenvolvimento da Teoria da Comunicação, a partir da necessidade de otimizar o uso do canal de transmissão para o telefone e o telégrafo (lembramos aqui o papel singular de Graham Bell, Claude Shannon e o próprio Dudley), permitirá o desenvolvimento do Vocoder por Homer Dudley (1939), do Bell Labs. Através dele, o som da fala era codificado pelo analisador (o emissor), que separava os componentes devidos à fonte sonora (pregas sonoras ou ruídos turbulentos) dos devidos às cavidades supraglotais. Um número reduzido de parâmetros podia assim ser transmitido pelo canal para ser decodificado à recepção. Para a decodificação, um oscilador de relaxação ou um gerador de ruído aleatório eram usados, segundo a informação dada pelo transmissor (quer se tratasse, respectivamente, de som produzido exclusivamente por vibração das pregas vocais ou não). Um circuito de controle de ressonância simulava o trato vocal. É justamente um dispositivo semelhante ao receptor do Vocoder que constituirá um segundo sintetizador elétrico, o Voder, ou *Voice Demonstrator* (Dudley, Riesz & Watkins 1939. Cf. também Dudley & Tarnoczy 1950), que foi demonstrado nas Feiras Mundiais de Nova York e

São Francisco em 1939 (fig. 3; note a impressionante semelhança entre essa figura com aquela acima dela, figura 2).

Circuitos como o Voder ou o modelo elétrico de Dunn (1950) simulavam a transdução articulatório-acústica realizada por dois elementos fundamentais: a fonte sonora (fonte de som das pregas ou o ruído produzido pelo estreitamento de alguma região supraglotal) e o filtro (efeito produzido pelas características ressoantes das cavidades do trato vocal). É imediato perceber do diagrama do Voder (Fig. 5) a equivalência entre circuito elétrico e regiões do aparelho fonador: o oscilador de relaxação simula o movimento das pregas vocais, o gerador de ruído aleatório, as turbulências de sons fricativos e oclusivas surdas e o controle de ressonância (realizado por dez filtros ressoadores), simula o trato vocal. É interessante notar a semelhança entre as teclas do Voder para interromper o sinal (t-d, p-b e k-g 'stops') e simular as oclusivas, e as alavancas da máquina de von Kempelen (cf. fig. 1). Como na máquina de Faber (fig. 2), o pedal também controlava a altura melódica.

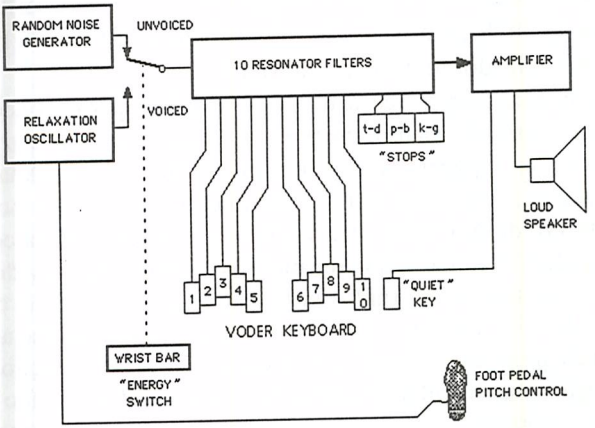

Figura 5: Diagrama do VODER, reproduzido do site <http://www.haskins.yale.edu/haskins/HEADS/SIMULACRA/voder.html> com a permissão de Philip Rubin

A partir do circuito elétrico desenvolvido por Dunn (1950) e mais tarde por aquele de Stevens e colegas (1953), em que a fala sintética é produzida a partir de controle manual, Rosen (1958) desenvolverá o primeiro circuito para a realização de síntese articulatória de forma automática, o *Dynamic Analog of the Vocal Tract*, ou DAVO (Klatt 1987).

O desenvolvimento da Síntese Articulatória nos Estados Unidos continuará a se desenvolver ininterruptamente, contando com a presença de pesquisadores europeus e japoneses (como os suecos Gunnar Fant, Rolf Carlson, Björn Granström, e os japoneses Osamu Fujimura, Shinji Maeda, E. Matsui, Noriko Umeda, entre outros).

A formulação matemática precisa de uma relação articulatório-acústica por Dunn (1950) e a Teoria das Perturbações de Chiba & Kajiyama (1941) vão colocar em cheque a Teoria das Cavidades Ressoantes de Helmholtz e conduzir à moderna Teoria Acústica de Produção da Fala, que apresenta uma formulação mais completa no trabalho de Gunnar Fant (1960). O desenvolvimento dessa teoria foi possível graças aos resultados dos primeiros experimentos com leitura de espectrogramas (vide seção 2.2), bem como aos avanços matemáticos dos séculos anteriores (Teoria das Equações Infinitesimais de Newton e Leibniz e a Análise Espectral de Joseph Fourier; cf. Dahan-Dalmedico & Peiffer 1986, para detalhes sobre a história da matemática), que conduziram a equações que expressam a relação entre as dimensões dos tubos no trato vocal e os formantes, isto é, que estabelecem uma relação articulatório-acústica. Sendo assim, o advento da espectrografia e da nova Teoria Acústica de Produção da Fala aliado ao advento da computação digital serão responsáveis pelo desenvolvimento e pelo interesse em Fonética Acústica, inclusive por lingüistas como Martin Joos (cf. Mattingly 1999).

2.2. Sintetizadores acústicos

A manipulação de parâmetros como os formantes, visíveis nos traçados dos primeiros espectrógrafos (Koenig, Dunn & Lacy 1946, Potter 1945), construídos nos Estados Unidos, e a pronúncia dos espectrogramas por uma máquina, tornarão possíveis os primeiros experimentos de percepção usando a fala sintética. A leitura apropriada dos espectrogramas, procurando identificar nos parâmetros acústicos *invariança* (via parâmetros passíveis de discretização) e *variabilidade* (devido à coarticulação – já apontada por Potter, Kopp & Green 1947 e Joos 1948 –, e entendida como o resultado acústico da produção simultânea dos segmentos da fala. Como exemplo de fenômeno coarticulatório temos o arredondamento do /t/ e do /d/ na palavra "tudo", devido à presença simultânea do arredondamento do /u/ naqueles segmentos) foi possível graças ao trabalho de Potter et al. 1947, cujo livro, *Visible Speech*, leva o título da obra de Alexander Melville Bell (1867 apud Tillmann 1994), esta última propondo um alfabeto

universal, um precursor do atual Alfabeto Fonético Internacional.[14] O *Pattern Playback* (Cooper, Liberman, Borst & Gerstman 1951)[15] é um exemplo digno de nota desse tipo de máquina, tendo em vista os experimentos que dela se serviram por cerca de quinze anos (cf. Mattingly 1999).

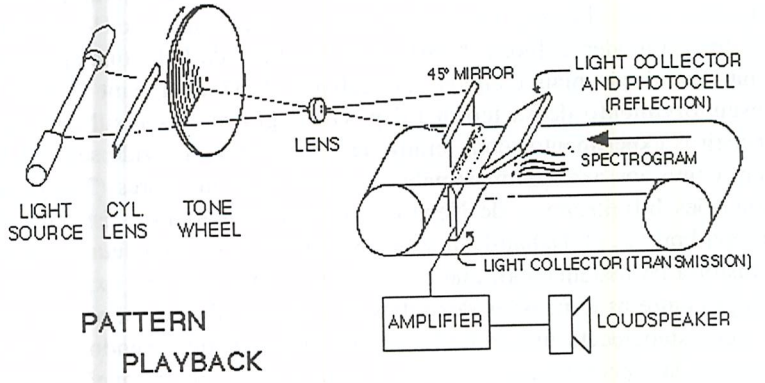

Figura 6: Pattern Playback, reproduzido do site <http://www.haskins.yale.edu> com a permissão de Philip Rubin

O funcionamento do Pattern Playback pode ser descrito sucintamente a partir da fig. 6. A fonte de luz à esquerda do dispositivo passa por um conjunto formado por um disco (que modula a luz à freqüência da escala do espectrograma), um jogo de lentes e um espelho a 45 graus, para então incidir sobre um espectrograma real ou estilizado, apresentado sobre a esteira do dispositivo, que se desloca com a taxa de elocução de uma gravação original ou sintética. O traçado do espectrograma é, no entanto, feito com tinta branca, para que os diferentes níveis de reflexão da luz, ao atingir os diversos matizes de branco, gerem os correspondentes níveis de corrente elétrica pela transdução realizada por uma célula foto-elétrica (descrição ligeiramente modificada a partir daquela em Cooper et al. 1951).

A estilização progressiva de espectrogramas por tentativa e erro, a partir de um espectrograma real, e a audição dos resultados sonoros através do Pattern Playback permitiram a investigação dos parâmetros acústicos mais relevantes para a percepção dos diferentes segmentos de fala,

possibilitando assim o estabelecimento sistemático das primeiras relações acústico-auditivas. Foi usando justamente o Pattern Playback que trabalhos como os de Delattre, Liberman, Cooper & Gerstman (1952), de Cooper, Delattre, Liberman, Borst & Gerstman (1952) e de Delattre, Liberman & Cooper (1955) demonstraram a importância dos dois primeiros formantes para a identificação das vogais, a do terceiro formante para a identificação de consoantes como o /r/ retroflexo americano e tanto da localização dos picos de amplitude no instante da soltura das oclusivas quanto dos movimentos de formantes na transição consoante-vogal para a identificação das mesmas: "these rapid changes are heard as important distinguishing characteristics of the sound stream and may indeed serve as a principal cue for the perception of the consonant-vowel combination – the syllable or 'half-syllable', as the case may be (Joos 1948, p. 122)." (Cooper et al. 1951). A necessidade da presença simultânea de alguns desses parâmetros fonético-acústicos é reconhecida não somente como forma de garantir a robustez da comunicação, mas também como forma de assegurar a inteligibilidade da fala (ibidem).

Todos esses estudos (Cooper et al. 1952, por exemplo) apontam para a possibilidade de realização de Síntese de Fala a partir do conhecimento dos valores dos parâmetros acústicos em zonas estáveis da progressão dos mesmos ao longo de um enunciado (por exemplo, os três primeiros formantes na região média de uma vogal plena), bem como dos movimentos transicionais dos mesmos parâmetros (por exemplo, o movimentos dos três primeiros formantes durante a transição de ataques consonantais simples ou complexos para uma vogal). A síntese que se fundamenta na descrição completa dos parâmetros fonético-acústicos para todo e qualquer enunciado é denominada Síntese (Acústica) Paramétrica (ou Síntese por Regras ou ainda Síntese por Formantes). Explorações tecnológicas que também contribuíram para o desenvolvimento de conhecimento fonético-acústico dizem respeito justamente à construção, também na década de 1950, de dois sintetizadores acústicos completamente elétricos, os primeiros em sua categoria a serem controlados automaticamente: o *Orator Verbis Electrics I* (OVE I), de Gunnar Fant (1953) e o *Parametric Artificial Talker* (PAT), de Walter Lawrence (1953). Ambos simulam os três primeiros formantes a partir de circuitos distintos: o primeiro com filtros analógicos organizados em série e o segundo, com filtros dispostos em paralelo. Os dois sintetizadores vão mesmo "dialogar"

durante uma conferência em 1956 no Massachussets Institute of Technology (MIT) (cf. Klatt 1987, p.742). A preferência por um sintetizador acústico com circuito paralelo se estabelecerá após Holmes ter conseguido, em 1972, produzir enunciados sintéticos utilizando uma versão aperfeiçoada do PAT em relação aos quais os ouvintes não conseguiam distinguir diferenças em relação a enunciados naturais[16] (Holmes 1973, sem esquecer de ouvir esses mesmos exemplos sonoros no site <http://www.cs.indiana.edu/rhythmsp/ASA/partA.html>, acesso em 24 jun. 2001).

Detalhes precisos do conhecimento fonético-acústico necessário para a Síntese de Fala já estão presentes desde a década de 1950, com trabalhos como o de Liberman, Ingemann, Lisker, Delattre e Cooper (1959). A partir da década de 1960, o trabalho de aperfeiçoamento de sintetizadores paramétricos prosseguirá com o desenvolvimento de circuitos e algoritmos para computadores digitais, como as regras para modificação de valores de freqüência fundamental (o correlato acústico da freqüência de vibração das pregas vocais) e para implementação das transições dos formantes (os valores estáticos são tabelados) aplicadas a um sintetizador paralelo por Holmes, Mattingly e Shearme (1964).

Logo cedo, esses sintetizadores permitiram a realização de experimentos com fala sintética, obtida a partir dos mesmos, e não mais com manipulações manuais como as obtidas com o Pattern Playback (cf. Fry, Abramson, Eimas & Liberman 1962). A maior figura envolvida no trabalho de desenvolvimento de sintetizadores e, mais tarde, sistemas de síntese paramétricos (vide seção 3) é certamente aquela de Dennis Klatt, que dedicou os vinte anos finais de sua vida para a Síntese de Fala (cf. Stevens 1991, 1992). Klatt é responsável pelo desenvolvimento de um sintetizador com circuitos em série e em paralelo juntamente com um algoritmo de regras de modificação de parâmetros fonético-acústicos, cujos comandos, escritos em FORTRAN, foram publicados para serem largamente usados pela comunidade científica (Klatt 1980). A síntese paramétrica é considerada por todos esses pesquisadores como a "true synthesis" (expressão usada em Liberman et al. 1959), por oposição a um outro tipo de Síntese Acústica, a eles contemporânea: a Síntese (Acústica) Concatenativa.

Também refletindo a partir da observação de espectrogramas, Harris (1953) efetua experimentos de percepção com a edição de trechos de fala gravados em fita magnética. Esse artigo exemplifica o quanto de

trabalho puramente mecânico está relacionado ao tipo de empresa da Síntese Concatenativa.[17] Esse método de Síntese de Fala fundamenta-se na possibilidade de justapor, de concatenar segmentos sonoros mínimos em algum sentido (adequadamente referidos por Harris como "building blocks") para a obtenção (ou "construção", como o termo de "building blocks" conduz a pensar) de uma frase qualquer. Duas possibilidades se abrem ao espírito: as unidades concatenantes podem ser palavras ou podem ser fones (segmentos acústicos correspondendo à realização de um fonema e sendo quase-isomórficos ao mesmo). A primeira opção era impossível na época de Harris, por razões tecnológicas relacionadas à impossibilidade de armazenamento de milhões de formas lexicais e a rápida recuperação das mesmas para a síntese, sem falar da variabilidade fonético-acústica, em princípio *ad infinitum*, pelo fato de as palavras serem pronunciadas sob condições prosódicas as mais diversas. A segunda possibilidade foi aquela testada por Harris. Ele demonstra habilmente que a concatenação de todas as consoantes de *onset* (como o /p/ de "pik") do inglês a uma rima proveniente de um ambiente único (o que ele usou foi o /ik/ de "kik"), produzindo, portanto, sílabas sintéticas como "p-ik" (com o /p/ editado da sílaba "pik" pronunciada naturalmente é o /ik/ editado da sílaba "kik" natural), gera problemas de inteligibilidade em 34 ouvintes: a maioria deles disseram ter ouvido "tik", ao escutarem a sílaba sintética p-ik.

Experiências como esta revelam que não é possível interromper o movimento transicional dos formantes na fronteira entre consoante e vogal, isto é, a transição CV precisa ser preservada em sua integralidade. De onde a proposta de unidades concatenantes que preservem essa transição. Essas unidades, chamadas em um primeiro momento de díades (Peterson, Wang & Sivertsen 1958) serão mais conhecidas por *difones*. O difone é definido como "segmento acústico que se estende da região estável de um fone para a região estável do próximo fone" (ligeiramente modificado de Peterson et al. 1958, p. 739). Mais recentemente, generalizou-se essa definição para considerar três ou mais fones, identificando-os como trifones e, genericamente, polifones. Assim, para sintetizar a palavra "bola" [b•l?], precedida e seguida de silêncio (representado aqui por "_") é necessário concatenar cinco difones: "_b", "b• ", "• l", "l?" e "?_". Mas se considerarmos que essa mesma palavra pode ser enunciada em diversas posições numa frase e que os parâmetros prosódicos clássicos, como a duração, a freqüência fundamental e a amplitude, variam muito segundo essa posição, seria necessário gravar os

difones sob condições prosódicas distintas, para obter naturalidade na pronúncia de uma simples palavra. Foi essa justamente a idéia inicial de Peterson et al (1958): suas díades são, na verdade, uma classe de difones gravados em condições prosódicas distintas. Por isso calculam que cerca de 8000 díades seriam necessárias para sintetizar frases do inglês americano. Juntando-se a isso o fato de que algumas unidades deveriam ser sílabas CVC (trifones), esse número seria ainda bem maior. Ora, nas décadas de 1960 a 1980, a dificuldade de memória física para o armazenamento de um tal número de unidades, bem como para a rápida recuperação das mesmas para a síntese de fala, era impensável.

Somente o advento de uma nova técnica permitirá um reganho de interesse pela Síntese Concatenativa. Essa técnica, desenvolvida por Moulines e Charpentier (1990) recebeu o nome de PSOLA, de *Pitch Synchronous Overlap and Add*. Como o nome indica, um polifone pré-gravado, com os períodos glotais devidamente marcados (de onde *pitch synchronous*), tem sua duração e freqüência fundamental atualizados para atender aos valores especificados à entrada do sistema de síntese concatenativa, a partir do contexto lingüístico apropriado. Os valores são ajustados ao mesmo tempo em que se concatenam os polifones necessários para realizar a frase a ser sintetizada. Essa concatenação não é feita pela justaposição de polifones consecutivos, mas entrelaçando parcialmente os mesmos (de onde *overlap and add*): os valores de amplitudes do sinal acústico na região de entrelaçamento são obtidos via média ponderada dos valores de amplitudes de cada um dos trechos dos polifones sendo concatenados.

Da simples manipulação de parâmetros manual ou automaticamente, a partir do conhecimento meramente fonético da produção de som, a Síntese de Fala vai deixar de se identificar com a construção de "sintetizadores" para adentrar a era dos Sistemas de Síntese de Fala, a partir da década de 1970.

3. Máquinas falantes como instrumentos lingüísticos

Os anos 1970 vão conhecer a necessidade da formulação de regras de natureza lingüística para a realização da passagem de uma sucessão de fonemas, obtidos de uma correspondência grafema-fonema a partir de um texto escrito ou de uma correspondência sintaxe-fonologia a partir de uma representação semântica de uma mensagem, para uma seqüência de

parâmetros acústicos variando continuamente no tempo. Acrescidos aos sintetizadores explicitados acima, esses algoritmos desenvolvidos para computadores digitais (observar aqui o desenvolvimento da Síntese de Fala atrelado ao desenvolvimento tecnológico) vão constituir os primeiros Sistemas de Síntese de Fala. Do que acaba de ser dito, pode-se identificar, segundo a natureza da representação à entrada do sistema, duas classes principais: os sistemas *Text-to-Speech* (TTS), isto é, a partir do texto escrito, e os sistemas *Concept-to-Speech* (CTS), a partir de entrada conceitual. A primeira classe de sistemas simula a leitura de um texto em voz alta. A segunda simula o mecanismo completo de produção de fala (se excluirmos, por um momento, os aspectos dialógicos).

Os algoritmos que manipulam representações fonológicas utilizam regras de produção e reconhecem-se devedores, nesse aspecto, do Componente Fonológico de *The Sound Pattern of English, SPE* (Chomsky & Halle 1968). As regras de atribuição de acento em SPE foram usadas por Sharon Hunnicutt e Francis Carroll em seu algoritmo de conversão grafema-fonema, em 1972. No entanto, esses algoritmos vão também explicitar como modificar valores de parâmetros fonético-acústicos a partir de entrada simbólica binária (saída do componente estritamente fonológico) ou escalar (saída do subcomponente fonético do componente fonológico), uma preocupação que não faz parte do *cahier de charges* do SPE (que considera uma Fonética não-sistêmica universal, isto é, um mero executor mecânico de especificações simbólicas. Vide também Albano, Barbosa, Gama-Rossi, Madureira & Silva 1998 para uma crítica a essa visão).

3.1. Sistemas de síntese a partir do texto escrito (TTS)

Para uma melhor compreensão das diversas etapas necessárias à realização de síntese de fala a partir do texto escrito, referir-se à figura 7. A etapa de pré-processamento, que não será detalhada aqui, consiste na obtenção de uma representação fonológica para permitir a pronúncia por extenso de abreviações que ocorrem nos textos escritos em decorrência do uso de símbolos especiais (grandezas físicas, moeda, sinais matemáticos), siglas, abreviaturas e números. É por vezes um algoritmo complexo, tendo em vista que o conjunto das regras de pronúncia dessas abreviações apresenta exceções (compare U.T.I. com ITA e IEL), ou

necessita mesmo de informações discursivas (por exemplo, em "Ele só pode chegar em 2/6" , por extenso: "Ele só pode chegar em dois de junho", e não "[...] em dois sextos").

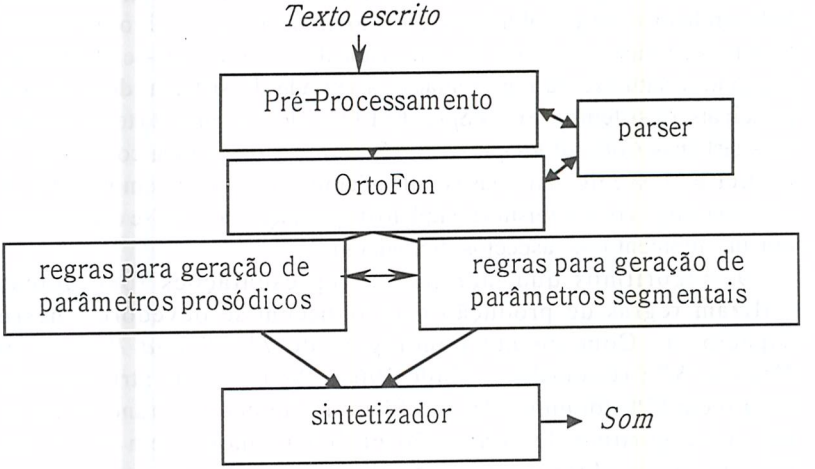

Figura 7: Diagrama geral de um sistema de síntese de fala a partir do texto escrito

O algoritmo OrtoFon realiza a passagem da representação em grafemas (ortográfica) para uma representação fônica (fonemas, arquifonemas ou alofones). É realizado na forma de um conjunto de regras de produção, e requer muitas vezes um *parser* (palavra derivada da expressão latina *pars orationis*), que fornece informações de natureza morfológica ou morfossintática (veja o caso de "Eu *piloto* o avião pela manhã, como todo *piloto* da Força Aérea").

A partir da representação fônica são necessárias regras para a geração de parâmetros fonéticos acústicos ou articulatórios, sejam eles de natureza prosódica ou segmental, como se verá mais adiante, no contexto histórico. Em um sistema de síntese de fala, o sintetizador passa a ser apenas a etapa final da conversão texto-fala. Esse dispositivo toma as especificações numéricas obtidas pelas regras anteriores e as converte em som gerando valores de formantes e parâmetros prosódicos acústicos ao longo do tempo através de fórmulas relacionando as posições de articuladores com os mesmos (sintetizador articulatório), ou converte especificações diretamente acústicas em som (sintetizador paramétrico).

É importante observar que um sistema de síntese concatenativo não precisa gerar parâmetros segmentais, mas recupera e concatena trechos

de som pré-gravados em um inventário de unidades, sejam elas polifones ou demi-sílabas (essa última preserva a transição CV, mas, diferentemente do difone, inclui toda a rima da sílaba de onde foi extraída. Vide Fujimura e Lovins 1978), modificando os parâmetros prosódicos clássicos a partir de regras de um modelo prosódico. Para realizar a concatenação, pode usar a técnica PSOLA como sintetizador ou uma técnica de busca estatística de unidades gravadas sob diversas condições prosódicas (cf. Campbell & Black 1997 a respeito dessa busca que minimiza a discrepância sintagmática entre os trechos concatenados e maximiza a verossimilhança paradigmática do trecho correspondendo a uma dada entrada lingüística). Outra técnica possível é o modelo híbrido ou *harmonic + noise* (Boëffard & Violaro 1994), que permite um ajuste mais fino dos valores dos parâmetros prosódicos de duração e de freqüência fundamental (a técnica PSOLA só permite ajuste para valores múltiplos do período glotal).

Os componentes fundamentais da síntese TTS são o *OrtoFon* e o modelo prosódico. No que diz respeito aos sistemas com sintetizadores paramétricos, de início foram usadas regras muito simples de conversão grafema-fonema para o inglês americano, como as de Carlson e Granström (1975).[18] Trabalhos como esse, que descrevem regras de reescrita para a pronúncia de todos os grafemas do inglês, bem como atribuem acento a partir de um conjunto de regras cíclicas, tiveram como precursores o algoritmo de Mattingly (1968), o primeiro para a obtenção de valores dos parâmetros prosódicos clássicos a partir de informação lingüística. No entanto, a necessidade do detalhamento da informação morfológica para a pronúncia adequada do inglês americano conduzirá ao desenvolvimento de um dicionário com cerca de 12000 morfemas, que integrará o sistema MITalk (Allen, Hunnicutt, Carlson & Granström 1979 e Allen, Hunnicutt & Klatt 1987). Se a decomposição em morfemas é satisfatória, a pronúncia dos mesmos dá a pronúncia das palavras primitivas; se não, um algoritmo de regras trata dos casos pendentes. Os morfemas foram obtidos a partir da análise do *Brown Corpus*, um corpus com mais de um milhão de palavras gravado na década de 1960. O advento do MITalk foi possível graças aos trabalhos de Lee (1969), Klatt (1970) e Allen (1973). O trabalho de Klatt (1970) também permitiu o desenvolvimento do sistema paramétrico Klattalk (Klatt 1981, 1982), que foi comercializado em 1983 pela Digital Corporation e recebeu o nome de DECTalk.

No que se refere à Síntese Concatenativa, a partir dos trabalhos já citados de Harris (1953) e Peterson et al. (1958), surgem os primeiros

sistemas empregando díades e difones, como o sistema de Dixon e Maxey (1968), para o inglês, o de Leipp e colegas (1968), para o francês, além do trabalho sobre o japonês, de Saito e Hashimoto (1968).O sistema de Olive (1977), também para o inglês, apresentou bons resultados e deu origem ao sistema da AT&T Bell Labs. Também merece destaque o trabalho de Browman (1980), que utiliza outra unidade concatenante, a demi-sílaba. Essa é definida no trabalho pioneiro de Fujimura e Lovins (1978). Um ganho de naturalidade, segundo os autores, parece advir da opção por essa unidade, ao invés do tradicional difone ou mesmo do polifone. Dados em português brasileiro (PB) o confirmam, sobretudo em rimas VN, isto é, contendo uma vogal nasalizada (Albano & Aquino 1997). Essa maior naturalidade, juntamente com o fato de que o acento frasal incide primordialmente na rima (cf. Vaissière 1983 e Barbosa 1996, para o PB), justificou seu uso na Síntese Concatenativa do PB (Barbosa, Violaro, Albano, Simões, Aquino, Madureira & Françozo 1999).

Algoritmos ainda mais complexos integram o sistema de síntese articulatória desenvolvido por Coker, Umeda e Browman (1973) que contou com a experiência do primeiro sistema de síntese de fala para o inglês (desenvolvido por japoneses), aquele de Matsui, Susuki, Umeda e Omura (1968). O sistema de Coker et al. (1973) empregava algoritmos para realizar a análise sintática (um *parser*), para a atribuição de pausas (silenciosas e subjetivas, via alongamento de final de constituinte e variação de altura) e acento, para o cálculo de duração de vogais, para a atribuição da freqüência fundamental e detalhes de variação alofônica.

A partir de modelos como o de Mattingly (1968) e o de Coker et al. (1973), outros modelos prosódicos vão surgir e se multiplicar. Para a obtenção da duração, destaca-se o modelo segmental de Klatt (1973, 1979), descrito sucintamente em Klatt (1987), com regras obtidas a partir de análise fonética exaustiva de variáveis lingüísticas e extralingüísticas que afetam a duração de um simples segmento (Lehiste 1970, Klatt 1976b). Regras para a obtenção da duração de unidades do tamanho ou acima da sílaba, por considerar o nível prosódico como primeiro, vão também ser propostas (Kohler 1986, para o alemão). Autores como Campbell e Isard (1991), para inglês britânico, partem de um nível prosódico, mas geram a duração de unidades do tamanho da sílaba a partir do paradigma de aprendizado automático, via redes neurais. Barbosa e Bailly (1994), para o francês, seguem a mesma linha, mas já têm uma preocupação com uma via lingüisticamente mais plausível (cf. também Barbosa 2001). Para uma

revisão de modelos de duração ver o artigo de Carlson (1991) e o capítulo 2 da tese de Barbosa (1994).

Um modelo entoacional, isto é, um algoritmo para a geração de valores de freqüência fundamental via regras, foi pioneiramente desenvolvido por Mattingly (1966) para o inglês britânico e adaptado posteriormente para o inglês americano (Mattingly 1968). Vários estudos descritivos sobre a entoação do inglês britânico (cf. Armstrong & Ward 1931) e do inglês americano (cf. Pike 1945) o precederam. Curvas entoacionais básicas como "fall", "rise" e "fall-rise" eram associadas à última sílaba proeminente de um sintagma. Um outro trabalho importante em síntese da entoação que se seguirá é aquele de Pierrehumbert (1981), que vai dar origem ao sistema ToBI (*Tone and Break Indices*), que propõe uma notação abstrata para a curva de freqüência fundamental em inglês americano. Para o japonês, o modelo de comandos de Fujisaki e Nagashima (1969) para a geração dos contornos entoacionais teve vários seguidores (por exemplo, Bailly 1989).

No que diz respeito aos modelos segmentais e modelos prosódicos, é preciso considerar que, para reproduzir o mecanismo de produção da fala, a interação entre os dois tipos de modelos é crucial (de onde a seta de duplo sentido na fig. 7). Os modelos aqui apresentados de forma sucinta procuram levar essa interação em conta, de alguma maneira. As regras de obtenção de duração de Klatt (1987), por exemplo, especificam um percentual variável de aumento de duração de um segmento pré-pausal em função da informação sintática e eurrítmica subjacente. Mas a verdade é que os segmentos se modificam mais drasticamente do que a mera modificação de parâmetros prosódicos levaria a crer: os valores dos formantes também se alteram ao hiperarticularmos uma sílaba em condição de fronteira prosódica forte. Regras que implementam esse tipo de interação prosódia-segmentos podem ser desenvolvidas para sistemas de síntese paramétricos ou articulatórios. Mas em princípio não para sistemas concatenativos, pois isso exigiria a gravação prévia de unidades concatenantes sob diversas condições prosódicas. Essa situação vem sendo revertida a partir dos anos 1990, através de um retorno à idéia de gravação de trechos de sinal acústico sob condições prosódicas variadas (Sagisaka 1988, Black & Campbell 1995 e Campbell & Black 1997), retorno esse estimulado pelo insucesso em obter síntese com elevada naturalidade usando a técnica PSOLA, referida acima. Esse modo de se fazer Síntese Concatenativa tem recebido o nome de *Corpus Synthesis* (Síntese via

Corpus). Para uma crítica ao mesmo, ver Barbosa (2001). Apesar do sucesso evidente da Síntese via Corpus (como no último Eurospeech 1999; vide Barbosa 1999a para uma avaliação), continua-se a usar a técnica PSOLA, como no recente projeto MBROLA (Dutoit, Pagel, Pierret, Bataille, Vrecken 1996). Essas considerações nos levam a pensar até que ponto os modernos sistemas TTS simulam o mecanismo de produção da fala.

Klatt (1976a) e Umeda (1976) vão apontar quais são e como devem funcionar os algoritmos para que os sistemas de síntese de fala simulem o inteiro mecanismo de produção de fala: "An additional possible motivation for creating a synthesis program is to define a functional model of human sentence generation. [...] a functional model can be extremely valuable in detecting gaps and inconsistencies in phonological and acoustic-phonetic characterizations of English." (Klatt 1976a, p. 391). A realização de um sistema que opera a partir da representação conceitual de uma mensagem será efetuada por Young e Fallside (1979) para uma aplicação específica, a recuperação de informação em um banco de dados.

3.2. Sistemas de síntese a partir do conceito (CTS)

O sistema de Young e Fallside (1979) é uma aplicação para recuperação de informação de um banco de dados sobre fornecimento de água. A informação é armazenada sob a forma de uma função contendo um operador (o verbo) e seus argumentos (sujeito, objetos, entre outros). A estrutura de superfície de uma sentença é construída a partir da gramática transformacional apresentada por Chomsky (1957). Utilizam um sintetizador concatenativo. A geração sonora é portanto efetuada a partir de informações de um componente semântico-sintático.

Esse tipo de sistema de síntese vem sendo usado com sucesso em Sistemas Automáticos de Diálogo (cf. por exemplo Bruce et al. 1995), no quadro do que veio a ser conhecido como Interação Homem-Máquina. O sistema de diálogo possui um subsistema de Reconhecimento de Fala. O operador solicita uma informação ao sistema, que a reconhece, representando-a sob a forma de uma estrutura lingüística abstrata, de natureza semântico-sintática. O uso de um OrtoFon, através da transformação dessa estrutura em texto escrito, além de ser completamente desnecessário, introduziria toda a série de ambigüidades contidas na relação escrita-fala. Por isso o subsistema de síntese usado nesses sistemas é sempre um CTS. Os sistemas de diálogo necessitam de

informação discursiva e pragmática em profundidade para poder interagir convenientemente com o operador humano.

Os sistemas CTS também são usados em Tradução Automática (cf. Hutchins & Somers 1992 para uma introdução à área). Nesses sistemas, o reconhecimento de fala na língua-fonte é seguido da construção de uma representação estrutural em uma Interlíngua ou já na língua-alvo, e do uso de um sistema CTS para a obtenção da fala sintética na língua-alvo (veja um exemplo de aplicação de tradução fala-fala entre o coreano e o japonês em Lee et al. 1995).

3.3. Primeiros sistemas de síntese em sueco, francês e português brasileiro

Os surgimentos do espectrógrafo e do computador digital nos Estados Unidos condicionaram o início da Síntese de Fala, sobretudo da síntese paramétrica e articulatória, nesse país. Os primeiros sistemas de síntese digitais são também americanos. Como vimos, no entanto, pesquisadores europeus e japoneses instalados nos Estados Unidos (como o francês Pierre Delattre e o japonês Osamu Fujimura), ou visitando instituições como o Bell Labs ou o MIT (como o sueco Gunnar Fant), ou ainda trabalhando em seus países com o inglês (como os japoneses do artigo de Matsui et al. 1968), contribuíram decisivamente para o desenvolvimento da área: não é possível desenvolvimento tecnológico sem conhecimento básico (os europeus se expressando em inglês, francês e alemão lideravam o conhecimento fonético na Europa no século XIX e fonético e fonológico no início do XX).

Os trabalhos com outras línguas que contribuíram de forma pioneira para a Síntese de Fala já foram apresentados. A multiplicação de publicações descrevendo sistemas de síntese em diversas línguas é um fato notório, estimulado pelas exigências das agências e indústrias financiadoras, sobretudo as companhias telefônicas. Um panorama dos sistemas nessas línguas exigiria um trabalho à parte. Destacamos aqui os primeiros trabalhos com o sueco, francês e português brasileiro (PB). A escolha da primeira língua se deve à importância da figura de Gunnar Fant, contando para isso com informação do Projeto Smithsoniano de História da Síntese de Fala, ainda em andamento (Maxey 2001). O francês é descrito pela contribuição singular dos pesquisadores franceses para a Síntese de Fala, bem como por possuirmos mais informações a respeito. Os primeiros trabalhos com o PB dispensam explicação.

O Instituto Real de Tecnologia (KTH, na sigla em sueco) de Estocolmo participou, na figura de Gunnar Fant, com o OVE I e seus aperfeiçoamentos no OVE II e III, da Síntese de Fala do inglês americano, durante a permanência de Fant no MIT. Ao que parece, sua primeira contribuição para o sueco parece ter sido em 1959 (Fant 1959). Quanto à síntese articulatória, somente com a tese de Liljencrants (1985) é que o KTH começa a explorar a área. O trabalho de Båvegård (1996) exemplifica a complexidade computacional da síntese articulatória, que constitui o estado-da-arte em Síntese de Fala no cenário internacional.

Em francês, são pioneiros os trabalhos de René Carré e colegas (1970), da *École Nationale Supérieure de Technologie* em síntese paramétrica. Em síntese concatenativa, a partir do trabalho pioneiro de pesquisadores do LIMSI em Paris (Leipp, Castellengo & Liénard 1968 e Leipp, Castellengo, Sapaly & Liénard 1968), desenvolveram-se trabalhos em outros institutos, como em Grenoble (Émerard 1977). Nessa cidade, o *Institut de la Communication Parlée* (ICP) foi pioneiro na pesquisa em Síntese Audiovisual (vide seção 4) através da contribuição sem precedentes de Christian Benoît (vide por exemplo Benoît, Mohamadi & Kandel 1994 e Le Goff, Guiard-Marigny, Benoît 1995). É também lugar de importantes desenvolvimentos em síntese articulatória (Bailly, Laboissière & Schwartz 1991). Em termos de modelos de geração automática da entoação os pioneiros são os trabalhos de Martin (1976) e Émerard (1977), a partir do quadro teórico e experimental fornecido por Dellattre (1966). O modelo prosódico de Bailly (1989), com elaborado tratamento de informação semântica, sintática e rítmica, merece nota não somente pela naturalidade que confere à síntese como também pela contribuição desse pesquisador do ICP à área de Síntese de Fala como um todo. Nesse modelo, a duração segue o modelo seminal de Klatt (1979). Também fundamentado nesse último é o modelo de geração automática de duração de Bartkova & Sorin (1987). O modelo prosódico de Barbosa e Bailly (1994), desenvolvido a partir do trabalho inicial de Campbell & Isard (1991), apresenta a vantagem da geração automática da pausa de forma integrada à geração de unidades do tamanho da sílaba, preservando a silabicidade e acentuação da frase francesa.

Em PB, o primeiro sistema de síntese é concatenativo (usando uma codificação por parâmetros LPC[19]), implementado por Egashira (1992) sob a supervisão de Fábio Violaro (Egashira & Violaro 1991), coordenador

do Laboratório de Processamento Digital de Fala (LPDF) da Faculdade de Engenharia Elétrica e Computação da Unicamp. Em conjunto com o Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE) do Instituto de Estudos da Linguagem (Unicamp), coordenado por Eleonora Albano e Edson Françozo, essa primeira equipe de engenheiros e lingüistas apresentou um projeto temático previsto em princípio para seis anos (1994 a 1999) à Fapesp denominado "Processamento de Texto e Sinal Acústico em Português Brasileiro: uma Interface Lingüística-Engenharia para a Ciência e Tecnologia da Fala", que inaugura e incentiva a pesquisa em Ciência e Tecnologia de Fala no Brasil.[20] A necessidade de um trabalho de maior envergadura, envolvendo profissionais das áreas concernidas, é resultado da constatação da atividade muito incipiente e dispersa em sistemas de síntese e reconhecimento de fala no Brasil (cf. o panorama de Violaro 1993). A implementação de um segundo sistema de síntese concatenativo do PB, batizado de Aiuruetê (Barbosa et al. 1999), com ambiente computacional implementado por Simões (1999) é um dos frutos desse trabalho comum. É resultado direto de resultados de pesquisa em Fonética Acústica, do desenvolvimento de um OrtoFon (Albano & Moreira 1996) e da montagem criteriosa de um inventário de polifones (Albano & Aquino 1997), a partir do princípio da demi-sílaba, todos eles realizados no LAFAPE, bem como da implementação, no LPDF, de técnicas como a PSOLA e o modelo híbrido de Violaro. Quanto aos modelos prosódicos, o sistema integrou recentemente o modelo de ritmo de Barbosa (1997). A entoação carece de modelos automáticos, mas os estudos fonético-acústicos de Madureira e colegas (1994, 1997, 1999) fornecem material suficiente para começarmos um.

O Aiuruetê foi demonstrado publicamente por mim, pela primeira vez em 4 de maio de 2001, na PUC-SP, por ocasião da abertura do XI InPLA, embora alguns exemplos em fita cassete tenham sido apresentados a um público de falantes do português brasileiro e europeu em Budapeste, por ocasião do Eurospeech'99 (vide apêndice e Barbosa 1999a), em setembro de 1999. O espectrograma de uma frase por ele produzida vem a seguir (fig. 8) e pode ser comparada com o exemplo com fala natural anterior (fig. 4). Os movimentos de formantes para as vogais e os componentes freqüenciais nos intervalos consonantais são muito semelhantes, mesmo que o locutor que forneceu sua voz ao Aiuruetê seja de dialeto (pernambucano) distinto daquele do locutor do exemplo com fala natural.

A síntese paramétrica em PB tem sido feita programando-se por tentativa e erro um sintetizador de Klatt (cf. Chiquito 1993, 1996). Ressente-se portanto da falta de transferência de conhecimento fonético-acústico lingüístico, que vem crescendo em uma taxa mais elevada apenas a partir do início dos anos 1990, com a montagem dos primeiros laboratórios de fonética acústica no Brasil (cf. Albano 1999).

Quanto à Síntese Articulatória, a exigência de equipamentos caros e de profissionais com formação na área ainda remete esse trabalho para uma etapa ulterior.

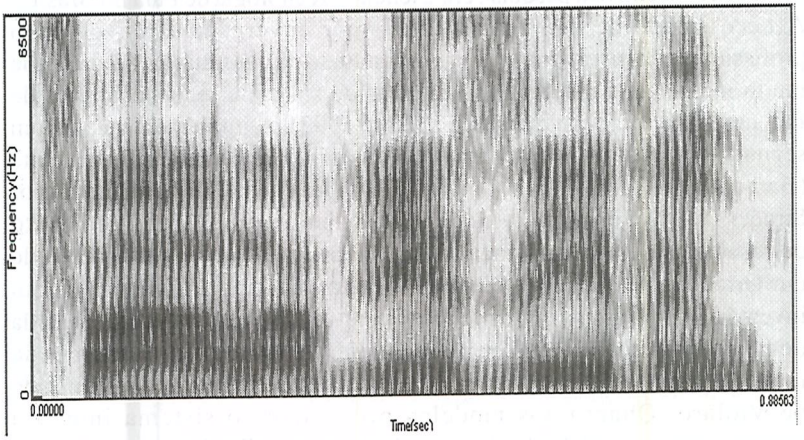

Figura 8: Espectrograma da frase "Fala visível", produzida pelo Aiuruetê. O enunciado sintético pode ser ouvido em <http://www.lafape.iel.unicamp.br/Docentes/Plinio/plinio_perfil.htm>

3.4. Usando máquinas falantes para testar componentes da gramática fônica: um exemplo

Como já demonstraram Mattingly (1971) e Klatt (1976a), podemos usar as máquinas falantes como um laboratório de testes para uma determinada teoria lingüística (desde que possa ser expressa de forma algorítmica), avaliando os componentes semântico, sintático (para produção da fala, nos sistemas CTS e para leitura, nos sistemas TTS) e fônico (segmental e prosódico). Dois exemplos da pesquisa com modelos de ritmo são apresentados.

A concepção teórica da estruturação rítmica do enunciado subjacente a esse modelo (justificada em PB por estudos fonético-acústicos como o

de Barbosa 1996) considera o papel primordial das unidades do tamanho da sílaba como um quadro para a especificação das durações dos segmentos, durações essas que seriam um subproduto da duração das unidades superiores. Por isso o modelo automático de ritmo que desenvolvemos para a síntese do PB (Barbosa 1997) gera em um primeiro momento a duração de uma unidade entre dois *onsets* de vogal consecutivos, isto é, unidade $VC(C_0)$, chamada de GIPC (de grupo inter-*perceptual-center*. Vide Barbosa (2000) para uma revisão sobre o isocronismo na fala que culminou com a noção de *perceptual-center*), a partir do aprendizado de *gestalten* rítmicas, para então distribuir essa duração entre os segmentos que a constituem. Para gerar a duração de um GIPC, o modelo leva em conta a natureza da vogal do mesmo, e não a das consoantes; sendo assim, os GIPCs [as] e [az] teriam a mesma duração, se inseridos no mesmo contexto lingüístico. Para a atribuição das durações dos segmentos dessas duas unidades, utiliza-se um modelo estatístico que atribui mais duração a segmentos mais longos em média. Assim, as durações das vogais [a] em [as] e [az] são distintas: visto que [s] tem duração em média maior que [z], sobra menos duração para ser atribuída ao [a] de [as]. Esse resultado final na atribuição das durações está de acordo com a pesquisa fonético-acústica lingüística: em PB, as vogais são mais curtas (longas) quando seguidas de segmentos não-vozeados (vozeados). Esse aspecto da influência contextual na duração segmental em PB é também partilhado por outras línguas como o francês e o inglês, mas é uma característica lingüística, pois em línguas como o checo e o polonês o efeito não se dá.

A atribuição de duração na síntese do PB para a implementação do acento se dá a partir de informação lingüística e de taxa de elocução. Fundamentados em pesquisa fonético-acústica em francês e PB e em propostas teóricas ligadas a sistemas dinâmicos (cf. Barbosa; 2001), consideramos o acento como a culminação de um movimento subjacente de acentuação que vai aumentando progressivamente a duração de unidades do tamanho da sílaba até o acento. Suponhamos agora que uma determinada proposta teórica, fundamentada na Fonologia Métrica considerasse a representação métrica do sintagma "Ele guarda" na frase "Ele guarda a sela do cavalo numa prateleira de uma antiga cela" como possuindo um peso 4 na sílaba "guar-". Que a tônica "e-" tivesse o peso 3 enquanto que as pós-tônicas "-le" e "-da" tivessem peso 1 (de fato a representação métrica habitual, de inspiração mattosiana, é assim). De nossa parte,

consideremos que pesos maiores determinassem valores maiores de duração para as sílabas.

A inspeção dos valores de duração dessas sílabas na frase pronunciada por um locutor (fig. 9) em três taxas de elocução revela uma situação mais complexa: à medida que a taxa de elocução aumenta (de lenta para normal e para rápida), a duração da sílaba na posição 2 (pós-tônica "-le") se torna maior que a da tônica "e-" (posição 1), invertendo o padrão esperado pela representação métrica. Isso decorre devido à interação entre um sistema lingüístico abstrato e um mecanismo de produção sujeito a leis da física (como a inércia). O padrão obtido abaixo é exatamente aquele obtido por um sistema de síntese do PB (cf. Barbosa et al. 1999) que se sirva dos pressupostos teóricos apresentados em nosso trabalho. Um modelo exclusivamente fundamentado na representação métrica não obteria frases com durações naturais.

Além disso, no caso do francês, efetuamos um teste de percepção (Barbosa & Bailly 1994) para avaliar dois padrões de acentuação: um obtido pelo aumento contínuo da duração dos GIPCs até a posição de acento contra um obtido pela maior duração da sílaba acentuada apenas. Os resultados comprovaram uma preferência dos sujeitos superior a 85% pelo primeiro padrão quando perguntados sobre qual das frases parece a mais natural.

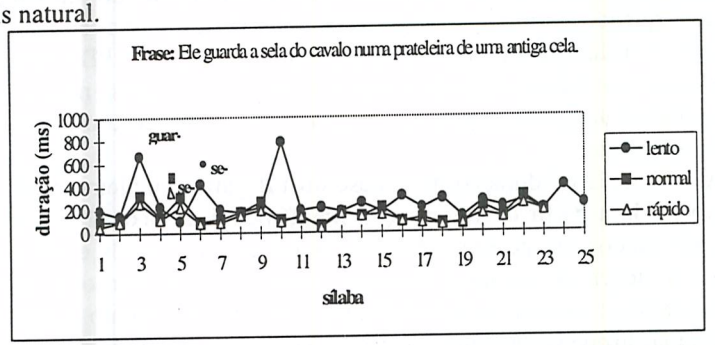

Figura 9: Contornos duracionais para a frase "Ele guarda a sela do cavalo numa prateleira de uma antiga cela."

A geração automática da estruturação rítmica do PB apresentada em Barbosa (1999b) e fundamentada em estudo fonético-lingüístico, complementa os exemplos acima como também defende uma integração entre ciência e tecnologia de fala. Como se percebe das considerações aqui apresentadas, os argumentos cartesianos de que máquinas não

possuem conhecimento de algo ou não respondem a partir de uma interação comunicativa (seção 1) não mais se aplicam às máquinas falantes de hoje. Os sistemas de síntese de fala que integram modelos lingüísticos e de produção complexos são uma resposta à primeira objeção. Os Sistemas Automáticos de Diálogo são uma resposta à segunda.

4. Conclusão: por um humanismo *éclairé*

A recente pesquisa em Síntese de Fala abriu seus horizontes gestuais e incluiu, com a Síntese Audiovisual, os movimentos faciais (principalmente boca, mas também olhos e parte superior da bochecha), produzidos em sincronismo com o som. O número e a qualidade das publicações em Síntese Audiovisual tem crescido, bem como a área de pesquisa tem expandido ainda mais seus horizontes, pois as cabeças falantes se transformaram em figuras animadas de corpo inteiro (ver, por exemplo, o impressionante site do grupo de Síntese Multimodal do KTH, disponível em <http://www.speech.kth.se/multimodal>).

A emoção também é alvo da pesquisa em síntese (cf. Murray & Arnott 1993 para uma revisão e os trabalhos de Murray & Arnott 1996 e Greasley et al. 1995). Reproduzir as modificações nos parâmetros fonético-acústicos advindas da pronúncia de uma frase sob um determinado estado emocional é o objetivo da pesquisa em emoção sintética. Ela tem empregado atores para as investigações laboratoriais. A pesquisa experimental com frases obtidas a partir de um controle maior dos fatores cognitivos, lingüísticos e extralingüísticos em jogo na expressão de uma determinada emoção real encontra evidentemente o sério problema ético de dispor uma pessoa a um determinado estado emocional.

Quando a pesquisa em Síntese de Fala conseguir obter fala sintética dotada de um certo colorido emocional com naturalidade indistinta da humana, aquilo que distinguirá a fala natural da sintética dirá respeito ao especificamente humano em nós e encontrará incontornavelmente a noção de sujeito e a espinhosa questão da consciência. É evidente que a pesquisa em primatologia com antropóides, bem como a paleontologia, pesquisando como se deu a hominização, bem como o aporte filosófico relacionado a essas questões desempenharão um papel de extrema relevância no tratamento desse tema.

Acreditamos termos exposto que, em um certo sentido, as máquinas falantes são instrumentos lingüísticos da mesma forma que dicionários e gramáticas: ferramentas para preencher lacunas de nosso conhecimento

imperfeito no que diz respeito às línguas (é claro que dicionários e gramáticas desempenham e desempenharam um papel muito mais importante do que o papel restrito que aponto aqui, como no processo de gramaticalização e afirmação das línguas vernáculas frente ao latim, mas não é esse o aspecto para o qual quero chamar a atenção). As lacunas que a máquina falante pode ajudar a preencher dizem respeito à abrangência das teorias lingüísticas e dos modelos de produção de fala como formas de explicação do falar.

Acreditamos também termos mostrado que a construção de máquinas falantes envolve uma atividade que é necessariamente trans- inter- e pluridisciplinar, necessitando da colaboração de profissionais de diferentes áreas e com formação inter- e pluridisciplinar que se disponham a entender os problemas e métodos das disciplinas outras, aquelas em que não se formaram primeiramente. Para começar essa aventura do conhecimento, é preciso ter o tipo de atitude do humanista renascentista universal: a tolerância em relação a métodos e técnicas de disciplinas mais afeitas às Ciências Humanas bem como àquelas mais afeitas às Ciências Naturais. Um exemplo desse tipo de atitude é dado por Albano (2001), que aborda no seu recente livro a relação entre Fonética e Fonologia propondo não mais uma separação estanque, pregada como necessária (os avanços em Fonologia talvez tenham demonstrado que o fora) pela Escola de Praga, mas uma integração que implica uma reavaliação da relação entre simbolismos do discreto e simbolismos do contínuo.

Se num certo sentido a experimentação é parte importante da pesquisa em Síntese de Fala, aproximando-a de um certo neo- ou pós-positivismo, essa é apenas a etapa instrumental, visto que a reflexão hipotético-dedutiva e a abertura a novos pressupostos teóricos constituem justamente aquilo que faz avançar o conhecimento teórico nas ciências da fala e, de forma mais abrangente, nas ciências da cognição. Além disso, a abertura interdisciplinar que o trabalho em Síntese exige é completamente contrária à concepção positivista de Comte, que pregava uma separação estrita entre as disciplinas.

A Síntese de Fala é portanto um lugar exemplar em que disciplinas podem e devem se reconciliar, num certo sentido como a tradução é o lugar de reconciliação entre as línguas para Benjamin (1971). A reconciliação busca atingir uma meta comum, qual seja, entendermos melhor o que e quem somos.

Agradecimentos

Ao CNPq (Bolsa de Produtividade em Pesquisa número 350382/98-0, vinculada ao projeto de número 524110/96-4), e à FAPESP, pelo Auxílio-Pesquisa *Jovem Pesquisador em Centro Emergente* número 95/09708-6. Esse trabalho também está associado ao Projeto Temático "Integrando Parâmetros Contínuos e Discretos em Modelos do Conhecimento Fônico e Lexical", número 01/00136-2, ainda em julgamento. Agradecemos a Sandra Madureira pela leitura e sugestões.

APÊNDICE: Tateamento das Ciências da Fala pelas vias das reuniões científicas

Embora encontre suas raízes junto às primeiras incursões do homem sobre os mecanismos de produção (no Ocidente, destacam-se os precisos estudos anatômicos de Leonardo da Vinci) e percepção da fala, a área de Ciência ou Ciências da Fala (em inglês e francês os termos equivalentes são *speech science* e *sciences de la parole*), vem se delineando mais precisamente a partir da articulação pioneira entre a Fonética, Acústica e a Fisiologia da Fala no primeiro congresso de Ciências Fonéticas em 1932, que contou com a participação da jovem Fonologia. Hoje se encontra no trabalho coletivo e conjunto de lingüistas, foneticistas,[21] acusticistas, fisiologistas, psicólogos experimentais, engenheiros elétrico-eletrônicos e cientistas da computação, tendo ganhado força a partir da colaboração pioneira de Roman Jakobson, Gunnar Fant e Morris Halle no *Preliminaries to Speech Analysis*[22] (1952), e encontrado sua via definitiva através do uso incontornável da espectrografia (Koenig, Dunn & Lacy 1946) e da computação digital (a partir dos anos 1960-1970), como meios para testar e modelar o conhecimento multifacetado da fala (cf. Boë 1997).

Atualmente, três grandes congressos internacionais, alguns congressos nacionais de grande público (no Japão e na Austrália) e diversos congressos-satélite e *workshops* (em Síntese de Fala e Síntese Audio-visual) reúnem os profissionais acima, bem como um novo profissional com formação interdisciplinar, o *cientista da fala* (*speech scientist, chercheur en parole*), para discutir suas pesquisas e os rumos científicos e tecnológicos da área. Dos três congressos internacionais, os dois primeiros são bianuais e patrocinados pela Associação Internacional de Comunicação Falada (*ISCA*, da sigla em inglês), antiga Associação Européia de Comunicação Falada (*ESCA*): os Eurospeech, acrônimo de *European Conference on Speech Communication and Technology*, desde 1989, e

os ICSLP, sigla de *International Conference on Spoken Language Processing*, desde 1990.[23] A terceira série de congressos se dá a cada quatro anos desde 1932 (com interrupções durante a Segunda Guerra), os ICPhS ou *International Congress of Phonetic Sciences*.

É importante salientar que o primeiro *workshop* internacional sobre Síntese de Fala se deu em 25 de maio de 1990, na cidade alpina de Autrans, França, organizado por Christian Benoît e Gérard Bailly, pesquisadores do ICP (os dois seguintes foram organizados em New Paltz, Estados Unidos, em 1994, e Mokonk, próximo a Sidney, em 1998. O mais recente foi no Atholl Palace Hotel Perthshire, próximo a Glasgow e Edinburgo, de 29 de agosto a 1º de setembro de 2001).

Quanto à Síntese Audiovisual, as reuniões científicas *Audio Visual Speech Processing* se iniciaram em Bonas (França), em 1995, e continuaram nos encontros em Rodes, em 1997, em Sidney, em 1998, e finalmente em Santa Cruz (Estados Unidos), em 1999.

Notas

[1] Cf. Calvin & Bickerton (2000) para uma discussão aprofundada sobre o que é a linguagem humana e como ela está armazenada em nossas estruturas neuronais, além da diferença entre essa estrutura e a de animais próximos como o bonobo, ou chimpanzé anão.

[2] A pergunta associada a essas, qual seja, como compreendemos o que dizem os falantes de nossa língua faz parte da pesquisa em Reconhecimento e Compreensão de Fala.

[3] O autor deu uma contribuição abrangente à área de Fonética a partir de um trabalho de vinte anos em Síntese de Fala, somente interrompido por sua morte precoce, em dezembro de 1988.

[4] Bem distante dos preconceitos presentes na história da educação dos surdos (cf. Saks 1999).

[5] Preferi apresentar aqui o texto da edição de 1824 (Descartes 1824), tal como se encontra no endereço *internet* da Association des Bibliophiles Universels (ABU), dado que a ortografia original (Descartes 1637) não apresenta leitura evidente para quem não conhece as relações grafema-fonema do francês escrito do século XVII. Na versão original, vê-se que Descartes usou itálico na palavra "automates". Da cópia da ABU tirei os colchetes indicando a paginação, para acrescentar os meus, assinalando os trechos que focalizo neste estudo. A paginação do texto é dada logo após a citação.

[6] Beaune define o autômato como "une machine porteuse du principe interne de son mouvement qui, en conséquence, garde inscrits en ses composants matériels ou ses actions, l'illusion, le rêve ou la feinte de la vie" (1980, p. 7).

[7] Para essa apresentação o flautista havia sido fantasiado como um fauno.

[8] Outras realizações, como os cinco tubos que produziam sons de cinco vogais, do russo Kratzenstein, que ganhou um concurso da Academia Real de Saint Petersburg em 1781, e a cabeça falante do abade francês Mical, apresentada em 1778, ambos referidos pelo próprio Kempelen (cf. Pompino-Marschall 1991, p.183) nem de longe se comparam à máquina do barão húngaro, louvada pela produção das consoantes.

[9] É também o autor de uma célebre fraude: um autômato vestido de turco que supostamente jogava xadrez. Após a descoberta de que havia um homem escondido em uma porção oca da

mesa que sustentava o tabuleiro, deve ter sido difícil acreditar em sua máquina falante (Dudley & Tarnoczy 1950). O pato "digeridor" de Vaucanson também não digeria propriamente, mas havia um mecanismo de troca da comida que ingeria por bolinhas de pão pintadas (Doyon & Liaigre 1967).

[10] Vide Pompino-Marschall (1991) para detalhes sobre testemunhos dessa turnê.

[11] Penso que para uma máquina falante nada mais é necessário do que um pulmão, uma glote e uma boca. [Tradução minha.]

[12] Em um anúncio sobre uma apresentação da máquina em Londres, em 1783, se lê: "Besides this Automaton [o jogador de xadrez], there is another Machine to be seen, not less wonderful and new. Such a Contrivance was attempted in several Ages, without any Success, and its Performance reputed quite impossible. This Machine now is brought into Execution, consisting in a small Organ, *which having the Voice of a Child between five and six Years of Age*, speaks a great many Words very distinctly, when played by its inventor." [*itálico* meu] (apud Pompino-Marschall 1991, p. 200). O anúncio deve ser de autoria do próprio Kempelen.

[13] A analogia elétrica da variação de pressão sonora com a variação de uma grandeza elétrica como a corrente parece ter sido feita pela primeira vez por Alexander Graham Bell (cf. Mattingly 1999).

[14] Notar aqui a diferença clara entre a observação dos espectrogramas, produzidos por uma máquina, através dos quais se procura conjugar elementos discretizáveis e não-discretizáveis, integrados em um "padrão" de descrição, e o alfabeto de Melville Bell, estipulando um conjunto de símbolos discretos para a representação do som de fala. É claro que a diferença nos tratamentos está relacionada com o avanço tecnológico que permitiu o contraste de conceitos de produção de fala e aqueles de percepção de fala. Albano (no prelo) discute o assunto com muita propriedade.

[15] O dispositivo de Schott (1948), do Bell Labs, também é um *pattern playback* e é anterior àquele do Haskins Labs.

[16] É um teste equivalente ao Teste de Turing, para as Máquinas Pensantes.

[17] E por isso considerada marginal por aqueles que propuseram a Síntese Paramétrica (embora também tenham feito tentativas de concatenação de unidades para a síntese).

[18] Que evoluirão para regras que conduzirão ao Infovox SA-101, 1983, sistema sueco multilíngüe de fala (apud Klatt 1987).

[19] De *Linear Predictive Coding*, técnica de redução de dados que permite codificar o sinal de fala separando a contribuição da fonte sonora (pregas vocais) daquela do filtro (trato vocal). Um gerador de ruído ou de padrão periódico simulando a fonte sonora e alguns poucos valores (os coeficientes do filtro LPC) representando o aporte dos formantes, exatamente como no Voder da figura 5, permite a recuperação completa do sinal.

[20] O Laboratório de Processamento de Sinais, coordenado por José Chiquito, participou em um primeiro momento da equipe através do trabalho com um sintetizador paramétrico.

[21] Que se reconhecem ou não na chamada Fonética Lingüística, que tem como figura emblemática Peter Ladefoged. Ver particularmente o livro de Fromkin 1985, a ele dedicado, e Ladefoged 1971.

[22] Um marco para o desenvolvimento da Fonologia moderna, com a proposta da Teoria dos Traços Distintivos, inspirada diretamente da Teoria da Informação de Shannon (1951) apud Jakobson et al. (1952), na área da Engenharia de Telecomunicações.

[23] As duas séries de congressos receberam recentemente a co-denominação, após a mudança da ESCA para ISCA, de *Interspeech Events*.

Referências Bibliográficas

ALBANO, Eleonora C. O português brasileiro e as controvérsias da Fonética atual: pelo aperfeiçoamento da Fonética Articulatória. **D.E.L.T.A. n.** 15 (Número Especial), 1999, p. 23-51.

_____. **O gesto e suas bordas**: esboço de fonologia acústico-articulatória do português brasileiro. Campinas: Mercado de Letras, 2001.

_____. A pulsação sob a letra: pela quebra de um silêncio histórico no estudo de som de fala. **Cadernos de Estudos Lingüísticos.** (No prelo).

_____; AQUINO, Patrícia A. Linguistic criteria for building and recording units for concatenative speech synthesis in Brazilian Portuguese. **Proceedings of Eurospeech'97**. Rhodes, Grécia, 1997,v.2, p. 725-728.

_____; BARBOSA, P.; GAMA-ROSSI, A.; MADUREIRA, S., SILVA, A. A interface fonética-fonologia e a interação prosódia-segmentos. **Estudos Lingüísticos XXVII** - (Anais do XLV Seminário do Grupo de Estudos Lingüísticos do Estado de São Paulo-GEL'97). 1998, p. 135-143.

_____; MOREIRA, A. A. Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese. **Proceedings of the International Conference on Spoken Language Processing**. V. 3, p. 1708-1711, out. 1996.

ALLEN, J. Speech synthesis from unrestricted text. **Speech Synthesis: Benchmark Papers in Acoustics**. FLANAGAN, James L.; RABINER, L. R. (Eds.). Stroudsburg, Pennsylvania: Dowden, Hutchinson & Ross, Inc., 1973.

_____; HUNICUTT, S.; CARLSON, Rolf; GRANSTRÖM, Bjorn. MITalk-79: The MIT Text-to-Speech System. **J. Acoust. Soc. Am.**, 1979, Suppl. 1 65, S130.

_____; HUNICUTT, S.; KLATT, Dennis H. **From text to speech: the MITalk system**. Cambridge, Mass. Cambridge University Press, 1987.

ARMSTRONG, L. E.; WARD, I. C. **A Handbook of English Intonation**. 2 ed. Cambridge, Inglaterra: Cambridge University Press, 1931.

AUROUX, S. (Dir.) Les Notion Philosophiques. JACAOB, André (Dir.) **Encyclopédie Philosophique Universelle**. Paris: Presses Universitaires de France, 1990.

BÅVEGÅRD, M. Towards an articulatory speech synthesiser: model development and simulations. **Speech, Music and Hearing Quaterly Progress and Status Report** 1. 1996, p.1-15.

BAILLY, Gérard. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. **Speech Communication** 8, 1989, p.137-146.

_____. LABOISSIÈRE, Rafael, SCHWARTZ, Jean-Luc. Formant trajectories as audible gestures: an alternative for speech synthesis. **Journal of Phonetics** 19(1), 1991, p. 9-23.

BALPE, Claudette. Jacques Vaucanson, mécanicien et monteur d'automates. **La Revue**. n· 20. Musée des arts et métiers. Set. 1997. Disponível em: <http://bose.cnam.fr/museum/revue/ref/r20a05.html>. Acesso em 4 jun. 2001.

BARBOSA, Plínio A. **Caractérisation et génération automatique de la structuration rythmique du français**. 1994. Tese inédita Doutorado. Grenoble, França: INPG/ICP.

_____. At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration: emphasis on segmental duration generation. **Cadernos de Estudos Lingüísticos**. n. 31, 1996. p. 3353.

_____. A model of segment (and pause) duration generation for Brazilian Portuguese text-to-speech synthesis. **Proceedings of the Fifth European Conference on Speech Communication and Technology**. Rhodes, Grécia, v. 5, p. 2655-2658, set. 1997.

_____. **Relatório Científico**. 1999a. (Referente a participação no *Sixth European Conference on Speech Communication and Technology* em Budapeste, possível via auxílio recebido via solicitação nº 768/99 ao FAEP/Unicamp).

_____. Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia de fala. SCARPA, Ester (Org.). **Estudos de Prosódia**. Campinas: Editora da Unicamp, 1999, p. 21-52.

_____. 'Syllable-timing in Brazilian Portuguese': uma crítica a Roy Major. **D.E.L.T.A.**, 16 (2), 2000, p. 369-402.

_____. Generating Duration from a Cognitively Plausible Model of Rhythm Production. **Proceedings of the Seventh European Conference on Speech Communication and Technology (Eurospeech 2001)**. Aalborg, Dinamarca. Set. 2001.

_____; BAILLY, G. Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, 15 (1-2), 1994, p. 127-137.

_____; VIOLARO, Fábio; ALBANO, Eleonora; SIMÕES, Flávio; AQUINO; Patrícia; MADUREIRA, Sandra; FRANÇOZO, Edson. Aiuruetê: a High-Quality Concatenative Text-to-Speech System for Brazilian Portuguese with Demisyllabic Analysis-Based Units and a Hierarchical Model of Rhythm Production. **Proceedings of the Sixth European Conference on Speech Communication and Technology**, Budapeste, Hungria. V. 5, p. 2059-2062, set. 1999.

BARTKOVA, K.; SORIN, Christel. A model of segment duration for speech synthesis in French. **Speech Communication** n. 6, 1987, p. 245-260.

BEAUNE, Jean-Claude. **L'Automate et ses mobiles**. Paris: Flammarion, 1980.

BELL, Alexander Melville. **Visible speech: universal alphabetics or self-interpreting physiological letters for the writing of all languages in one alphabet**. Nova York: Trübner, 1867.

BENJAMIN, Walter. La Tâche du traducteur. **Mythes et violence**. Tradução: Maurice de Gandillac. Paris: Denoël, 1971.

BENOÎT, Chistian; MOHAMADI, T.; KANDEL, Sonia. Audio-visual intelligbility of French speech in noise. **Journal of Speech and Hearing Research** n. 37, 1994, p. 1195-1203.

BLACK, Alan; CAMPBELL, Nick. Optimising selection of units from speech databases for concatenative synthesis. **Proceedings of the Fourth European Conference on Speech Communication and Technology (Eurospeech 1995)**, v. 1, 1995, p. 581-584.

BOË, Louis-Jean. Sciences phonétiques et relations forme/substance: un siècle de ruptures, négociations et réorganisations. **Histoire Épistémologie Langage.** SHESL, PUV. n. 19/1, 1997, p. 5-41.

BÖEFFARD, Olivier; VIOLARO, Fábio. Using a hybrid model in a text-to-speech system to enlarge prosodic modifications. **International Conference on Spoken Language Processing** (ICSLP '94), Yokohama, Japão, 1994, p. 727-730.

BROWMAN, Cathy P. (1980) Rules for demisyllable synthesis using Lingua, a Language Interpreter. **Proceedings Int. Conf. Acoust. Speech Signal Process.** 1994, p. 561-564.

BRUCE, Gösta; GRANSTRÖM, Björn; FILIPSON, Marcus; GUSTAFSON, Kjell; HORNE, Merle; HOUSE, David; LASTOW, Birgitta; TOUATI,

Paul. Speech synthesis in spoken dialogue research. **Proceedings of the Fourth European Conference on Speech Communication and Technology (Eurospeech 1995)**. V. 2, 1995, p.1169-1172.

CALVIN, W., BICKERTON, D. **Lingua ex Machina**: Reconciling Darwin and Chomsky with the human brain. Cambridge, Mass.: The MIT Press, 2000.

CAMPBELL, Nick W.; ISARD, S. D. Segment durations in a syllable frame. **Journal of Phonetics**, n. 19, 1991, p. 37-47.

_____.; BLACK, Alan. Prosody and the Selection of Source units for Concatenative Synthesis. In VAN SANTEN, J.P.H.; SPROAT, R.W. OLIVE, J.P. and Hirschberg, J. (Eds.). **Progress in Speech Synthesis**. New York: Springer-Verlag, 1997, p. 279-292.

CARLSON, Rolf. Duration models in use. **Proceedings of the XII International Congress of Phonetic Sciences**. V.1, 1991, p. 243-246.

_____; GRANSTRÖM, Björn. A phonetically-oriented programming language for rule description of speech. **Speech Communication.** n. 2. 1975, p. 245-253.

CARRÉ, René; BEAUVIALA, Jean-Pierre; PAILLÉ, Jean. Filters for Formant synthesizers. **IEEE Trans. on Audio and Electro**. AU-18 (3), 1970, p. 300-303.

CATER, John P. **Electronically Speaking: Computer Speech Generation**. Howard M. Sams & Co, 1983.

CHIBA, T.; KAJIYAMA, M. **The vowel – its nature and structure**. Tokyo: Phonetic Society of Japan, 1941.

CHIQUITO, J. G. Síntese de Fala usando o sintetizador de formantes de Klatt. **XI Simpósio Brasileiro de Telecomunicações**. 1993.

_____. Interface entre processamento de texto e de sinal para a síntese de fala por regras. **Anais do VII Simpósio Brasileiro de micro-ondas e optoeletrônica e XIV Simpósio Brasileiro de Telecomunicações**, v.1, 1996, p. 355-360.

CHOMSKY, Noam. **Syntactic Structures**. Haia: Mouton, 1957.

_____; HALLE, Morris. **The Sound Pattern of English**. Nova York: Harper and Row, 1968.

COKER, Cecil H.; UMEDA, Noriko; BROWMAN, Cathy P. Automatic synthesis from ordinary English text. **IEEE Trans. Audio Electroacoust**. AU-21, 1973, p. 293-297.

COOPER, Franklin S.; DELLATTRE, Pierre C.; LIBERMAN, Alvin M.; BORST, John M; GERSTMAN, Louis J. Some experiments on the perception of synthetic speech sounds. **J. Acoust. Soc. Am.** n. 24., 1952, p. 597-606.

_____; LIBERMAN, Alvin M.; BORST, John M., GERSTMAN. "The interconversion of audible and visible patterns as a basis for research in the perception of speech". **Proc. Natl. Acad. Sci.**, 1951, n. 37, p. 318-325.

DAHAN-DALMEDICO, Amy; PEIFFER, Jeanne. **Une histoire des mathématiques:** routes et dédaless. Paris: Éditions du Seuil, 1986.

DELATTRE, Pierre. Les dix intonations de base du français. **French Review** 40, Illinois: American Association of Teachers of French, 1966, p. 1-14.

_____; LIBERMAN, Alvin M.; COOPER, Franklin S. Acoustic loci and transitional cues for consonants. **J. Acoust. Soc. Am.** n. 27. 1955, p. 769-773.

_____; LIBERMAN, Alvin M.; COOPER, Franklin S.; GERSTMAN, Louis J. An experimental study of the acoustic determinants of vowel color; observations on One- and Two-Formant vowels synthesized from spectrographic patterns. **Word**, n. 8., 1952, p. 195-210.

DESCARTES, René [1637] **Discours de la méthode:** pour bien conduire sa raison, et chercher la vérité dans les sciences. (Corpus des œuvres de philosophie en langue française.) Paris: Fayard, 1987.

_____. ([1824] s.d.) **Discours de la méthode:** pour bien conduire sa raison, et chercher la vérité dans les sciences. COUSIN, Victor (Ed.) **Œuvres de Descartes**. V. 1, Paris. Disponível em: < http://abu.cnam.fr/ >. Acesso em: 6 jun. 2001. Copista: CUBAUD, Pierre.

DIXON, N. R.; MAXEY, H. D. Terminal analog synthesis of continuous speech using the diphone method of segment assembly. **IEEE Trans. Audio Electroacoust**. AU-16, 1968, p. 40-50.

DOYON, André; LIAIGRE, Lucien. **Jacques Vaucanson, mécanicien de génie**. Paris: Presses Universitaires de France, 1967.

_____; LIAIGRE, Lucien. Automate. **Encyclopedia Universalis**. V.2, 1985, p.1172-1177.

DUDLEY, Homer. The Vocoder. **Bell Laboratories Record**, n. 18, 1930, p.122-126.

_____; RIESZ, R. R.; WATKINS, S. S. A. A synthetic speaker. **Journal of the Franklin Institute**, n. 227. 1939, p. 739-764.

_____; TARNOCZY, T. H. The Speaking Machine of Wolfgang von Kempelen. **J. Acoust. Soc. Am.** 22 (2), 1950, p. 151-166.

_____. Fundamentals of Speech Synthesis. **J. Audio Engr. Soc.** 3. 1955, p. 170-185.

DUNN, H. K. The calculation of vowel resonance, and an electrical vocal tract. **J. Acoust. Soc. Am.**, n. 22. 1950, p. 740-753.

DUTOÎT, T.; PAGEL, V.; PIERRET, N.; BATAILLE, F.; VRECKEN, O. van der. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. **Proceedings of the International Conference on Spoken Language Processing.** V. 3, 1966, p. 1393-1396.

EGASHIRA, Francisco. **Síntese de Voz a partir de texto para a língua portuguesa.** Dissertação (Mestrado). Campinas: Universidade Estadual de Campinas, 1992.

_____; VIOLARO, Fábio. Síntese de Voz a partir de texto para a língua portuguesa. **Simpósio Brasileiro de Telecomunicações**. São Paulo: EDUSP, 1991.

ÉMERARD, Françoise. **Synthèse par diphones et traiement de la prosodie**. Tese (Doutorado) [3e cycle]. Université de Grenoble, França, 1977.

FANT, Gunnar. Speech Communication Research. **Ingl. Vetenskaps Akad. Stockholm**. Suécia, 24, 1953, p. 331-337.

_____. Acoustic analysis and synthesis of speech with applications to Swedish. **Ericsson Technics** 1, 1959.

_____. **Acoustic theory of speech production**. Haia: Mouton. 1960.

FLANAGAN, James L. **Speech analysis, synthesis and perception**. Berlim: Springer-Verlag, 1965.

_____. Voices of Men and Machines. **J. Acoust. Soc. Am**. 51, 1972, p. 1375-1387.

_____; RABINER, L. R. (Eds.). **Speech synthesis**. Stroudsburg: Dowden, Hutchinson, and Ross, Inc. (Benchmark Papers in Acoustics), 1973.

FROMKIN, Victoria (Ed.). **Phonetic Linguistics**: Essays in honor of Peter Ladefoged. New York: Academic Press, 1985.

FRY, Dennis B.; ABRAMSON, Arthur S.; EIMAS, Peter D.; LIBERMAN, Alvin M. The identification and discrimination of synthetic vowels. **Language and Speech**. n. 5, 1962, p. 171-189.

FUJIMURA, Osamu; LOVINS, J. Syllables as Concatenative Phonetic Elements. **Syllables and Segments**, BELL, A.; HOOPER, J. B. (Eds.). NewYork: North Holland, 1978, p.107-120.

FUJISAKI, Hiroya; NAGASHIMA, S. Synthesis of pitch contours of connected speech. **Annual Rept. Engr. Res. Inst.** 28, Univ. de Tóquio, Tóquio, 1969, p. 53-60.

GREASLEY, Peter; SETTER, Jane; WATERMAN, Mitch; SHERRARD, Carol; ROACH, Peter; ARNFIELD, Simon; HORTON, David Representation of prosodic and emotional features in a spoken language database. **Proceedings of the XIIIth International Congress of Phonetic Sciences.** V. 1, 1995, p. 242-245.

HARRIS, Cyril M. A Study of the Building Blocks in Speech. **J. Acoust. Soc. Am.** 25, 1953, p. 962-969.

HELMHOLTZ, H. (1877) **On the sensations of tone.** 4ª ed. New York: New York Dover. 1954.

HOLMES, J. N. The influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer. **IEEE Trans. Audio Electroacoust.** AU-21, 1973, p. 298-305.

_____. MATTINGLY, Ignatius; SHEARME, J. N. Speech Synthesis by Rule. **Language and Speech** n. 7. 1964, p. 127-143.

HUTCHINS, W. J.; SOMERS, H. L. **An Introduction to Machine Translation.** New York: Academic Press, 1992.

JAKOBSON, Roman; FANT, Gunnar; HALLE, Morris [1952]. **Preliminaries to Speech Analysis.** 11ª ed. Cambridge, EUA: The MIT Press, 1976.

JOOS, M. Acoustic Phonetics. **Language Monograph** n. 23. Baltimore, 1948.

KEMPELEN, Wolfgang von [1791]. **Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Machine.** Degen, J. V. (Ed.). Vienna, 1970.

KLATT, Dennis H. Synthesis of stop consonants in initial position. **J. Acoust. Soc. Am.** Suppl. 1 47, S93, 1970.

_____. Durational characteristics of prestessed word-initial consonant clusters in English. **Research Laboratory of Electronics QPR 108.** Cambridge, MA: MIT Press, 1973, p. 253-260.

_____. Structure of a phonological rule component for a synthesis-by-rule program. **IEEE Transactions on Acoustic, Speech, and Signal Processing.** V. 24 (5), 1976a, p. 391-398.

_____. Linguistic uses of segmental duration in English: acoustic and perceptual evidence. **J. Acoust. Soc. Am.** 59, 1976a, p. 1208-1221.

_____. Synthesis by rule of segmental durations in English sentences. LINDBLOM, Bjorn, ÖHMAN, S. (Eds.). **Frontiers of Speech Communication Research**. New York: Academic Press. 1979, p. 287-300.

_____. Software for a cascade/parallel formant synthesizer. **J. Acoust. Soc. Am.** 67. 1980, p. 971-995.

_____. A text-to-speech conversion system. **Proc. AFIPS Office Automation Conference**. 1981, p. 51-61.

_____. The Klattalk text-to-speech conversion system. **Proceedings of the Int. Conf. Acoust. Speech Signal Process**. 1982, p. 1589-1592.

_____. Review of text-to-speech conversion for English. **J. Acoust. Soc. Am.** 82 (3). 1987 p. 737-793.

KOENIG, W.; DUNN, H. K.; LACY, L. Y. The Sound Spectrography. **J. Acoust. Soc. Am**. V. 18, 1946, p. 21-32.

KOHLER, Klaus J. Invariance and variability in speech timing: from utterance to segment in German. PERKELL, J.; KLATT, D. H. (Eds.). **Invariance and variability in speech processes**. Erlbaum Hillsdale, 1986, p. 268-298.

KÖSTER, J.-P. **Historische Entwicklung von Syntheseapparaten**. Hamburgo: Helmut Buske Verlag Hamburg, 1973.

LADEFOGED, Peter. **Preliminaries to Linguistic Phonetics**. Chicago: University of Chicago Press, 1971.

LA METTRIE, Julien Offroy de [1748]. **L'Homme-machine**. Paris: Mille et une nuits. 2000.

_____. [1751]. **Oeuvres philosophiques**. 2 v. (Corpus des œuvres de philosophie de langue française, Paris: Fayard). 1987.

LAWRENCE, Walter. The synthesis of speech from signals which have a low information rate. JACKSON, W.; (Ed.). **Communication Theory**. Londres: Butterworths. 1953, p. 460-469.

LEE, F. F. Reading Machine: from text to speech. **IEEE Trans. Audio Electroacoust**. AU-17. 1969, p. 275-282.

LEE, Young-Jik; KIM, Young-Sum; LEE, Jung-Chul; RYOO, Joon-Hyung; YANG, Jae-Woo. Korean-Japanese Speech translation system for hotel reservation Korean Front Desk side. **Proc. of the Fourth European Conference on Speech Communication and Technology (Eurospeech 1995)**. V. 2, 1995, p. 1197-2000.

LE GOFF, Bertrand; GUIARD-MARIGNY, Thierry; BENOÎT, Christian Read my lips... and may jaw! How intelligible are the components of a speaker's face? **Proceedings of the Fourth European Conference on Speech Communication and Technology (Eurospeech 1995)**. V. 1, 1995, p. 291-298.

LEHISTE, Ilse **Suprasegmentals**. Cambridge, MA: MIT Press, 1970.

LEIPP, E.; CASTELLENGO, M.; LIÉNARD, J. S. La Synthèse de la parole à partir de digrammes phonétiques. **Fifth International Congress on Acoustics**. Tokyo, 1968.

_____; SAPALY, J.; LIÉNARD, J. S. Structure physique et contenu sémantique de la parole. **Revue d'Acoustique** 3-4. 1968.

LIÉNARD, J. S.; TEIL, D.; CHOPPY, C.; RENARD, G.; SAPALY, J. Diphone synthesis of French: vocal response and automatic prosody from the text. **Proceedings Int. Conf. Speech Signal Process**. ICASSP-77, 1977, p. 560-563.

LIBERMAN, A. M.; INGEMANN, F.; LISKER, L.; DELATTRE, P. C.; COOPER, F. S. Minimal rules for synthesizing speech. **J. Acoust. Soc. Am.** 31, 1959, p. 1490-1499.

LILJENCRANTS, J. **Speech synthesis with a reflection-type line analog**. Tese (Doutorado, inédita). Estocolmo Instituto Real de Tecnologia (KTH), 1985.

MADUREIRA, S. Pitch patterns in Brazilian Portuguese: an acoustic-phonetic analysis. **Proceedings of the Fifth Australian International Conference on Speech Science and Technology**. Perth, Austrália, v.1, 1994, p. 156-161.

_____; BARBOSA, P.A.; FONTES, M.; SPINA, D.; CRISPIM, K. Post-stressed syllables in Brazilian Portuguese as markers. **Proceedings of the XIVth International Congress of Phonetic Sciences**. V. 2, San Francisco, EUA. p. 917-920, ago. 1999.

MADUREIRA, S.; FONTES, M. Fundamental contours in Brazilian Portuguese words. BOTINIS, A; KOUROUPETROGLOU, G.; CARAYIANNIS, G. (Eds.). **Proceedings of the ESCA workshop Intonation: Theory, Models and Applications**. Atenas, Grécia. Univ. of Athens, set. 18-20, 1997, p. 211-214.

MARTIN, Philippe. Synthèse par règles de l'intonation de la phrase. **Actes des Septièmes Journées d'Étude sur la Parole**. GALF, Nancy, 1976, p. 207-213.

MATSUI, E.; SUZUKI, T.; UMEDA, N.; OMURA, H. Synthesis of fairy tales using an analog vocal tract. **Proceedings of the Sixth International Congress on Acoustics**. Tóquio, Japão, B159-162, 1968.

MATTINGLY, Ignatius. Synthesis by rule of prosodic features. **Language and Speech** 9, 1966, p. 1-13.

_____. Synthesis-by-Rule of General American English. Supplement to **Status Report on Speech Research**. Haskins Laboratories New Haven , CT, 1968, p. 1-223.

_____. Synthesis by rule as a tool for phonological research. **Language and Speech** 14. 1971, p. 47-56.

_____. Speech Synthesis for Phonetic and Phonological Models. SEBEOK, T.A. (Ed.). **Current Trends in Linguistics**. Haia: Mouton, v. 12, 1974.

_____. A short history of Acoustic Phonetics in the U. S. OHALA, John J. et al. (Eds). **A Guide to the history of the phonetic sciences in the United States**. Berkeley: University of California.1990, p. 1-6.

MAXEY, H. D. Smithsonian Speech Synthesis History Project. (Division of Information Technology and Society, National Museum of American History), 2001. Disponível em: <http://www.mindspring.com/~dmaxey/ssshp/ss_home.htm>. Acesso em: 28 jun. 2001.

MONCEL, T. du. **Le Téléphone, le microphone et le phonographe**. Paris, 1880.

MOULINES, E.; CHARPENTIER, F. Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. **Speech Communication**. 9 (5/6), 1990, p. 453-467.

MURRAY, Iain R.; ARNOTT, John L. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. **J. Acoust. Soc. Am**. 93(2), 1993, p. 1097-1108.

_____; ARNOTT, John L. Synthesizing emotions in speech: is it time to get excited? **Proceedings of the International Conference on Spoken Language Processing**. V. 3, 1996, p. 1816-1819.

OLIVE, Joseph. Rule synthesis of speech from diadic units. **Proc. Int. Conf. Acoust. Speech Signal Process**. ICASSP-77, 1997, p. 568-570.

ONDREJOVIC, Slavomír. Wolfgang von Kempelen and his "Mechanism of Human Speech", 1996. Tradução do eslovaco. Disponível em:

<http://www.slovakradio.sk/kultura/expstudio/kempe.html>. Acesso em: 4 jun. 2001.

PETERSON, G. E.; WANG, W. S. Y.; SIVERTSEN, E. Segmentation techniques in speech synthesis. **J. Acoust. Soc. Am.** 30 (8), 1958, p. 739-742.

PIERREHUMBERT, Janet. Synthesizing intonation. **J. Acoust. Soc. Am.** 70 (4), 1981, p. 985-995.

PIKE, Kenneth L. **The Intonation of American English**. Ann Arbor: University of Michigan Press, 1945.

POMPINO-MARSCHALL, Bernd. Wolfgang von Kempelen und seine Sprechmaschine. **Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München**. 29, 1991, p. 181-252.

POTTER, R. K. Visible patterns of sounds. **Science**, 102, 1945, p. 463-470.

_____; KOPP, G. A.; GREEN, H. C. **Visible speech**. New York: van Nostrand, 1947.

ROSEN, G. A Dynamic Analog Speech Synthesizer. **J. Acoust. Soc. Am.** 30, 1958, p. 201-209.

RUBIN, Philip; GOLDSTEIN, Louis. The Pattern Playback. Disponível em: <http://www.haskins.yale.edu/haskins/MISC/PP/pp.html>. Acesso em: 11 jun. 2001.

_____; VATIKIOTIS-BATESON, Eric. Talking Heads. s. d. Disponível em: <http://www.haskins.yale.edu/haskins/HEADS/contents.html>. Acesso em: 28 jun. 2001.

RUSSEL, G. O. **The Vowel**. Columbus: Ohio State University Press, 1928.

SAGISAKA, Yoshinori. Speech synthesis by rules using an aotinal selection on non-uniform synthesis units. **IEEE Conf. on Acoust., Seepch and Signal Proc.** S14-8, 1988, p. 679-682.

SAKS, Oliver. **Vendo Vozes: uma viagem ao mundo dos surdos**. Tradução: Laura Teixeira Motta. São Paulo: Companhia das Letras, 1999.

SAITO, S.; HASHIMOTO, S. Speech synthesis system based on interphoneme transition unit. **Proc. Intern. Congr. Acoust.** B-5-12, Tóquio, Japão, 1968.

SCHOTT, L. O. A Playback for Visible Speech. **Bell Laboratories Record**. v. 26, 1948, p. 333-339.

SHANNON, C. E. The Redundancy of English. FOERSTER, H. von (Ed.). **Cybernetics. Trans. of the 7th Conference**, 1951.

SIMÕES, Flávio O. **Implementação de um sistema de Conversão Texto-Fala para o Português do Brasil**, 1999. Dissertação (Mestrado) - FEEC. Universidade Estadual de Campinas.

STEVENS, Kenneth N. The contribution of speech synthesis to Phonetics: Dennis Klatt legacy. **Proceedings of the XIIth International Congress of Phonetic Sciences**. V.1, 1991, p. 28-37.

_____. Speech synthesis methods: homage to Dennis Klatt. BAILLY, Gérard, BENOÎT, Christian (Eds.). **Talking Machines: theories, models, and designs**. Amsterdam: Elsevier Publishers. 1992, p. 3-6.

_____; KASOWZKY, S.; FANT, G.; An electrical analog of the cocal tract. **J. Acoust. Soc. Am.** 25, 1992, p. 734-742.

STEWART, J. Q. An electrical analog of the vocal organs. **Nature**, 110, 1922, p. 311-312.

TILLMANN, H. G. Phonetics, early modern, especially instrumental and experimental work. ASHER, R. E.; SIMPSON, J. M. Y. (Eds.). **The Encyclopedia of Language and Linguistics**. Oxford: Pergamon Press. V. 6, 1994, p. 3082-3095.

TRAUNMÜLLER, Hartmut. Wolfgang von Kempelen's and the subsequent speaking machines. 2000. Disponível em: <http://www.ling.su.se/staff/hartmut/kemplne.htm>. Acesso em: 4 jun. 2001.

UMEDA, Noriko. Linguistic rules for text-to-speech synthesis. **Proceedings of the IEEE**, 64 (4), 1976, p. 443-451.

VAISSIÈRE, J. Language-independent prosodic features. **Prosody: models and measurements**. CUTLER: A.; LADD, D.R. (Eds.). Berlim: Springer-Verlag, 1983, p. 53-66.

VIOLARO, Fábio. Panorama de investigações em processamento de fala no Brasil. **Actas do Primeiro Encontro de Processamento da Língua Portuguesa Escrita e Falada**. Lisboa, p. 183-193, fev. 1993.

YOUNG, S. J.; FALLSIDE, F. Speech synthesis from concept: a method for speech output from information systems. **J. Acoust. Soc. Am.** 66 (3), 1979, p. 685-695.