IDENTIFICAÇÃO DE VOGAIS: ASPECTOS ACÚSTICOS, ARTICULATÓRIOS E PERCEPTUAIS

RICARDO MOLINA DE FIGUEIREDO

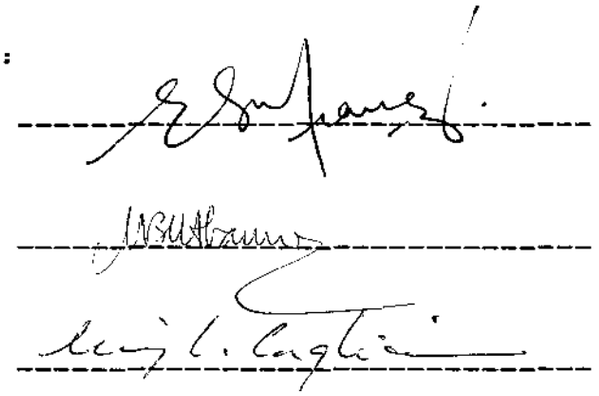
Dissertação apresentada ao Departamento de Lingüística do Instituto de Estudos da Linguagem da Universidade Estadual de Campinas como requisito parcial para a obtenção do título de mestre em Lingüística.

Este exemplar é a redação final da tese defendida por Ricardo Molina de Figueiredo e aprovada pela Comissão Julgadora em 15/03/90.

Campinas, fevereiro de 1990

Orientador: Profª Drª Eleonora Cavalcante Albano

Banca Examinadora:



Campinas, de de 1990

RESUMO

Discussão de questões relacionadas à percepção de sons lingüísticos, com ênfase na identificação de vogais em diversas situações experimentais. Vários níveis de representação são examinados, desde as transformações iniciais realizadas pelo sistema auditivo periférico até a participação dos centros cognitivos, destacando-se o aspecto interativo do processamento. A eficiência de pistas acústicas específicas para a identificação de vogais é discutida e recentes sugestões de novos parâmetros caracterizadores são avaliadas, examinando-se , eventualmente, a possibilidade de utilização de algumas dessas pistas em sistemas automáticos de reconhecimento de fala.

PREFÁCIO

O estudo da percepção da fala, até há pouco tempo atrás, estava praticamente restrito aos lingüistas e a alguns psicólogos. Tem ficado cada vez mais evidente, entretanto, que uma maior compreensão dos processos perceptuais envolvidos na codificação do sinal acústico depende de uma atividade interdisciplinar mais ampla, incluindo também a cooperação de matemáticos, engenheiros, cientistas da fala e audição e pesquisadores da Inteligência Artificial. Esse fato se reflete no grande número de trabalhos mais recentes assinados conjuntamente por investigadores dessas diferentes disciplinas. Esse esforço conjunto tem se mostrado bastante fecundo e vem acenando com resultados que prometem, para um futuro não muito distante, soluções para antigos e renitentes problemas na área de percepção de fala. Vários fatores contribuíram para esse avanço:

- Novas técnicas e estratégias foram desenvolvidas para estudar as diversas fontes de variabilidade acústico-fonética, envolvendo o uso de bancos de dados extensos compilados a partir de grupos heterogêneos de falantes. Análises estatísticas desses dados têm permitido estimativas quantitativas da ocorrência de diversos fenômenos da fala e o aprimoramento de algoritmos de normalização capazes de neutralizar boa parte da variabilidade.

- Muitos pesquisadores voltaram sua atenção para o estudo das manifestações acústicas associadas aos efeitos da coarticulação. Embora a coarticulação imponha elementos adicionais de variação, tem ficado cada vez mais claro que são exatamente os elementos transientes que podem fornecer pistas acústicas mais consistentes. Experimentos recentes têm demonstrado que a informação sobre um segmento fonológico espraia-se também ao longo de segmentos vizinhos, exigindo uma ampliação da janela de análise. Embora a formalização desse conhecimento tenha encontrado sua melhor expressão no âmbito da descrição dos fenômenos articulatórios (especialmente no contexto da "fonologia não-linear" e da "action-theory"), vários estudos têm sido bem sucedidos na determinação de parâmetros acústicos dinâmicos, particularmente no que diz respeito à caracterização de vogais.

- Na última década houve um grande número de trabalhos focalizando o processamento e codificação dos, sinais de fala pelo sistema auditivo periférico. Essa pesquisa tomou duas direções: Por um lado procurou-se identificar certas propriedades no padrão de disparos das fibras nervosas auditivas, associando esses padrões a certas características acústicas dos sons de fala. Por outro lado, vários pesquisadores desenvolveram modelos de base psico-física, incorporando dados psico-acústicos na descrição da filtragem que parece ser reealizada pelo sistema auditivo. O principal objetivo dessa pesquisa é obter

representações espectrais mais adequadas, compatíveis com as capacidades reais da audição humana.

- Novos modelos de processamento foram desenvolvidos, incorporando a participação de centros cognitivos. Esses modelos assumem que, além dos problemas básicos inerentes ao reconhecimento acústico-fonético, a compreensão da mensagem lingüística envolve o acesso a uma variedade de fontes de informação que o ouvinte possui, na condição de usuário de uma língua natural. Além da informação puramente física contida na forma de onda, o ouvinte recorre ao conhecimento estrutural que possui da língua nativa; certas propriedades seqüenciais do Léxico, por exemplo, são usadas ativamente em interação com o dado sensorial, de forma a permitir decisões acústico-fonéticas seguras, mesmo quando há considerável ambigüidade no sinal acústico.

O corpo principal do presente trabalho se propõe a comentar algumas teorias e descobertas empíricas associadas a essas recentes linha de pesquisa, enfatizando os processos subjacentes à identificação de vogais. As duas primeiras seções tratam de temas menos específicos relacionados a propriedades gerais dos sistemas vocálicos, destacando de que forma aspectos de produção/percepção restringem a organização interna desses sistemas.

fndice

1

ASPECTOS BÁSICOS DA PERCEPÇÃO

Os primeiros estudos sistemáticos sobre as vogais ocuparam-se principalmente com a sua reprodução artificial. Em 1769, a Academia Imperial de São Petesburgo propôs as seguintes questões para a atribuição de seu prêmio anual:

1) Qual a natureza e características dos sons das vogais a,e,i,o,u, que as fazem diferir uma da outra?

2) É possível construir um instrumento que, como o registro *vox humana* de um órgão, possa expressar exatamente o som dessas vogais?

O prêmio foi ganho por Kratzenstein. Ele construiu cinco cavidades ressonantes que, quando excitadas por uma palheta vibrante, simulavam as cinco vogais. As formas dos ressoadores, segundo o relato de Kratzenstein, teriam sido baseadas na conformação do trato vocal humano, embora a semelhança não pareça ser tão evidente para nossa visão atual (v. figura 1.1)

Quase na mesma época, von Kempelen, em Viena, conseguia uma imitação mais bem sucedida das vogais. Seu aparelho (v. figura 1.2) consistia de um ressoador cônico de couro maleável cuja forma podia ser ajustada manualmente para cada vogal. A excitação era fornecida através de uma palheta colocada em vibração por umfole. A máquina de von Kempelen era bastante sofisticada em

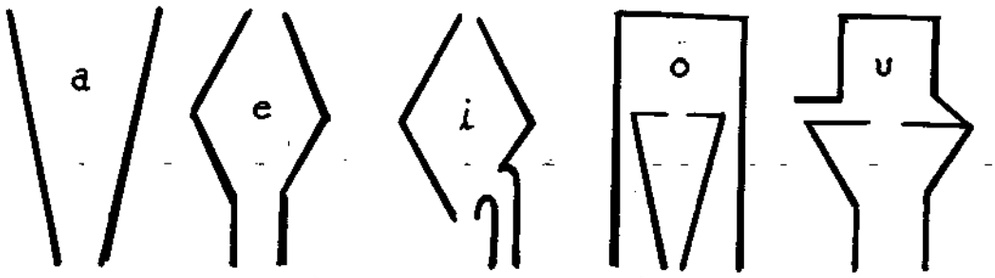


FIGURA 1.1

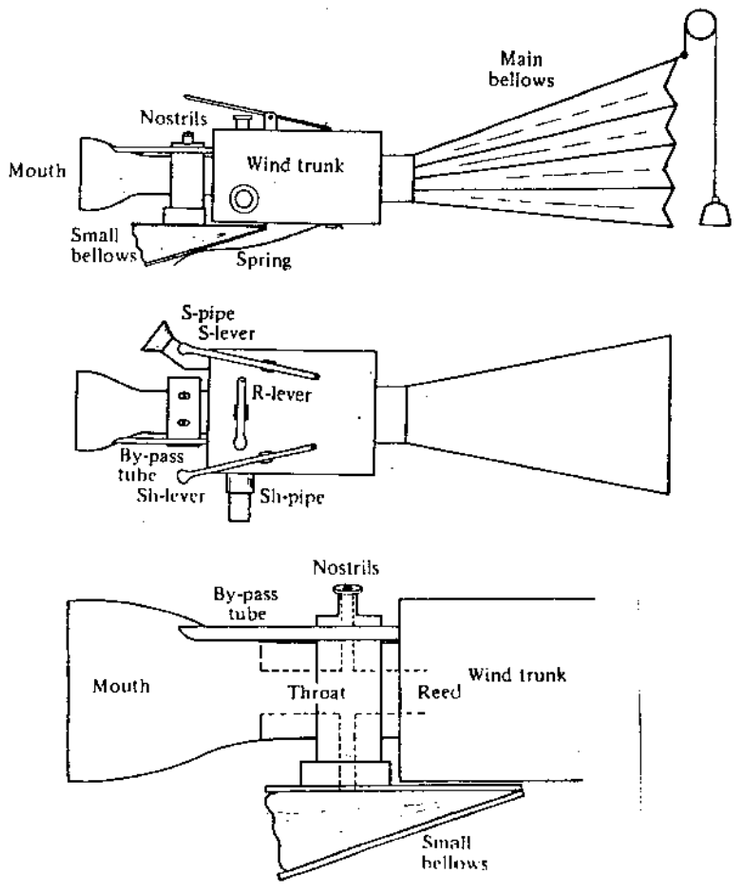


FIGURA 1.2

alguns aspectos e incluía também mecanismos para a produção de várias consoantes. Certos detalhes eram particularmente engenhosos, como o pequeno fole que, na produção de plosivas sonoras - como /b/, por exemplo - simulava a expansão do trato, permitindo a continuação da vibração da palheta (v.seção 5).

A próxima investigação importante foi conduzida por Willis, perto de 1830. Willis realizou uma série de experimentos com tubos uniformes de comprimento variável, usando uma palheta vibrante como fonte de excitação. Possivelmente devido à configuração do experimento, termina por concluir que a qualidade característica das vogais depende de dois sons INDEPENDENTES:o das cordas vocais (correspondendo à palheta) e o de uma ressonância do trato vocal (correspondendo ao tubo uniforme). Willis acreditava que essa ressonância única (para cada vogal) seria a mesma para homens, mulheres e crianças.

Wheatstone, em 1837, constrói uma réplica da máquina de fala projetada por von Kempelen (seguindo instruções escritas deixadas pelo próprio Kempelen). Experimentos baseados no funcionamento desse aparelho levaram-no a conclusões opostas às de Willis. Wheatstone atribuiu a qualidade vocálica à INTERAÇÃO entre as frequências componenciais produzidas pela vibração das cordas vocais e as frequências de ressonânca do trato. Segundo essa teoria, as frequências de ressonância amplificam harmônicos específicos da vibração glotal - daí o nome, usado por alguns, de "Teoria Harmônica". Esse ponto de vista foi mais tarde

desenvolvido por Helmholtz, que também pressupõe uma certa acomodação do trato à freqüência fundamental.

A noção de FORMANTE, tal como hoje a conhecemos, só surgiu realmente com o trabalho de Hermann, no final do século XIX. Hermann retoma algumas sugestões implícitas no trabalho de Willis, afirmando que não precisa haver qualquer relação entre a vibração das cordas vocais e as frequências de ressonância associadas ao trato supralaríngeo. Sendo assim, as cordas vocais operariam apenas como um agente para excitar as frequências das cavidades do trato; um único sopro de ar produzido pela glote é suficiente para colocar o ar dessas cavidades em movimento vibratório e não há necessidade de haver qualquer tipo de ajuste entre a freqüência fundamental e as ressonâncias, ou seja, essas frequências seriam "enarmônicas" uma em relação à outra.

Embora alguma controvérsia em torno da superioridade da teoria "Harmônica" ou da "Enarmônica" tenha animado o círculo acadêmico no final do século XIX, o ponto de vista de Hermann prevaleceu e ficou cada vez mais estabelecido que as vogais poderiam ser caracterizadas por ressonâncias fixas ou "formantes", e que o modo de vibração da fonte seria independente dessas ressonâncias.

Apesar de algumas contribuições importantes (Lloyd 1890, Paget 1923, Crandall 1925 e Fletcher 1929)[1] não houve avanços significativos na teoria das vogais por alguns anos. Com o advento do osciloscópio - na década de 30 - e do espectrógrafo -

na década de 40 - novas perspectivas se abriram para a pesquisa fonética. Imagens mais precisas da forma de onda e medidas exatas das frequências de ressonância podiam ser obtidas diretamente, facilitando a comparação entre as características de diferentes vogais e de diferentes falantes. O lingüista Martin Joos, trabalhando com o equipamento espectrográfico dos Laboratórios Bell, preparou uma importante monografia, "Acoustic Phonetics" (Joos 1948), que lançou as bases da moderna teoria das vogais. Joos pretendia encontrar as ressonâncias invariantes previstas por estudos anteriores, mas seus resultados confirmavam apenas parcialmente essas previsões. Ele observou uma considerável variação inter-falante no padrão de formantes de vogais que, no entanto, eram percebidas como tendo a "mesma" qualidade. Joos conclui que o espaço acústico onde se distribuem as vogais tem uma natureza relativística, no sentido de serem as frequências absolutas dos formantes menos importantes, de um ponto de vista perceptual, do que as relações estabelecidas no interior do espaço vocálico de um falante particular. Segundo Joos, o ouvinte avalia a qualidade de uma vgal referenciando-a a outras vogais produzidas pelo mesmo falante, ou seja, o ouvinte "calibra" sua percepção em função do falante.

É fácil constatar que os valores absolutos dos formantes não podem ser os determinantes últimos das vogais. Basta examinar trabalhos que levantem grande número de dados referentes às frequências de formantes das vogais de uma língua (por exemplo:

Peterson e Barney 1952, para o inglês americano e Behlau 1984, para o Português Brasileiro - cujos dados examinaremos mais adiante). O que se observa nesses levantamentos é que, se traçarmos um gráfico cruzando os dois primeiros formantes, haverá uma sobreposição considerável entre categorias vocálicas vizinhas. Na maioria dos casos, a inclusão do terceiro formante, produzindo um gráfico tri-dimensional, pouco contribuirá para melhorar a classificação.

Ladefoged 1967 relata um experimento onde se procura testar a realidade do relativismo das freqüências dos formantes, segundo a sugestão implícita em Joos (1948). Ele usa como estímulos seis versões sintetizadas da sentença "Please say this word is: b_t", preenchendo a lacuna com vogais sintetizadas com os seguintes formantes:

Palavra Teste	F1	F2
A	375	1700
B	450	1700
C	575	1700
D	600	1300

Além disso fez variar também o nível médio geral de F1 e F2 ao longo de toda a sentença introdutória. Ladefoged verifica que a interpretação da palavra estímulo pode variar dependendo do nível de frequência estabelecido para a sentença precedente. A

palavra teste A, por exemplo, foi identificada como *bit* em 87% dos casos quando precedida pela versão da sentença introdutória na qual F1 varia em uma faixa mais alta de frequência. A mesma palavra teste A é, no entanto, identificada como *bɛt* em 90% dos casos se a sentença introdutória tem F1 variando em faixa mais baixa de frequência.

O que se observa aqui é um efeito da ADAPTAÇXO, um fenômeno bastante conhecido em Psicologia Perceptual e que ocorre também em outras dimensões sensoriais (visão cromática, paladar, avaliação de pesos, comprimentos, etc.). Se o F1 médio precedente é alto o ouvinte avalia o F1 da palavra teste como "mais baixo" e identifica a palavra como *bit*; se o F1 precedente movimenta-se em uma faixa baixa de frequência, o efeito da adaptação é exatamente o oposto, ocorrendo um deslocamento perceptual que dirige a avaliação na direção de uma vogal mais baixa, articulatoriamente, e o ouvinte tende a escutar a palavra teste como *bɛt*. Ainsworth (1975), também usando estímulos sintetizados, chega a resultados semelhantes aos acima descritos (ver também Dechovitz 1977; Ladefoged e Broadabend 1957).

O fenômeno de adaptação pode não estar restrito às influências exercidas pelo ambiente acústico imediato. Há alguma evidência indicando que a estrutura global do sistema vocálico do grupo linguístico ou dialetal ao qual pertence o ouvinte também condicione, em certa medida, os julgamentos categoriais que ele venha a realizar. Ladefoged (1967) verifica que a palavra teste

D, por exemplo, provoca um número quase igual de respostas *bat* e *but* para os sujeitos escoceses, mas 99% de respostas *but* para os sujeitos ingleses. Parece que, no sistema vocálico dos escoceses, a vogal em *bat* está mais próxima da palavra teste D [2].

Vogais Quânticas

É preciso encarar a influência do ambiente acústico com certa cautela. Em primeiro lugar devemos considerar que há alguma contra-evidência experimental. Strange *et al.* (1976), por exemplo, não encontraram influência significativa das vogais precursoras sobre a percepção da vogal teste. Não podemos esquecer, sobretudo, que altas taxas de identificação podem ser obtidas para vogais (incluindo seções apenas dessas vogais) mesmo quando vozes de diferentes falantes são aleatoriamente misturadas (Assmann *et al.* 1982). Outro fato a considerar é que os efeitos de adaptação, tais como verificados nos experimentos em Ladefoged (1967), parecem ter certos limites naturais: o ouvinte, influenciado pelo padrão de formantes precedente, pode eventualmente avaliar /i/ por /e/, ou vice-versa, mas não confundirá, por exemplo, /i/ com /a/, por mais que se manipule o nível geral de freqüência da sentença precursora.

Não se trata, portanto, de um "relativismo absoluto". Se esse fosse o caso, cairíamos, sem dúvida, na armadilha que Nearey (1989) chama de "the bootstrap problem", isto é, se cada vogal é

relativa a cada outra vogal, como poderíamos, afinal, entrar no sistema? (v. também Peterson 1961).

Uma saída parcial para esse impasse pode ser encontrada em certas restrições acústicas e articulatórias dos sistemas vocálicos. O fato é que determinados sons parecem ocupar uma espécie de lugar privilegiado dentro dos sistemas. As vogais /i, a,u/, por exemplo, são encontradas na grande maioria das línguas (Maddieson 1986; Lindblom 1986). Essas categorias vocálicas são universalmente favorecidas aparecendo de modo mais estável e praticamente independente do tamanho (número de elementos) do sistema.

Stevens 1972, em um trabalho bastante conhecido, sugere que algumas vogais possuem uma espécie de natureza quântica, i.e., as manobras articulatórias envolvidas na produção dessas vogais podem ser razoavelmente flexibilizadas sem que haja, necessariamente, uma alteração radical na saída acústica do sinal. A tese geral de Stevens é que a relação entre os parâmetros que descrevem a articulação e a saída acústica não é simplesmente LINEAR; desse modo, assim como podem ocorrer configurações articulatórias onde um deslocamento dos articuladores pouco altera o aspecto acústico, também há situações onde uma pequena mudança na articulação está associada a uma mudança acústica brusca. Um dos exemplos dados por Stevens focaliza a produção de uma vogal cuja qualidade seja próxima a um /i/; a partir de uma determinada posição, um mínimo aumento da

constrição provocará a turbulência suficiente para a produção de uma fricativa [3]

Percepção Categorial

Stevens 1972 sugere ainda que as transformações impostas ao sinal acústico pelo mecanismo auditivo podem acentuar o caráter de "platô" das regiões nas quais a sensitividade à mudança articulatória é baixa. Haveria, dessa forma, uma certa cumplicidade entre os mecanismos gerador e auditivo no sentido de que o falante seleciona sons com atributos acústicos bem definidos e o sistema auditivo está, por outro lado, predisposto a perceber sons com tais atributos de modo categorial.

O fenômeno conhecido como Percepção Categorial parece confirmar essa sugestão de Stevens, pelo menos para alguns sons. Alguns experimentos desenvolvidos pelos Laboratórios Haskins nas décadas de 50 e 60 revelaram que a capacidade perceptual para discriminar alguns pares de sons era MENOR que a capacidade para identificar, com rótulos,esses mesmos sons (Liberman *et al.* 1957). Esses resultados pareciam ir contra as expectativas, já que a maioria dos testes psico-perceptuais indicavam que dimensões sonoras isoladas, tais como frequência, timbre, intensidade, podiam ser discriminadas com muito mais precisão do que identificadas (v. p.ex. Miller 1956; Pollack 1952).

Esses primeiros estudos, facilitados pela técnica então recente do *pattern playback*, verificaram que em certas fronteiras fonéticas, como por exemplo entre as sílabas /ba/-/da/, ocorre uma espécie de salto perceptual descontínuo - ou seja, "quantizado" - assim que a transição de F2 atinge uma determinada inclinação no estímulo sintetizado. Estudos posteriores com outros contrastes consonantais observaram que o fenômeno se repetia também em contínuos /sonoro-não sonoro/, /fricativo-africado/, /palatal-não palatal/, etc, embora com menor intensidade do que para consoantes plosivas (para uma vasta bibliografia nessa área ver Repp 1984).

O mesmo tipo de teste, entretanto, quando realizado com contínuos representando sons vocálicos apresentou resultados diferentes dos obtidos com estímulos consonantais. Quando o conjunto de estímulos é, por exemplo, extraído de uma interpolação entre os extremos /i/-/I/ (fazendo-se variar linearmente os valores dos três primeiros formantes), embora ocorra um pico na função de discriminação localizado na região central do contínuo, o índice de discriminação é muito maior do que no caso das consoantes, isto é, os sujeitos são capazes de discriminar razoavelmente bem dois sons vocálicos mesmo que os estímulos estejam próximos a um dos extremos do contínuo (Fry *et al.* 1962; Cutting 1977). Para as vogais, portanto, ao contrário das consoantes, a percepção parecia operar de forma mais contínua do que propriamente categorial.

Essa associação mais estreita de uma percepção do tipo categorial com as consoantes foi um dos argumentos chave para o desenvolvimento da TEORIA MOTORA (Liberman *et al.* 1967). Segundo essa abordagem, a percepção e o controle articulatório envolveriam os mesmos processos neurológicos (ou processos intimamente inter-relacionados). Quando categorias fonéticas diferentes estão associadas a gestos articulatórios essencialmente discretos - como no caso de consoantes plosivas, por exemplo, - a percepção dos estímulos de um contínuo físico abrangendo essas categorias será categorial. Por outro lado, se variações articulatórias contínuas entre categorias fonéticas são possíveis - como no caso das vogais - então espera-se que a percepção também se efetue de uma forma mais ou menos contínua, isto é, com alta discriminabilidade em pares de estímulos próximos, independentemente da posição ocupada por esses estímulos a longo do contínuo. Em suma: a Teoria Motora considera a percepção categorial como um reflexo direto da organização articulatória.

Acústico ou Fonético?

Explanações de natureza puramente psico-perceptual foram esboçadas para o fenômeno da percepção categorial, principalmente com base no funcionamento da memória a curto termo (MCT). Alegou-se que as diferenças intra-categoriais entre duas ocorrências de

sons vocálicos permaneceriam mais tempo na MCT - em comparação com as consoantes -, de tal modo que o sujeito, durante o teste perceptual, seja capaz de fazer comparações mais exatas entre os estímulos. A discriminação de vogais envolveria, portanto, processos perceptuais baseados fundamentalmente em informação ACúSTICA e não predominantemente FONÉTICA como para as consoantes (Cf. Cutting 1977).

A disputa teórica que emergiu desses experimentos girou principalmente em torno de uma possível MODULARIDADE do mecanismo perceptual. A base lógica da Teoria Motora, por exemplo, tornava conveniente, senão necessária, uma divisão dos sons em duas classes gerais: os que são fala e os que não são. Consequentemente, seria também conveniente postular uma especificidade dos mecanismos perceptuais envolvidos no processamento de fala. As questões mais importantes que nortearam essa discussão podem ser assim resumidas:

- existem argumentos suficientes para postular dois modos distintos de percepção: o acústico (ou auditivo) e o fonético?
- em que medida, e de que modo, esses módulos se inter-relacionam durante o processo perceptual?
- seria o módulo fonético uma evolução adaptativa a partir de um modo perceptual basicamente auditivo mais primitivo?

Essas questões talvez não tenham uma resposta simples e direta e a discussão tem se mantido bastante viva até hoje, conduzida principalmente à sombra da polêmica gerada em torno das teses gerativistas (v.p.ex. Fodor 1983, e o comentário de Albano 1987). O exame de algumas pesquisas experimentais pode, no entanto, lançar alguma luz sobre essa problemática.

Uma das linhas de pesquisa concentrou-se no estudo dos mecanismos perceptuais utilizados por crianças em fase pré-linguística. Algumas dificuldades metodológicas inerentes a esse tipo de teste foram superadas através de uma série de engenhosos experimentos onde as respostas de crianças entre 1-4 meses de idade foram aferidas pelo batimento cardíaco e ritmo de sucção registrado através de um sensor conectado a um bico "chupeta". Presumiu-se que a criança apresentaria alterações desses parâmetros quando percebesse que um determinado estímulo era "novo" em relação ao anterior, repetidamente apresentado. Dessa forma foi possível verificar, sem recurso a respostas diretas, a capacidade de discriminação. Esses experimentos revelaram que os infantes apresentavam um comportamento bastante parecido ao dos adultos, no que diz respeito à percepção categorial de sons nos contínuos /ba/-/pa/,/ba/-/da/ e /bæ/-/dæ/ (Eimas 1974; Eimas *et al.* 1971; Morse 1972).

Outra linha de pesquisa enfocou a percepção categorial em animais. Kuhl e Miller (1975;1978) observaram que chinchilas dividiam o contínuo de VOT (*Voice Onset Time*) entre plosivas

alveolares, representando um contraste sonoro-não sonoro, da mesma forma que humanos, isto é, categorialmente. Morse e Snowdon (1975), usando estímulos extraídos do contínuo /bæ/-/dæ/-/gæ/, relatam que macacos Rhesus demonstram razoável discriminação intercategorial e baixa sensitividade a diferenças intracategoriais.

Os experimentos com animais têm de suplantar sérios problemas metodológicos (aferição da resposta, p.ex.), e a validade dos resultados tem sido alvo de alguma controvérsia (Cf. Repp 1984). O pequeno número de trabalhos nessa área parece refletir essas dificuldades laboratoriais. De qualquer forma, pelo menos um trabalho mais recente demonstrou com sucesso a existência de percepção categorial para alguns sons em primatas (Kuhl e Padden 1983).

Os resultados dos experimentos com crianças de colo e com não-humanos sugerem que o fenômeno da percepção categorial não pode ser atribuído - pelo menos para alguns sons - apenas à experiência lingüística. Essa stuação levou alguns autores a especular quanto à possível pré-existência, filogenetica- e ontogeneticamente, de um modo auditivo elementar de percepção, a partir do qual evolui o módulo fonético especializado. Segundo esse ponto de vista, as categorias linguísticas seriam essencialmente psico-acústicas em sua natureza, não existindo, a rigor, a necessidade de postular um modo fonético independente de percepção. A fala, afinal, não seria um fenômeno tão especial, do

ponto de vista sonoro, como alguns preferem acreditar.

Alguma evidência suportando essa hipótese foi fornecida por estudos que observaram percepção categorial para estímulos não-linguísticos. Cutting e Rosner 1974, fazendo variar o tempo de ataque de uma onda "dente-de-serra" em incrementos de 10 milisegundos entre os extremos de 0 a 80 ms obtêm com esses estímulos funções de discriminação muito similares às observadas para sílabas como /ba/-/da/, por exemplo. Os itens com ataque inferior a 40 ms foram identificados como *pluck* em praticamente 100% das respostas; estímulos com ataque maior que 40 ms foram identificados como *bow*: apenas o estímulo situado no centro do contínuo, com ataque de 40 ms, foi interpretado de forma ambígua, recebendo 40% de identificações *pluck* e 60% de *bow* [4]. Além disso, na comparação direta de pares de estímulos, os ouvintes tendem a considerar DIFERENTES apenas os ítens que se encontram em lados opostos do conjunto de estímulos (estímulos com 10 e 30 ms de ataque são considerados IGUAIS, mas estímulos com 30 e 50 ms - embora tenham a mesma distância física - são considerados diferentes) (v. também Cutting *et al.* 1976).

O mais interessante em relação a esse tipo de experimento é que o parâmetro tempo de ataque não é uma pista apenas para a distinção de sons quase-musicais como os usados no teste acima descrito, mas é também importante para estabelecer contrastes linguísticos. A distinção entre uma africada (como /t∫a/) e uma fricativa (como /∫a/) baseia-se quase que exclusivamente no

ataque muito mais gradual da segunda. Quando um conjunto dessas sílabas gerado sinteticamente com os mesmos tempos de ataque dos estímulos não linguísticos é submetido ao mesmo tipo de teste perceptual, os ouvintes produzem padrões de identificação e discriminação muito semelhantes aos produzidos para os ítens pluck vs. bow (Cutting e Rosner 1974) (v.também Miller *et al.* 1976).

Ora, se a mesma pista isolada - tempo de ataque - pode ser usada para distinguir categorias tanto dentro quanto fora da fala, seria razoável supor que pelo menos algumas das distinções binárias sobre as quais a fala é construída poderiam ser baseadas em distinções binárias AUDITIVAS, e aí teríamos, uma vez mais, evidência em favor da prevalência de um modo auditivo de percepção. Jusczyk *et al.* 1977 repetem o experimento *pluck vs bow* com crianças de dois meses e também observam um padrão categorial nas respostas. Pesquisas mais recentes enfocando a discriminação de sons não-linguísticos por crianças de cerca de oito meses têm verificado que o sistema auditivo dos infantes responde a esses estímulos de modo semelhante ao do adulto. Esses resultados têm sido interpretados como um suporte à tese de que a percepção da fala pode estar diretamente ligada a habilidades psico-acústicas básicas (v. Aslin 1989).

O ponto crucial da distinção auditivo/fonético passou a ser para alguns pesquisadores, uma questão de transformação psico-fisiológica, sendo o mecanismo perceptual concebido como um

sistema de análise basicamente auditiva com diferentes níveis de processamento de crescente complexidade (v.p.ex. Blechner *et al.* 1976). Para esse grupo a equivocada postulação de modos "especiais" de processamento para a fala deve-se, principalmente, ao conhecimento ainda incompleto dos fatores acústicos relevantes que subjazem aos sons de fala. A possível distinção auditivo/fonético se revelaria, assim, menos de essência do que de grau.

Essa posição, entretanto, não descarta a possível existência de mecanismos neurais especializados seletivamente a sinais acústicos particulares, embora não necessariamente associados à fala. Esses "detetores binários", tal como foram chamados, passaram a ser a explicação, com base neurológica, da percepção categorial. E mais, esses mecanismos, como indicavam as experiências com crianças de colo, pareciam ser um equipamento inato do aparelho perceptual.

A existência de mecanismos neurais ("microprocessadores") específicos para alguns sons não é surpreendente, já que estudos com animais diferentes do *Homo Sapiens* têm repetidamente demonstrado que mecanismos similares estão presentes em seus cérebros. Técnicas eletrofisiológicas, cuja aplicação é impossível em humanos, já isolaram mecanismos neurais que respondem a sinais específicos de interesse para os animais, em particular sinais acústicos usados para a comunicação. Mesmo animais simples como o grilo parecem estar equipados com unidades

neurais que codificam informação sobre os elementos rítmicos do canto de acasalamento desses insetos (Hoy e Paul 1973).

Especialmente interessantes são os experimentos com anfíbios anuros. Rãs e sapos são os animais mais simples que produzem sons por meio de uma fonte laríngea e um trato vocal supralaríngeo. As vocalizações desses animais são produzidas de forma comparável à dos primatas: as cordas vocais na laringe vibram rapidamente emitindo pequenos jatos de ar no interior do trato supralaríngeo, que age como um filtro acústico modulador. Estudando um tipo de rã da América do Norte, a *Rana Catesbiana*, Capranica 1965 observa a existência de diferentes tipos de vocalização, cumprindo funções específicas (acasalamento, manutenção territorial, alarme para o grupo, etc.), associadas a propriedades acústicas distintas e, obviamente, a diferentes manobras articulatórias. A vocalização associada ao acasalamento, em especial, assemelha-se a uma vogal, de tal modo que foi possível utilizar um sintetizador de fala para produzir os estímulos artificiais nos experimentos com a *Rana Catesbiana*. Esse tipo de vocalização produz frequências de formantes nas faixas de 200 e 1400 Hz, aproximadamente. Os testes mostraram que as rãs só respondem aos coaxos sintetizados se há uma concentração de energia em uma ou nas duas faixas de frequência dos formantes que caracterizam o coaxo natural. A presença de energia acústica em outras frequências tem o efeito de inibir as respostas do animal.

Examinando diretamente as células neurais do mesmo tipo de rã, Frishkopf e Goldstein 1963, em um estudo eletrofisiológico, observam dois tipos diferentes de células auditivas: uma com sensitividade máxima para frequências entre 1000 e 2000 Hz e outra respondendo maximamente na região entre 200 e 700 Hz. As unidades que respondem às frequências mais baixas são inibidas se é acrescentado um formante extra na faixa de 500 Hz.

Experiências com primatas também revelaram a existência de diferentes tipos de célula neural. Registrando a atividade elétrica de células isoladas no córtex auditivo do macaco-de-cheiro (*Saimiri Sciureus*) desperto, Wollberg e Newman 1972 verificam que algumas unidades neurais respondem à maioria das vocalizações com propriedades acústicas complexas, enquanto outras unidades respondem apenas a alguns tipos de sinal [5].

A existência de unidades neurais sintonizadas para alguns sons que ocorrem na fala pode parecer contra-evidência à hipótese da natureza essencialmente não fonética do mecanismo perceptual. O que deve ser entendido, no entanto, é que esses detetores talvez funcionem como analisadores de traços ACúSTICOS e responderão da mesma forma se o sinal disparador aparece em um contexto de fala ou não. Por outro lado, seria ingênuo acreditar que níveis superiores de codificação não interfiram de algum modo com esses detetores. Na verdade, na maioria dos casos, a experiência linguística parece perturbar o funcionamento de algumas unidades neurais de decisão binária, quase sempre no

sentido de PIORAR o seu desempenho.

Repp 1984 relata alguns experimentos focalizando a influência do *background* linguístico nas funções de discriminação e identificação. Sons no contínuo /ra/-/la/ apresentam, por exemplo, um pico de identificação nos testes feitos para falantes nativos de inglês americano, enquanto os sujeitos japoneses, submetidos aos mesmos estímulos, são incapazes de discriminar adequadamente pares de sílabas, independentemente da distância física entre elas ao longo do contínuo. No entanto, se os F3 desses mesmos estímulos forem apresentados em isolamento, dissolvento assim a impressão de som de fala, americanos e japoneses obtêm resultados similares. Esse resultado é sugestivo, pois a única pista diferenciando as sílabas sintetizadas no contínuo /ra/-/la/ era exatamente a transição de F3.

É como se a capacidade física de discriminar lá estivesse, embora o falante, de alguma forma influenciado pela estrutura fonológica particular de sua língua, acabe tomando uma decisão cognitiva - e não puramente perceptual - perdendo assim parte de sua capacidade linguística de distinguir. Experimentos com crianças em fase pré-linguística mostram, por exemplo, que elas são sensíveis a certas distinções que não são fonêmicas em sua futura língua. Crianças norte-americanas são capazes de discriminar o contraste /pré-vozeado - não vozeado/, enquanto os rebentos espanhóis discriminam sem dificuldade a oposição /aspirado - não aspirado/, embora esses contrastes não ocorram

nos sistemas fonológicos de suas línguas nativas respectivas (Repp 1984).

Essa sensibilidade infantil a certas distinções parece ser perdida na fase adulta a não ser que os traços específicos tornem-se associados a uma distinção fonológica. Essa situação evoca uma curiosa INdistintividade adquirida, uma espécie de DESaprendizado imposto pelo *background* linguístico particular. Essa perda de capacidade distintiva pode, de fato, ocorrer bem cedo no desenvolvimento do indivíduo: Werker 1982 (*apud* Repp 1984), em um raro estudo longitudinal sobre percepção categorial, relata o desaparecimento de algumas distinções não-fonêmicas já aos 8-10 meses, uma fase quando, sintomaticamente, certos segmentos começam a emergir no balbucio.

Outro fator que parecia influenciar a performance dos sujeitos em testes de percepção categorial estava relacionado ao fenômeno da adaptação seletiva. Já vimos, em Ladefoged 1967, que o ambiente acústico imediatamente precedente pode influenciar a classificação de uma vogal, ao ponto mesmo de confundir certas fronteiras categoriais (julgar um intencionado /i/ como /e/, ou vice-versa, dependendo do nível geral de F1 da sentença introdutória). Alguns pesquisadores diretamente ligados à questão da percepção categorial tentaram averiguar, em experimentos menos informais do que em Ladefoged 1967, em que medida ocorreriam desvios sistemáticos das fronteiras categoriais em situação pós-adaptativa; além disso - e talvez o mais importante - havia

interesse em verificar, caso existissem efeitos pós-adaptativos, qual a natureza do processo perceptual envolvido. Como se sabe, efeitos pós-adaptativos são observáveis em relação a várias dimensões sensoriais; esses efeitos, entretanto, podem estar associados a níveis de processamento bem distintos. Efeitos na visão cromática, por exemplo, ocorrem em regiões periféricas; o mesmo não ocorre com a avaliação subjetiva de medidas de comprimento, onde os desvios perceptuais estão certamente associados a mecanismos de ordem cognitiva.

Antes de analisarmos os experimentos focalizando efeitos da adaptação para sons, será ilustrativo examinar certas características do mesmo fenômeno para a visão cromática, uma área de pesquisa onde parece haver mais acordo quanto à base neurofisiológica dos efeitos de adaptação (Cf. Mc.Collough 1965; Blakemore e Campbell 1969). Além disso, certos paralelos entre as duas dimensões podem ajudar a compreender o que provavelmente ocorre com a audição.

Os efeitos provocados pela adaptação na visão têm a vantagem de ser facilmente verificáveis. Se uma pessoa olhar fixamente para uma mancha de cor azul por cerca de 15-30 segundos e então, repentinamente, fixar a visão em uma parede branca bem iluminada, ela verá uma mancha amarela com o mesmo contorno da original azul. Esse efeito é conhecido como "pós-imagem cromática" e é melhor entendido em termos de mecanismos envolvendo processos OPONENTES. O azul é interpretado como uma cor OPOSTA ao amarelo,

pelo menos em algum estágio da análise fotocromática subsequente à excitação ds cones da retina. Amarelo e azul parecem estar ligados às mesmas células, mas uma das cores estabelece essa ligação de um modo excitatório, enquanto a outra, oposta, o faz de modo inibitório. Fixar uma mancha ou figura de cor azul por vários segundos causa uma fadiga no sistema visual de tal modo que, quando submetido a um estímulo neutro - a cor branca - o sujeito perceberá, por um breve período de tempo, esse estímulo como sendo da cor oposta - no caso, o amarelo.

O efeito ocorre também reciprocamente: fixar uma mancha amarela produzirá uma sensação de azul quando o sujeito fixar um campo pós-adaptativo neutro.

A situação experimental para a adaptação a estímulos sonoros é necessariamente diferente daquela dos experimentos com a visão. Como o sinal auditivo extingue-se rapidamente, é preciso apresentá-lo repetidamente ao sujeito - em certos casos de 100 a 200 vezes -, de modo a reavivar continuamente a "imagem" perceptual na memória. Se o estímulo adaptativo (aquele previamente repetido para o sujeito) é, por exemplo /da/ e um estímulo neutro perto da fronteira entre /ba/ e /da/ é apresentado ao sujeito imediatamente após a estimulação adaptativa, ele perceberá esse estímulo neutro (que, sem a adaptação seria classificado como "ambíguo") como um bom exemplar de /ba/ (Cutting 1977). Tudo parece ocorrer como se /ba/ e /da/, a exemplo do que acontece com o amarelo e o azul, fossem

"opostos" um ao outro.

Dois aspectos, entretanto, distinguem o efeito de adaptação com estímulos de fala do efeito com cores. Em primeiro lugar, na audição, o efeito é muito mais duradouro: enquanto a pós-imagem cromática dura, no máximo, cerca de 30 segundos (McCollough 1965), o desvio perceptual em um contínuo /ba/-/da/, por exemplo, pode permanecer até por algumas horas (Eimas *et al.* 1973). A segunda diferença diz respeito ao *locus* do efeito no sistema perceptual; na visão, o efeito de adaptação NÃO se transfere de um olho para o outro, isto é, se a mancha azul for exposta apenas ao olho direito, não será criada uma imagem cromática pós-adaptativa no olho esquerdo. Isso demonstra claramente que o *locus* do efeito cromático é bastante periférico, ou, pelo menos, muito perto da retina. Por outro lado, o efeito de adaptação com estímulos de fala TRANSFERE-SE de um ouvido para o outro e, geralmente, mantendo a mesma magnitude (Eimas *et al.* 1973), o que indica, em contraposição ao fenômeno visual, a existência de processos mais ou menos centrais [6].

Esses resultados com sons de fala pareciam indicar que a base da adaptação era mais fonética que auditiva, ou seja, que os deslocamentos perceptuais poderiam ser atribuídos a níveis de processamento hierarquicamente superiores. Mas a evidência não é conclusiva. Cutting 1977 rebate a hipótese de que os desvios perceptuais pós-adaptativos reflitam diretamente critérios cognitivos de decisão; ele argumenta que a adaptação afeta não

apenas a identificação (deslocando o limite categorial) mas TAMBÉM a discriminação (afetando proporcionalmente a performance do sujeito em discriminar ítens ao longo do contínuo), o que sugere uma natureza essencialmente auditiva para o fenômeno. A transferibilidade do efeito adaptativo de um ouvido para o outro tampouco serve como evidência suficiente para se pensar o contrário, já que isso indica apenas que o *locus* físico do processamento encontra-se em algum ponto DEPOIS da convergência das fibras nervosas originárias do ouvido, o que pode ser um ponto tão baixo no sistema, como o complexo olivário superior, ou tão alto quanto o córtex (Cf. Denes e Pinson 1973).

Os resultados dos experimentos baseados no paradigma da adaptação seletiva não aconselham que se assuma uma posição radical quanto à natureza linguística ou não linguística dos possíveis "detetores neurais". As evidências apresentadas por esses estudos são quase sempre inconclusivas e passíveis de interpretações dúbias. Um exemplo típico dessa situação são os resultados obtidos por Ades 1974. Ele procura verificar se os efeitos pós-adaptativos para um determinado som linguístico são transferíveis entre contextos diferentes, isto é, se esses efeitos permanecem mesmo que se mude, por exemplo, o ambiente vocálico ou a posição do som na sílaba. Ades observa que a adaptação com /de/ desloca a fronteira categorial entre /bæ/-/dæ/ quase da mesma forma como ocorre com /dæ/. Por outro lado, ele também relata que a adaptação com /dæ/ não tem nenhum efeito

sobre o contínuo /æb/-/æd/. Ora, a situação aqui é bastante curiosa: o efeito parece fonético o suficiente para se transferir através de algumas diferenças acústicas tais como aquelas entre ambientes vocálicos distintos, mas não fonético bastante para se transferir para uma outra posição silábica.

Os resultados de Ades 1974 não são fáceis de interpretar. Aquele que deseja dar conta dos dados em termos puramente auditivos poderia alegar que a diferença entre as duas condições se deve ao fato de haver menos dimensões acústicas em comum entre as duas sílabas com a consoante em posição diferente do que entre duas sílabas onde apenas as vogais são diferentes; no primeiro caso, apesar do núcleo vocálico comum, as transições dos formantes percorrem trajetórias diferentes em cada sílaba. Uma explanação em termos lingüísticos requer talvez um pouco mais de engenhosidade, mas é sempre possível: basta postular, por exemplo, um nível silábico de processamento; uma hipótese suportada por alguma evidência experimental (Savin e Bever 1970; Wood e Day 1975; Fujimura 1975).

Uma forma menos ortodoxa, e mais sensata, de abordar a distinção fonético/auditivo é admitir que os dois tipos de processamento interagem na percepção. Na verdade não há grande desacordo quanto a aceitar a existência de níveis de codificação hierarquicamente organizados, relacionados a diferentes transformações do sinal, progressivamente na direção de uma representação mais "simbólica" nos níveis superiores [7]. Há

vantagens no processamento em vários níveis e esse é um princípio geral de organização cognitiva, válido para outros sistemas organizados hierarquicamente (Deutsch 1982). Se essa hierarquia espelha apenas uma gradação crescente de complexidade, ou se torna patente, em algum nível, uma alteração qualitativa mais específica que justifique a atribuição do rótulo "linguístico" (ou "fonético") parece ser uma questão secundária, pelo menos para algumas áreas de pesquisa. Não há, por outro lado, qualquer prejuízo em fazer referência à distinção fonético/auditivo desde que haja razoável acordo quanto ao significado geral desses termos. O fato é que essa distinção tem inegável valor heurístico para algumas linhas de pesquisa e não pode ser simplesmente descartada. É preciso considerar também que estudos focalizando a identificação bi-auricular reforçam a tese da realidade psicológica da distinção fonético/auditivo, revelando frequentemente prevalência do ouvido direito para os estímulos de fala (Shankweiler e Studdert-Kennedy 1967) e do ouvido esquerdo para sons não-linguísticos (Kimura 1964). Talvez, no futuro, se atingirmos uma maior compreensão dos fatores acústicos relevantes constitutivos da fala, haverá uma menor ênfase na postulação de modos "especiais" de processamento associados à fala. Atualmente, no entanto, a abordagem mais adequada é a que concilia os dois modos, fonético e auditivo, e procura examinar, com base no ainda limitado conhecimento que possuímos do sinal acústico, quais as formas de interação desses hipotéticos níveis de

codificação.

Nesse contexto, uma das questões mais importantes diz respeito à simultaneidade ou não dos diferentes processos perceptuais. Como interagem - se é que interagem - os níveis de codificação, serial ou paralelamente, ou ambos? [8] Alguns experimentos investigaram a relação entre os processos fonéticos e auditivos baseando os testes em tarefas exigindo atenção seletiva para estímulos que variam ao longo de duas dimensões simultaneamente. Observou-se que, quando ambas as dimensões são lingüísticas - como por exemplo, consoante plosiva inicial e vogal em sílabas CV - a atenção seletiva para qualquer dimensão é prejudicada (isto é, aumenta o tempo de reação para a identificação) se a outra dimensão varia aleatoriamente (Wood e Day 1975). O mesmo padrão de interferência simétrica também ocorre se as duas dimensões são não-linguísticas, como por exemplo, altura melódica e intensidade (Wood 1975). O resultado mais interessante é obtido, entretanto, quando uma dimensão é lingüística e a outra não; nesse caso surge uma interferência assimétrica: os tempos de reação para a identificação de consoantes plosivas em sílabas CV aumentam significativament se a altura melódica varia aleatoriamente, mas os tempos de reação para a identificação da altura melódica NÃO se modificam com a variação da consoante (Day e Wood 1972).

Esses resultados, antes de mais nada, suportam a distinção fonético/auditivo. A conseqüência mais importante, no entanto, é

que o padrão de interferência assimétrica resultante dos experimentos de Day e Wood 1972 sugere um modelo de processsamento SERIAL: a dimensão não-linguística sendo processada ANTES da dimensão lingüística..

Esse ponto de vista é, no entanto, ameaçado por outra evidência experimental. Wood 1974 verifica que quando AMBAS as dimensões, linguística e não-linguística (consoante plosiva e altura melódica em sílabas CV), variam REDUNDANTEMENTE, os sujeitos as identificam mais rapidamente do que na condição onde apenas uma dimensão varia. Esse fenômeno, chamado de "ganho de redundância" (Garner e Felfoldy 1970) é um forte argumento contra o modelo serial: se o processamento da altura melódica já estivesse completado antes de começar o processamento da consoante, como sugere o modelo serial, o sujeito não seria capaz de utilizar informação redundante sobre a consoante de forma a diminuir o tempo de resposta para a identificação da altura melódica. Um modelo PARALELO, assumindo que o processamento das duas dimensões se sobrepõem, ou são simultâneos, parece explicar melhor os fatos.

Modelos de processamento estritamente seriais ou paralelos não dão conta, entretanto, do quadro como um todo. Blechner er al. 1976, também utilizando estímulos bi-dimensionais, com as duas dimensões não-lingüísticas (tempo de ataque e intensidade), mas com maior controle das condições experimentais, verificou que os sujeitos possuem um certo grau de liberdade quanto ao tipo de

processamento que utilizam em diferentes condições. Observou-se que as estratégias de processamento (serial ou paralelo) podem ser opcionais para algumas condições, mas obrigatórias para outras. Além disso, um dos testes realizados nesse estudo revelou ainda um padrão de interferência assimétrica: a variação da intensidade interfere com o processamento do tempo de ataque, embora este não tenha virtualmente qualquer efeito no processamento da intensidade, um padrão observado, até então, apenas em estímulos mistos, onde uma dimensão é lingüística e a outra não (Day e Wood 1972; Wood 1974,1975).

Se aceitarmos os resultados de Blechner *et al.* 1976 como evidência, será preciso reconsiderar certas questões. A interferência assimétrica na percepção de estímulos bi-dimensionais não-lingüísticos indica que há mais de um nível de processamento atuando. Não parece prudente, pois, agrupar indistintamente todas as propriedades acústicas que não determinam pistas lingüísticas, associando-as a um nível específico de processamento. Nessa mesma direção apontam os experimentos que verificaram percepção categorial também para sons não lingüísticos, como já vimos anteriormente em relação à distinção *pluck vs. bow* (Cutting e Rosner 1974).

Da mesma forma, não seria razoável associar os sons "lingüísticos", como um todo, a uma estratégia comum de processamento; já vimos, quando examinamos algumas pesquisas sobre percepção categorial, que estímulos vocálicos e

consonantais não produzem perfis idênticos de resposta: a percepção de vogais envolve provavelmente mecanismos menos específicos, ou especializados, que a percepção de consoantes. Cutting 1977 sugere que a percepção mais contínua das vogais é uma consequência de sua maior força de representação na memória auditiva. Se a hipótese de Cutting está correta, é razoável esperar que perturbações no funcionamento da MCT devam alterar o perfil das respostas categoriais.

Modificações na Percepção Categorial

O papel específico da memória auditiva foi focalizado em uma série de experimentos onde se observou as alterações nas funções de discriminação e identificação sob condições desfavoráveis para a memória auditiva. Há várias formas de dificultar o trabalho da memória, entre elas:

- redução temporal do estímulo
- inclusão de intervalo de silêncio entre os estímulos
- inclusão de interferência (ruído branco ou um som qualquer) entre os estímulos a comparar.

Esse tipo de *design* experimental produziu resultados diferentes para vogais e consoantes. Examinaremos a seguir alguns desses trabalhos.

Pisoni 1973 intercala intervalos variáveis de silêncio (de 0 a 2 segundos) entre apresentações de estímulos extraídos dos contínuos /i/-/I/, /bæ/-/dæ/ e /ba/-/pa/. Ele observa, dentro das expectativas, uma queda na discriminação mais ou menos proporcional ao intervalo de silêncio inserido. No entanto, o efeito global é muito pequeno para os contínuos consonantais, sendo relevante apenas para as vogais. A ausência de efeito nas consoantes não surpreende, já que a discriminação intra-categorial desses segmentos é baixa mesmo em condições favoráveis; a alteração das taxas de discriminação para as vogais sugere, entretanto, que a MCT pode ter um papel importante na percepção desses sons.

Fujisaki e Kawashima 1969, 1970 incluem nos testes perceptuais uma condição onde a vogal interferente /a/ imediatamente sucede o primeiro elemento do par vocálico a ser discriminado - no caso, vogais sintéticas no contínuo /i/-/e/. Essa condição, se comparada à condição sem vogal interferente, provoca um sensível aumento da percepção categorial, isto é, a fronteira categorial se torna mais nítida para o sujeito, concomitantemente à uma redução da capacidade de discriminar estímulos do mesmo lado do contínuo. Repp *et al.* 1979 obtêm resultados semelhantes para o contínuo /i/-/I/-/ɛ/, usando a vogal /y/ como contexto interferente.

A similaridade entre a vogal interferente e os estímulos-teste também influi nas respostas. Inserindo dois contextos

interferentes distintos, /a/ ou /ɛ/, entre segmentos extraídos do contínuo /i/-/ɪ/, Pisoni 1975 verifica que, embora ocorra a esperada diminuição na performance com o contexto /a/, os escores de discriminação foram significativamente mais baixos quando a vogal interferente era /ɛ/, um som foneticamente mais próximo dos estímulos a serem comparados.

O nível e a seletividade da atenção também podem afetar consideravelmente o padrão de processamento. Estímulos consonantais fisicamente muito próximos que, em condições normais, não seriam discriminados, podem vir a sê-lo se os sujeitos são orientados no sentido de aumentar seu nível de atenção (Barclay 1972).

Lane 1965 demonstra que a capacidade de perceber categorialmente pode, em certa medida, ser adquirida através de treinamento. Ele observa que, após um treinamento simples, os sujeitos produzem respostas categoriais no julgamento de sons complexos não lingüísticos (padrões espectrográficos de fala reproduzidos no sentido inverso) [9].

O conjunto de resultados da série de experimentos relacionados à memória, atenção seletiva e treinamento indica que os mecanismos perceptuais são bastante flexíveis. O peso relativo de cada modo perceptual é continuamente reajustado de acordo com as propriedades do estímulo e a especificidade da tarefa. Uma baixa razão sinal/ruído, ou um corte na informação espectral, tal como ocorre na conversação telefônica, tende a

exigir um maior grau de atenção seletiva orientada para certas características do sinal acústico que, de outro modo seriam irrelevantes [10]. Por outro lado, se o conjunto de pistas momentaneamente acessível é mais rico, como no caso da interação face-a-face, quando se torna possível utilizar informação visual suplementar através da observação direta dos movimentos de alguns articuladores, pode haver um certo relaxamento dos mecanismos auditivos de diferenciação fina.

O mecanismo perceptual parece naturalmente predisposto a integrar vários tipos de pista, sejam fatores absolutos ou relacionais. Na percepção de vogais, especificamente, tanto propriedades intrínsecas do segmento (padrão de formantes, por exemplo),quanto aspectos extrínsecos ao segmento (relação com ambiente acústico precedente, efeitos do contexto consonantal, etc.) podem estar em jogo durante o processamento. Não se trata, entretanto, de uma redução a uma fórmula simples do tipo "quanto mais pistas melhor" - embora isso seja parcialmente verdadeiro -, mas sim que certas pistas talvez tenham efeito perceptual ótimo apenas no contexto de outras pistas, e vice-versa, ocorrendo uma espécie de reforço mútuo que facilita a decodificação. É preciso considerar também que o ouvinte, além da informação física codificada no sinal acústico, possui um conhecimento mais ou menos detalhado da estrutura da língua (em vários níveis: fonológico, lexical, etc.) que é ativamente usado em conjunção com o *input* sensorial de modo a desenvolver uma representação

eficiente da mensagem.(v. seção 8)

Uma das limitações inerentes à situação laboratorial, em testes perceptuais, diz respeito à impossibilidade prática de controlar todas essas fontes de informação que, durante o processamento de fala real, estão ao alcance do ouvinte. Por outro lado, paradoxalmente, é exatamente a possibilidade de isolar artificialmente certos parâmetros que permite, nos testes, um acesso mais direto a mecanismos perceptuais específicos: o que seria da ciência biológica se as reações fisiológicas observadas *in vitro* não pudessem ser consideradas?

Os testes perceptuais têm, afinal, uma história a contar, e tanto mais rica será a informação que deles extraímos quanto menos ignorarmos as restrições peculiares a cada *design* experimental. Os experimentos focalizando o fenômeno da percepção categorial indicam que (apesar de eventuais reorientações do padrão de processamento devidas a fatores como deficit de memória, nível de atenção seletiva, e treinamento) as vogais estão, em geral, mais estreitamente associadas a processos perceptuais de natureza contínua (ou analógica), em contraposição ao modo predominantemente categorial (ou digital) típico de (pelo menos alguns) segmentos consonantais.Para avaliar corretamente esses resultados é preciso considerar, no entanto, que os segmentos utilizados nesse tipo de teste são sons sintetizados, e que os estímulos intermediários são meras interpolações artificialmente produzidas ao longo do contínuo entre os

extremos; embora esses estímulos extremos simulem estruturas de formantes equivalentes aos sons de fala real, as interpolações linearmente espaçadas entre eles não refletem obrigatoriamente padrões espectrais com base articulatória viável (Cf. Lindblom e Sundberg 1971). Outro ponto a considerar é que, embora a razão principal para a percepção mais contínua dessas vogais sintetizadas isoladas seja, sem dúvida, sua alta discriminabilidade inerente e boa retenção auditiva, não é menos verdade que a homogeneidade acústica que confere essas vantagens perceptuais não é muito típica das vogais tal como ocorrem na fala natural. Não é muito fácil estimar até que ponto vogais isoladas sintetizadas perdem sua naturalidade, já que isso depende em grande parte da qualidade da síntese e do contraste vocálico focalizado, mas é plausível supor que, pelo menos em alguns casos, essas vogais artificiais cheguem a sofrer alguma deterioração maior quanto à sua qualidade como som de fala, o que poderia desestimular modos mais fonéticos de processamento.

O fato é que vogais estáveis (sintetizadas ou não) omitem uma série de pistas presentes na situação de fala real. Pesquisas recentes têm ressaltado a importância de outras pistas que não a estrutura de formantes, tais como: freqüência fundamental (Syrdal e Gopal 1986; Traunmüller 1988), duração intrínseca (Crystal e House 1988), o aspecto transicional às margens da seção vocálica (Strange 1989), etc.(O exame mais detalhado da eficácia de cada uma dessas pistas será realizado mais adiante (v. seções 5 e 6).

2
MACRO-ESTRUTURA DOS SISTEMAS VOCÁLICOS

Mesmo admitindo que a maioria dos testes psico-perceptuais não reflete toda a realidade da fala natural, é inegável que o conjunto de experimentos anteriormente examinado (v. seção 1) indica que a percepção de vogais aproxima-se mais de um modo "auditivo" de processamento, ou seja, qualidades vocálicas acusticamente muito próximas podem ser discriminadas mesmo que não representem pontos do sistema fonológico na língua nativa do ouvinte. Esse estado de coisas nos leva a especular que - pelo menos no que diz respeito à capacidade de análise fina do sistema perceptual - seria razoável esperar que os sistemas fonológicos naturais possuíssem um grande número de vogais. O exame de alguns inventários representativos, descrevendo um grande número de sistemas fonológicos de várias línguas do mundo, revela, contudo, uma realidade bem diferente.

Maddieson 1986, em um estudo estatístico sobre a base de dados UPSID (UCLA Phonological Segment Inventory Database), contendo um total de 317 línguas diferentes, relata que, embora o número total de segmentos (vogais + consoantes) possa variar bastante - de um mínimo de 11 (ROTOKAS e MURA) um máximo de 141 (XU) -, 70% das línguas analisadas possuem de 20 a 37 segmentos. A média do número de vogais em cada sistema é, no entanto,

consideravelmente menor do que a do número de consoantes: média do número de vogais = 8.7, do número de consoantes = 22.8. Ainda que o número absoluto de vogais tenda a ser maior em sistemas com grande número de elementos, a razão vogais/consoantes diminui ainda mais nos sistemas mais extensos: é como se existisse uma restrição maior ao crescimento dos sistemas vocálicos, se comparados aos sistemas consonantais.

Existem, é claro, casos excepcionais em ambas as direções; as línguas **HAIDA** (46 cons., 3 vog.), **JAQARU** (38 cons., 3 vog.) e **BURUSHASKI** (38 cons., 5 vog.) são exemplos de sistemas com grande desequilíbrio em favor das consoantes. Pequenos sistemas consonantais conjugados a sistemas vocálicos extensos são os menos prováveis, embora pareça haver uma tendência nessa direção nas línguas da Nova Guiné, mesmo assim a razão vogal/consoante mal se aproxima de 1: **PAWAIAN** (10 cons., 12 vog.), **DARIBI** (13 cons., 10 vog.) e **FASU** (11 cons., 10 vog.) são exemplos típicos.

Crothers 1978 (*apud* Lindblom 1986) faz um levantamento dos sistemas vocálicos de 209 línguas, e classifica os segmentos em 37 categorias:

anterior	central	posterior	
i ü	ɨ u̇	ɯ u	alta
I Ü	ɨ̈ U̇	ɯ̈ U	alta-baixa
e ö̞	ɘ ɵ̞	ɘ̈ o	média-baixa
E ö	ə Ö	ɘ̈ O	média
ɛ ɜ	ɛ̇	ʌ ɔ	média-baixa
æ	æ	ɐ	baixa-alta
ä	a ä	ɑ ɒ	baixa

(a segunda coluna de cada grupo corresponde às arredondadas)

Uma transcrição larga permite visualizar melhor a distribuição tipológica dos sistemas. Desse modo, na tabela 2.1 (adaptada de Lindblom 1986), /I/, /ɯ/ e /ü/, por exemplo, são representados como /i/, /u/ e /i/, respectivamente [1].

nº de vogais no sistema	% de línguas	qualidades vocálicas (transcrição larga)
3	13.2	i a u
4	7.5	i ɛ a u
	5.1	i ɨ a u
5	31.6	i ɛ a ɔ u
	2.9	i ɨ ɛ a o
6	16.6	i ɨ ɛ a ɔ u
	4.0	i e ɛ ɔ o u
7	8.0	i ɨ e a ə o u
	6.3.	i e ɛ a ɔ o u
9	4.0	i ɨ e ɛ a ə ɔ o u

A norma, segundo os dados de Crothers 1978, é de sistemas com 5-6 vogais; cerca de 80% das línguas possuem 6 ou menos vogais básicas, uma média ainda mais baixa que a relatada por Maddieson 1986. O mais notável nesses inventários, entretanto, não é o número absoluto de itens em cada sistema, mas sim, certas regularidades e hierarquias que podemos observar. Praticamente todos os sistemas - independentemente do número total de elementos - possuem o triângulo vocálico /i/-/a/-/u/ (ou vogais muito próximas dessas qualidades prototípicas). É como se os sistemas crescessem "em torno" dessas vogais polares. Há, em geral, um favorecimento de vogais mais periféricas; o número de vogais interiores, como por exemplo /ü/, /ɨ/ e /ʊ/, nunca excede a quantidade de vogais periféricas (anteriores ou posteriores). Vogais médias só ocorrem se existem também vogais altas e baixas. Todas as línguas têm vogais altas (pelo menos uma). Em geral, o número de distinções na dimensão **ALTURA** é igual ou maior que o número de distinções na dimensão **ANTERIORIDADE** sendo que as vogais anteriores tendem a possuir mais subdivisões de **ALTURA** do que as posteriores.

Os testes perceptuais focalizando a discriminação de pares vocálicos, como já vimos, não indicam qualquer predisposição do mecanismo auditivo-perceptual para o favorecimento de faixas de frequência particulares ou combinações de formantes com um certo padrão. As assimetrias observadas na distribuição das qualidades

não devem, pois, ser motivadas diretamente por algum tipo de restrição do aparelho perceptual, pelo menos no que diz respeito ao processamento em níveis mais periféricos. É preciso considerar que a distribuição não homogênea das qualidades vocálicas pode estar, pelo menos parcialmente, condicionada a certas restrições de ordem articulatória: na fala, afinal, não importa apenas aquilo que se pode ouvir, mas também aquilo que se é capaz de dizer.

É oportuno, nesse momento, retornar às formulações de Stevens 1972 (v.seção 1). Como vimos, Stevens sugere que, no interior do espaço articulatório, existem certas regiões "platô", onde pequenas alterações dos articuladores não modificam significativamente a saída acústica; a linguagem, diz ele, "seeks out these regions, as it were, and from them assembles an inventory of phonetic elements that are used to form the code for communication by language". As vogais /i/, /a/ e /u/ estão associadas a posições articulatórias extremas, e representam típicos "platôs", segundo a concepção de Stevens 1972; as condições articulatórias peculiares desses segmentos parecem, de fato, estar relacionadas à sua presença constante nos sistemas fonológicos das línguas do mundo. A posição privilegiada das vogais /i/, /a/ e /u/ dentro do sistema pode ser avaliada através do estudo dos processos de aquisição. Olmsted 1971 (*apud* Lieberman 1984:204) analisou a estrutura fonética das palavras produzidas por 100 crianças entre 15 e 54 meses.

Observou-se que as vogais /i/, /u/ e /a/ eram menos sujeitas a erros (substituição por outra qualidade) do que as outras vogais (com exceção de /ə/ que, em inglês é diferenciada por duração das outras vogais "não-quânticas"). Essas vogais são também adquiridas precocemente em relação às demais: já aos três meses /i/, /a/ e /u/ aparecem (Buhr 1980).

Certas características mais globais da geometria do trato explicam também algumas assimetrias observadas nos inventários de Crothers 1978 e Maddieson 1986. É preciso levar em conta que os articuladores não são infinitamente elásticos, nem totalmente independentes. É evidente que esses articuladores possuem grande mobilidade e é mesmo possível realizar configurações envolvendo um razoável antagonismo muscular (v.p.ex. Lindblom *et al.* 1977), mas a economia interna do sistema de produção tende a favorecer movimentos nos quais se busca uma otimização dos recursos articulatórios. O exame de radiografias laterais do trato revela, por exemplo, que a posição da mandíbula na produção de vogais, é otimizada, no sentido em que coopera com a língua, prevenindo assim uma excessiva deformação desta (Lindblom e Sundberg 1971).

Condições Sensório-Motoras

É provável que mecanismos sensoriais distintos estejam subjacentes a diferentes articuladores ou a diferentes tipos de movimento de cada articulador. Sabe-se que a extensão da área

cortical associada a um determinado órgão está relacionada ao grau de precisão e de controle fino exigido pelo tipo de movimento executado por esse órgão. Através da estimulação elétrica de pontos específicos do cérebro é possível demonstrar, por exemplo, que os movimentos do polegar podem ser provocados a partir de áreas corticais bem maiores do que as associadas ao movimento da perna (Kandel e Schwartz 1981: *apud* Lindblom e Lubker 1985:169). Nesse contexto, interessaria ao foneticista verificar em que medida certos aspectos do trato vocal são favorecidos em termos de representação cortical e - o mais importante - se a possível vantagem sensorial de alguma região ou tipo de movimento afeta de algum modo a macro-estrutura dos sistemas fonológicos.

Essa indagação esbarra, entretanto, em algumas dificuldades. O percurso percorrido por um estímulo entre os órgãos sensórios periféricos e o sistema nervoso central envolve uma série de transformações psico-neurofisiológicas cujas particularidades o estágio atual da pesquisa não permite compreender inteiramente. Nesse delicado campo de pesquisa, tanto complicações técnicas quanto inevitáveis embaraços de ordem ética costumam se interpor entre o pesquisador e seu objeto, dificultando o acesso direto aos mecanismos neurológicos envolvidos nas transformações. O quadro se complica ainda mais quando se trata do estudo da motricidade associada à linguagem em geral, uma ação prenhe de complexas derivações funcionais e perpassada por uma dimensão

simbólica que a diferencia radicalmente de outros comportamentos como andar, mastigar, manipular objetos, etc. - embora algumas dessas ações possam, eventualmente, exigir um controle motor ainda mais preciso do que a fala.

O excesso de fatores complicadores aconselha que as questões mais importantes sejam reformuladas em termos mais específicos e que se desenvolva experimentos cujos resultados permitam interpretar, por via indireta, as relações e transformações existentes entre os diversos níveis de representação. Uma importante contribuição nesse sentido é oferecida pela série de experimentos sensoriais conduzidos por Lindblom e Lubker 1985 (LL85, daqui em diante). Os autores partem da premissa de que as fibras nervosas aferentes não são "high fidelity recorders" (Mountcastle 1975), pois atuam, em geral, acentuando certos traços do estímulo e negligenciando outros. Dessa forma, a informação física sobre os movimentos articulatórios, tal como fornecida por medidas de laboratório, podem ser - e provavelmente são - diferentes da informação usada no cérebro ao gerar e monitorar esses movimentos. A questão básica aqui é tentar estabelecer uma transformação que relacione, quantitativamente, a escala física com uma escala sensorialmente relevante.

No domínio auditivo, várias escalas, derivadas de experimentos perceptuais, já foram sugeridas para descrever a variação de freqüencia em termos subjetivos (Mel, Bark, Koenig, etc.). No domínio articulatório, entretanto, não existe uma

escala que permita, ainda que tentativamente, uma calibração neuropsicológica válida ds movimentos da fala. De modo a estabelecer uma relação quantitativa entre os planos subjetivo e objetivo, LL85 recorrem às sugestões de Stevens 1975. Stevens propõe um princípio psicofísico genérico estabelecendo que a relação entre a magnitude física do estímulo e a sensação subjetiva (seja qual for a dimensão sensorial envolvida) pode ser descrita por uma função exponencial expressa por uma fórmula do tipo:

$$y = lx^k$$

onde y = magnitude da sensação

l = constante

x = magnitude do estímulo

k = expoente

O estudo de LL85 restringe-se aos movimentos da mandíbula e da língua em tarefas razoavelmente simples. Os testes com sons lingüísticos baseiam-se na produção de sílabas /jV/ com a lacuna vocálica sendo preenchida pelo conjunto de vogais periféricas /i, e,ɛ,æ,a,ɑ,o,u/. Como referencial para todos os sujeitos (n=7) foi usada a sílaba /jɛ/, sendo o sujeito requisitado a atribuir um valor arbitrário relativo à posição, ou ao movimento, requerido para a produção dessa sílaba. Cada uma das sílabas com as demais vogais era então produzida, em ordem aleatória, cabendo ao

sujeito atribuir números que avaliassem a posição ou o movimento dessas produções em relação ao valor anteriormente atribuído ao referencial /jɛ/.

Nos testes focalizando o movimento lingual os sujeitos avaliam os deslocamentos em duas condições: "língua como um todo" e "ponto específico da língua". No segundo caso um fino estilete metálico estimulava a língua do sujeito em um ponto a aproximadamente 2 cm do ápice; a estimulação era repetida periodicamente durante o teste e os sujeitos foram instruídos a se concentrar nesse ponto para suas estimativas do deslocamento da língua.

Todos os sujeitos de LL85 apresentaram altas correlações (média de .91) entre as estimativas subjetivas e os deslocamentos medidos do movimento mandibular. Para os movimentos da língua as correlações são consideravelmente mais baixas (média de .62) e variam bastante de sujeito para sujeito [2]. Verificou-se também uma alta correlação entre as estimativas subjetivas feitas com base na "língua como um todo" e em "ponto específico da língua", indicando que os sujeitos, subconscientemente, julgam o deslocamento da estrutura muscular inteira apenas com base no deslocamento da região anterior da língua.

Os resultados de LL85 sugerem que: (a) os movimentos mandibulares são avaliados subjetivamente com maior precisão do que os movimentos da língua, e (b) a região anterior da língua parece ser sensorialmente privilegiada.

Ora, esses fatos são consistentes com as assimetrias distribucionais que observamos nos inventários de Crothers 1978 e Maddieson 1986, onde se verifica uma tendência ao favorecimento de distinções fonéticas envolvendo a região **ANTERIOR** do trato, assim como um maior número de qualidades distribuídas ao longo da dimensão **ABERTURA**. É razoável supor, portanto, que certas restrições sensório-motoras condicionem, ao menos parcialmente, a macro-estrutura dos sistemas vocálicos.Hardcastle 1970 já havia, na verdade, hipotetizado uma relação direta entre a diversidade dos sons da fala - não apenas as vogais - e a variedade dos mecanismos sensórios a longo do trato; ele notara que as línguas privilegiam a parte anterior do trato, e dá como exemplo a distribuição não uniforme do ponto de articulação na série de fricativas não-sonoras, com base na tabela da Associação Fonética Internacional.

O Espaço Articulatório

Os experimentos de LL85 nos obrigam a revisar certas noções tradicionais sobre o espaço articulatório. Examinemos mas de perto os movimentos investigados nesse estudo, sob uma ótica mais quantitativa. A figura 2.1 mostra graficamente as relações entre o movimento mandibular e o movimento da língua. O gráfico superior representa as medidas físicas objetivas da abertura mandibular *vs.* deslocamento da língua medido em relação à

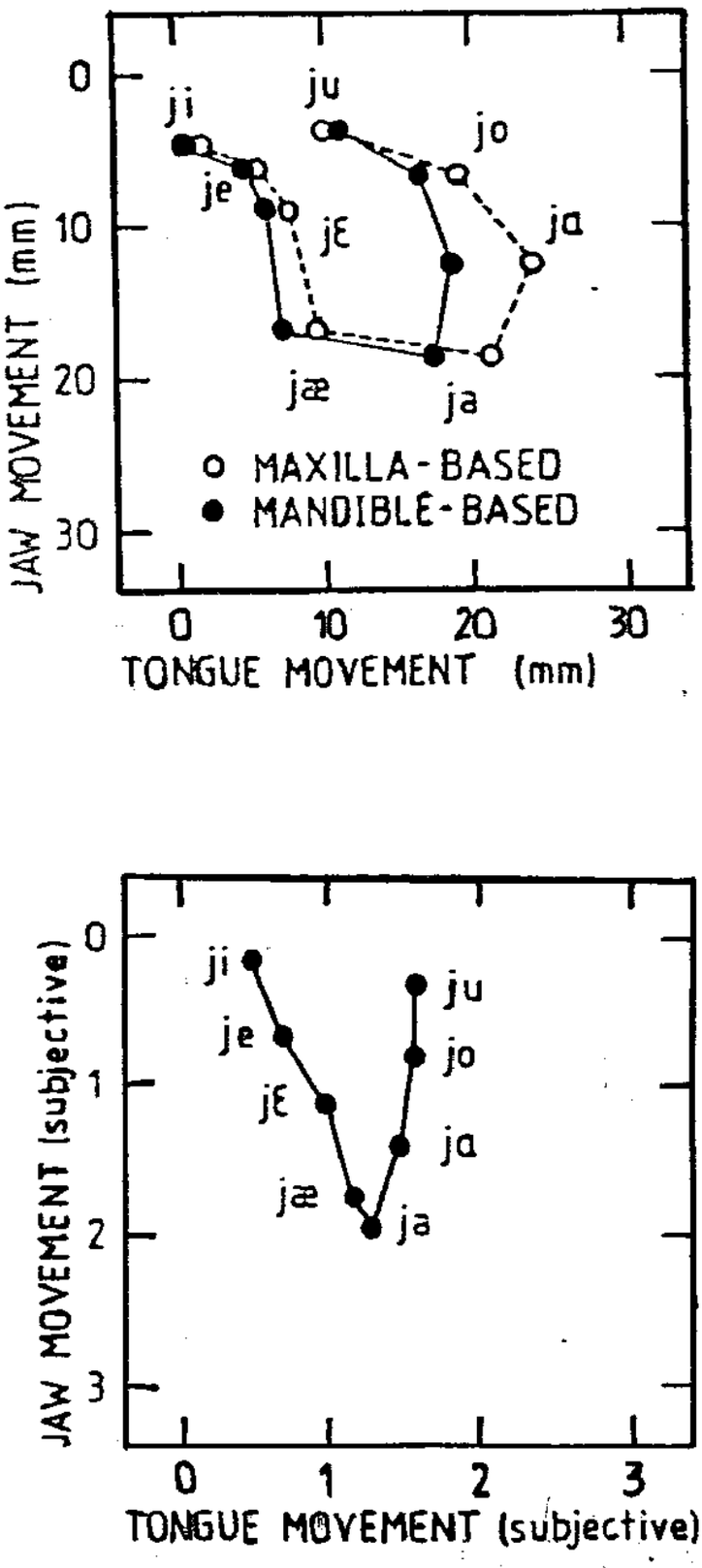

JAW MOVEMENT (mm)
TONGUE MOVEMENT (mm)
0
10
20
30
ji
ju
je
jo
jɛ
jɑ
jæ
ja
MAXILLA-BASED
MANDIBLE-BASED
JAW MOVEMENT (subjective)
TONGUE MOVEMENT (subjective)
1
2
3

FIGURA 2.1

mandíbula e ao maxilar. O gráfico inferior cruza os mesmos movimentos mas baseia-se nas estimativas subjetivas [3].

A dimensão "abertura", definida tanto pela série anterior /i,e,ɛ,æ/, quanto pela posterior /u,o, ,a/ aparece mais estendida do que as distâncias articulatórias no eixo antero-posterior, definidas, por exemplo, entre /i/-/u/ ou /ɛ/-/o/; parece haver mais "espaço" para distinções na dimensão abertura do que para contrastes anterior/posterior.

Até que ponto essas representações espelham uma realidade psicológica? Se há mesmo um estreitamento no eixo horizontal motivado por restrições sensório-motoras, devemos esperar uma menor incidência, nas línguas do mundo, de vogais com altura máxima, tais como /y,ü,ɨ,ø/. De fato, consultando Crothers 1978, observamos que apenas 24.2% do total das línguas analisadas possuem uma vogal entre /i/ e /u/, enquanto a grande maioria - 72.7% - não tem qualquer vogal nesse espaço. Inversamente, a ausência de qualidades vocálicas intercaladas entre /i/-/a/ e entre /a/-/u/ constitui a exceção, não a regra: apenas 16.7% e 26.8% das línguas não possuem vogais intercaladas nas séries anterior e posterior, respectivamente [4].

Essa realidade contrasta com representações articulatórias tradicionais, como por exemplo, o quadrilátero das vogais cardinais proposto por Daniel Jones em 1917 (Abercrombie 1967:152). Na verdade, nunca ficou muito claro qual o critério usado por Jones para definir seu esquema: auditivo,

articulatório, ou ambos? Mesmo os defensores de Jones não conseguem entrar em acordo quanto à motivação principal da distribuição das vogais cardinais. Abercrombie 1967 diz que essas vogais "are established not in articulatory but primarily in auditory terms" (pg. 153) "...the eight cardinal vowels, therefore, are eight equally-spaced auditory points forming a kind of scale (grifo do autor) of vowel quality" (pg. 154). Mais adiante, no entanto, o mesmo Abercrombie afirma: "The fact is that they all have both an articulatory and an auditory aspect... the eight cardinal vowels, therefore, in whatever way they were established, are used as articulatory postures for descritive purposes" (pg. 156) (grifo nosso). Outro defensor do modelo de Jones admite que a equidistância auditiva "may be a property ascribed to cardinal vowels solely by their originator" (Ladefoged 1967:99).

O esquema de Jones não pretende, é claro, mapear fielmente o espaço articulatório, tratando-se antes de uma representação estilizada desse espaço. Mas seria errôneo, por outro lado, concebê-lo como totalmente abstrato, já que para sua elaboração, Jones parece ter-se baseado em princípios relativamente concretos - pelo menos do ponto de vista articulatório - cujos reflexos devem continuar presentes, diagramaticamente, na representação gráfica final, tal como a conhecemos (da mesma forma como a projeção bidimensional de uma superfície esférica - um mapa geográfico convencional, por exemplo - não altera

substancialmente as relações internas de distância). Se medirmos os lados do trapézio das vogais cardinais encontraremos as seguintes quantidades em uma escala arbitrária:

distância V_1 - V_8 (série superior) = 5
distância V_1 - V_4 (série anterior) = 4.5
distância V_5 - V_8 (série posterior)= 4

(com base no diagrama em Jones 1918:37)

A distorção em relação aos resultados de LL85 é evidente: a **MAIOR** distância aqui - a série superior - é a **MENOR** segundo as medidas objetivas e subjetivas de LL85 [5].

O Espaço Acústico-Perceptual

Os resultados de LL85, suportados pelos dados de Maddieson 1986 e Crothers 1978, indicam uma prevalência das distinções na dimensão "abertura" no espaço articulatório subjetivo. Em que medida, porém, essa tendência encontra paralelo com o espaço acústico-perceptual? Examinemos alguns dados do Português Brasileiro (PB), extraídos de Behlau 1984 [6].

A tabela 2.2 mostra as freqüencias médias dos dois primeiros formantes das sete vogais orais do PB para três grupos de falantes (homens, mulheres e crianças).

		/a/	/ɛ/	/e/	/i/	/ɔ/	/o/	/u/	VF_1	VF_2	VF_2-F_1
homens	F1	807	699	563	398	715	558	400	2.02	2.18	4.39
	F2	1440	2045	2339	2456	1201	1122	1182			
mulheres	F1	956	769	628	425	803	595	462	2.24	2.38	4.98
	F2	1634	2480	2712	2984	1317	1250	1290			
crianças	F1	1086	902	698	465	913	682	505	2.33	2.45	5.91
	F2	2873	3243	3637	3980	2793	2823	2667			

TABELA 2.2

Expressando as variações de cada um desses formantes em termos da razão entre os valores máximos e mínimos para cada grupo, obteremos os parâmetros VF_1 e VF_2, também representados na tabela. Em todos os grupos VF_2 é maior que VF_1. Tradicionalmente, F2 é considerado um correlato direto do movimento ântero-posterior da língua, e F1 um parâmetro primariamente controlado pelo grau de abertura mandibular (Fant 1960; Lindblom e Sundberg 1971). De modo a checar essa relação, efetuamos um teste de correlação, (método de Pearson) a partir dos dados de Behlau 1984. De modo a permitir uma comparação quantitativa, foram atribuídos valores *dummy* para cada vogal, tanto na dimensão "abertura" (ou "altura") quanto na dimensão "anterioridade" [7]:

	altura	anterioridade
/a/	-1	0
/ɛ/	0	1
/e/	0	1
/i/	1	1
/ɔ/	0	-1
/o/	0	-1
/u/	1	-1

As correlações para cada subgrupo de falantes foram:

	F1 * altura	F2 * anterioridade
	(r)	(r)
homens	-.84	.91
mulheres	-.81	.92
crianças	-.80	.93

(em todos os casos $p < 0.0001$)

Como se pode observar, as altas correlações indicam que existe, de fato, uma relação estreita entre F1 e F2 e as dimensões articulatórias "altura" e "anterioridade", respectivamente [8].

Voltando à tabela 2.2 verificamos que , em termos acústico-perceptuais, a dimensão "abertura" cobre uma faixa **MENOR** que a dimensão "anterioridade" (VF_2 maior que VF_1). Se, em lugar de F2, utilizarmos a diferença entre o segundo e o primeiro formante - um outro parâmetro freqüentemente associado às distinções antero-

posteriores (v.p.ex. Ladefoged 1975) - a variação na dimensão "abertura" torna-se, comparativamente, ainda menor, como podemos constatar comparando na tabela 2.2 os valores de VF_2-F_1 e VF_1.

Ora, esses resultados são exatamente opostos aqueles obtidos por LL85 em relação à percepção do movimento articulatório; na representação acústico-perceptual, a dimensão "anterioridade" (expressa por VF_2 ou VF_2-F_1) é que parece favorecida, contrariando assim as tendências que observamos nos sistemas vocálicos em geral. Essa discrepância parece indicar que há algo errado com nossa representação acústica do espaço vocálico.

Um trabalho que nos ajudará a esclarecer esse ponto é o de Liljencrants e Lindblom 1972. Eles tentam desenvolver um modelo numérico cujo objetivo é predizer a macro-estrutura fonética dos sistemas vocálicos. A hipótese inicial assume que a distribuição das qualidades vocálicas obedece a um princípio de contrastividade máxima que pode ser definida com base em um espaço acústico-perceptual. A partir do modelo de Lindblom e Sundberg 1971, que reflete os graus de liberdade naturais do trato vocal, é possível derivar um espaço acústico que represente os limites reais de uma vogal articulatoriamente POSSíVEL. Para facilitar a computação das distâncias, esse espaço foi representado bi-dimensionalmente, em termos de M1 e M2', ou seja, o primeiro formante e o segundo formante corrigido em relação ao terceiro formante, expressos em unidades MEL [9].

Os contrastes perceptuais foram definidos por Liljencrants e Lindblom 1972 como a distância euclidiana entre dois pontos do sistema. Dessa forma, a distância perceptual D_{ij} entre as duas vogais hipotéticas i e j será expressa matematicamente como:

$$D_{ij} = [(M1_i - M1_j)^2 + (M2'_i - M2'_j)^2]^{1/2}$$

Segundo a hipótese inicial, os sistemas vocálicos tendem a maximizar os contrastes perceptuais entre os elementos. Desse modo, o programa de computador deve encontrar configurações para cada sistema de n vogais, onde o somatório das distâncias entre todos os pares de elementos seja máximo, ou seja:

$$\sum_{1}^{m} D_k \text{ ---> maximizado}$$

onde: D_k é a distância euclidiana entre o k-gésimo par de vogais

m = número de pares de vogais no sistema (m=n(n-1)/2, onde n é o número total de vogais)

O estudo de Liljencrants e Lindblom informa mais pelos erros do que pelos acertos nas previsões. O algoritmo funciona bastante bem para sistemas de até 6 vogais, gerando previsões compatíveis com a realidade das línguas naturais. A partir de 7 vogais, no

entanto, o modelo produz um número de vogais altas maior do que a expectativa. Para sistemas de 7 e 8 vogais, o algoritmo prevê duas vogais intercaladas entre /i/ e /u/, e para sistemas de 9 a 12 vogais, 3 qualidades são geradas no mesmo intervalo.

É claro que não é impossível encontrar línguas que eventualmente correspondam a essas previsões. Mas são, sem dúvida, casos excepcionais. No levantamento de Crothers 1978, por exemplo, apenas 1% das línguas observadas possui duas vogais na série superior, entre /i/ e /u/, e nenhuma possui três vogais nessa mesma posição. Liljencrants e Lindblom 1972, com base em dados de Trubetzkoy 1929, Hockett 1955 e Sedlak 1969, também não encontraram exemplos com três qualidades intercaladas entre /i/ e /u/.

Os erros de previsão do algoritmo de Liljencrants e Lindblom, antes de mais nada, reforçam nossa suspeita de que a contrastividade perceptual não é o único determinante da macro-estrutura dos sistemas fonéticos em geral e dos sistemas vocálicos em especial. Maddieson 1986 observa, com procedência, que se "maximização de distintividade" fosse o princípio único segundo o qual os sistemas vocálicos são construídos, deveríamos esperar que o conjunto de vogais mais freqüente não fosse /i,e,a, o,u/, mas sim algo como /i,ẽ,a̤,o̰,uˤ/, onde cada vogal não difere das demais apenas quanto à qualidade, mas também por ser plana, nasalizada, aspirada, laringalizada e faringalizada. Também com as consoantes ocorre algo semelhante. Cliques, lembra Maddieson,

são altamente "salientes", no entanto, as poucas línguas que possuem esse tipo de som usam séries múltiplas de cliques, em vez de explorar esse traço de modo a realizar oposições mais contrastivas entre, por exemplo, um clique dental e uma plosiva velar. O que parece ocorrer é que certas dimensões de contraste são usadas perferencialmente em relação a outras, de uma maneira que não está relacionda diretamente à "saliência fonética".

É provável que a organização articulatória de uma língua obedeça a um princípio genérico de "menor esforço" para produzir as oposições dentro do sistema (v. p.ex. a sinergia articulatória no sistema conjugado língua/mandíbula :Lindblom e Sundberg 1971). Tal princípio poderia talvez explicar a prevalência das distinções na dimensão altura vocálica; Liljencrants e Lindblom 1972 sugerem que, se as vogais forem produzidas com a língua mais próxima da configuração "neutra" - isto é, com menor tensão muscular - haverá um ENCOLHIMENTO do espaço acústico modificando principalmente o segundo formante (v. Lindblom e Sundberg 1971: 1176). A conseqüência é que as vogais difeririam principalmente em termos de F1, ou seja, em grau de abertura. O favorecimento de F1 nos contrastes vocálicos oferece algumas vantagens perceptuais; Lindblom 1975 (*apud* Lindblom 1986) argumenta que, se os sistemas vocálicos desenvolveram margens de segurança garantindo a diferenciação perceptual na comunicação sob condições adversas de ruído (a situação mais freqüente da vida real), seria de esperar que houvesse efetivamente uma maior

exploração dos contrastes baseados na variação de F1, já que essa ressonância é mais intensa e, portanto, mais resistente a qualquer tipo de interferência ambiental [10].

Outra limitação do algoritmo usado por Liljencrants e Lindblom 1972 diz respeito à desconsideração de fatores relacionados ao contexto fonético; os cálculos foram efetuados com base apenas nas propriedades intrínsecas de cada ponto do sistema. Os próprios autores reconhecem que o modelo poderia ser melhorado se fosse levado em conta que os princípios de diferenciação perceptual e de redução/otimização da energia articulatória operam não só paradigmaticamente como também sintagmaticamente. É possível, dizem eles, que como conseqüência de condições sintagmáticas para a otimização da estrutura silábica, haja um favorecimento de segmentos vocálicos e consonantais que maximizem as distâncias perceptuais não apenas dentro de cada classe separadamente (vogal ou consoante), mas também no interior do conjunto inteiro. Esse ponto de vista assume que mudanças espectrais maiores tendem a facilitar o processamento perceptual e o reconhecimento. Dessa forma, não basta avaliar os contrastes perceptuais realizáveis exclusivamente dentro do sistema vocálico, mas também considerar os efeitos coarticulatórios; é possível, por exemplo, que vogais e consoantes sejam escolhidas de tal forma que a extensão das transições em combinações CV ou VC arbitrárias seja maximizada (Lindblom *et al.* 1984). Um exemplo extremo desse princípio pode

ser observado em algumas línguas Caucasianas, cujo sistema vocálico pode se restringir a apenas uma ou duas oposições, mas, compensatoriamente, possuem um sistema consonantal extremamente rico (Halle 1970) [11].

A importância dos aspectos transicionais na percepção de vogais tem sido enfatizada em vários estudos experimentais (Verbrugge *et al.* 1976; Strange *et al.* 1976; Gottfried e Strange 1980). Esses estudos verificaram que vogais em contextos CVC são identificadas mais acuradamente do que vogais isoladas produzidas pelo mesmo grupo de falantes, apesar de uma considerável ambigüidade nas freqüencias dos formantes (ver, no entanto, Macchi 1980, Assmann *et al.* 1982 e Rakerd *et al.* 1984 para resultados um pouco diferentes)(a importância dos aspectos dinâmicos na caracterização de vogais será examinada mais cuidadosamente na seção 6).

3

CLASSIFICAÇÃO AUTOMÁTICA BASEADA NOS FORMANTES: O PROGRAMA DISCRIM

Os erros de previsão do modelo de Liljencrants e Lindblom 1972 (v. seção 2) deixam entrever que uma representação exclusivamente baseada nas freqüências dos formantes não reflete adequadamente a realidade perceptual. Tivemos a oportunidade de constatar esse fato através de algumas análises estatísticas a partir de dados de Behlau 1984. Nesse trabalho foram medidos os três primeiros formantes das vogais orais e nasais do Português Brasileiro /a,ɛ,e,i,ɔ,o,u,ɐ̃,ẽ,ĩ,õ,ũ/, no contexto pVs, produzidas por 90 falantes divididos em tres grupos: homens, mulheres e crianças.

Para os cálculos estatísticos foram utilizados programas integrantes do sistema SAS instalado no computador VAX-785 na Universidade Estadual de Campinas. O algoritmo de classificação que se mostrou mais eficiente para os dados em questão foi o programa DISCRIM. Esse procedimento calcula funções discriminantes lineares para classificar variáveis quantitativas. O critério de classificação baseia-se nas matrizes de covariância no interior de cada classe: cada observação a ser testada é colocada na classe da qual tem a menor distância no espaço multidimensional definido pelas variáveis numéricas. É importante

observar que esse algoritmo, na verdade, não classifica propriamente as observações, mas antes testa a probabilidade de uma determinada observação pertencer a uma das categorias pré-estabelecidas. O output do programa pode ser expresso em termos do percentual de acertos, isto é, um valor que indique, proporcionalmente, a quantidade de itens corretamente classificados (para informações mais detalhadas consultar SAS User's Guide: Statistics cap. 16).

A figura 3.1 mostra o percentual de acertos obtidos por DISCRIM com base nas freqüências dos três primeiros formantes, expressas em Hz e em uma transformação logarítmica ($\log_{10}(F_n)$). A linha superior do gráfico mostra, também em termos percentuais, os acertos dos sujeitos no teste perceptual de identificação realizado por Behlau 1984. Não foram incluídas as vogais nasais porque cálculos preliminares verificaram que, com base apenas nas freqüências dos três primeiros formantes, é praticamente impossível distinguir, por meio de DISCRIM, o traço nasalidade. A classificação, dessa forma, acaba sendo uma repartição quase aleatória entre as categorias orais e nasais equivalentes [1]. É importante observar que, também no teste perceptual realizado por Behlau 1984, houve um grande número de erros envolvendo pares oral/nasal, o que indica pouca saliência do traço nasalidade para os ouvintes, pelo menos no tipo de contexto (logatomas) em que foram apresentadas essas vogais nasais.

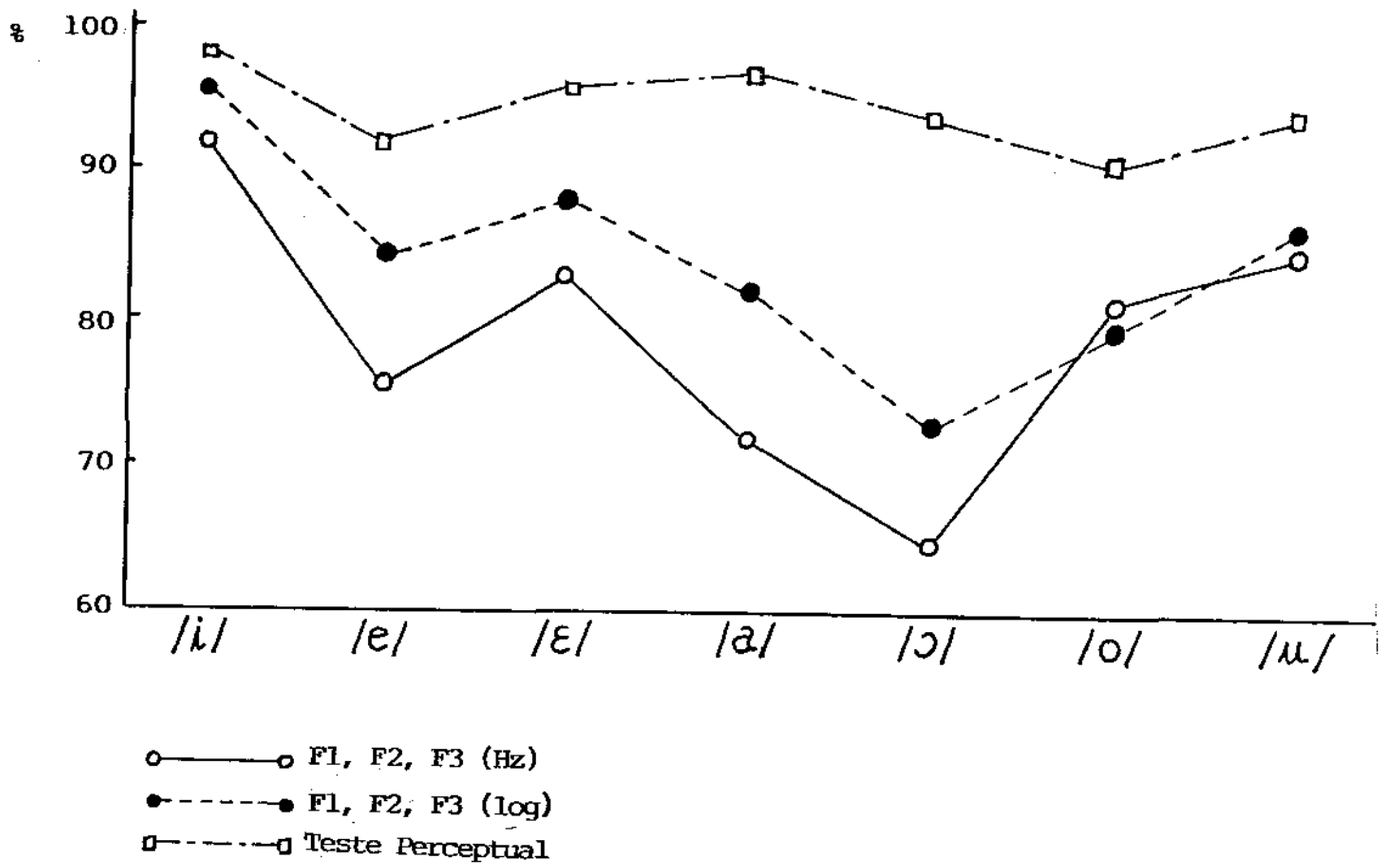


FIGURA 3.1

Examinando a figura 3.1, verifica-se que a performance do algoritmo usado por DISCRIM não é uniforme: algumas vogais são classificadas mais eficientemente do que outras. As vogais /i/ e /u/ têm os maiores índices de acerto, o que parece reforçar a noção de que essas vogais desempenham realmente um papel especial na estruturação dos sistemas vocálicos (Stevens 1972; Jakobson e Waugh 1979:92 ff.; Lieberman 1984:159 ff.).

Comparando-se os acertos de DISCRIM com os acertos no teste perceptual (linha superior do gráfico), observa-se que há alguma discrepância entre os dois tipos de resultado. A vogal /u/, por exemplo, é a segunda melhor discriminada pelo programa, mas é apenas a quarta (depois de /i/, /a/ e /ɛ/) no teste perceptual. Por outro lado, apenas 70% das vogais /a/ foram corretamente classificadas por DISCRIM, embora essa vogal tenha o segundo menor índice de erros perceptuais (2.9%), sendo superada somente pela "supervogal" /i/ (Cf. Lieberman 1984:161), que produziu apenas 1.98% de erros no teste perceptual.

O não paralelismo entre os resultados do teste perceptual e os obtidos por intermédio de DISCRIM indica, de acordo com nossas expectativas, que uma especificação simples em termos de formantes é inadequada, ou seja, a percepção de vogais - pelo menos de algumas vogais -, mesmo em isolamento, envolve outros parâmetros além das ressonâncias características. É possível que, por exemplo, a inclusão das amplitudes relativas dos formantes entre as variáveis paramétricas produzisse em espaço acústico mais semelhante ao espaço perceptual (Bernstein 1981), melhorando assim a classificação automática.

Dificuldades de mensuração talvez estejam relacionadas a alguns resultados. As medidas de Behlau 1984 foram feitas com base em espectrogramas de banda larga; embora vogais produzidas fora de um contexto frasal sejam razoavelmente estáveis, nem sempre é possível aferir com segurança o centro das freqüências

da ressonância, já que, particularmente no caso das vogais posteriores, os picos de amplitude nos espectrogramas de seção podem não estar muito bem definidos. A baixa discriminabilidade da vogal /ɔ/, por exemplo, pode estar relacionada a esse tipo de dificuldade; Ladefoged 1967 observa que essa vogal é geralmente difícil de especificar em espectrograma, e as freqüências dos formantes acabam sendo muito arbitrárias. É provável que medições mais exatas melhorassem a discriminação automática para algumas vogais. De qualquer forma, é preciso considerar que certas qualidades parecem ser mesmo muito próximas perceptualmente; a maioria dos erros perceptuais relacionados às vogais /a/ e /ɔ/ consistiu exatamente da confusão entre as duas categorias, o que está de acordo com observações anteriores em testes perceptuais com o inglês americano e britânico (Peterson e Barney 1952; Ladefoged 1967).

Outro ponto a considerar é a própria eficiência do algoritmo de classificação. O programa **DISCRIM** assume que cada categoria possui uma distribuição normal multivariada, ou seja, o espaço n-dimensional criado pelas n variáveis numéricas é considerado homogêneo; dessa forma, as distâncias utilizadas como critério de classificação são diretamente comparáveis seja qual for a orientação do vetor no interior desse espaço. Há boas razões para crer que essa assunção não seja inteiramente verdadeira, pelo menos para algumas regiões desse espaço.

Já comentamos anteriormente que certas restrições relacionadas à organização articulatória, ao controle sensório-motor e à experiência lingüística podem influenciar consideravelmente os mecanismos de produção/percepção. Desse modo, espaços criados a partir de parâmetros extraídos do sinal de fala dificilmente apresentarão propriedades homogêneas em toda sua extensão. No caso específico da classificação automática baseada em formantes, não parece muito razoável presumir - como faz **DISCRIM** - que F1, F2 e F3 variem "livremente". Embora o programa estabeleça uma métrica baseada na CO-variância dos parâmetros de controle, informação adicional quanto ao peso relativo e características distribucionais de cada uma das variáveis certamente produzirá resultados mais promissores. O algoritmo pode ser aperfeiçoado, por exemplo, na medida em que incorpore regras probabilísticas derivadas da observação de um grande número de dados da língua particular à ser analisada. Dessa forma, o cômputo das distâncias poderia ser feito não mais (ou apenas) em relação ao centro virtual de cada classe mas (também) levando em conta as diferentes densidades em cada região do espaço.

A figura 3.2 ajudará a esclarecer esse ponto. Para facilitar a visualização, representamos as distâncias em um espaço bidimensional; os princípios gerais, no entanto, são válidos para qualquer número de dimensões. Na figura 3.2, o ponto p representa o item hipotético a ser classificado em uma das classes

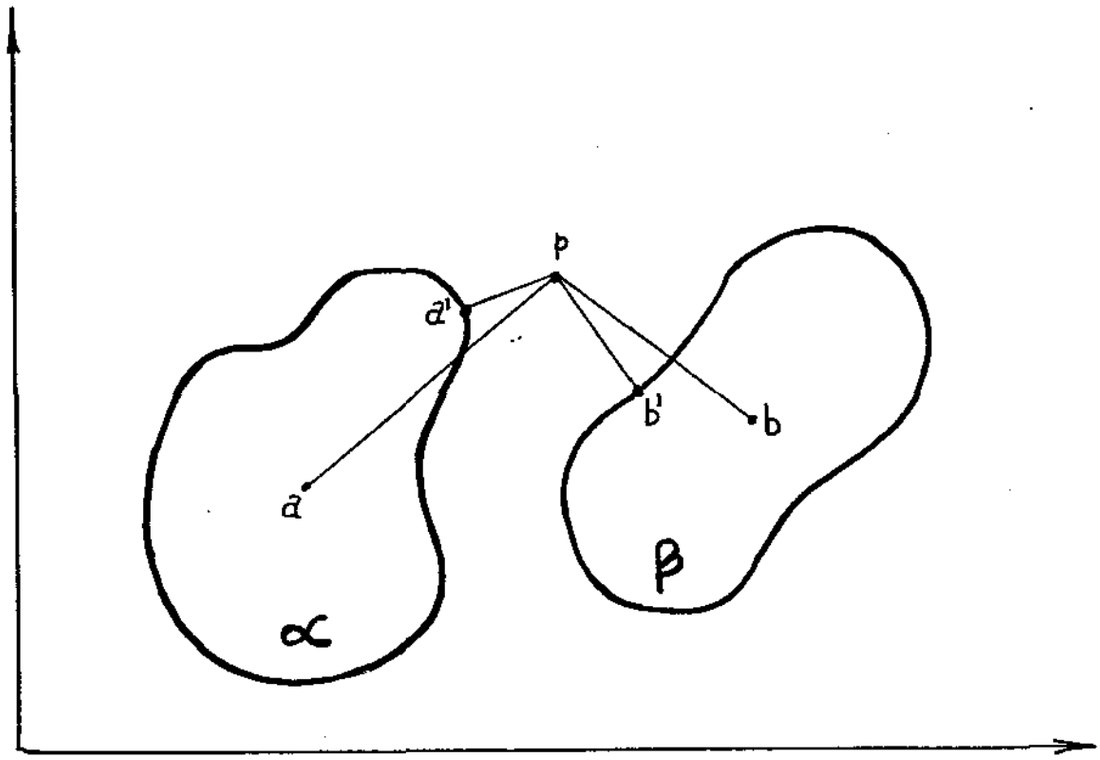


FIGURA 3.2

hipotéticas α ou β . Se a medida de distância for tomada em relação ao centro de cada conjunto, o ponto p seria classificado na categoria β , já que $\overline{pb} < \overline{pa}$. Se, no entanto, aferirmos a distância em relação às margens de cada conjunto, p seria classificado na categoria α, pois $\overline{pa}' < \overline{pb}'$.

É importante ressaltar que a representação feita através da figura 3.2 é bastante esquemática. Os limites de cada categoria não refletem diretamente a distribuição real dos elementos de cada categoria, mas antes a densidade dessa distribuição em várias direções com base em algum índice de dispersão (desvio

padrão, por exemplo). A figura ilustra, portanto, uma possibilidade de decisão tomada em relação a níveis idênticos de dispersão para as duas categorias (a mesma quantidade de unidades de desvio padrão, por exemplo). É preciso não perder de vista, no entanto, que uma classificação desse tipo depende de decisões prévias, mais ou menos arbitrárias, quanto à representatividade do *corpus* utilizado. Essa métrica certamente será mais eficiente quando se parte de um vocabulário limitado e/ou um conjunto conhecido de falantes. Sistemas automáticos com capacidade de adaptação, no entanto, já foram sugeridos; esses sistemas possuem um módulo "supervisor" que checa as decisões de classificação e atualiza os parâmetros a cada nova entrada de dados, reformulando os níveis de distribuição em relação aos quais novas decisões serão tomadas (Pal *et al.* 1980).

Escolha da Escala de Freqüência

Um fator que influi consideravelmente no desempenho do algoritmo de classficação é a unidade escalar - usada para representar as freqüências ds formantes. Como se pode observar na figura 3.1, todas as vogais, com exceção de /o/, foram classificadas mais eficientemente por DISCRIM quando os formantes são expressos em uma escala logarítmica - no caso específico da vogal /a/ há um ganho expressivo de cerca de 10%. Esse resultado reflete o fato, bastante conhecido, de que a escala de freqüência

em Hz não é muito realista, do ponto de vista psico-perceptual: a diferença absoluta em Hz entre duas freqüências em uma faixa alta não provoca o mesmo efeito perceptual que a mesma diferença em uma faixa baixa.

A transformação logarítmica povoca uma compressão nas faixas muito altas de freqüência e faz com que a representação se aproxime mais da realidade perceptual. Como a transformação é mais efetiva nas freqüências altas, é possível que a pequena influência do uso de escalas log na discriminação de /o/ e /u/ esteja relacionada ao fato de essas vogais terem freqüências de ressonância tipicamente baixas.

O uso de escalas log contribui também no sentido de diminuir diferenças entre os grupos de falantes (homens, mulheres e crianças) provocando um efeito normalizador que equaliza a dispersão nos planos F1 X F2 e F2 X F3 (Lennig e Hindle 1977; Kent e Forner 1979).

Outras transformações do eixo das freqüências têm sido sugeridas. As escalas adotadas são, em geral, derivadas de testes psico-físicos. A escala MEL (Stevens e Volkmann 1940; *apud* Nearey 1989), de largo uso, foi baseada em tarefas envolvendo fracionamento de intervalos de altura melódica e julgamentos subjetivos de diferenças de altura melódica em sons senoidais. A escala sugerida por Koenig (1949; *apud* Miller 1989) e a escala MEL TÉCNICO (Fant 1973) são aproximações da escala MEL original.

Mais recentemente tem se tornado freqüente o uso de representações em termos de unidades **BARK** (v. Zwicker 1961; Zwicker e Terhardt 1980). Essa medida é baseada em testes de mascaramento auditivo e em estudos que estimam a largura de banda dos filtros auditivos. A escala **BARK** foi originalmente desenvolvida para dar conta de algumas propriedades relacionadas à resolução de freqüência no aparelho auditivo, e não propriamente como uma representação de altura melódica (pitch), como a escala **MEL**, por exemplo. É possível, no entanto, através de uma transformação matemática relativamente simples, construir uma escala em unidades **BARK** a partir da função original de *banda crítica* proposta por Zwicker 1981 (Moore e Glasberg 1983) [2]

Alguns pesquisadores tentaram desenvolver escalas *ad hoc* para freqüências de formantes, especificamente projetadas com a finalidade de otimizar alguma propriedade de representação vocálica. Traunmüller 1981, 1988 propõe uma modificação da escala **BARK** na região abaixo de 250 Hz de modo a dar conta de certas relações entre F1 e F0. Nearey 1978 e Lennig 1978 (*apud* Nearey 1989) tentam construir escalas artificiais que conduzam à uma variância mais homogênea para as medidas dos formantes. Foulkes 1961, embora não diretamente procupado com a construção de uma escala padrão, demostra que uma transformação matemática das coordenadas em um plano **F1 X F2** (com base em dados de Peterson e Barney 1952) pode fazer com que as fronteiras categoriais se tornem menos complicadas, aproximando-se, em alguns casos de

linhas retas.

Embora essas escalas e transformações possam eventualmente aumentar a eficiência de alguns esquemas de reconhecimento, a vantagem, quando existe, parece ser mínima em relação a uma simples escala logarítmica, e é discutível se, levando em conta a complexidade adicional, justifica-se sua utilização (Nearey 1989). No caso das escalas *ad hoc* há ainda o agravante de não existir, em geral, uma motivação psico-física para sua elaboração. Além disso, esses ajustes são feitos com base em conjuntos específicos de dados procurando obter um máximo de eficiência do algoritmo de classificação; não é certo, entretanto, que essa eficiência seja mantida na análise de dados com diferente distribuição (vogais de outra língua ou dialeto, por exemplo).

Mesmo escalas perceptualmente relevantes (**MEL**, **BARK**, etc) não oferecem vantagens significativas em algumas aplicações, e sua superioridade sobre escalas mais tradicionais (**Hz** ou **log**) tem sido constestada por alguns pesquisadores (Peterson 1961; Miller 1989).

Na análise dos dados de Behlau 1984 verificamos que o uso das escalas **MEL**, **BARK** e **KOENIG** produziu resultados semelhantes ao da escala **log** quanto ao desempenho do programa **DISCRIM**; apenas a escala tradicional em **Hz** mostrou-se nitidamente menos eficaz para a classificação automática. É preciso considerar, no entanto, que há alguma evidência sugerindo que a escolha do algoritmo de

classificação pode interagir com a escala utilizada; Hillenbrand e Gayvert 1987 relatam que mesmo uma representação em Hz produz altas taxas de identificação se for aplicada na classificação uma análise discriminante que não presuma a homogeneidade das matrizes de dispersão.

Escalas logarítmicas, apesar de derivadas por meio de uma transformação puramente matemática, refletem também uma realidade perceptual. O sistema musical, que depois da linguagem falada pode ser considerado como o código sonoro hierarquizado mais complexo, baseia-se em relações logarítmicas; todas as escalas musicais, de todas as culturas, embora possam diferir quanto à organização dos intervalos internos, têm a iotava como módulo, o que implica um escalamento do tipo log. Harris 1960 demostra que os ouvintes escalam a altura melódica de acordo com freqüências log - uma escala MEL só é obtida quando se exige o escalamento de intervalos muito grandes. Harris observa que o uso de escalas MEL resultaria em formas musicais estranhas, baseadas em intervalos consideravelmente maiores que o semitom; além disso, essa música seria necessáriamente monofônica, já que escalas MEL não admitiriam relações harmônicas - pelo menos não do modo como são hoje concebidas [3].

Outra vantagem da representação logarítmica é que, nas medições espectrais tem-se a oportunidade de avaliar distâncias e velocidade de transição em termos que podem ser facilmente comparados a medidas de outros sons (v. Miller 1989).

4

NORMALIZAÇÃO

Transformações mais adequadas do eixo das freqüências e a implementação de algoritmos de classificação mais sofisticados podem, sem dúvida, contribuir para um melhor desempenho dos sistemas de reconhecimento automático de vogais. No entanto, ao se optar por uma representação em termos de formantes é praticamente inevitável que o sistema, em algum momento, se defronte com o clássico problema da superposição parcial de categorias. O fato é que, especialmente quando a população a ser analisada é diferenciada em termos de sexo e faixas etárias, existirão alguns pontos sobre os quais não será possível tomar uma decisão segura de classificação com base apenas nas freqüências absolutas dos formantes. Em um espaço tridimensional definido por **F1 X F2 X F3**, o /ɔ/ intencionado por um adolescente pode ocupar o mesmo lugar de um /a/ de um adulto, ou o /ɛ/ de uma mulher pode estar ao lado de um /e/ produzido por uma criança de 8 anos. Esse tipo de dificuldade, no entanto, não parece existir para ouvintes humanos: identificamos com a mesma eficiência vogais produzidas por adultos e crianças.

Vogais "Calibradoras"

Já examinamos acima (v. seção 1) alguns experimentos que demonstram que a categorização de algumas vogais pode ser alterada sistematicamente através da manipulação da freqüência média dos formantes na sentença anterior ao estímulo teste (Ladefoged e Broadabend 1957; Ladefoged 1967; Ainsworth 1975). Esses resultados podem ser interpretados sob uma ótica meramente psicofísica, no sentido em que refletiriam um princípio perceptual mais ou menos genérico de adaptação seletiva. À luz da presente discussão, entretanto, é possível ver aí a atuação de mecanismos mais sofisticados: o que parece ocorrer é uma estimativa, por parte do ouvinte, das dimensões do trato vocal do falante, com base nas tessituras dos formantes no ambiente fonético precedente. O experimento relatado por Dechovitz 1977 ajudará a esclarecer esse ponto; em vez de manipulações sintéticas do sinal, Dechovitz utiliza como estímulos-teste sílabas extraídas da fala real de um homem adulto embutidas em uma frase produzida por uma criança de 9 anos. Os desvios perceptuais observados são similares as obtidos com fala sintetizada: o ouvinte "erra" a qualidade vocálica intencionada porque situa o som produzido pelo adulto em relação ao espaço vocálico estimado para a criança.

Esses resultados são reveladores, mas certamente não explicam tudo. O fato é que ouvintes humanos não precisam ouvir

uma sentença completa de modo a poder estimar as dimensões hipotéticas do trato vocal do falante e assim ajustar a "balística" perceptual. Lieberman 1984 sugere que algumas vogais talvez sirvam como "calibradores naturais" do sistema; essa função seria desempenhada principalmente pela "supervogal" /i/. Ele observa que essa vogal é a que produz menor número de identificações erradas nos testes perceptuais realizados por Peterson e Barney 1952 (o mesmo ocorre com dados do PB em Behlau 1984). As vogais /i/ de todos os falantes - homens, mulheres ou crianças - são sempre identificadas corretamente em virtude da posição especial que ocupam no espaço acústico. As figuras 4.1 e 4.2 ilustram esse ponto.

A figura 4.1 mostra os limites de cada categoria em um plano F1 X F2 para todos os falantes de Behlau 1984 (30 homens, 30 mulheres e 30 crianças). como se pode observar, o conjunto das vogais /i/ praticamente não se sobrepõe a qualquer outra categoria [1].

A figura 4.2 apresenta os valores médios para cada vogal/grupo de falantes em um espaço tridimensional definido por F1 X F2 X F3 expressos em escala logarítmica ($\log_{10}(F_n)$). Fica evidente a posição privilegiada da vogal /i/. A situação é bem diferente em outras áreas do gráfico: observe-se, por exemplo, a proximidade do /a/ masculino com os /ɔ/ dos dois outros grupos [2].

Também a distribuição global da energia acústica no espectro de /i/ favorece o reconhecimento dessa vogal. Delattre *et al.*

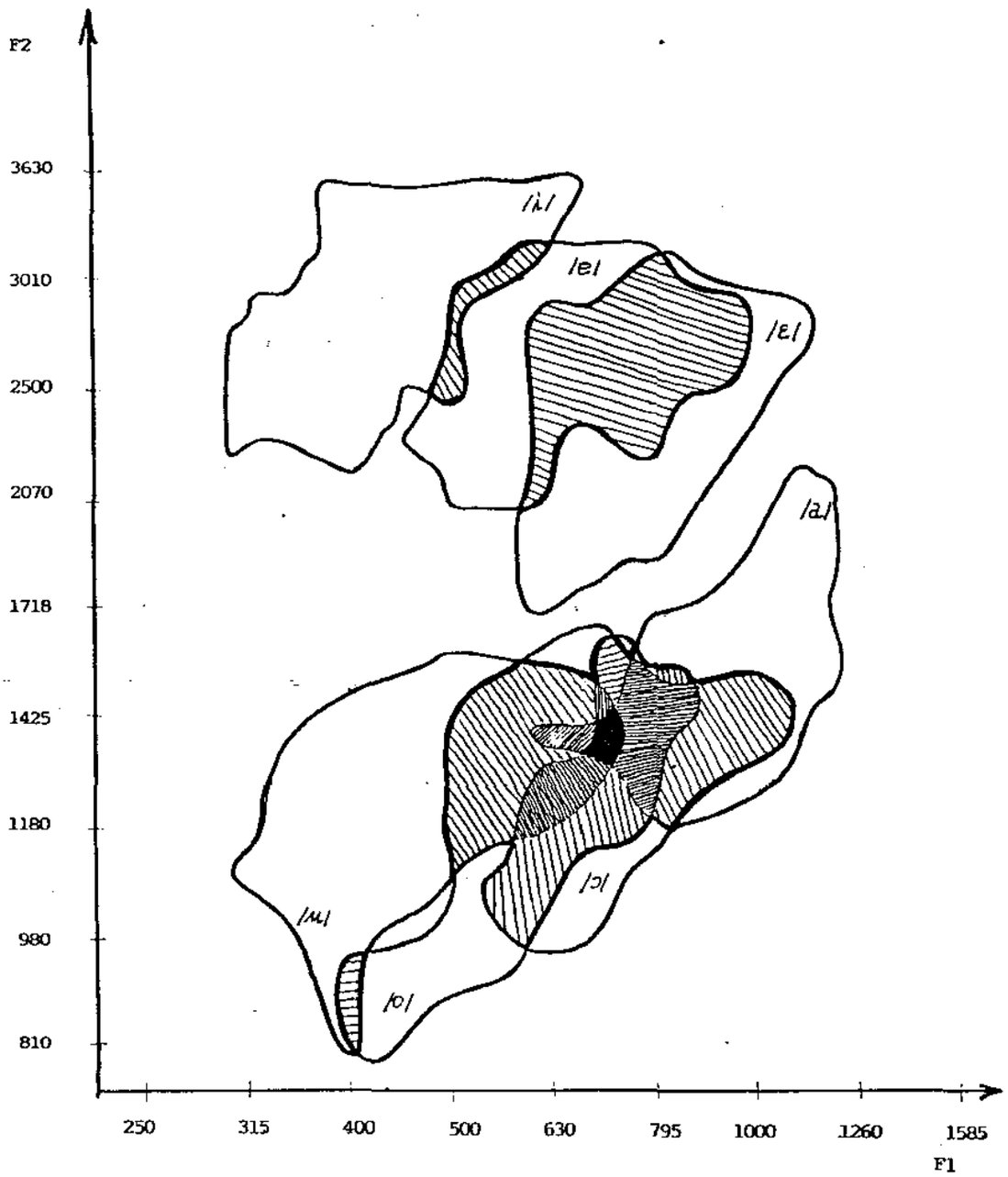

F2
3630
3010
2500
2070
1718
1425
1180
980
810
/i/
/e/
/ɛ/
/a/
/ɔ/
/ʊ/
/o/
250
315
400
500
630
795
1000
1260
1585
F1

FIGURA 4.1

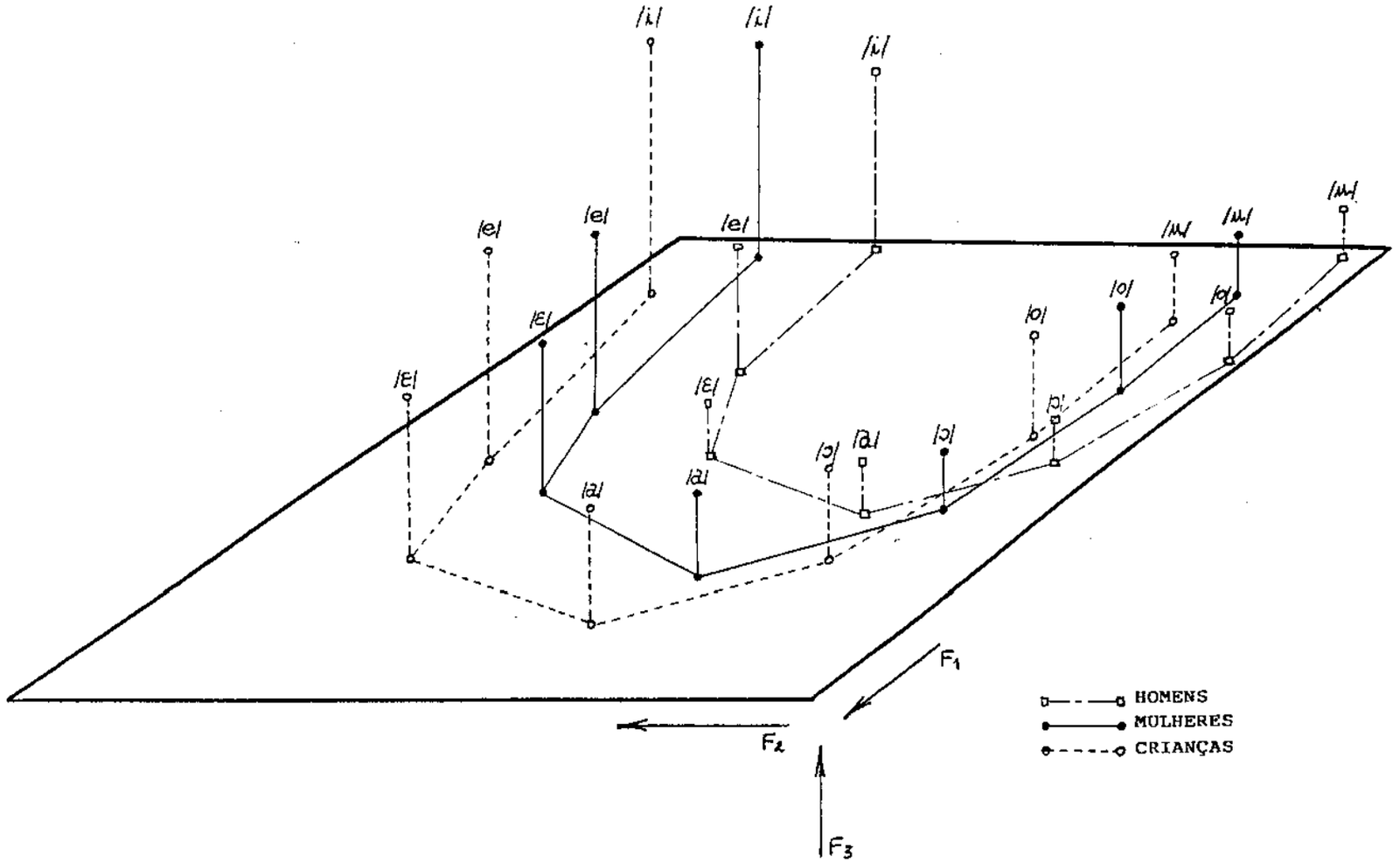


FIGURA 4.2

1952, produzindo aproximações de um e dois formantes para várias vogais, observam que /i/ é a qualidade mais facilmente identificável nas duas condições (89% e 100% de acertos, respectivamente). Eles sugerem que a típica concentração de energia acústica em frequências altas basta para caracterizar essa vogal; dessa forma,"a single formant near the normal position of the second formant seemed to produce the /i/ color rather well" (pg.232). Da mesma forma, Carlson *et al.* 1975, nas suas aproximações de vogais com apenas dois formantes (embora os valores absolutos difiram razoavelmente daqueles encontrados por Delattre *et al.* 1952) observam que há uma descontinuidade entre o F2' (ou seja, o F2 perceptualmente ajustado como melhor aproximação para uma vogal de dois formantes) de /i/ e o F2' da vogal mais próxima [3].

É interessante observar que a posição da vogal /i/ no espaço acústico parece também privilegiada mesmo se considerarmos as distribuições de cada formante separadamente. Aplicando o programa **DISCRIM** nos dados de Behlau 1984 (apenas vogais orais) observamos que, utilizando como variáveis paramétricas cada um dos formantes em isolamento, ou apenas os dois primeiros formantes, obtínhamos, em todos os casos, as maiores taxas de classificação correta para a vogal /i/, como demonstra a figura 4.3.

Assim como /i/, também a vogal /u/ ocupa uma posição extrema, tanto do ponto de vista articulatório quanto acústico.

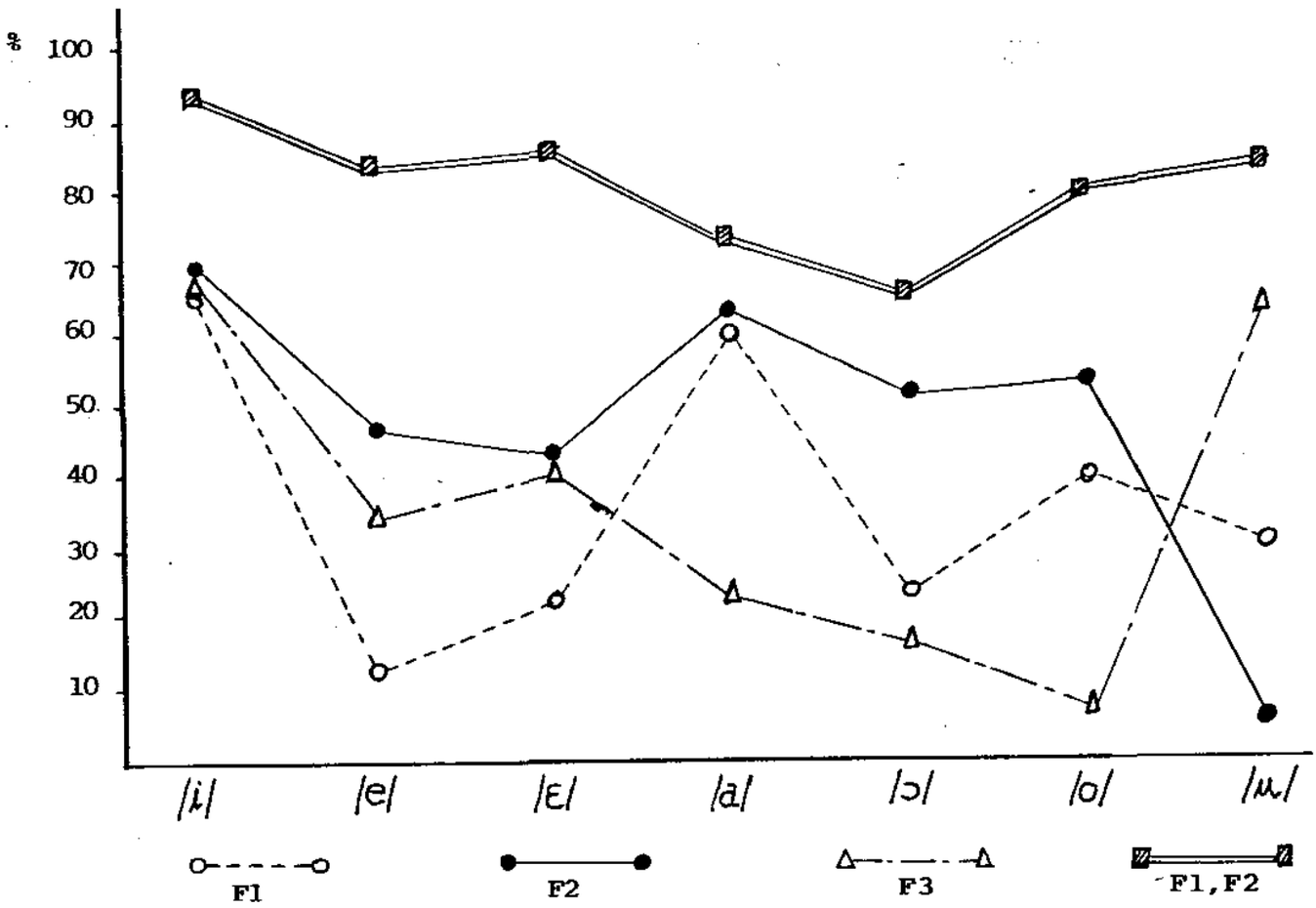


FIGURA 4.3

Em geral, a vogal /u/ é bem identificada nos testes perceptuais, embora provoque, regularmente, um maior número de confusões que /i/ (Peterson e Barney 1952; Fairbanks e Grubb 1961; Fowler e Shankweiler 1978; Strange *et al.* 1976; Ryalls e Lieberman 1982). Nos testes perceptuais de Behlau 1984 a vogal /u/ não é tão bem identificada, aparecendo depois de /i/, /a/ e /ɛ/; é preciso considerar, no entanto, que grande parte dos erros perceptuais relacionados a /u/ resultam de confusões com o equivalente nasal

/u/.

A aparente vantagem perceptual das vogais /i/ e /u/ serviu de base para o algoritmo de normalização desenvolvido por Gerstman 1968. De modo a compensar a variação inter-falante, ele propõe coeficientes de normalização que poderiam ser extraídos apenas a partir de F1 e F2 das vogais /i/ e /u/ de cada falante.

O algoritmo de Gerstman 1968 presume, no entanto, que o ouvinte deve adiar seu julgamento quanto à qualidade de uma vogal até ter acesso a ocorrências de /i/ ou /u/. De modo a contornar essa limitação, Nearey 1978 (v. Lieberman 1984:163 ff. para um resumo) sugere uma solução mais consistente com as respostas perceptuais de ouvintes humanos. O algoritmo de normalização proposto por Nearey assume apenas que as vogais produzidas por um dado falante podem ser especificadas se o ouvinte tem acesso a pelo menos **UMA** vogal identificada. O algorítmo pressupõe que, se o ouvinte conhece a identidade de uma vogal qualquer, ele é capaz de derivar as dimensões do trato que a produziu. Com base nessa informação inicial seria possível ajustar o mecanismo perceptual para o reconhecimento dos outros pontos do sistema vocálico daquele falante particular. Lieberman 1984 lembra que a freqüente utilização de expressões estereotipadas como "aberturas" de conversação pode ser mais do que um mero rito social, e talvez sirvam como sinais de normalização: "a listener hearing these openers knows the intended phonetic targets; thus any stereotyped opener can serve as a calibrating signal for vocal tract

normalization" (pg. 166).

Proporções entre Formantes

Alguns pesquisadores têm sugerido que se for usada uma representação baseada na proporção (razão) entre os formantes é possível eliminar, ou pelo menos reduzir consideravelmente, diferenças na descrição acústica de vogais relacionadas à variação inter-falante. A idéia não é nova e origina-se do trabalho de Lloyd 1890 (v.Shoup e Pfeifer 1976). Lloyd chamou sua teoria de "The relative Resonance Theory" e afirmava, em suma, que a qualidade vocálica dependia dos intervalos entre as ressonâncias e não de seus valores absolutos. Modificações da teoria original de Lloyd reaparecem esporadicamente na literatura (Potter e Steinberg 1950; Peterson 1961; Broad 1976; Kent 1979 e Miller 1989), embora a referência ao trabalho de Lloyd seja freqüentemente omitida (v. no entanto Miller 1989).

Não há dúvida que, dentro de certos limites, as vogais retêm sua qualidade se as razões entre os formantes forem mantidas. Isso pode ser verificado através de um experimento bastante simples: basta acelerar ou retardar a velocidade de reprodução de um gravador de fita. Essa técnica banal de compressão/expansão temporal da cadeia de fala faz com que todas as freqüências subam ou desçam proporcionalmente ao aumento/redução da velocidade de reprodução sem alterar, todavia, as razões entre as freqüências

dos formantes. Se a alteração de velocidade de reprodução não for muito grande, a fala permanece compreensível e as vogais identificáveis. Klumpp e Webster 1961, usando esse procedimento, observam que um aumento de 1.5 em relação à velocidade da gravação original mantem ainda cerca de 90% da inteligibilidade de um sinal de fala.

Uma especificação simples apenas em termos das proporções entre os formantes não consegue, no entanto, dar conta do quadro como um todo. A teoria é incapaz de esclarecer algumas questões básicas:

- Porque a transposição das proporções da estrutura de formantes de uma vogal só é possível dentro de certos limites?
- Porque algumas vogais perceptualmente distintas têm razões de formantes similares?
- Porque diferentes ocorrências de uma mesma vogal podem ter diferentes razões de formantes?

O simples exame da figura 4.4 revela algumas deficiências da teoria. Como se observa, os dados para uma determinada vogal, embora tendam a se agrupar ao longo de linhas representando razões F2/F1 constantes, produzem ainda uma dispersão considerável. Também se nota que algumas qualidades distintas distribuem-se ao longo de linhas similares.

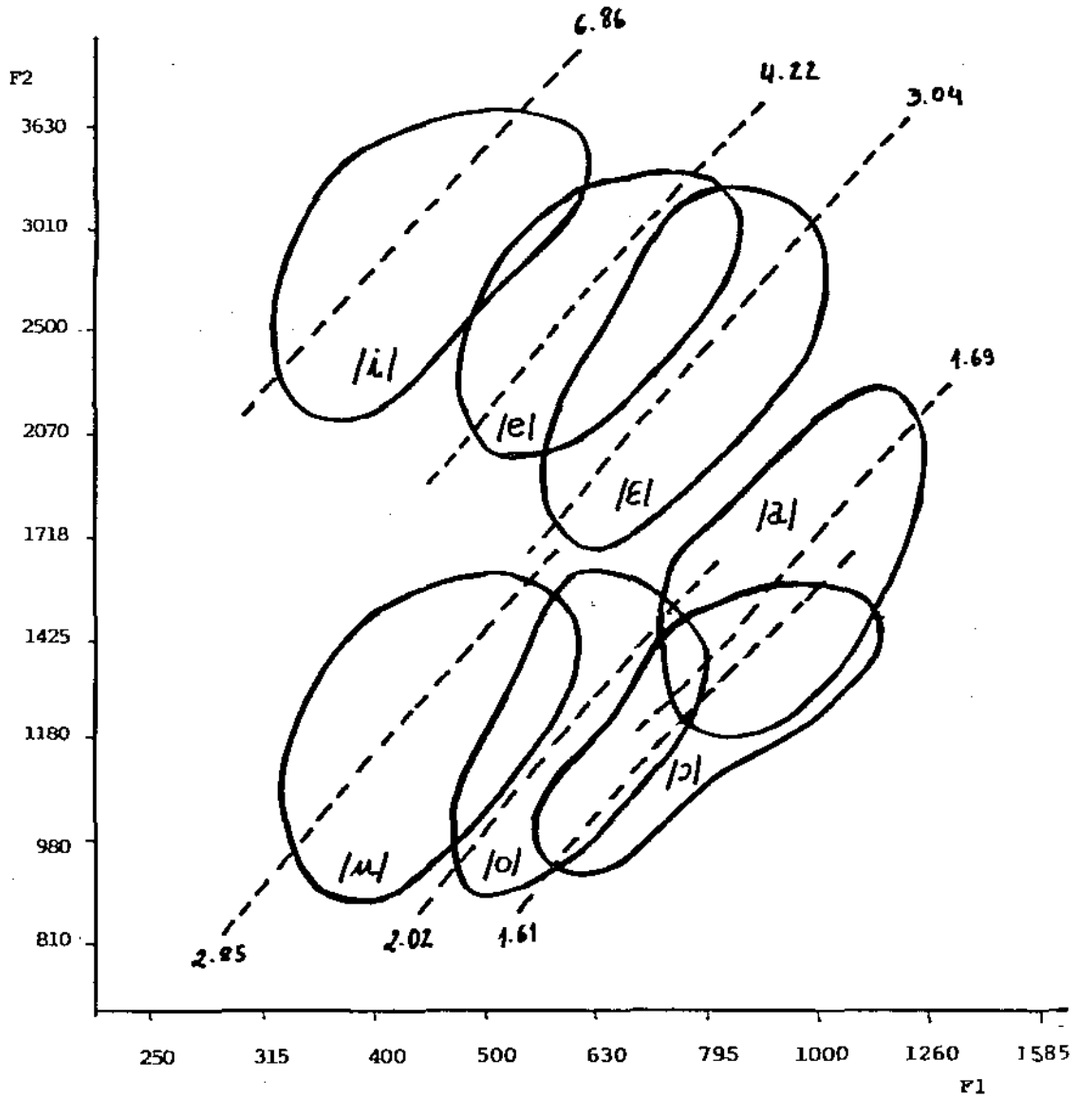

F2
3630
3010
2500
2070
1718
1425
1180
980
810
250
315
400
500
630
795
1000
1260
1585
F1
6.86
4.22
3.04
1.69
2.85
2.02
1.61
/i/
/e/
/ɛ/
/a/
/ɔ/
/o/
/u/

FIGURA 4.4

Diferenças no Padrão de Formantes Relacionadas a Sexo e Idade

A maior dificuldade para um modelo baseado exclusivamente nas proporções entre os formantes e, de forma geral, para qualquer algoritmo de normalização, é o fato de existirem certas diferenças anatômicas entre diferentes grupos de falantes - relacionadas principalmente a sexo e idade - que não permitem supor uma uniformidade de relação entre os padrões de formantes para vogais equivalentes. Os formantes das mulheres e das crianças não sobem todos proporcionalmente em relação ao padrão masculino. O trato vocal feminino (e provavelmente o de crianças até a fase pré-pubertal) não difere do masculino apenas quanto às dimensões globais, mas também quanto a certas proporções internas entre as diversas partes; enquanto a parte anterior do trato é praticamente igual à dos hommens, a faringe feminina (e a das crianças) é significativamente mais curta (Fant 1980). Além disso, observações diretas usando tomografia computadorizada verificaram que existem também diferenças mais específicas quanto às proporções de algumas estruturas internas da parte posterior do trato; no trato feminino a laringo-faringe e a orofaringe são relativamente mais curtas que a porção média da laringe, em comparação com as mesmas medidas do trato masculino (Sundberg *et al.* 1987).

O principal efeito dessas assimetrias anatômicas é a impossibilidade de fixar um fator simples de escala capaz de

relacionar, de forma geral, as estruturas de formantes de homens e mulheres, ou de homens e crianças. Um algoritmo como o de Gerstman 1968 (v.acima), por exemplo, que re-escala linearmente todas as frequências entre dois extremos arbitrariamente pré-determinados (0 e 999), embora consiga assimilar boa parte da variação intersubjetiva, não é capaz de eliminar completamente a superposição parcial entre algumas categorias.

O método de Gerstman pressupõe que os padrões de formantes individuais para cada vogal convergem para um ponto comum, isto é, que cada padrão pode ser entendido como uma compressão/expansão de um modelo prototípico comum. A realidade, infelizmente, não é tão simples. Examinando a figura 4.5, que mostra os aumentos precentuais para cada vogal/formante de mulheres e crianças em relação aos valores médios do grupo masculino (a partir dos dados em Behlau 1984), pode se notar que as diferenças relativas variam tanto em função da vogal quanto do formante. É interessante observar, no entanto, que o padrão de aumento para os dois grupos (mulheres e crianças) é bastante parecido (desconsiderando os valores absolutos) como indica o paralelismo das linhas em cada formante.

Tentativas de sintetizar imitações convincentes da voz feminina são fatalmente mal-sucedidas se o padrão de formantes é apenas um re-escalamento uniforme do padrão masculino. Multiplicar os valores masculinos por um fator simples produz uma fala sintetizada inteligível mas pouco natural (v. Klatt 1977).

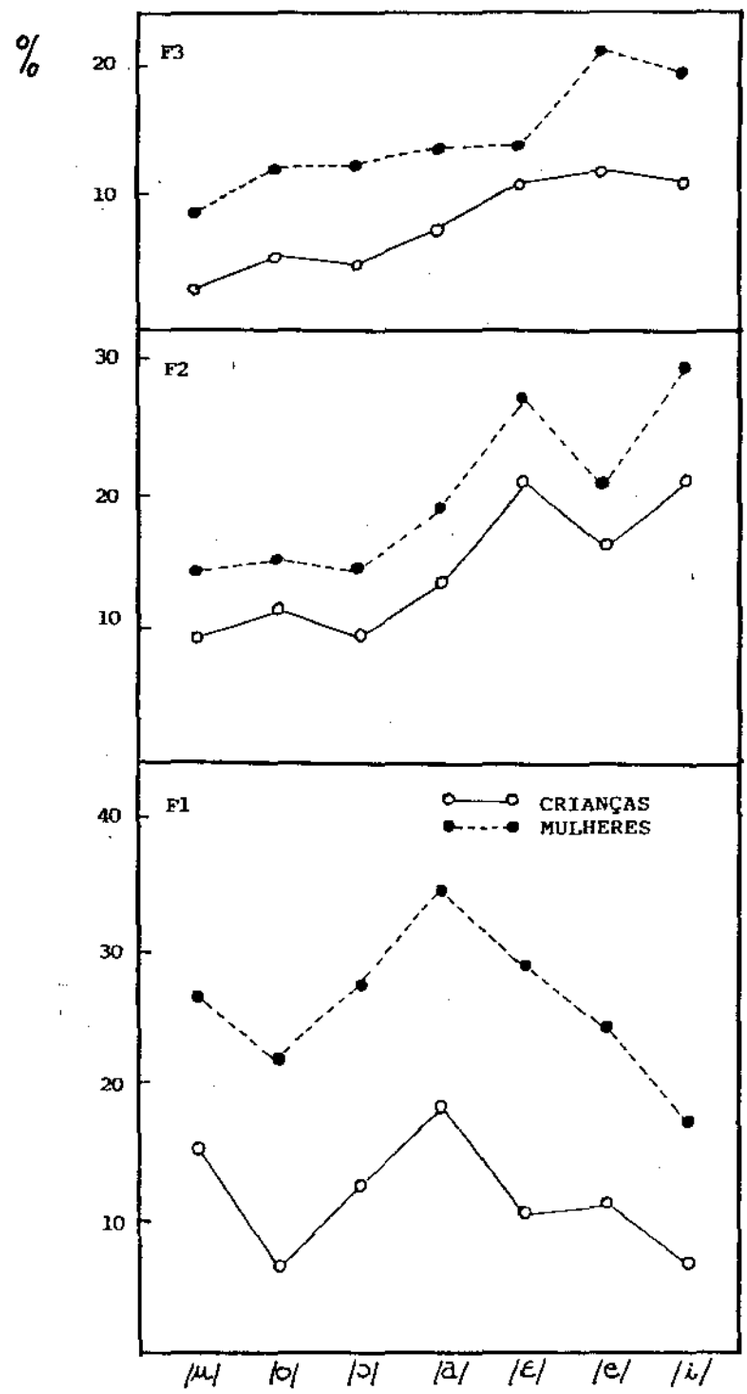


FIGURA 4.5

Fant 1973;1980 sugere que a não uniformidade da relação entre os grupos masculino e feminino pode estar relacionada às diferentes vinculações de formantes específicos com cavidades específicas. Se o formante tem relação com a faringe, haveria uma tendência a ocorrer uma maior alteração, já que essa região do trato é, como vimos acima, onde as diferenças anatômicas entre homens e mulheres (ou crianças) são mais acentuadas - é o caso, por exemplo, do F2 de /i/. Não existindo uma vinculação direta do formante com a região posterior do trato, as diferenças tendem a ser menores, é o caso de F1 e F2 de /u/, por exemplo.

Outro aspecto a considerar são as possíveis compensações que os falantes estão aptos a realizar para tentar minimizar diferenças perceptuais. Essa possibilidade já foi observada por Helmholtz 1857: "What is lacking to the childish and female mouth in capacity can be easily replaced by narrower closure of the opening, so that the resonance can still be as deep as in the larger male mouth" (*apud* Scripture 1902:408). Fant 1973 observa que formantes produzidos através de configurações de duplo ressoador, como F1 e F2 de vogais posteriores arredondadas, são menos dependentes do tamanho geral do trato; nessa situação, um trato menor poderia - como sugerira Helmholtz - ser compensado por um maior arredondamento e/ou aumento da constrição. A impossibilidade de utilizar esses mecanismos compensatórios em vogais abertas como /a/, diz Fant, explicaria os maiores aumentos percentuais observados nos formantes femininos dessas vogais,

especialmente F1, que é mais dependente das dimensões globais do trato.

Essas hipóteses, admite Fant 1980, são um tanto especulativas, já que há pouco conhecimento a respeito dos mecanismos compensatórios utilizados. A escassez de dados derivados de observações radiográficas diretas e as limitações impostas pela postura pouco natural da cabeça exigida para esse tipo de observação dificultam a pesquisa e aconselham cautela nas conclusões. Além disso, Fant 1980 adverte quanto ao fato de as diferenças não uniformes entre homens e mulheres encontrarem um paralelo com padrões similares observados na comparação de vozes de cantores com registros de baixo e de tenor; tais diferenças não poderiam ser apenas uma conseqüência automática de diferentes escalas anatômicas e parecem revelar compensações segundo critérios ainda não muito bem compreendidos.

A possível influência de fatores sócio-lingüísticos complica ainda mais o quadro. Fant 1980 ressalta que talvez algumas compensações ocorram, não para minimizar diferenças perceptuais, mas sim, pelo contrário, para **MARCAR** contrastes entre diferentes grupos de sexo e idade. Traunmüller 1988, estudando a realização das vogais (japonês) por diferentes grupos de sexo e idade, observa que o grupo feminino tende a produzir qualidades mais periféricas que o grupo masculino, uma diferença similar à observada entre vogais acentuadas e reduzidas. Traunmüller conclui que "there are...convincing reasons to believe thar the

phonetic quality of women's and men's renditions of the same vowel phonemes is not precisely the same" (pg. 22).

Deve também ser levado em conta que uma explanação da não uniformidade entre os padrões masculino e feminino baseada nas diferentes condições anatômicas e nas vinculações dos formantes à cavidades específicas do trato não poderia ter validade geral; o fato é que uma associação estrita formante/cavidade, na maior parte dos casos, é apenas ideal, devido às muitas interações entre os ressoadores. Essa associação só é viável em casos especiais, quando o grau de constrição é bastante acentuado, produzindo no mínimo uma razão 6:1 entre as áreas seccionais de cada cavidade, de modo a criar ressoadores mais ou menos independentes (Sundberg 1974).

5
FREQÜÊNCIA FUNDAMENTAL

A dificuldade de extrair parâmetros invariantes com base apenas nas freqüências dos formantes (absolutas ou relativas) tem levado alguns pesquisadores a incluir outras variáveis na descrição acústica das vogais. Nos últimos anos tem sido largamente enfatizada a importância da freqüência fundamental na caracterização de vogais (Traunmüller 1981, 1988; Syrdal e Steele 1985; Syrdal e Gopal 1986; Miller 1984, 1989). A questão, no entanto, é controversa e merece um exame mais atento, já que há bons argumentos e evidência experimental tanto contra como a favor dessa abordagem.

FO Intrínseco

Um dos argumentos usados para justificar a inclusão de FO na descrição acústica é a possível existência de uma relação sistemática entre a qualidade de uma vogal e seu "FO intrínseco" (v., por exemplo, Traunmüller 1988:5). De fato, Lehiste e Peterson 1961, medindo a freqüência fundamental de vogais em contexto CVC (medição no pico da seção vocálica) produzidas por vários falantes em um contexto frasal "Say the word...again", observam que, mantidos aproximadamente os mesmos padrões

acentuais e entoacionais, há uma variação quase regular do F0 intrínseco negativamente correlacionada com o primeiro formante da vogal. Assim, para todos os falantes, vogais altas fechadas (ou seja, com F1 baixo) como /i/ e /u/, tendem a estar associadas com um F0 intrínseco mais elevado do que vogais abertas (ou seja, com F1 alto) como /a/ ou /ɔ/. Outros estudos verificaram o mesmo fenômeno em vogais de outras línguas, como o dinamarquês (Peterson 1978) e o alemão (Antoniadis e Strube 1981), embora a magnitude das diferenças entre vogais varie um pouco de língua para língua.

Lehiste 1970 atribui essa variação a certas condições articulatórias características de cada vogal: a posição elevada da língua nas vogais altas como /i/ e /u/ tende a provocar, por sinergia muscular, um levantamento concomitante da laringe, aumentando assim o estiramento e tensão das cordas vocais e, conseqüentemente, reduzindo seu período vibratório.

Em geral, o abaixamento da laringe associa-se a uma diminuição da freqüência fundamental (Ohala 1972; Rossi e Autesserre 1981). Alguns cantores, cuja tessitura natural é um pouco mais baixa que a normalmente exigida para um registro típico de tenor, podem compensar a diferença adotando uma postura articulatória que mantém a laringe levantada durante o canto (Laver 1980:28).

Explanações de natureza aerodinâmica também já foram sugeridas. Atkinson 1973 relaciona a elevação da freqüência

fundamental nas vogais fechadas ao acoplamento acústico provocado pela proximidade de F1 e F0 nessas vogais. Lieberman 1977:94 afirma que "there is aerodynamic coupling between the first formant frequency and the glottal waveform. Low first formants result in a somewhat higher fundamental frequency of phonation, whereas a higher first formant frequency yelds a lower F0".

Influência do Contexto Consonantal

Embora haja uma relação entre o F0 intrínseco e a qualidade vocálica, as diferenças só podem ser estabelecidas em termos de valores médios, quando um número considerável de enunciados são comparados. Examinando mais de perto o ambiente consonantal das palavras-teste CVC, Lehiste e Peterson 1961 verificaram que o comportamento da freqüência fundamental medida na vogal intermediária sofria forte influência do tipo de consoante imediatamente anterior: em geral, F0 mais altos ocorriam após consoantes não-sonoras, enquanto F0 consideravelmente mais baixos estavam asociados a vogais precedidas por consoantes sonoras.

O mais interessante é que a influência do contexto consonantal precedente na variação do valor máximo de F0 medido na vogal neutraliza em parte os efeitos da qualidade vocálica no F0 intrínseco. A tabela II em Lehiste e Peterson 1961:421, mostra os valores referentes aos picos de F0 em função da vogal e da consoante anterior. Se, com base nos dados dessa tabela,

aferirmos a variação percentual em termos da relação entre os valores máximo e mínimo para o FO médio das diferentes vogais (excluindo ditongos) medidos em vários contextos consonantais, vamos obter um valor de 12.96%, ou seja:

Variação perceptual máxima inter-vogal =
[(FO/i/ / FO/ /) - 1] x 100 = [(183/162) - 1] x 100 = 12.96%

Examinando agora a variação em função da consoante anterior obtemos - segundo um cálculo semelhante - os seguintes valores para cada vogal:

$V_{/i/}$ = 20.24%	$V_{/a/}$ = 17.40%
$V_{/I/}$ = 10.36%	$V_{/ɔ/}$ = 20.68%
$V_{/ɛ/}$ = 15.89%	$V_{/U/}$ = 14.83%
$V_{/æ/}$ = 16.66%	$V_{/u/}$ = 23.75%
$V_{/ə/}$ = 20.68%	$V_{/ɜ/}$ = 13.92%

(onde $V_{/v/}$ é a variação percentual máxima da vogal v em função de diferentes consoantes precedentes)

Ora, apenas a vogal /I/ apresenta uma variação **MENOR**, em função do contexto consonantal, do que a variação global intervocálica. Isso parece indicar que a consoante imediatamente anterior tem um maior peso na determinação do máximo de FO medido na vogal do que a própria qualidade dessa vogal. Essa constatação

impõe, é claro, certas limitações à postulação de uma relação simples e direta entre a qualidade de uma vogal e seu F0 intrínseco.

Acompanhando a variação no valor absoluto de F0, Lehiste e Peterson 1961 observaram um padrão regular no movimento de F0 ao longo da palavra-teste em função do contexto consonantal: após consoantes não-sonoras, o pico de F0 ocorria imediatamente após a realização da oclusão, ou seja, já no *onset* vocálico; nas vogais precedidas por consoantes sonoras, a freqüência fundamental subia lentamente, alcançando o valor máximo aproximadamente no centro da seção vocálica da palavra-teste CVC. Resultados semelhantes foram obtidos em outros estudos (House e Fairbanks 1953; Mohr 1971; Löfqvist 1975). Alguma evidência experimental derivada de testes perceptuais foi utilizada para sugerir que o padrão do movimento de F0 no *onset* vocálico era o fator decisivo para a distinção do traço "sonoridade" ma consoante em sílabas CV (Haggard *et al.* 1970, 1981).

Assim como no caso de F0 intrínseco de vogais, diferentes hipóteses foram sugeridas para explicar o fenômeno. Ladefoged 1967 esboça uma explanação aerodinâmica, observando que a produção de plosivas sonoras caracteriza-se por uma queda de F0 durante o período de oclusão, devido ao decréscimo da pressão sub-glotal. Dessa forma, imediatamente após a realização da oclusão, a freqüência fundamental tem um valor baixo no *onset* vocálico e adquire um padrão ascendente à medida que a pressão se

estabiliza. Por outro lado, em consoantes não-sonoras - diz Ladefoged - o maior fluxo de ar através da glote logo após a oclusão, cria uma força de Bernoulli acima do normal, aumentando a taxa de vibração das cordas vocais.

Outros pesquisadores tentaram explicar as mudanças de F0 no *onset* da vogal em função da tensão vertical da laringe. Segundo essa abordagem, nas plosivas sonoras as cordas vocais estariam mais frouxas, de modo a facilitar o vozeamento, ocorrendo o oposto para as plosivas frouxas, de modo a facilitar o vozeamento, ocorrendo o oposto para as plosivas não-sonoras; esses estados se estenderiam inercialmente ao longo da vogal adjacente modificando o padrão da freqüência fundamental (Hombert *et al.* 1979; Ohde 1984). Essa hipótese é consistente com dados que revelam uma posição mais baixa da laringe e do osso hióide nas oclusivas sonoras (Kent e Moll 1969; Ewan e Krones 1974).

As duas explanações não são mutuamente excludentes. O abaixamento da laringe não tem apenas o efeito de diminuir a tensão nas cordas vocais: se a laringe não abaixasse nas oclusivas sonoras, a pressão transglotal se equalizaria em cerca de 4-20 milisegundos e o pulso glotal seria interrompido, cessando o vozeamento. É sabido, no entanto, que os falantes conseguem produzir oclusivas sonoras com cerca de 80-100 ms de duração; isso só é possível porque o abaixamento da laringe aumenta a capacidade do trato vocal de absorver o fluxo de ar transglotal, adiando assim a equalização da pressão [1].Em alguns

casos pode haver alguma dificuldade para executar as manobras necessárias; Kent e Moll 1969 observam que a oclusiva /g/ freqüentemente apresenta falhas na barra de vozeamento. Eles sugerem que essa instabilidade parece estar relacionada à dificuldade de manter a laringe abaixada com a posição muito elevada da língua exigida para a produção de /g/, já que a língua e o sistema muscular que controla a laringe apresentam um certo grau de inter-dependência.

A questão do FO intrínseco não é de fácil equacionamento. A principal dificuldade é conseguir descrever as diversas interações entre a vogal e a consoante anterior. Como vimos acima, existem pelo menos dois fatores atuantes: os efeitos de co-articulação e a influência da própria vogal. Provavelmente a avaliação do peso relativo de cada um desses fatores na determinação do FO intrínseco depende também da consideração de variáveis suprasegmentais, muitas vezes negligenciadas. Não é surpreendente, pois, que alguns estudos não confirmem um padrão muito regular no comportamento da freqüência fundamental em função da consoante anterior. Umeda 1981, por exemplo, baseando-se em dados de leitura fluente, verifica que a direção do movimento de FO no *onset* vocálico "is not as reliable a voicing indicator of the preceding consonant as it is in isolated utterances" (pg.354). Também o pico de FO de diferentes vogais não revela um padrão previsível. Segundo Umeda, o assim chamado "FO intrínseco" é obscurecido por outros fatores na situação de

leitura fluente; o fenômeno seria um artefato resultante de condições experimentais baseadas na produção de palavras ou sentenças isoladas (v., no entanto, Ladd e Silverman 1984, para uma resposta a Umeda 1981).

Ohde 1984, mesmo usando sentenças isoladas como referência, não encontra evidência suportando uma dicotomia simples, movimento ascendente vs. descendente de F0 no *onset* vocálico como um invariante acústico correlacionado com o traço sonoridade. Os dados de Ohde 1984 revelam que quedas substanciais de F0 ocorrem no início da vogal, independentemente da qualidade sonora ou não-sonora da consoante precedente (v. também Silverman 1986).

Interação F0/Formantes

Na fala real há uma multiplicidade de fatores influindo no comportamento de F0: padrão rítmico-acentual, contorno entoacional, estrutura sintática, acento lexical, contraste tema/rema, etc. (v. por exemplo Bolinger 1986; Cruttenden 1986; Halliday 1967; Crystal 1969). A dificuldade em estabelecer um modelo que consiga integrar as influências combinadas desses diferentes contextos torna problemática a inclusão da freqüência fundamental na descrição das vogais. Por outro lado, não seria prudente descartar totalmente as perturbações de F0 ao nível segmental. A discrepância entre alguns resultados experimentais talvez indique apenas a atuação de variáveis não controladas

(Steele 1986). Não se deve esquecer que avaliações feitas com grande número de dados regularmente indicam uma tendência a um FO intrínseco mais elevado nas vogais altas, quando são considerados os valores médios. Há indícios de que as perturbações segmentais de FO sejam um efeito involuntário de natureza articulatória que, na fala natural podem ser compensados, mas apenas em uma certa medida, através de *feedback* auditivo. É significativo que, havendo algum impedimento de monitoração, como no caso dos deficientes auditivos, os efeitos tendam a ser maiores, e na mesma direção observada em falantes normais. (Bush 1981, *apud* Ternström *et al.* 1988). Outro aspecto a considerar diz respeito à grande sensibilidade do sistema auditivo para pequenas variações da freqüência fundamental; o limite mínimo de discriminabilidade perceptual para variações de FO é bastante baixo, da ordem de 0.3 Hz para vogais isoladas (Flanagan e Saslow 1958; Klatt 1973), e de cerca de 5 Hz para variações na região do acento nuclear em sentenças simples (Harris e Umeda 1987) [2].

As considerações acima revelam uma grande complexidade na interação entre os níveis segmental e suprasegmental na determinação da freqüência fundamental de uma vogal. É importante levar em conta, entretanto, que a validade do parâmetro "FO intrínseco" não é condição necessária para justificar a inclusão de FO entre as freqüências características de uma vogal (embora esse argumento já tenha sido usado como evidência indireta: v. Traunmüller 1988:5). Desconsiderando as particularidades de cada

teoria, para os pesquisadores que incluem F0 na descrição acústica de vogais (por exemplo: Syrdal e Gopal 1986; Traunmüller 1988; Miller 1989) o importante é menos o valor absoluto de F0 do que certas relações entre F0 e F1. Para a linha de pesquisa auto-intitulada "abordagem tonotópica", a distância crítica entre F0 e F1, expressa em unidades BARK (ou seja, a "distância tonotópica") é o parâmetro fundamental para a distinção do traço "abertura" em vogais; dessa forma, vogais fechadas como /i,I,u,U/ caracterizam-se por um intervalo MENOR que 3 BARK entre F1 e F0 (Syrdal e Gopal 1986, v. tabela III, pg. 1091). Já para Miller 1989, a freqüência fundamental entra na composição de um fator de normalização chamado pelo autor de "sensory reference" (SR) matematicamente definido como:

$$SR = 168\ (GMF0 / 168)^{1/3}$$

(onde GMF0 é a média geométrica do F0 atual - apenas no interior da seção vocálica; o valor fixo 168 e o expoente 1/3 são empiricamente determinados)

Com base em SR e na freqüência do primeiro formante, Miller define então um dos parâmetros para a descrição das vogais: y=log(F1/SR) (os outros parâmetros são: x=log(F3/F2) e z=log(F2/F1)).

Existem, contudo, algumas dificuldades, tanto de ordem teórica quanto empírica, para sustentar uma relação estreita

entre F0 e F1, mesmo que o valor absoluto da freqüência fundamental não esteja em jogo. No que diz respeito ao sistema de produção, tal colocação é problemática do ponto de vista da independência fonte/filtro: embora haja um certo acoplamento entre a fonte glotal e o trato vocal, essas interações são, em geral, desprezadas para todos os efeitos práticos, e o sistema é considerado como apresentando características lineares (Markel e Gray 1976). Modelos clássicos de produção (Fant 1960, por exemplo) consideram a interação irrelevante em função da impedância muito pequena do trato vocal em relação à impedância glotal.

Quanto ao aspecto perceptual, a postulação de uma ligação estreita entre a freqüência fundamental e os formantes colide com a distinção tradicional entre traços fonatórios e articulatórios, vistos geralmente como relacionados a propriedades perceptuais essencialmente independentes (Ladefoged 1971). Testes perceptuais revelam que os ouvintes não são capazes de distinguir vogais sintetizadas com e sem simulação da interação fonte/filtro (Nord *et al.* 1986).

Os primeiros experimentos de Dudley com o VODER e o VOCODER, combinando informação espectral derivada da fala real com espectros de fonte arbitrários foram considerados como evidência da independência perceptual entre as características da fonte e do filtro (v. "The carrier nature of speech": Dudley 1940).

Há outras indicações de uma relativa independência fonte/filtro no domínio perceptual. A fala em atmosferas carregadas de Hélio gasoso altera consideravelmente as freqüências dos formantes, duplicando-as em alguns casos; a freqüência fundamental, todavia, permanece praticamente inalterada. Apesar da modificação substancial nas relações entre FO e os formantes, a inteligibilidade da fala em ambientes com Hélio continua alta (Beil 1962; Morrow 1971).

Dentro de certos limites, alterações na velocidade de reprodução de fala em gravadores de fita também modificam as relações entre FO e a estrutura de formantes, sem perdas substanciais de inteligibilidade (Klumpp e Webster 1961).

Nearey 1989 observa que vozes pouco comuns como a do marinheiro Popeye dos desenhos animados podem apresentar um padrão anômalo na relação FO/formantes. Medindo esses parâmetros na voz de Jack Mercer (o dublador original de Popeye), Nearey verifica que os formantes equivalem às médias de um trato infantil, enquanto a freqüência fundamental encontra-se **ABAIXO** da média habitual para homens adultos (com base nos valores médios de Peterson e Barney 1952).

Apesar da ausência de vibração glotal periódica, a fala sussurrada é bastante inteligível, embora menos que a fala normal (Kallail e Emanuel 1984).

Outra evidência em favor da independência relativa entre FO e os formantes vem da observação das línguas tonais: se a

"distância tonotópica" entre F1 e F0 fosse um aspecto decisivo para caracterizar algumas vogais, F1 deveria acompanhar o movimento do contorno tonal de modo a evitar a ditongação. Nas línguas acentuais, também não é provável que F1 cubra realmente a mesma faixa de variação do contorno entoacional (em alguns casos mais de uma oitava: v. Lieberman 1967) de modo a manter invariante a distância de F0. Gottfried e Chew 1986, em um estudo sobre vogais cantadas, relatam que vogais produzidas com mudança de uma oitava na freqüência fundamental (de 130 para 260 Hz) mostram um aumento de apenas 10% em F1.

No plano neurológico, há pelo menos um caso individual onde se demonstrou que os gestos orais e laríngeos são dissociados. Gandour e Windsor 1988 relatam um distúrbio afásico no qual o paciente apresentava uma deficiência seletiva exclusivamente no sistema fonatório. Esse aspecto foi dramaticamente evidenciado quando se utilizou uma eletro-laringe que permitiu ultrapassar a região prejudicada. Falando com a laringe artificial, o paciente conseguia comunicar-se normalmente; a articulação supralaríngea mostrou-se normal e a apraxia dos comportamentos de fala desapareceu.

Flanagan 1972 descreve uma eletro-laringe que funciona através do contato externo com um ponto da garganta próximo da localização das cordas vocais. O controle da freqüência fundamental é efetuado manualmente por meio de um botão variável. Nessas condições, apesar do controle de F0 ser bastante precário,

a fala é inteligível, indicando que a relação FO/Formantes não é crítica.

Apesar da instabilidade da relação FO/formantes e das dificuldades teóricas envolvidas, existe, entretanto, um conjunto de evidências apontando efeitos de FO na percepção de vogais, embora a maior parte das observações tenha como ponto de partida dados de vogais sintetizadas. Lehiste e Meltzer 1973 cruzaram os FO médios de homens, mulheres e crianças com os padrões de formantes dos mesmos grupos, com base nos valores médios de cada grupo segundo os dados de Peterson e Barney 1952; dessa forma, para cada uma das dez vogais resultaram nove possíveis combinações de FO + Formantes. Os testes perceptuais indicaram as seguintes taxas de identificação:

		FORMANTES		
		homens	mulheres	crianças
FO	homens	76%	54%	44%
	mulheres	77%	82%	77%
	crianças	43%	43%	68%

Em um experimento similar, Ryalls e Lieberman 1982 também demonstram que os erros dos ouvintes são afetados por mudanças na freqüência fundamental. Estímulos com estruturas de formantes baseadas nos valores médios masculinos produzem um aumento significativo nos erros de identificação se a freqüência

fundamental é elevada de 135 Hz (média normal masculina) para 250 Hz; para mudanças na direção oposta - de 135 para 100 Hz - não há alteração no padrão de respostas. Para as vogais baseadas nas médias dos formantes do grupo feminino, os erros perceptuais aumentam se F0 é alterado de 185 Hz (média normal feminina) para 250 Hz; se a fundamental desce para 100 Hz, os erros também aumentam, embora menos que nos estímulos com F0=250 Hz.

Os dois estudos indicam que, em geral, disparidades maiores entre F0 e Formantes resultam em um aumento dos erros perceptuais. Observa-se, no entanto, uma assimetria nos resultados: um decréscimo de F0 em relação ao valor normal tem efeito menos prejudicial na percepção do que um acréscimo. Ryalls e Lieberman 1982 argumentam que essa diferença pode estar relacionada à possibilidade de fazer uma amostragem mais densa do espectro quando a fundamental é baixa, o que permite ao ouvinte uma extração de formantes mais precisa.

Alguns experimentos evidenciam efeitos de F0 na categorização de vogais sintetizadas. Miller 1953 demonstra que a duplicação de F0 (de 144 para 288 Hz) provoca um deslocamento na categorização de algumas vogais perto de fronteiras perceptuais, mesmo se o envelope espectral é mantido constante. Miller estima um deslocamento de 80 Hz (cerca de 16%) na fronteira de F1 entre /ɷ/ e /ʌ/ para uma alteração de uma oitava em F0; na fronteira /I/ - /ɛ/ o deslocamento de F1 é de 30 Hz (cerca de 6%) para a mesma alteração de F0. Carlson *et al*. 1975 relatam uma relação

monotônica entre F1 e F0 de modo a manter a identidade fonética com F0 crescente. Traunmüller 1981 apresenta uma série de experimentos explorando a relação entre F0 e F1 na percepção; ele conclui que as fronteiras perceptuais definidas por F1 - isto é, na dimensão "abertura" - são fortemente afetadas por F0. Seus resultados indicam um aumento de cerca de 28% em F1 na faixa de 500 Hz correspondendo a uma elevação de uma oitava na freqüência fundamental (de 130 para 260 Hz). Vários outros estudos apontam na mesma direção (Fujisaki e Kawashima 1968; Slawson 1968; Ainsworth 1975).

Há pouca evidência experimental quanto à importância perceptual de F0 na identificação de vogais produzidas naturalmente. Parte dessa evidência vem da observação de vogais produzidas no canto. O já citado estudo de Gottfried e Chew 1986, embora tenha observado um acréscimo de apenas 10% em F1 para um salto de oitava em F0, verificou, por outro lado, um aumento no índice de erros perceptuais proporcional ao aumento da freqüência fundamental da voz de contra-tenor usada na produção dos estímulos. Carlson *et al.* 1975 ressaltam que em registros extremamente altos - uma voz de soprano, por exemplo - pode haver uma perda substancial da identidade fonética da vogal cantada [3].

Alguma evidência indireta da importância de F0 na caracterização de vogais naturais pode ser derivada de trabalhos que focalizam a distribuição de um grande número de vogais. Foulkes 1961, através de uma série de transformações matemáticas

dos dados de Peterson e Barney 1952, verifica que a inclusão de um fator de correção baseado no F0 medido em cada vogal diminui consideravelmente a sobreposição de categorias e aumenta as identificações corretas automáticas. Assmann *et al.* 1982, trabalhando com um modelo estatístico de reconhecimento de padrões, relatam que a inclusão da freqüência fundamental entre os parâmetros de controle aumenta a correlação entre as previsões do modelo e os julgamentos subjetivos de um grupo de ouvintes.

F0 como Fator de Normalização

O papel real desempenhado pela freqüência fundamental na percepção de vogais, como demonstram as evidências freqüentemente conflitantes expostas acima, é uma questão que, pelo menos no estágio atual da pesquisa, parece estar longe de ser inteiramente resolvida. Assim como em relação a outros aspectos da percepção de vogais - e de sons de fala em geral - é provável que as dificuldades tenham como causa principal as quase inevitáveis restrições contextuais características de cada *design* experimental. Como já comentamos anteriormente, diferentes tarefas podem acionar diferentes estratégias perceptuais; sendo assim, não é tão surpreendente a eventual incompatibilidade de alguns resultados. A divergência, no entanto, pode ser apenas aparente, e resultados conflitantes podem estar, afinal, "corretos", dependendo do ponto de vista adotado.

No caso específico da influência de F0 na percepção de vogais, é preciso ressaltar que o mecanismo perceptual pode extrair vários tipos de informação desse parâmetro. No entanto, não é necessário - e nem seria possível - que essas informações estejam TODAS presentes em cada vogal realizada. Um parâmetro como o "F0 intrínseco", por exemplo, pode eventualmente ser mascarado por variações entoacionais mais amplas, o que não significa que o fenômeno, de forma geral, inexista ou seja irrelevante em todas as situações. O mesmo poderia ser dito em relação à distância crítica F1-F0 postulada pela "abordagem tonotópica" acima discutida: o fato de variações entoacionais muito bruscas impossibilitarem localmente a manutenção da distância acústica F1-F0 exigida pela teoria não elimina a eficácia da pista em condições menos extremas.

Devido à magnitude das variações suprasegmentais, é pouco provável que encontremos associações invariantes entre a freqüência fundamental e a qualidade vocálica. Há um outro tipo de informação derivada de F0 que pode, porém, ser usada pelo ouvinte com alguma segurança. Embora não exista uma vinculação anatômica necessária entre o trato e a conformação das cordas vocais, a freqüência fundamental média é, em geral, um bom indicador das dimensões globais do trato supralaríngeo, correlacionando-se negativamente com essas dimensões [4]. O ouvinte sabe disso e é capaz de usar essa informação para estabelecer uma referência perceptual com efeito de normalização.

A relação entre F0 e as dimensões do trato é evidenciada em estudos que analisam um grande número de vogais produzidas por homens, mulheres e crianças. Em geral, observa-se, para cada vogal isoladamente, uma correlação positiva entre os valores da freqüência fundamental e os valores dos formantes (Potter e Steinberg 1950; Fujisaki e Kawashima 1968). Segundo Fujisaki e Kawashima 1968, a relação entre F0 e o n-ésimo formante pode ser aproximada por uma equação linear:

$$F_n = a_n\ (F0 + b_n)\ Hz,$$

(onde a_n e b_n são constantes ajustadas para cada vogal)

Infelizmente Behlau 1984 não publicou os valores da freqüência fundamental para cada vogal observada no seu estudo sobre o PB. Foram feitas, entretanto, medidas de F0 para a emissão estável da vogal /ä/ de cada um dos 90 falantes (Behlau 1984: 78-86). Como esse era o único dado disponível, consideramos, para os cálculos abaixo descritos, essas medidas como representativas do F0 médio de cada falante. Com base nesse F0 médio estimado e nos valores médios dos formantes de cada falante [5], foi realizada uma análise de variância (ANOVA) modelando a variável GRUPO em três classes (feminino, masculino e crianças). A tabela 5.1 apresenta os resultados.

modelo: Grupo	Graus de liberdade:2		
VARIÁVEL DEPENDENTE	r^2	F	p<
FO_m	.752	131.94	.0001
$F1_m$	.652	81.73	.0001
$F2_m$	.691	97.56	.0001
$F3_m$	.643	48.70	.0001

TABELA 5.1

Os dados da tabela 5.1 indicam que, de acordo com a expectativa, FO é o parâmetro que melhor explica a variância inter-grupo (maior valor de r^2). No entanto, observa-se que também as médias individuais dos formantes são estatisticamente significativas para caracterizar os diferentes grupos. Curiosamente, o menor valor de r^2 está associado a $F3_m$, indo contra a expectativa de que o terceiro formante varia relativamente pouco de vogal para vogal, estando mais regularmente correlacionado com as dimensões do trato (Cf.Fant 1960; Fujisaki e Kawashima 1968). Esse resultado um tanto inesperado pode estar relacionado, em parte, à não-homogeneidade do grupo de crianças em Behlau 1984. esse grupo era constituído de 15 crianças do sexo masculino e 15 do sexo feminino, com faixa etária variando entre 8 e 12 anos de idade. Dificilmente, nessas condições, encontraríamos uma distribuição uniforme no interior desse grupo: uma menina de 12 anos, por exemplo, já apresenta um

padrão de formantes semelhante ao de uma mulher adulta (Kent e Forner 1979). Também o grupo masculino apresenta alguns problemas. É geralmente reconhecido que homens e mulheres diferem quanto ao processo de maturação física, ou seja, indivíduos do sexo feminino atingem mais cedo a conformação física estável da fase adulta (Waber 1980). Behlau 1984:29 relata que a faixa etária dos indivíduos adultos foi delimitada entre 18 e 45 anos de idade. É provável que, no grupo masculino, essa faixa seja demasiadamente ampla para produzir uma distribuição homogênea.

A relativa eficiência da freqüência fundamental como previsor das dimensões do trato pode ser evidenciada nos dados de Behlau 1984, pelas correlações rezoavelmente altas entre F0 e as médias individuais de cada formante. Os coeficientes de correlação com $F1_m$, $F2_m$ e $F3_m$ são respectivamente, .69, .70 e .70 ($p<.0001$ em todos os casos).

Como a média de cada formante também parece explicar parte da variação inter-grupo (v.tabela 5.1), é possível que parâmetros mistos, considerando tanto F0 quanto F_n médios, sejam mais eficientes para caracterizar diferentes grupos de falantes. De modo a testar essa hipótese, definimos algumas variáveis onde diferentes pesos foram tentativamente atribuídos a cada um dos parâmetros isolados. Os resultados de ANOVA para essas combinações lineares, expostos na tabela 5.2 .indicam que é possível aumentar o valor de r^2. Como o método de atribuição de pesos foi absolutamente informal, é provável que um ajuste mais

rigoroso produza resultados mais significativos.

Modelo:Grupo Graus de Liberdade:2

VARIAVEL DEPENDENTE	r^2	F	p<
k_1	.785	99.07	.0001
k_2	.791	102.42	.0001
k_3	.802	109.64	.0001

onde $k_1 = (FO_m + F1_m + F2_m + F3_m)/4$

$k_2 = (3 \times FO_m + 2 \times F3_m + F2_m + F1_m)/7$

$k_3 = (4 \times FO_m + 2 \times F3_m + 2 \times F2_m + F1_m)/9$

TABELA 5.2

Sumário

Resumindo, diríamos que há vários parâmetros relacionados com FO que podem ter alguma influência na percepção de vogais, embora haja alguma controvérsia quanto à natureza e universalidade de algumas dessas pistas. Assim como ocorre com a estrutura de formantes, parte da informação vem de propriedades intrínsecas da vogal, e parte de fatores extrínsecos, ou seja, distribuídos ao longo do ambiente fonético precedente. Como

parâmetro intrínseco, F0 contribui de duas maneiras, em parte inter-relacionadas:

1) na medida em que se observa uma variação quase-regular - se considerados valores médios de um número razoável de dados - entre o valor absoluto de F0 e a qualidade da vogal: vogais fechadas têm, em geral, F0 intrínseco mais alto que vogais abertas.
2) na medida em que F1 interage com F0, sendo a distância acústica entre esses dois parâmetros uma das pistas para distinções na dimensão "abertura vocálica".

Como fator de normalização extrínseco, a freqüência fundamental média provavelmente funciona como uma referência perceptual para a avaliação das dimensões do trato do falante.

A divergência entre alguns estudos não deve ser encarada como prova da inconsistência de F0 como freqüência característica das vogais. Os contra-exemplos baseiam-se geralmente em casos excepcionais e apenas indicam que o mecanismo perceptual humano está apto a processar fala mesmo em condições anômalas, com base em outros parâmetros não diretamente relacionados à freqüência fundamental. Vozes como a do marinheiro Popeye, vozes alteradas pela atmosfera de Hélio ou vozes produzidas por excitação de eletro-laringe estão longe de corresponder a um padrão normal, e essa estranheza é facilmente percebida pelos nossos ouvidos.

O uso de F0 no processo de normalização será melhor entendido como uma OPÇÃO perceptual, uma pista eficiente em algumas situações, mas não necessária, nem suficiente para a identificação de vogais. Quando a tarefa específica envolve, por exemplo, o reconhecimento muito rápido de monossílabos, a informação derivada de F0 parece ser irrelevante (v. Summerfield e Haggard 1975), talvez pelo fato de o ouvinte não ter tempo suficiente para avaliar a freqüência fundamental.

6

ASPECTOS DINÂMICOS

As abordagens analisadas até aqui ao longo deste trabalho, embora possam divergir quanto a alguns pontos específicos, possuem uma característica comum: pressupõem que uma configuração espectral estática preserva a informação acústica essencial para a identificação de vogais. Cada vogal é, afinal, uma forma canônica, representada como um ponto em um espaço acústico multidimensional, cujas coordenadas são os dois ou três primeiros formantes. Como esse tipo de representação produz inevitavelmente sobreposições de categorias - especialmente em grupos heterogêneos de falantes - a informação derivada da estrutura de formantes deve ser de alguma forma compensada - ou "normalizada" - para neutralizar variações inter-falante. Em geral, a sobreposição de categorias, para esses modelos, é encarada como uma projeção errada no espaço errado: se certas transformações no padrão de formantes e - para alguns - em F0 forem aplicadas, com base em algum fator intrínseco ou extrínseco de normalização, acredita-se que limites categoriais mais definidos serão obtidos.

Esse paradigma pode ser razoavelmente eficiente quando se trata de vogais produzidas em isolamento ou em certos contextos controlados. Na análise da fala real, entretanto, surgem várias dificuldades para uma especificação estática das vogais. O

simples exame de um espectrograma de um trecho de fala produzido em velocidade normal ou rápida revela que algumas sílabas podem não conter uma porção suficientemente estável passível de ser claramente identificada como representativa da vogal (mesmo que não se trate de ditongação). Nessas situações nos defrontamos com o clássico problema da segmentação, ou seja: em que ponto devemos realizar a seção espectral que traduza a representação paramétrica mais adequada do segmento vocálico intencionado pelo falante?

"Target Undershoot"

Estudos clássicos como os de Lindblom 1963 e Stevens e House 1963 demonstraram que, na fala fluente, quando vogais são coarticuladas com consoantes, os alvos ("targets") acústicos/articulatórios canônicos freqüentemente não são atingidos. Embora a evidência apresentada baseie-se em dados de apenas um falante, Lindblom 1963 relata variações de até 70% nos formantes de uma vogal, atribuíveis a efeitos de coarticulação com as consoantes plosivas circundantes em sílabas CVC. Stevens e House 1963 mostram que vogais coarticuladas em sílabas CVC, combinadas com 14 consoantes diferentes, apresentam trajetórias de formantes em contínua mudança, cujos máximos de freqüência não coincidem com as medidas em vogais isoladas.

Lindblom e Studdert-Kennedy 1967, em outro estudo clássico, sugerem que os ouvintes interpretam os *targets* de vogais coarticuladas em função da direção e velocidade das transições marginais. Eles utilizam sílabas CVC geradas sinteticamente e demonstram que as fronteiras perceptuais entre/I/ e /ʊ/ se deslocam em função da direção da transição de F2 e da duração silábica total. Os autores, com base nesses resultados e em resultados de estudos anteriores, generalizam esses efeitos, propondo um modelo que prevê um déficit (*undershoot*) na produção, que seria compensado pelo ouvinte através de um superávit (*overshoot*) perceptual. Em outras palavras: na fala normal - e especialmente na fala rápida - os efeitos de coarticulação impedem que as posições articulatórias canônicas correspondentes a uma vogal sejam alcançadas, de modo a produzir uma estrutura de formantes típica da vogal intencionada; o ouvinte, no entanto, seria capaz - segundo os autores - de neutralizar essa distorção através de uma extrapolação das transições, recompondo assim, virtualmente, a representação acústica prototípica da vogal intencionada.

Essas diferenças sistemáticas nos padrões acústicos dinâmicos podem ser potencialmente importantes do ponto de vista perceptual: diferentes vogais, eventualmente sobrepostas no espaço acústico definido pelos formantes reais, talvez pudessem ser diferenciadas em termos das trajetórias dos formantes definidas pelas transições.

Alguns estudos verificaram que, de fato, vogais coarticuladas eram melhor identificadas que vogais produzidas em isolamento, mesmo se o grupo de ouvintes fosse constituído de sujeitos foneticamente pouco sofisticados, e apesar de uma considerável ambigüidade acústica nas freqüências dos formantes medidas no ponto de maior aproximação ao *target* (Strange *et al.* 1976; Gottfried e strange 1980). Verbrugge et al. 1976 observam que trechos silábicos, ou ainda menores, extraídos da fala fluente, são melhor identificados se parte do contexto circundante é também incluído no sinal [1].

Segundo o modelo proposto por Lindblom 1963 e Lindblom e Studdert-Kennedy 1967, os ouvintes deveriam compensar, de forma quase regular, os *undershoots* articulatórios/acústicos inerentes à produção de vogais coarticuladas. A questão da correspondência complementar entre produção e percepção, prevista pelo modelo do *target undershoot*, tem sido alvo, porém, de alguma controvérsia. Experimentos baseados no paradigma de Lindblom/Studdert-Kennedy, embora revelem efeitos sistemáticos do contexto consonantal na identificação de vogais em sílabas CVC, não confirmam, entretanto, a mesma magnitude das compensações perceptuais observadas em Lindblom 1963 (Nearey 1989). Há vários fatores que modificam o grau de *undershoot* articulatório na fala fluente. O modelo original (Lindblom 1963) previa que os *targets* poderiam ser determinados em função da duração do segmento e da direção das regiões transicionais. Verificou-se, porém, que os falantes

são relativamente livres para variar o grau de *undershoot*, de forma independente da duração; na fala rápida, os *targets* acústicos/articulatórios podem ser atingidos apesar do encurtamento das vogais. Alguns falantes são capazes de reprogramar a organização articulatória, aumentando a velocidade dos movimentos dos articuladores, de modo a diminuir ou eliminar a redução vocálica (Gay 1978; Engstrand 1988). Diferentes indivíduos usam diferentes estratégias de coarticulação; na verdade, essas características pessoais são, em alguns casos, tão marcantes, que podem servir como um parâmetro razoavelmente eficiente para a identificação do falante em sistemas automáticos (Nolan 1983).

O estilo de fala parece também influir no grau de redução; quando os sujeitos são instruídos explicitamente no sentido de realizar as seqüências teste com a pronúncia mais clara possível, os valores dos formantes aproximam-se dos *targets* canônicos, mesmo que essa fala hiperarticulada seja produzida rapidamente (Lindblom e Moon 1988). Essa constatação é compatível com a hipótese de que parte da redução vocálica ocorrendo em fala rápida talvez **NÃO SEJA** compensada perceptualmente pelos ouvintes e implique simplesmente uma perda parcial de contraste fonético (Koopmans-van Beinum,1980;*apud* Nearey 1989).

Invariância no Movimento

De modo a examinar mais de perto a relevância perceptual das fontes dinâmicas de informação, Strange *et al.* 1983 desenvolveram uma técnica posteriormente usada em uma série de experimentos focalizando os efeitos de coarticulação. Eles modificam sistematicamente sílabas CVC produzidas naturalmente e então digitalizadas. Uma das manipulações cria sílabas onde a seção vocálica central mais estável é substituída por um intervalo de silêncio. Testes perceptuais revelaram que esses estímulos são bem identificados pelos ouvintes, embora mantenham apenas as transições inicial e final, sem nenhuma informação do *target*. O reconhecimento das vogais permanece alto mesmo quando 90% da seção central é retirada e substituída por igual duração de silêncio ou ruído neutro, deixando apenas um ou dois períodos de F0 nas margens silábicas, e os erros perceptuais, quando ocorrem, consistem em confusões com a categoria vocálica espectralmente mais próxima da vogal intencionada (Parker e Diehl 1984).

Esses resultados podem ser interpretados à luz de duas teorias bastante diferentes, embora não necessariamente excludentes. Uma possível explanação, seguindo o modelo de Lindblom 1963, sugeriria que as transições mantidas nas margens das sílabas com silêncio central fornecem informação suficientemente robusta para que o ouvinte seja capaz de extrair o *target* intencionado. Uma hipótese alternativa diria que a

informação dinâmica nas vogais é complementar, mas essencialmente distinta da informação estática definida pelos possíveis *targets*. A segunda hipótese inspira-se na concepção das vogais como GESTOS; segundo esse ponto de vista, as vogais são EVENTOS articulatórios essencialmente dinâmicos que manifestam uma organização de forças sobre os articuladores que é única para cada uma das vogais de um dialeto. A interação entre essas forças dá origem a um estilo particular de movimento articulatório e a um padrão acústico dinâmico correspondente; esse padrão veicula informação basicamente diferente da informação presente nos *targets* do núcleo silábico. Essa colocação insere-se no paradigma da "action theory" e tem encontrado sua melhor expressão nos trabalhos de Fowler 1980, 1986, 1987 (v. também Browman e Goldstein 1986; Johnson 1987).

De modo a testar a força explanatória de cada uma das hipóteses acima descritas, Rakerd e Verbrugge 1987 projetaram um experimento baseado na construção de estímulos HíBRIDOS, onde são combinadas a porção inicial de sílabas /bVb/ faladas por uma mulher, com a porção final de sílabas equivalentes faladas por um homem, sendo os dois trechos separados por um intervalo de silêncio. Segundo a hipótese "extração de target" (Lindblom 1963), um estímulo desse tipo provocaria certamente grande confusão perceptual: como as dimensões dos tratos masculino e feminino são diferentes, os *targets* também seriam diferentes, e a sílaba apresentaria, portanto, informação conflitante para o

ouvinte (se de fato é informação *target* a veiculada pelas transições). Por outro lado, segundo a hipótese "percepção holística do evento" defendida por Fowler, uma discrepância entre os *targets* atuais ou extrapolados dos dois falantes não deve causar, necessariamente, grandes distúrbios perceptuais:como os falantes, apesar das diferenças anatômicas, pertencem ao mesmo grupo dialetal, espera-se que produzam uma determinada vogal com um estilo comum de movimento articulatório. A teoria prevê que o ouvinte é capaz de integrar perceptualmente esses "gestos" comuns, e a precisão da identificação dessas sílabas híbridas deve ser tão alta quanto para sílabas com silêncio central produzidas por um único falante. De fato, Rakerd e Verbrugge 1987 verificam que não há diferença significativa no reconhecimento dos dois tipos de estímulo, ambos produzem o mesmo percentual de erros de identificação. Esses resultados são mais compatíveis com o modelo "holístico" de percepção; no entanto, é preciso considerar que a eliminação da informação do núcleo vocálico mais estável provoca, em geral, um aumento das confusões perceptuais, indicando que a informação do tipo *target* TAMBÉM tem sua parcela de importância. no processo perceptual.

Outra propriedade dinâmica analisada em pesquisas recentes é a mudança espectral intrínseca: já foi observado que vogais isoladas - pelo menos no inglês americano e canadense - raramente apresentam uma trajetória estável dos formantes, mesmo que se trate de supostos monotongos (Strange 1989; Nearey 1989). Nearey

e Assmann 1986 sugerem que as terminações das trajetórias dos formantes em vogais isoladas (isto é, não coarticuladas) podem servir como pistas dinâmicas para a identificação. Eles realizam um experimento onde vogais naturalmente produzidas são sinteticamente manipuladas e divididas em duas porções de 30 ms: uma centrada em um ponto a 24% da duração total do segmento (que chamaremos de A) e outra centrada a 64% (que chamaremos de B). A partir desses elementos, três tipos de estímulo são construídos: A-B, A-A e B-A, onde as porções inicial e final da vogal são combinadas inserindo-se um intervalo central de silêncio com 10 ms de duração. Os resultados mostraram que os estímulos A-B, que mantinham a ordem natural das terminações da trajetória de cada formante, produzem menos erros que as condições A-A e B-A, e são tão bem identificados quanto as vogais originais.

Esses resultados sugerem que o ouvinte, de alguma forma, faz uso da informação derivada da variação do padrão espectral ao longo da vogal. É importante ressaltar que esse aspecto não está diretamente relacionado aos efeitos de coarticulação e nem ao processo de extração de *targets*, e parece constituir uma pista autônoma. Resta saber em que medida o efeito persiste em contextos envolvendo coarticulação e se o fenômeno tem um caráter universal, não se restringindo apenas ao inglês.

Duração Intrínseca

A duração intrínseca é outro aspecto dinâmico que tem uma relação quase-sistemática com a qualidade vocálica. Tem sido observado que, em geral, vogais abertas tendem a ser mais longas (Peterson e Lehiste 1960). Essa relação é freqüentemente atribuída a características inerciais do movimento da língua e, especialmente, da mandíbula: a maior duração das vogais abertas seria uma conseqüência direta das trajetórias articulatórias mais extensas envolvidas na sua produção (Lehiste 1970).

A importância perceptual da duração segmental foi evidenciada em alguns experimentos: a neutralização da informação relativa à duração do núcleo silábico em sílabas CVC sintetizadas faz aumentar o número de erros de identificação da vogal (Strange *et al.* 1983; Strange 1987). Embora o fenômeno tenha sido mais extensivamente estudado no inglês, há trabalhos que verificam o efeito perceptual da duração intrínseca em várias línguas diferentes (v. Delgado Martins 1986).

Não temos conhecimento de dados para o PB; no Português de Portugal, porém, há alguma evidência estatística confirmando a tendência geral de uma menor duração intrínseca associada às vogais fechadas: Delgado Martins 1975, com base nos valores médios extraídos da produção de várias vogais por um único falante, relata a seguinte ordem decrescente de duração intrínseca:

/ɛ/>/ɔ/ = /o/>/a/>/e/>/u/>/i/

Assim como já observamos em relação ao "FO intrínseco" (v. seção 5), também a duração intrínseca sobre influências tanto do contexto fonético quanto das variações rítmico/acentuais impostas pelo nível suprasegmental. Desde o trabalho clássico de Fry 1955 sobre o acento fonético, sabe-se, por exemplo, que a duração é um dos correlatos mais constantes da tonicidade: mantidos fixos os demais fatores, uma vogal acentuada é mais longa do que uma vogal não acentuada.

O contexto consonantal também afeta a duração da vogal precedente. O alongamento relativo das vogais antes de consoantes sonoras já foi observado há algum tempo (House e Fairbanks 1953; Peterson e Lehiste 1960; House 1961), e pode ser considerado um traço quase universal, presente em línguas bastante diferentes (v. Kluender *et al.* 1988; v.,no entanto, Crystal e House 1988, abaixo, para alguma evidência conflitante).

Crystal e House 1988 apresentam resultados parciais no bojo de um ambicioso projeto focalizando a duração de vogais no inglês americano. Esse estudo baseia-se em um grande número de dados (30 falantes, cerca de 4000 vogais produzidas); além disso, destaca-se dos demais estudos pelo fato de ter sido exercido um maior controle sobre os diversos efeitos capazes de interferir na duração segmental. Crystal e House confirmam a existência de características duracionais inerentes diferenciando grupos de

vogais na dimensão "abertura". Consistentemente com observações anteriores, eles verificam também que vogais acentuadas tendem a ser mais longas do que vogais não acentuadas, e, mais especificamente, que pelo menos 3 graus de tonicidade (acento primário, acento secundário, e ausência de acento) diferem quanto à duração. Curiosamente, colidindo com a farta evidência existente, os dados de Crystal e House não mostram claramente um aumento da duração vocálica em função do traço sonoridade na consoante subseqüente. Alguns efeitos relacionados ao ponto e modo de articulação da consoante pós-vocálica também foram observados, mas devido às interações com o fator tonicidade, os resultados são de difícil interpretação: vogais precedendo consoantes labiais, por exemplo, tendem a ser mais longas do que vogais antes de alveolares ou velares. Mas se as condições de tonicidade são controladas, o quadro torna-se mais confuso: vogais não acentuadas não se diferenciam em termos de duração em função da qualidade da consoante subseqüente, e vogais acentuadas são regularmente mais curtas antes de velares e mais longas antes de labiais e alveolares. Além dos fatores acima descritos, Crystal e House relatam ainda uma influência relacionada à organização lexical/sentencial da fala: a presença de um limite de palavra ou de uma pausa após uma vogal parece provocar um aumento na duração dessa vogal, embora o efeito não tenha a mesma magnitude para todas as qualidades vocálicas.

A multiplicidade de fatores segmentais e suprasegmentais influindo na duração vocálica torna problemática a inclusão desse parâmetro em sistemas de reconhecimento automático. Embora exista bastante evidência acústica e perceptual quanto à eficácia do parâmetro duração relacionado a diferentes usos lingüísticos, inexiste, por enquanto, um modelo suficientemente abrangente capaz de integrar os diversos modos de interação entre esses usos. Os poucos estudos abordando diretamente a questão esgotam-se em uma mera descrição de efeitos isolados, chegando, no máximo, a descrever interações entre dois fatores: tonicidade X duração, contexto consonantal X duração, etc. (v. Klatt 1976; Port 1981; Crystal e House 1982,1988).

Na verdade, mesmo se conhecidas certas restrições contextuais, a situação de fala real acrescenta outros elementos de variação que tornam o quadro ainda mais complexo. Na fala rápida, por exemplo, as pistas duracionais sofrem um processo de neutralização que praticamente elimina certas oposições (Gay 1978). Outro problema diz respeito à variação inter-falante, que em alguns casos pode ser maior do que as diferenças relacionadas à influência do contexto fonético (Crystal e House 1982).

Embora o atual estágio da pesquisa não permita ainda o desenvolvimento de esquemas "bottom-up" baseados nas características de duração, algumas dificuldades podem ser parcialmente superadas em sistemas automáticos de reconhecimento. A exigência de um estilo mais cuidadoso de fala e/ou uma fase

prévia de treinamento da máquina para um falante específico podem aumentar a eficiência da pista "duração intrínseca", especialmente se a influência do contexto fonético é controlada (Deng *et al.* 1989).

Postura vs. Gesto

A inclusão de fatores dinâmicos inerentes à produção de vogais promoverá, certamente, uma melhora significativa nos sistemas de reconhecimento automático. As consequências dessa especificação dinâmica das vogais ultrapassam, no entanto, as necessidades pragmáticas mais imediatas, e inserem-se em uma discussão teórica mais abrangente, que aponta para um rompimento com algumas noções tradicionais sobre a estrutura da fala. Mais especificamente, surgem algumas dificuldades em relação ao modelo segmentalista proposto no conhecido trabalho de Jakobson, Fant e Halle 1951, cujas diretrizes teóricas orientaram grande parte da pesquisa fonológica na segunda metade do século. Segundo esse modelo, um segmento fonológico é representável em termos de um "pacote" de traços distintivos, sendo os contrastes definidos pela oposição entre um ou mais traços; presume-se que cada um desses traços tenha um correlato articulatório e/ou acústico que deve estar presente **EM ALGUM MOMENTO** de modo a caracterizar o segmento. Esses conjuntos de traços simultâneos constituiriam unidades elementares temporalmente coerentes (pelo menos em um

certo intervalo de tempo, mesmo que esse intervalo seja curto), funcionando basicamente como "blocos de construção" da cadeia de fala; cada um desses segmentos - ou fonemas - possuiriam propriedades inerentes suficientes para garantir sua substanciação como unidades físicas independentemente realizáveis.

Estaria fora do escopo do presente trabalho discutir exaustivamente as limitações dos modelos pós-Jakobsonianos. Alguns pontos, entretanto, merecem ser examinados mais de perto, já que a aceitação não-crítica do segmentalismo pode nos levar a procurar invariâncias onde elas não existem de fato. O recente desenvolvimento de técnicas radiográficas computadorizadas, envolvendo baixa exposição à radiação, permitiu um exame bastante minucioso dos movimentos articulatórios durante a produção de fala. Uma das descobertas mais reveladoras dessa pesquisas é que a atividade articulatória não tem uma relação simples com a cadeia segmental acusticamente definida. Observa-se freqüentemente uma defasagem do pico de atividade de um articulador em relação ao segmento supostamente caracterizado por essa atividade específica. Fujimura 1980, observando radiograficamente a produção da sílaba /mowst/, verifica que a constrição labial para o *glide* /w/ manifesta seu máximo durante o período não-sonoro do *cluster* consonantal /st/. Vaissière 1988, em um meticuloso estudo radiográfico do movimento do velum em consoantes nasais, relata que a abertura máxima do conduto para a

cavidade nasal freqüentemente ocorre no segmento vocálico que precede a consoante, isto é, antes da oclusão concomitante com o murmúrio nasal; além disso, Vaissière observa que a altura absoluta do velum é menos importante do que a DIREÇÃO do movimento, ou seja, o traço [+/-nasal] não é definido pela abertura vélica mas antes pela relação entre essa abertura e a posição imediatamente anterior.

Essas constatações são bastante surpreendentes para uma abordagem que modela a fala como uma sequência linearmente concatenada de segmentos. O fato é que a caracterização do segmento fonológico em termos de uma configuração articulatória ESTÁTICA não consegue dar conta daquilo que se observa na fala real; a fala é um evento essencialmente dinâmico, que se realiza e desenvolve NO tempo. Negligenciar o aspecto temporal parece ter sido o grande equívoco das descrições baseadas em sistemas de traços distintivos. O quadro teórico da fonologia gerativa logo se deparou com uma série de evidências conflitantes; dificuldades surgiram, principalmente, em conexão com a escolha adequada do conjunto de traços, com a não-ortogonalidade entre alguns desses traços e com decisões sobre a dimensão binária ou *n*-ária de alguns traços (v. Liljencrants e Lindblom 1972; Ohala 1985). Embora bastante esforço tenha sido canalizado para a solução desses problemas, especialmente no sentido de estabelecer sistemas de traços cada vez mais "universais", a formulação básica permaneceu irretocada e as freqüentes discrepâncias entre

a descrição formal e a atividade real dos articuladores ficaram por conta dos perniciosos efeitos da "performance". De acordo com a perspectiva gerativista, o mecanismo articulatório - devido a limitações físicas do trato vocal - inevitavelmente distorce a realização dos segmentos discretos ideais: na passagem da mente para o mundo, os fonemas abstratos seriam - para usar a metáfora de Hockett (*apud* Fowler 1987) - "moídos" pelos efeitos da coarticulação na fala real.

Ora, parece contra-intuitivo caracterizar a atividade articulatória como fundamentalmente destrutiva: a comunicação pela fala, afinal, é altamente eficiente, e seria, no mínimo, extravagante imaginar que sua produção fosse tipicamente imperfeita. Mesmo admitindo a possibilidade de uma representação estática através de um conjunto de traços, restaria saber como interpretar os eventos transientes intermediários que ligam os supostos estados canônicos: seriam esses eventos - como parece supor o paradigmaa gerativista - irrelevantes, ou seja, não integrantes do enunciado originalmente intencionado pelo falante? Se isso fosse verdadeiro, deveríamos esperar que a melhor informação para a identificação dos segmentos fosse fornecida pelas regiões mais estáveis - isto é, menos coarticuladas - do sinal acústico. O fato é que - como vimos em alguns experimentos acima descritos - a informação espectral extraída do centro temporal de uma vogal parece ser **MENOS** eficiente do que a informação dinâmica derivada das transições às margens dessa

vogal (v. por exemplo, Strange *et al.* 1976, 1983). Algo semelhante ocorre com a identificação de consoantes; estudos confrontando a informação espectral no momento da liberação da explosão com a informação fornecida pelas transições dependentes da vogal em sílabas CV demonstram que os ouvintes quase sempre identificam as consoantes plosivas com base nas transições, e não com base no espectro no momento da explosão (Blumstein *et al.* 1982; Walley e Carrell 1983). Um estudo recente (Diehl *et al.* 1987) sugere que o reconhecimento da consoante inicial em sílabas /(b,d,g)Vs/ é dependente da vogal, e que uma porção mais ou menos extensa da trajetória dos formantes da vogal precisa ser avaliada pelo ouvinte antes que as consoantes possam ser identificadas com segurança - na verdade uma porção que vai bem além dos movimentos transicionais mais abruptos imediatamente subseqüentes à explosão.

A eficiência da informação dinâmica contradiz as teorias que representam os segmentos como *targets* canônicos. O desenvolvimento recente de abordagens que não limitam a análise a "instantâneos fotográficos" da atividade articulatória aponta para uma fonologia menos centrada em propriedades abstratas inefáveis e mais preocupada com as potencialidades reais do trato vocal. É notável que diferentes linhas de pesquisa desenvolvidas na última década indiquem um retorno à caracterização articulatória (Browman e Goldstein 1986; Fowler 1986; Fujimura 1988), procurando aí as invariâncias que pareciam impossíveis de

serem detectadas a partir exclusivamente do sinal acústico. Em parte, verifica-se nesses enfoques uma reafirmação de alguns princípios da teoria motora, na medida em que o processo perceptual é interpretado basicamente como uma recuperação da organização articulatória (especialmente no âmbito da "action theory"; v. Browman e Goldstein 1986; Fowler 1987). A diferença em relação à teoria motora original é que o evento articulatório essencial não é mais um configuração estática, mas sim um conjunto coordenado de movimentos de vários articuladores.

Essas relações coordenadas entre os articuladores podem ser melhor entendidas à luz de um princípio geral de "eqüifinalidade" (Fowler 1987); segundo esse princípio, alguns sistemas (entre eles o sistema articulatório) tendem a estabelecer uma organização interna que permite alcançar um objetivo comum a partir de condições iniciais variáveis, e por meio de trajetórias também variáveis. A atuação de tal princípio poderia explicar, por exemplo, porque vogais produzidas com impedimento do movimento mandibular são acusticamente normais, ou quase normais, já a partir do primeiro pulso de F0 (Lindblom e Sundberg 1971; Lindblom *et al.* 1979; Fowler e Turvey 1980); nesses casos, a impossibilidade de utilizar informação através de *feedback* auditivo sugere que não é propriamente um processo de adaptação articulatória que está em jogo, mas antes um controle **COORDENATIVO** dos diferentes articuladores, de tal modo que, mesmo com o bloqueio da mandíbula, a relação língua-palato

característica da vogal intencionada possa ser aproximada já em um primeiro momento.

Evidências suportando o princípio de eqüifinalidade podem ser obtidas também através da observação da fala normal. Sussmann *et al.* 1973 relatam que a posição da mandíbula durante a oclusão em uma consoante bilabial varia com a altura da vogal precedente ou subseqüente; em plosivas bilabiais produzidas no contexto de vogais baixas a mandíbula contribui **MENOS** do que os lábios para a realização da oclusão, em comparação com a produção da consoante no contexto de vogais altas. Esses resultados demonstram como a articulação é capaz de coordenar as exigências, às vezes conflitantes, de diferentes segmentos fonológicos; a organização dos movimentos articulatórios é definida em função da tarefa específica a ser realizada - no caso a oclusão bilabial -, e o que existe em comum entre as diferentes realizações não é exatamente uma determinada configuração articulatória, ou uma **POSTURA**, mas antes uma intenção, ou um **GESTO**, que só pode ser satisfatoriamente descrito na medida em que se amplia o domínio temporal da análise para além dos limites do segmento fonológico.

O quadro teórico esboçado pela fonologia não-linear de base articulatória oferece, sem dúvida, a possibilidade de uma interpretação mais realista da atividade motora da fala. No entanto, a elegância na exposição dos conceitos mais fundamentais não encontrou ainda um paralelo com a descrição formal dos parâmetros relevantes; embora tenha havido algum progresso na

descrição dos mecanismos articulatórios (v. especialmente Browman e Goldstein 1986), as relações entre a produção e a realidade acústico-perceptual são estabelecidas de forma quase sempre oblíqua. Se a percepção é, realmente, a extração de informação gestual, é preciso explicitar quais elementos do sinal acústico permitem a recuperação da organização articulatória, caso contrário a descrição fica resumida à dimensão fisiológica, sem maiores implicações lingüísticas.

Parece que, dentro do campo da fonologia, vivenciamos um daqueles períodos onde novas descobertas experimentais obrigam uma imediata revisão dos cânones até então quase universalmente aceitos pela comunidade científica, ou um período que anuncia - na terminologia de Kuhn 1962 - uma "mudança de paradigma". Tipicamente, esses momentos delineam uma atitude científica que, sob o peso das evidências, deve rejeitar "aquilo que não pode ser", sem ter desenvolvido, todavia, o aparato formal adequado para descrever os novos fenômenos. Enquanto a pesquisa não avançar mais nessa direção, continuaremos usando informação do tipo *target* nos sistemas de reconhecimento automático, seja como fonte única de informação, seja como dado complementar à informação dinâmica eventualmente passível de ser extraída do sinal acústico.

7
O ESPECTRO COMO UM TODO

A grande maioria dos estudos sobre vogais aceita de forma quase inconteste que as freqüências dos formantes (estáticas ou em movimento) são a representação acústica mais adequada para uma vogal. No entanto, já que a estrutura de formantes não está diretamente presente no sinal acústico, é razoável questionar se os ouvintes, em algum estágio do processo perceptual, realmente efetuam uma transformação do sinal capaz de derivar essa estrutura. Essa discussão não é nova, e já esteve na base da controvérsia em torno das teorias "harmônica" e "enarmônica", defendidas, no século XIX, por Helmholtz e Hermann, respectivamente (v. Seção 1).

Alguns estudos propuseram que as distinções lingüísticas são feitas com base nas propriedades do espectro global, sem qualquer referência direta ao padrão de formantes subjacente. Segundo esse ponto de vista, os picos espectrais locais são as pistas mais efetivas para a percepção de vogais. Uma das formas de testar essa hipótese consiste em realizar uma análise do espectro total com base em filtros de banda relativamente estreita - freqüentemente de 1/3 de oitava [1]. O resultado desse tipo de análise é uma representação n-dimensional do sinal de fala, onde cada uma das dimensões é o *output* de um dos n filtros. A partir

dessa tranformação inicial, é possível re-sintetizar o sinal; essa re-síntese pode utilizar as n dimensões originalmente extraídas ou, alternativamente, realizar uma redução para um menor número de dimensões, por meio de procedimentos estatísticos adequados (análise de fatores, análise de componente principal, etc.) [2]. Testes perceptuais utilizando estímulos assim construídos verificam que a informação fonética essencial é preservada, mesmo com uma considerável redução do número de dimensões (Pols *et al.* 1969; Klein et al. 1970; Li et al. 1972,1973), embora possa haver alguma perda de informação paralingüística, principalmente em relação à identificação do falante e à expressão emocional (Pols 1975). Como nenhuma das transformações realizadas envolve a extração de formantes, os resultados desses experimentos têm sido interpretados como suporte para a hipótese de que as distinções lingüísticas essenciais podem ser feitas com base apenas nas características gerais do espectro.

São os Formantes Psicologicamente Reais?

Não fica totalmente claro, no entanto, se os dados espectrais citados nos estudos acima são realmente independentes do padrão de formantes subjacente. Pols *et al.* 1969, por exemplo, relatam que os dois primeiros fatores extraídos pela análise de componente principal correspondem aproximadamente a F1 e F2. Na

verdade, existe uma certa vinculação entre a configuração global do espectro e as freqüências absolutas dos formantes; pelo menos para sons sem a presença de zeros espectrais - como é o caso de vogais não-nasalisadas - é possível prever, com pequena margem de erro, o envelope espectral a partir das freqüências de ressonância (Fant 1960).

Existem vários argumentos contra a hipótese da percepção baseada no espectro como um todo. Um dos problemas diz respeito à influência de fatores ambientais. As características locais e globais do espectro estão sujeitas a alguma distorção em função, por exemplo, da forma, tamanho e nível de absorção acústica do ambiente onde o sinal de fala se propaga; até mesmo a posição da fonte sonora em relação ao ouvinte pode afetar algumas características do espectro acústico (Liebermann 1984:155). Apesar dessas perturbações, a fala permanece inteligível em todas as condições: se o contraste lingüístico tivesse de ser derivado do espectro "bruto", seria preciso que o mecanismo perceptual fosse capaz de adaptar-se rapidamente a todos esses efeitos ambientais (o que é improvável, embora seja uma hipótese que não deve ser totalmente descartada).

A existência de um estágio perceptual onde, de algum modo, são extraídos os formantes é evidenciada em um experimento conduzido por Remez *et al.* 1981; nesse estudo, os padrões de formantes de uma sentença real são representados por um conjunto de ondas senoidais, cujas freqüências e amplitudes correspondem

às dos formantes derivados do sinal original. Apesar dessa radical redução de informação espectral, alguns ouvintes são capazes de interpretar esses estímulos como fala, e identificar as sentenças originalmente enunciadas. Os dados de Remez *et al.* são consistentes com o fato de sermos capazes de perceber como fala alguns sinais emitidos por pássaros "falantes". Pássaros como o mainá, por exemplo, imitam a fala humana produzindo sinais acústicos nos quais os dois primeiros formantes são simulados através de ondas senoidais; só percebemos esses sinais como fala porque eles possuem energia acústica na região dos formantes (Lieberman 1984:156).

Identificação da Nasalidade

Embora pareça não haver dúvidas quanto ao fato de que os formantes desempenham um papel importante na determinação da qualidade vocálica, é preciso não perder de vista que essa não é a única informação acústica capaz de ser extraída pelo sistema auditivo. Existem outras características espectrais que são fundamentais para a identificação de alguns sons. Esse parece ser o caso das vogais nasalisadas. Já comentamos anteriomente (v. seção 3) que é praticamente impossível distinguir pares vocálicos oral/nasal com base apenas na estrutura de formantes; embora algumas vogais - especialmente /a/ - sofram uma considerável mudança na qualidade fonética quando nasalisadas, é improvável

que esquemas de reconhecimento automático sejam capazes de identificar o traço nasalidade a partir dessa informação. Há alguma evidência indicando que a alteração nas freqüências dos formantes (especialmente a variação em F1) também não é um fator decisivo na percepção de vogais nasalisadas por ouvintes humanos (Wright 1986).

A caracterização acústica da nasalidade envolve modificações espectrais de vários tipos. Além de uma eventual alteração do padrão de formantes, a nasalização de vogais pode estar associada

- ao aparecimento de ressonâncias específicas
- à presença de zeros espectrais
- à queda geral de intensidade (especialmente na região de F1)
- ao alargamento da largura de banda dos formantes

A detecção automática dessas pistas (e de outras possíveis não relacionadas) envolve uma série de dificuldades. O primeiro problema é saber QUAL pista está presente, já que qualquer um desses traços é, normalmente, suficiente para emprestar uma "cor nasal" à vogal (Cagliari 1977). Outra dificuldade, mais específica, diz respeito à localização dos zeros. Sendo a cavidade nasal praticamente indeformável, cada falante produzirá zeros espectrais mais ou menos fixos; esses zeros idiossincráticos só aparecerão claramente no espectro se não houver coincidência, ou muita proximidade, com uma freqüência de

ressonância (nesse caso o efeito se reduzirá a uma queda de intensidade do formante próximo ao zero). Além disso, o mínimo espectral normalmente associado com o zero pode não se desenvolver a ponto de ficar visível quando a área de acoplamento com a cavidade nasal é pequena (Fant 1960:149).

A caracterização acústica da nasalidade é uma questão ainda não satisfatoriamente resolvida. Exames mais detalhados da literatura verificam, em geral, bastante divergência quanto aos fatores acústicos supostamente relevantes (v.p.ex. Cagliari 1977). A incompatibilidade entre alguns resultados deve-se, em parte, ao fato de a nasalidade poder ser produzida por diferentes processos. A participação da cavidade nasal, por exemplo, ao contrário do que habitualmente se acredita, não é uma condição necessária para a realização de sons com colorido nasal (Laver 1980); sem o acoplamento com o ressoador nasal os zeros espectrais não ocorrem (ou são quase imperceptíveis) e as consequências acústicas podem estar condicionadas apenas a um efeito geral de *damping*, envolvendo o enfraquecimento e alargamento de banda de algumas ressonâncias. Sistemas de reconhecimento automático de fala baseados exclusivamente na detecção de proeminências e zeros espectrais são ineficazes para identificar esse tipo de nasalidade; por outro lado, é provável que uma análise usando informação do espectro global seja mais robusta, já que tanto a interação com zeros quanto os efeitos de *damping* refletem-se através de uma distribuição de energia mais

uniforme ao longo do espectro acústico (Wright 1986). Chen 1986 descreve um sistema recentemente desenvolvido que detecta a presença do componente de baixa freqüência característico do murmúrio nasal comparando as energias do *output* de filtros passa-banda de 100-350 Hz e 350-850 Hz.

Vantagens da Representação Global

Além dos problemas relacionados à identificação do traço nasalidade, outro aspecto que deve ser considerado é a dificuldade prática de reconstituir a trilha (*tracking*) de formantes em sistemas automáticos. Na análise de fala contínua, o aparecimento de pseudo-formantes, a intensidade muito fraca de alguns formantes, e a possibilidade de rotular erroneamente máximos espectrais (confundir F3 com F2, por exemplo, é um erro típico), entre outros problemas, complicam a extração da trilha de formantes (v. Li *et al*. 1973b; McCandless 1974; Miller 1984). Mesmo que alguns algoritmos possam ser desenvolvidos para contornar alguns desses problemas, sua aplicação exige normalmente um certo custo computacional, impossibilitando o processamento em tempo real. Na verdade, até o presente momento, não temos notícia de algoritmos "infalíveis" para a extração da trilha de formantes; análises mais detalhadas exigem, freqüentemente, intervenção manual para corrigir algumas distorções (v. Zue e Lamel 1986).

Por outro lado, uma representação baseada na configuração global do espectro pode ser obtida com bastante rapidez. É claro que a informação assim obtida será restrita a certas propriedades gerais do envelope espectral. No entanto, o conhecimento da distribuição da energia acústica ao longo do espectro é, quase sempre, suficiente para estabelecer algumas distinções básicas envolvendo traços como grave vs. agudo, compacto vs. difuso, vocálico vs. consonantal, fricativa forte vs. fricativa fraca, etc. Essas oposições não garantem, obviamente a identificação direta de todos os segmentos possíveis, mas - como examinaremos mais cuidadosamente na seção 8 - uma classificação fonética "larga", quando complementada por regras *top-down*, é, em muitos casos, suficiente para que o sistema tome a decisão correta, ou, pelo menos, para que se reduza consideravelmente a probabilidade de erro.

Representações Acústicas mais Realistas

O recente avanço na compreensão das transformações efetuadas pelo sistema auditivo periférico tem permitido o desenvolvimento de técnicas de análise espectral que modelam mais realisticamente o processamento humano de fala. Uma série de importantes estudos usando animais procuraram descrever, em termo bastante precisos, como alguns sinais de fala são codificados no sistema auditivo periférico [3]. Esses estudos examinaram a atividade nervosa no

sistema auditivo durante a resposta a estímulos simples de fala, tais como vogais estáveis e consoantes plosivas em sílabas CV (v. p.ex. Moore e Cashin 1974,1976; Kiang 1980; Delgutte 1980). A observação do padrão de disparos nos aferentes nervosos permite identificar algumas propriedades que correspondem, de um modo direto, a propriedades acústicas importantes, ou a certas características das vogais (Sachs e Young 1979) e das consoantes (Miller e Sachs 1983). Essa promissora linha de pesquisa tem sugerido novas formas de representar o sinal de fala ("neurogramas", "cócleagramas", "espectrogramas neurais", etc.) que, presumivelmente, forneceriam informação mais relevante quanto às dimensões perceptuais subjacentes.

A incorporação de dados obtidos em testes psicofísicos também pode contribuir para o desenvolvimento de representações perceptualmente mais realistas. Lindblom 1986, baseando-se em alguns aspectos universais bem conhecidos da audição humana, propõe um espectro transformado obtido através de filtragens que consideram as propriedades auditivas abaixo relacionadas e comentadas:

- **RESOLUÇÃO DE FREQÜÊNCIA:** os filtros auditivos aumentam em largura em função da freqüência (Moore e Glasberg 1983; Greenberg 1988); a análise, portanto, deve evitar filtragens com banda fixa. Uma possível solução é projetar filtros simulando as "bandas críticas" definidas pela função BARK (v. Zwicker 1961).

- CARACTERÍSTICAS ASSIMÉTRICAS DE MASCARAMENTO: freqüências baixas mascaram freqüências altas de um modo mais eficaz do que o oposto (Greenberg 1988); sendo assim, um filtro ideal deve ter um decaimento mais suave na direção das freqüências mais altas, ou seja, o centro do filtro não coincide com o centro da largura de banda.

- RESPOSTA NÃO-LINEAR DE FREQÜÊNCIA: a relação entre AUDIBILIDADE (*loudness*) e freqüência não é linear; o nível de pressão sonora de dois sons com diferentes freqüências, mas com o mesmo nível aparente de AUDIBILIDADE pode diferir consideravelmente, dependendo da separação de freqüência entre os dois sons. Mais especificamente: uma determinada intensidade (em dB) em uma freqüência mais alta tem um efeito perceptual maior do que a mesma intensidade em uma freqüência mais baixa. Quando correções para essas não-linearidades são aplicadas ao *output* do filtro, o principal efeito é uma redução da contribuição dos componentes de baixa freqüência.

- AUDIBILIDADE (*loudness*): a partir de aproximadamente 40 *fons* [4], o efeito perceptual de AUDIBILIDADE duplicado a cada aumento de 10 *fons*, enquanto para valores entre 0-40 *fons* o mesmo aumento provoca efeitos mais reduzidos. Dessa forma, uma diferença de x *fons* entre vales espectrais equivale a uma diferença de sonoridade bem menor do que a mesma diferença de x *fons* entre dois picos. Incorporando esse tipo de informação ao modelo é possível representar formalmente a idéia de que picos

espectrais carregam mais peso perceptual do que vales.

Dentre as correções acima sugeridas, a mais relevante parece ser o ajuste da largura de banda dos filtros de análise. Tanto a análise realizada pelo espectrógrafo analógico tradicional, quanto técnicas digitais como a Predição Linear (LPC) (Atal e Hanauer 1971; Makhoul 1975; Markel e Gray 1976) presumem larguras de banda fixas. Esse procedimento pode provocar, todavia, algumas distorções; por exemplo: formantes que fazem parte de um complexo de ressonâncias de alta freqüência nem sempre são individualmente resolvidos (devido à maior largura de banda dos filtros auditivos nessa região), como pode fazer supor uma representação espectrográfica em bandas fixas [5].

Embora não existam dúvidas quanto ao fato de que as larguras dos filtros auditivos aumentam em função da freqüência, há alguma discordância quanto às larguras efetivas. Várias estimativas já foram feitas com base em dados psicofísicos; essas estimativas podem variar consideravelmente, indo de 1/2 a 1/10 de oitava (v. Moore e Glasberg 1983, para uma resenha desses estudos) [6].As faixas sugeridas mais freqüentemente estão no interior da área hachurada da figura 7.1 (extraída de Kewley-Port e Luce 1984); o limite superior é definido pela escala BARK, baseada no conceito de bandas críticas (Zwicker 1961; Zwicker *et al.* 1979), enquanto o limite inferior, de 1/6 de oitava, é similar à proposta de Patterson 1976. A largura constante presumida pela análise LPC é

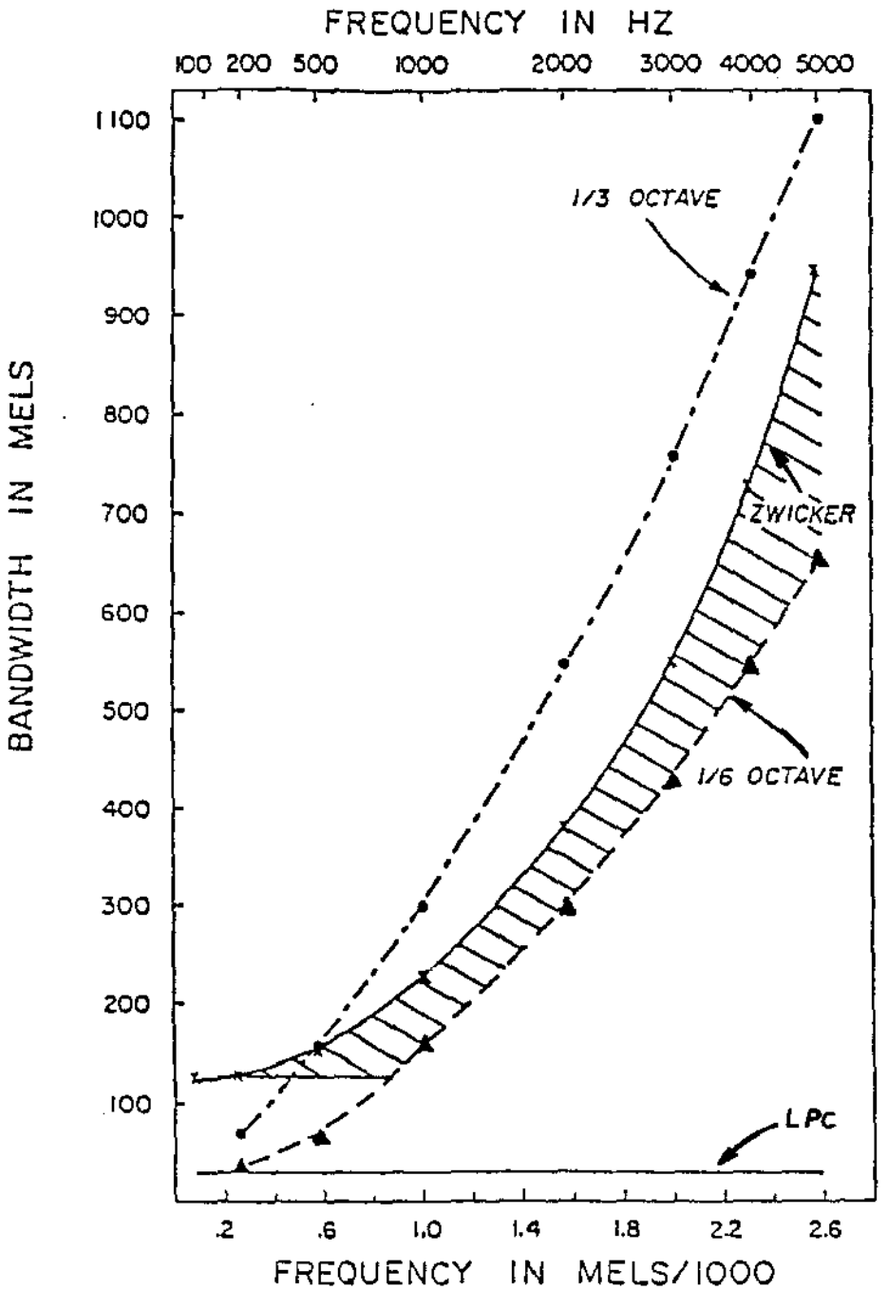

FREQUENCY IN HZ
100 200 500 1000 2000 3000 4000 5000
BANDWIDTH IN MELS
1100
1000
900
800
700
600
500
400
300
200
100
1/3 OCTAVE
ZWICKER
1/6 OCTAVE
LPc
.2 .6 1.0 1.4 1.8 2.2 2.6
FREQUENCY IN MELS/1000

FIGURA 7.1

representada pela linha horizontal na parte inferior da figura; como se pode observar, a análise LPC, apesar de seu amplo uso no processamento de fala, não caracteriza apropriadamente a filtragem no sistema auditivo. O mesmo poderia ser dito do espectrógrafo analógico convencional, embora a largura de banda utilizada por esse tipo de aparelho no modo "banda larga" (cerca de 300 Hz) pareça ser mais adequada do que a presumida na análise LPC.

A figura 7.2 (também em Kewley-Port e Luce 1984) dá uma idéia do efeito da incorporação de filtros variáveis na análise. A parte superior da figura mostra um espectro corrente tridimensional da sílaba /bu/ produzido por LPC com larguras de banda constantes; o segundo espectro, para a mesma sílaba, mostra os resultados com uma filtragem usando larguras de banda que aumentam com a freqüência. O exame das duas representações revela algumas diferenças importantes; com a filtragem variável fica evidente, por exemplo, a diminuição do peso relativo dos componentes de baixa freqüência e a atenuação dos picos no espectro transformado (os picos ficam menos "agudos").

O desenvolvimento de representações espectrais mais compatíveis com a realidade auditiva periférica permitirá, certamente, a implementação de novos algoritmos que poderão melhorar o desempenho dos sistemas de reconhecimento automático. É importante não esquecer, contudo, que essa é apenas uma faceta de um problema bem mais complexo; avanços mais significativos

dependem também de uma maior compreensão dos mecanismos auditivos centrais que integram o *input* sensorial, assim como dos processos de decisão cognitivo-perceptuais que interpretam lingüisticamente os padrões inicialmente extraídos do estímulo. Examinaremos alguns desses aspectos na próxima seção.

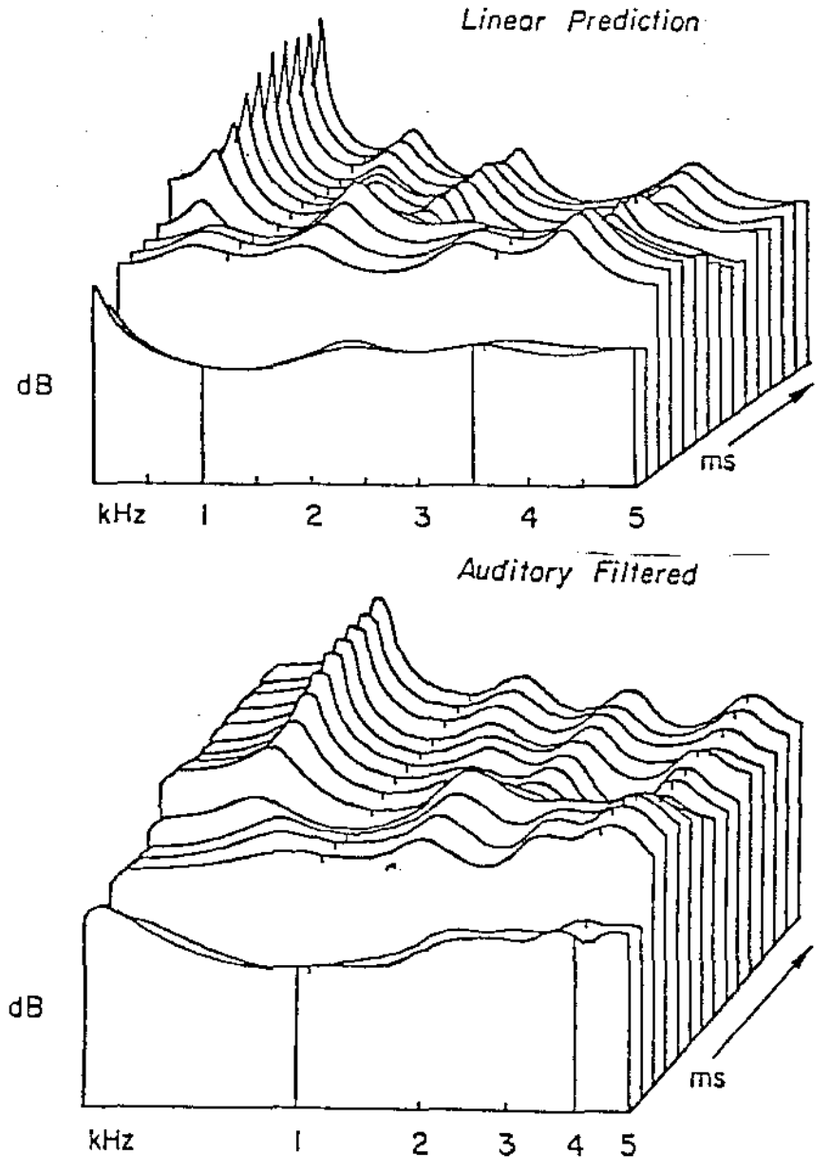


FIGURA 7.2

8

INFORMAÇÃO "TOP-DOWN"

A maior parte da pesquisa sobre percepção da fala desenvolvida ao longo das últimas décadas dirigiu seus esforços para a compreensão da percepção de fonemas e sílabas em isolamento ou em contexto constante. Essa estratégia, obviamente redutora, justifica-se, em parte, se considerarmos a enorme complexidade envolvida na percepção e interpretação da fala fluente; a dificuldade de modelar a interação entre os diversos níveis de processamento obriga o experimentador a limitar drasticamente a influência contextual, sob o risco de perder o controle das variáveis relevantes. Mas esse procedimento, infelizmente, acarreta um risco ainda maior: ao desbastar as arestas do seu objeto, de modo a tornar mais interpretável o resultado do experimento, o pesquisador talvez esteja suprimindo exatamente aqueles aspectos que são mais importantes na situação de fala real, ou seja - como na advertência do dito popular -, é preciso ter cuidado para, junto com a água, não jogar fora também a criança. É claro que experimentos onde há controle das características do estímulo sempre revelam alguma faceta do processo perceptual; o fato, no entanto, de o mecanismo perceptual ser **CAPAZ** de realizar algo não autoriza generalizações, no sentido de assumir - como ocorre freqüentemente - que comportamentos semelhantes ocorrerão em

outras situações, especialmente quando a presença de informação mais saliente e/ou diretamente acessível garante a eficiência do processamento.

Examinemos o caso específico da identificação de vogais. Como já discutimos ao longo do texto, a produção de vogais na fala fluente está sujeita a diversas fontes de variabilidade: o *output* acústico pode variar, entre outros fatores, em função do falante (sexo, idade, grupo dialetal, etc.), dos efeitos da coarticulação (o problema do "target undershoot"), da interação com o nível prosódico (padrão rítmico/acentual, contorno entoacional), do modo de fonação (fala gritada, sussurrada, etc.) e da velocidade e estilo de fala (hipoarticulada, hiperarticulada). Além da variabilidade inerente à produção, certas condições de transmissão também podem, eventualmente, modificar sensivelmente o sinal original; o posicionamento e distância da fonte sonora em relação ao ouvinte, o nível de ruído ambiental, a absorção acústica do ambiente, o corte espectral na recodificação telefônica, entre outros fatores, fazem com que na vida real - longe das condições ideais do laboratório - o sinal acústico já sofra algumas transformações (pelo menos em relação a certas propriedades globais do envelope espectral) antes mesmo de chegar ao sistema auditivo periférico.

Apesar das inúmeras fontes de variação, o mecanismo perceptual humano demonstra uma notável habilidade em reconstituir a mensagem lingüística intencionada: aquilo que

parece quase impossível de ser formalmente descrito através de um modelo teórico constitui, todavia, a mais corriqueira das atividades humanas: interpretar sinais de fala. O que garante essa eficiência da percepção humana? Dificilmente encontraremos a resposta examinando estímulos descontextualizados; experimentos com sons lingüísticos isolados tendem a enfatizar pistas SUFICIENTES, mas negligenciam a INTERAÇÃO entre essas pistas. Por outro lado, na percepção da fala fluente, o ouvinte tem acesso a um conjunto de informações interrelacionadas que facilitam a identificação; o ouvinte não toma decisões com base apenas nas propriedades fonético-acústicas do estímulo, mas também recorre a todo e qualquer conhecimento estrutural que possui da língua. Informações de ordem fonológica, lexical, sintática, semântica e pragmática são integradas na análise perceptual, assegurando o reconhecimento eficaz das unidades lingüísticas, mesmo sob condições adversas de produção e transmissão. Um sistema de reconhecimento automático que não considere essas interações fatalmente se defrontará, em algum momento, com impasses irresolvíveis. O fato é que, em alguns casos, por mais sofisticado que seja o classificador fonético do sistema automático, será impossível tomar uma decisão segura a partir exclusivamente da análise acústica do sinal. Esse tipo de dificuldade tende a ocorrer com incômoda freqüência no reconhecimento de fala rápida e/ou hipoarticulada (especialmente em sílabas átonas); nesses casos pode haver um grau muito elevado

de redução, neutralizando as características distintivas do segmento vocálico. Embora o exame dos aspectos transicionais às margens da vogal possa resolver algumas ambigüidades, não é certo que esse procedimento garanta uma identificação correta em todas as situações.

Antes, porém de examinarmos a implementação de processos decisórios *top-down* em sistemas automáticos, será conveniente discutir mais detalhadamente alguns modelos que tentam descrever os diferentes níveis de análise da percepção humana.

Serial vs. Paralelo

Um modo mais ou menos simples de modelar o processamento humano de fala é concebê-lo como um sistema de diferentes níveis: o sinal de fala passaria por várias transformações sucessivas, desde o *input* sensorial até as camadas superiores, interpretativas. Teríamos, portanto, uma seqüência de processos, cada um deles definindo um nível de representação: auditivo, fonético, fonológico, lexical, semântico, sintático, pragmático (etc...?). Embora possa haver algum desacordo quanto à regulação interna de cada um desses níveis, não há como negar que aí estão configuradas dimensões psicologicamente reais. Há, no entanto, formas bastante diferentes de encarar as RELAÇÕES entre esses níveis. Segundo o modelo SERIAL - que praticamente dominou a Psicolingüística nos anos 70 (v. especialmente Fodor *et al.* 1974)

-, o fluxo de informação entre os diversos componentes do sistema corre apenas em uma direção ("de baixo para cima"), sem possibilidade de retroalimentação entre níveis distintos; a percepção é vista assim como um processo essencialmente HIERARQUIZADO, onde cada componente (ou módulo) é autônomo em suas operações - o processo de análise, disparado pela informação *bottom-up*, é afetado somente por fontes de informação no interior de cada módulo.

Uma forma alternativa de encarar a questão é através de um modelo de processamento PARALELO interativo. Esse ponto de vista assume que o sistema é mais flexivelmente estruturado, de tal modo que a análise desenvolvida em um nível possa, em princípio, afetar as operações em outro nível; dessa forma, a informação, ao contrário do modelo anterior, não se propaga unidirecionalmente, permitindo que decisões de níveis superiores ajustem ou corrijam decisões de níveis inferiores (v. Marslen-Wilson 1980; Elman e McClelland 1980; Zwitserlood 1989).

Há bastante evidência experimental suportando a hipótese do modelo pararelo interativo. Hemdal e Hugues 1967, em um teste de identificação de vogais em sílabas CVC, observam que os índices de acerto são significativamente maiores quando a combinação CVC produz uma palavra existente na língua nativa do ouvinte, o que sugere que os sujeitos fazem uso do significado lexical para tomar decisões fonéticas - mesmo em palavras apresentadas em isolamento [1].

A identidade lexical do estímulo pode afetar a decisão a respeito de um traço específico. Ganong 1980 constrói um contínuo sistetizado entre GIFT-KIFT, separando estímulos intercalados linearmente entre os extremos sonoro/não-sonoro. Ele verifica que a fronteira categorial tende para o extremo GIFT, ou seja, os ouvintes identificam um maior número de pontos do contínuo como GIFT, que é uma palavra real. Exatamente o oposto ocorre quando o contínuo é GISS-KISS, onde o primeiro termo, não sendo uma palavra da língua, é escolhido com menos freqüência.

Aquilo que "ouvimos" depende, em parte, de uma certa expectativa em relação ao que acreditamos que o falante provavelmente dirá, isto é, tendemos a "ouvir" sons consistentes com uma conversação que faça sentido. É oportuna a observação de Lieberman 1984:218: "anyone who has ever attempted to make a transcription of the proceedings of a conference knows that it is often impossible to understand what a speaker is saying unless one is familiar with the topic under discussion". É por essa razão que, na tradução simultânea, o intérprete deve travar conhecimento prévio com o tema da palestra ou conversação a ser traduzida, caso contrário sua performance será prejudicada.

A influência do contexto semântico na percepção do detalhe fonético já foi verificada experimentalmente há algum tempo. Miller *et al*. 1951 relatam que palavras apresentadas em listas arbitrárias são menos inteligíveis do que as mesmas palavras contextualizadas em uma sentença normal.

Dependendo da informação contextual anterior, a representação em um nível pode antecipar a codificação em níveis mais básicos; decisões a nível lexical, por exemplo, são às vezes tomadas antes mesmo que a informação acústico-fonética esteja completa. Em um estudo onde se exigia dos sujeitos que reproduzissem o mais rápido possível trechos de prosa fluente, Marslen-Wilson 1975 verifica que os tempos de latência podem ser extremamente curtos, em muitos casos cerca de 250 ms - o que equivale ao atraso de apenas uma sílaba. Em várias ocasiões, portanto, os ouvintes são capazes de reproduzir palavras antes do término de sua produção.

Marslen-Wilson e Tyler 1980, em um estudo detalhado sobre a estrutura temporal da percepção temporal da fala, obtêm resultados mais significativos controlando o tipo de sentença onde a palavra *target* aparece. Três contextos de prosa foram usados:

(a) sentença sintatica- e semanticamente normal;

(b) sentença com estrutura sintática gramatical, mas semanticamente anômala;

(c) pseudo-sentença composta por uma ordem randômica de itens lexicais.

Foram medidos os tempos de resposta (identificação da palavra *target*) nessas diferentes condições. Os autores verificam que a reação em (c) é mais lenta que em (b) e em (b) mais lenta que em (a), sendo a diferença entre (a)-(b) maior que entre (b)-(c).

Além disso observou-se que, nas condições (a) e (b), os tempos de latência diminuem à medida que a palavra a ser identificada aparece em posições mais próximas ao fim da frase. Esses resultados demonstram que tanto a informação semântica precedente quanto o conhecimento de estrutura sintática da sentença contribuem para acelerar o processo de decisão lexical, sendo que as restrições semânticas parecem ser potencialmente mais efetivas. Fica claro também que quanto MAIS informação (de qualquer tipo) mais rápida é a resposta.

Unidades Perceptuais

Alguns testes perceptuais baseados no paradigma da velocidade de resposta já observaram que os tempos de reação para a identificação de palavras são geralmente menores do que para o reconhecimento de fonemas (Marslen-Wilson e Tyler 1980) ou das sílabas constituintes das mesmas palavras (Foss e Swinney 1973). Outros estudos relatam que estruturas silábicas são percebidas antes que os fonemas componentes (Savin e Bever 1970; Wood e Day 1975). Esses estranhos resultados parecem nos colocar no "mundo dos espelhos" da Alice de Lewis Carrol, onde curiosamente, o julgamento antecede o crime (Cf. Studdert-Kennedy 1976). Esse problema nos remete a uma antiga discussão na área de percepção da fala, que é a determinação da unidade mínima da análise perceptual. Afinal, qual é o tamanho da "molécula" perceptual?

Até bem recentemente havia pouco desacordo quanto a aceitar que o sinal de fala, em algum estágio do processamento, era internamente representado como uma seqüência de segmentos discretos (fonemas) e traços distintivos (embora sempre tenha havido alguma discussão sobre como deveria ser a descrição exata desses traços: articulatória, acústica, ou ambas). Alguns problemas surgem, entretanto, quando observamos que existe uma discrepância entre as entidades abstratas da análise lingüística tradicional e os parâmetros acústicos e articulatórios, ou seja, o fonema só pode ser adequadamente descrito (acústica- e/ou articulatoriamente) se incorporarmos informação espraiada para além das fronteiras do segmento fonológico (v.seção 6 para uma discussão mais detalhada). Seriam os fonemas e os traços distintivos, portanto, unidades fictícias, fruto da rica imaginação de alguns lingüistas? Existem bons argumentos para admitir que a sílaba, por exemplo, talvez fosse uma unidade mais realista - afinal, falamos "em sílabas", e não fonema-a-fonema, e são as sílabas que veiculam o padrão rítmico-acentual. Por outro lado, já se demonstrou que tanto o fonema quanto o traço distintivo são psicologicamente reais; Fromkin 1971, analisando um grande número de erros de metátese, observa que os falantes podem inverter não só palavras e sílabas, como também fonemas (far-more--> mar-fore) e traços (dear blue--> glear plue). Embora os dados aqui refiram-se a características de produção não é fora de propósito admitir que, se o falante tem controle independente

sobre a unidade envolvida no erro, então essas unidades também podem ser independentemente percebidas.

Muita da discussão sobre a unidade perceptual - básica ou "natural" poderia ser evitada se uma distinção mais exata do nível de análise fosse definida. A unidade de processamento não é a mesma em todos os níveis da análise e pode variar conforme a atenção do ouvinte é dirigida para diferentes aspectos da mensagem lingüística. Uma unidade é mais ou menos "primária" em função do nível de processamento exigido pela tarefa específica apresentada ao ouvinte; Fujisaki *et al.* 1986 verificam que menos de 40% das sílabas apagadas em um contexto sentencial são notadas pelos ouvintes, enquanto esse número cresce para 70% no contexto de palavras em isolamento. Isso sugere que a sílaba é provavelmente a unidade perceptual a nível da palavra, mas a nível da sentença essa unidade talvez seja a palavra.

Admitindo-se que a palavra é o percepto elementar ao nível da sentença, não parecerá tão estranho que os tempos de reação em tarefas de reconhecimento de fonemas sejam, em geral, maiores do que em tarefas envolvendo a identificação de palavras. Na detecção de fonemas isolados não há como usar informação contextual e a decisão depende exclusivamente da análise fonético-acústica (Marslen-Wilson e Tyler 1980).

Para um defensor do modelo estritamente serial alguns resultados acima descritos poderiam parecer um tanto paradoxais, já que uma representação lexical só seria acessada, no modelo

serial, após ter-se completado o processamento fonético acústico. Do ponto de vista de um modelo paralelo interativo, no entanto, diferentes níveis de análise são processados simultaneamente, ou seja, o sistema está "on line", integrando constantemente informações de várias fontes (fonética, lexical, semântica, sintática, etc) [2].

Embora haja algumas diferenças específicas entre os diversos modelos paralelos já propostos (v.Zwitserlood 1989), é geralmente aceito que o processo de decisão lexical envolve duas fases: o **ACESSO LEXICAL** e a **SELEÇÃO/INTEGRAÇÃO**. O acesso envolve a intermediação de um conjunto inicial contendo todas as formas compatíveis com algum trecho do *input* sensorial inicial. Esse conjunto de "palavras candidatas" - também chamado de **COHORT** - deve ser reduzido até que reste apenas uma palavra. Essa redução ocorre na fase de seleção/integração, onde é testada a adequação das propriedades semânticas, sintáticas e pragmáticas das palavras candidatas; quando essas propriedades são compatíveis com uma representação a nível sentencial, o ouvinte está apto a tomar uma decisão lexical. É importante frisar que o termo "fase" não implica uma seqüência temporal; o processo de seleção/integração pode correr paralelamente com a análise sensorial e a ativação do léxico (v. Zwitserlood 1989).

Suponhamos, por exemplo, que o ouvinte tenha reconhecido, por meio da análise fonético-acústica, o fragmento "capi". Na fase de acesso, todas as representações lexicais compatíveis com

o *input* sensorial serão ativadas: capital, capitão, capivara, capitólio, etc. Com base na informação contextual precedente, o processo de seleção/integração decide, se possível, qual o item mais adequado; se as restrições contextuais forem suficientes a decisão pode, em alguns casos, ser tomada a partir apenas do fragmento (o que pode superar dificuldades relacionadas com pronúncia deficiente, ruído ambiental, etc.). Continuemos com o nosso exemplo, supondo que o fragmento reconhecido apareça nos seguintes contextos:

I	II
(a) Eles	(a) a perda
(b) Os investidores lamentavam	de seu CAPI...
(c) Os soldados	(b) a morte

e, para simplificar, que o léxico só tenha duas entradas compatíveis: CAPITÃO e CAPITAL. Os contextos I_a + II_b, I_c + II_a e I_c + II_b conduziriam, quase inequivocamente à decisão "capitão". I_b + II_a levaria provavelmente à seleção de "capital". I_a + II_a admitiria as duas soluções e seria preciso mais informação *bottom-up* para decidir. I_b + II_b cria uma situação interessante; aqui tanto "capital" quanto "capitão" seriam compatíveis, mas apenas metaforicamente (o que dá uma idéia do tipo de dificuldade que pode surgir no controle da informação contextual em sistemas de reconhecimento automático de fala).

Reconhecimento de Palavras em Sistemas Automáticos

Em sistemas automáticos nem sempre é possível identificar fonemas com total segurança a partir apenas do sinal acústico. Em alguns casos a análise acústica necessária para uma distinção (por exemplo entre /a/ e /ɔ/ ou /m/ e /n/) tem de ser muito detalhada e a probabilidade de erro é grande. Além disso, uma classificação fonética fina, mesmo quando possível, exige normalmente um certo custo computacional. Para todos os efeitos, portanto, é conveniente reduzir a necessidade de análises acústicas pormenorizadas.

Para um sistema de reconhecimento de fala, no entanto, é menos importante identificar fonemas do que unidades linguísticas maiores, como morfemas, palavras ou mesmo sentenças. Como vimos anteriormente em relação à percepção humana, o espaço formado pelas hipóteses interpretativas do sinal gerado pela análise fonético-acústica pode ser substancialmente reduzido pela interação com outras fontes de informação: restrições fonológicas, sintáticas, semânticas e pragmáticas, assim como o conhecimento que o falante tem da estrutura do Léxico podem resolver ambiguidades e acelerar o processamento - em alguns casos permitindo que a decisão lexical antecipe-se à análise fonética.

O grande problema é que a formalização e automatização do conhecimento estrutural que o falante tem da língua envolve,

obviamente, dificuldades gigantescas. Embora a nível fonológico e sintático muitas dessas dificuldades possam ser superadas (pelo menos há algum otimismo quanto a isso), o controle efetivo do contexto semântico/pragmático não passa, por enquanto, de uma possibilidade mais ou menos remota. É preciso não perder de vista, contudo, que para um bom número de aplicações a variabilidade da informação contextual de alto nível pode ser bastante limitada; um sistema pode ser projetado, por exemplo, para o ambiente "carta comercial", onde há um grande probabilidade de ocorrerem construções sintáticas regulares e efeitos de significado não-metafóricos. Na verdade, no atual estágio da pesquisa, muitas outras restrições costumam ser impostas: vocabulário limitado, número limitado de falantes, exigência de treinamento prévio com o(s) falante(s), exigência de fala lenta e bem articulada, ausência de ruído ambiental, etc.

Apesar das dificuldades e limitações contextuais, algumas pesquisas da última década têm conseguido resultados encorajadores, especialmente no reconhecimento de palavras isoladas. Tal avanço não teria sido possível sem o trabalho seminal de Shipman e Zue 1982. Nesse estudo - quase um "ovo-de-Colombo" - Shipman e Zue demonstram que uma análise acústica "bruta" (mas confiável) em termos de classes fonéticas bem largas é capaz de promover uma redução substancial do número de palavras candidatas. Uma representação com apenas seis fonemas é sugerida:

[VOGAL], [STOP], [LíQUIDA/GLIDE],
[NASAL], [FRICATIVA/GLIDE], [FRICATIVA FRACA]

Usando essa classificação larga, o léxico é estruturado em conjuntos de palavras com a mesma representação - as palavras "pato" e "gota", por exemplo, pertenceriam ao mesmo conjunto (ou COHORT) definido pela representação larga [STOP] [VOGAL] [STOP] [VOGAL]. No inglês, Shipman e Zue verificam que, para um léxico de 20000 palavras, há uma média de apenas 35 palavras partilhando a mesma sequência fonética larga; a classe mais extensa tem cerca de 200 palavras, ou seja, apenas 1% do Léxico. Esses números, teoricamente, podem variar em função da estrutura particular do Léxico: estudos posteriores com outras línguas mostram, entretanto, resultados não muito distantes dos obtidos para o inglês (v. Billi *et al.* para o italiano e Vernooij et al. para o holandês).

O objetivo de um sistema de reconhecimento de fala é, no entanto, identificar palavras, e não localizar o COHORT ao qual a palavra *target* pertence. Dessa forma, após a classificação fonética larga, será necessário definir sub-classes fonéticas que permitam encontrar a palavra correta entre as palavras candidatas integrantes do *cohort*. Mas será que voltamos ao ponto de partida, isto é, para identificar a palavra é preciso, afinal, realizar uma análise acústica detalhada? Não é bem assim. Consideremos os seguintes pontos:

(1) a classificação larga (que é rápida e segura) já diminuiu decisivamente o espaço de procura - só isso a justificaria, pelo menos como primeiro passo.

(2) a questão da eficiência da classificação larga depende em grande parte do tamanho do Léxico; em pequenos vocabulários as restrições sequenciais são, muitas vezes, suficientes para localizar a palavra correta.

(3) cada *cohort* apresenta certas propriedades estruturais que podem ser exploradas para reduzir a necessidade de análises acústicas finas.

Examinemos mais detalhadamente o ponto (3). Como o tamanho de qualquer *cohort*, mesmo em Léxicos extensos, é sempre relativamente pequeno, não é muito complicado estabelecer, com base nas propriedades intrínsecas do *cohort*, um conjunto de estratégias que permita a divisão do *cohort* em palavras individuais. Um exemplo (adaptado de Vernooij *et al.* 1989) ajudará a esclarecer esse ponto. Suponhamos um *cohort* imaginário ([STOP] [VOGAL] [FRICATIVA] [VOGAL]) composto de apenas três itens:

1) /kaza/ = "casa"

2) /kasa/ = "caça"

3) /pasu/ = "passo"

A partir das propriedades estruturais desse pequeno *cohort* é possível derivar as estratégias possíveis que permitiriam discriminar todos os itens:

A) $(z,s)_{1,2} + (k,p)_{1,3} + (k,p)_{2,3}$

B) $(z,s)_{1,2} + (k,p)_{1,3} + (a,u)_{2,3}$

C) $(z,s)_{1,2} + (z,s)_{1,3} + (k,p)_{2,3}$

D) $(z,s)_{1,2} + (z,s)_{1,3} + (a,u)_{2,3}$

E) $(z,s)_{1,2} + (a,u)_{1,3} + (k,p)_{2,3}$

F) $(z,s)_{1,2} + (a,u)_{1,3} + (a,u)_{2,3}$

onde $(a,b)_{i,j}$ significa: para discriminar as palavras *i* e *j*, a distinção entre os fonemas *a* e *b* deve ser feita.

Qualquer uma dessas estratégias permite separar inequivocamente as palavras do *cohort*, mas algumas parecem mais adequadas. Vernooij *et al*. 1989 sugerem dois critérios básicos para selecionar a estratégia ótima:

(1) escolher a estratégia cujo conjunto de distinções pode ser mais facilmente realizado pela análise acústica (o conhecimento prévio do desempenho do sistema pode, por exemplo, determinar um peso de "risco" para cada distinção específica)

(2) escolher a estratégia que envolve um menor número de distinções fonéticas.

Esses critérios podem ser combinados, evitando, por exemplo, a escolha de uma estratégia que, apesar de incluir poucas distinções, encontre dificuldades para implementar essas distinções. O importante é que se levarmos em conta esses

critérios será possível, em um grande numero de casos - especialmente se o *cohort* não for muito extenso e/ou complexamente estruturado -, distinguir palavras no interior do *cohort* com base em sub-classes cujo número é menor do que os fonemas da língua; no nosso exemplo, se escolhermos a estratégia D ou F, serão suficientes as distinções entre os traços sonoro/não sonoro (diferencia /z/ de /s/) e compacto/difuso (diferencia /a/ de /u/). O exemplo dado baseia-se, obviamente, em um *cohort* extremamente reduzido, mas os princípios gerais valem para qualquer tamanho de Léxico.

Vejamos um exemplo mais realista. Suponhamos que o analisador fonético tenha fornecido como *output* inical a sequência larga: [STOP] [VOGAL] [LíQUIDA] [VOGAL].
Consultando o *Novo Dicionário da Língua Portuguesa* (Buarque de Holanda Ferreira, 1ª edição), que contém cerca de 10^5 palavras, encontramos 35 palavras (se nenhuma nos escapou) que se encaixam nessa classificação larga. Como palavras de 4 fonemas são muito frequentes, podemos estimar que a ordem de grandeza desse subconjunto do Léxico deve estar entre 10^3 - 10^4 palavras. Dessa forma, a redução conseguida apenas com a classificação larga é, no mínimo, de aproximadamente 97% em relação ao subconjunto de palavras com 4 fonemas.

Para diminuir ainda mais o espaço de procura, seria conveniente dividir a classe [VOGAL] em subclasses que incluam, cada uma, grupos de fonemas frequentemente confundidos na análise

fonética fina. Para isso é necessário conhecer previamente o desempenho do sistema. Suponhamos que nosso algoritmo de classificação seja o programa DISCRIM (v. seção 3). A tabela 8.1 mostra, em termos absolutos e percentuais, as classificações certas e erradas produzidas por DISCRIM com base nos dados de 90 falantes de PB (Behlau 1984), usando F1, F2 e F3 (em escala log) como parâmetros de controle.

NUMBER OF OBSERVATIONS AND PERCENTS CLASSIFIED INTO VOG:

FROM VOG	ɛ	ɔ	a	e	i	o	u	TOTAL
ɛ	79	0	0	11	0	0	0	90
	87.78	0.00	0.00	12.22	0.00	0.00	0.00	100.00
ɔ	0	74	8	0	0	8	0	90
	0.00	82.22	8.89	0.00	0.00	8.89	0.00	100.00
a	0	13	74	0	0	3	0	90
	0.00	14.44	82.22	0.00	0.00	3.33	0.00	100.00
e	11	0	0	78	1	0	0	90
	12.22	0.00	0.00	86.67	1.11	0.00	0.00	100.00
i	0	0	0	5	85	0	0	90
	0.00	0.00	0.00	5.56	94.44	0.00	0.00	100.00
o	0	9	2	0	0	72	7	90
	0.00	10.00	2.22	0.00	0.00	80.00	7.78	100.00
u	0	0	0	0	0	9	48	57
	0.00	0.00	0.00	0.00	0.00	15.79	84.21	100.00
TOTAL	90	96	84	94	86	92	55	597
PERCENT	15.08	16.08	14.07	15.75	14.41	15.41	9.21	100.00
PRIORS	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.0955	

TABELA 8.1

Observa-se que a grande maioria dos erros envolve confusões entre qualidades próximas (/ɛ/ por /e/, /a/ por /ɔ/, etc). Se

reunirmos as classes que mais se confundem obteremos uma classificação menos larga do que apenas [VOGAL] sem, contudo, aumentar muito a possibilidade de erro. As três subclasses mais convenientes são [i+e+ɛ], [a+ɔ] e [o+u]. A classe [i+e+ɛ] não se confunde com nenhuma das outras; há ainda alguma confusão entre as duas classes restantes, que produzem 22 erros:

/ɔ/ por /o/ ou vice-versa --> 17 erros
/o/ por /a/ ou vice-versa --> 5 erros

Percentualmente, no entanto, esses 22 erros representam apenas 3.68% de classificações equivocadas:

[22(=n_o erros) / 597(=n_o classificações realizadas)] x 100 = 3.68

o que parece garantir uma boa margem de segurança para classificar, via **DISCRIM**, em termos das 3 categorias sugeridas.

A figura 8.1 mostra como uma classificação nessas 3 categorias subdividiria o *cohort* inicial.

Como se observa, com essa nova classificação tríplice (que parece bastante confiável) o *cohort* mais extenso contém apenas 8 elementos - (bala, bola, gala...) - e a maioria tem 4 ou menos. A média é de 3.5 palavras por *cohort*, o que significa que usando apenas 3 categorias largas (em vez dos 7 fonemas) foi possível reduzir a procura em mais 90% (de 35 palavras no *cohort* inicial

para uma média de 3.5).

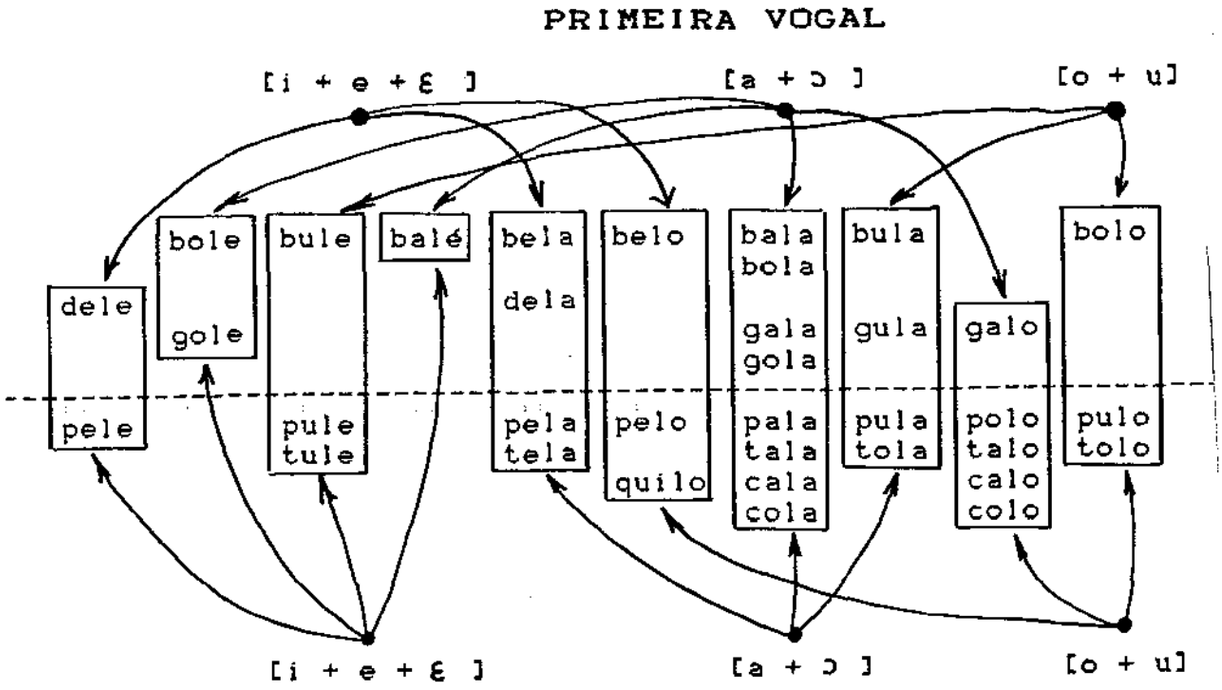


FIGURA 8.1

Se o analisador acústico for capaz de distinguir o traço [+/-sonoro] na consoante inicial (linha tracejada) o conjunto máximo será de apenas 4 elementos; sendo possível especificar a consoante inicial a decisão será, no máximo, entre apenas duas palavras candidatas.

É importante ressaltar que a dificuldade envolvida na decisão depende das características intrínsecas de cada *cohort*. Para dividir {dele, pele}, por exemplo, a distinção do traço [+/-sonoro] na consoante inicial já é suficiente. Já em {bole, gole}

o traço [+/-sonoro] é irrelevante, sendo preciso diferenciar de outro modo a consoante inicial - uma estratégia segura poderia ser, por exemplo, comparar as direções das transições de F2 (ascendente em /bɔ/ e descendente em /gɔ/; Cf. Delattre *et al.* 1955). No *cohort* (balé), de um elemento, a decisão correta depende apenas da classificação vocálica larga. A divisão de um *cohort* em itens individuais estará, portanto, associada a um conjunto ótimo de estratégias, no sentido de haver uma exploração máxima das características estruturais do *cohort*, minimizando assim a necessidade de análises acústicas detalhadas.

É claro que o método de acesso lexical a partir de classificação fonética larga presume que o sistema **SAIBA** que o *input* é uma palavra, porisso é particularmente apropriado para reconhecimento de palavras isoladas. No entanto, se houver cuidado na produção, no sentido de fornecer pistas prosódicas suficientes para a segmentação em palavras, as decisões podem ser até facilitadas: se a palavra *target* aparece contextualizada em uma sentença, é possível acrescentar algumas restrições que reduzam o tamanho do *cohort* determinado pela análise acústica larga; critérios como frequência de ocorrência da palavra na língua (ou, mais especificamente, no universo estabelecido), adequação sintática e aceitabilidade semântica podem ser combinados para reduzir o espaço de procura ou aumentar a probabilidade de uma identificação correta (v. Shigenaga *et al.* *1986; Stock et al. 1989*).

COMENTÁRIO FINAL

A natureza basicamente expositiva do presente trabalho faz com que não caiba aqui, propriamente, qualquer tipo de conclusão. Faremos, portanto, apenas um levantamento de algumas hipóteses/indícios sugeridos pelo conjunto de dados examinados.

1- o mecanismo perceptual humano é extremamente flexível e adaptável a diferentes estímulos e tarefas; resultados experimentais de testes psico-perceptuais devem, portanto, ser vistos com extrema cautela, especialmente no caso de sons linguísticos apresentados em isolamento. As questões relacionadas à percepção da fala fluente diferem substancialmente daquelas ligadas à identificação de fonemas ou traços, pois inevitavelmente envolvem a participação do sistema cognitivo do ouvinte. Um modelo adequado precisa considerar a interação de várias fontes de informação, desde as transformações iniciais efetuadas a nível do sistema auditivo periférico até os processos centrais que lidam com unidades mais estruturadas.

2- na identificação/interpretação de sons linguísticos o ouvinte, aparentemente, não dá a mesma atenção a todos os aspectos do

sinal de fala; o processo perceptual será melhor entendido como uma série de OPÇÕES, ou seja, o ouvinte, para tomar uma decisão, só utiliza aquelas propriedades do sinal que são informativas em algum ponto no tempo.

3- no caso específico das vogais as pistas acústicas podem ser classificadas em duas grandes divisões que se entrecruzam. A primeira divisão relaciona-se com o aspecto temporal: algumas propriedades podem ser extraídas "instantaneamente" - e são, portanto, estáticas - e outras exigem uma janela maior de análise - são dinâmicas. A segunda divisão separa as caracterísiocas inerentes ao segmento fonológico (intrínsecas) daquelas que dependem de algum tipo de relação com outros trechos do sinal (extrínsecas). O quadro abaixo resume as características mais importantes segundo essa subdivisão.

	ESTÁTICO	DINÂMICO
INTRÍNSECO	-Padrão de Formantes -F0 Intrínseco -"Distância F1-F0" -Amplitude relativa dos formantes	-Duração Intrínseca -Mudança espectral intrínseca
EXTRÍNSECO	-Padrão de Formantes referido a outras vogais do falante -F0 médio	-"gesto" coarticulado

(é preciso considerar, no entanto, que na fala fluente essas propriedades podem ser consideravelmente alteradas pela interação

com o nível suprasegmental e alguns dos constrastes observáveis em estímulos isolados podem ser eventualmente neutralizados)

4- certos aspectos dos segmentos vocálicos talvez não possam ser caracterizados apenas com base nas pistas acústicas descritas no item 4. Esse parece ser o caso da nasalidade, provavelmente associada a características globais do espectro.

5- apesar da alta discriminabilidade dos estímulos vocálicos em isolamento, os sistemas fonológicos, via-de-regra, possuem um número reduzido de vogais, isto é, embora o mecanismo perceptual seja capaz de realizar distinções sub-fonêmicas bastante finas, os pontos do sistema vocálico mantém entre si uma distância acústico-perceptual MAIS do que suficiente. O fato é que "distância perceptual" ou "saliência fonética" não são os únicos fatores que influem na organização do sistema; a estrutura interna de um sistema fonológico é também condicionada por restrições de produção (articulatórias e sensório-motoras) que fazem com que o espaço fonológico não seja homogêneo em todas as direções - a prevalência de vogais anteriores e de distinções na dimensão "abertura" parecem ser consequências de algumas dessas restrições. Seria incompleto, no entanto, considerar o sistema vocálico em separado: vogais não ocorrem isoladamente na fala real. As restrições articulatórias valem também , e principalmente, para a estruturação a nível silábico. Lindblom *et*

al. sugerem que o conjunto de sílabas de uma língua obedece a um princípio geral de "menor esforço" articulatório, sem contudo comprometer a saliência perceptual, ou seja, as combinações CV e VC de uma língua tenderiam a ser de tal forma que garantissem "sufficient perceptual differences at acceptable articulatory costs" (Lindblom *et al.* 1984:193). É claro que cada língua encontra sua própria solução otimizada de "economia interna", mas de forma geral, a maior incidência de sistemas com mais consoantes do que vogais sugere que esse arranjo é mais conveniente segundo um critério genérico de "custo/benefício" articulatório e perceptual.

6- procedimentos estatísticos relativamente simples são capazes de classificar com bom índice de acerto as vogais orais do PB com base apenas nas frequências dos três primeiros formantes, mesmo que o conjunto de dados tenha sido produzido por um grupo heterogêneo quanto a sexo e idade. Algumas vogais provocam mais confusões (especialmente /a/ e /ɔ/), outras (especialmente /i/) são particularmente bem identificadas. Como as vogais foram produzidas em contexto fixo /pVs/ não imerso em sentença, resta saber em que medida esses resultados se manteriam para vogais extraídas da fala fluente, onde pode haver um grau considerável de redução vocálica.

7- embora a distinção vogal/consoante seja, intuitivamente, a bifurcação mais óbvia dos sons de fala, não é certo que, do ponto de vista acústico, seja possível estabelecer um limite seguro entre as duas categorias: na fala fluente a identificação de um segmento fonológico (vogal ou consoante) depende da informação extraída das regiões transicionais. Resultados experimentais recentes têm enfatizado que essa informação dinâmica transicional não é meramente um efeito "destrutivo" da coarticulação e talvez se constitua no aspecto acústico mais consistente e invariante para a caracterização de vogais. Ao contrário das teorias *target* tradicionais, que encaram esses fatores transientes como portadores de ruido, é razoável supor que a sobreposição de fonemas devida à coarticulação é até perceptualmente vantajosa, na medida em que faz com que a fala seja mais resistente ao ruído, dispersando a informação de cada fonema sobre um maior intervalo de tempo.

NOTAS

Seção 1

1) Algumas referências usadas nessa curta retrospectiva histórica foram obtidas indiretamente, já que vários textos originais não eram acessíveis. As fontes consultadas foram: Helmholtz 1877, Scripture 1902, Fletcher 1953, Ladefoged 1967, Stevens e House 1961, Flanagan 1972, Linggard 1985 e Jenkins 1987.

2) A constatação de que o distanciamento psicológico entre dois sons de um sistema fônico não tem uma relação simples e direta com a "realidade acústica" não é um fato novo. Sapir 1925, embora sem evidência experimental, já sugerira que as características particulares de um sistema fonológico podem determinar, em certa medida, as avaliações perceptuais dos usuários desse sistema.
Vários estudos inter-lingüísticos têm reafirmado a impossibilidade de se postular um espaço perceptual universal em relação aos sons da fala. Falantes de diferentes línguas nativas (ou mesmo dialetos, como em Ladefoged 1967) respondem ao mesmo estímulo de forma mais ou menos diferente; análises estatísticas que possibilitam a simulação de espaços perceptuais n-dimensionais mostram que também fatores etários influem na geometria desses espaços. Parece que adultos e crianças organizam seus espaços vocálico-perceptuais segundo padrões diferentes, não só no que diz respeito às distâncias relativas entre os elementos do sistema, mas também quanto à própria dimensionalidade da percepção: Butcher 1982 verifica, por exemplo, que crianças tendem a ouvir vogais em três dimensões, enquanto os adultos estruturam sua percepção em apenas duas dimensões. É possível que, durante o processo de amadurecimento da aquisição lingüística, o ouvinte elimine gradativamente alguns aspectos redundantes do sinal de fala e integre dimensões não-ortogonais. Isso talvez explique a redução dimensional no espaço perceptual adulto.

3) É preciso considerar, no entanto, que certas situações atípicas podem alterar um pouco esse quadro. Se houver, por exemplo, um aumento substancial na velocidade e volume do ar, é possível obter as condições de turbulência para a produção de fricativas mesmo com aberturas consideráveis.

4)Foram usados os termos originais *pluck* e *bow* por não haver uma tradução razoável. Como se trata de rótulos mais ou menos arbitrários, cuja única finalidade é aferir a possível resposta categorial do sujeito, o significado é de importância secundária. De qualquer forma, esses termos estão associados a modos de interpretação musical em instrumentos de arco (violino, violoncelo, etc) e correspondem aos ataques "stacatto" e "legato", respectivamente.

5)Hoy e Paul 1973, Capranica 1965, Frishkopf e Goldstein 1963 e Wollberg e Newman 1972 *apud* Lieberman 1977, 1984

6)Carlson *et al.* 1975 relatam um experimento bi-auricular que, embora não baseado em um paradigma de adaptação seletiva, suporta a tese de um processamento não-periférico das vogais. Eles apresentam aos sujeitos várias combinações de formantes diferentemente distribuídas entre os dois ouvidos; por exemplo: F1 no ouvido esquerdo, F2,F3,F4 no direito e vice-versa; F1 e F2 no esquerdo, F3 e F4 no direito, e vice-versa, etc. Verificou-se que a alternância não produz diferenças perceptuais significativas. Isso indica que, em algum estágio, mais ou menos central, há uma soma de componentes auditivos.

7)Aceitaremos provisoriamente que a organização do processamento se dá de forma hierarquizada, já que os níveis de análise devem ser ao menos parcialmente sucessivos, de modo a preservar a ordem temporal. Na seção 8 veremos que essa hierarquia não é, no entanto, muito rígida.

8)Deve ser frisado que estamos falando das transformações iniciais do *input* sensorial, e as possibilidades de processamento serial devem ser entendidas, por enquanto, nesse âmbito (acústico-fonético). Mais adiante (seção 8) examinaremos a hipótese de um modelo paralelo mais amplo, incluindo a interação com níveis superiores (lexical, sintático, semântico, etc).

9)Os resultados de Lane têm sido contestados. Réplicas com estímulos similares aos que ele usou observaram que o treinamento afeta apenas alguns sujeitos (Cf. Cutting 1977). Além disso, Studdert-Kennedy *et al.* 1970, em um perspicaz comentário sobre os dados de Lane, observam que há uma importante distinção entre as curvas de discriminação de Lane e as curvas produzidas pela percepção categorial "natural". Embora surjam fronteiras categoriais mais ou menos claras depois do treinamento,

a capacidade de discriminação entre estímulos próximos do mesmo lado do contínuo não parece se alterar significativamente, permanecendo no mesmo nível (alto) anterior ao treinamento.

10)A adaptação a baixas razões sinal/ruído não depende apenas de uma reorientação da atenção. Na verdade o mecanismo autitivo está equipado para essa adaptação; a análise feita ao nível da cóclea não é linear, isto é, a forma dos filtros não permanece invariante em função do nível de pressão sonora. As respostas da membrana basilar - suas características de filtragem - se ALARGAM para os níveis de alta amplitude. Essa característica permite que a fala seja inteligível mesmo a taxas bem baixas da razão sinal/ruído; tão baixas quanto 5 dB, enquanto aparelhos artificiais já têm problemas em realizar análises espectrais com 25 dB de razão sinal/ruído (Cf. Greenberg 1988).

Seção 2

1)O recurso a uma transcrição larga não interfere na análise, já que o objetivo aqui é apenas demonstrar a distribuição das qualidades segundo critérios mais ou menos amplos de espacialização articulatória.

2)Um dos sujeitos, entretanto, não revelou correlações significativas para o movimento da língua. Curiosamente, esse sujeito era um dos três foneticistas experientes participantes do experimento. É estranho que LL85 não comentem esse fato, já que há algumas sugestivas implicações nele embutidas. Embora seja uma questão a ser examinada mais cuidadosamente, é possível que certas idéias pré-concebidas a respeito do movimento da língua durante a produção de vogais tenha, de algum modo, influenciado os julgamentos introspectivos desse foneticista. Vale lembrar, apenas a título de analogia, que ao testar a precisão da transcrição de contornos entoacionais por dois foneticistas experientes, Lieberman 1965 observa que as transcrições, apesar de consistentes inter- e intra-subjetivamente, não correspondiam ao movimento real de FO. Ao impor, entretanto, os mesmos padrões entoacionais sobre um timbre fixo equivalente à vogal /a/, verificou-se que as transcrições apresentavam mudanças consideráveis na direção de uma maior exatidão objetiva. Lieberman conclui, com base nesses resultados, que "the linguists' ears were remarkably good as long as they did not hear the message. A natureza e magnitude dos efeitos da experiência lingüística sobre testes sensoriais estão

longe de serem inteiramente determinadas, mas o fenômeno, por si só, nos previne, mais uma vez, quanto a certas dificuldades inerentes aos testes psicolingüísticos. Parece ser uma regra geral que quanto mais "lingüística" se configura a tarefa, menos "objetividade" é possível nas avaliações. Testes preliminares realizados por LL85 envolvendo tarefas não-lingüísticas (avaliação de áreas de círculos, de movimento vertical do antebraço e de movimentos mandibulares sem produção de som) apresentaram correlações mais altas que os testes onde se exigia a produção real de sons de fala. Os próprios sujeitos, relatando o grau de dificuldade associado a cada movimento específico ordenam as tarefas do seguinte modo (em ordem crescente de dificuldade auto-avaliada):

movimento do antebraço - mov. mandibular (silente) - mov. mandibular (fala) - mov. da língua (fala).

3)Poder-se-ia objetar, quanto ao gráfico baseado nas medidas subjetivas, que os experimentos de LL85 não exigiram dos sujeitos uma comparação direta entre os movimentos da mandíbula e da língua. Seria falso, portanto, segundo essa objeção, equalizar em uma mesma representação as duas unidades subjetivas. Os autores lembram, entretanto, que as estimativas foram realizadas com base em valores **RELACIONAIS**, não absolutos, isto é, os sujeitos não atribuem um valor escalar fixo a cada realização, mas sim um número que indicasse **PROPORCIONALMENTE** a quantidade de movimento em relação a uma sílaba referencial (/jɛ/). A informação relevante é, pois, a **RAZÃO** entre os valores estimados, o que parece viabilizar a comparação direta entre os dois tipos de movimento.

4)Esses dados são consistentes com observações referentes ao processo e aquisição e vogais por crianças. Buhr 1980 relata que crianças com idade abaixo de dois anos têm uma tessitura relativamente estreita de variação em F2. A habilidade em usar variações de F2, refletindo um controle do movimento lingual antero-posterior, parece ser desenvolvida **MAIS TARDE** do que o controle de F1, relacionado aos movimentos da língua e da mandíbula no eixo vertical.

5)O modelo de Jones não é de fácil legitimação. Butcher 1982 demonstra, com argumentação bem sólida, que os princípios básicos norteadores do sistema, tais como arrolados por Abercrombie 1967:154 são de difícil sustentação, ou seja, as vogais cardinais seriam:

- arbitrariamente selecionadas
- exatamente determinadas e invariáveis
- periféricas
- auditivamente equidistantes

Butcher demole cada um desses pressupostos e nos deixa uma pergunta no ar: para que serve afinal o diagrama das vogais cardinais?

Mesmo que admitamos uma função meramente didática para o modelo não é certo que a qualidade exata de cada vogal cardinal seja facilmente transmissível. Descrições escritas, como insistia o próprio Jones, não são eficientes para o aprendizado (Abercrombie 1967:155). Segundo Abercrombie 1967:155, palavras-chave extraídas de línguas naturais "are quite useless for learning to pronounce them [as vogais cardinais]". Gravações também não seriam eficientes, pois "most people would not be able to learn to pronounce them simply from hearing them...", embora, admite Abercrombie, "exceptionally gifted individuals might perhaps manage it..." (grifo nosso). A maioria das pessoas, conclui ele, "need explicit instruction...under the direct supervision of a competent teacher". A questão aqui é julgar a validade de um sistema que depende de certos dotes "excepcionais" do aprendiz e da supervisão pessoal de um mestre "iniciado" (um foneticista inglês, presumivelmente...). O curioso é que mesmo foneticistas (ingleses!) treinados na tradição das vogais cardinais podem variar consideravelmente quanto à sua própria percepção dessas qualidades ao longo de curtos períodos de tempo (Laver 1965; apud Butcher 1982:64). É preciso considerar também em que medida o uso não crítico do modelo em salas de aula pode conduzir à cristalização de noções errôneas sobre as realidades articulatória e acústico-perceptual.

6) A qualidade basicamente periférica das vogais orais do PB permite uma comparação com os dados de LL85. O modo de produção dessas vogais (em ambiente fonético fixo) e o uso de valores médios contribui também para eliminar alguns fatores perturbadores de variação contextual e subjetiva.

7) É possível que se coloque alguma objeção quanto aos valores arbitrários atribuídos. Os valores traduzem apenas a classificação tradicional em vogais altas (/i/,/u/), médias (/e/,/ɔ/,/o/) e baixas (/a/), na dimensão "altura", e em anteriores (/i/,/e/,/ɛ/), centrais (/a/) e posteriores (/ɔ/,/o/,/u/), na dimensão "anterioridade". O fato de não termos utilizado mais valores intermediários tem dois motivos. Em primeiro

lugar, não é muito seguro afirmar que um /e/ é, invariavelmente "mais alto" que um /ɛ/, por exemplo. Em segundo lugar - o que é mais importante para o cálculo em questão - aumentar a quantidade de valores escalares não alteraria em praticamente nada o resultado da correlação.

8)As menores correlações entre F1 e a dimensão "altura" parecem indicar que nessas distinções existem outros fatores envolvidos. Um desses fatores pode ser a necessidade de manter uma distância crítica entre F1 e F0 nas vogais altas anteriores (Cf. Syrdal e Gopal 1986; Traunmüller 1988). Esse assunto será discutido adiante, na seção 5.

9)A correção de F2 em relação a F3 de modo a obter uma representação com apenas duas ressonâncias não parece prejudicar muito a percepção das qualidades vocálicas, embora haja uma pequena diminuição de informação para vogais altas anteriores (Carlson et al. 1975). Além disso, já se demonstrou que é possível derivar F3 a partir de F2 e F1 com base em procedimentos matemáticos relativamente simples, cujos eventuais desvios de previsão encontram-se abaixo do limite perceptual de discriminação para F3 (Sato et al. 1982).

10)Lieberman 1967 sugere, de forma genérica, que os códigos lingüísticos aparentemente evitam exigir demais do sistema auditivo humano, operando sempre com uma "margem de erro". Seria particularmente importante conhecer essa margem de erro, de modo a interpretar corretamente o papel dos vários parâmetros acústicos na percepção da fala. Maior atenção a essa questão talvez evitasse algumas discussões estéreis sobre a maior ou menor eficácia dessa ou daquela pista para a percepção de alguma dimensão lingüística. É preciso considerar, no entanto, que, *a priori*, não existe informação totalmente redundante no sinal de fala; determinado tipo de informação pode ser irrelevante para um nível de codificação, mas relevante para outro. Certas características da fonte glotal, por exemplo, talvez não tenham importância capital nas oposições segmentais, mas são informativas ao nível da codificação paralingüística (para uma discussão menos técnica e mais filosófica dessa questão, ver Granger 1968, 1971).

11)Na verdade, essas línguas podem possuir variantes sub-fonêmicas. Halle 1970 relata que KABARDIAN, uma língua caucasiana, tem cerca de 17 qualidades. No entanto, todas são variantes posicionais de duas vogais básicas

/a/ e /ə/.

Seção 3

1)Quando o *input* para o programa DISCRIM inclui também as categorias nasais, apenas as vogais /a/, /ɛ/ e /ɔ/ atingem um índice de classificações corretas nitidamente acima de 50%. No caso das vogais /ɛ/ e /ɔ/ isso se explica pelo fato de não haver equivalentes nasais para essas vogais abertas no PB. Já para a vogal /a/, o menor número de confusões oral/nasal parece estar relacionado à .mudança de qualidade fonética que ocorre quando essa vogal é nasalisada, envolvendo uma centralização na direção do chuá; desse modo, o equivalente nasalizado de /a/, será melhor representadocomo /ɐ̃/, e não /ã/. É interessante notar que, no teste perceptual realizado por Behlau 1984, são exatamente as vogais /a/, /ɛ/ e/ɔ/ aquelas que são menos confundidas com categorias nasais próximas (v. Tabela em Behlau 1984:38).

2)É notável que a maioria ds trabalhos utilizando unidades BARK não explicitem esse ponto. Traunmüller 1981:1465, por exemplo, fala de "...a scale representing critical bands with unit width (1 BARK), wich may also be considered a tonality scale" (grifo nosso). A concepção original de Zwicker 1961 era, no entanto, expressar "the natural division of the audible range by the ear", com cada "banda crítica" tendo uma certa largura, dentro da qual o ouvido integraria qualquer frequência. Não se trata, pois, *stricto sensu*, de uma escala "tonal". É por esse motivo que não faz muito sentido falar em valores "absolutos" em BARK, já que o importante são apenas as DIFERENÇAS expressas nessa unidade.

3)Harris se refere, é claro, à música tonal tradicional. Muitos compositores abandonaram, já há algum tempo, qualquer tipo de apoio tonal para suas obras, seja não respeitando as regras tradicionais da Harmonia e Contraponto, seja eliminando - de forma ainda mais radical - o semitom como unidade escalar básica para suas obras. Para alguns críticos, a desconsideração de certas propriedades auditivas naturais condena esse tipo de música ao esquecimento da História (Hindemith 1937; Pleasants 1955). Essa pode ser uma posição esteticamente reacionária; no entanto, é notável a pouca receptividade da música não tonal por parte do público em geral. É provável que a impopularidade da música não baseada em relações harmônicas tonais venha, em parte, da dificuldade que tem o ouvido em assimilar estruturas sonoras que não reflitam a série dos harmônicos

naturais, já que o mapeamento das freqüências a nível da cóclea parece ser melhor descrito em termos de relações logarítmicas (Greenwood 1961).

Seção 4

1)A ausência de interseção de /i/ com outras categorias é particularmente notável, já que optou-se aqui por uma representação dos limites reais de cada classe, isto é, cada área fechada inclui todas as observações para as categorias por elas definidas. Vale observar, no entanto, que houve bastante cuidado no sentido de traçar os limites de cada classe, levando em conta as diferentes densidades nas várias regiões, de modo a não sobrecarregar demais certas áreas do gráfico.

2)Lieberman 1984:163, com base nos valores médios de F1 e F2 para os dados e Peterson e Barney 1952, verifica também uma grande proximidade entre o /a/ dos adolescentes e o /ɔ/ dos adultos; ele observa que essa sobreposição é consistente com o grande número de confusões perceptuais entre /a/ e /ɔ/ nos testes de Peterson e Barney 1952. Também nos dados de Behlau 1984 existe um grande número de erros envolvendo essas duas vogais. No entanto, há algumas diferenças entre os dois estudos. Lieberman parece não ter levado em conta que, como relatam Peterson e Barney, as baixas taxas de identificação para /a/ e /ɔ/ resultam primariamente do fato de alguns membros do grupo de falantes, e muitos do grupo de ouvintes, falarem dialetos que não diferenciam essas duas qualidades. Esse não parece ser, entretanto, o caso em Behlau 1984, já que, dentro de cada grupo, as distâncias /a/-/ɔ/ não são particularmente pequenas.

3)Embora /i/ seja de fato a vogal com F2 mais alto e maior distância entre F1 e F2, essas não são as únicas pistas efetivas para seu reconhecimento. Estudos recentes têm ressaltado, por exemplo, a importância da manutenção de uma distância crítica menor que 3 BARK entre F1 e F0 para a caracterização de vogais altas, entre elas /i/ (Syrdal e Gopal 1986; Traunmüller 1988). Os próprios Delattre et al. 1952 relatam que, na aproximação de dois formantes para /i/ "progressive reductions in the intensity of the first formant resulted finally...in a destruction of vowel color" (pg. 230), o que indica que F1, de alguma forma, também cumpre algum papel.

Seção 5

1)Riordan 1980 rejeita essa hipótese com base no fato de que o abaixamento da laringe também é observado nas oclusivas nasais. Com a participação das cavidades nasais, diz Riordan, não haveria necessidade de evitar o aumento brusco da pressão supra-glotal, logo o abaixamento da laringe deve cumprir outra função. A observação de Riordan é interessante, mas é preciso considerar que o abaixamento da laringe nas consoantes sonoras talvez tenha se tornado um ato reflexo involuntário que é executado sempre que o traço sonoridade está presente, independentemente da co-presença do traço nasalidade. Além disso, deve ser levado em conta que movimentos articulatórios raramente cumprem uma única função, e é possível que o abaixamento da laringe nas nasais esteja associado a exigências específicas desse tipo de segmento.

2)Essa alta sensibilidade a micro-perturbações de F0 obriga modelos realistas de síntese entoacional a incluir, apesar das dificuldades, regras relacionadas com efeitos segmentais na variação de F0 (v. "componente micro-prosódico" em Thorsen 1980 e "variação micro-melódica" em DiCristo e Hirst 1986).

3)No caso especial da mudança de qualidade em vogais cantadas é preciso levar em conta a possível presença de outros fatores não diretamente relacionados com a relação F0/Formantes. Cantores profissionais adotam posturas articulatórias típicas que forçam o aparecimento de uma ressonância característica na região entre 2500-3000 Hz (o "singing formant"). Essas manobras parecem envolver o abaixamento da laringe e a eventual expansão do ventrículo de Morgagni e têm, provavelmente, o objetivo de conseguir um timbre "brilhante", especialmente nas vogais posteriores (v. Sundberg 1970,1974 para uma interpretação acústico-articulatória desse fenômeno; v. também Nolan 1983:151 para contra-evidência à hipótese do abaixamento da laringe na produção do "singing formant").

4)Existem, é claro, casos onde o F0 médio não é compatível com as dimensões do trato - como na voz do Popeye -, mas trata-se de casos excepcionais.

5)As médias de cada formante foram calculadas com base apenas nas vogais orais, com as frequências expressas em Hz. Como para algumas vogais /u/ faltavem medidas de F3,

retirou-se essa vogal do cálculo do F3 médio.

Seção 6

1)Algumas hipóteses alternativas explicando porque vogais isoladas são percebidas com menor precisão foram esboçadas. A argumentação baseava-se principalmente em dois pontos: (1) vogais isoladas, em geral, não ocorrem na fala, enquanto sílabas CVC (em inglês) são freqüentes e, não raramente, representando morfemas; e (2) algumas vogais (/I,ɛ,^,U/) não podem ocorrer em posição silábica final (em inglês americano e canadense), dessa forma elas seriam fonologicamente pouco apropriadas para testes em isolamento (Macchi 1980; Diehl *et al.* 1981; Assmann *et al.* 1982; Rakerd *et al.* 1984). No entanto, apesar da inegável validade dessas objeções, nenhum estudo experimental provou que vogais isoladas são **MELHOR** identificadas que vogais coarticuladas(Cf. Strange 1987;1989).

Seção 7

1)A escolha da largura de banda de 1/3 de oitava não é arbitrária. Experimentos perceptuais examinando a capacidade dos ouvintes em separar os 5 primeiros harmônicos de um tom complexo, indicam que o poder analisador de freqüência do ouvido humano é comparável a filtros passa-banda com cerca de 1/3 de oitava (Plomp 1975). Alguns estudos, entretanto, utilizam bandas mais estreitas, chegando a transformações de até 35 dimensões (Li *et al.* 1972;1973a). Há alguma controvérsia quanto à largura de banda que melhor simularia os filtros auditivos (Moore e Glasberg 1983); o assunto será tratado mais detalhadamente adiante.

2)Uma análise estatística de fatores ou de componente principal é, basicamente, uma técnica de redução de dados que procura novas "direções" através do espaço original; essas novas direções, ou "fatores", são combinações lineares das dimensões originais (no caso específico, o *output* dos filtros), e tentam dar conta, tanto quanto possível, da variância entre as classes.

3)Experimentos envolvendo medidas nos terminais nervosos causam prejuízos irreversíveis no aparelho auditivo, o que invabiliza a observação direta de humanos. A habilidade em formar uma representação espectral, entretanto, não é exclusiva do homem. Sabe-se que a caacidade humana de resolução de freqüências é equivalente a de outros mamíferos; dessa forma, os

resultados FÍSICOS com base na observação de alguns animais (especialmente gatos) não devem diferir, teoricamente, dos dados hipotéticos para humanos (Greenberg 1988).

4)O *fon* é uma unidade que quantifica a sensação subjetiva da intensidade sonora, sendo uma medida, portanto derivada de experimentos psicofísicos. Definiu-se como referencial um som de 1 kHz com *n* dB. Na freqüência de 1 kHz, portanto, as unidades *fon* coincidem com as medidas de dB. Em outras freqüências só eventualmente ocorrem coincidências, já que a sensibilidade auditiva varia em função da freqüência. Para um som de 1 kHz, 10 *fons* equivalem a 10 dB, mas para um som de 100 Hz, por exemplo, essa mesma sensação auditiva só seria produzida com uma pressão sonora de 45 dB aproximadamente (v. Jeans 1937:228). Deve ser observado que a medida em *fons* é relativa e exprime apenas a equivalência auditiva entre um determinado som e ou outro som de 1 kHz a *n* dB. Existe uma outra unidade, o *sone*, que já foi sugerida como medida ABSOLUTA da sensação subjetiva: um número 2n de *sones* deve dar a sensação subjetiva de um som duas vezes mais intenso (Nepomuceno 1977). Optamos pela tradução AUDIBILIDADE (para *loudness*) para não haver confusão com SONORIDADE, que seria a dimensão expressa em *sones*. Alguns autores, no Brasil, entretanto, falam de SONORIDADE, mesmo quando a unidade é o *fon* (v.p.ex. Pauli et al. 1980).

5)Experimentos perceptuais com vogais artificiais de 2 formantes confirmam isso. Já se verificou que vogais altas anteriores (/i,y,e/, por exemplo) sintetizadas soam naturalmente se o segundo e o terceiro formantes são substituídos por um formante único localizado em algum ponto entre essas duas ressonâncias (Delattre et al. 1952; Carlson *et al.* 1975), sugerindo que o ouvido deve integrar perceptualmente máximos espectrais em regiões de alta freqüência.

6)A discrepância entre as estimativas pode parecer estranha, já que todas as avaliações são baseadas em experimentos psicofísicos. É importante notar, no entanto, que esses experimentos podem utilizar diversos tipos de estímulo (sons senoidais, sons complexos, ruído, etc.) ou envolver tarefas diferentes (reconhecimento de sons com mascaramento, escalamento de freqüências, *loudness* de sons complexos, etc). É possível que as características de filtragem do sistema auditivo periférico sejam adaptáveis ao tipo de tarefa e de estímulo, o que coloca uma dificuldade adicional para

o modelamento da audição (v. também nota 10 à seção 1).

Seção 8

1)A influência do contexto lexical na identificação e reconhecimento de fonemas deve ser considerada em testes perceptuais, sob o risco de haver distorção. Esse pode ter sido o caso em Behlau 1984 em relação ao grande número de erros perceptuais para estímulos nasalisados; como as vogais a serem reconhecidas eram apresentadas no contexto /pVs/, as combinações com vogais nasalisadas produziam logatomas, o que pode ter dificultado a identificação.

2)Na verdade, a própria noção de "nível de representação" deve ser encarada com algum cuidado. Marslen-Wilson e Tyler 1980 observam que um modelo interativo presume que a representação, a cada momento, contenha TODA a informação até então obtida. Dessa forma, se uma palavra pode ser identificada, então ela é obrigatoriamente percebida como PALAVRA; na medida em que seqüências de palavras possam ter uma descrição estrutural, isso se tornará automaticamente parte de sua representação perceptual; no momento em que uma INTERPRETAÇÃO possa ser alcançada, então as palavras serão percebidas dentro desse novo contexto. A análise do *input* "must propagate as far as its properties permit. The extent to which a given input can propagate will determine the 'level of representation' at which it becomes perceptually available" (Marslen-Wilson e Tyler 1980:66).

B I B L I O G R A F I A

ABREVIATURAS:

JASA = Journal of Acoustical Society of America
JSHR = Journal of Speech and Hearing Research
J.Phon = Journal of Phonetics
J.Mem.Lg. = Journal of Memory and Language
L.Speech = Language and Speech
Perc.Psych. = Perception & Psychophysics

ABERCROMBIE,D. 1967.*Elements of General Phonetics*, Aldine.Atherton,NY

ADES,A.E. 1974."How phonetic is selective adaptation?Experiments on syllable position and vowel environment",*Perc.Psych.* 16:61-66

AINSWORTH,W. 1975."Intrinsic and extrinsic factors in vowel judgements", in Fant e Tatham(eds)1975:103-113

ALBANO,E. 1987."Modulado contra modular:Contribuicao ao debate do inatismo", manuscrito, UNICAMP

ANTONIADIS,Z. e H.W. STRUBE 1981."Untersuchungen zum 'intrinsic pitch' deutscher Vokale", *Phonetica* 38:277-290

ASLIN,R.N. 1989."Discrimination of frequency transitions by human infants", *JASA* 86:582-590

ASSMANN,P., T. NEAREY e J. HOGAN 1982."Vowel identification:Ortographic, perceptual and acoustical aspects", *JASA* 71:975-989

ATAL,B.S. e S.L. HANAUER 1971."Speech analysis and synthesis by linear prediction of the speech wave", *JASA* 50:637-655

ATKINSON,J.E. 1973."Intrinsic F0 in vowels:Physiological correlates", *JASA* 53:346

BARCLAY,J.R. 1972."Noncategorical perception of a voiced stop:a replication", *Perc.Psych.* 11:269-273

BEHLAU,M.S. 1984.*Uma Análise das Vogais do Português Brasileiro Falado em São Paulo*, Tese de Mestrado.Escola Paulista de Medicina

BEIL,R. 1962."Frequency analysis of vowels produced in a helium rich atmosphere", *JASA* 34:347-349

BERNSTEIN, J. 1981."Formant-based representations of auditory similarity among vowel-like sounds", *JASA* 69:1132-1144

BILLI,R., G. MASSIA e F. NESTI 1986."Word preselection for large vocabulary speech recognition", *Proc. ICASSP*,Tokyo,65-68

BLAKEMORE,C. e F.W. CAMPBELL 1969."On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images", *Journal of Physiology* 203:237-260

BLECHNER,M.J.,R.S. DAY e J.E. CUTTING 1976."Processing two dimensions of nonspeech stimuli:The auditory-phonetic distinction reconsidered", *Journal of Experimental Psychology:Human Perception and Performance* 2:256-266

BLUMSTEIN,S., E. ISAACS e J. MERTUS 1982."The role of gross spectral shape as a perceptual cue to place of articulation in initial stop consonants", *JASA* 72:43-50

BOLINGER,D. 1986.*Intonation and Its Parts*, Edward Arnold,London

BROAD,D.J. 1976."Toward defining acoustic phonetic equivalence for vowels", *Phonetica* 33:401-424

BROWMAN,C.P. e L.M. GOLDSTEIN 1986."Towards an articulatory phonology", in Ewen e Anderson(eds) 1986:219-252

BUHR,R.D. 1980."The emergence of vowels in an infant", *JSHR* 23:75-94

BUTCHER,A. 1982."Cardinal vowels and other problems", in Crystal(ed) 1982:50-72

BUTTERWORTH,B., B. COMRIE e O. DAHL (eds) 1984.*Explanations for Language Universals*, Mouton

CAGLIARI,L.C. 1977.*An Experimental Study of Nasality with Particular Reference to Brazilian Portuguese*, PhD thesis, Univ. Edinburgh

CAPLAN,D.(ed) 1980.*Biological Studies of Mental Processes*,MIT press

CARLSON,R., G. FANT e B. GRANDSTRÖM 1975."Two-formant models, pitch and vowel perception",in Fant e Tatham(eds) 1975:55-82

CLYNES,M.(ed) 1982.*Music Mind and Brain*,Plenum press,NY

CRUTTENDEN,A. 1986.*Intonation*,Cambridge

CRYSTAL,D. 1969.*Prosodic Systems and Intonation in English*, Cambridge

CRYSTAL,D.(ed) 1982.*Linguistic Controversies*,Edward Arnold,NY

CRYSTAL,T.H. e A.S. HOUSE 1982."Segmental durations in connected speech signals:Preliminary results", *JASA* 72:705-716

CRYSTAL,T.H. e A.S. HOUSE 1988."The duration of American-English vowels:An overview", *J.Phon.* 16:263-284

CUTTING,J.E. 1977."The magical number two an the natural categories of speech and music", in Sutherland(ed) 1977:1-33

CUTTING,J.E., B.S. ROSNER e C.F. FOARD 1976."Perceptual categories for musiclike sounds:Implications for speech perception", *Quarterly Journal of Experimental Psychology* 28

CUTTING,J.E.e B.S. ROSNER 1974."Categories and boundaries in speech and music", *Perc.Psych.* 16:564-570

DAVID,E.E. e P.B. DENES 1972.*Human Communication:A Unified View*,McGraw-Hill,NY

DAY,R.S. e C.C. WOOD 1972."Interactions between linguistic and nonlinguistic processing", *JASA* 51:79

DECHOVITZ,D. 1977."Information conveyed by vowels:A negative finding", JASA sup.1,61:S39

DELATTRE,P.C., A.M. LIBERMAN e F.S. COOPER 1955."Acoustic loci and transitional cues for consonants", *JASA* 27:769-773

DELATTRE,P.C., A.M. LIBERMAN, F.S. COOPER e L.J. GERSTMAN 1952."An experimental study of the acoustic determinants of vowel color:Observations of one- and two-formant vowels synthesized from spectrographic patterns, in FRY(ed) 1976:221-237

DELGADO MARTINS,M.R. 1975."Vogais e consoantes do Português:Estatística de ocorrência,duracão e intensidade", *Boletim de Filologia* XXIV:1-11,Lisboa

DELGADO MARTINS,M.R. 1986.*Sept Études sur la Perception*, I.N.I.C.,Lisboa

DELGUTTE,B. 1980."Representation of speechlike sounds in the discharge patterns of auditory-nerve fibers", *JASA* 68:843-857

DENES,P.P. e E.N.PINSON 1973.*The Speech Chain:The Physics and Biology of Spoken Language*,Anchor Books,NY

DENG,L., M. LENIG e P. MERMELSTEIN 1989."Use of vowel duration information in a large vocabulary word recognizer", *JASA* 86:540-548

DEUTSCH,D. 1982."Organizational processes in music", in Clynes(ed) 1982:119-136

DIEHL, R.L, K.R. KLUENDER, D.J. FOSS, E.M. PARKER e M.A. GERNSBACHER 1987. "Vowels as islands of reliability", *J.Mem.Lg.* 26:564-573

DIEHL,R.L., S.B. McCUSTER e L.S. CHAPMAN 1981."Perceiving vowels in isolation and in consonantal context", *JASA* 68:239-248

DUDLEY,H. 1940."The carrier nature of speech", in Flanagan e Rabiner(eds) 1973:22-42

DiCRISTO,A. e D.J HIRST 1986."Modelling French micromelody:Analysis and Synthesis", *Phonetica* 43:11-30

EIMAS,P.D. 1974."Auditory and linguistic processing of cues for place of articulation by infants", *Perc.Psych.* 16:513-521

EIMAS,P.D., E.R. SIQUELAND, P. JUSCZYC e J.M. VIGORITO 1971."Speech perception in infants", *Science* 171:303-306

EIMAS,P.D., W.E. COOPER e J.D. CORBIT 1973."Some properties of linguistic feature detectors", *Perc.Psych.* 13:247-252

ELMAN,J.L. e J.L. McCLELLAND 1984."Speech perception as a cognitive process: The interactive activation model",in Lass(ed) 1984:337-374

ENGSTRAND,O. 1988."Articulatory correlates of stress and speaking rate in Swedish VCV utterances, *JASA* 83:1863-1875

EWEN, C.J. e J.M. ANDERSON(eds) 1986.*Phonology Yearbook* 3

FAIRBANKS,G. e P. GRUBB 1961."A psychological investigation of vowel formants", *JSHR* 4:203-219

FANT,G. 1960.*Acoustic Theory of Speech Production*, The Hague:Mouton

FANT,G. 1973.*Speech Sounds and Features*,Cambridge

FANT,G. 1980."The relations between area functions and the acoustic signal", *Phonetica* 37:55-86

FANT,G. e M. TATHAM(eds) 1975.*Auditory Analysis and Perception of Speech*, Academic press,London

FLANAGAN, J.L. e M.G. SASLOW 1958."Pitch discrimination for synthetic vowels", *JASA* 30:435-442

FLANAGAN,J.L. 1972."Voices of men and machines", *JASA* 51:1375-1387

FLANAGAN,J.L. e L.R. RABINER(eds) 1973.*Speech Synthesis*, Dowden,Hutchinson e Ross, Pennsylvania

FLETCHER,H. 1953.*Speech and Hearing in Communication*, Krieger,NY,1972

FODOR,J.A. 1983.*The Modularity of Mind*,Cambridge

FODOR,J.A., T.G. BEVER e M.F. GARRETT 1974.*The Psychology of Language*, McGraw-Hill,NY

FOSS,D.J. e D.A. SWINNEY 1973."On the psychological reality of the phoneme: Perception,identification and consciousness", *Journal of Verbal Learning and Verbal Behavior* 12:246-257

FOULKES,J.D. 1961."Computer identification of vowel types", *JASA* 33:7-11

FOWLER,C.A. 1980."Coarticulation and theories of extrinsic timing", *J.Phon.* 8:113-133

FOWLER,C.A. 1986."An event approach to the study of speech perception from a direct-realist percspective", *J.Phon.* 14:3-28

FOWLER,C.A. 1987."Perceivers as realists,talkers too:Commentary on papers by Strange,Diehl et al. and Rakerd and Verbrugge", *J.Mem.Lg.* 26:574-587

FOWLER,C.A. e D.P. SHANKWEILER 1978."Identification of vowels in speech and non-speech contexts", *JASA* 63, sup. 1:S4A

FOWLER,C.A. e M. TURVEY 1980."Immediate compensation for bite block speech", *Phonetica* 37:306-326

FROMKIN,V.A. 1971."The nonanomalous nature of anomalous utterances", *Language* 47:27-52

FROMKIN,V.A.(ed) 1985.*Phonetic Linguistics*, Academic press,NY

FRY,D.B. 1955."Duration and intensity as physical correlates of linguistic stress", *JASA* 27:765-768

FRY,D.B. 1976.*Acoustic Phonetics*,Cambridge

FRY,D.B., A.S. ABRAMSON, P.D. EIMAS e A.M. LIBERMAN 1962."The identification and discrimination of synthetic vowels", *L.Speech* 5:171-189

FUJIMURA,O. 1975."Syllable as a unit of speech recognition",*IEEE trans. ASSP* 23:82-87

FUJIMURA,O. 1980."Modern methods of investigation in speech production", *Phonetica* 37:38-54

FUJIMURA,O.(ed) 1988.Articulatory Organization:From Phonology to speech signals,*Phonetica* sup. 45:77-83

FUJISAKI,H. e T. KAWASHIMA 1968."The roles of pitch and higher formants in the perception of vowels", *IEEE trans. Audio Electroacoustics*, AU-16:73-77

FUJISAKI,H., K. HIROSE, H. UDAKAWA e N. KANEDERA 1986."A new approach to continous speech recognition based on considerations on human processes of speech perception", *Proc. ICASSP*,Tokyo,1959-1962

GANDOUR,J. e J. WINDSOR 1988."Selective impairment of phonation:A case study", *Brain and Language* 35:313-339

GANONG,W.F. 1980."Phonetic categorization in auditory word perception",*Journal of Experimental Psychology:Human Perception and Performance* 6:110-125

GAY,T. 1978."Physiological and acoustic correlates of perceived stress", *L.Speech* 21:347-353

GERSTMAN,L.J. 1968."Classification of self-normalized vowels", *IEEE trans. Electroacoustics* AU-16:78-80

GOTTFRIED,T. e S. CHEW 1986."Intelligibility of vowels sung by a countertenor", *JASA* 79:124-130

GOTTFRIED,T.L. e W. STRANGE 1980."Identification of coarticulated vowels", *JASA* 68:1626-1635

GRANGER,G.G. 1968.*Filosofia do Estilo*, Perspectiva,SP,1974

GRANGER,G.G. 1971."Langue et systémes formels", *Langages* 21:71-87

GREENBERG,S. 1988."Acoustic transduction in the auditory periphery", *J.Phon.* 16:3-17

GREENWOOD,D.D. 1961."Auditory masking and the critical band", *JASA* 33:484-501

HAGGARD,M., Q. SUMMERFIELD e M. ROBERTS 1981."Psychoacoustical and cultural determinants of phoneme boundaries:Evidence from trading F0 cues in thew voiced-voiceless distinction", *J.Phon.* 9:49-62

HAGGARD,M., S. AMBLER e M. CALLOW."Pitch as a voicing cue", *JASA* 47:613-617

HALLE,M. 1970."Is Kabardian a vowelless language?", *Foundations of Language* 6:95-103

HALLIDAY,M.A.K. 1967."Notes on transitivity and theme in English", *Journal of Linguistics* 3:199-244

HARRIS,J.D. 1960."Scaling of pitch intervals", *JASA* 32:1575-1581

HARRIS,M.S. e N. UMEDA 1897."Difference limens for fundamental frequency contours in sentences", *JASA* 81:1139-1145

HELMHOLTZ,H. 1877.*On the Sensation of Tone*, Dover,NY,1954

HEMDAL,J.F. e G.W. HUGHES 1967."A feature based computer recognition program for the modeling of vowel perception", in Wathen-Dunn(ed) 1967:440-453

HILLENBRAND,J. e R. GAYVERT 1987."Speaker-independent vowel classification based on fundamental frequency and formant frequencies", *JASA* sup.1,81:S93

HINDEMITH,P. 1937.*The Craft of Musical Composition*, Schott & Co.,London

HOMBERT,J.M., J.J. OHALA e W.G. EWAN 1979."Phonetic explanations for the development of tones", *Language* 55:37-58

HOUSE,A.S. 1961."On vowel duration in English", *JASA* 33:1174-1178

HOUSE,A.S. e G. FAIRBANKS 1953."The influence of consonant environment upon the secondary acoustical characteristics of vowels", *JASA* 35:84-92

HOUSE,A.S. e G. FAIRBANKS 1953."The influence of consonantal environment upon the secondary acoustic characteristics of vowels", *JASA* 25:105-113

JAKOBSON,R., G. FANT e M. HALLE 1951.*Preliminaires to Speech Analysis:The Distinctive Features and Their Correlates*, MIT press,1976

JAKOBSON,R., e L. WAUGH 1979.*The Sound Shape of Language*, Indiana

JEANS,J. 1937.*Science and Music*, Dover,NY
JENKINS,J.J. 1987."A selective history of issues in vowel perception", *J.Mem. Lg.* 26:542-549

JOHNSON,N.F. 1987."A tutorial symposium on dynamic conceptions of vowel perception:An introduction", *J.Mem.Lg.* 26:539-541

JONES,D. 1918.*An Outline of English Phonetics*,Cambridge,1975

JOOS,M. 1948."Acoustic Phonetics", *Language* sup. 24:1-136

JUSCZYC, P.W., B.S. ROSNER, J.E. CUTTING, C.F. FOARD e L.B. SMITH 1977. "Categorical perception of nonspeech sounds by 2-month-old infants", *Perc. Psych.* 21:50-54

KALLAIL,K. e F. EMANUEL 1984."An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects", *J.Phon.* 12:175-186

KENT,R.D. 1979."Isovowel lines for the evaluation of vowe formant structure in speech disorders", *Journal of Speech and Hearing Disorders* 44:513-521

KENT,R.D. e K.L. MOLL 1969."Vocal-tract characteristics of the stop cognates", *JASA* 46:1549-1555

KENT,R.D. e L.L. FORNER 1979."Developmental study of vowel formant frequencies in an imitation task", *JASA* 65:208-217

KEWLEY-PORT,D. e P.A. LUCE 1984."Time-varying features of initial stop consonants in auditory running spectra:A first report", *Perc.Psych.* 35:353-360

KIANG,N.Y.S. 1980."Processing of speech by the auditory nervous system", *JASA* 68:830-835

KIMURA,D. 1964."Left-right differences in the perception of melodies", *Quarterly Journal of Experimental Psychology* 16:355-358

KLATT,D.H. 1973."Discrimination of fundamental frequency contours in synthetic speech:Implications for models of speech perception", *JASA* 53:8-16

KLATT,D.H. 1976."Linguistic uses of segmental duration in English:Acoustic and perceptual evidence", *JASA* 59:1208-1221

KLATT,D.H. 1987."Review of text-to-speech conversion for English", *JASA* 82:737-793

KLEIN,W., R. PLOMP e L.C.W. POLS 1970."Vowel spectra,vowel spaces and vowel identification", *JASA* 48:1000-1009

KLUENDER,K.R., R.L. DIEHL e B.A. WRIGHT 1988."Vowel-length differences before voiced and voiceless consonants:An auditory explanation", *J.Phon.* 16:153-169

KLUMPP,R.G. e J.C. WEBSTER 1961."Intelligibility of time-compressed speech", *JASA* 33:265-267

KUHL,P.K. e D.M. PADDEN 1983."Enhanced discriminability at the phonetic boundaries for the place feature in macaques", *JASA* 73:1003-1010

KUHL,P.K. e J.D. MILLER 1975."Speech perception by the chinchilla:Voiced-voiceless distinction in alveolar plosive consonants", *Science* 190:69-72

KUHN,T.S. 1962.*A Estrutura das Revolucões Cientificas*, Perspectiva,SP,1978

LADD,D.R. e K.E.A. SILVERMAN 1984."Vowel intrinsic pitch in connected speech", *Phonetica* 41:31-40

LADEFOGED,P. 1967.*Three Areas of Experimental Phonetics*, Oxford

LADEFOGED,P. 1971.*Preliminaires to Linguistic Phonetics*, Univ.Chicago press

LADEFOGED,P. 1975.*A Course in Phonetics*, Harcourt Brace Jovanovich,NY

LADEFOGED,P. e D.E. BROADABENT 1957."Information conveyed by vowels", *JASA* 29:98-104

LANE,H. 1965."Motor theory of speech perception:A critical review", *Psychological Review* 72:275-309

LASS,N.(ed) 1976.*Contemporary Issues in Experimental Phonetics*, Academic press,NY

LASS,N.J.(ed) 1984.*Speech and Language:Advances in Basic Research and Practice*,Academic press

LAVER,J. 1980.*The Phonetic Description of Voice Quality*,Cambridge

LEHISTE,I. 1970.*Suprasegmentals*,Cambridge

LEHISTE,I. e D. MELTZER 1973."Vowel and speaker identification in natural and synthetic speech", *L.Speech*. 16:356-364

LEHISTE,I. e G.E. PETERSON 1961."Some basic considerations in the analysis of intonation", *JASA* 33:419-425

LENNIG,M. e D. HINDLE 1977."Uniform scaling as a method of vowel normalization", *JASA* 62:S26A

LI,K-P., G.W. HUGHES e A.S. HOUSE 1973(a)."Speech reconstituted from spectra of reduced dimensionality:A study of intelligibility", *JASA* 53:S329A

LI,K-P., G.W. HUGHES e T.B. SNOW 1973(b)."Segment classification in continous speech", *IEEE trans. Audio and Electroacoustics*, AU-21:50-57

LIBERMAN,A.M., F.S. COOPER, D.S. SHANKWEILER e M. STUDDERT-KENNEDY 1967. "Perception of the speech code", *Psychological Review* 74:431-461

LIBERMAN,A.M., K.S. HARRIS, H.S. HOFFMAN e B.C. GRIFFITH 1957."The discrimination of speech sounds within and across phoneme boundaries", *Journal of Experimental Psychology* 54:358-368

LIEBERMAN,P. 1965."On the acoustic basis of the perception of intonations by linguists", *Word* 21:40-54

LIEBERMAN,P. 1967.*Intonation,Perception,and Language*, MIT press

LIEBERMAN,P. 1977.*Speech Physiology and Acoustic Phonetics:An Introduction*, MacMillan Pub.,NY

LIEBERMAN,P. 1984.*The Biology and Evolution of Language*,Harvard

LILJENCRANTS,J. e B. LINDBLOM 1972."Numerical simulation of vowel quality systems:The role of perceptual contrast", *Language* 48:839-862

LINDBLOM,B. 1963."Spectrographic study of vowel reduction", *JASA* 35:1773-1781

LINDBLOM,B. 1986."Phonetic universals in vowel systems", in Ohala e Jaeger(eds) 1986:13-44

LINDBLOM,B. e J. LUBKER 1985."The speech homunculus and a problem of phonetic linguistics", in Fromkin(ed) 1985:169-192

LINDBLOM,B. e J. SUNDBERG 1971."Acoustical consequences of lip,tongue,jaw and larynx movement", *JASA* 50:1166-1179

LINDBLOM,B. e M. STUDDERT-KENNEDY 1967."On the role of formant transitions in vowel recognition", *JASA* 42:830-843

LINDBLOM,B. e S-J. MOON 1988."Formant undershoot in clear and citation-form speech", *Perilus* VIII:20-33

LINDBLOM,B., J. LUBKER e T. GAY 1977."Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation", *J.Phon* 5

LINDBLOM,B., P. MacNEILAGE e M. STUDDERT-KENNEDY 1984."Self-organizing processes and the explanation of phonological universals", in Butterworth et al. (eds) 1984:181-203

LINGGARD,R. 1985.*Electronic Synthesis of Speech*, Cambridge

LöFQVIST,A. 1975."Intrinsic and extrinsic F0 variations in Swedish tonal accents", *Phonetica* 31:228-247

MACCHI,M.J. 1980."Identification of vowels spoken in isolation versus vowels spoken in consonantal context", *JASA* 68:1636-1642

MADDIESON,I. 1986."The size and structure of phonological inventories", in Ohala e Jaeger(eds) 1986:105-123

MAKHOUL,J. 1975."Linear prediction:A tutorial review", *Proc. IEEE* 63:561-580

MARKEL,J.D. e A.H. GRAY Jr. 1976.*Linear Prediction of Speech*, Springer Verlag, NY

MARSLEN-WILSON,W. 1975."Sentence perception as an interactive parallel process", *Science* 189:226-228

MARSLEN-WILSON,W. e L.K. TYLER 1980."The temporal structure of spoken language understanding", *Cognition* 8:1-71

MILLER,G.A. 1956."The magical number seven,plus or minus two:Some limits on our capacity for processing information", *Psychological Review* 63:81-97

MILLER,G.A., G.A. HEISE e W. LICHTEN 1951."The intelligibility of speech as a function of the context of the test materials", *JASA* 41:329-335

MILLER,J.D. 1984."Auditory processing of the acoustic patterns of speech", *Arch.Otolaryngol.* 110:154-159

MILLER,J.D. 1989."Auditory-perceptual interpretation of the vowel", *JASA* 85:2114-2134

MILLER,J.D., C.C. WEIR, R.E. PASTORE, W.J. KELLY e R.J. DOOLING 1976. "Discrimination and labelling of noise-buzz sequences with varying noise-lead times:An example of categorical perception", *JASA* 60:410-417

MILLER,M.I. e M.B. SACHS 1983."Representation of stop consonants in the discharge patterns of auditory-nerve fibers", *JASA* 74:502-517

MILLER,R.L. 1953."Auditory tests with sythetic vowels", *JASA* 25:114-121

MOHR,B. 1971."Intrinsic variations in the speech signal", *Phonetica* 23:65-93

MOORE,B. e B. GLASBERG 1983."Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", *JASA* 74:750-753

MOORE,T.J. e J.L. CASHIN 1974."Response patterns of cochlear nucleus neurons to excerpts from sustained vowels", *JASA* 56:1565-1576

MOORE,T.J. e J.L. CASHIN 1976."Response of cochlear-nucleus neurons to synthetic speech", *JASA* 59:1443-1449

MORROW,C. 1971."Speech in deep-submergence atmospheres", *JASA* 50:715-728

MORSE,P.A. 1972."The discriminatios of speech and nonspeech stimuli in early infancy", *Journal of Experimental Child Psychology* 14:477-492

MORSE,P.A. e C.T. SNOWDON 1975."An investigation of categorical speech discrimination by rhesus monkeys", *Perc.Psych.* 17:9-16

McCANDLESS,S.S. 1974."An algorithm for automatic formant extraction usin linear prediction spectra", *IEEE trans, Acoustic Speech Signal Processing* 22:135-141

McCOLLOUGH,C. 1965."Color adaptation of edge-detectors in the human visual system",*Science* 149:1115-1116

NEAREY,T. 1989."Static,dynamic and relational properties in vowel perception", *JASA* 85:2088-2113

NEAREY,T.M. e P.F. ASSMANN 1986."Modeling the role of inherent spectral change in vowel identification", *JASA* 80:1297-1308

NEPOMUCENO,L.X. 1977.*Acústica*, Edgard Blucher,SP

NOLAN,F. 1983.*The Phonetic Bases of Speaker Recognition*, Cambridge

NORD,L., T.V. ANANTHAPADMANABHA e G. FANT 1986."Perceptual tests using an interactive source filter model and considerations for synsthesis strategies", *J.Phon.* 14:401-404

OHALA,J.J. 1972."How is pitch lowered?", *JASA* 52:124

OHALA,J.J. 1985."Around flat", in Fromkin(ed) 1985:223-241

OHALA,J.J. e J.J.JAEGER(eds) 1986.*Experimental Phonology*,Academic press,NY

OHDE,R.N. 1984."Fundamental frequency as an acoustic correlate of stop consonant voicing, *JASA* 75:224-230

PAL,S., A. DATTA e D. DUTTA MAJUMDER 1989."A self-supervised vowel recognition system",*Pattern Recognition* 12:27-34

PARKER,E.M. e R.L. DIEHL 1984."Identifying vowels in CVC syllables:Effects of inserting silence and noise", *Perc.Psych.* 36:369-380

PATTERSON,R.D. 1976."Auditory filter shapes derived with noise stimuli", *JASA* 59:640-654

PAULI,R.V., F.C. MAUAD e H.P. HEILMANN 1980. *Física:Ondas,Acústica,ótica*, E.P.U.,SP

PETERSEN,R. 1978."Intrinsic fundamental frequency of Danish vowels", *J.Phon.* 6:177-189

PETERSON, G.E. e H.L. BARNEY 1952."Control methods used in a study of the vowels", *JASA* 24:175-184

PETERSON,G. 1961."Parameters of vowel quality", *JSHR* 4:10-29

PETERSON,G. e I. LEHISTE 1960."Duration of syllable nuclei in English", *JASA* 32:693-703

PISONI,D.B. 1973."Auditory and phonetic memory codes in the discrimination of consonants and vowels", *Perc.Psych.* 13:253-260

PLEASANTS,H. 1955.*The Agony of Modern Music*, Simon & Schuster,NY

PLOMP,R. 1975."Auditory analysis and timbre perception", in Fant e Tatham(eds) 1975:7-22

POLLACK,I. 1952."The information of elementary multidimensional auditory displays", *JASA* 26:155-158

POLS,L.C. 1975."Analysis and synthesis of speech using a broad-band spectral representation", in Fant e Tatham(eds) 1975:23-36

POLS,L.C., W. van der KAMP e R. PLOMP 1969."Perceptual and physical space of vowels sounds", *JASA* 46:458-467

PORT,R.F. 1981."Linguistic timing factors in combination", *JASA* 69:262-274

POTTER, R.K. e J.C. STEINBERG 1950."Toward the specification of speech", *JASA* 22:807-820

RAKERD,B. e R.R. VERBRUGGE 1987."Evidence that the dynamic information for vowels is talker independent in form", *J.Mem.Lg.* 26:558-563

RAKERD,B., R.R VERBRUGGE e D.P. SHANKWEILER 1984."Monitoring for vowels in isolation and in a consonantal context", *JASA* 76:27-31

REMEZ,R.E., P.E. RUBIN, D.B. PISONI e T.O. CARRELL 1981."Speech perception without traditional speech cues", *Science* 212:947-950

REPP,B. 1984."Categorical perception:Issues,methods,findings",in Lass(ed) 1984:243-335

RIORDAN,C.J. 1980."Larynx height during English stop consonants", *J.Phon.* 8:353-360

ROSSI,M. e D. AUTESERRE 1981."Movements of the hyoid and the larynx and the intrinsic frequency of vowels", *Phonetica* 9:233-249

RYALLS,J. e P. LIEBERMAN 1982."Fundamental frequency and vowel perception", *JASA* 72:1631-1634

SACHS.M.B. e E.D. Young 1979."Encoding of steady-state vowels in the auditory nerve:Representation in terms of discharge rate", *JASA* 66:470-479

SAPIR,E. 1925."Os padrões sônicos na linguagem", trad. de Mattoso Câmara Jr. em *Lingüística Como Ciência*,Liv. Acadêmica,RJ, 1969:79-99

SATO,S., M. YOKOTA e H. KASUYA 1982."Statistical relationships among the first three formant frequencies in vowel segments in continous speech", *Phonetica* 39:36-46

SAVIN,H.B. e T.G.BEVER 1970."The nonperceptual reality of the phoneme",*Journal of Verbal Learning and Verbal Behavior* 9:295-302

SCRIPTURE,E.W. 1902.*The Elements of Experimental Phonetics*, Edward Arnold, London,1973

SHANKWEILER,D.P. e M.STUDDERT-KENNEDY 1967."Identification of consonants and vowels presented to the left and right ears", *Quarterly Journal of Experimental Psychology* 19:59-63

SHIGENAGA,M., Y. SEKIGUCHI, T. YAGISAWA e K. KATO 1986."A Speech recognition system for continously spoken Japanese sentences-SPEECH YAMANASHI", *Trans. IEEE*-Japan 69:675-683

SHIPMAN,D.S e V.W. ZUE 1982."Properties of large lexicons:Implications for advanced isolated word recognition systems", *Proc. ICASSP*:546-549

SHOUP,J.E. e L.L. PFEIFER 1976."Acoustic characgteristics of speech sounds", in Lass(ed) 1976:172-224

SILVERMAN,K.E.A. 1986."F0 segmental cues depend on intonation:The case of the rise after voiced stops", *Phonetica* 43:76-91

SLAWSON,A.W. 1968."Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency", *JASA* 43:87-101

STEELE,S. 1986."Interaction of vowel F0 and prosody", *Phonetica* 43:92-105

STEVENS, K.N. e A.S. HOUSE 1963."Perturbation of vowel articulations by consonantal context:An acoustical study", *JSHR* 6:111-128

STEVENS,K.N. 1972."The quantal nature of speech:Evidence from articulatory-acoustic data", in David e Denes(eds) 1972:51-66

STEVENS,K.N. e A.S. HOUSE 1961."An acoustical theory of vowel production and some of its implications", in Fry(ed) 1976:52-74

STOCK,O., R.FALCONE e P. INSINNAMO 1989."Bidirectional charts:A potential technique for parsing spoken natural language sentences", *Computer,Speech and Language* 3:219-237

STRANGE,W. 1987."Information for vowels in formant transitions", *J.Mem.Lg.* 26:550-557

STRANGE,W. 1989."Evolving theories of vowel perception", *JASA* 85:2081-2087

STRANGE,W., J.J. JENKINS e T.L. JOHNSON 1983."Dynamic specification of coarticulated vowels", *JASA* 74:695-705

STRANGE,W., R.R. VERBRUGGE, D.P. SHANKWEILER e T.R. EDMAN 1976."Consonantal environment specifies vowel identity", *JASA* 60:213-224

STUDDERT-KENNEDY,M. 1976."Speech Perception",in Lass(ed) 1976:243-293

STUDDERT-KENNEDY,M. , A.M. LIBERMAN, K.S. HARRIS e F.S. COOPER 1970."Motor theory of speech perception:A reply to Lane's critical review", *Psychological Review* 77:234-249

SUMMERFIELD,A.Q. e M.P. HAGGARD 1975."Vocal tract normalisation as demonstrated by reaction times", in Fant e Tatham(eds) 1975:115-141

SUNDBERG,J. 1970."Formant structure and articulation of spoken and sung vowels", *Folia Phoniatrica* 22:28-48

SUNDBERG,J., C. JOHANSSON, H. WILBRAND e C. YTTERBERG 1987."From sagittal distance to area:A study of transverse, vocal tract cross-sectional area", *Phonetica* 44:76-90

SUNDBERG,J. 1974."Articulatory interpretation of the 'singing formant'", *JASA* 55:838-844

SUTHERLAND,N.S.(ed) 1977.*Tutorial Essays in Psychology*, Hillsdale,NJ:Erlbaum Assoc.

SYRDAL,A. e H. GOPAL 1986."A perceptual model of vowel recognition based on the auditory representation of American English vowels", *JASA* 79:1086-1100

SYRDAL,A. e S. STEELE 1985."Vowel F1 as a function of speaker fundamental frequency", *JASA* sup. 1,78:S56

TERNSTRÖM,S., J. SUNDBERG e A. COLLDÉN 1988."Articulatory F0 perturbations and auditory feedback", *JSHR* 31:187-192

THORSEN,N.G. 1980."A study of the perception of sentence intonation:Evidence from Danish", *JASA* 63:1014-1030

TRAUNMÜLLER,H. 1981."Perceptual dimension of openess in vowels", *JASA* 79:1086-1100

TRAUNMÜLLER,H. 1988."Paralinguistic variation and invariance in the characteristic frequencies of vowels", *Phonetica* 45:1-29

UMEDA,N. 1981."Influence of segmental factors on fundamental frequency in fluent speech", *JASA* 70:350-355

VAISSIERE,J. 1988."Prediction of velum movement from phonological specifications", *Phonetica* 45:122-139

VERBRUGGE,R.R., W. STRANGE, D.P. SHANKWEILER e T.R. EDMAN 1976."What information enables a listener to map a talker's vowel space?", *JASA* 60:198-212

VERNOOIJ,G-J., G. BLOOTHOOFT e Y. van HOLSTEIJN 1989."A simulation study on the usefulness of broad phonetic classification in automatic speech recognition", *Proc. IEEE*:85-88

WABER,D.P. 1980."Maturation:Thoughts on renewing an old acquaitanceship", in Caplan(ed) 1980:8-26

WALLEY,A. e T.CARRELL 1983."Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants", *JASA* 73:1011-1022

WATHEN-DUNN,W. 1967.*Models for the Perception of Speech and Visual Form*, MIT press

WOOD,C.C. 1974."Parallel processing of auditory and phonetic information in speech perception", *Perc.Psych.* 15:501-508

WOOD,C.C. 1975."Auditory and phonetic levels of processing in speech perception:Neurophysiological and information processing analyses", *Journal of Experimental Psychology:Human perception and Performance* 1:3-20

WOOD,C.C. e R.S.DAY 1975."Failure of selective attention to phonetic segments in consonant-vowel syllables",*Perc.Psych.* 17:346-350

WRIGHT,J.T. 1986."The behavior of nasalized vowels in the perceptual vowel space", in Ohala(ed) 1986:45-67

ZUE,V.W. e L.F. LAMEL 1986."An expert spectrogram reader:A knowledge-based approach to speech recognition", *Proc. ICASSP*, Tokyo,1197-1200

ZWICKER,E. 1961."Subdivision of the audible frequency range into critical bands (Frequenzgruppen)", *JASA* 33:248

ZWICKER,E. e E. TERHARDT 1980."Analytical expressions for critical band rate and critical bandwidth as a function of frequency", *JASA* 68:1523-1525

ZWICKER,E., E. TERHARDT e E. PAULUS 1979."Automatic speech recognition using psychoacoustic models", *JASA* 65:487-498

ZWITSERLOOD,P. 1989."The locus of the effects of sentential-semantic context in spoken-word processing", *Cognition 32:25-64*