IDENTIFICAÇÃO DE FALANTES: ASPECTOS TEÓRICOS E METODOLÓGICOS

RICARDO MOLINA DE FIGUEIREDO

Tese apresentada ao Departamento de Lingüística do Instituto de Estudos da Linguagem da Universidade Estadual de Campinas, como requisito parcial para a obtenção do grau de Doutor em Ciências.

Orientador: Profª Drª Eleonora Albano

CAMPINAS - 1994

Este exemplar é a redação final da tese defendida por Ricardo Molina de Figueiredo e aprovada pela Comissão Julgadora em 30/03/94.

Profª Drª Eleonora Cavalcante Albano

T/UNICAMP

F469i

Figueiredo, Ricardo Molina de
Identificação de Falantes:
Aspectos Teóricos e Metodológicos/
Ricardo Molina de Figueiredo -
Campinas (SP : s.n.), 1994

Orientador: Profª Drª Eleonora Albano
Tese (Doutorado), Instituto de Estudos
da Linguagem.

1. Lingüística - Fonética Acústica -
Identificação de Falantes

Ficha catalográfica elaborada pela Biblioteca do IEL-UNICAMP

BANCA EXAMINADORA

AGRADECIMENTOS

à Eleonora, minha orientadora, pelas sempre pertinentes observações.

ao Fortunato, por ter vislumbrado a importância dessa linha de pesquisa.

à Helena pela revisão, sugestões e apoio moral.

ao Arnoldo, pelas informações estatísticas.

ao Nagle, pela sua habilidade em apontar detalhes que me passaram desapercebidos.

ao Sílvio, pelas sugestões e o imprescindível auxílio na impressão.

à Beth, à Lúcia e à Ilda, pelo apoio logístico.

ao pessoal do Laboratório de Fonética do IEL, Adelaide, Agnaldo, Beth, Márcio, Mário e Patrícia, por terem permitido que eu utilizasse com tanta freqüência o disputado horário do DSP-5500.

aos participantes dos experimentos, Zaldo, Edson, Agnaldo, José Pedro, Walter, Márcio, Domenico, e os gêmeos José Alexandre e José Ricardo, pela paciência em se submeter a tarefas tão monótonas.

RESUMO

O presente trabalho pretende examinar a eficiência de diversos parâmetros acústicos na Identificação de Falantes. Nos experimentos analisou-se um conjunto básico de 8 falantes, adultos do sexo masculino, com idades entre 22 e 45 anos. Em alguns casos incluiu-se a análise de mais dois falantes, gêmeos idênticos, de modo a examinar instrumentalmente as diferenças entre vozes perceptualmente muito semelhantes. Os parâmetros estudados foram: Formantes Vocálicos, Freqüência Fundamental, Espectro de Longo Termo, Velocidade de Fala, Consoantes Nasais e VOT (*Voice Onset Time*). Discutiu-se também a eficiência da inspeção visual de espectrogramas na Identificação de Falantes, um tema especialmente relevante para o modelo forense, e que tem provocado grande controvérsia nas últimas décadas.

Índice

PREFÁCIO

A pesquisa na área de Identificação de Falantes tem progressivamente se firmado como um ramo legítimo da Fonética. Tem ficado cada vez mais claro que considerar as diferenças entre falantes como um ruído que deve ser abstraído dos dados, de modo a acessar o código lingüístico, é uma abordagem por demais simplificadora; para compreender adequadamente os processos de percepção e produção, é preciso avaliar conjuntamente os dois tipos de informação (falante e código lingüístico). Assim, tanto o foneticista tradicional quanto o pesquisador voltado para a Identificação de Falantes, deverão dar conta de duas questões básicas, mas que se colocam de forma um pouco diferente para cada um; o primeiro concentra sua atenção nos seguintes problemas: (1) como o ouvinte normaliza diferentes sinais acústicos produzidos por diferentes falantes, percebendo-os como a mesma unidade lingüística e (2) como o mesmo indivíduo pode usar diferentes estratégias articulatórias para produzir saídas acústicas equivalentes; já ao segundo interessa saber: (1) como diferentes ouvintes percebem diferentes sinais acústicos como sendo produzidos pelo mesmo falante e (2) como diferentes indivíduos, embora empregando estratégias articulatórias equivalentes, podem produzir saídas acústicas diferentes. As questões básicas, obviamente, se completam; tomando a liberdade de parodiar Jakobson (1976:71), poderíamos dizer que o *falante* (assim como o fonema) também é a "invariante nas variações".

Independentemente do objeto específico de estudo (falante ou código lingüístico), o pesquisador estará sempre envolvido com a avaliação de variações **dentro** de variações. Se o interesse maior é o código lingüístico, o foco de atenção

será nos aspectos acústicos que possuem alta variabilidade intra-falante, enquanto para o estudo dos determinantes de identidade interessam os traços onde a variância inter-falante é maior do que a variância intra-falante. Apesar da diferença de enfoque, os resultados experimentais serão mutuamente complementares, na medida em que, em qualquer caso, será ampliado o conhecimento do âmbito de variação do parâmetro estudado. Nesse sentido, acreditamos que os resultados dos experimentos realizados ao longo deste trabalho extrapolem o campo da Identificação de Falantes, e possam ser úteis para outras linhas de pesquisa.

O trabalho está organizado da seguinte forma. As duas primeiras seções tratam de aspectos mais gerais relacionados à Identificação de Falantes. A seção 1 tenta situar o tema dentro do panorama dos estudos lingüísticos, discutindo os diferentes paradigmas (Identificação *vs.* Verificação) e configurando com mais detalhes a situação forense, um campo de aplicação cujas especificidades impõem dificuldades ainda maiores ao pesquisador.

A seção 2 examina as diversas condições experimentais que podem alterar, de algum modo, o desempenho de ouvintes no reconhecimento ou discriminação de falantes. Embora a ênfase nessa seção tenha sido a dimensão perceptual, a discussão serve de base para a definição de diferentes paradigmas experimentais, que também poderiam ser empregados na avaliação de procedimentos automáticos.

Na seção 3 são descritos o *corpus* utilizado nos experimentos e a metodologia básica de gravação e análise. As particularidades de cada experimento são descritas em cada seção específica, abordando cada parâmetro acústico isoladamente (seções 4 a 9).

Nas seções 4 a 9 encontra-se o corpo principal do trabalho, onde é apresentada uma série de experimentos visando avaliar a eficiência de diferentes parâmetros acústicos na Identificação de falantes. São estudados aspectos

relacionados aos Formantes Vocálicos, Freqüência Fundamental, Espectro de Longo Termo, Velocidade de Fala, Consoantes Nasais e VOT (*Voice Onset Time*). Cada um desses fatores acústicos é discutido à luz dos resultados encontrados na literatura especializada, em contraponto com nossos próprios resultados. De modo a inserir um fator de variabilidade intra-falante, estudamos, para a maior parte dos parâmetros acústicos, a influência da velocidade de emissão no comportamento do parâmetro. Sempre que possível incluiu-se também uma análise comparativa de dois gêmeos idênticos. Diferentes estratégias estatísticas foram empregadas para avaliar os dados, de acordo com a natureza de cada parâmetro. É importante ressaltar que nosso objetivo não é exatamente desenvolver técnicas automáticas de Identificação, mas antes estabelecer um referencial para que se possa avaliar o potencial de cada parâmetro.

A seção 10 trata da possibilidade de identificar falantes através da inspeção visual de espectrogramas, uma discussão que tem gerado bastante polêmica, já há algum tempo. No nosso entender, existem diversos mal-entendidos cercando a questão: nem o espectrograma pode ser comparado a uma "impressão digital" (como querem seus defensores), nem o exame espectrográfico carece de qualquer objetividade (como afirmam alguns de seus oponentes). Mostraremos, na seção 10, exemplos extraídos de alguns casos forenses reais, onde fica claro que a análise espectrográfica convencional pode prestar grande auxílio, na medida em que concentra uma grande quantidade de informação.

A falta de pesquisa sistemática na área, entre nós, nos fez optar por uma perspectiva mais abrangente, cujo principal objetivo é abrir um campo de possibilidades dentro da pesquisa experimental voltada à Identificação de Falantes. Cada um dos parâmetros acústicos aqui estudados mereceria, certamente, um tratamento mais detalhado; esperamos, no entanto, que as informações aqui

obtidas, embora não esgotem cada aspecto específico, possam servir de base para futuras pesquisas.

SEÇÃO 1: ASPECTOS GERAIS

1.1) *Introdução*

A capacidade humana de reconhecer indivíduos apenas pela voz manifesta-se cotidianamente nas mais variadas situações, inclusive sob condições pouco favoráveis, envolvendo diferentes tipos de distorções e limitações do sinal acústico, tais como transmissões telefônicas, presença de ruído intenso, simultaneidade de vozes, etc. Embora faça parte do senso comum reconhecer que cada indivíduo possui uma "voz", ou um "jeito de falar" que lhe são característicos, a questão foi largamente negligenciada pelas correntes lingüísticas dominantes, especialmente as de inspiração Saussureana.

A perspectiva imposta por Saussure, reduzindo o individual às contingências de ordem social, permeou grande parte dos estudos lingüísticos no nosso século, com raras exceções. Um dos primeiros a perceber a importância dos aspectos individuais foi o lingüista Edward Sapir; reconhecendo a ordem determinada pela dimensão social, mas, ao mesmo tempo, estabelecendo um contraponto com a expressão da individualidade, Sapir (1927:65) afirma:

> *"a sociedade tem os seus padrões, as suas maneiras pré-estabelecidas de proceder (...) ao passo que o indivíduo tem o seu próprio método de se servir desses padrões especiais da sociedade, ajeitando-os a seu modo para fazê-los propriedade particular, sua e não de outrem".*

Sapir observa que os traços individuais podem se manifestar em diversos "níveis", tais como entonação, ritmo, velocidade, qualidade de voz, vocabulário, etc., prevenindo, contudo, que, até então, não se havia dado um *status* científico à questão (uma observação que, de uma certa forma, permanece válida até hoje).

A questão da individualidade foi tratada também por Firth. Criticando a excessiva influência de Saussure, Firth (1950) coloca o estudo dos aspectos individuais no mesmo nível que os aspectos sociais da linguagem:

> *"we may assume that any social person speaking in his own personality will behave systematically, since experienced language is universally systemic (...) we must not expect to find one closed system (...) but we may apply systematic categories to the statement of the facts (...) stating them by the spectrum of linguistic techniques (...) the study of one person at a time seem to me amply justified as a scientific method"* (Firth 1950:187; grifo do autor).

Dentro de perspectiva semelhante, encontramos as idéias de Hockett. Importante nesse sentido é a noção de *idioleto*, proposta pelo autor:

> *"the totality of speech habits of a single person at a given time constitutes an idiolect"* (Hockett 1958:321).

De um certo modo, Hockett inverte os pressupostos Saussureanos, colocando o idioleto como base do sistema:

> *"...in the last analysis a language is observable only as a collection of idiolects (...) speaking is not collective behavior"* (Hockett 1958:321-2).

Hockett vislumbra duas abordagens possíveis: a *Lingüística Descritiva*, que ignora diferenças inter-pessoais e inter-grupos, e a *Lingüística Sincrônica*, que incluiria a primeira, além de outros tipos de investigação, particularmente a *Dialetologia Sincrônica*, esta focalizando as diferenças sistemáticas entre indivíduos e grupos. Em trabalho mais recente, Hockett aborda a mesma questão, por outro ângulo:

> *"Some investigators ally natural language with logic and mathematics, and place the foundations of all three in an ideal world of pure logic. Others see language as a feature of everyday human conduct, and believe that the foundations of our discipline must be empirical in the same way as are those of biology and physics (...) I simply announce that [my view] is the second"* (Hockett 1987:1).

A dicotomia indivíduo/grupo, antiga na Lingüística, está na base das discussões a respeito da possibilidade de identificar falantes. Garvin e Ladefoged (1963), em uma das primeiras tentativas de estabelecer um quadro de referência descritivo, cruzam a distinção indivíduo/grupo com outra, não menos polêmica, que é a oposição orgânico *versus* adquirido (ou aprendido). Como características orgânicas no nível individual, Garvin e Ladefoged incluem aqueles aspectos diretamente relacionados à estrutura dos órgãos vocais de um determinado falante, especialmente a laringe e as cavidades nasais; fatores orgânicos também poderiam - segundo os autores - estar associados a características de grupos, ou classes de indivíduos, definidos em termos de raça, idade, sexo, etc. Características de grupo

adquiridas estariam relacionadas a condições sócio-culturais e regionais, enquanto os traços adquiridos idiossincráticos podem ser classificados em duas categorias:

> *"(1) individual variation within a particular single group pattern; (2) idiosyncratic speech patterns due to the use of a mixture of social and/or regional varieties of speech by a given individual"* (Garvin e Ladefoged 1963:195).

A dicotomia orgânico/adquirido tem permeado a maioria das discussões sobre Identificação de Falantes, admitindo-se que essa divisão espelha bem as origens das diferenças inter-falante (v. p. ex.: Glenn e Kleiner 1968; Wolf 1972; Atal 1976; Bricker e Pruzansky 1976). Em geral, postula-se que as pistas mais confiáveis são aquelas dependentes de características orgânicas invariantes, em oposição aos aspectos relacionados a padrões aprendidos. Não é simples, entretanto, estabelecer uma relação direta entre as características anatômicas individuais e as particularidades da saída acústica, na medida em que, ao se comparar diferentes falantes, os possíveis efeitos diretos da anatomia individual estarão parcialmente ocultos por outras fontes de variação inter- e intra-falante (sócio-cultural, afetiva, etc). Além disso, é preciso considerar que, embora as características orgânicas do falante limitem a variação em uma determinada dimensão, essa informação mescla-se necessariamente com a informação lingüística que eventualmente explore essa mesma dimensão.

Por outro lado, é indiscutível que a conformação fisiológica do aparelho vocal restringe o âmbito de variação de um parâmetro particular. Assim, por exemplo, a extensão do trato vocal determina, em uma certa medida, o valor médio dos

formantes (especialmente os formantes altos), a massa e comprimento das cordas vocais determinam a faixa de variação de F0 para um determinado falante, etc.

Embora a plasticidade do trato permita, a princípio, uma ampla gama de variação de um parâmetro acústico qualquer, em situações "normais" (sem a presença de disfarce, imitação, fala gritada, condições físicas anômalas, etc) a capacidade total do sistema de produção não é esgotada, ficando a variação naturalmente limitada ao que poderíamos chamar de faixa "ótima" de utilização para o falante particular. As fontes lingüísticas de variação que atuam dentro dessa faixa otimizada não mascaram totalmente a informação do falante, caso contrário não seríamos capazes de, perceptualmente, reconhecer ou discriminar diferentes falantes. Além disso é preciso considerar que a variação de um determinado parâmetro acústico será preferencialmente avaliada a partir de uma prévia sub-categorização de ordem lingüística, limitando o campo de variação, por exemplo, à realização de um certo fonema em um certo contexto fonético, etc.

Uma determinada característica fisiológica deixará uma marca tão mais invariante no sinal acústico quanto menores forem as possibilidades de suas propriedades intrínsecas serem modificadas por ações articulatórias. A configuração das cavidades nasais, por exemplo, está relacionada a aspectos quase invariantes do espectro de sons nasais (gripes e resfriados excluídos); um certo grau de afastamento entre os incisivos superiores pode criar um pico de energia característico no espectro de fricativas anteriores, etc. Por outro lado, um aspecto não diretamente orgânico, resultado de um padrão adquirido, não será, necessariamente, menos estável. Muitas ações articulatórias estão de tal forma cristalizadas que se tornam transparentes para o falante, permanecendo invariantes, mesmo com mudança de registro (informal para formal, por exemplo).

Ao longo do presente trabalho faremos eventualmente referência à oposição orgânico/adquirido, entendendo-a, entretanto, não como uma dicotomia, mas antes como a expressão de dimensões complementares, cuja distinção só pode ser estabelecida em termos graduais. O mais importante, nos parece, é não perder de vista a complexidade subjacente aos traços de fala dependentes de falante, ou seja, não perder de vista a própria Linguagem. Na verdade, a eficiência da maioria dos parâmetros acústicos potencialmente úteis para a Identificação de Falantes está parcialmente condicionada a fatores específicos da língua e limitada por exigências de ordem fonológica. A variação intra-individual dos formantes de uma vogal, por exemplo, está naturalmente limitada pela estrutura do sistema vocálico da língua nativa do falante; se a vogal em questão situa-se em uma região mais densa do sistema, haverá menos "espaço" para a variação intra-subjetiva, sob o risco de essa variação ultrapassar uma fronteira categorial. Um exemplo típico desse fenômeno ocorre em Irlandês, com relação à assimilação entre segmentos adjacentes. A existência em Irlandês de, pelo menos, 3 fonemas laterais contrastantes minimiza efeitos de coarticulação desses sons com vogais adjacentes, ao contrário do Inglês que, não possuindo a mesma diversidade fonêmica em laterais, permite uma variação intra-falante considerável (v. Nolan 1983:116). O mesmo pode ocorrer com respeito à nasalização; em Chinantec há contrastes entre vogais orais, levemente nasalizadas e fortemente nasalizadas (Ladefoged 1971:34), o que, provavelmente, restringe a faixa de variação intra-falante dos graus de nasalidade nesse segmentos vocálicos.

1.2) *O Modelo Forense*

A Identificação de Falantes encaixa-se no quadro geral que engloba os problemas de reconhecimento de padrão e pode ser considerada um exemplo de identificação pessoal biométrica. Esse termo serve para diferenciar técnicas que baseiam a identificação em certas características intrínsecas do indivíduo; nessa categoria estariam também incluídas outras técnicas tais como: impressões digitais, padrões de íris e retina, estrutura genética, etc. Uma diferença importante em relação a essas técnicas precisa, no entanto, ser estabelecida. O sinal de fala deve ser entendido como uma função complexa que envolve não apenas aspectos anatômicos, como também fatores sócio-culturais e ambientais; o sinal acústico gerado pelo falante não fornece diretamente informação anatômica detalhada - pelo menos de uma forma explícita. Isso distingue a Identificação de Falantes da Identificação de Impressões Digitais, já que esta se vale de características físicas estáticas, enquanto a primeira (assim como a Grafotécnica) está mais fortemente relacionada a traços dinâmicos de performance, que dependem de uma **ação** no tempo.

Existem limitações inerentes à natureza do sinal de fala e sua relação com o falante. Para avaliar esses limites é preciso compreender de que modo a informação específica do falante está codificada no sinal de fala. O sinal de fala é uma conseqüência direta dos mecanismos articulatórios, os quais são determinados pelo aparelho vocal e controle neurológico. Assim, há duas fontes possíveis de informação de falante: as características físicas e estruturais do trato vocal e o controle neuro-sensorial do sistema cérebro/articuladores. Essa informação inerente ao falante é veiculada no sinal de fala juntamente com outras informações, incluindo-se aí não só a mensagem lingüística como também o estado emocional, o estado de saúde, sexo do falante, idade, peso, altura, etc (v. Laver e Trudgill 1979).

As características do sinal de fala são primariamente determinadas pela mensagem lingüística. Os fatores inerentes ao falante podem ser entendidos como pertencendo às mensagens secundárias (para- ou extra-lingüísticas) e estão codificados como variações não-lingüísticas da mensagem lingüística básica. Assim, a informação útil para identificação do falante veicula-se indiretamente no sinal de fala, como um efeito colateral do processo articulatório; de uma certa forma, a informação inerente ao falante pode ser vista como um "ruído" aplicado sobre a mensagem lingüística básica. A principal dificuldade na Identificação de Falantes relaciona-se, pois, ao fato de não existir traços de fala (ou transformações de traços) dedicados exclusivamente a veicular informação discriminadora de falante.

No entanto, o fato é que diferentes indivíduos tipicamente apresentam características no sinal de fala que são bastante particulares. A experiência pessoal de cada um demonstra a grande habilidade humana em reconhecer pessoas pela voz, mesmo em situações bastante adversas (baixa razão sinal/ruído, limitação de banda, etc); o grande desafio que se coloca para o cientista da fala é estabelecer um modelo que reproduza essa habilidade (sem que, no entanto, precise necessariamente simular os mesmos processos humanos). Esse desafio tem motivado, nas últimas décadas, um grande número de estudos na área da identificação de falantes, uma área que recebeu um extraordinário impulso com o desenvolvimento de sistemas de processamento digital de sinal. Com a multiplicidade de estudos e de enfoques, a questão genérica do reconhecimento de falantes adquiriu nuances e dividiu-se em aplicações e sub-problemas específicos, entre eles o modelo forense. Discutiremos abaixo alguns paradigmas relacionados ao reconhecimento de falantes, tentando situar mais adequadamente o modelo forense.

A proliferação de estudos na área produziu uma concomitante expansão terminológica. Assim, fala-se de Identificação, Verificação, Discriminação, Autenticação de voz ou de falante (*speaker* ou *talker*), sem que, necessariamente,

haja referência a tarefas especificamente diferentes. Em geral aceita-se que o termo genérico **Reconhecimento de Falantes** englobe todos os processos de decisão (humanos ou automatizados) que utilizam traços do sinal de fala para determinar se uma pessoa é o falante de um dado enunciado (Atal 1976). No presente trabalho utilizaremos com mais freqüência o termo Identificação de Falante, já consagrado na literatura, fazendo, entretanto, a ressalva de que essa denominação não é inteiramente adequada ao modelo forense (v. discussão abaixo).

Dois paradigmas têm sido mais claramente delimitados na literatura: a Identificação e a Verificação de Falantes. Embora freqüentemente se pense nessa classificação apenas como a expressão de aplicações diferentes, cada um dos termos está relacionado a problemas e pressupostos específicos. A Identificação de Falantes consiste em atribuir um enunciado produzido por um falante desconhecido a um indivíduo dentro de um grupo de N falantes; o processo de decisão tem, portanto, N saídas possíveis se o grupo é fechado (i.e., se sabemos que o falante desconhecido pertence ao grupo) e N+1 saídas se o grupo é aberto (se o falante desconhecido não pertence obrigatoriamente ao grupo resta mais uma alternativa: o enunciado não se associa a nenhum dos falantes do grupo). Na situação clássica de Verificação, o problema consiste em examinar se o enunciado produzido pelo falante desconhecido foi produzido por um, e apenas um, determinado falante; o processo de decisão reduz-se assim a uma escolha binária, com apenas duas saídas possíveis.

Doddington (1985) tratou formalmente a distinção entre Identificação e Verificação, realizando uma simulação a partir de n medidas hipotéticas com distribuição normal. A taxa esperada de erro (falsa aceitação ou falsa rejeição) foi calculada em função do tamanho da população e do número de medidas compondo cada vetor. Doddington observa que o tamanho da população é um fator crítico para o paradigma da Identificação, com a probabilidade de decisão errada tendendo a 1 para populações indefinidamente grandes, independentemente da dimensionalidade

do vetor de traços utilizado; esse resultado é esperado, já que a Identificação supõe comparações com enunciados-referência de todos os falantes do grupo, fazendo com que o índice de erros cresça rapidamente em função do tamanho da população. O resultado mais interessante relatado em Doddington (1985) é que o desempenho da Verificação de Falantes permanece satisfatório mesmo para populações grandes, com a probabilidade de decisão errada tendendo assintoticamente a um valor limite.

Aparentemente a dificuldade em resolver problemas de Verificação é menor. Isso, entretanto, não é totalmente verdadeiro. Há uma dificuldade com a Verificação que não está presente na Identificação para grupos fechados; a Verificação depende muito mais de uma compreensão abrangente da variabilidade dos traços de fala empregados na discriminação. Enquanto na Identificação para grupos fechados basta determinar qual referência é **mais próxima** da amostra desconhecida, na Verificação há o problema mais genérico de julgar qual é **suficientemente próxima**. No primeiro caso, os limites de variabilidade estão pré-definidos pela própria população, enquanto no segundo caso, a decisão deve ser tomada em função de limiares absolutos pré-estabelecidos.

A situação mais difícil é a Identificação para grupos abertos; nesse caso, a alternativa adicional de rejeição (o enunciado desconhecido pode não estar associado a um membro do grupo) exige - assim como na Verificação - uma caracterização estatística segura dos parâmetros de fala.

Embora o modelo forense esteja freqüentemente associado ao paradigma de Identificação, o mais correto parece ser associar a situação forense típica com a Verificação, como, acertadamente, sugere Doddington (1985) - mais especificamente, diríamos, com a Verificação para grupos indefinidamente grandes. Na maior parte dos casos forenses o que se espera do perito é uma decisão binária: a voz questionada **é** ou **não é** a voz do suspeito. Existindo mais de um suspeito, caracteriza-se um múltiplo problema de Verificação. Há apenas uma situação

forense onde se pode falar propriamente de Identificação: quando existe um conjunto de suspeitos e há a certeza de que a voz questionada pertence a um desses suspeitos; nesse caso recaímos na Identificação para grupos fechados.

Um complicador adicional para a situação forense é a possibilidade de os falantes não serem cooperativos. Esse aspecto distingue o problema forense das aplicações mais freqüentes da Verificação Automática; em sistemas de acesso restrito, o falante **quer** ser reconhecido. Dificilmente esperaríamos o mesmo de um suspeito (culpado ou não!). Assim, enquanto o impostor no paradigma da Verificação Automática é potencialmente um imitador, a expectativa no modelo forense é que nos defrontemos eventualmente com o problema do disfarce.

A questão colocada pelo disfarce é muito mais complexa do que a da imitação. Enquanto o imitador deve realizar a tarefa de aproximar-se de um padrão de referência particular, aquele que disfarça a voz precisa apenas afastar-se aleatoriamente de seu próprio padrão habitual. Vários estudos têm demonstrado que o imitador dificilmente é bem sucedido. Hall e Tosi (1975) relatam que ouvintes não treinados diferenciam auditivamente o imitador em 75% dos casos. Sistemas automatizados de verificação de falante também são relativamente resistentes à imitadores (Luck 1969; Hair e Rekieta 1972), embora já se tenha observado que imitadores profissionais consigam uma taxa de aceitação ligeiramente maior do que impostores casuais nesses sistemas (Lummis e Rosenberg 1972). Mesmo que sejam capazes de modificar o padrão de formantes e o F0 médio, imitadores profissionais não conseguem, em geral, atingir os valores da voz alvo (Endres et al. 1971; Hall e Tosi 1975). Na verdade, mesmo as tentativas de simulação de um determinado sotaque não passam desapercebidas para a maioria dos falantes nativos do dialeto imitado (Tate 1979).

O disfarce, por outro lado, pode dificultar consideravelmente a identificação. Já se verificou que a identificação através da audição e/ou leitura espectrográfica cai

consideravelmente se uma das vozes está disfarçada (Reich e Duke 1979; Reich et al. 1976; Hollien et al. 1982). O padrão de formantes e o F0 médio podem ser sensivelmente alterados pelo disfarce (Endres et al. 1971), assim como o espectro de longo termo (Doherty 1975; Hollien e Majewski 1977). Alguns parâmetros temporais, entretanto, parecem ser relativamente resistentes ao disfarce (Johnson et al. 1984).

Um atenuante para a situação forense é o fato de o disfarce ser bastante saliente perceptualmente; Reich e Duke (1979) relatam que mesmo ouvintes destreinados conseguem reconhecer tentativas de disfarce em cerca de 90 % dos casos (v. também Reich 1981). O fato é que a voz disfarçada soa quase sempre pouco natural para um ouvinte atento; além disso é extremamente difícil manter um padrão de disfarce coerente ao longo de um longo trecho de fala, o que exigiria um controle muito exato dos mecanismos articulatórios, uma tarefa que apenas poucos artistas profissionais e foneticistas experimentados conseguiriam realizar com êxito. Na verdade, disfarce e imitação são raramente observados em casos forenses reais (Cf. Ladefoged 1984), já que, na maior parte dos casos, durante a gravação questionada, o envolvido não tem consciência de que está sendo gravado. Tentativas de disfarce durante a gravação para a coleta do padrão de confronto também são improváveis, visto que qualquer alteração pareceria saliente àqueles que acompanharam o suspeito desde o início do processo (agentes policiais, advogados, etc).

Uma dificuldade mais realista do que o disfarce consciente é a possibilidade de haver uma mudança de estilo de fala não intencional durante a gravação da amostra de confronto. O contexto situacional nesses casos provoca geralmente algum constrangimento no suspeito, com eventuais alterações em alguns parâmetros acústicos, mais freqüentemente relacionadas a aspectos suprasegmentais, tais como freqüência fundamental e velocidade de fala. Nossa experiência com casos forenses

tem demonstrado, entretanto, que é possível, com um pouco de habilidade, contornar esse tipo de dificuldade, estabelecendo um clima mais informal durante a gravação, através da colocação de temas que motivem o suspeito. O problema não é muito diferente do encontrado nas pesquisas sócio-lingüísticas, onde, em geral, uma maior informalidade é obtida do mesmo modo; Labov recomendava aos pesquisadores que evocassem na conversação temas relacionados a alguma situação de perigo na qual o informante já estivera eventualmente envolvido. Estratégias semelhantes podem ser empregadas na situação forense, quase sempre com sucesso, desde que não haja limitação de tempo para a gravação (voltaremos a discutir a questão do disfarce e da imitação na seção 10).

O número de variáveis que podem estar presentes em casos forenses é grande; fatores como qualidade da gravação, duração do material gravado, marcas particulares de um determinado falante, etc, fazem com que cada caso deva ser examinado à luz de suas próprias características. A intervenção humana aqui é fundamental, e não parece sensato vislumbrar - pelo menos para futuro próximo - sistemas de decisão automática para a aplicação forense (o que não inviabiliza, entretanto, o emprego de procedimentos estatísticos em fases intermediárias da análise).

O problema da "objetividade" das técnicas de Identificação de Falantes surge com bastante freqüência no contexto da situação forense. A questão é geralmente colocada de forma vaga: "Qual a **probabilidade** de acerto...?". Acreditamos que não é possível ainda (e, provavelmente, tampouco será um dia) definir exatamente essa "probabilidade", já que, em virtude da multiplicidade de condições, é impossível adotar métodos estandardizados para todos os casos. Qualquer foneticista, por mais experiente que seja, terá pouco a dizer sobre uma gravação de baixa qualidade com poucos segundos de duração; a situação pode ser bem diferente quando se trata de

gravações longas e/ou de boa qualidade. As dificuldades não parecem muito diferentes das encontradas em várias outras técnicas de identificação, tais como a Grafotécnica, identificação visual, ou mesmo o exame de impressões digitais, técnica esta cuja eficiência, apesar da mística de "infalibilidade" que a cerca, também depende fortemente da qualidade do material colhido, quase sempre fragmentário - o que impõe dificuldades análogas às observadas em gravações de baixa qualidade.

SEÇÃO 2: A ABORDAGEM PERCEPTUAL

2.1) *Identificação de Falantes: Três Abordagens*

A pesquisa de laboratório no campo da Identificação de Falantes já vem de cerca de seis décadas, com diferentes abordagens. Hecker (1971) reconhece 3 métodos básicos: (1) através da escuta ; (2) por meio de sistemas automatizados e (3) pela inspeção visual de espectrogramas, aos quais chamaremos de *abordagens perceptual, automática* e *espectrográfica*, respectivamente (ao longo da presente serão discutidos alguns experimentos relacionados à abordagem perceptual. As possibilidades de identificar falantes por meio de sistemas automáticos serão discutidas no corpo principal do trabalho, nas seções onde examinamos separadamente diferentes parâmetros acústicos - seções 4 a 9; a abordagem espectrográfica será discutida separadamente na seção 10).

A classificação dos tipos de abordagem proposta por Hecker (1971) tem sido largamente aceita como quadro de referência básico em vários trabalhos posteriores (v. p. ex.: Bricker e Pruzansky 1976; Atal 1976; Rosenberg 1976; Doddington 1985; Hollien 1990). Nolan, entretanto, em um dos mais abrangentes estudos sobre Identificação de Falantes, discute essa classificação tripartite, propondo uma divisão em apenas duas categorias, que chama de (1) *technical speaker recognition* e (2) *naive speaker recognition* (Nolan 1983:7). De acordo com a argumentação do autor, a diferença básica entre as abordagens deve ser colocada em termos do emprego ou não de *técnicas analíticas* para resolver o problema, independentemente do fato de serem essas técnicas adquiridas por humanos ou programadas automaticamente. Nolan afirma que a divisão entre os métodos automáticos e o exame de espectrogramas é meramente contingente, já que, para um observador treinado, é

possível realizar medidas confiáveis com base em espectrogramas, que poderiam servir como *input* para estratégias posteriores de decisão automática (ou seja, uma espécie de sistema híbrido, semi-automático). Outro aspecto ressaltado por Nolan a respeito da categorização original de Hecker (1971), é o fato de, nessa divisão, não haver lugar para o reconhecimento *técnico* de falantes por meios auditivos, ou seja:

> *the application of auditory techniques acquired through phonetic training to making decisions about the identity of speech samples* (Nolan 1983:7).

Embora os comentários de Nolan sejam pertinentes, sua divisão do problema em apenas duas categorias, dependendo apenas se

> *only normal everyday human abilities are exploited or whether specialised techniques - aural, visual, or electronic - are brought to bear* (Nolan 1983:8),

não é menos problemática do que a classificação de Hecker (1971). A delimitação que faz Nolan entre habilidades perceptuais "normais" e "especializadas" no domínio da audição não parece ser conveniente. A habilidade de ouvintes "*naive*" (como os rotula Nolan) em identificar/discriminar falantes pode variar bastante inter-subjetivamente, como veremos ao discutirmos alguns experimentos na seção 2.3; não deveríamos esperar algo muito diferente para "especialistas". O que desejamos frisar é que em termos de **habilidade auditiva** em identificar falantes não parece possível estabelecer uma descontinuidade entre ouvintes ***naive*** e especializados, independentemente de como esses últimos rotulam suas próprias habilidades. É bem conhecida a dificuldade de foneticistas treinados em produzir transcrições consistentes entre si e/ou correlacionadas com o dado físico objetivo (Cf. Lieberman

1965; Butcher 1982); no caso do reconhecimento de falantes não seria surpreendente encontrarmos opiniões divergentes de diferentes foneticistas a respeito da identidade de um determinado falante, em análises apenas auditivas (Hollien 1990 relata alguns casos forenses onde isso efetivamente ocorreu; v. também Ladefoged 1978). É evidente que um foneticista (bem) treinado pode **localizar** aspectos do enunciado potencialmente relevantes para a identificação do falante (embora essa habilidade em destacar aspectos importantes também não deva ser homogênea para qualquer grupo de foneticistas, independentemente do grau de treinamento), mas esse procedimento deve, necessariamente, ser sucedido pela análise instrumental, caso contrário cairemos no perigoso terreno das "opiniões", cuja validade poderá depender de fatores mais contingentes (atribuições oficiais ou - o que é pior - auto-proclamações de *expertising*) do que propriamente científicos. Se a análise técnico-instrumental é fundamental - e com isso Nolan parece concordar - , então não é preciso manter a divisão *naive*/especialista no nível das habilidades auditivas.

No que diz respeito ao exame de espectrogramas, a perspectiva de Nolan é mais adequada. Nesse caso temos, evidentemente, que estabelecer uma divisão mais clara entre uma abordagem técnica e uma abordagem *naive*. No segundo caso, a identificação será o mero resultado do confronto visual de padrões gráficos, enquanto o exame técnico pressupõe um certo nível de estruturação. Observe-se que há uma distinção fundamental com relação à habilidade auditiva em reconhecer falantes, já que no último caso trata-se de capacidade **naturalmente** adquirida, parte inerente de esquemas internos de representação, que já se manifestam desde muito cedo (uma criança de colo reconhece a voz da mãe muito antes de processar informação no nível lingüístico). Não seria ilícito fazer uma analogia com os princípios do gerativismo e supor uma "competência" dos ouvintes para reconhecer falantes; na verdade, na decodificação do sinal de fala, a representação do falante parece preceder necessariamente o processamento lingüístico, na medida em que o

ouvinte deve "calibrar" seu sistema perceptual, ajustando-o para o espaço vocálico particular do falante (Joos 1948; Johnson 1987; Nearey 1989).

2.2) *Abordagem Perceptual: Esboço de um Quadro Conceitual*

A multiplicidade de abordagens experimentais focalizando os aspectos perceptuais do reconhecimento de falantes exige que sejam traçadas algumas diretrizes prévias, de modo a situar adequadamente os diferentes objetivos experimentais. Bricker e Pruzansky (1976) propõem um esquema linear simples, que concebe o processo de reconhecimento de falantes como um seqüência de estágios que conduz a uma resposta. Cada fase do processo está relacionada a um tipo específico de informação do falante; em cada estágio, essa informação é transformada, antes de passar ao próximo estágio. A natureza puramente linear do esquema é discutível, mas, para os nossos objetivos (mapear os diversos procedimentos experimentais), essa questão é secundária. A figura 2.1 apresenta o esquema de Bricker e Pruzansky (1976:298); no primeiro nível (dentro dos retângulos) estão representadas as fases do processo de reconhecimento, cada uma delas relacionada à forma de informação do falante (segundo nível) e ao elemento operacional que deve ser examinado para obter informações a respeito dessa fase do processo.

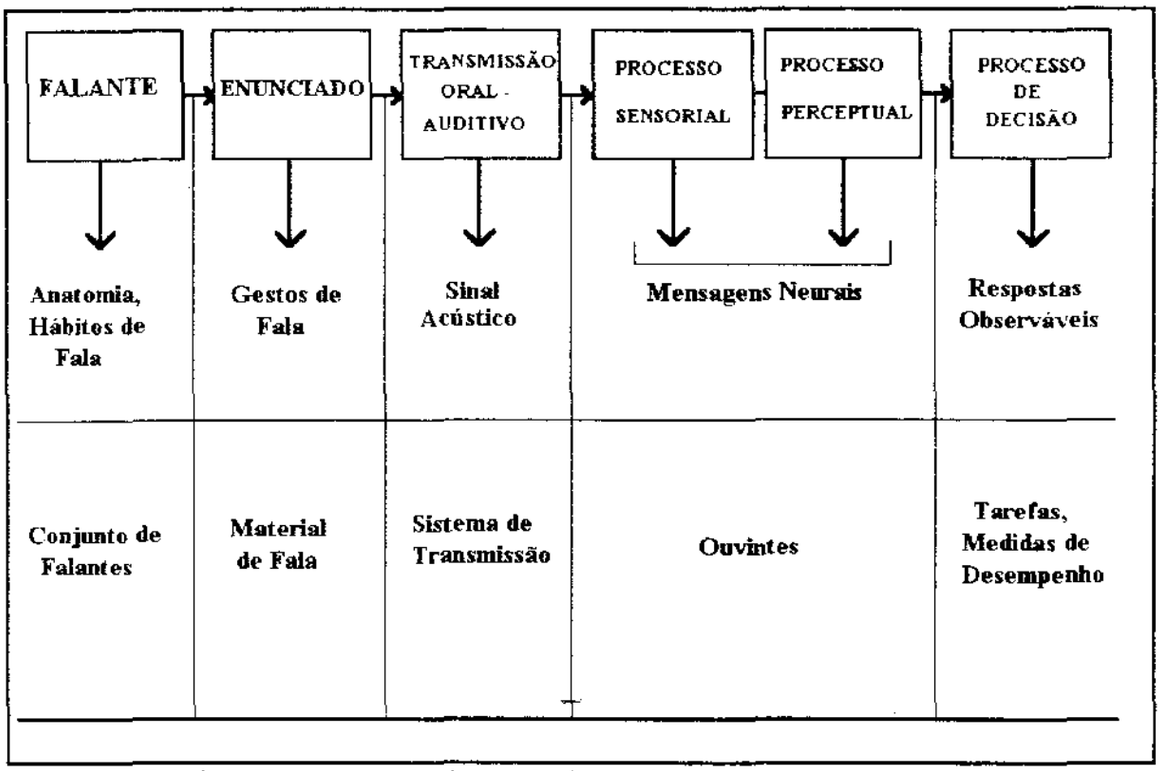


FIGURA 2.1: Esquema relacionando fases do processo de reconhecimento de falantes (auditivo) com o tipo de informação do falante e com os elementos operacionais abordados em experimentos (adaptado da figura 9.1, em Bricker e Pruzansky 1976:298)

O esquema da figura 2.1 destaca os principais determinantes das condições experimentais que devem ser fixadas para o estudo de cada aspecto do processo perceptual. As características básicas do **conjunto de falantes**, por exemplo, podem ser delimitadas de vários modos, com maior ou menor homogeneidade quanto a sexo, idade, raça, procedência, etc. No que diz respeito ao **material de fala**, podem ser estudadas as influências do tamanho e conteúdo do enunciado, estilos de produção, etc. Os efeitos de diferentes **condições de transmissão** podem ser avaliados através de diversas manipulações do sinal (limitação de banda, adição de ruído, etc). A variabilidade inter-**ouvintes** na habilidade em identificar/discriminar falantes pode ser examinada avaliando-se o efeito de treinamento específico no

desempenho. Finalmente, o tipo de **tarefa** pode envolver identificação, discriminação, avaliação escalar de similaridade, etc.

Na seção 2.3 serão discutidos experimentos que focalizam diversos aspectos do processo de reconhecimento de falantes através da audição, seguindo as linhas mestras do esquema na figura 2.1. Eventualmente, alguns experimentos podem explorar dois ou mais aspectos ao mesmo tempo (influência do treinamento no reconhecimento de falantes em diferentes condições de transmissão; influência do tamanho do enunciado e condições de transmissão nas avaliações de peso/altura, etc).

2.3) *Abordagem Perceptual: Procedimentos Experimentais*

2.3.1) *Características do Conjunto de Falantes*

Somos capazes não só de identificar falantes particulares, como também de extrair informação relacionada a características gerais de grupos de falantes, tais como sexo, idade, constituição física, etc. A precisão desses julgamentos pode variar bastante, dependendo da característica específica e, não raramente, encontramos resultados experimentais divergentes. Discutiremos a seguir alguns experimentos que focalizam a capacidade perceptual de inferir do sinal de fala certas características de grupos de falantes.

2.3.1.1) *Constituição Física*

Alguns traços vocais estão fortemente relacionados a diferenças anatômicas entre falantes. Em uma certa medida, esses traços não podem ser alterados livremente pelo falante. Fatores como o volume e comprimento das cavidades oral e

nasal, impõem limites à faixa de variação dos formantes; a massa e dimensões das cordas vocais delimitam a variação de F0; a capacidade pulmonar limita a faixa de amplitude, etc. Em geral, há uma correlação entre as dimensões do sistema fonador e certas características físicas globais. Assim, espera-se que um indivíduo alto e de compleição forte possua um trato vocal longo, cordas vocais volumosas e um grande volume respiratório, aspectos que se refletirão acusticamente em termos de médias baixas de formantes e freqüência fundamental e média alta de volume de fala (*loudness*).

A experiência cotidiana demonstra que ao falarmos com alguém desconhecido, sem que haja informação visual - pelo telefone, por exemplo - , teremos certas expectativas a respeito da constituição física do nosso interlocutor, provavelmente em decorrência das pistas vocais comentadas no parágrafo anterior. Mas até que ponto nossas estimativas estão próximas da realidade? Uma série de estudos tentou abordar mais objetivamente essa questão, de modo a verificar uma possível relação sistemática entre as avaliações subjetivas e os aspectos físicos reais. Bonaventura (1935; *apud* Laver e Trudgill 1979:9) avalia julgamentos subjetivos baseados em associações entre vozes e fotografias, verificando a existência de uma correlação significativa; a precisão dos julgamentos, entretanto, varia em função do tipo físico básico, definido segundo a classificação de Kretschmer (1925): os julgamentos são mais exatos para os tipo *pícnico* do que para os tipos *leptossômico* e *atlético* (v. também Fay e Middleton 1940 e Moses 1941 para resultados semelhantes) [1].

Lass e Harvey (1976) realizam um experimento mais controlado de associação voz-fotografia. Nesse estudo, os sujeitos avaliam fotos de duas pessoas diferentes (sempre apresentadas em pares) e devem julgar a qual das duas pertence o estímulo vocal; os testes são fechados, ou seja, há sempre uma resposta correta. As fotos são apresentadas em dois tipos de formato: corpo inteiro e apenas rosto. Lass e

Harvey verificam um acerto maior que o acaso, sendo que as fotos de corpo inteiro obtêm um índice um pouco maior. Os autores sugerem que essa diferença pode estar, em parte, relacionada a inferências extraídas de traços vocais que reflitam de algum modo o *status* sócio-econômico do falante, que o ouvinte, nas fotos de corpo inteiro, poderia associar ao vestuário. De qualquer forma, a existência de uma correlação significativa para as fotografias de rostos leva os autores a concluir que

> *there are apparently perceptual cues in the voice which reflect physical features conveyed in photographs of speakers* (Lass e Harvey 1976:1235).

Lass e seu grupo realizaram pesquisa extensa examinando julgamentos de peso e altura em diversas condições experimentais. Em Lass e Davis (1976) os sujeitos julgam peso/altura em um paradigma de múltipla escolha; Lass *et al.* (1978a) examinam avaliações através de estimativas diretas; Lass *et al.* (1980a) estudam julgamentos comparativos de peso/altura em amostras de fala apresentadas aos pares; Lass e Colt (1980) comparam julgamentos de peso/altura através de amostras de fala e através de fotos, constatando que as diferenças nas estimativas visuais e auditivas são pequenas; Lass *et al.*(1979a) examinam a influência do tamanho do enunciado (vogais isoladas, monossílabos, dissílabos e sentenças) nas avaliações de peso/altura, verificando que a complexidade fonética influi pouco na acuidade dos julgamentos; Lass *et al.* (1980b) utilizam estímulos com voz normal e sussurrada, observando que os resultados não diferem significativamente para as duas condições; Lass *et al.* (1979b) estudam o efeito de alterações temporais, utilizando amostras de fala comprimida e retrogradada, verificando que no segundo caso há uma queda significativa no desempenho dos juízes; Lass *et al.*(1980c) investigam o efeito de diferentes tipos de filtragem (*low-pass* e *high-pass*),

constatando que diferentes porções do espectro contêm pistas acústicas suficientes para avaliações aproximadas de peso/altura; Lass *et al.* (1980d) testam a consistência das estimativas dos sujeitos, verificando que o erro médio nas avaliações de peso/altura praticamente não se altera em quatro seções espaçadas de um dia, indicando que não há um efeito significativo de aprendizado.

Embora os experimentos do grupo de Lass tenham indicado que os ouvintes conseguem estimar com uma certa precisão o peso e a altura de falantes apenas com base em amostras de fala, esses resultados têm sido alvo de controvérsia. Cohen *et al.*(1980) criticam a metodologia empregada por Lass, sugerindo que os resultados obtidos seriam um sub-produto de técnicas estatísticas inapropriadas. Gunter e Manning (1982) observam que as estimativas de peso/altura só se aproximam das medidas reais se é feita uma comparação entre as **médias** (média real *versus* estimada, como nos experimentos de Lass), um procedimento que absorve grande parte dos erros de estimativa de cada sujeito; Gunter e Manning verificam que, ao se empregar *scores* de diferença (estimativa do sujeito *menos* medida real), os resultados mostram uma ampla variação nas estimativas, tanto intra- quanto inter-subjetivamente. Em um estudo recente, van Dommelen (1993) reavalia os mesmos dados empregados nos experimentos de Lass, submentendo-os a procedimentos estatísticos mais rigorosos, para constatar a inexistência de um relação significativa entre as estimativas de peso/altura dos ouvintes e as medidas reais; van Dommelen, no entanto, observa, ao contrário de Gunter e Manning (1982), que as estimativas, embora não correspondam à realidade, são altamente consistentes **entre** os ouvintes, ou seja, aparentemente, os ouvintes associam um **estereótipo** de peso/altura a algum fator acústico - muito provavelmente F0, afirma van Dommelen. Na verdade, parece não existir uma relação *de fato* entre F0 médio e peso/altura; Künzel (1989) examina essa possibilidade, estudando um grupo de 183 falantes (105 homens e 78

mulheres), e não verifica uma correlação significativa entre F0 médio e peso/altura reais dos falantes.

Muitos experimentos deverão ser realizados antes que se chegue a uma conclusão a respeito da capacidade de estimar corretamente peso e altura apenas a partir de pistas vocais. No entanto, no nosso entender, o aspecto mais importante é o fato de os ouvintes terem uma expectativa coerente, isto é, as mesmas pistas vocais conduzem ao mesmo tipo de inferência para a maioria dos ouvintes. Seria ingênuo reconhecer nessa reação *cliché* um ato arbitrário, já que deve ser fruto da experiência cotidiana em observar que **na média** certos aspectos vocais (F0 médio, faixa de freqüência dos formantes, etc.) correlacionam-se com certos aspectos físicos gerais. Obviamente, a variedade dos fatores metabólicos influenciando peso e altura reais dificulta a observação de relações significativas em experimentos limitados ao laboratório, mas é provável que tais relações sejam mais evidentes se forem estudados conjuntos de falantes **suficientemente** grandes.

2.3.1.2) *Sexo*

A identificação do sexo do falante está, obviamente, fortemente vinculada ao F0 médio. O comprimento das cordas vocais é, em média, 20 % menor nas mulheres (Wu e Childers 1991). Essa diferença anatômica faz com que o F0 das mulheres seja, em média, cerca de 80 % mais alto do que o dos homens. Mas F0 não é a única pista para a identificação do sexo do falante. Na verdade, a faixa de variação inter-falante do F0 médio em uma população grande não define limites precisos entre os sexos. Com base nos dados apresentados em Behlau (1984) para o F0 médio de vogais isoladas, podemos verificar uma faixa de variação de 78 - 161 Hz para os homens e 121 - 327 Hz para as mulheres; há, portanto, uma interseção entre os dois conjuntos de medidas.

A possibilidade de existirem outras pistas diferentes de F0, que seriam utilizadas pelos ouvintes para identificar o sexo do falante, foi examinada em diversos estudos empregando sons sem informação glotal. Schwartz (1968) descreve um experimento onde os sujeitos devem identificar o sexo dos falantes (adultos, n=18: 9 homens, 9 mulheres) a partir de produções isoladas das fricativas não-sonoras /f, θ, s, S/, observando-se acertos estatisticamente significativos de 93 % e 90 % para /s/ e /S/, respectivamente; a análise espectrográfica mostrou que os espectros de /s,S/ das mulheres tinham o ponto de maior concentração de energia um pouco mais alto (5500 e 2300 Hz nos homens *versus* 6500 e 3000 Hz nas mulheres, para /s/ e /S/, respectivamente). Esse deslocamento para freqüências mais altas, conclui Schwartz, seria uma conseqüência do menor tamanho do trato vocal feminino; a dificuldade em identificar o sexo dos falantes com /f,θ/ se deveria ao espectro relativamente plano desses sons. Ingemann (1968) conduz um experimento semelhante com maior número de fricativas, observando que, à medida que diminui a porção do trato à frente da constrição, mais difícil se torna a identificação correta do sexo do falante, isto é, fricativas mais anteriores veiculam menos informação quanto às dimensões do trato individual.

Schwartz e Rine (1968) neutralizam a informação glotal empregando vogais /i/ e /a/ sussurradas; apenas 400 milisegundos da seção central de cada vogal são usados para construir os estímulos. Os autores verificam um índice de 100 % de acertos nas identificações de sexo com /a/ e 95 % com /i/.

Brown e Feinstein (1975) utilizam uma laringe elétrica com freqüência fixa de 120 Hz, de modo a uniformizar o F0 de um grupo de falantes (adultos, n=20: 10 homens, 10 mulheres). Os autores verificam que vozes masculinas com concentração de energia espectral nas freqüências baixas e vozes femininas com energia concentrada nas freqüências altas são facilmente identificadas; falantes com energia

espectral mais centralizada provocam um maior número de identificações erradas de sexo, sugerindo que a informação espectral é importante para a decisão.

Os resultados dos experimentos acima descritos indicam que a identificação do sexo do falante prescinde de informação relacionada a F0. Em parte, a informação espectral extraída pelo ouvinte vem do menor tamanho do trato feminino, mas isso não explicaria tudo, já que em alguns casos espera-se que a diferença no tamanho **total** do trato seja pequena, ou eventualmente inexista. O fato é que não é apenas o tamanho total do trato que difere em homens e mulheres, mas também certas proporções entre regiões do trato. A razão dos comprimentos da faringe e do trato inteiro são diferentes entre os sexos: os homens têm a faringe proporcionalmente mais longa do que as mulheres (Fant 1973; 1980). Sundberg *et al.* (1987) observam, através de tomografia computadorizada, que, mesmo considerando-se a faringe isoladamente, há diferenças nas proporções relativas da laringofaringe e da orofaringe entre homens e mulheres. Essas distinções anatômicas fazem com que o *F-Pattern* feminino apresente **razões** entre formantes diferentes do masculino, ou seja, os formantes das mulheres não podem ser calculados apenas como um re-escalamento linear, a partir das freqüências dos formantes masculinos. Assim, mesmo nos casos em que o tamanho total do trato é equivalente, há informação suficiente para identificar o sexo do falante.

Há alguns aspectos relacionados à fonte glotal, além do F0 médio, que também diferenciam homens e mulheres. As estruturas laríngeas dos dois sexos diferem anatomicamente em diversos aspectos (Titze 1987), o que traz conseqüências quanto às características da forma de onda glotal. Holmberg *et al.*(1988) observam que a vibração das cordas vocais das mulheres apresenta menor velocidade de fechamento e menor duração da fase fechada; esses aspectos refletem-se acusticamente na forma de um envelope espectral da fonte mais abrupto, emprestando à voz feminina uma certa qualidade *breathy*. A reprodução dessas

características é uma condição fundamental para a naturalidade na síntese de vozes femininas (Klatt e Klatt 1990; Karlsson 1992).

Algumas diferenças espectrais entre homens e mulheres parecem estar relacionadas a aspectos de absorção acústica. Childers e Wu (1991) verificam que as larguras de banda dos formantes vocálicos de mulheres são consideravelmente mais largas do que no espectro de vogais produzidas por homens. Os autores sugerem que essas diferenças estariam associadas, nas mulheres, a maiores perdas térmicas e de radiação, embora não expliquem porque essas perdas são maiores nas mulheres.

Alguns estudos apontam diferenças entoacionais entre homens e mulheres, indicando, para as mulheres, uma maior variabilidade de F0, expressa não só em termos de uma maior faixa de variação, como também pelo uso de um maior número de tipos de contorno entoacional (Brend 1971; Smith 1979; Abe 1980; Graddol e Swann 1983; Garret e Healey 1987). Esse aspecto, provavelmente, tem determinações sócio-culturais, embora alguns fatores anatômicos também possam ter alguma influência (Cf. Graddol e Swann 1983). O mesmo parece ocorrer com aspectos de fluência; Coleman (1971:576) sugere que

> *it is possible that males and females differ in some learned speech characteristics such as rate or juncture.*

Com efeito, já se observou que as mulheres falam, em média, mais rápido que os homens (Kaiser 1940; Elert 1964: *apud* Karlsson 1992). Aparentemente, algumas diferenças no plano da organização temporal do enunciado são definidas bem cedo; Günzburger *et al.*(1987) verificam que os ouvintes só identificam corretamente o sexo de crianças na pré-puberdade (7-8 anos) se os estímulos são sentenças - para vogais isoladas a identificação fica perto do acaso. Os autores argumentam que,

nessa faixa de idade, não existem diferenças anatômicas entre os sexos que pudessem fornecer pistas de F0 e formantes (por esse motivo a identificação do sexo é impossível com vogais isoladas), já quando a identificação baseia-se em sentenças, o ouvinte pode se valer de pistas rítmico-temporais, derivadas - segundo os autores - de hábitos articulatórios culturalmente determinados. Não é certo, entretanto, que inexista alguma influência de aspectos anatômicos; Kuehn e Moll (1976) sugerem que a fala mais rápida das mulheres pode estar associada ao menor tamanho de seus órgãos vocais, o que provocaria movimentos mais rápidos entre os *targets* articulatórios.

2.3.1.3) *Idade*

A avaliação da idade do falante parece ser uma das mais importantes dimensões psicológicas no processo de percepção de uma voz particular (Walden *et al.*1978). Experimentos perceptuais têm demonstrado que os ouvintes conseguem estimar razoavelmente bem a idade de um falante exclusivamente através de informação vocal. Ptacek e Sander (1966) verificam que em 99 % dos casos os ouvintes diferenciam acertadamente vozes de indivíduos em dois grupos de idade (menos de 35 *versus* mais de 65 anos). Shipp e Hollien (1969) relatam uma correlação alta (r=.88; p<.01) entre a idade cronológica real e a idade estimada pelos juízes, para um grupo de 175 falantes, abrangendo uma faixa de 70 anos (v. também Helfrich 1979, para resenhas de outros estudos perceptuais).

Alterações fisiológicas no nível da laringe são a principal causa de modificações no sinal acústico em função da idade. Durante a adolescência, o rápido aumento da massa vibrante provoca um decréscimo significativo no F0 médio, na fase da "mudança de voz". O F0 médio, no entanto, continua decrescendo, embora com menor velocidade, mesmo após a maturidade (Endres *et al.* 1971). Hollien e

Shipp (1972) observam um aumento de F0 após 65 anos, mas esse resultado parece estar relacionado ao método sincrônico que utilizaram: como as gerações mais jovens tendem a ter compleição mais robusta, o efeito observado pode ser ilusório, relacionado antes à constituição física do que à idade. O único estudo longitudinal sobre variações de F0 em função da idade, observou um decréscimo constante de F0, mesmo após os 65 anos (Endres *et al.* 1971). Pode existir, no entanto, uma variação inter-falante considerável dentro de grupos da mesma idade, em função da condição física geral (Ramig e Ringel 1983)

O aumento das micro-perturbações de F0 em falantes mais idosos, tanto no domínio da freqüência quanto da amplitude (*jitter* e *shimmer*), também é um aspecto freqüentemente observado (Helfrich 1979; Ramig e Ringel 1983; Linville 1988). Essas perturbações podem estar relacionadas a problemas de coordenação no sistema nervoso central (Helfrich 1979), mas fatores de ordem fisiológica também devem contribuir; Ferreri (1959; *apud* Hollien e Shipp 1972), estudando as laringes de falantes idosos (60 anos ou mais) relata uma perda de elasticidade nas estruturas laríngeas, além do acúmulo de tecido gorduroso e a eventual existência de edemas crônicos [2].

2.3.2) *Material de Fala*

Um dos aspectos mais importantes em experimentos perceptuais de reconhecimento de falantes é o próprio material de fala utilizado como estímulo. Nesse sentido, devem ser examinados os efeitos da duração e conteúdo do material apresentado aos ouvintes. Um dos primeiros experimentos abordando esse aspecto (Pollack *et al.*1954), utiliza um conjunto de vozes familiares aos ouvintes (trechos extraídos de fala corrente), e examina o efeito da duração do estímulo, cruzando essa variável com uma série de outros fatores: tamanho do conjunto de falantes,

condições de filtragem, voz sussurrada e apresentação de duas vozes simultâneamente. Os principais resultados desse estudo são: (1) a performance dos ouvintes melhora rapidamente até amostras de um segundo; a partir daí há uma tendência à estabilização; (2) para obter o mesmo índice de acertos com voz sussurrada é necessário empregar uma amostra de fala cerca de três vezes maior do que com voz normal; (3) quando duas vozes são apresentadas simultâneamente, também é preciso aumentar o tamanho da amostra para obter índices iguais aos obtidos com vozes isoladas; nesse tarefa, porém, parece haver uma grande variação na proficiência dos ouvintes (v. seção 2.3.4); (4) para um pequeno conjunto de falantes, o índice de identificações corretas altera-se pouco em função de diferentes padrões de filtragem (*low-* e *high-pass*); para grupos maiores a performance cai consideravelmente na condição filtrada ; (5) a duração do estímulo *per se* é menos importante do que o repertório de fala: a mera repetição do mesmo estímulo 3 ou 4 vezes não aumenta os acertos, mas a apresentação de um estímulo com mais informação segmental (mesmo que seja, em termos absolutos, mais curto do que as repetições de um mesmo trecho) aumenta os acertos.

Bricker e Pruzansky (1966) realizam um experimento utilizando amostras de fala de vários tipos: vogais e sílabas CV extraídas de fala corrente, palavras monossilábicas, palavras *non sense* dissilábicas e sentenças. Os autores observam que a identificação aumenta com o aumento do número de fonemas do estímulo, mesmo que seja controlada a duração; assim, sílabas CV produzem mais acertos do que vogais, mesmo sendo mais curtas do que as últimas, palavras monossilábicas produzem mais acertos do que sílabas CV, mesmo que sejam mais curtas que as últimas, etc.

Em uma simulação de situação forense, Künzel (1990) utiliza 5 falantes (adultos, sexo masculino), que lêem 2 textos de diferentes durações (4 e 16 segundos). Uma segunda amostra de cada um dos falantes, de 60 segundos, é

apresentada aos ouvintes para que se familiarizem com as vozes. Um dos falantes do grupo é selecionado para ser o "criminoso". Grupos de ouvintes devem então decidir (8 e 30 dias após a escuta da amostra de familiarização), com base nas amostras de 4 e 16 segundos, qual dos falantes "suspeitos" é o "criminoso". Künzel observa que a amostra mais longa (16 segundos) provoca um maior índice de identificações corretas.

Compton (1963) emprega como estímulo seções de diferentes durações extraídas de vogais /i/ sustentadas. Compton verifica que apenas 25 milisegundos são suficientes para obter identificações corretas acima do acaso (36 % de acertos); o índice de identificações corretas cresce para 65 % nas amostras de 750 ms. Amostras submetidas a diversas condições de filtragem (*low-* e *high-pass*) devem ser, em geral, mais longas, de modo a obter o mesmo índice de acertos das amostras não-filtradas; a filtragem *high-pass*, no entanto, afeta menos a identificação do que a *low-pass*.

Emmorey *et al.* (1984) examinam, a partir de um conjunto de vozes famosas (políticos, artistas, etc.) a influência do tamanho e conteúdo do enunciado nos índices de reconhecimento. Três tipos de estímulo foram utilizados: vogais isoladas, palavras isoladas e textos de 2 segundos, observando-se índices de acertos de 34 %, 40 % e 61 %, respectivamente. Ao contrário de Bricker e Pruzansky (1966; já comentado acima), não surgiu uma relação consistente entre o número de fonemas (nas palavras ou nos textos) e o número de identificações corretas; as autoras argumentam que a diferença pode estar relacionada ao fato de, em Bricker e Pruzansky (1966) terem sido usadas vozes familiares aos ouvintes, enquanto aqui foi empregado um conjunto de vozes famosas, embora não expliquem porque essas duas tarefas envolveriam processamentos diferentes.

Alguns experimentos perceptuais têm indicado que o conteúdo segmental específico também influencia a identificação, embora os resultados não sejam

totalmente consistentes nos diversos estudos. Stevens *et al.*(1968) verificam, informalmente, que palavras contendo vogais anteriores são mais eficientes para o reconhecimento auditivo de falantes. Por outro lado, La Rivière (1975) observa um favorecimento em vogais baixas /a,æ/. Avaliações estatísticas da eficiência de diferentes vogais,. com base em medidas de formantes, também não têm relatado resultados consistentes (Cf. Goldstein 1976 *versus* Paliwal 1984) [3].

A questão da importância da duração e conteúdo do estímulo não é fácil de ser tratada. O fato é que, mesmo para grupos de falantes teoricamente "homogêneos" não há como se garantir que os ouvintes estejam utilizando pistas semelhantes para todos os falantes. Alguns aspectos podem ser mais salientes para um falante do que para outro. Assim, se a pista relevante está relacionada a características da fonte glotal (perturbações de F0, por exemplo), é provável que a inclusão de mais fonemas no estímulo influencie menos do que se a pista saliente for relacionada a aspectos articulatórios. É preciso considerar também que a saliência de uma determinada pista pode variar também entre os ouvintes. Todos esses complicadores dificultam uma avaliação objetiva dos resultados experimentais.

2.3.3) *Condições do Canal de Transmissão*

A experiência cotidiana demonstra que o reconhecimento de falantes através do canal telefônico impõe um maior grau de dificuldade. A limitação de banda imposta pela transmissão telefônica, restringindo o sinal a uma faixa de 350 -3500 Hz aproximadamente, é a principal responsável pela queda na performance, embora outros fatores também possam influir, tais como o ruído de quantização e distorção harmônica (a magnitude desses efeitos, no entanto, depende de fatores como o *path* da conexão telefônica e o tipo de aparelho utilizado: Cf. Ichikawa *et al.* 1978).

Künzel (1990) verifica experimentalmente que estímulos apresentados através do canal telefônico produzem um menor índice de acertos, em um paradigma de decisão perceptual binária para pares de falantes (o ouvinte deve decidir se as duas amostras pertencem ou não ao mesmo falante). Schmidt-Nielen e Stern (1985) observam que, mesmo para o reconhecimento de vozes familiares, a filtragem *narrow-band* faz o índice de erros crescer de 12 para 31 %, em comparação com o sinal não filtrado.

O efeito de diversos tipos de filtragens *low-* e *high-pass* em interação com a duração do enunciado já foi abordado em 2.3.2, onde pudemos observar que, em geral, é necessário um maior conteúdo segmental para que os sinais filtrados produzam o mesmo índice de acertos dos sinais integrais.

O efeito de diferentes tipos de ruído e distorção ambientais na identificação perceptual de falantes ainda não mereceu estudo sistemático. Há, entretanto, uma série de estudos sobre a inteligibilidade da fala em condições adversas, que podem oferecer algum subsídio para a discussão do nosso tema.

Helfer e Huntley (1991) estudam os efeitos do ruído e da reverberação na inteligibilidade da fala, destacando os tipos de segmento consonantal mais afetados em cada condição. Sob ruído, a percepção das nasais é pouco prejudicada, enquanto plosivas e fricativas são freqüentemente confundidas (na maior parte dos casos, plosivas são confundidas entre si, ou erroneamente identificadas como fricativas). Com a presença de reverberação, o quadro é invertido, sendo os segmentos nasais mais prejudicados. Quanto ao ponto de articulação, Helfer e Huntley verificam que as bilabiais e dentais sofrem uma maior degradação do que as alveolares e palatais, tanto sob ruído quanto sob reverberação. Como regra geral, a reverberação afeta mais sensivelmente a percepção de traços de baixa freqüência e o ruído prejudica os traços de alta freqüência (Cf. Gelfand e Silman 1979; Helfer 1992).

Cox *et al.* (1987) realizam uma série de testes de inteligibilidade de fala a partir de amostras de conversação simulada, mais próximas a situações reais. Quatro ambientes diferentes são também simulados: sala de estar normal, sala de aula e dois eventos sociais com diferentes características de ruído e reverberação. Os autores observam que, efetivamente, os ambientes mais ruidosos, ou com mais reverberação, prejudicam sensivelmente a inteligibilidade, mesmo com fala fluente (onde o ouvinte pode utilizar pistas contextuais *top-down*). Uma análise estatística baseada nos erros específicos dos sujeitos revelou que os aspectos que mais contribuíram para a perda de inteligibilidade foram (em ordem decrescente de importância): confusões quanto ao ponto de articulação da consoante, vozeamento e tipo de vogal. O aspecto mais interessante nesse estudo é a observação de que diferentes falantes apresentam diferentes graus de suscetibilidade à degradação imposta pelo ruído e reverberação; há uma forte interação falante x traço, ou seja, determinados segmentos de determinados falantes são mais afetados pelo ruído/reverberação ambientais.

Um aspecto importante para a questão da identificação de falantes, diz respeito às modificações na produção de fala em ambientes ruidosos. Van Summers *et al.* (1988) estudam alterações na produção de palavras isoladas de dois falantes em quatro diferentes condições : silêncio e com ruído de 80, 90 e 100 dB, inserido através de fones de ouvido. Os autores relatam os seguintes resultados: (1) a amplitude média da fala aumenta proporcionalmente ao nível de ruído adicionado; (2) a duração segmental também aumenta, do mesmo modo que a amplitude; (3) o F0 médio aumenta significativamente para um falante, mas para o outro não há alteração significativa; (4) a inclinação (*tilt*) espectral diminui, ou seja, há um aumento da energia em freqüências mais altas. Uma análise perceptual mostrou que os estímulos produzidos originalmente sob ruído apresentam um maior grau de inteligibilidade para os ouvintes, especialmente se apresentados também sob ruído.

Já se observou que a fala sob ruído de 90 dB apresenta um aumento na freqüência do primeiro formante e uma redução no segundo formante (Juang 1991). A elevação de F1 pode estar relacionada ao aumento do esforço vocal, com conseqüente maior abertura mandibular (Cf. Traunmüller 1988). A redução na freqüência média de F2 não parece ter uma explicação simples, implicando uma postura mais posteriorizada da língua, o que pode ser uma conseqüência indireta da expansão dos movimentos mandibulares, já que há uma certa interação entre as duas estruturas (Cf. Lindblom e Lubker 1985).

A magnitude do aumento de amplitude na fala sob ruído parece estar relacionada com a configuração espectral do ruído. Sundberg *et al.* (1988) observam que o aumento de amplitude é inversamente proporcional ao nível de energia nas altas freqüências do ruído interferente.

As alterações da fala em função da presença de ruído e/ou reverberação são um aspecto importante para a o estudo da Identificação de Falantes, especialmente no paradigma forense, já que, nessa situação, as gravações são freqüentemente realizadas em condições adversas. Como já comentamos anteriormente, faltam, infelizmente, estudos sistemáticos sobre o tema. Os estudos acima discutidos indicam, no entanto, que devemos esperar modificações em alguns parâmetros acústicos, embora não seja possível avaliar objetivamente as conseqüências no nível perceptual.

2.3.4) *Ouvintes*

2.3.4.1) *Habilidade*

Faltam investigações sistemáticas quanto às diferenças entre os ouvintes, mas alguns estudos fornecem uma indicação da magnitude da variabilidade do desempenho inter-ouvinte. Stevens *et al.* (1968) testam diversos aspectos de performance em um experimento onde o ouvinte deve decidir, entre oito vozes, qual é igual à amostra teste; os autores observam que há diferenças consideráveis na habilidade dos ouvintes para essa tarefa, com índices de erro variando de 5 a 16%. Em um experimento envolvendo decisões do tipo *igual-diferente* em pares de vozes, Clarke e Becker (1969) verificam uma variação inter-ouvinte ainda maior nos índices individuais de erros: 8 a 26%. Em Bricker e Pruzansky (1966) as taxas de erro de 16 ouvintes na identificação de 10 falantes, a partir de dissílabos isolados, variou de 2 a 27 %. (v. também Rosenberg 1973; Hollien *et al.* 1982).

A habilidade relativa dos ouvintes parece estar, de algum modo, relacionada a determinadas propriedades do sinal. Abberton e Fourcin (1978) empregam estímulos de fala normal, sussurrada, além de diversas combinações do sinal laringográfico isolado com diferentes manipulações de aspectos rítmicos e espectrais (F0 + durações com e sem informação espectral, F0 + durações normalizadas para todos os falantes, F0 normalizado para todos os falantes + pistas duracionais não normalizadas, etc.); cada condição, portanto, ressalta determinadas características do sinal original. Abberton e Fourcin verificam que há uma variação significativa na habilidade dos ouvintes na identificação dos falantes, mas a variação não é homogênea para todas as condições, isto é, alguns ouvintes têm um desempenho melhor em algumas condições e pior em outras. Nesse sentido, também é interessante o estudo de Pollack *et al.* (1954), onde se verifica que há uma variação na proficiência dos ouvintes no reconhecimento de vozes familiares, mas apenas no teste envolvendo a apresentação de duas vozes simultâneamente.

Treinamento é um fator relevante para o desempenho dos ouvintes na identificação de falantes. Clarke e Becker (1969) observam que o grupo de ouvintes

que teve oportunidade de trabalhar previamente com um conjunto de vozes diferente do utilizado no teste obteve 67 % de acertos, enquanto o outro grupo, submetido pela primeira vez à tarefa, obteve apenas 58 % de respostas corretas. A experiência prévia com qualquer tipo de material de fala melhora a performance na detecção auditiva de tentativas de disfarce (Reich e Duke 1979; Reich 1981). Stevens *et al.* (1968) verificam que o tempo médio de decisão em testes de identificação auditiva de falantes cai consideravelmente à medida que o ouvinte se familiariza com a tarefa. O efeito do treinamento, entretanto, pode não ser homogêneo para todo o grupo treinado, com alguns ouvintes se beneficiando mais do que outros (Hollien *et al.* 1982).

O grau de familiaridade com os falantes é também um componente importante na avaliação do desempenho dos ouvintes na identificação. Hollien *et al.* (1982) examinam esse aspecto, utilizando estímulos em três diferentes condições: fala normal, com *stress* psicológico induzido (choques elétricos aleatórios) e sob disfarce. Os resultados desse estudo indicaram uma diferença significativa entre os dois grupos na identificação dos falantes, com o grupo familiarizado com os ouvintes obtendo um índice expressivamente maior de respostas corretas, independentemente da condição de produção. No mesmo estudo, Hollien *et al.* examinam o efeito da familiaridade com a língua falada (Inglês Americano), observando que o grupo (poloneses) que não conhecia o idioma obtém um maior índice de erros, embora o efeito aqui seja menos expressivo do que o observado com relação à familiaridade com o conjunto de falantes. É interessante notar, em Hollien *et al.* (1982), que na condição *disfarce* o efeito da familiaridade com o idioma do falante tende a desaparecer, isto é, os ouvintes poloneses não apresentam uma queda de performance tão acentuada nessa condição quanto os ouvintes ingleses. Aparentemente, as estratégias empregadas pelos ouvintes estrangeiros são tão efetivas para a condição *disfarce* quanto são para a condição fala normal. Esse

resultado é um tanto surpreendente, na medida em que há uma certa expectativa que os ouvintes não familiarizados com a língua direcionem sua atenção menos para detalhes articulatórios locais do que para aspectos de longo termo, sendo esses últimos mais passíveis de modificação sob disfarce (Cf. Hollien e Majewski 1977; Hollien 1990).

Outro aspecto que parece influenciar os níveis de identificação auditiva de falantes é a defasagem temporal entre as apresentações da amostra teste e da amostra referência. Resultados experimentais focalizando esse aspecto são difíceis de ser comparados, já que a performance aqui vai depender de uma série de fatores, tais como grau de familiaridade com as vozes, tamanho do conjunto de falantes, além da própria extensão de tempo entre as sessões. Um possível *design* experimental envolve a tarefa de, a partir da oitiva de um conjunto de vozes, decidir qual delas corresponde a uma amostra ouvida há um determinado espaço de tempo. Essa tarefa simula razoavelmente bem a situação forense de uma "testemunha auricular", que deve identificar um criminoso a partir de um grupo de suspeitos. Um dos primeiros estudos perceptuais sobre identificação de falantes, motivado por um famoso caso de seqüestro [4] (McGehee 1937) baseia-se nesse paradigma experimental. O procedimento utilizado exigia que os sujeitos ouvissem "ao vivo" a voz de um indivíduo (oculto por uma cortina opaca) e o identificassem, entre um grupo de falantes "suspeitos" algum tempo após, com o intervalo de tempo variando de apenas um dia até cinco meses (McGehee 1944 reproduz o mesmo tipo de experimento com vozes gravadas). McGehee observa que o índice de identificações corretas é bastante alto no dia imediatamente após a escuta da amostra referência, mas cai gradualmente em sessões mais afastadas no tempo (de 83 % com um dia de defasagem até apenas 13 % - perto do nível de acaso - cinco meses depois).

Pesquisas mais recentes não suportam os resultados de McGehee (1937, 1944), embora seja difícil uma comparação direta, já que os intervalos de tempo são

menores e pode haver a influência de outras variáveis, que discutiremos mais adiante. Hollien *et al.* (1983), examinando a performance de ouvintes na identificação auditiva de um "criminoso" em um evento simulado, observam que o índice de acertos é **maior** duas semanas do que um dia ou uma semana após o "crime". Künzel (1990), também empregando um paradigma de crime simulado, observa que a performance 8 dias após o evento é melhor do que no dia imediatamente posterior.

A influência do tempo decorrido na performance é um tema difícil de ser tratado experimentalmente. Se são utilizados grupos distintos para cada intervalo de tempo (como em Hollien *et al.* 1983), é preciso que esses grupos sejam suficientemente grandes, de modo a neutralizar eventuais efeitos relacionados ao grau de habilidade do ouvinte, um aspecto que, como já vimos acima, pode ter uma considerável variabilidade inter-subjetiva. Por outro lado, se o mesmo grupo realiza todos os testes (como em Künzel 1990), é possível que surja um efeito de aprendizado, isto é, a performance melhora com o tempo decorrido porque os ouvintes familiarizam-se com a tarefa.

Alguns estudos têm observado que o emprego de amostras não-contemporâneas do mesmo falante para o teste de identificação acarreta uma queda no desempenho (Rothman 1975; Papamichalis e Doddington 1984), mas esse *design* experimental também traz algumas dificuldades de interpretação, já que as amostras de fala comparadas diferem também quanto ao conteúdo, não sendo possível separar totalmente esse efeito da influência da não-contemporaneidade.

Ao longo desta seção pudemos observar que não só varia a habilidade do ouvinte em reconhecer falantes, como há também uma série de fatores que podem modificar de algum modo essa habilidade, dependendo da definição da tarefa específica e das condições experimentais. O principal objetivo prático no exame de

diferenças de habilidade inter-ouvintes é a definição de um conjunto de ouvintes suficientemente grande, de modo a neutralizar diferenças individuais e obter uma performance média estável. Há alguma evidência (Williams 1964; *apud* Bricker e Pruzansky 1976:301) indicando que grupos de 12 ouvintes produziriam respostas médias estáveis, mas - como vimos ao longo dessa discussão - a proficiência individual parece interagir fortemente com outros fatores, o que recomenda prudência antes de generalizar resultados de experimentos controlados.

2.3.4.2) *Fontes de Informação Utilizadas pelo Ouvinte*

O processo perceptual *per se* não pode ser examinado diretamente, mas apenas através de inferências, isto é, avaliando-se a força relativa de diferentes fontes de informação contidas no sinal. Diversas formas de filtragem/mascaramento podem ser aplicadas, de modo a isolar pistas ou, mais provavelmente, conjuntos de pistas. Estímulos assim construídos produzirão respostas seletivas, revelando, ao menos indiretamente, algumas estratégias perceptuais empregadas pelos ouvintes.

Vários procedimentos de manipulação do sinal original já foram empregados experimentalmente para evidenciar/mascarar determinados traços acústicos potencialmente importantes para a percepção: sinais laringográficos isolados (van Dommelen 1987) ou em conjunto com outras pistas prosódicas (Abberton e Fourcin 1978), voz sussurrada (La Rivière 1975), laringe elétrica com F0 fixo (Coleman 1973), filtragens *low-*, *high-* e *band-pass* com diferentes freqüências de corte (Pollack *et al.* 1954; Compton 1963; La Rivière 1975), fala retrogradada (van Lancker *et al.* 1985a), fala comprimida/expandida (van Lancker *et al.* 1985b), etc. A tabela 2.1 resume as principais técnicas de manipulação do sinal, destacando os aspectos acústicos que são evidenciados ou neutralizados em cada caso.

Filtragem/Mascaramento	**Pista(s) Mantida(s) Mais Importante(s)**	**Pista(s) Neutralizada(s) Mais Importante(s)**
Laringe Elétrica (F0 fixo) Voz Sussurrada	Aspectos Espectrais (FILTRO)	F0 médio Contornos de F0
Sinal Laringográfico *Low-Pass* (≤ 200 Hz)	F0 médio Contornos de F0	Aspectos Espectrais Padrões Articulatórios
Filtragens *band-pass*	Faixas Seletivas do Espectro	Faixas Seletivas do Espectro
Fala Comprimida ou Expandida (F0 mantido)	Aspectos de F0 Qualidade de Voz Freq. média de Formantes	Velocidade de Fala
Fala Retrogradada	F0 médio Qualidade de Voz Freq. média de Formantes	Padrões Articulatórios

TABELA 2.1: Principais efeitos de diversas técnicas de filtragem/mascaramento

Não discutiremos agora os resultados dos muitos experimentos que empregaram técnicas de filtragem/mascaramento, já que, se ainda não foram examinados anteriormente, certamente o serão ao longo do presente trabalho, quando abordarmos, separadamente, diversos parâmetros acústicos importantes para a Identificação de Falantes (seções 4 a 9). Mais interessante seria discutir alguns aspectos gerais que podemos depreender da análise do conjunto desses experimentos. O ponto mais relevante nos parece ser a própria validade dos métodos de filtragem/mascaramento. A questão aqui é saber em que medida o isolamento ou neutralização de um determinado traço afeta a percepção de outro. A experiência mais extensa com o estudo das pistas importantes para a identificação de segmentos de fala sugere que, na maior parte dos casos, existe uma certa interdependência entre as pistas, como se uma "potencializasse" a(s) outra(s). Analogamente, seria razoável ter o mesmo tipo de expectativa no que diz respeito à Identificação de Falantes.

Estendendo um pouco mais a analogia, também parece razoável admitir que, assim como no que se relaciona aos diferentes traços distintivos da **fala**, podemos esperar diferentes graus de complexidade na interação de pistas relevantes para diferentes **falantes**. Com efeito, há evidências indicando que esse é o caso; van

Lancker *et al.* (1985a,b) verificam que, após diversos tipos de manipulação do sinal (retrogradação, compressão e expansão), os efeitos no reconhecimento de vozes familiares aos ouvintes varia consideravelmente em função do falante, ou seja, algumas vozes são bem reconhecidas em qualquer condição (normal ou manipulada), enquanto outras tornam-se praticamente irreconhecíveis após a manipulação, embora sejam facilmente reconhecíveis na condição normal. Dito ainda de outra forma: um determinado parâmetro pode ser essencial para a caracterização de algumas vozes, mas totalmente irrelevante para outras. Como complicador adicional podemos certamente contar com alguma variação inter-ouvintes, já que a força relativa de uma pista ou complexo de pistas não será homogênea para todos os ouvintes (Cf. van Dommelen 1987).

2.3.5) *Especificação da Tarefa*

De um modo geral, podemos definir três tipos básicos de tarefa: *Discriminação*, *Reconhecimento* e *Identificação* . Por *discriminação* entendemos a tarefa que exige do ouvinte uma decisão binária do tipo *igual-diferente*, tomada a partir de um par de vozes apresentadas simultâneamente. *Reconhecimento* seria o processo análogo à experiência cotidiana de reconhecer vozes familiares sem acesso visual ao falante. *Identificação* envolve o confronto de uma amostra teste com um conjunto de *n* amostras referência, tendo o ouvinte que selecionar, entre essas amostras, qual corresponde à voz da amostra teste.

Para a tarefa de identificação podem existir modificadores importantes. O teste pode ser *aberto* ou *fechado:* no primeiro caso, a amostra teste não precisa corresponder **necessariamente** a qualquer uma das amostras referência (o fato de o ouvinte ser ou não informado quanto à natureza do teste seria uma condição adicional). Outra divisão na tarefa de identificação relaciona-se com a defasagem de tempo entre a apresentação das amostras teste e referência: as amostras podem ser apresentadas simultâneamente ou com um atraso de um determinado intervalo de tempo. Evidentemente, é possível combinar as duas divisões. Assim, podemos ter um teste *aberto* de identificação, envolvendo a apresentação de um conjunto de amostras teste alguns dias/semanas/meses após a amostra referência; nesse caso teríamos uma simulação da situação forense típica da "testemunha auricular" [5].

Há evidências de que *discriminação* e *reconhecimento* são processos independentes. Van Lancker e Kreiman (1985) estudam a performance dessas duas tarefas em três diferentes grupos: sujeitos normais, sujeitos com lesão cerebral no hemisfério esquerdo e sujeitos com lesão cerebral no hemisfério direito. Os seguintes resultados são observados: (1) nos sujeitos normais o desempenho na

discriminação (vozes não familiares aos ouvintes) é fracamente correlacionado com o desempenho no *reconhecimento* (vozes familiares) e nos sujeitos com lesão cerebral não há qualquer correlação; (2) o grupo com lesão no hemisfério direito tem mais habilidade na tarefa de *reconhecimento*, enquanto o grupo com lesão no hemisfério esquerdo tem mais habilidade para a *discriminação*. Com base nas tradicionais associações de especialidades com os hemisférios cerebrais, as autoras interpretam os resultados como uma indicação de que a *discriminação* envolve prioritariamente uma análise de **traços**, onde detalhes da voz são isolados e comparados, enquanto o *reconhecimento* dependeria mais fortemente da análise de um *pattern*, ou seja, um processo holístico.

É possível entender a *identificação* como uma sucessão de *discriminações*, especialmente em testes *abertos*. Mas, mesmo nesse caso, não é certo que o ouvinte tome sua decisão final a partir de decisões binárias par-a-par, sendo mais provável que estabeleça um balizamento perceptual em função do conjunto de amostras apresentado. Provavelmente a *identificação*, tal como foi aqui definida, seja um processo híbrido, envolvendo ao mesmo tempo análise de traços e percepção gestáltica.

SEÇÃO 3: MATERIAL E MÉTODOS

Ao longo deste trabalho serão descritos vários experimentos envolvendo o estudo da variabilidade inter- e intra-falante de diversos parâmetros acústicos, tais como formantes, freqüência fundamental, espectro de longo termo, etc. A presente seção descreve o material básico utilizado para as análises, assim como a metodologia empregada para a gravação. Detalhes específicos referentes a técnicas de medição e análise dos parâmetros acústicos serão abordados separadamente em cada seção individual, ao longo do trabalho.

3.1) *Falantes*

Para a maior parte dos experimentos utilizou-se um grupo de oito falantes, adultos (23-45 anos), do sexo masculino. Cinco falantes desse grupo são oriundos de cidades do interior do Estado de São Paulo (EN, ZP, WA, MS e DO), um de Mato Grosso (AG), um do Rio de Janeiro (R1/R2) e um de Recife (ZR).

Um dos falantes realizou duas gravações com uma defasagem de 4 meses. Na primeira gravação (onde está denominado como R1) esse falante estava sob forte estado gripal, com presença de secreção nas vias respiratórias superiores e conseqüente alteração qualitativa da voz. Na segunda gravação (onde está codificado como R2), sem qualquer vestígio de gripe, o mesmo falante apresentava sua qualidade de voz habitual. A inclusão das amostras de R1 e R2 permitirá a verificação de eventuais variações vocais em função da presença da enfermidade e, acumulativamente, em função da não-contemporaneidade das amostras.

Em alguns experimentos foram empregadas gravações de dois gêmeos idênticos (univitelinos), do sexo masculino, 23 anos de idade, naturais da cidade de Campinas, Estado de São Paulo (denominados JA e JR, ao longo do trabalho). A análise comparativa dos gêmeos JA e JR permitirá verificar em que medida a configuração anatômica impõe limites na variação de certos parâmetros acústicos, já que, presumìvelmente, gêmeos monozigóticos devem possuir características orgânicas bastante semelhantes, especialmente nessa faixa de idade.

3.2) *Material de Fala*

O grupo principal de 8 falantes (7 + R1/R2) leu um texto científico de 126 palavras (extraído de Vaz 1983; ver anexo, texto I). Optou-se por um texto de caráter neutro de modo a reduzir variações expressivas. O texto em questão é suficientemente longo para permitir a extração de aspectos de longo termo, sendo, ao mesmo tempo, compatível com o tamanho das amostras de fala normalmente encontradas na situação forense típica de Identificação de Falantes.

Todos os falantes do grupo principal leram o texto I em duas condições de velocidade de produção: normal e rápida. As instruções para cada tarefa foram as seguintes:

a) velocidade normal - "leia o seguinte texto da forma mais confortável e natural possível"

b) velocidade rápida - "leia o seguinte texto o mais rápido possível, procurando, entretanto, não prejudicar a inteligibilidade"

A leitura na velocidade rápida causou certas dificuldades para alguns falantes, provocando erros de hesitação e/ou omissão de fonemas e sílabas. Nesses casos a tarefa foi repetida até que a leitura se tornasse fluente e totalmente inteligível. Em vários experimentos foi feita uma análise comparativa das amostras do mesmo falante nas duas condições de velocidade, de modo a verificar a sensibilidade de determinados parâmetros a esse tipo de variação intra-falante.

Os gêmeos JA e JR leram um texto diferente dos falantes do grupo principal. O texto, de 238 palavras, foi extraído do noticiário esportivo de um jornal diário (v. anexo, texto II) e apresenta o mesmo caráter neutro do texto I. Os gêmeos JA e JR leram o texto II apenas na velocidade normal de produção (ver acima). Nos experimentos onde foi necessário comparar duas amostras de longo termo utilizou-se a primeira *versus* a segunda metade do texto II.

3.3) *Gravação*

As gravações foram realizadas em ambiente sem tratamento acústico especial, mantendo-se, entretanto, o nível de ruído de fundo mais baixo possível. Todos os falantes posicionaram-se no mesmo local na sala, guardando uma distância de aproximadamente 30 cm do microfone. O nível de gravação foi otimizado para cada falante, evitando-se *overflow*; o mesmo nível foi mantido para as gravações nas duas velocidades de produção. As gravações foram feitas em fita cassete normal (FUJI DR-I, IEC I/type I, 60 min.), em gravador marca Gradiente, modelo Esotech D-II, conectado a microfone dinâmico unidirecional Realistic, modelo 33984-C.

3.4) *Equipamento de Análise Acústica*

Para a maior parte das análises acústicas foi utilizado o Sonógrafo Digital da KAY Elemetrics DSP - 5500, do Laboratório de Fonética do Instituto de Estudos da Linguagem da Universidade Estadual de Campinas. Para algumas medidas empregou-se o sistema CSL (*Computerized Speech Lab*), também da KAY Elemetrics (modelo 4300), no Laboratório de Análise Sonora do Departamento de Medicina Legal da Universidade de Campinas.

Para a extração das medidas de amplitude e F0 através de LPC (*Linear Predictive Coding*), foi empregado o sistema da KAY Elemetrics, modelo 5635, acoplado a um micro-computador AT-386 (*clone* IBM).

As saídas impressas (espectrogramas, formas de onda, etc) foram produzidas em impressora térmica (KAY, modelo 5510), com 16 níveis de cinza e 75 DPI de resolução.

3.5) *Procedimentos Estatísticos*

Para os cálculos de natureza estatística foi utilizado o conjunto de programas BMDP, desenvolvido na Universidade da Califórnia. Esse pacote consiste de módulos independentes especificamente desenvolvidos para cada tipo de análise (distribuição, testes *t*, ANOVA, regressão, análise *cluster*, etc). Vários desses módulos foram utilizados ao longo do presente trabalho. Uma exposição mais detalhada de cada procedimento estatístico acompanhará a discussão de cada parâmetro acústico analisado.

SEÇÃO 4: FORMANTES (VOGAIS)

4.1) *Eficiência de medidas de Formantes para Identificação de Falantes*

Vários estudos têm verificado que as freqüências dos formantes de vogais produzidas por diferentes falantes apresentam uma considerável variabilidade inter-falante, mesmo em contextos fonéticos fixos (Peterson e Barney 1952; Behlau 1984). A constatação dessa variabilidade motivou uma série de estudos que procuram explicar de que modo - apesar das diferenças inter-individuais absolutas nos valores dos formantes - os ouvintes conseguem recuperar categorialmente o sistema vocálico da língua, independentemente do falante.

Lloyd (1890) já observara que as medidas de freqüência absoluta dos formantes não seriam suficientes para caracterizar vogais específicas (*apud* Shoup e Pfeifer 1976). Desde então, várias teorias têm sido propostas, baseadas em algum critério de normalização não diretamente dependente dos valores absolutos dos formantes. A maior parte desses estudos inspira-se no trabalho basilar de Joos (1948), que sugere ser a percepção das vogais baseada em uma espécie de calibração efetuada pelo ouvinte, onde a qualidade da vogal não é depreendida através de uma correspondência direta a um padrão fixo, mas sim a uma referência interna, estabelecida pelo próprio sistema vocálico do falante. Testes experimentais posteriores verificaram que a hipótese de Joos (1948) era, ao menos parcialmente, correta. Ladefoged (1967), observa que o julgamento categorial do ouvinte a respeito da qualidade da vogal em uma palavra-teste pode variar em função da freqüência média de F1 (sinteticamente manipulada) na sentença imediatamente precedente. Gerstman (1968) procura formalizar matematicamente o problema, normalizando os formantes de cada falante a partir dos máximos e mínimos individuais; assim, aos

valores máximo e mínimo de cada formante de cada falante atribuem-se os valores estandardizados 999 e 0 respectivamente, sendo todos os demais valores intermediários re-escalados linearmente com base nessa tessitura normalizada. Após a transformação, Gerstman obtém cerca de 97% de classificações corretas nas categorias vocálicas.

Outras abordagens incluem F0 na caracterização das vogais. Segundo esse ponto de vista, a percepção da qualidade vocálica não está vinculada apenas à estrutura de formantes, mas depende também de informação relacionada ao F0 local. Foulkes (1961), trabalhando com os dados de Peterson e Barney (1952), elabora uma transformação matemática que inclui o F0 local como fator de normalização, relatando um índice de 76 - 97 % de classificações automáticas corretas das categorias vocálicas, após a aplicação da transformação. Miller (1953), manipulando o F0 de vogais sintéticas, verifica que quando esse F0 é duplicado ocorre um deslocamento na categorização de algumas vogais, mesmo se o envelope espectral é mantido constante.

A questão da normalização perceptual no reconhecimento de vogais já foi tratada em estudos que incluem, além da informação de F0, transformações adicionais que utilizam relações internas entre os formantes e entre os formantes e F0. Miller (1989) propõe um modelo perceptual baseado em 3 coordenadas, x, y e z, onde x = Log (F3/F2), y = Log (F1/SR) e z = Log (F2/F1), sendo SR uma medida derivada da média geométrica local de F0. Syrdal e Gopal (1986) utilizam, em vez de razões, diferenças na escala BARK, baseada na divisão natural da faixa de freqüência pelo ouvido humano (Cf. Zwicker 1961); Syrdal e Gopal sugerem que as distâncias críticas entre formantes vizinhos, e entre F1 e F0, são usadas pelo ouvinte para classificar as vogais em categorias (v. também Traunmüller 1988) (para uma discussão mais extensa sobre a normalização perceptual de vogais ver Figueiredo 1990).

A existência de estratégias de normalização realizadas pelo ouvinte, no entanto, não invalida o fato de que esse mesmo ouvinte é capaz de extrair informação dependente do falante, apenas a partir dos valores absolutos das freqüências dos formantes. Na verdade, os modelos acima comentados, que pressupõem uma calibragem baseada em um referencial interno definido pelo sistema vocálico individual, sugerem que o ouvinte extrai primeiro informação do falante, antes de efetuar a normalização.

Vários experimentos têm demonstrado que estímulos contendo apenas informação relacionada ao trato vocal (filtro) permitem um índice de identificação do falante bem acima do acaso. Coleman (1973) utiliza uma laringe elétrica com F0 fixo, de modo a neutralizar a informação da fonte glotal, obtendo um acerto de 90 % na discriminação auditiva de um grupo de 20 falantes. Experimentos utilizando amostras consistindo de vogais isoladas em fala sussurrada (e, portanto, sem informação da fonte), têm verificado que as identificações auditivas baseadas nesses estímulos bastante limitados estão bem acima do acaso (La Rivière 1975; Tartter 1991).

Esses resultados sugerem que as freqüências dos formantes podem ser uma pista importante para a determinação da identidade do falante. A importância da estrutura de formantes na individualidade da voz é destacada em um engenhoso experimento conduzido por Kuwabara e Takagi (1991); nesse estudo, através da manipulação (re-síntese) de F0, formantes e larguras de banda de vogais em palavras *non-sense*, contatou-se que as alterações dos formantes provocam as distorções mais sensíveis no reconhecimento de vozes familiares aos ouvintes.

A informação dos formantes extraídos em núcleos vocálicos já foi utilizada em sistemas de Verificação Automática de Falante com sucesso. Paliwal (1984), a partir de um vetor quadri-dimensional composto pelas freqüências dos 4 primeiros

formantes de vogais em palavras /hVd/, obtém um alto índice de identificações automáticas corretas, por meio de uma métrica baseada na distância Euclidiana entre os vetores.

4.2) *Variações Intra-Falante no Espectro de Vogais*

Assim como qualquer outro aspectro acústico da fala, os formantes de um fonema vocálico não são estáveis ao longo de todas as condições de produção. Fatores como velocidade de emissão, contexto fonético e acentuação podem exercer uma influência considerável nas características espectrais de uma vogal. A extensão e regularidade dessas variações, no entanto, não são de fácil determinação, e os estudos experimentais focalizando esse tema não apresentam resultados totalmente consistentes entre si.

Uma das primeiras formulações a respeito da questão encontra-se em Lindblom (1963). Nesse clássico estudo, Lindblom sugere que existe uma relação quase sistemática entre a duração de uma vogal e suas características espectrais. Assim, em determinadas condições, um fonema vocálico será *reduzido*, afastando-se de sua realização canônica e aproximando-se de uma vogal neutra no centro do triângulo vocálico; teoricamente, a redução seria tanto maior quanto menor fosse a duração do núcleo vocálico. Lindblom introduz o conceito de *target undershoot*, segundo o qual, embora possa existir um déficit acústico, a "intenção" do falante subjacente à produção da vogal é sempre a mesma, independentemente das circunstâncias contextuais. Em um trabalho posterior (Lindblom e Studdert-Kennedy 1967) é introduzida a noção complementar de *perceptual overshoot*, segundo a qual o ouvinte compensaria perceptualmente o *undershoot* articulatório, de modo a recuperar o *target* intencionado - mas não realizado - pelo falante.

O modelo proposto por Lindblom (1963) prevê um certo grau de redução vocálica em taxas mais altas de emissão. Esse aspecto, entretanto, não é confirmado em outros experimentos. Gay (1978a), por exemplo, não observa alteração significativa nos dois primeiros formantes de vogais em fala rápida. Outro estudo (van Son e Pols 1990) observa uma pequena alteração em F1 na fala rápida (Holandês), mas cuja magnitude está próxima do limite de discriminação diferencial para esse formante.

Resultados posteriores mostraram que o modelo original de *target undershoot* era simples demais, e não se poderia estabelecer uma relação direta entre a duração da vogal e a redução, ou seja, a duração apenas não seria suficiente para prever o grau de *undershoot*. Na verdade, os falantes têm liberdade para variar o grau de *undershoot*, independentemente da duração efetiva da vogal. Desde que instruídos para isso, os falantes podem atingir os alvos articulatórios e acústicos, apesar das durações curtas das vogais (em função da maior velocidade de fala e/ou de condições de tonicidade) (Engstrand 1988; Lindblom e Moon 1988; Moon 1991).

Há evidências de que os ajustamentos realizados pelo falante na fala rápida sejam alcançados através de estratégias articulatórias diferentes, dependentes do falante. Sonoda (1987), estudando a dinâmica articulatória dos movimentos mandibulares em função de mudanças na velocidade de fala, observa que, em palavras *non-sense* do tipo /$V_1V_2V_3$/, um dos falantes ajusta consideravelmente a velocidade dos movimentos mandibulares na fala rápida, enquanto o outro não, o que causa um maior grau de *undershoot* para o segundo falante.

O controle das variações impostas pelos efeitos co-articulatórios é uma das maiores dificuldades para o uso de medidas de formantes para a Identificação de Falantes, especialmente se considerarmos que esses efeitos podem se difundir até mesmo a partir de segmentos não imediatamente adjacentes (Öhman 1966). Por outro lado, a existência de variação intra-falante em um parâmetro qualquer não

impede sua utilização, desde que a variação **inter-falante** seja de maior magnitude, sendo esse o princípio fundamental para a validação de uma determinada medida a ser aplicada no reconhecimento de falantes (Wolf 1972; Atal 1976; Rosenberg 1976). No que diz respeito às freqüências de formantes vocálicos, há evidências de que esse seja o caso; Broad e Fertig (1970) observam que as repetições de uma vogal em vários contextos fonéticos pelo mesmo falante definem no espaço acústico um *cluster* mais denso do que aquele produzido pelas variações inter-falante.

4.3) *Medidas derivadas dos Formantes Vocálicos: Análise Estatística*

A presente seção descreve um experimento onde se procura investigar a eficiência de medidas derivadas dos formantes de vogais selecionadas como indicadores da identidade do falante. Além das freqüências foram também medidas a amplitude de cada formante e as inclinações (*slopes*) definidas por pares ou conjuntos de formantes.

4.3.1) *Material e Métodos*

Para essa fase do experimento foram empregadas amostras do grupo principal de falantes (n=9; 7 + R1/R2), baseadas na leitura do texto I (ver anexo) em duas condições de velocidade de produção (para maiores detalhes metodológicos ver seção 3).

4.3.1.1) *Seleção de Vogais*

Foram selecionadas apenas as vogais tônicas (n=75; texto I, v. anexo) no nível lexical. Tal procedimento evita em parte o efeito neutralizador da redução vocálica nas sílabas átonas, o que poderia inserir um fator adicional de dispersão para a caracterização de espaços vocálicos pessoais. Além disso, o uso exclusivo de vogais tônicas aproxima o experimento da situação forense real, onde, freqüentemente, as limitações do meio (gravador, ruído ambiental, etc) permitem acesso seguro apenas a medidas baseadas em trechos com maior razão sinal/ruído.

As vogais selecionadas foram classificadas segundo uma transcrição larga em oito grupos:

/a/, /ɛ/, /e/, /i/, /O/, /o/, /u/, /ã/

Os exemplos abaixo mostram algumas palavras do texto com as respectivas vogais lexicalmente tônicas tal como classificadas pelo critério utilizado:

/a/	reativid/a/de, c/a/da, fix/a/r, determin/a/do, etc
/ɛ/	c/ɛ/lulas, ad/ɛ/re, p/ɛ/le, etc
/e/	tamb/e/m, difer/e/ntes, s/e/r, /e/ste, etc
/i/	fabr/i/cam, f/i/gado, s/i/, segu/i/nte, etc
/O/	linf/O/citos, antic/O/rpos, c/O/pias, etc
/o/	/o/utros, recept/o/res, cl/o/nes, /o/ito, etc
/u/	/u/m, estrut/u/ras, d/u/as, conj/u/nto, etc
/ã/	s/ã/ngue, membr/ã/na, gar/ã/ntem, subst/ã/ncias, etc

Observe-se que, com exceção das vogais /a/-/ã/, não se fez a distinção nasal-não nasal, mesmo quando o contexto fonético implica nasalização, como em cl/o/ne ou difer/e/nte. Há dois motivos para esse procedimento: em primeiro lugar, sabe-se que a nasalidade afeta mais pronunciadamente a vogal aberta /a/, provocando, em geral, uma centralização dessa vogal, quando nasalizada; as demais vogais têm alterações espectrais menos relevantes quando em contexto nasal (Cf. Wright 1986); além disso, a criação de categorias adicionais para explicitar a condição de nasalidade em todas as vogais, formaria - em função da própria estrutura do *corpus* disponível - grupos com número muito reduzido de elementos.

Para a seleção das vogais utilizadas no experimento não houve preocupação em obter um número igual de observações para cada qualidade. O único critério foi a seleção de vogais em sílabas tônicas (lexicalmente). Esse procedimento, embora implique um certo desbalanceamento estatístico, reflete em parte a própria distribuição das categorias vocálicas na língua, aproximando o experimento da

situação forense real, onde nunca é possível obter amostras equilibradas das diferentes qualidades vocálicas.

O número de observações para cada grupo vocálico, totalizando 75 casos (em **cada** condição de produção: velocidade normal *vs.* rápida), está listado abaixo:

VOGAL	n=
/a/	12
/ɛ/	9
/e/	15
/i/	11
/O/	8
/o/	9
/u/	6
/ã/	5
total	75

4.3.1.2) *Medidas de Freqüências de Formantes*

A determinação de uma medida "exata" para um formante não é uma questão trivial. Procurou-se aqui apenas adotar uma metodologia que garantisse um critério válido de comparação entre os falantes. Considerou-se como medida de formante (freqüência e amplitude) aquela tomada com base no espectro de seção de banda larga (BW=300Hz) a partir de uma janela de aproximadamente 10-20 ms localizada em uma região do núcleo vocálico considerada como *target*. Definiu-se como *target*, prioritariamente, o ponto onde, dependendo da vogal, localiza-se o máximo ou mínimo apropriado de um determinado formante, considerando a posição da vogal no espaço F1 X F2. Assim, as realizações das vogais /a,O,E,ã/ foram medidas no ponto de máximo F1, as vogais /u,o/ no ponto de mínimo F2, e as vogais /i,e/ no ponto de máximo F2. Com esse critério procurou-se encontrar um ponto ideal mais próximo da qualidade "canônica" da vogal em questão (Cf. van Son e Pols 1990).

Em alguns (poucos) casos, entretanto, não foi possível utilizar o método de máximos/mínimos de F1 e F2. Especialmente em ambientes envolvendo ditongos (como "oito"), tritongos ("iguais") ou líquidas ("células"), pode ser difícil determinar um *target* , devido à natureza dos movimentos transicionais. Nesses casos optou-se por um dos seguintes métodos (nessa ordem de prioridade): (a) localizar o ponto de máxima energia no núcleo vocálico, (b) localizar a região mais estacionária do núcleo vocálico ou (c) localizar a região mais central do núcleo vocálico. Eventualmente, mais de uma dessas condições pôde ser atendida. Convém ressaltar que, em todos os casos houve sempre monitoração auditiva para avaliar a qualidade vocálica, além de um acompanhamento visual simultâneo no espectro no tempo (espectrograma tradicional de banda larga) de modo a checar o centro exato da ressonância.

Mesmo com os cuidados tomados nas medidas dos formantes, é preciso levar em conta que o espectro de seção de banda larga não reflete diretamente os picos da função de transferência. Esses picos só podem ser obtidos por aproximação, e o erro será tanto maior quanto maior for a freqüência fundamental do falante (para uma discussão extensa do assunto v. Fant 1962, Lindblom 1962).

Em geral, o erro máximo para medidas feitas diretamente no espectro de seção de banda larga é F0/2, mas deve ser menor que isso na maior parte dos casos. Para falantes do sexo masculino teremos, pois, um erro **máximo** de aproximadamente 50-60 Hz, o que é considerável para o F1 de vogais altas , mas praticamente desprezível para o F2 da maioria das vogais e para todos os formantes acima de F3 de todas as vogais. Avaliou-se que o erro implícito no procedimento utilizado não distorceu as medidas como um todo, já que o número considerável de observações tende a anular essa variação aleatória, fazendo com que o erro **médio** seja bem menor do que 50-60 Hz. Nas vogais altas, para diminuir o erro, foram

utilizados espectros de banda estreita, de modo a calcular aproximadamente o ponto exato de F1 a partir dos componentes harmônicos individuais.

Embora correndo o risco da circularidade, é preciso abordar as medidas de formantes com alguma expectativa em relação às regiões onde devem ser encontrados os formantes de cada vogal, de modo a evitar a inclusão de formantes "espúrios", isto é, picos espectrais que não correspondem às freqüências de ressonância "naturais" de uma vogal. Esses picos adicionais podem surgir em conseqüência da influência de zeros glotais provocados pelo acoplamento fonte/filtro. Embora esses picos espectrais extras possam, por si só, estabelecer diferenças entre falantes, eles não foram considerados no contexto do presente experimento, em função de sua inerente instabilidade.

4.3.1.3) *Medidas de Amplitude dos Formantes*

Foram medidas, diretamente do espectro de seção de banda larga (v. 4.3.1.2) as amplitudes em dB de cada um dos formantes visíveis no espectro. O ponto exato para a aferição da amplitude do pico espectral foi exatamente o mesmo que o usado para aferir a freqüência dos formantes (v. 4.3.1.2).

As medidas de amplitude extraídas diretamente do espectro de seção não podem ser diretamente comparáveis entre falantes, já que dependem em parte de fatores ligados a características do meio, tais como: nível de entrada do gravador, nível de entrada do sonógrafo durante a aquisição e distância falante/microfone (não controlada rigorosamente no experimento). Além disso, o mais relevante quanto às amplitudes dos formantes são seus valores **relativos**. De modo a preservar as relações entre as amplitudes dos diferentes formantes, e ao mesmo tempo estandardizar as medidas para diferentes falantes efetuou-se uma normalização onde

o nível do primeiro formante (L1) foi igualado a 60 dB para todos os falantes e todas as vogais. Os demais níveis foram então ajustados em relação a essa referência fixa de 60 dB. Assim, se L1 é a amplitude do primeiro formante, então:

$L1_{normal}$= 60 dB

$L2_{normal}$= (60-L1+L2) dB

Ln_{normal}= (60-L1+Ln) dB

4.3.1.4) *Slopes de retas interpoladas aos picos espectrais*

É provável que as diferenças entre amplitudes de diferentes formantes no espectro de uma vogal reflita certos aspectos individuais do falante. O decaimento global do espectro da fonte, a eventual presença de zeros glotais, características particulares de *damping* em certas faixas de freqüência, etc, refletem-se de algum modo no espectro.

A simples diferença entre amplitudes talvez não seja um parâmetro ideal, na medida em que as freqüências dos formantes envolvidos não são representadas. Uma medida melhor é um número que represente a queda dB/oitava entre os picos. Um modo simples de obter essa medida é achar a tangente do ângulo da reta que une os dois picos espectrais em questão; se o eixo das freqüências está expresso em escala logarítmica de base 2, essa tangente dá diretamente a queda em dB/oitava entre os dois picos. A figura 4.1 ilustra esse procedimento. Na figura, a tangente de α, fornece a queda em dB/oitava entre F2 e F3.

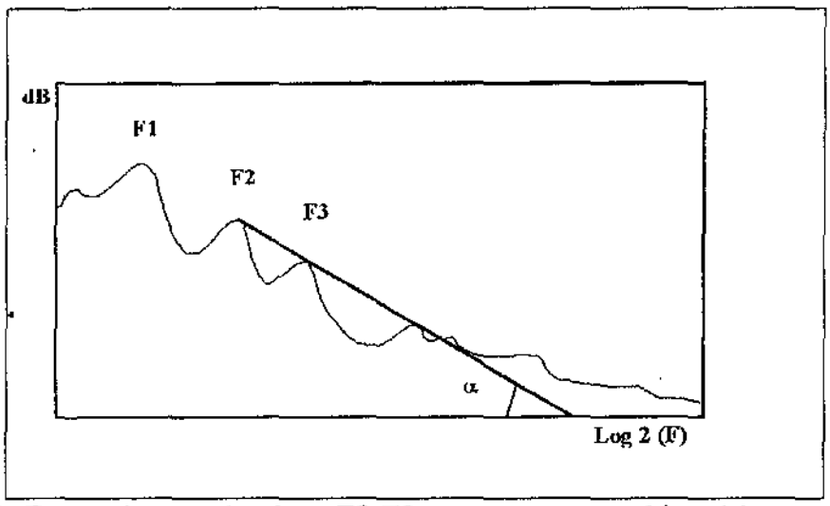


FIGURA 4.1: Determinação do *slope* F2-F3 em um espectro hipotético; o slope é dado pela tangente de α, com a abscissa expressa em escala LOG (LOG_2 [F_{Hz}]).

Dessa forma foram obtidos os *slopes* das retas passando por todos os pares de formantes no espectro. Assim:

Sl_{12} = *slope* da reta que passa por F1 e F2

Sl_{13} = *slope* da reta que passa por F1 e F3

Sl_{mn} = *slope* da reta que passa por Fm e Fn

Além desses *slopes*, foram também calculados os *slopes* das retas de regressão (método dos quadrados mínimos) ajustadas a mais de dois formantes. A figura 4.2 ilustra o procedimento. Na figura, a tangente de β, a reta ajustada a F1,F2 e F3 dá a queda em dB/oitava. Assim:

Sl_{123} = *slope* da reta ajustada a F1,F2 e F3

Sl_{234} = *slope* da reta ajustada a F2,F3 e F4

$Sl_{1,2,3,4}$ = *slope* da reta ajustada a F1,F2,F3 e F4

$Sl_{k,l,m,n}$ = *slope* da reta ajustada a Fk,Fl,Fm e Fn

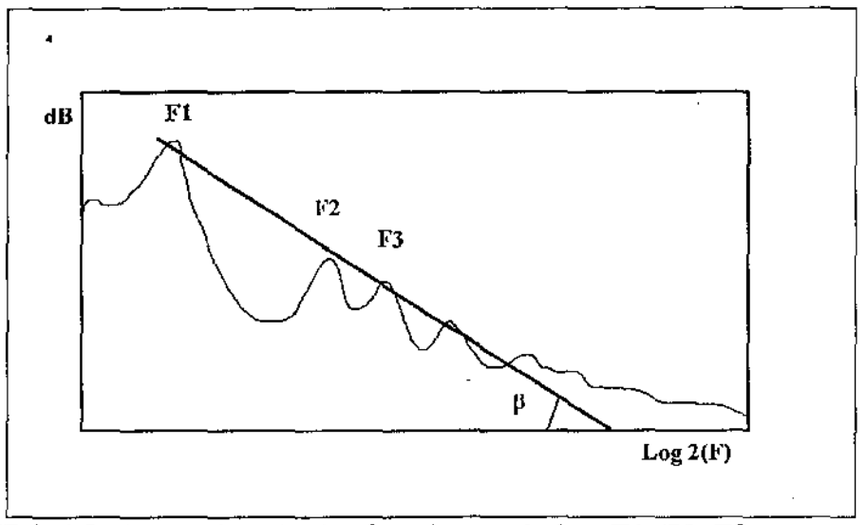


FIGURA 4.2: Determinação do *slope* interpolado a F1, F2, F3 em um espectro hipotético. O *slope* é dado pela tangente do ângulo β, com a abscissa expressa em escala LOG (LOG_2 [F_{Hz}]).

4.3.2) *Resultados*

A tabela 4.1 apresenta resultados de estatística descritiva básica (programa BMDP-2D) para as variáveis contínuas. Nessa tabela todos os casos do grupo principal de falantes foram incluídos (n=9; 7+R1/R2), agrupando-se assim todas as condições (velocidade de produção, falante e qualidade da vogal). A tabela apresenta os seguintes dados estatísticos:

Média aritmética (**MED**)

Mediana (**MEDN**) (valor que divide 50% da distribuição)

Moda (**MODA**) (valor mais freqüente)

Skewness (**SK**) (grau de assimetria vertical da distribuição)

Curtose (**KU**) (grau de achatamento da distribuição)

Desvio-padrão (**DP**)

Na tabela 4.1 foram incluídos os resultados a partir das freqüências de formantes após uma transformação logarítmica (Flog1, Flog2, etc). Para maior clareza, esses resultados foram reconvertidos para Hz.

	MED	MEDN	MODA	DP	SK	KU
F1	448.0	440.0	320.0	140.7	.38	-.94
F2	1559.4	1520.0	1040.0	439.2	.03	-1.11
F3	2492.7	2480.0	2520.0	184.2	.25	.44
F4	3577.6	3580.0	3680.0	234.6	-.33	.41
Flog1	426.0	439.8	319.8	---	-.03	-1.14
Flog2	1488.8	1520.1	1038.3	---	-.35	-1.00
Flog3	2469.5	2469.4	2503.9	---	.01	.10
Flog4	3565.8	3565.9	3666.0	---	-.59	1.22
L1	59.3	59.0	58.0	4.5	-.38	.74
L2	45.1	45.0	43.0	7.8	-.26	-.28
L3	38.3	39.0	39.0	8.1	-.31	-.23
L4	40.7	41.0	41.0	8.5	-.26	-.39
Sl12	-8.3	-7.5	não única	3.7	-1.10	2.08
Sl13	-8.6	-8.3	-6.0	3.3	-.35	-5.10
Sl14	-6.2	-6.0	0.0	2.8	-.37	-.12
Sl23	-8.6	-9.4	0.0	12.0	.87	4.62
Sl123	-8.6	-8.3	não única	3.3	-.50	-.05
Sl234	-3.3	-3.7	0.0	8.2	.27	1.67
Sl1234	-7.0	-6.8	não única	3.0	-.10	1.95

TABELA 4.1: Estatística descritiva básica das principais variáveis contínuas, agrupando todos os falantes (grupo principal; n=9: 7 + R1/R2) e todas as variáveis categoriais (vogal e velocidade de emissão).

Podemos observar, na tabela 4.1, que as distribuições relativas aos formantes não se afastam fortemente da normalidade, tanto na assimetria (*Skewness*) quanto no achatamento (Curtose). A transformação em escala LOG não altera significativamente as características das distribuições; as menores médias observadas na escala LOG são decorrência da diminuição da assimetria positiva em relação à escala Hz. As distribuições de F3 e F4 tendem a ser leptocúrticas, isto é, com muitos valores concentrados em uma região estreita (Cf. Lutz 1983); esse aspecto está relacionado ao fato de os formantes altos sofrerem uma menor variação em função da qualidade da vogal (lembramos que nessa tabela todas as vogais foram reunidas).

É interessante observar que as médias globais dos formantes, se aproximam de uma vogal "neutra", com espaçamento de cerca de 1000 Hz entre os formantes contíguos. Essa vogal sintética, /ə/, corresponderia aproximadamente às ressonâncias naturais de um tubo reto com cerca de 17.6 cm de comprimento, da mesma ordem de grandeza do trato oral de um homem adulto (Fant 1960). Da mesma forma, observamos que o *slope* médio entre F1 e F4 (SL14), fica em torno de -6 dB/oitava, o decaimento previsto para o espectro da vogal neutra, segundo a teoria acústica da produção da fala (Fant 1960).

A tabela 4.2 dá os resultados de todas as variáveis contínuas para cada falante do grupo principal (n=9; 7 + R1/R2), reunindo ainda todas as vogais e as duas condições de velocidade de produção. De modo a verificar a eficiência de cada uma das variáveis contínuas para a distinção dos falantes do grupo, foi realizada uma análise de variância (ANOVA, programa BMDP-7D) para cada variável separadamente. Os resultados de ANOVA estão listados nas tabelas 4.3a e 4.3b; o valor de F, juntamente com o nível de significância (p<) dão uma estimativa da força de cada variável contínua em explicar a variância inter-falante (um maior valor de F indica que a variável em questão varia significativamente inter-falantes). Também

nas tabelas 4.3a e 4.3b são apresentados os resultados de um teste estatístico de comparação de médias (método *Student-Newman-Keuls*, programa BMDP-7D); nesse teste, as médias individuais são comparadas para cada par de falantes e linhas horizontais unem os falantes que **não** são significativamente diferentes quanto à variável em questão (os falantes são ordenados em médias crescentes, da esquerda para a direita.

A tabela 4.3a indica que apenas F2 (tanto em escala Hz quanto em LOG) não estabelece qualquer diferença significativa entre os falantes. Esse resultado está relacionado ao fato de F2 ser o principal determinante da qualidade vocálica; como todos as categorias vocálicas estão reunidas, a variância intra-falante em função da vogal praticamente anula a variância inter-falante (mais adiante examinaremos o efeito isolado da vogal na variância).

F1, embora também seja um fator importante na determinação da qualidade vocálica, apresentou um valor de F significativo, embora não muito expressivo. Nesse sentido, é interessante examinar as *funções de sensitividade*, definidas em Fant (1980). Essas funções estabelecem uma relação entre a energia associada a um formante e as diferentes regiões do trato; no caso de F1, a energia está distribuída mais ou menos por igual ao longo do trato como um todo (com a possível exceção de /u/; v. Fig. 12 em Fant 1980:72), fazendo com que F1 seja, em grande parte, dependente da extensão total do trato, um aspecto anatômico que pode variar consideravelmente entre os falantes.

Fal. →	ZR	EN	R1	ZP	AG	WA	MS	R2	DO
F1	454.7	435.7	464.2	480.1	412.3	457.1	419.3	482.4	426.1
	148.7	131.7	126.1	137.5	127.7	151.4	160.0	131.1	133.4
F2	1524.1	1601.4	1582.2	1566.9	1583.1	1533.1	1568.3	1586.0	1491.8
	435.4	456.9	447.7	413.0	469.4	428.7	492.9	436.9	361.7
F3	2456.2	2638.6	2478.8	2337.6	2589.6	2553.9	2466.9	2493.7	2415.7
	181.7	148.6	146.1	164.3	124.6	183.3	198.5	157.7	151.8
F4	3645.1	3757.0	3785.2	3411.2	3473.0	3364.8	3588.1	3657.9	3510.3
	130.6	218.8	184.4	193.9	203.0	164.2	161.2	272.9	167.8
Flog1	431.4	416.4	446.0	461.1	392.9	432.0	391.0	464.3	405.3
Flog2	1458.2	1530.7	1513.8	1512.8	1509.6	1469.4	1485.8	1522.2	1448.1
Flog3	2449.0	2633.9	2474.6	2331.4	2586.9	2547.7	2459.2	2488.4	2411.9
Flog4	3643.2	3750.8	3779.6	3406.3	3468.2	3361.7	3585.6	3648.3	3506.9
Sl12	-8.43	-6.23	-10.91	-7.54	-10.34	-6.66	-7.91	-8.56	-7.97
	3.34	2.83	4.83	2.86	3.77	2.91	2.94	3.54	2.58
Sl13	-9.08	-6.00	-11.18	-9.33	-8.64	-6.08	-8.13	-10.11	-8.81
	2.83	2.77	3.04	3.26	2.73	2.69	2.87	3.32	2.61
Sl14	-5.81	-4.49	-6.55	-9.70	-8.13	-3.75	-4.75	-7.93	-5.72
	2.37	2.28	1.82	2.52	2.09	2.14	2.18	2.23	2.27
Sl23	-11.30	-3.7	-15.72	-11.32	-3.64	-3.45	-4.46	-14.38	-9.55
	9.91	9.53	9.90	9.45	16.42	8.76	17.01	8.11	7.49
Sl24	-1.62	-.996	-1.79	-12.43	-4.83	+1.29	+1.35	-7.35	-1.27
	7.03	5.40	5.12	5.63	6.53	6.82	7.3	7.08	6.38
SL34	+8.48	+4.53	+11.60	-10.47	-5.45	+9.98	+10.70	+1.31	+10.98
	10.49	10.74	10.31	11.12	13.12	10.44	8.27	16.26	26.0
SL123	-9.12	-6.13	-10.98	-9.28	-9.02	-6.27	-8.12	-9.99	-8.79
	2.94	2.64	3.33	3.33	2.49	2.58	2.84	3.58	2.74
SL234	-1.79	-1.13	-2.1	-11.55	-4.82	+.843	+1.09	-9.49	-1.67
	7.21	5.8	5.71	8.01	6.68	7.06	7.48	8.53	6.80
SL1234	-6.76	-5.04	-8.06	-9.53	-8.30	-4.41	-5.57	-8.61	-6.53
	2.25	2.26	2.45	2.96	2.01	2.37	2.51	2.72	2.76

TABELA 4.2: Médias (linha superior) e desvios-padrão (linha inferior) das principais variáveis contínuas para cada falante separadamente. As médias em LOG dos formantes foram reconvertidas em Hz, para maior clareza.

F1	F= 5.08	p< .0001
AG MS DO EN ZR WA R1 ZP R2		
F2	F= .98	NS
DO ZR WA ZP MS R1 AG R2 EN		
F3	F= 43.79	p< .0001
ZP DO ZR MS R1 R2 WA AG EN		
F4	F= 84.37	p< .0001
WA ZP AG DO MS ZR R1 EN R2		
Flog1	F= 6.26	p< .0001
MS AG DO EN ZR WA R1 ZP R2		
Flog2	F= .69	NS
DO ZR WA ZP MS R1 AG R2 EN		
Flog3	F= 44.96	p< .0001
ZP DO ZR MS R1 R2 WA AG EN		
Flog4	F= 79.83	p< .0001
WA ZP AG DO MS ZR R1 EN R2		

TABELA 4.3a: Resultados de ANOVA e do teste de comparação de médias *Student-Newman-Keuls* (BMDP-7D). Um valor estatisticamente significativo de F indica que a variável em questão explica parte da variabilidade inter-falante. No teste de comparação de médias, as linhas horizontais unem os falantes que **não** são estatisticamente distintos quanto à variável estudada (os falantes estão dispostos em ordem crescente de médias, da esquerda para a direita)

SL12	F= 30.33	p< .0001
R1 AG	R2 ZR DO MS ZP	WA EN
SL13	F= 47.70	p< .0001
R1 R2	ZP ZR DO AG MS	WA EN
SL14	F= 99.55	p< .0001
ZP AG R1	R2 ZR DO MS	EN WA
SL23	F= 26.27	p< .0001
R1 R2	ZP ZR DO MS EN AG	WA
SL24	F= 58.53	p< .0001
ZP R2 AG	R1 ZR DO EN	WA MS
SL34	F= 39.67	p< .0001
ZP AG R2 EN	ZR WA MS DO	R1
SL123	F= 42.51	p< .0001
R1 R2	ZP ZR AG DO MS	WA EN
SL234	F= 55.62	p< .0001
ZP R2 AG	R1 ZR DO EN	WA MS
SL1234	F= 73.33	p< .0001
ZP R2 AG	R1 ZR DO MS	EN WA

TABELA 4.3b: Resultados de ANOVA e do teste de comparação de médias *Student-Newman-Keuls* (BMDP-7D). Um valor estatisticamente significativo de F indica que a variável em questão explica parte da variabilidade inter-falante. No teste de comparação de médias, as linhas horizontais unem os falantes que **não** são estatisticamente distintos quanto à variável estudada (os falantes estão dispostos em ordem crescente de médias, da esquerda para a direita)

Os formantes mais altos, F3 e F4, apresentam valores altos de F. Esse resultado está dentro das expectativas, na medida em que esses formantes variam pouco com a vogal (especialmente F4) e estão mais fortemente associados a características anatômicas do trato. F3 e F4 permitem uma separação razoavelmente eficiente dos falantes, formando 6 grupos distintos (em 9 possíveis). No entanto, as duas produções não-contemporâneas do mesmo falante (R1/R2) são estatisticamente distintas em F4; a média desse formante sofre uma queda de 3.5% na segunda gravação (R2; v. tabela 4.2), quando o falante já não se encontrava mais sob estado gripal (como na primeira gravação, codificada como R1). De acordo com Fant (1980:72),

> *F4 of all vowels has a substantial peak of energy located in the larynx tube.*

Assim, é possível que a inflamação no nível da laringe tenha, de algum modo, afetado o comportamento de F4, o único formante que sofreu alteração significativa em função do estado gripal.

A transformação em escala LOG praticamente não alterou os resultados de ANOVA (valores de F). Os grupos estatisticamente distintos, no entanto são um pouco diferentes no primeiro e terceiro formantes, se comparados com os obtidos a partir da escala Hz. No terceiro formante, por exemplo, as médias dos falantes WA e AG são significativamente diferentes quando é utilizada a escala LOG, enquanto, na escala Hz, os mesmos falantes são reunidos em um único grupo.

A tabela 4.3b apresenta os resultados para os *slopes* calculados a partir das amplitudes relativas de dois ou mais formantes. Todos os *slopes* têm valores significativos de F. Os valores mais expressivos são dos *slopes* que incluem as

amplitudes de F1 e F4 (SL14, F= 99.55, p< .0001; SL1234, F= 73.33, p< .0001). Os *slopes* que incluem as amplitudes de F2 e/ou F3 têm em geral um maior desvio-padrão intra-falante (v. tabela 4.2), provavelmente em função da maior variabilidade inter-vogais. Aparentemente, SL14 reflete bem o *tilt* espectral médio do falante, um aspecto que pode fornecer alguma informação a respeito de certas características da fonte, como o grau de *breathiness*, por exemplo (Hammarberg *et al.* 1986; Klatt e Klatt 1990; Childers e Lee 1991). Mais adiante (v. seção 6), estudaremos a variabilidade inter-falante de *slopes* extraídos do espectro de longo termo (ELT), uma medida mais estável do que os decaimentos obtidos a partir das amplitudes relativas dos formantes. É importante ressaltar, entretanto, que parâmetros baseados no *tilt* espectral são extremamente sensíveis às condições do meio de transmissão/captação do sinal, sendo seu emprego condicionado à possibilidade de reproduzir essas condições no momento da comparação entre os enunciados teste e referência, uma limitação que torna mais viável a utilização desses parâmetros nos modelos de Verificação Automática de Falantes, já que, na situação forense típica, nem sempre é possível simular exatamente as mesmas condições da gravação original.

4.3.2.1) *Efeito da Velocidade de Emissão na Variação dos Parâmetros Acústicos*

A presente seção examinará as eventuais alterações dos parâmetros acústicos estudados, em função das diferentes condições de velocidade de emissão (normal *vs.* rápida; para uma descrição da metodologia ver seção 3)

A tabela 4.4 resume os resultados de uma análise de variância (ANOVA; BMDP-7D) comparando os grupos velocidade NORMAL *versus* RÁPIDA. O teste examina apenas o efeito isolado da velocidade, integrando todos os falantes e todas as vogais.

Para os formantes, o único efeito estatisticamente significativo foi em F4 ($F=5.25$; $p<.02$), observando-se um pequeno aumento de cerca de 1% na freqüência média, na velocidade rápida (os efeitos para os formantes em escala LOG foram semelhantes aos em escala Hz e não estão representados na tabela). F1 sofre um aumento médio de cerca de 2.5%, embora não significativo, estatisticamente. Um pequeno aumento de F1 na fala produzida em taxas mais rápidas também foi observado em van Son e Pols (1990), para o Holandês; nesse estudo, os autores sugerem que essa variação de F1 pode estar vinculada ao maior esforço vocal envolvido na produção da fala rápida, observado no falante examinado. Sabe-se que variações no esforço vocal podem alterar consideravelmente alguns movimentos articulatórios e, conseqüentemente, os formantes de vogais (Schulman 1989; uma maior abertura mandibular teria, de fato, o efeito de aumentar F1 (Fant 1960). No caso do grupo de falantes aqui estudado, podemos observar que, assim como relatado em van Son e Pols (1990), há um aumento consistente na amplitude média durante a produção da fala rápida; esse aspecto pode ser constatado examinando-se as amplitudes não normalizadas dos formantes (L1, L2, L3, L4) na tabela 4.4, onde se verifica um aumento médio, estatisticamente significativo, de cerca de 3 dB na velocidade rápida (lembramos que o nível de entrada de gravação foi mantido igual para as duas condições de velocidade, assim como a distância falante/microfone).

O fato de o aumento de amplitude na fala rápida ser da mesma magnitude para todos os formantes, parece indicar que o maior esforço vocal provoca um aumento igualmente distribuído sobre todo o espectro; a ausência de diferenças significativas entre os níveis normalizados (resumidos como LNnorm, na tabela 4.4) e entre qualquer um dos *slopes* (resumidos como SLmn) apontam na mesma direção.

A ausência de distinções significativas para os *slopes* indica que o *tilt* espectral não se altera substancialmente na fala rápida. Pelo menos no grupo de falantes aqui estudado, o maior esforço vocal, portanto, não parece estar associado a

uma alteração importante no nível da onda glotal (*strained voice*), já que uma modificação dessa natureza implicaria um *slope* menos abrupto (Cf. Kitzing 1986). A estabilidade das medidas de decaimento espectral em diferentes condições de velocidade de emissão, aliada à alta variância inter-falante (v. tabela 4.3b), torna esse tipo de parâmetro extremamente atraente para a Identificação de Falantes (desde que mantidas as mesmas condições de transmissão/captação do sinal). Mais adiante (v. seção 6) estudaremos a variação de *slopes* extraídos de diversas faixas do espectro de longo termo (ELT).

VAR.↓	V.↓	n =	média	D.P.	F =	p <
F1	N	665	443.1	143.6	NS	NS
	R	661	453.1	137.7		
F2	N	656	1559.9	453.3	NS	NS
	R	659	1558.8	425.2		
F3	N	631	2494.6	182.0	NS	NS
	R	625	2490.7	186.6		
F4	N	617	3562.3	236.2	5.25	.02
	R	614	3592.9	232.3		
SLmn	-	-	-	-	NS	NS
L1	N	665	58.1	4.46	98.37	.0001
	R	661	60.5	4.3		
L2	N	656	43.8	7.9	40.00	.0001
	R	659	46.5	7.6		
L3	N	631	37.0	8.4	33.23	.0001
	R	625	39.6	7.6		
L4	N	617	39.2	8.5	43.62	.0001
	R	614	42.3	8.1		
LNnorm	-	-	-	-	NS	NS

TABELA 4.4: Resultados de ANOVA (BMDP-7D) testando os efeitos da velocidade de emissão nas variáveis contínuas (NS = não significativo).

4.3.2.2) *Efeitos de Interação entre Variáveis Categoriais*

Nosso *corpus* foi estruturado de forma a possuir apenas três variáveis classificatórias: FALANTE, VELOCIDADE DE EMISSÃO e VOGAL. Será importante examinar se existe algum efeito de interação entre essas classes. O programa BMDP-2V cria modelos de análise de variância que permitem avaliar efeitos isolados de variáveis categorias juntamente com possíveis interações entre essas variáveis. A tabela 4.5 resume resultados de testes estatísticos para efeitos isolados e interações entre FALANTE X VELOCIDADE, FALANTE X VOGAL e VOGAL X VELOCIDADE. É importante observar que a influência de cada categoria deve ser entendida apenas dentro de cada modelo; assim, por exemplo, o valor de F para FALANTE (isolado) é naturalmente diferente para cada modelo combinado (FALANTE/VELOCIDADE e FALANTE/VOGAL). Constam da tabela os formantes e SL14, o *slope* que se mostrou mais eficiente para a separação dos falantes.

Na tabela 4.5, podemos verificar que, entre os formantes, F1 e F2 não apresentam interação significativa FALANTE X VELOCIDADE, nem efeitos isolados de VELOCIDADE, nesse modelo (FALANTE X VELOCIDADE). Há um pequeno efeito isolado de FALANTE para F1 ($F=3.5$; $p< .0005$), confirmando o já observado na tabela 4.3a. F3 e F4 revelam um efeito considerável de FALANTE, também dentro das expectativas, mas há também um efeito significativo da interação FALANTE X VELOCIDADE, indicando que a variação desses formantes, em função da velocidade, não se dá de forma homogênea para todos os falantes do grupo estudado.

VAR	F1	F2	F3	F4	SL14
Falante	3.5/.0005	NS	45.8/.0001	87.6/.0001	100.9/.0001
Velocidade	NS	NS	NS	7.8/.005	NS
Fal. X Vel.	NS	NS	3.8/.0002	3.9/.0001	3.1/.002
Falante	23.8/.0001	6.5/.0001	57.2/.0001	111.3/.0001	119.2/.0001
Vogal	1071.0/.0001	1381.0/.0001	48.4/.0001	67.8/.0001	70.3/.0001
Fal. X Vogal	4.5/.0001	3.5/.0001	6.3/.0001	3.2/.0001	3.3/.0001
Vogal	804.6/.0005	1255/.0001	32.4/.0001	35.7/.0001	35.6/.0001
Velocidade	6.1/.02	NS	NS	3.8/.05	NS
Vog. X Vel..	3.8/.05	4.9/.0003	NS	NS	NS

TABELA 4.5: Resultados de Análise de Variância (BMDP-2V) incluindo no modelo tanto os efeitos isolados quanto as possíveis interações entre variáveis categoriais. Em cada linha, o primeiro número dá o valor de F e o segundo, após a barra inclinada, o nível de significância (NS=não significativo).

As figuras 4.3 e 4.4 mostram graficamente as médias de F3 e F4, respectivamente, nas duas condições de velocidade, para cada falante separadamente. Na figura 4.3 fica claro que as alterações em F3, em função da velocidade variam entre os falantes, não só quanto à magnitude mas também quanto à própria direção da mudança. Assim, por exemplo, os falantes ZP, DO e WA sofrem apenas um pequeno aumento de F3 na velocidade rápida, enquanto nos falantes R1, MS e ZR se verifica uma diminuição, de maior magnitude, na mesma condição. As duas produções do falante R1/R2, curiosamente, apresentam efeitos opostos na velocidade rápida: aumento em R2 e diminuição em R1; esse aspecto pode estar relacionado às diferentes condições de ressonância, em função da presença do estado gripal na primeira gravação (R1), mas a causa específica do efeito observado não é de fácil explicação. Para F4, as diferenças inter-falantes são menores. Na figura 4.4 verificamos que os falantes ZR, R1, R2, MS, EN e DO têm aumentos pequenos de F4 na velocidade rápida, WA apresenta um aumento mais

expressivo, AG uma diminuição de cerca de 100 Hz, enquanto o falante ZP praticamente não altera F4 em função da velocidade de emissão.

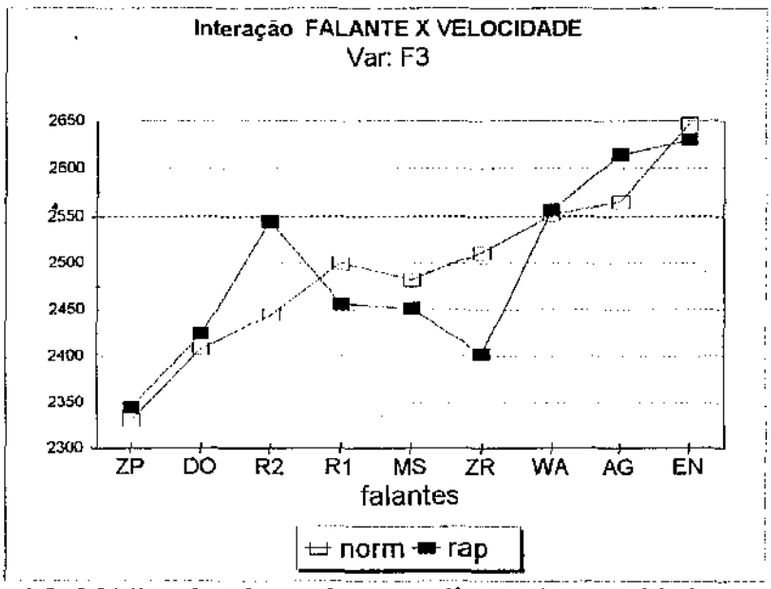


FIGURA 4.3: Médias de F3 nas duas condições de velocidade de emissão (normal *vs.* rápida), para cada falante separadamente. Podemos observar que o efeito varia entre os falantes não só em magnitude, mas também na direção da mudança.

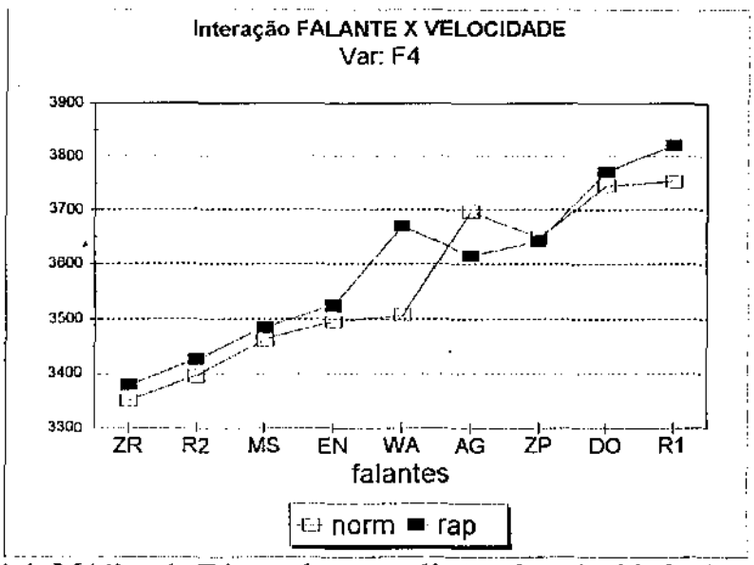


FIGURA 4.4: Médias de F4 nas duas condições de velocidade de emissão (normal *vs.* rápida), para cada falante separadamente. Podemos observar que o efeito varia entre os falantes não só em magnitude, mas também na direção da mudança.

O aumento de F4, observado na maioria dos falantes, na velocidade rápida, pode estar associado a um levantamento da laringe (Fant 1960:84), em função de uma postura geral mais tensa nessa condição de produção. No caso de AG, que apresentou a tendência oposta, é interessante ressaltar que esse falante praticamente não alterou a velocidade de emissão, apesar das instruções explícitas (v. seção 7). A inconsistência nos resultados relativos a F3 são mais difíceis de explicar, especialmente se confrontados com os resultados de F4. Nolan (1983:170) relata que um aumento de F3 é observado nas qualidades de voz *laringe abaixada, palatalizada, dentalizada* e *velarizada,* e uma diminuição de F3 ocorre nas qualidades *faringalizada, laringe levantada, retroflex* e *open rounding.* Se aceitarmos um levantamento de laringe associado ao aumento de F4, não é possível explicar o aumento de F3 para a maioria dos falantes. Essa possibilidade, entretanto, é compatível com o observado no falante ZR; esse falante apresentou o maior grau de diminuição em F3, na velocidade rápida, concomitantemente com um aumento em F4; observou-se também que esse mesmo falante produziu a maior variação de velocidade intra-falante entre as duas condições (v. seção 7), provavelmente associada a uma postura laríngea bastante diferenciada (levantamento de laringe?), um aspecto que, no falante ZR, se refletiu nas características do *tilt* espectral, modificando sensivelmente a configuração do ELT (v. seção 6), e até mesmo no *slope* extraído a partir de F1-F4 (SL14; v. figura 4.5).

É possível que a divergência inter-falantes nos efeitos da velocidade sobre F3 tenha alguma relação com o grau de *undershoot* articulatório associado a diferentes estratégias adaptativas efetuadas por cada falante quando produzindo fala rápida. Já comentamos anteriormente (v. seção 4.2) que diferentes falantes adaptam a velocidade de seus movimentos articulatórios em diferentes graus (Cf. Sonoda 1987). O grau de variação de F3 aumenta, considerando a distância da constrição em relação à glote, à medida em que diminui a área da constrição (Fant 1960:84); assim,

deslocamentos linguais relacionados com *undershoots* articulatórios podem provocar efeitos significativos em F3, especialmente em vogais altas (Nolan 1983:176). É importante observar, no entanto, que, no que diz respeito a F1 e F2 vimos que não parece ter havido *undershoot* significativo, mas esse resultado pode ter sido, de algum modo, distorcido pela grande variabilidade desses formantes em função da qualidade vocálica, já que, ao examinarmos o efeito da velocidade de emissão nos formantes consideramos todas as vogais em conjunto. Um efeito significativo poderia, entretanto, aparecer em F3, sendo esse formante menos dependente da qualidade vocálica. Mais adiante examinaremos os efeitos da velocidade em cada vogal separadamente (para constatar que, de fato, houve uma compressão do espaço vocálico na velocidade rápida; v. figura 4.8).

A variável contínua SL14 (*slope* F1-F4), além do efeito isolado de falante (já examinado anteriormente; v. tabela 4.3b), indica também uma interação significativa FALANTE X VELOCIDADE (F=3.1; p< .002). A figura 4.5 apresenta graficamente as diferenças inter-falantes no comportamento de SL14 em função da velocidade de emissão. Para a maior parte dos falantes SL14 varia pouco em função da velocidade de emissão. Há, no entanto, algumas divergências: ZP, ZR e WA apresentam um decaimento espectral mais suave na fala rápida, sinalizando uma voz mais "tensa" (*strained*), ou "hiper-funcional" (Cf. Kitzing 1986); R1/R2, EN e DO mostram a tendência oposta, mas com variações de pequena magnitude; MS e AG têm pouca ou nenhuma alteração nesse parâmetro. Apesar do efeito de interação estatisticamente significativo FALANTE X VELOCIDADE, SL14 é relativamente estável nas diferentes velocidades de emissão, indicando que medidas baseadas no *tilt* espectral podem ser bastante eficientes para a Identificação de Falantes; esse aspecto é confirmado na seção 6, onde examinaremos a eficiência de medidas derivadas do ELT.

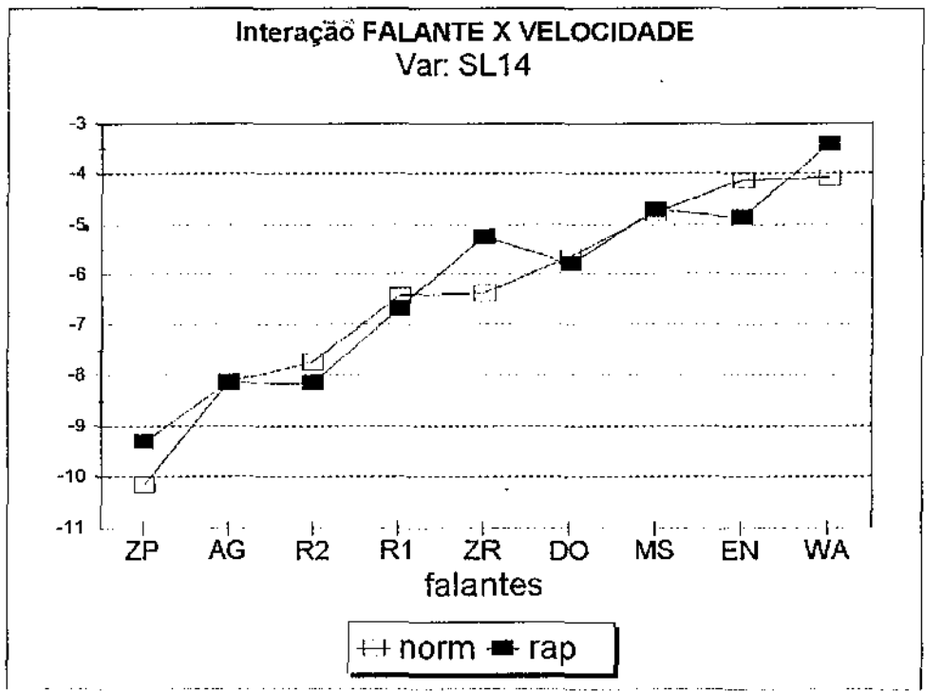


FIGURA 4.5: Médias de SL14 (*slope* F1-F4) nas duas condições de velocidade de emissão (normal *vs.* rápida), para cada falante separadamente. Podemos observar que o efeito varia entre os falantes não só em magnitude, mas também na direção da mudança.

No modelo VOGAL X VELOCIDADE aparecem dois efeitos relevantes de interação, em F1 e F2 ($F= 3.8$; $p<.05$, e $F= 4.9$; $p< .0003$; v. tabela 4.5). As figuras 4.6 e 4.7 mostram graficamente as médias de F1 e F2, para cada categoria vocálica separadamente, nas duas condições de velocidade de emissão.

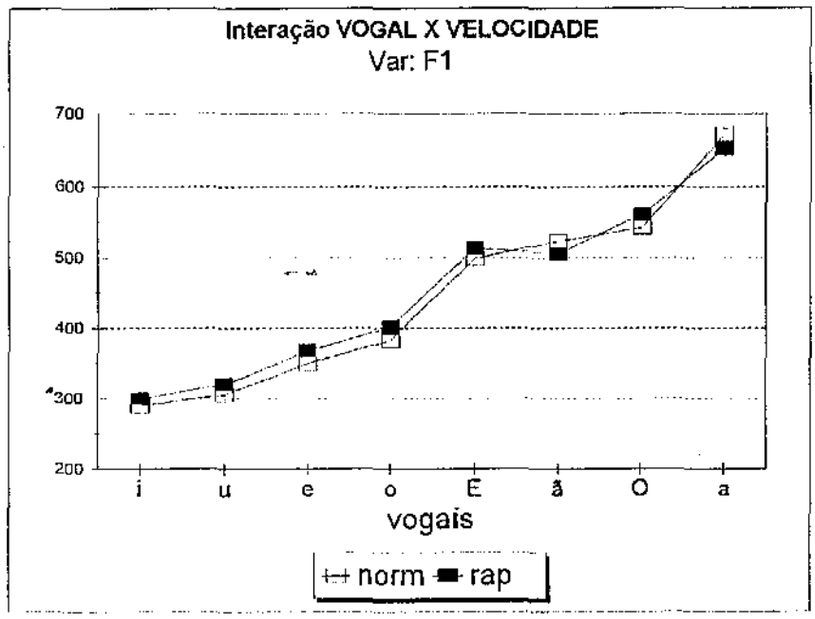


FIGURA 4.6: Médias de F1 para cada categoria vocálica, nas duas condições de velocidade (normal *vs.* rápida). Apenas as vogais baixas /a/ e /ã/ apresentam uma diminuição de F1 na velocidade rápida, enquanto as demais têm um aumento.

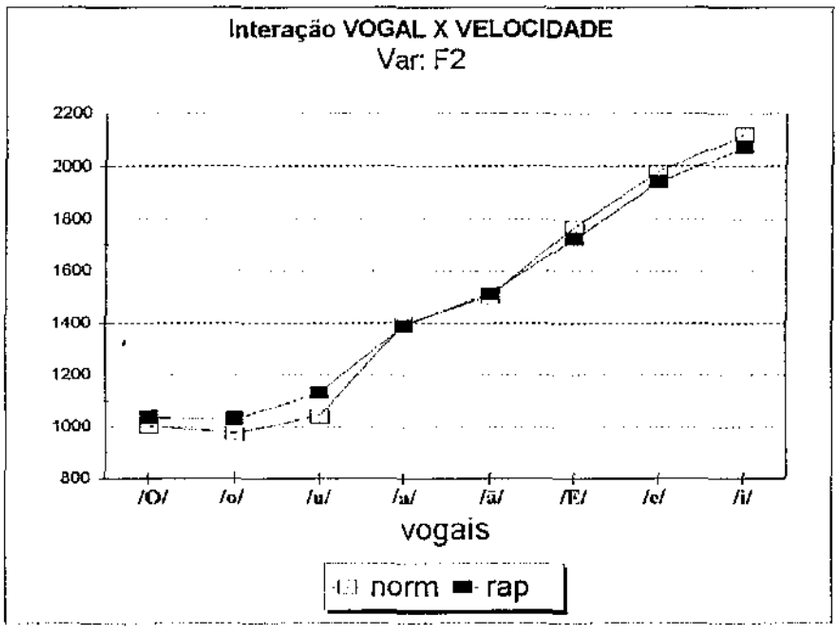


FIGURA 4.7: Médias de F2 para cada categoria vocálica, nas duas condições de velocidade (normal *vs.* rápida). As vogais posteriores têm um aumento de F2 na velocidade rápida, enquanto as anteriores apresentam a tendência inversa. As vogais mais centrais /a/ e /ã/ praticamente não se alteram.

Podemos observar, na figura 4.6, que a variação de F1 em função da velocidade não ocorre de forma homogênea para todo o grupo de vogais. As vogais mais centrais /a/ e /ã/ apresentam uma tendência inversa às demais vogais, diminuindo F1 na velocidade rápida. Da mesma forma, para F2, verificamos, na figura 4.7, que as vogais anteriores /ɛ, e, i/ têm uma diminuição na freqüência de F2 na velocidade rápida, as vogais posteriores /O, o, u/ têm um aumento, e as vogais centrais /a/ e /ã/ praticamente não se alteram. Os resultados combinados de F1 e F2 indicam um encolhimento do polígono vocálico definido no espaço F1 X F2, como ilustra a figura 4.8. A figura 4.8 deixa claro que houve um *undershoot* articulatório na velocidade rápida, fazendo com que as vogais sofressem um pequeno grau de centralização. Esse efeito está certamente relacionado a posturas linguais menos extremas, mais próximas da posição neutra. A redução vocálica observada encaixa-se no modelo de *target undershoot* (Lindblom 1963), que prevê uma relação entre a duração vocálica e o grau de redução (v. seção 4.2). Discutimos anteriormente (v. seção 4.3.2.1) a possibilidade de o aumento de F1 na velocidade rápida estar associado a uma maior abertura mandibular, em decorrência do maior esforço vocal verificado nessa condição (Cf. van Son e Pols 1990:1692). Examinando agora a variação de F1 em função da velocidade, vogal a vogal, vemos que essa hipótese é improvável, pois o aumento de F1 não ocorre para as vogais abertas /a, ã/. É possível admitir, entretanto, que tenha havido uma maior abertura mandibular **média** na velocidade rápida, ou seja, maior para todas as vogais, com exceção de /a,ã/, onde a abertura mandibular teria sido menor do que na velocidade normal.

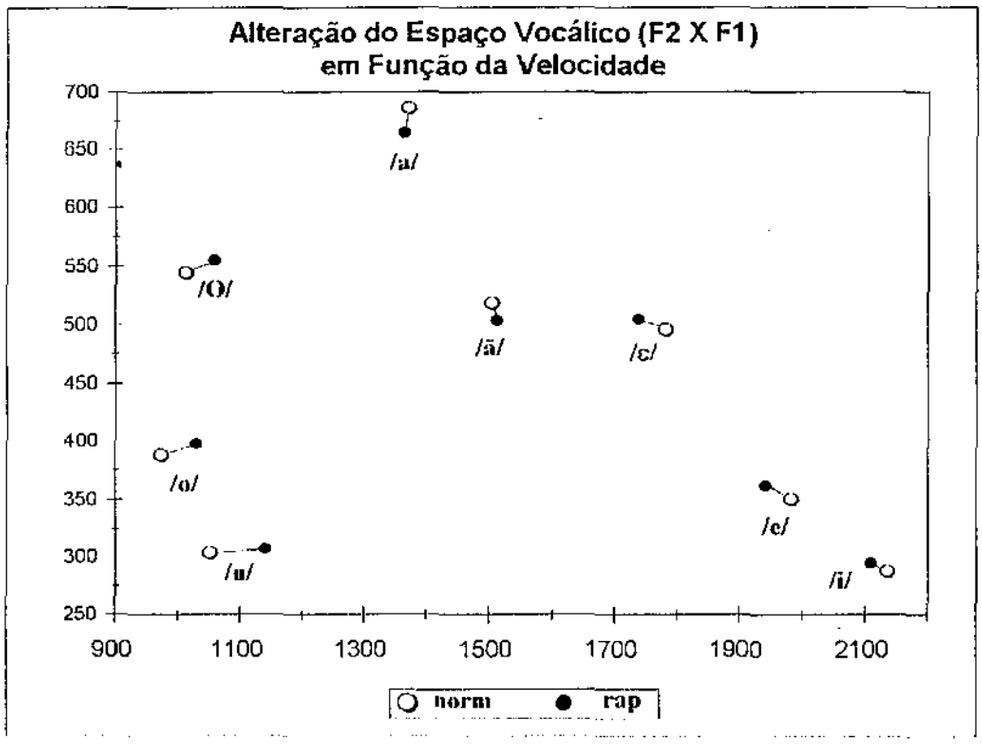


FIGURA 4.8: Espaço vocálico definido por F2 X F1 (F2 na abscissa e F1 na ordenada). Há um encolhimento do polígono vocálico na velocidade rápida, indicando um pequeno *undershoot* articulatório nessa condição.

Ainda na tabela 4.5, no modelo FALANTE X VOGAL, verificamos que FALANTE, isoladamente, tem efeitos significativos em todos os formantes, inclusive F2. Esse resultado pode parecer contraditório com o apresentado anteriormente, na tabela 4.3a, onde F2 não mostra efeito significativo. A diferença explica-se pelo fato de no presente modelo (FALANTE X VOGAL) a variância ser explicada levando em conta também a variância das categorias vocálicas; assim,

apesar de a variabilidade de F2 estar fortemente vinculada às distinções vocálicas (v. na tabela 4.5 o alto valor de F para VOGAL, isoladamente: F= 1381.0; p< .0001), a variabilidade inter-falante pode tornar-se significativa, se for considerada a co-variância da categoria VOGAL.

A comparação dos efeitos isolados VOGAL e FALANTE, na tabela 4.5, dá uma idéia da influência relativa de cada fator. Vemos que o peso da influência de cada formante, na distinção de falantes, se dá na seguinte ordem: F4 > F3 > F1 > F2 (F= 111.3, 57.2, 23.8, 6.5; p< .0001, para todos). Os efeitos de VOGAL, dentro das expectativas, apresentam um quadro quase invertido: F2 > F1 > F4 > F3 (F=1381.0, 1071.0, 67.8, 48.4; p<.0001 para todos).

Todos os formantes, no modelo FALANTE X VOGAL apresentam efeitos de interação estatisticamente significativos entre as duas variáveis categoriais. Esse resultado indica que a variação inter-falante de cada formante não se dá de forma homogênea para todas as vogais, ou seja, alguns falantes podem ter um determinado formante relativamente mais alto para uma vogal, mas mais baixo para outra. Esse aspecto ficará claro no exame das figuras 4.9 - 12, que mostram, graficamente, as médias individuais de cada falante, para cada vogal (integrando as duas condições de velocidade). Para estabelecer um referencial, todos os gráficos nas figuras 4.9 - 12 foram ordenados colocando os falantes em ordem crescente do valor do formante analisado para a vogal /a/. A tabela 4.6 apresenta numericamente as médias individuais, desvios-padrão e o número de ocorrências de cada categoria.

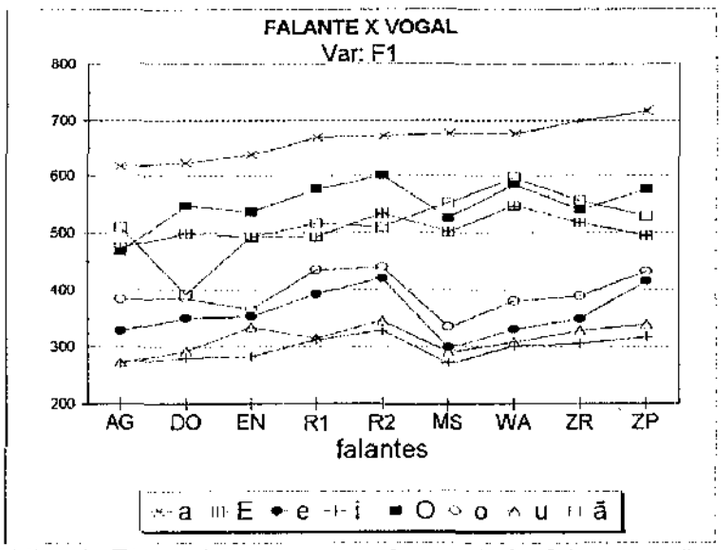


FIGURA 4.9: Médias de F1 por falante, para cada vogal. Os falantes estão ordenados, da esquerda para a direita, segundo a ordem dos valores de F1 para a vogal /a/. O gráfico permite constatar o efeito estatisticamente significativo FALANTE X VOGAL. Se não existisse interação, todas as linhas seriam exatamente paralelas.

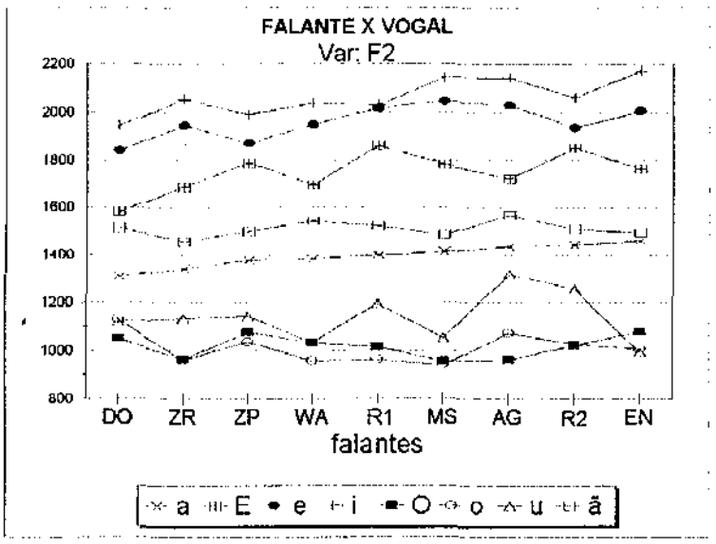


FIGURA 4.10: Médias de F2 por falante, para cada vogal. Os falantes estão ordenados, da esquerda para a direita, segundo a ordem dos valores de F2 para a vogal /a/. O gráfico permite constatar o efeito estatisticamente significativo FALANTE X VOGAL. Se não existisse interação, todas as linhas seriam exatamente paralelas.

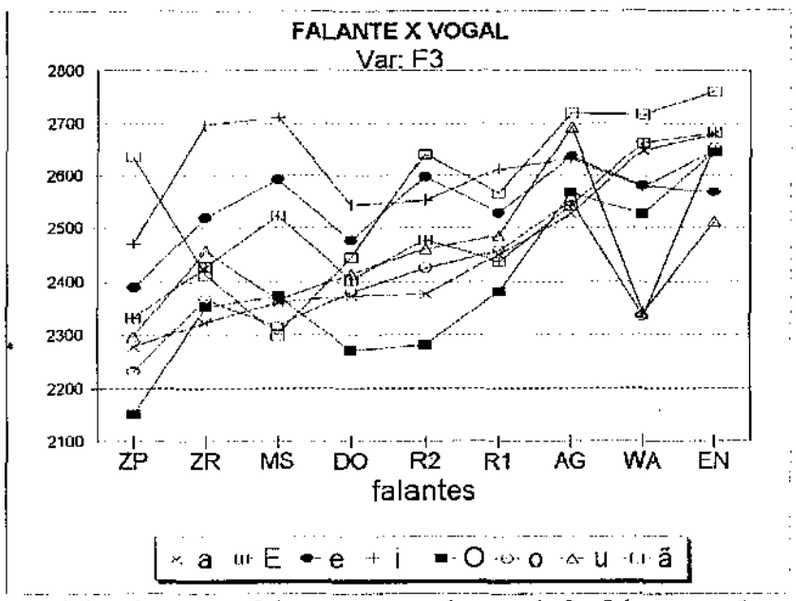


FIGURA 4.11: Médias de F3 por falante, para cada vogal. Os falantes estão ordenados, da esquerda para a direita, segundo a ordem dos valores de F3 para a vogal /a/. O gráfico permite constatar o efeito estatisticamente significativo FALANTE X VOGAL. Se não existisse interação, todas as linhas seriam exatamente paralelas.

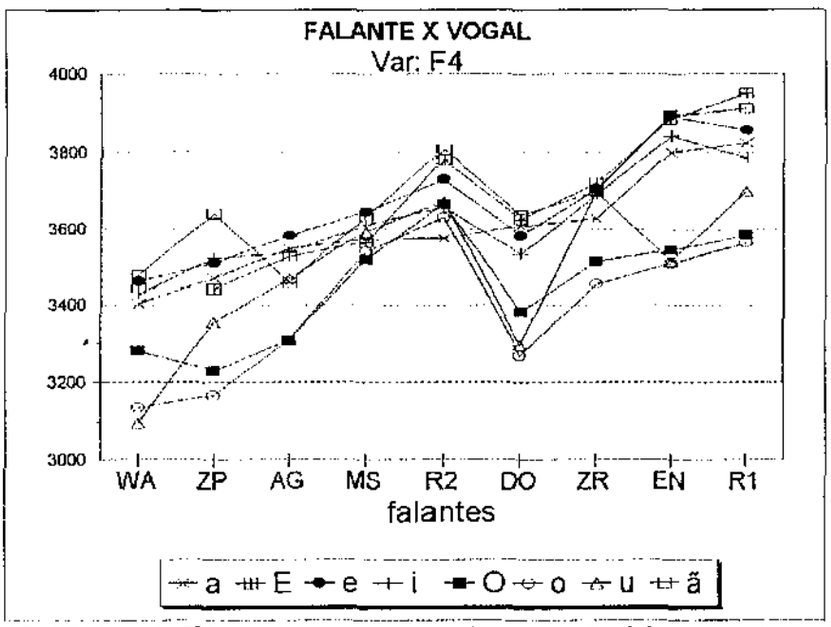


FIGURA 4.12: Médias de F4 por falante, para cada vogal. Os falantes estão ordenados, da esquerda para a direita, segundo a ordem dos valores de F4 para a vogal /a/. O gráfico permite constatar o efeito estatisticamente significativo FALANTE X VOGAL. Se não existisse interação, todas as linhas seriam exatamente paralelas.

Fal.	F_n	/a/	/ɛ/	/e/	/i/	/O/	/o/	/u/	/ã/	Tot.
ZR	F1	700	518	349	300	540	377	323	556	454
	dp	47	48	51	34	29	53	40	99	149
	n=	24	18	30	22	16	16	12	10	148
	F2	1342	1687	1943	2078	960	947	1080	1452	1524
	dp	83	126	106	144	58	116	211	53	435
	n=	24	18	30	22	16	15	11	10	146
	F3	2323	2427	2520	2696	2352	2392	2464	2412	2456
	dp	92	119	124	188	94	211	198	82	182
	n=	24	18	30	20	16	15	10	10	143
	F4	3628	3720	3704	3687	3512	3453	3713	3700	3645
	dp	94	85	68	89	73	152	147	74	131
	n=	24	18	30	22	16	15	12	10	147
EN	F1	638	493	353	283	536	372	333	492	436
	dp	58	53	49	15	30	38	82	96	131
	n=	24	18	30	22	16	16	12	10	148
	F2	1457	1764	2003	2193	1080	1003	993	1492	1601
	dp	109	129	194	186	86	125	143	103	457
	n=	24	18	30	22	16	14	11	10	145
	F3	2680	2682	2568	2648	2645	2666	2510	2760	2639
	dp	132	97	135	185	122	149	94	158	149
	n=	24	18	30	20	16	14	12	10	144
	F4	3800	3882	3890	3845	3542	3504	3477	3892	3757
	dp	131	44	118	70	88	141	401	68	219
	n=	24	18	30	22	16	15	12	10	147
R1	F1	663	493	395	317	576	431	317	516	464
	dp	43	51	54	25	41	32	12	43	126
	n=	24	18	30	22	16	16	11	10	148
	F2	1402	1860	2015	2084	1017	960	1107	1520	1582
	dp	89	77	106	206	109	81	198	88	448
	n=	24	18	30	22	16	15	9	10	147
	F3	2382	2478	2597	2553	2282	2427	2460	2640	2479
	dp	99	64	61	131	95	171	48	50	146
	n=	24	18	29	17	16	15	8	9	138
	F4	3824	3951	3856	3767	3585	3563	3700	3912	3785
	dp	115	130	116	169	99	170	191	141	184
	n=	22	18	29	22	16	13	8	5	138

TABELA 4.6 (1ª parte): Médias, desvios-padrão (dp) e número de ocorrências (n=) de cada formante de cada vogal, para cada falante separadamente (condicões de velocidade integradas)

Fal.	F_n	/a/	/ɛ/	/e/	/i/	/O/	/o/	/u/	/ã/	Tot.
ZP	F1	708	507	413	317	581	422	335	536	480
	dp	63	38	45	24	40	61	47	29	138
	n=	24	18	30	22	16	16	11	10	147
	F2	1383	1743	1883	2137	1077	1031	1136	1464	1567
	dp	83	124	162	288	76	92	138	87	413
	n=	24	18	30	21	16	14	9	10	142
	F3	2275	2321	2412	2523	2135	2235	2297	2509	2338
	dp	72	88	121	152	98	89	74	221	164
	n=	24	18	26	17	16	15	8	9	137
	F4	3469	3441	3508	3522	3229	3152	3355	3640	3411
	dp	178	90	138	133	106	88	156	215	194
	n=	24	12	26	15	14	15	8	5	116
AG	F1	619	474	325	270	467	371	282	510	412
	dp	40	26	60	32	22	37	26	75	128
	n=	24	18	29	22	16	17	12	10	148
	F2	1424	1772	2028	2162	962	1022	1153	1568	1583
	dp	254	108	159	192	52	130	270	70	469
	n=	24	18	28	22	16	17	12	10	147
	F3	2537	2542	2636	2601	2567	2587	2551	2720	2590
	dp	127	75	61	158	47	140	228	86	125
	n=	24	18	29	18	16	12	9	10	136
	F4	3545	3535	3574	3498	3306	3306	3340	3458	3473
	dp	165	244	195	178	116	87	207	193	203
	n=	22	18	29	20	16	14	6	10	135
WA	F1	677	547	331	299	584	395	282	574	457
	dp	38	49	55	37	26	65	26	56	151
	n=	24	18	30	22	16	16	12	10	148
	F2	1382	1698	1949	2064	1034	951	1063	1530	1533
	dp	68	136	173	155	61	107	229	73	429
	n=	24	18	30	21	16	16	12	10	147
	F3	2649	2664	2581	2576	2525	2322	2333	2717	2554
	dp	152	166	115	149	75	112	86	267	183
	n=	24	18	30	19	16	16	12	8	143
	F4	3402	3443	3460	3417	3281	3218	3094	3436	3365
	dp	82	60	69	75	118	205	240	135	164
	n=	24	18	30	21	16	18	11	10	148

TABELA 4.6 (2ª parte): Médias, desvios-padrão (dp) e número de ocorrências (n=) de cada formante de cada vogal, para cada falante separadamente (condicões de velocidade integradas)

Fal.	F_n	/a/	/ɛ/	/e/	/i/	/O/	/o/	/u/	/ã/	Tot.
MS	F1	676	501	300	266	527	334	278	552	419
	dp	73	54	42	22	41	66	39	111	160
	n=	24	18	30	21	16	18	12	5	144
	F2	1413	1783	2032	2192	956	939	1000	1516	1568
	dp	222	97	180	100	62	105	217	92	493
	n=	24	18	29	21	15	18	10	10	145
	F3	2362	2524	2591	2712	2432	2317	2420	2346	2467
	dp	119	94	118	163	271	120	228	179	198
	n=	24	18	28	12	16	18	10	10	136
	F4	3573	3562	3644	3607	3532	3538	3576	3666	3588
	dp	120	189	245	163	171	180	134	174	161
	n=	24	45	18	21	16	18	11	10	146
R2	F1	679	524	414	325	602	440	333	538	482
	dp	31	61	55	36	31	72	33	42	131
	n=	24	18	28	22	16	18	12	10	148
	F2	1428	1814	1974	2135	1028	1022	1145	1516	1586
	dp	87	112	150	205	57	129	282	69	437
	n=	24	18	30	21	15	17	11	10	36
	F3	2425	2457	2585	2576	2383	2457	2442	2596	2494
	dp	126	105	132	156	144	207	120	111	158
	n=	24	16	29	15	15	15	9	10	133
	F4	3575	3780	3740	3681	3670	3619	3469	3807	3658
	dp	276	330	246	253	50	266	444	299	273
	n=	16	6	18	17	16	17	9	6	105
DO	F1	624	499	351	278	547	383	295	393	426
	dp	41	54	69	31	35	71	35	103	133
	n=	24	18	28	22	16	18	12	9	147
	F2	1311	1582	1831	1983	1050	1126	1093	1512	1492
	dp	82	66	131	203	52	145	177	59	362
	n=	24	18	30	22	16227	18	12	10	150
	F3	2372	2402	2475	2543	1	2365	2415	2448	2416
	dp	93	91	60	279	54	120	146	125	152
	n=	24	18	30	19	16	18	11	10	146
	F4	3607	3626	3585	3540	3380	3267	3307	3648	3510
	dp	95	98	80	70	76	184	130	90	168
	n=	24	18	30	22	16	17	12	10	149

TABELA 4.6 (3ª parte): Médias, desvios-padrão (dp) e número de ocorrências (n=) de cada formante de cada vogal, para cada falante separadamente (condicões de velocidade integradas)

Fal.	F_n	/a/	/ɛ/	/e/	/i/	/O/	/o/	/u/	/ã/	Tot.
Tot.	F1	665	506	359	295	551	392	311	518	448
	dp	57	52	65	35	50	65	46	87	141
	n=	216	162	265	197	144	151	107	84	1326
	F2	1394	1745	1961	2114	1019	1001	1086	1508	1560
	dp	140	132	165	200	81	128	213	82	439
	n=	216	162	267	194	142	144	100	90	1315
	F3	2445	2500	2551	2602	2399	2409	2433	2569	2492
	dp	177	151	125	187	197	192	161	203	184
	n=	216	160	266	152	143	138	89	87	1256
	F4	3602	3658	3661	3623	3452	3401	3450	3681	3577
	dp	192	228	197	192	176	234	310	222	235
	n=	201	144	250	182	142	142	89	81	1231

TABELA 4.6 (4ª parte): Médias, desvios-padrão (dp) e número de ocorrências (n=) de cada formante de cada vogal, para o total de falantes (condicões de velocidade integradas)

Nas figuras 4.9 - 12, podemos verificar que os valores relativos inter-falante de cada formante variam consideravelmente em função da vogal. Se não existisse qualquer interação FALANTE X VOGAL, as linhas referentes às médias de cada categoria vocálica seriam exatamente paralelas. Na figura 4.9, por exemplo, podemos observar que o falante MS tem valores médios de F1 baixos para as vogais fechadas, mas valores relativamente altos para as abertas. Para F2, os efeitos de interação são menores, como já verificamos estatisticamente na tabela 4.5; mas ainda é possível observar desvios individuais relevantes: o falante AG, por exemplo, apresenta um valor excepcionalmente alto de F2 /u/, enquanto o falante EN, apresenta uma média relativamente baixa apenas nessa vogal. O grande cruzamento de linhas no gráfico referente a F3 (figura 4.11) ressalta o maior efeito de interação VOGAL X FALANTE; para alguns falantes, F3 pode variar bastante entre as diferentes categorias vocálicas (ZP e MS, por exemplo), enquanto para outros a faixa de variação é mais estreita (DO e AG, por exemplo). No quarto formante

(figura 4.12) as diferenças inter-individuais na dispersão inter-vocálica são também evidentes; o falante MS, por exemplo, forma um *cluster* compacto, variando pouco F4 entre as vogais, enquanto o falante ZP apresenta a tendência inversa. É interessante destacar, na figura 4.12, que, apesar da dispersão inter-vocálica, é evidente a potencialidade de F4 para a separação de falantes em grupos distintos; esse tipo de informação, embora não possa ser usada como determinante da identidade, pode ser extremamente útil para a exclusão de um determinado suspeito na aplicação forense (o falante WA, por exemplo, não seria confundido com os falantes MS, R1/R2, ZR e EN).

A pré-categorização das qualidades vocálicas, mesmo sob a forma de transcrição larga aqui utilizada, acrescenta informação importante, na medida em que permite particularizar regiões do espaço vocálico que, eventualmente, reflitam características idiossincráticas. A figura 4.13 reúne as médias de cada falante separadamente, para cada vogal (duas condições de velocidade integradas), graficamente dispostas em um espaço bidimensional definido por F2 X F1 (F2 na abscissa e F1 na ordenada); cada qualidade vocálica está separada por uma linha fechada reunindo as médias de todos os falantes, com símbolos diferentes para cada falante. Podemos observar, na figura 4.13, que as categorias vocálicas aqui definidas não se sobrepõem, com exceção de /i/ e /e/, onde a pequena sobreposição é causada principalmente pela posição muito alta do /e/ dos falantes AG e MS.

A formação de *clusters* vocálicos bem definidos sugere que a não-distinção oral-nasal não introduziu variação adicional significativa aos dados; a formação das classes separadas /a/ e /ã/, por outro lado, mostrou-se necessária, com a resultante formação de *clusters* independentes no espaço F2 X F1. Nesse sentido é interessante observar que a freqüente associação de um F1 mais alto para vogais nasalizadas (se comparadas com seu correspondente não-nasalizado) (Cf. Kent e Read 1992) não se confirma aqui. Esse aspecto foi examinado em Wright (1986), onde se verifica que

para a vogal /ã/ a previsão não é válida, ocorrendo o oposto, ou seja, uma diminuição de F1, emprestando uma qualidade mais centralizada à vogal nasalizada.

A figura 4.13 ressalta diferenças inter-falantes interessantes. Observe-se, por exemplo, como o polígono vocálico do falante DO é menos periférico do que o dos demais falantes ou como o falante AG tem uma realização bastante centralizada de /u/, enquanto EN tem um valor baixo de F2 para a mesma vogal; a posição de /ã/, do falante DO, é totalmente divergente, em relação aos demais falantes [1]; para o falante EN, as realizações das categorias classificadas como /o/ e /u/ são muito próximas, no espaço F2 X F1, enquanto são distantes para o falante AG, etc.

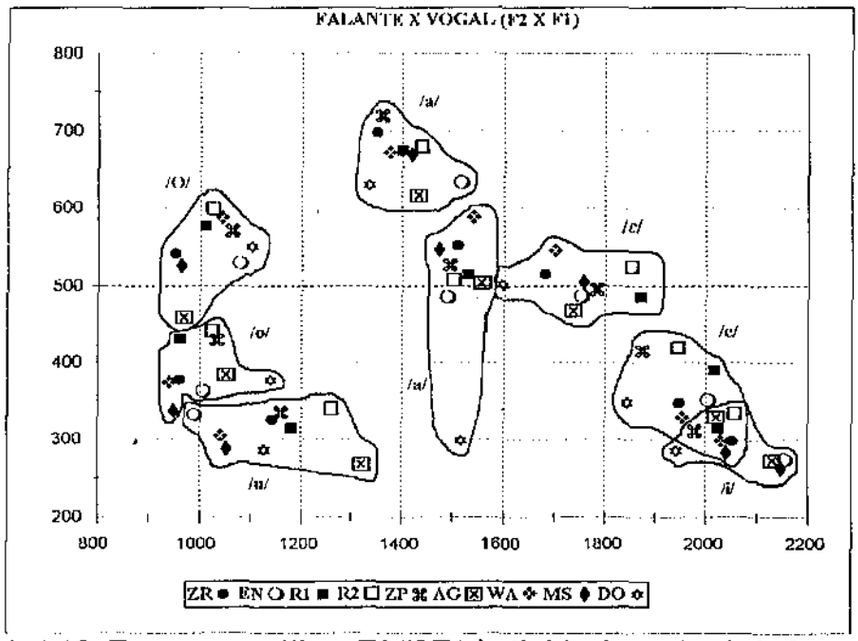


FIGURA 4.13: Espaços vocálicos F2 X F1 individuais; cada símbolo representa a média de um falante em cada categoria vocálica.

Além dos formantes, a variável SL14 (*slope* F1-F4) também mostrou efeitos significativos no modelo FALANTE X VOGAL. A influência de FALANTE é maior do que a de VOGAL (F= 119.2, 70.3, respectivamente, ambos com p< .0001; v. tabela 4.5). Há um efeito de interação significativo, ilustrado pela figura 4.14. O gráfico sugere que, apesar da variação inter-vocálica, o parâmetro em questão tem um âmbito de variação inter-falante considerável.

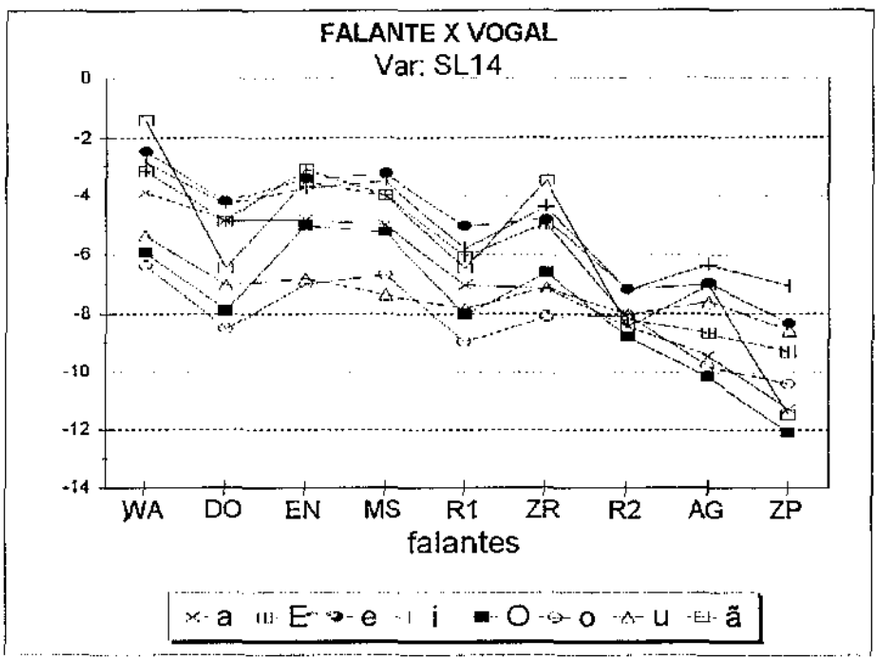


FIGURA 4.14: Médias de SL14 por falante, para cada vogal. Os falantes estão ordenados, da esquerda para a direita, segundo a ordem dos valores de SL14 para a vogal /a/. O gráfico permite constatar o efeito estatisticamente significativo FALANTE X VOGAL. Se não existisse interação, todas as linhas seriam exatamente paralelas.

4.3.2.3) *Eficácia de diferentes categorias vocálicas na Identificação de Falantes*

Sistemas de Identificação e Verificação de falantes podem empregar procedimentos dependentes ou independentes do conteúdo segmental. Em geral, procedimentos dependentes do material segmental são mais adequados nas situações onde o falante é cooperativo, tal como é esperado no paradigma de Verificação (Rosenberg 1976; Doddington 1985). No entanto, mesmo na situação forense, existem casos onde a questão da cooperação do falante não se coloca e é possível utilizar procedimentos dependentes do material segmental [2]. Nessas situações, pode ser importante saber quais fonemas veiculam mais eficientemente informação específica do falante; essa informação é relevante, por exemplo, para a construção da(s) sentença(s) de confronto. A eficácia de diferentes vogais e, mais especificamente, de diferentes formantes dessas vogais na Identificação ou Verificação de falantes já foi abordada em uma série de trabalhos (Bricker e Pruzansky 1966; Stevens et al. 1968; LaRivière 1975; Goldstein 1976; Paliwal 1984; Tartter 1991). Os resultados relatados, entretanto, não são convergentes; Stevens *et al.* (1968) observam que a identificação auditiva é mais eficiente se o falante produz uma palavra contendo vogal alta anterior enquanto LaRivière (1975), também estudando julgamentos auditivos, verifica um favorecimento das vogais /a/ e /æ/; Goldstein (1976) relata que F1 da vogal central /ə/ teve o menor valor de F em uma comparação de dez falantes, enquanto Paliwal (1984) verifica, para um grupo semelhante de falantes, que F1 /ə/ tem um alto valor de F. É provável que a inconsistência nos resultados esteja relacionada - ao menos em parte- a questões metodológicas (avaliação perceptual *versus* estatística) e a particularidades dos diferentes conjuntos de vogais ou dos sistemas vocálicos nos quais se basearam os experimentos. Além desses fatores, há alguma evidência sugerindo que em testes de

avaliação perceptual, o favorecimento de uma qualidade vocálica pode estar relacionado a características idiossincráticas do ouvinte (Bricker e Pruzansky 1966).

A principal limitação dos estudos acima citados é o emprego de estímulos baseados em contextos controlados (palavras ou sentenças fixas isoladas), o que pode ter contribuído para diminuir a variância intra-falante. Esse paradigma experimental, apesar de válido para o modelo de Verificação, é pouco compatível com a situação forense típica, onde é preciso avaliar também a variabilidade relacionada aos diferentes contextos fonéticos encontrada normalmente na fala fluente.

A presente seção pretende examinar a eficácia de diferentes vogais extraídas de leitura fluente (texto I; grupo principal de falantes, gêmeos excluídos). Os testes estatísticos foram baseados no conjunto total de vogais, integrando as medidas das duas condições de velocidade. A tabela 4.7 mostra os resultados de uma análise de variância para verificar a significância das diferenças entre os grupos (falantes), realizada através do programa BMDP-7D. BMDP-7D não exige grupos com o mesmo tamanho e possui testes robustos para verificar diferenças entre grupos mesmo que as variâncias dos grupos sejam desiguais (o que se espera para conjuntos de dados não balanceados, como o utilizado no presente experimento). BMDP-7D executa, em primeiro lugar, uma análise de variância *standard*, fornecendo o valor de F e o nível de significância. É realizado também o teste *Levene*, para verificar a hipótese de igualdade de variâncias entre os grupos; todas as vezes em que essa hipótese foi rejeitada, utilizou-se alternativamente o valor mais robusto de F fornecido pelo teste *Brown-Forsythe*, que não pressupõe igualdade de variâncias entre os grupos (nesses casos o valor de F está expresso em itálico na tabela 4.7). Testes *post hoc* de comparação múltipla de médias foram realizados (método *Tukey*)

para examinar quais pares de falantes foram considerados diferentes para cada uma das variáveis (cada variável aqui é um determinado formante de uma determinada vogal).

A tabela 4.7 dá o valor de F e o respectivo nível de significância correspondentes às análises de variância. A tabela indica também quantos pares de falantes foram corretamente separados, isto é, considerados como possuindo médias significativamente diferentes pelo procedimento de comparação *Tukey*, a um nível de significância menor que .05 (coluna **p.c.**,na tabela 4.7). Nas comparações entre R1 e R2, como se trata de produções não contemporâneas do mesmo falante, consideramos uma **não-distinção** entre as médias como um "acerto". A tabela 4.7 mostra também, para cada falante, o número de pares significativamente diferentes incluindo esse falante, com base na comparação de médias de cada variável (isto é, de cada formante de cada uma das vogais).

F1											
Vog.	**F=**	**p.c.**	**ZR**	**EN**	**R1**	**ZP**	**AG**	**WA**	**MS**	**R2**	**DO**
/a/	*9.7* ***	14	3	2	2	4	5	2	2	3	5
/ɛ/	*3.3* **	4	0	1	2	0	1	3	0	1	0
/e/	*16.3* ***	21	4	3	6	6	3	3	6	7	4
/i/	*12.1* ***	17	2	3	5	4	5	2	5	5	3
/O/	*23.5* ***	21	4	5	4	4	8	4	5	6	2
/o/	*6.1* ***	7	1	1	2	1	1	0	3	5	0
/u/	*3.3* **	4	0	1	1	1	0	0	3	2	0
/ã/	*4.2* **	8	1	0	2	1	1	1	1	2	7
Tot.	*5.1* ***	8	0	0	2	3	3	0	2	4	2
F2											
Vog.	**F=**	**p.c.**	**ZR**	**EN**	**R1**	**ZP**	**AG**	**WA**	**MS**	**R2**	**DO**
/a/	*2.7* *	2	0	1	1	0	0	0	0	1	1
/ɛ/	*9.7* ***	12	2	1	5	2	1	2	1	4	6
/e/	*5.9* ***	9	0	1	3	3	2	0	2	2	5
/i/	*2.7* **	3	0	1	1	0	0	0	1	1	2
/O/	*8.1* ***	11	3	3	1	3	3	1	4	1	3
/o/	4.5 ***	5	1	0	2	0	0	1	1	1	4
/u/	NS	1	0	0	1	0	0	0	0	1	0
/ã/	NS	2	1	0	1	0	1	0	0	1	0
Tot.	NS	1	0	0	1	0	0	0	0	1	0

TABELA 4.7 (1ª parte): Eficiência relativa de cada categoria vocálica para a separação de falantes (par a par). O valor de F é um indicador genérico da força de cada variável (F_n de uma determinada vogal) (* = p<.05; ** = p<.01; *** = p<.001). A coluna **p.c.** indica o número total de pares distintos pela variável em questão. As colunas correspondentes a cada falante indicam quantas vezes um determinado falante foi corretamente separado de outro falante (no caso das produções de R1/R2 considerou-se um "acerto" quando a variável **não** distingue o par R1/R2).

F3											
Vog.	**F=**	**p.c.**	**ZR**	**EN**	**R1**	**ZP**	**AG**	**WA**	**MS**	**R2**	**DO**
/a/	*37.6* ***	23	3	7	5	5	8	7	3	5	3
/ɛ/	*23.3* ***	22	3	7	4	6	5	7	4	4	4
/e/	*12.6* ***	15	2	2	3	7	3	2	2	3	6
/i/	2.3 *	1	0	0	1	0	0	0	0	1	0
/O/	*25.4* ***	22	4	6	6	7	5	4	4	4	4
/o/	*11.4* ***	15	2	7	3	4	5	2	2	3	2
/u/	*2.7* *	3	0	0	1	1	2	1	0	1	0
/ã/	*8.8* ***	14	4	4	3	1	3	3	5	2	3
Tot.	*43.8* ***	27	4	7	6	8	6	7	4	6	6
F4											
Vog.	**F=**	**p.c.**	**ZR**	**EN**	**R1**	**ZP**	**AG**	**WA**	**MS**	**R2**	**DO**
/a/	*18.9* ***	20	4	7	6+	4	3	7	3	2+	4
/ɛ/	*19.5* ***	22	6	6	7	5	4	5	3	4	4
/e/	*31.0* ***	26	6	7	7	5	5	7	4	6	5
/i/	*18.6* ***	21	5	7	6	4	4	5	3	5	3
/O/	*33.2* ***	25	5	5	5	6	5	5	5	8	6
/o/	*15.5* ***	18	2	3	5	5	3	5	4	5	4
/u/	*6.5* ***	8	2	1	3	0	0	5	1	2	2
/ã/	*9.4* ***	14	2	4	5	0	4	5	3	3	2
Tot.	*81.0* ***	28	6	7	6+	6	6	7	6	5+	7

TABELA 4.7 (2ª parte): Eficiência relativa de cada categoria vocálica para a separação de falantes (par a par). O valor de F é um indicador genérico da força de cada variável (F_n de uma determinada vogal) (* = p<.05; ** = p<.01; *** = p<.001). A coluna **p.c.** indica o número total de pares distintos pela variável em questão. As colunas correspondentes a cada falante indicam quantas vezes um determinado falante foi corretamente separado de outro falante (no caso das produções de R1/R2 considerou-se um "acerto" quando a variável **não** distingue o par R1/R2). O sinal **+** indica que nessa variável, as produções do mesmo falante R1/R2 foram estatisticamente diferentes.

O valor de F, na tabela 4.7, expressa a razão entre a variância inter-falante e a variância intra-falante; quanto maior esse valor, portanto, mais eficaz é o parâmetro acústico examinado. Comparando os resultados com aqueles obtidos a partir do conjunto inteiro de vogais (tabela 4.3a; seção 4.3.2), vemos que, para os dois primeiros formantes, a distintividade inter-falante tende a aumentar nas análises de variância baseadas em cada uma das categorias vocálicas separadamente. Por outro lado, para os formantes altos a variabilidade cai consideravelmente para algumas vogais isoladas - especialmente F3 das altas /i/ e /u/ e F4 de /u/ e /ã/. Os resultados para /u/ são discutíveis, já que há poucos exemplos dessa vogal no *corpus* estudado. O baixo valor de F para F3/i/, entretanto, parece estar relacionado a uma alta variabilidade intra-falante, em decorrência, provavelmente, de efeitos de coarticulação com diferentes contextos fonéticos. O desvio-padrão de F3/o/ de cada falante isolado também tende a ser alto, mas as distribuições individuais são suficientemente afastadas para permitir um número razoável de distinções entre falantes (para uma avaliação dos desvios-padrão individuais para cada vogal/falante, ver tabela 4.6).

É interessante observar que algumas variáveis separam mais eficientemente determinados falantes. O falante DO, por exemplo, é significativamente diferente de todos os demais, com exceção do falante EN, na variável F1/ã/; observe-se ainda que DO foi o **único** falante considerado diferente nessa variável (v. no entanto nota 1). O falante AG é diferente de todos nas variáveis F1/O/ e F3/a/, e o falante R2 distingue-se significativamente de todos os demais (com exceção de si próprio em amostra não contemporânea: R1) na variável F4/O/. Por outro lado, alguns falantes tendem a ser sistematicamente confundidos em todas as variáveis; os falantes ZR e MS, por exemplo, só são significativamente diferentes em F1/i/ ($p<.01$), F1/e/ ($p<.05$) e F4/ɛ/ ($p<.05$).

As amostras não contemporâneas do falante R1-R2 foram consideradas significativamente iguais para todas as variáveis, com exceção de F4/a/ (v. tabela 4.7, células assinaladas com +). Na primeira coleta (R1), onde o falante estava sob forte estado gripal, a média de F4/a/ foi consideravelmente mais alta: 3824 Hz, contra 3575 Hz na segunda coleta (R2), já em condições normais de saúde. Na verdade, com exceção das vogais /O/ e /o/, todas as demais têm F4 ligeiramente mais alto na condição gripal (R1). As funções de "sensitividade" (*sensitivity*) para perturbações na área seccional em diversas regiões do trato vocal, dadas em Fant (1980:71), indicam que o F4 de todas as vogais é sensível a pequenas perturbações na laringe. As previsões do modelo proposto por Fant (1960; v. nomograma pg. 82) indicam que F4 é relativamente estável para todos os pontos de constrição, sofrendo um considerável aumento apenas para redução de área seccional na baixa faringe. É possível que a alteração em F4/a/ no falante R1 esteja relacionada à existência de um estado inflamatório afetando a área seccional na região da laringe e/ou faringe.

Stevens (1972, 1989) sugere que a relação entre os parâmetros articulatórios e a saída acústica não é linear; existiriam algumas configurações articulatórias - diz Stevens - onde uma variação articulatória dentro de um certo âmbito não altera significativamente os aspectos acústicos relevantes. No caso específico das vogais, essas regiões seriam os pontos ao longo do trato vocal onde um estreitamento provocaria um valor máximo ou mínimo de um ou mais de um dos dois ou três primeiros formantes:

> *when a constriction is located in the vicinity of one of these points, the frequencies of the formants are relative insensitive to modifications in the position of the constriction. At other constriction locations, the formant frequencies are much more sensitive to changes in the constriction position* (Stevens 1989:15).

As vogais polares /i/, /a/ e /u/ ajustam-se naturalmente a essa condição. A vogal /i/ aproxima maximamente F2 e F3, e as vogais /a/ e /u/ aproximam maximamente F1 e F2, a segunda com arredondamento adicional. Os pontos de constrição dessas vogais caracterizariam regiões "platô" (Stevens 1972) acusticamente mais estáveis. A existência quase universal dessas vogais quânticas nos sistemas vocálicos das diferentes línguas é consistente com a formulação de Stevens (1972,1989) (Cf. Lindblom 1986).

Parece razoável inferir que a menor suscetibilidade acústica a pequenas alterações no ponto de constrição faria com que a configuração espectral das vogais quânticas /i,a,u/ fosse menos sujeita à variação em função de efeitos coarticulatórios. Assim, a variabilidade intra-falante dos formantes de um conjunto de vogais em diferentes contextos fonéticos deveria ser menor para essas qualidades vocálicas do que para vogais não polares. Mais especificamente, poderíamos esperar que especialmente os formantes maximamente próximos fossem mais estáveis (F1, F2 de /a,u/ e F2, F3 de /i/). Um conjunto de medidas com menor dispersão para cada falante deveria produzir uma separação mais nítida entre falantes nos testes estatísticos de comparação de médias de formantes específicos (e, conseqüentemente, um valor mais alto de F). Examinando os valores de F das análises de variância para cada formante de cada vogal, verificamos que as vogais polares, ao contrário da expectativa, parecem separar os falantes menos

eficientemente do que as vogais médias. A tabela 4.8 mostra, para cada formante, a ordem decrescente do valor de F correspondendo a cada vogal.

Fn	Ordem decrescente de F
F1	/O/>/e/>/i/>/a/>/o/>/ã/>/ɛ/>/u/
F2	/ɛ/>/O/>/e/>/o/>/i/>/a/>/ã/>/u/
F3	/a/>/O/>/ɛ/>/e/>/o/>/ã/>/u/>/i/
F4	/O/>/e/>/ɛ/>/a/>/i/>/o/>/ã/>/u/

TABELA 4.8: Força relativa de cada variável (F_n de uma determinada vogal) na distintividade de falantes, avaliada pelo valor de F (razão entre a variância inter-falante e a variância intra-falante).

A informação fornecida pela tabela 4.8 é particularmente relevante para F2, já que esperamos observar aí, mais notadamente, efeitos de coarticulação em função dos diferentes contextos fonéticos. Segundo a teoria quântica, a expectativa seria encontrar uma maior estabilidade acústica nas vogais /a/ e /i/ nesse formante; o que ocorre, entretanto, é exatamente o oposto, com as vogais médias /E,O,e,o/ produzindo, aparentemente, distribuições de F2 menos dispersas. A prevista estabilidade de F3/i/ também não é confirmada.

É importante ressaltar que as observações a respeito da maior ou menor estabilidade acústica das diferentes vogais não podem ser consideradas conclusivas, em virtude da natureza do *corpus* aqui estudado. Como já comentamos anteriormente, não houve controle dos contextos fonéticos; assim, uma determinada vogal aparece em um conjunto particular de contextos, não diretamente comparável aos conjuntos de contextos das outras vogais. É possível que um conjunto de dados balanceado quanto a esse aspecto produzisse resultados um pouco diferentes. As tendências aqui observadas, entretanto, são consistentes com os resultados obtidos

por Paliwal (1984) para o Inglês Britânico; nesse estudo são examinadas onze vogais (/ə, Λ, u, U, o, O, a, æ, E, I, i/), verificando-se que para qualquer um dos formantes isoladamente, ou para os quatro primeiros formantes em conjunto, as vogais médias e/ou centrais (especialmente /ə/ e /Λ/) têm um melhor desempenho na identificação de falantes (n=10) do que as vogais "quânticas" /i,a,u/.

4.3.3) *Comparação de Medidas de Formantes entre Gêmeos*

As vozes de gêmeos univitelinos guarda entre si, geralmente, uma grande semelhança. Testes perceptuais indicam que ouvintes com bom desempenho na tarefa de discriminar falantes não conseguem, na maior parte dos casos, distinguir corretamente vozes de gêmeos idênticos (Rosenberg 1973). A similaridade entre as configurações anatômicas dos gêmeos imprime, certamente, uma similaridade nas saídas acústicas, mas é preciso considerar também que, salvo casos excepcionais, os gêmeos são criados no mesmo ambiente sócio-cultural e devem possuir, portanto, um vasto conjunto comum de padrões adquiridos.

Apesar da similitude no nível perceptual, alguns sistemas de Reconhecimento Automático de Falante não confundem gêmeos idênticos (Lummis 1973; Doddington 1985), indicando que existem pistas não salientes para ouvintes humanos. Duas possibilidades não exclusivas podem ser aqui consideradas: (1) o ouvinte simplesmente não concentra sua atenção nas pistas relevantes e/ou (2) a magnitude da diferença é muito pequena e está abaixo do limite perceptual de discriminação diferencial (mas não abaixo do limite de um sistema automatizado).

De modo a examinar mais de perto a variação entre gêmeos univitelinos, estudaremos nesta seção um conjunto de medidas de formantes, extraídos exclusivamente de vogais lexicalmente tônicas, com base na leitura do Texto II (ver

anexo) (para detalhes da metodologia de gravação e preparação do *corpus*, ver seção 3).

As figuras 4.15 e 4.16 mostram os espaços vocálicos definidos por F1 × F2 e F3 × F4, respectivamente; o falante JA está representado pelos retângulos vazios e o falante JR pelos retângulos cheios. No espaço definido pelos dois primeiros formantes existe uma grande sobreposição entre os dois gêmeos, um pouco menor apenas na região das vogais centrais. Nos formantes mais altos o padrão global é menos definido mas percebe-se que a coincidência é menor do que no espaço F1 × F2.

Uma forma de avaliar o grau de diferença entre os formantes dos gêmeos JA/JR é através de um teste *matched-t* (BMDP-3D). Nesse procedimento, cada observação de cada falante é comparada diretamente com uma observação equivalente do outro falante, par a par, separadamente. A tabela 4.9 apresenta as médias e desvios-padrão de cada formante, para cada gêmeo separadamente, o valor de *t*, a diferença média, desvio-padrão da diferença (JR–JA), o número de casos onde o valor de JR foi maior do que JA, o número de casos com diferença zero e, finalmente, o total de casos computados para cada formante.

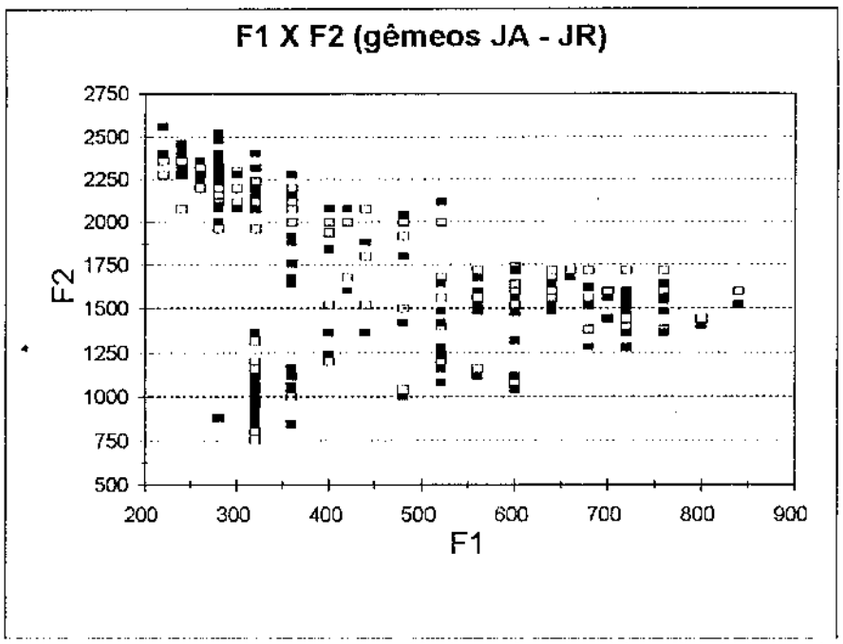


FIGURA 4.15: Espaços vocálicos F1 X F2 dos gêmeos JA e JR (JA=retângulos vazios e JR=retângulos cheios)

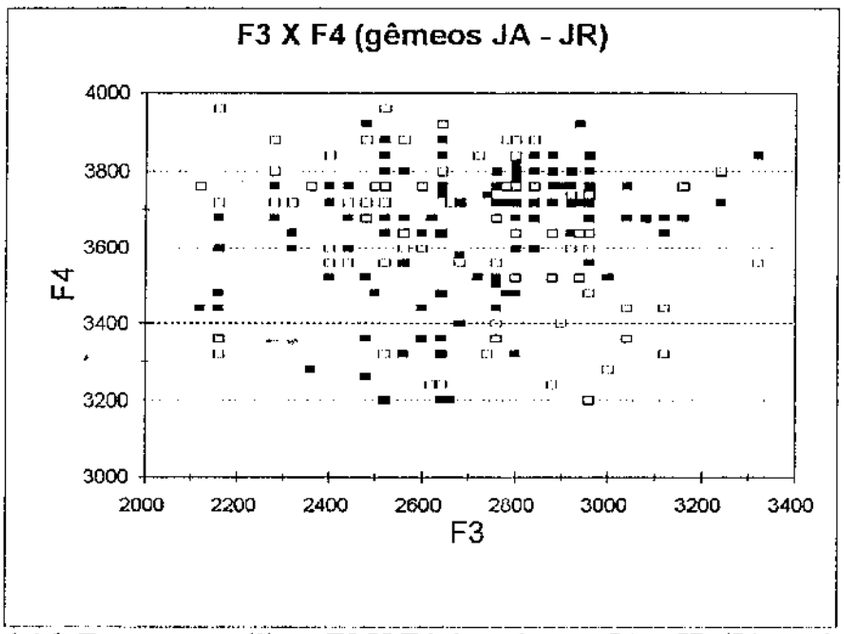


FIGURA 4.16: Espaços vocálicos F3 X F4 dos gêmeos JA e JR (JA=retângulos vazios e JR=retângulos cheios)

		F1	F2	F3	F4
JA	**med.**	460.14	1716.86	2571.80	3635.33
	D.P.	172.81	445.89	249.71	202.50
JR	**med.**	441.47	1708.66	2689.02	3639.17
	D.P.	169.82	482.46	265.21	190.93
matched-t		-3.79	-.98	7.09	.45
p<		.0002	NS	.0001	NS
dif. (JR–JA)		-18.67	-11.19	118.83	7.22
D.P. (dif.)		57.41	132.06	183.60	176.61
JR > JA		30	54	90	60
dif = 0		34	11	8	10
Total		136	134	120	119

TABELA 4.9: Médias, desvios-padrão de formantes, *matched-t*, diferença média (JR –JA), desvio-padrão da diferença par-a-par, número de casos onde o valor de JR é maior que JA, número de casos com diferença zero e total de casos computados.

Verificamos, na tabela 4.9, que as médias dos gêmeos são bastante próximas, especialmente F2 e F4. As médias de F1 e F3, no entanto, apesar de próximas, são significativamente diferentes, como indica o valor de t . O exame das diferenças *matched* sugere que alguns pares de medidas (JA *versus* JR) podem ter valores consideravelmente diferentes, já que o desvio-padrão das diferenças é relativamente alto, embora o valor absoluto da diferença média - com exceção de F3 - seja baixo. Quanto à direção das diferenças, não há padrão em F2 e F4. Já em F1 e F3, há uma tendência a um dos gêmeos ter um valor maior que o outro (JA em F1 e JR em F3; ver última linha da tabela 4.9).

A grande proximidade das médias dos formantes dos gêmeos JA/JR sugere que pode estar em jogo um condicionamento de natureza fisiológica, fazendo com que, independentemente das diferenças locais, a conformação do trato, presumivelmente semelhante nos dois gêmeos, determine a faixa de variação de cada formante (pelo menos para um número considerável de observações, como parece ser o caso aqui). É importante ressaltar que as diferenças estatisticamente

significativas em F1 e F3 só aparecem no teste *matched* (ao realizarmos a análise de variância tradicional, comparando apenas as distribuições dos dois falantes - programa BMDP-7D - não observamos diferença significativa em F1 e F3).

Como o teste *matched* indicou diferenças em F1 e F3 realizamos uma nova análise de variância incluindo no modelo a interação FALANTE X VOGAL. Os resultados desse teste estão na tabela 4.10, e indicam que existe interação significativa FALANTE X VOGAL, tanto em F1 quanto em F3. Um teste posterior de comparação de médias revelou que apenas a vogal /ã/ apresentava diferenças significativas entre os dois gêmeos, em F1 e F3 (ver tabela 4.11).

F1			
	G.L.	F	p<
FALANTE	1	10.20	.0016
VOGAL	7	362.75	.0001
FAL. X VOGAL	7	2.71	.01
F3			
	G.L.	F	p<
FALANTE	1	23.72	.0001
VOGAL	7	12.15	.01
FAL. X VOGAL	7	2.54	.01

TABELA 4.10 Resultados de ANOVA incluindo no modelo a interação FALANTE X VOGAL (apenas gêmeos JA e JR).

F1/ã/	**n=**	**média**	**D.P.**
JA	12	608.3	95.5
JR	12	505.0	59.8
F3/ã/	**n=**	**média**	**D.P.**
JA	12	2456.0	92.9
JR	12	2784.0	110.87

TABELA 4.11: Médias e desvios-padrão de F1 e F3 da vogal /ã/. Apenas essa vogal é significativamente diferente (F1 e F3) nos dois gêmeos.

A diferença em F1 e F3 da vogal nasalizada /ã/ entre os gêmeos JA e JR não é de fácil interpretação. A princípio, não devemos esperar nesse caso uma diferença nas configurações anatômicas dos tratos nasal e oral, que pudesse justificar as diferenças espectrais observadas. É provável que os gêmeos JA e JR possuam, portanto, diferentes estratégias para a produção da qualidade nasal, envolvendo, por exemplo, diferentes graus de abertura do *velum* , um aspecto que, modificando as razões de abertura do acoplamento nasal/oral, pode alterar consideravelmente o grau de nasalização e, conseqüentemente, a configuração espectral (ver também seção 8, mais adiante, onde são comparadas produções da consoante nasal /n/ dos gêmeos JA e JR, em especial a figura 8.2).

A comparação dos formantes vocálicos dos gêmeos JA e JR indica que, efetivamente, a configuração anatômica impõe limites bastante definidos para a variação desses parâmetros. Mesmo incluindo a vogal /ã/, que apresentou uma divergência mais acentuada, as diferenças percentuais entre as médias dos formantes ficam em torno de 2 - 4.5 % , um valor da mesma ordem de grandeza do limite mínimo de discriminação diferencial para os formantes em questão (Flanagan 1972). É importante ressaltar, entretanto, que a similaridade entre os formantes só aparece claramente quando consideramos os valores médios; a comparação caso-a-caso (*matched-t*) mostrou que podem existir variações consideráveis entre os gêmeos (mesmo desprezando as vogais /ã/), ou seja, o componente anatômico idiossincrático não se manifesta diretamente sob a forma de valores absolutos locais, mas sim como uma limitação genérica na variação que um falante pode impor ao seu aparelho vocal.

4.4) *Formantes: Comentário Final*

O emprego das freqüências de formantes para a Identificação de Falantes envolve algumas dificuldades. Um dos problemas é estabelecer critérios seguros para a medição dos formantes. Os movimentos transicionais, especialmente em fala rápida, dificultam a determinação de um ponto exato para realizar a medida. Além disso, é preciso considerar que a recuperação dos picos da função de transferência (na verdade a informação mais relevante para estabelecer uma correspondência com a configuração articulatória do falante) depende de uma aproximação, que será tanto menos exata quanto mais alto for o F0 do falante. Ainda está para ser avaliado o efeito acumulativo dessas pequenas distorções na eficiência de medidas *target* de formantes para a Identificação de Falantes, mas, de modo geral, os resultados do experimento descrito nesta seção indicam que esse tipo de medida fornece informação dependente do falante, embora não se possa considerar essa informação como determinante inequívoca de identidade.

Outro tipo de medida, que assimile as próprias transições, talvez seja mais eficiente do que medidas baseadas em valores *target* estáticos (Cf. Goldstein 1976). Uma possibilidade aqui seria a definição de uma quantidade que expressasse aspectos dinâmicos da trajetória de um formante ao longo do segmento vocálico, como, por exemplo, a velocidade de mudança, dada pela primeira derivada. Uma abordagem mais sofisticada poderia estabelecer um gradiente que incorporasse a movimentação de dois ou mais formantes simultâneamente (Olivier, Com. Pessoal). Perceptualmente, há evidências de que, mesmo em segmentos vocálicos curtos, o ouvinte usa informação temporal para identificar o falante; Bricker e Pruzansky (1966) apresentam estímulos que são excertos de vogais com cerca de 100 milisegundos, verificando que, quando os excertos são apresentados em gravação

retrogradada há maior dificuldade para identificar o falante, indicando que os ouvintes, de algum modo, usam informação disponível apenas no curso temporal normal do segmento vocálico.

A utilização de pistas dinâmicas derivadas do *tracking* de formantes em sistemas automáticos exigiria o desenvolvimento de algoritmos para a extração das freqüências exatas ao longo do enunciado. Esse é um problema antigo que já motivou uma série de trabalhos, sem que, contudo, se tivesse chegado a uma solução totalmente satisfatória (Schaefer e Rabiner 1970; Olive 1971; Markel 1972; McCandless 1974). Trabalhos mais recentes, no entanto, têm relatado resultados promissores, através do uso de algoritmos mais robustos, apontando a possibilidade de extração do *tracking* de formantes até mesmo com a presença de ruído de fundo (Niederjohn e Lahat 1985).

A informação dinâmica extraída do *tracking* de formantes tem sido um dos elementos mais utilizados em esquemas de verificação automática de falantes, normalmente após algum tipo de alinhamento temporal (*time-warping*), onde os padrões (teste e referência) são alinhados com base em eventos selecionados do enunciado, geralmente definidos por inflexões em F1 ou F2 (Das e Mohn 1971; Goldstein 1976; Gomes 1993).

As dificuldades em extrair o *tracking* de formantes podem ser contornadas através do emprego dos coeficientes obtidos por meio da análise LPC. Esses coeficientes são computacionalmente mais simples de serem obtidos e fornecem praticamente a mesma informação que a análise de formantes. Em um certo sentido, a análise de formantes pode ser considerada como uma extensão da análise LPC (Rosenberg 1976). Um dos primeiros sistemas eficientes de Verificação Automática de Falantes, desenvolvido nos Laboratórios Bell, utilizou com sucesso os contornos de dois coeficientes LPC (A_4 e A_8, para uma análise de ordem 8) em substituição ao *tracking* de formantes, empregando essa informação juntamente com contornos de

F0 e amplitude (Rosenberg e Sambur 1975; v. também: Wakita 1975; Wood 1978; Bogner 1981; Furui 1981).

SEÇÃO 5: FREQÜÊNCIA FUNDAMENTAL

5.1) *Eficiência de Medidas de Freqüência Fundamental (F0) para Identificação de Falantes*

Todos já tiveram a experiência de identificar vozes familiares em cômodos adjacentes, pelo som irradiado através das paredes. Nessas situações, a filtragem *low-pass* exercida pelas paredes praticamente elimina a informação no nível segmental, restando apenas pistas de natureza prosódica relacionadas à variação de F0 e da intensidade. Essa experiência corriqueira indica que há uma boa parte de informação da identidade do falante veiculada apenas por parâmetros derivados de F0.

Vários experimentos têm confirmado a importância perceptual de F0 no reconhecimento de falantes. Abberton e Fourcin (1978) gravam sinais laringográficos que são apresentados isoladamente, eliminando assim qualquer informação supra-glotal (filtro); esses estímulos limitados, veiculando apenas informação da fonte, permitem, no entanto, um índice médio de reconhecimento de cerca de 70 %. Van Dommelen e Win (1987) obtêm, com o mesmo tipo de material, um índice ainda mais alto de acertos (72-84 %). Compton (1963), em um experimento de discriminação, apresenta aleatoriamente pares de falantes (n=9) produzindo vogais /i/ submetidas a vários tipos de segmentação e filtragem a 15 ouvintes familiarizados com as vozes, verificando que os falantes mais confundidos entre si são, sistematicamente, aqueles com F0 mais próximo.

As variações de F0 na fala parecem ser um aspecto perceptualmente bastante saliente. Talvez por esse motivo essa seja uma característica sistematicamente explorada por imitadores profissionais, que procuram aproximar-se do F0 médio do

imitado, embora, não consigam uma coincidência exata (Endres *et al.* 1971; Hall e Tosi 1975). Por outro lado, a saliência perceptual de F0 faz com que esse seja um parâmetro mais vulnerável a tentativas de disfarce (Doherty e Hollien 1978). Com efeito, a experiência forense nos tem revelado que, não raramente, durante a coleta de voz-padrão, o suspeito procura alterar algumas características de F0, em geral tendendo a uma média mais baixa e a uma menor variabilidade [1].

É provável que a atenção especial que damos a F0, na percepção, seja condicionada também por outros fatores, não diretamente ligados à necessidade de reconhecer um falante específico. Sabe-se, por exemplo, que o ouvinte utiliza informação entoacional (continuidade dos contornos de F0) para separar perceptualmente diferentes vozes simultâneas (Brokx e Nooteboom 1982; Chalikia e Bregman 1989). Informação de F0 também pode servir como auxiliar na normalização de vogais, agindo como um calibrador do sistema, ajustando o espaço vocálico perceptual para um falante particular (Johnson 1990). Embora nos dois exemplos (separação de vozes e normalização de vogais) não se possa falar propriamente de um mecanismo de reconhecimento do falante, é evidente que está em jogo um processo de construção de uma **representação** do falante.

Contornos de F0 produzidos por diferentes falantes, para a mesma sentença, apresentam, geralmente, uma maior variância inter-falante do que intra-falante (Jassem e Kudela-Dobrogowska 1980). A informação extraída de contornos entoacionais já foi utilizada com relativo sucesso em alguns sistemas experimentais de verificação automática de falantes; a comparação automática de padrões entoacionais é quase sempre precedida de um ajuste temporal (*time warping*) que normaliza a previsível variação duracional do enunciado em diferentes amostras (Atal 1972; Lummis 1973; Das e Mohn 1971).

Embora muitos experimentos apontem o F0 médio como o parâmetro mais importante para a identificação auditiva de falantes, em comparação com outro tipo de medida (espectro, amplitude, *speech rate*, etc) (v. p.ex. Compton 1963; Clarke e Becker 1969; Matsumoto *et al.* 1973; Abberton 1976), o mesmo não ocorre quando o critério de decisão é estatístico. Markel *et al.* (1977), com base nas razões de variância inter/intra-falante, verificam que parâmetros espectrais são mais importantes do que parâmetros derivados de F0. Doherty e Hollien (1978) avaliam a força relativa de 3 vetores, constatando que a performance, em ordem decrescente de eficiência é a seguinte: (1) LTS (consistindo das médias de longo termo das saídas de um banco de 23 filtros de 1/3 de oitava), (2) SFF (consistindo de 2 elementos baseados em F0: F0 médio e desvio padrão de F0) e (3) ST (2 elementos: tempo de fonação + taxa de articulação).

É possível que a divergência entre o desempenho de F0 na identificação auditiva e nos procedimentos automáticos exista por conta de uma representação deficiente nos vetores baseados em F0. Na verdade é difícil inferir diretamente dos testes perceptuais quais aspectos relacionados a F0 estão sendo efetivamente usados pelo ouvinte, especialmente quando são usados estímulos de fala fluente. Mais difícil ainda é quantificar pistas potencialmente relevantes na percepção. O emprego, por exemplo, de uma medida genérica como o desvio padrão para representar a variabilidade entoacional tende a neutralizar diferenças individuais e deixa de capturar uma série de características perceptualmente relevantes no curto termo. Hollien (1991:240) comenta que o desempenho insatisfatório das primeiras abordagens se deve ao uso de um número muito reduzido de parâmetros; resultados mais encorajadores, diz Hollien, têm sido obtidos com vetores envolvendo cerca de 30 parâmetros, incluindo medidas que incorporam variações mais locais de F0.

Existem basicamente dois tipos de medidas que podem, de algum modo, expressar a variabilidade de um contorno de F0: (1) as que levam em conta a

organização temporal e (2) as que são definidas sem considerar esse aspecto. Índices genéricos de variância (desvio padrão, coeficiente de variabilidade, etc) pertencem, obviamente, à segunda categoria, assim como medidas mais sofisticadas como os momentos mais altos da distribuição de F0 [2]. Descrições dessa natureza desprezam a seqüência exata das variações de F0 em função do tempo: o mesmo conjunto de medidas de F0 pode ser rearranjado em qualquer ordem aleatória sem que se alterem a média, o desvio padrão ou momentos mais altos da distribuição (Cf. Atal 1976:471).

Uma medida relativamente simples, e que captura a informação temporal, são as diferenças de várias ordens calculadas a partir dos valores da curva de F0. A curva formada pela diferença de primeira ordem consiste nas diferenças entre cada par adjacente de valores na curva original de F0; a diferença de segunda ordem consiste na diferença entre cada par adjacente na curva de primeira ordem, e assim por diante. Mead (1974; *apud* Nolan 1983:126) calculou diferenças até a quarta ordem, utilizando 5 curvas, incluindo o contorno original de F0; os 4 primeiros momentos foram calculados para cada uma dessas curvas, totalizando 20 parâmetros para cada falante.

A eficiência de parâmetros como as diferenças e os momentos depende muito da confiabilidade das medidas originais, já que erros eventuais tendem a ser amplificados após esse tipo de transformação (Markel *et al.* 1977). Essa pode ser uma limitação prática importante, já que os algoritmos conhecidos para extração de F0 têm um certo grau de falibilidade (Cf. Kent e Read 1992:78 ff).

Medidas derivadas das microperturbações de F0 (*jitter* e *shimmer* [3]) são freqüentemente utilizadas como um método de classificação de vozes com algum tipo de patologia laríngea, ou na monitoração durante o tratamento (Fritzell e Fant 1986). Embora essas aperiodicidades de F0 não tenham sido usadas para a Identificação de Falantes, é possível que sejam potencialmente eficazes nesse

sentido, visto que valores anômalos de *jitter* ou *shimmer* estão associados a determinadas qualidades de voz (v. nota 3). A extração de medidas efetivas de microperturbações de F0, no entanto, exige condições ideais de captação do sinal e, para maior confiabilidade, a fonação sustentada de vogais por cerca de 2 segundos (Titze *et al.* 1987), uma condição que dificulta seu uso no modelo forense [4]. Recentemente, o desenvolvimento de algoritmos mais robustos tem apontado resultados encorajadores na extração de medidas de micro-perturbações de F0 mesmo sob ruído (especialmente de alta-freqüência; v. Johnson *et al.* 1990:233).

5.2) *Variações Intra-Falante*

Assim como outros parâmetros acústicos extraídos da fala, F0 interage com uma série de fatores de ordem lingüística, além de ser uma das pistas mais importantes para a veiculação dos aspectos afetivo/emocionais. Todos esses componentes contribuem, em maior ou menor grau, para definir o âmbito da variação intra-falante de parâmetros relacionados com F0. Uma das questões mais importantes para o uso eficiente de F0 como indicador de identidade é avaliar essa variação intra-falante, analisando as causas e a extensão dos efeitos provocados por diferentes condições (lingüísticas e não-lingüísticas). A presente seção pretende discutir alguns fatores que podem influir diretamente na variação intra-falante de F0.

5.2.1) *Variação em Amostras Não-Contemporâneas*

Já se verificou que amostras de fala colhidas em diferentes momentos podem apresentar variações de F0. Garret e Healey (1987) observam flutuações do F0 médio ao longo do dia, efetuando 3 medidas, uma pela manhã (8 - 9 h), uma segunda no começo da tarde (11.30 - 12.30 h) e a última no fim da tarde (16 - 17 hs), a partir de leituras de um mesmo texto por falantes adultos do sexo masculino e feminino; eles verificam que os homens apresentam um aumento pequeno - mas estatisticamente significativo - da manhã para a tarde, enquanto as mulheres não apresentam um padrão regular na variação ao longo do dia. Brown *et al.* (1976) estudam o mesmo tipo de variação, mas para um intervalo de tempo maior, compreendendo medidas ao longo de 5 dias. Brown *et al.* verificam uma variação média de cerca de 15 Hz para 8 dos 18 sujeitos do grupo; alguns falantes chegam a variar mais de 20 Hz, enquanto outros praticamente não alteram seu F0 médio. Sambur (1975b), analisando dados colhidos ao longo de 3.5 anos, verifica que o F0 médio pode variar consideravelmente ao longo das diversas medidas nesse espaço de tempo. Markel e Davis (1979), testando um esquema de verificação automática de falante, realizam gravações com intervalo de 2-3 semanas entre sessões, ao longo de 3 meses, usando como material conversação espontânea (entrevistas com tema livre); cada falante (n=17) participa de 10 sessões, em média, de 13 minutos cada, totalizando cerca de duas horas de material gravado. Markel e Davis verificam, da mesma forma que os outros estudos, que há uma pequena variação no F0 médio ao longo das diferentes sessões, cuja magnitude, porém, depende do falante. Markel e Davis estudam também o comportamento do desvio padrão, observando que existe uma dependência entre essa medida e a quantidade de material usado, mais

especificamente uma relação diretamente proporcional entre o valor do desvio padrão de F0 e o inverso da raiz quadrada do número de *frames* da amostra utilizada.

Atkinson (1976) estuda a variação inter- e intra-falante de contornos de F0 em diferentes produções ao longo de um ano, para um grupo de 5 falantes do sexo masculino, a partir de uma única sentença (*Bev loves Bob*), com diferentes acentuações frasais em forma declarativa e interrogativa. Atkinson verifica que, embora as formas gerais dos contornos para cada tipo de sentença sejam similares, existem traços característicos do falante relacionados ao *timing*, extensão dos movimentos de F0 nos picos acentuais e F0 médio. Atkinson verifica que, na média, a variabilidade inter-falante é um pouco maior do que a variabilidade intra-falante, mas há casos onde o oposto pode ocorrer (alguns falantes variam consideravelmente seus contornos ao longo das diversas amostras). Atkinson (1976:441) sugere que a variação intra-falante pode ser de dois tipos, uma associada a um "componente estático",

> *that shifted the entire contour up or down by changing the average F0,*

e um "componente dinâmico",

> *[that] consists of random interleavings that affect the relative shape within a contour.*

Segundo Atkinson, mudanças absolutas em F0 (componente estático) atuariam em um nível paralingüístico, veiculando informação sobre o sexo, idade, identidade e estado afetivo/emocional do falante, enquanto o componente dinâmico estaria mais fortemente relacionado com os traços lingüísticos. Essa afirmação de

Atkinson é discutível, e até mesmo um tanto contraditória, já que no mesmo artigo, como comentamos acima, o autor reconhece a existência de

> *idiosyncratic differences in timing, degree of F0 rise or fall (...) which are consistent for a given speaker* (Atkinson 1976:441).

Assim, o componente dinâmico, definido por Atkinson, parece também veicular alguma informação do falante. Com efeito, alguns sistemas automáticos já empregaram com sucesso características dinâmicas de contornos de F0 para a formação de vetores de identificação, utilizando, geralmente, essa informação conjugadamente com outras pistas (v. p.ex: Atal 1972; Lummis 1973; Rosenberg e Sambur 1975; Doddington 1985).

Atkinson observa que o componente estático depende basicamente do valor inicial de F0, no *onset* da fonação, afetado por uma variável aleatória determinada por uma combinação de diversas condições no momento do *onset*, incluindo variações na pressão sub-glotal, diferenças no volume de ar nos pulmões e diferentes estados dos músculos associados à laringe. Esse aspecto vai de encontro à hipótese sugerida por Crystal (1969:143), onde se afirma que

> *For any speaker, the first prominent syllable of a tone-unit is articulated at or around a stable pitch level for the majority of his tone-units.*

É interessante observar, ainda com respeito ao trabalho de Atkinson, que as distribuições de F0 de dois falantes podem ser consideravelmente diferentes, apesar de o F0 médio ser eventualmente muito próximo (compare-se, por exemplo, os falantes GK e JA - Atkinson 1976: figura 2, pg. 442 -, onde, apesar das médias

próximas - 103 e 114 Hz, respectivamente - as distribuições são marcadamente platocúrtica, para GK, e leptocúrtica, para JA). Informações mais detalhadas sobre as distribuições, como o grau de *skewness* e curtose podem, portanto, ser fundamentais para diferenciar falantes com F0 médios próximos ou semelhantes.

Variações ao longo de intervalos de tempo maiores também já foram estudadas. Helfrich (1979) resenha uma série de experimentos comparando os valores de parâmetros relacionados a F0 em diferentes faixas de idade. Os resultados relatados por Helfrich são consistentes, indicando que o F0 médio, para homens e mulheres, decresce até os 20 anos de idade e estaciona; para os homens, no entanto, há uma tendência a um novo acréscimo a partir dos 65 anos. Helfrich associa a queda do F0 médio na maturidade à calcificação na estrutura da laringe e à diminuição da elasticidade cartilaginosa e muscular. Mais difícil de explicar é o aumento do F0 médio após os 65 anos, nos falantes do sexo masculino; Hollien e Ship (1972), constatando o mesmo tipo de fenômeno, associam o aumento do F0 médio na velhice a diversos fatores: (a) atrofia do sistema nervoso central; (b) aumento da pressão sangüínea; (c) alterações no sistema respiratório e (d) mudanças endócrinas e musculares. Helfrich (1979) comenta, pertinentemente, que o resultado pode ser apenas um subproduto do *design* experimental: todos os experimentos analisados baseiam-se em dados sincrônicos, sendo a comparação feita entre grupos contemporâneos com diferentes idades. A média relativamente maior nos falantes mais idosos relacionar-se-ia com o fato de que o tamanho médio do corpo (especialmente dos homens) cresceu nas últimas décadas; assim, a maior massa na região da laringe faria com que o F0 médio dos mais jovens fosse relativamente menor. A observação de Helfrich é suportada pelos resultados de Endres *et al.* (1971), um dos raros (se não o único) estudo longitudinal enfocando a variação de F0 ao longo de diferentes fases da vida de um indivíduo; com efeito, analisando o

comportamento do F0 médio ao longo de 15 anos para falantes com diferentes idades iniciais, Endres *et al.* verificam que, para todas as faixas de idade inicial, há um decréscimo constante do F0 médio, mesmo após os 60 anos de idade.

A variabilidade de F0 também parece diminuir com a idade. Helfrich (1979) observa que a maior parte dos estudos indica um decréscimo do desvio padrão a partir da meia-idade. O estudo longitudinal de Endres *et al.* (1971:1844) verifica o mesmo fenômeno:

> *The individual distribution curves become more narrow with increasing age, i.e., the speakers seem to lose the ability to vary the fundamental frequency of their voices.*

Helfrich (1979) sugere que a diminuição da variabilidade dos contornos de F0 na velhice não deve estar associada exclusivamente a aspectos orgânicos, mas também a fatores de natureza sócio-cultural (maior isolamento social, por exemplo).

Linville (1988), estudando as micro-perturbações de F0 em dois grupos de mulheres (jovens *vs.* idosas) produzindo fonações sustentadas de vogais /i,a,u/, relata que há um aumento significativo no *jitter* no grupo mais idoso. Linville observa que, curiosamente, as falantes mais jovens diferem marcadamente das mais velhas quanto ao padrão de *jitter* em relação às 3 vogais testadas; as falantes mais jovens apresentam níveis mais baixos de *jitter* na vogal /a/, enquanto as mais velhas apresentam o maior grau de instabilidade exatamente nessa mesma vogal. Linville esboça uma explicação muscular para o fenômeno, sugerindo que as vogais altas (especialmente /i/) envolveriam ajustamentos articulatórios mais pronunciados, afetando diretamente as cordas vocais, fazendo com que essas vogais se tornem mais resistentes aos efeitos da idade quanto à instabilidade de F0.

O aumento das micro-perturbações de F0 na idade avançada está provavelmente associado a problemas de coordenação do sistema nervoso central (Helfrich 1979).

5.2.2) *Variações em Função do Estado Afetivo-Emocional*

Grande parte da variabilidade intra-falante de parâmetros relacionados a F0 se dá por conta da necessidade de expressar, através da fala, estados afetivos/emocionais. Diderot (1767; *apud* Fónagy 1981) já observara que a entonação seria a "expressão natural dos estados da alma". Williams e Stevens (1972) sugerem que F0 é particularmente apropriado para veicular conteúdos emocionais, visto que, especialmente em línguas não tonais, há poucos aspectos lingüísticos que dependem diretamente de F0; assim, com exceção de marcas acentuais e a sinalização de fronteiras sintáticas, o falante seria relativamente livre para variar F0 [5].

Um grande esforço de pesquisa já foi dedicado na observação de variações de parâmetros relacionados a F0 em função de modificações no estado afetivo/emocional. Um estudo pioneiro, nesse sentido, foi realizado por Lieberman e Michaels (1962), onde, através de manipulações do sinal original foram criados estímulos sem informação segmental contendo apenas diferentes combinações de pistas prosódicas (contorno de F0, F0 + modulação de amplitude, F0 fixo + amplitude, etc); Lieberman e Michaels verificam que a informação dada pelo contorno isolado de F0 é suficiente para que os ouvintes reconheçam grande parte das atitudes intencionadas pelos falantes. Fónagy (1978) em um experimento parecido, utilizando gravações laringográficas, obtém resultados semelhantes.

Com relação às emoções "simples", há em geral convergência quanto às modificações observadas em F0 nos diferentes estudos. A manifestação vocal de

"alegria" (*joy*), por exemplo, parece estar sempre vinculada a um aumento no F0 médio e na variabilidade do contorno, em comparação com o estado "neutro"; o mesmo pode não ocorrer com estados emocionais mais complexos, tais como "ironia", "desprezo", etc, quando os resultados nem sempre são coincidentes (v. p. ex. Fónagy e Magdics 1963; Zlatoustova e Kedrova 1987; Nushikian 1987). O problema, em geral, é a dificuldade em definir rótulos não ambíguos para esses estados afetivos mais complexos (tanto para o falante - um ator, na maior parte dos casos - quanto para os juízes ouvintes).

Uma forma de contornar o problema com a rotulação dos estados emocionais é basear o julgamento em categorias mais abstratas, utilizando a técnica do *Diferencial Semântico*, desenvolvida por Osgood *et al.* (1957). Segundo essa técnica, os ouvintes julgariam cada estímulo atribuindo valores a um conjunto de escalas bi-polares representando diferentes dimensões psicológicas; uma análise estatística posterior criará, a partir do perfil de respostas dos juízes, um conjunto de fatores, agrupando cada um algumas das dimensões originais. Um dos primeiros trabalhos a usar o diferencial semântico para interpretar as respostas dos ouvintes foi realizado por Uldall (1961); nesse experimento aplicou-se sinteticamente 16 tipos diferentes de contorno de F0 a um conjunto de sentenças, controlando-se a tessitura, o F0 no final da sentença, a forma geral do contorno e relações acentuais entre sílabas fracas e fortes. Após a análise multi-dimensional Uldall resume as respostas a apenas 3 fatores: *pleasant/unpleasant, authoritative/submissive* e *strong/weak*, verificando que alguns aspectos de F0 estão fortemente correlacionados a essas dimensões: uma maior tessitura de F0, por exemplo, está associada a altos valores de *strong* e *authoritative*, enquanto uma tessitura estreita associa-se ao fator *unpleasant*.

Fónagy e Bérard (1972) realizam um experimento, também baseado em respostas em diferencial semântico. A partir de uma única frase, "*Il est huit heures*",

são sugeridos 26 diferentes contextos situacionais a uma atriz, que deve produzir a entonação adequada àquela situação. Os estímulos assim produzidos foram avaliados por um grupo de sujeitos (n=80) em um diferencial semântico de 9 dimensões; os sujeitos deveriam também descrever brevemente que tipo de situação o estímulo sugeria. Fónagy e Bérard verificam que os dois tipos de resposta podem variar bastante entre os ouvintes para situações complexas; por outro lado, as situações que evocam emoções elementares (medo, cólera, etc) apresentam uma maior convergência nas respostas. O mais relevante nesse estudo é a constatação de uma grande diversidade nas realizações da sentença proposta, cuja simplicidade é notável. No que diz respeito especificamente a F0, observou-se uma grande variedade de formas de contorno, tessitura e F0 médio [6]. Deve ser ressaltado, entretanto, que a mensagem emocional aqui não depende apenas do contorno de F0, mas também de alterações na qualidade de voz, amplitude, etc.

5.2.3) *Efeitos da Presença de Stress Psicológico* [7]

A determinação de pistas vocais relacionadas a situações envolvendo algum tipo de *stress* psicológico tem motivado, nas últimas décadas, uma série de estudos, em função, principalmente, das possibilidades de aplicação militar (detecção de *stress* em pilotos de prova, astronautas, paraquedistas, etc; Cf. Disner 1982). Na área forense o interesse concentra-se na possível aplicação desse tipo de informação no desenvolvimento do tão polêmico "detector de mentiras" (Hollien *et al.* 1987; Disner 1982), mas a questão é relevante também no sentido de avaliar a extensão da variação de certos parâmetros vocais (especialmente relacionados a F0) em situações envolvendo *stress*; assim, pode ser importante verificar o tipo e a magnitude de alterações vocais no suspeito, ao ser submetido à coleta de voz-padrão (uma situação

que, em geral, envolve alguma pressão psicológica), ou determinar se a voz questionada carrega traços que apontem a presença de *stress*.

Do ponto de vista fisiológico, o corpo responde a situações envolvendo *stress* através da ativação da pituitária e da liberação de adrenalina, de modo a produzir hormônios cuja função principal é proteger preventivamente o corpo de maiores prejuízos. Esse processo provoca alterações significativas na pulsação, pressão arterial, respiração, tensão muscular, além de outras mudanças menos aparentes, mas que possuem, todas, o potencial de afetar sensivelmente o delicado mecanismo da laringe (Levine 1971). A sensibilidade da laringe aos efeitos do *stress* faz dos parâmetros derivados de F0 os principais candidatos a alterações significativas nessa situação.

A principal dificuldade nessa área de pesquisa é a própria simulação em laboratório de situações que provoquem *stress* psicológico. Existem aqui duas abordagens: (1) *stress* provocado por execução de tarefa complexa; (2) *stress* relacionado a comportamento mentiroso [8].

Um exemplo da estratégia (1) é dado pelo estudo de Hecker *et al.* (1968). Nesse experimento, o *stress* é induzido através da realização de operações aritméticas sob pressão de tempo; para garantir um maior envolvimento com a tarefa, estabeleceu-se uma remuneração para cada operação corretamente resolvida. Os autores verificam vários efeitos relacionados a F0, mas há uma considerável variação inter-falante na entensão e até mesmo na direção desses efeitos. O F0 médio, por exemplo, pode sofrer um acréscimo ou decréscimo sob *stress*; a tessitura para a maioria dos falantes amplia-se sob *stress*, mas alguns sujeitos apresentam o quadro oposto. Os efeitos mais consistentes foram o aumento dos índices de *jitter* e *shimmer* e a presença de alguns períodos glotais alongados e irregulares no final de grupos tonais.

O estudo das pistas vocais para a detecção de comportamento mentiroso (abordagem 2) causa certamente mais problemas para o pesquisador do que a indução de *stress* por tarefa complexa. A grande dificuldade aqui é configurar uma situação que efetivamente simule uma situação real, garantindo que o sujeito esteja suficientemente envolvido no experimento de modo a produzir uma resposta satisfatória. Vários projetos experimentais já foram testados, com diferentes níveis de compromisso do sujeito com a tarefa, criando assim, diferentes níveis de *stress* no sujeito (v. Disner 1982). Os resultados desses estudos, apesar da diversidade das condições experimentais, apontam para um efeito consistente, mais especificamente um aumento no F0 médio durante a tentativa de fraude e, em alguns casos, um aumento na variabilidade de F0 (v. p.ex. Ekman *et al.* 1976; Streeter *et al.* 1977; Scherer *et al.* 1985). Scherer *et al.* (1985) sugerem que podem estar presentes dois processos, um de natureza fisiológica, envolvendo um maior tensionamento das cordas vocais em função do maior *stress* na condição mentirosa e outro, de natureza psicológica, envolvendo uma estratégia de *self-presentation*, onde o falante empresta maior expressividade à sua fala (talvez pela tentativa de dissimulação) aumentando assim o F0 médio e o desvio padrão.

Alguns (raros) estudos se valeram da possibilidade de estudar os efeitos do *stress* na voz a partir de eventos da vida real, aproveitando gravações realizadas quando os falantes se encontravam diante de um perigo real iminente, ou em contextos situacionais envolvendo grande pressão psicológica. Um clássico estudo com esse tipo de material é Williams e Stevens (1972). Nesse trabalho é analisada a voz do locutor de rádio que descrevia a chegada do dirigível *Hindenburg*, quando, repentinamente, o aparelho explodiu. Os autores verificam que durante toda a narração do desastre há um aumento considerável tanto na mediana de F0 (de 166 para 196 Hz) quanto na extensão da variação de F0 (de 124-196 para 152-260 Hz),

embora a fala do locutor já apresentasse uma grande variabilidade de F0 antes do incidente (de acordo com o estilo habitual desses profissionais). Além dessas alterações, Williams e Stevens observam o aparecimento esporádico de irregularidades em F0, manifestadas como uma espécie de tremor nos espectrogramas de banda estreita. Os autores sugerem que essas irregularidades poderiam refletir a perda de controle fino da musculatura e a presença de um padrão respiratório anômalo, reações fisiológicas ao *stress* situacional.

No mesmo trabalho, Williams e Stevens comparam a gravação original do locutor com a performance de um ator ao qual se deu a oportunidade de estudar o texto transcrito da transmissão radiofônica original (sem contudo ter acesso à gravação original). O ator deveria simular a situação, como se estivesse descrevendo a catástrofe. As alterações em F0 na fala do ator foram na mesma direção das observadas no locutor original, mas ainda de maior magnitude (mediana de F0 aumenta de 138 para 222 Hz e extensão de F0 de 117-168 para 117-280 Hz). A grande coincidência entre os efeitos vocais do ator e do locutor coloca a possibilidade de estar em jogo algum tipo de estereótipo cultural, associando convencionalmente certos padrões de fala a situações características de *stress* (mesmo na produção do locutor, um profissional que, assim como o ator, deve ter um grande domínio desse tipo de convenção dramática).

Streeter *et al.*(1983) descrevem outro estudo envolvendo uma situação real. O material consiste na gravação de uma longa conversação telefônica entre um funcionário da Companhia de Eletricidade de New York e seu superior imediato, começando uma hora antes do *blackout* de 1977 e terminando aproximadamente no momento do *blackout* total. Um dos aspectos mais interessantes desses dados é a possibilidade de estudar os efeitos de um *stress* presumivelmente crescente, ao longo de um intervalo de tempo relativamente largo (e, portanto, com um vasto material de fala); essas características, de uma certa forma, aproximam o evento estudado do

contexto mais controlado do laboratório, sem, contudo, depender de simulações de comportamento. Os resultados relatados por Streeter *et al.* são um tanto surpreendentes, em uma primeira análise; as modificações nos parâmetros relacionados a F0 ao longo da uma hora de conversação apresentam direções opostas para os dois interlocutores: enquanto a média e desvio padrão de F0 decrescem gradativamente para o operador, o oposto ocorre com seu chefe. Os autores concluem que esses resultados vão contra a existência de uma reação única e estável a situações de *stress* psicológico (pelo menos no que diz respeito a medidas relacionadas com F0), levantando, entretanto, uma segunda alternativa: o funcionário subalterno, na medida em que transferia a responsabilidade de decisão para seu chefe, aliviava a pressão sobre si próprio. Os próprios autores reconhecem a natureza *ad hoc* dessa hipótese, mas, devido às características da situação, não é possível descartá-la totalmente.

Qualquer que seja a interpretação dos resultados de Streeter *et al.*(1983), o estudo ressalta os perigos de associar diretamente parâmetros acústicos (F0 ou qualquer outro) a estados internalizados de um indivíduo [9]. No que diz respeito aos efeitos em F0 em situações de *stress* (real, induzido ou simulado), os resultados de diversos experimentos indicam apenas, com segurança, que há uma **modificação** em relação aos padrões habituais do falante, mas nem sempre na mesma direção (embora a tendência predominante seja um acréscimo de F0 médio sob *stress*). Resultados mais confiáveis exigiriam um exame mais detalhado das condições experimentais, com maior controle (se for possível esse controle!) do **grau** de *stress*, da **autenticidade** ("real", simulado, +/- estereotipado, etc) e da **qualidade** (tentativa de fraude, medo, tensão, ansiedade, etc).

5.2.4) *Outras Variações*

A sensibilidade das estruturas laríngeas a diversos tipos de modificações metabólicas faz com que os parâmetros relacionados a F0 sofram algum tipo de variação em função de um grande número de fatores. Inflamações laríngeas provocadas por uma simples gripe podem ter um efeito considerável nas medidas de F0 (Rosenberg 1976:479; v. também, mais abaixo - seção 5.3, os resultados para o falante R1/R2). A quantidade e consistência do muco lubrificante na laringe podem afetar certas características vibratórias, tornando mais difícil o controle fino de F0; essas condições podem ocorrer durante estados de excitação sexual, em homens e mulheres, assim como em mulheres grávidas ou em estado pré-menstrual (Laver e Trudgill 1979:13). O próprio estado corporal geral pode afetar F0, na medida em que a fadiga modifica a elasticidade muscular; assim, o cansaço físico tende a diminuir o F0 médio e reduzir a variabilidade de F0 (Abe 1980).

Condições psicológicas transientes, tais como estados depressivos, podem alterar certas características de F0, em geral produzindo uma queda na variabilidade dos contornos entoacionais (Nilsonne *et al.* 1988) (v. também Chevrie-Muller *et al.* 1978;1985, para resenhas de vários estudos sobre os efeitos de distúrbios mentais em F0).

A produção de fala sob ruído está normalmente associada a um aumento no F0 médio; esse efeito pode estar relacionado a um aumento no esforço vocal nessas condições (van Summers *et al.* 1988).

De grande interesse para a aplicação forense é o estudo das alterações provocadas pela intoxicação alcoólica. Sabe-se que níveis mais altos de álcool no sangue perturbam, entre outros fatores neuro-motores, os controles aferente e eferente da laringe (Schenker 1982); assim, há uma expectativa de que aspectos

relacionados à vibração glotal sofram alguma alteração em indivíduos alcoolizados. Klingholz *et al.*(1988) estudam os efeitos vocais em níveis controlados de álcool no sangue (variando gradativamente de .05 % a .15 %, em passos de + .01 %, para cada indivíduo). Os sujeitos (n=11) leram um texto fixo (Alemão) sob todos os níveis graduais de intoxicação alcoólica, e diversos parâmetros acústicos foram medidos em cada condição (sóbrio e com .05%, .06% ... .15 %). O único efeito isolado estatisticamente significativo observado foi relacionado a modificações na distribuição de F0; a intoxicação alcoólica provocou um aumento considerável da variabilidade de F0 e um pequeno deslocamento para cima do valor modal de F0 (embora a média não tenha praticamente se alterado). Os autores observam que o álcool no sangue causa um inchaço nas cordas vocais, mas essa modificação deveria provocar um decréscimo de F0, que não foi observado; os autores sugerem que o aumento de massa nas cordas vocais pode ter sido compensado por uma maior tensão muscular, neutralizando assim os efeitos no F0 médio. Foi observada uma variação inter-falante na magnitude dos efeitos, provavelmente em função dos diferentes níveis individuais de resistência à absorção do álcool. É importante ressaltar que nesse experimento as taxas de intoxicação alcóolica são relativamente baixas; alterações mais expressivas em F0 devem ser esperadas com níveis mais altos de álcool no sangue.

Um interessante estudo envolvendo um caso forense real onde havia suspeita de intoxicação alcoólica é apresentado por Johnson *et al.* (1990). Nesse trabalho é analisada a fala produzida pelo comandante do *Exxon Valdez*, o navio petroleiro que, em março de 1989, provocou um grande desastre ecológico no Alasca. Durante a investigação surgiu a suspeita de que o comandante do navio estava alcoolizado no momento do acidente, um fator que poderia ter contribuído para o desenrolar dos fatos (e que poderia ter implicações no processo legal envolvendo as companhias seguradoras). A única evidência disponível eram as gravações das transmissões de

rádio realizadas entre o comandante e o posto da Guarda Costeira. Johnson *et al.* analisaram cinco amostras diferentes desses diálogos, gravadas em diferentes momentos: 33 horas antes do acidente, uma hora antes do acidente, imediatamente após o acidente, uma hora após o acidente e nove horas após o acidente. Além dessas amostras, foi utilizada, para efeito de confronto, a gravação de uma entrevista dada pelo comandante a uma estação de TV, um ano após o desastre. Para a extração de F0 foram utilizadas sempre as mesmas frases (*Exxon Valdez* e *thirteen and sixteen*), escolhidas em função de seu aparecimento freqüente na conversação, em posições comparáveis nos contextos sentenciais e discursivos. Medidas de F0 foram tomadas nas vogais das frases escolhidas e comparadas entre as diferentes amostras. Johnson *et al.* verificam um decréscimo gradual do F0 médio até o momento do acidente (amostras a -33 h, -1 h e imediatamente após); após o acidente o F0 médio torna a subir (+1 h, +9 h); o F0 médio da amostra mais distante do momento do acidente (- 33 h) está aproximadamente na mesma faixa do F0 médio observado na entrevista usada como controle. Observou-se também uma tendência a um aumento no nível de *jitter* nas amostras temporalmente mais próximas ao acidente. Johnson *et al.*, através dessas e de outras evidências (efeitos nas durações, mudanças no espectro de /s/, etc), sugerem que o comandante estaria realmente alcoolizado no momento do acidente. A possibilidade de que as alterações de F0 estivessem relacionada à presença de *stress* é descartada, já que os efeitos aparecem gradualmente **antes** do acidente. Outra possibilidade, associando o fenômeno observado ao aumento da fadiga (não vocal mas geral) também é improvável, pois as amostras 9 horas **após** o acidente, quando se esperaria encontrar um nível maior de fadiga corporal, são mais semelhantes à amostra tomada 33 h antes do acidente do que à amostra **durante** o acidente.

Os estudos sobre os efeitos da intoxicação alcoólica em F0 parecem apontar para várias direções, do mesmo modo que os experimentos envolvendo *stress*

psicológico, já comentados acima (v. 5.2.3). Klingholz *et al.* (1988) não observam alteração significativa no F0 médio, enquanto Trojan e Kryspin-Exner (1968) verificam um aumento significativo nessa medida em falantes alcoolizados. Vários estudos resenhados em Johnson *et al.* (1990) - assim como suas observações no caso *Exxon Valdez* - sugerem uma diminuição do F0 médio. Diferenças inter-subjetivas no mesmo experimento também podem ocorrer (Pisoni e Martin 1989; *apud* Johnson *et al.* 1990:219). Assim como no caso do *stress* psicológico, há uma grande dificuldade aqui em controlar os aspectos individuais internos que podem interagir com as condições controladas do experimento. O grau de resistência ao álcool, todos sabemos, varia bastante de indivíduo para indivíduo; assim, o controle exato da dosagem (mesmo em função do peso individual) não garante efeitos de mesma magnitude. Pode haver também uma interação com o perfil de personalidade do sujeito, de modo que a intoxicação alcoólica potencialize traços já existentes; desse modo, um indivíduo naturalmente extrovertido pode, sob o efeito do álcool, "amplificar" certos traços vocais, tais como o F0 médio e a variabilidade de F0, cujos valores acima da média estão normalmente associados à personalidade extrovertida (Cf. Scherer 1978;1979).

5.3) *Medidas de F0 : Análise Estatística*

Na presente seção estudaremos as variações intra- e inter-falante de conjuntos de medidas de F0. Dois tipos de distribuição serão analisadas: (1) medidas obtidas apenas nos núcleos de vogais lexicalmente tônicas e (2) medidas obtidas automaticamente via LPC.

Na situação forense é bastante comum termos como material para análise gravações de baixa qualidade, com nível alto de ruído de fundo ou outros tipos de distorção. Algoritmos para a extração automática de F0, mesmo em condições

ideais, podem cometer alguns erros (Kent e Read 1992:78 ff). No caso de gravações degradadas é praticamente inviável o emprego de métodos automatizados para extração de F0. Em fitas de baixa qualidade, no entanto, mesmo com relações críticas sinal/ruído, é possível medir F0 manualmente através de espectrogramas de banda estreita, especialmente em pontos de maior amplitude do sinal, como em núcleos de vogais lexicalmente tônicas; nessas regiões, os harmônicos ficam bem definidos, apesar do ruído de fundo, e o F0 pode ser medido pela distância entre dois ou - preferencialmente - mais harmônicos (Cf. Figueiredo *et al.*1993).

Uma questão que pode surgir em relação a esse método é se o conjunto de medidas assim obtido é representativo ou se insere o deslocamento de alguns parâmetros estatísticos, se compararmos os resultados com medidas extraídas automaticamente ao longo da amostra de fala inteira (e não apenas nos núcleos vocálicos). Essa possibilidade, pelo que temos conhecimento, nunca foi analisada em outros estudos. As seções 5.3.1 e 5.3.2 apresentarão resultados para os dois tipos de medida, de modo a que se possa avaliar as eventuais diferenças entre os dois métodos.

5.3.1) *Medidas de F0 nos Núcleos Vocálicos*

Nessa fase do experimento o F0 de cada vogal foi medido diretamente da forma de onda em um ponto localizado no máximo de amplitude de cada núcleo vocálico. Apenas as vogais lexicalmente tônicas foram selecionadas (para o critério de seleção das vogais e metodologia de gravação, ver seção 3). Mediu-se o intervalo de tempo entre dois picos, visualmente identificados na tela do sonógrafo DSP-5500 da KAY Elemetrics. Em nenhum caso houve dificuldade em determinar dessa forma os períodos. O valor exato de F0 foi então calculado como:

intervalo de tempo entre dois picos, visualmente identificados na tela do sonógrafo DSP-5500 da KAY Elemetrics. Em nenhum caso houve dificuldade em determinar dessa forma os períodos. O valor exato de F0 foi então calculado como:

$2 \div \Delta t$, sendo Δt o intervalo de tempo entre dois picos sucessivos [10]

A tabela 5.1 apresenta os resultados de estatística descritiva básica a partir de medidas de F0 obtidas tal como acima descrito. Nessa tabela estão reunidos todos os falantes do grupo principal (n=9; 7 + R1/R2), com resultados para as duas condições de velocidade de emissão (normal *vs.* rápida) separadamente e agrupadas. A tabela inclui também resultados para as mesmas medidas após transformação logarítmica ($LOG_2\ F0_{Hz}$), denominadas Flog0.

Velocidade Normal							
	média	**median**	**moda**	**D.P.**	**SK**	**KU**	**n =**
F0	130.8	130.0	140.2	27.2	.07	-.34	674
Flog0	127.9	130.0	140.2	-	-.50	.16	674
Velocidade Rápida							
	média	**median**	**moda**	**D.P.**	**SK**	**KU**	**n =**
F0	141.3 .	143.3	151.7	28.9	-.08	-.02	672
Flog0	138.1	143.3	151.7	-	-.67	.21	672
Velocidade Normal + Rápida							
	média	**median**	**moda**	**D.P.**	**SK**	**KU**	**n =**
F0	136.0	138.4	140.2	28.5	.02	-.19	1346
Flog0	132.9	138.4	140.2	-	-.56	.11	1346

TABELA 5.1: Estatística descritiva básica para F0 e Flog0 (= $LOG_2\ F0_{Hz}$). Para maior clareza os resultados obtidos a partir da escala LOG foram reconvertidos para Hz. SK é o grau de assimetria da distribuição no eixo vertical (*skewness*), e KU é a *Curtose*, ou seja, o grau de achatamento da distribuição. **n** é o número de medidas isoladas usadas em cada cálculo.

Pela tabela 5.1 podemos observar que a transformação logarítmica afeta alguns aspectos da distribuição. O grau de assimetria vertical (*skewness*) torna-se consideravelmente negativo, ou seja, a distribuição estende-se mais longamente para a direção dos valores mais baixos. Essa mudança é esperada, já que a escala LOG tende a diminuir a influência de valores anomalamente altos (Engelman 1990:138). Em termos absolutos, porém, a escala Hz produz distribuições mais próximas da normal, quanto ao *skewness*. Quanto ao grau de achatamento, ou *Curtose*, há uma tendência a distribuições mais concentradas (leptocúrticas) na escala LOG. De modo geral, as distribuições em Hz parecem mais próximas da normalidade e não há motivo para empregar transformações logarítmicas nesse caso. Nas medidas de tendência central (média, mediana e moda), a única diferença observada entre as duas escalas, obviamente, é na média, com uma pequena diminuição na escala LOG; a divergência, entretanto, fica em torno de 2 %, e pode ser considerada irrelevante para a variável em questão.

Uma informação relevante na tabela 5.1 é a indicação de um aumento considerável, de cerca de 9.5 % na média (Hz) das amostras na condição velocidade rápida de emissão. A tabela 5.2 mostra os resultados de uma análise de variância (ANOVA - BMDP-7D) testando a significância estatística dessa diferença. A tabela 5.2 apresenta também os intervalos de confiança (95%) para as duas condições (velocidade normal *vs.*rápida).

ANOVA (BMDP-7D)							
VAR.	**Grupo**	**G.L.**	**F=**	**p<**	*Levene F*	*p<*	**n=**
F0	velocid.	1	46.77	.0001	.29	.59	1350

95% CONFIDENCE INTERVALS

```
(norm)  L_____M_____U
(rapid)                                  L_____M_____U
        +---------+---------+---------+---------+---------+
     128.4     131.4     134.4     137.4     140.4     143.4
```

TABELA 5.2: Resultados de ANOVA (BMDP-7D) testando a diferença entre F0 na velocidade normal *versus* rápida (grupo principal de falantes; n=9). Os resultados apontam uma diferença estatisticamente significativa entre as duas condições de produção.

A tabela 5.2 indica um forte efeito de VELOCIDADE no comportamento de F0, ou seja, há um aumento estatisticamente significativo do F0 médio na velocidade rápida de emissão, quando considerado o grupo total de falantes (n=9; 7 + R1/R2). O teste *Levene* para igualdade de variâncias não é significativo (p<.59), indicando que há homogeneidade de variância entre os grupos (normal *vs.*rápida). Os intervalos de confiança dão uma estimativa gráfica da diferença entre as distribuições; quanto menor o intervalo, mais acuradas são as estimativas da média de um determinado grupo (Dixon *et al.*1990:200) [11]. Podemos observar que as médias nas duas condições aparecem bastante afastadas (apesar de a diferença **absoluta** não ser de grande magnitude), indicando um efeito consistente.

A tabela 5.3 apresenta resultados de estatística descritiva básica para cada falante separadamente, em cada condição de velocidade de produção. Podemos observar que, com exceção de R2, todos os falantes (inclusive R1) aumentam o F0 médio na velocidade rápida; o percentual de variação, no entanto, é consideravelmente diferente entre os falantes (os falantes ZR, DO e MS, por exemplo, modificam mais seu F0 médio do que o falante AG). Há uma diferença considerável no F0 médio das duas produções não-contemporâneas do falante R1/R2, onde se verifica um menor valor na primeira gravação (R1), quando o falante encontrava-se sob forte estado gripal; essa diferença confirma a observação de Rosenberg (1976:479), de que inflamações laríngeas podem ter um efeito significativo nas medidas de F0. Uma provável explicação para o fenômeno seria o aumento de massa das cordas vocais, em virtude do estado inflamatório; outra possibilidade, que não exclui a primeira, seria a perda de capacidade pulmonar, o que traria uma redução na pressão trans-glotal.

Falante ↓	Vel.	média	D.P.	SK	KU
ZR	N	112.0	17.2	.10	-.67
	R	134.5	15.7	-.77	.24
EN	N	132.5	15.6	.29	.46
	R	143.0	14.8	.23	-.47
R1	N	134.2	15.7	-.19	-.60
	R	137.7	15.0	-.38	-.41
ZP	N	143.4	21.2	.07	-.23
	R	150.3	19.4	.00	.45
AG	N	89.5	10.1	-.24	-.65
	R	90.4	8.9	-.25	-.46
WA	N	134.9	22.2	.51	-.20
	R	140.8	20.6	.39	-.31
MS	N	146.7	16.0	.15	-.99
	R	171.9	23.4	.18	-.79
R2	N	158.9	23.7	-.23	-.63
	R	157.3	21.4	-.18	-.58
DO	N	127.6	24.5	.22	.70
	R	147.3	22.9	-.11	-.50

TABELA 5.3: Estatística descritiva básica de F0 (Hz) para cada falante separadamente, em cada condição de velocidade de emissão (N=normal; R=rápida)

Ainda com respeito à tabela 5.3, verificamos que todos os falantes, com exceção de MS, apresentam um desvio padrão menor na velocidade rápida. A princípio, esse resultado sugere uma menor variabilidade nos contornos entoacionais, na velocidade rápida, mas a natureza dos dados (apenas vogais lexicalmente tônicas) não permite uma conclusão definitiva a esse respeito (a questão será discutida mais detalhadamente adiante, na seção 5.3.2).

Na tabela 5.3, os valores relacionados à configuração geral da distribuição, *skewness* (SK) e curtose (KU) apresentam um comportamento um tanto errático, de difícil interpretação. Alguns falantes (AG, WA, MS e R1/R2) mantêm padrões coerentes nas duas velocidades de emissão. Os demais falantes alteram SK ou KU ou ambos, no que diz respeito à **direção** do sinal; assim, o falante ZR, que apresenta uma pequena assimetria positiva na velocidade normal (SK = .10), passa a ter uma distribuição consideravelmente alongada para a esquerda na velocidade rápida (SK = -.77), enquanto o falante EN passa de uma distribuição leptocúrtica na velocidade normal (KU=.47) para uma platocúrtica na velocidade rápida (já o oposto ocorre com ZP).

Os resultados para SK são um pouco mais coerentes, já que apenas ZR e DO invertem o sentido da assimetria vertical (é interessante observar que esses falantes foram os que mais substancialmente alteraram a velocidade de fala na condição velocidade rápida). De modo geral, entretanto, *skewness* e curtose não parecem parâmetros muito estáveis, ao menos para o tipo de medida aqui empregada (F0 em núcleos de vogais lexicalmente tônicas); é preciso considerar que o número de medidas é pequeno em cada condição de velocidade (n=75), o que torna os valores de SK e KU pouco representativos. Mais adiante, na seção 5.3.2, examinaremos novamente esse aspecto com base em um maior número de medidas.

A tabela 5.4 examina a influência de FALANTE na variação de F0. Os resultados são de ANOVA (BMDP-7D). A análise foi realizada separadamente para cada condição de velocidade e reunindo as duas condições. A última linha da tabela 5.4 inclui no modelo a interação FALANTE X VELOCIDADE, de modo a que se possa avaliar a força relativa de cada efeito.

ANOVA (BMDP-7D)		
MODELO	**F =**	**p <**
FALANTE (normal)	79.4	.0001
FALANTE (rápida)	87.8	.0001
FALANTE (norm.+ rap.)	140.5	.0001
FALANTE	157.6	.0001
VELOCIDADE	91.5	.0001
FAL. X VELOC.	9.97	.0001

TABELA 5.4: Resultados de ANOVA (BMDP-7D). As duas primeiras linhas mostram o efeito isolado de FALANTE em cada velocidade. A última linha inclui no modelo a interação FALANTE X VELOCIDADE.

Na tabela 5.4, vemos que a variação inter-falante de F0 é altamente significativa, em qualquer das condições de velocidade. A última linha da tabela 5.4 indica que há uma interação significativa FALANTE X VELOCIDADE (F=9.97; p<.0001), ou seja, a variação de F0 em função da velocidade não se dá da mesma forma para todos os falantes (como já havíamos verificado informalmente na tabela 5.3). Observamos também que o efeito de FALANTE (F=157.6) é maior do que o de VELOCIDADE (F=91.5), indicando que, apesar da variação provocada pelas diferentes condições de velocidade, as distinções inter-falante permanecem; com efeito, quando agrupamos as duas condições de velocidade, o efeito isolado de FALANTE é ainda mais significativo (F= 140.5).

Para testar quais pares de falantes são significativamente distintos, realizamos um teste *student-t* de comparação de médias, comparando cada possível par de falantes do grupo (n=9; 7 + R1/R2). A tabela 5.5 mostra os resultados desse teste, dando o valor de *t*, e o nível de significância (p<), para cada par de falantes.

FAL.\	EN	R1	ZP	AG	WA	MS	R2	DO
ZR	-8.6 -2.3**	-9.2 NS	-9.8 -5.0	8.3 16.5	-7.8 -2.3**	-13.9 -10.7	-14.0 -5.7	-5.3 -3.8
EN	-	NS NS	-3.3 -3.4	19.2 24.3	NS NS	-6.2 -9.8	-7.9 -4.3	NS -2.1**
R1	-	-	-2.4** -4.4	19.8 19.8	NS NS	NS -10.4	-6.9 -5.2	NS -3.2
ZP	-	-	-	17.3 22.2	2.0** 2.6*	NS -5.9	-3.8 NS	3.6 NS
AG	-	-	-	-	-15.5 -18.4	-25.0 -26.9	-21.8 -20.8	-11.9 -19.2
WA	-	-	-	-	-	-4.0 -8.3	-5.9 -3.5	NS NS
MS	-	-	-	-	-	-	-2.8 4.5	5.7 6.6
R2	-	-	-	-	-	-	-	7.3 NS

TABELA 5.5: Resultados de testes *student-t* (BMDP-3D) comparando médias de F0 (medidas em núcleos vocálicos) entre falantes (n=9). O número superior (valor de *t*) refere-se à velocidade normal, e o inferior à rápida. Todas as comparações são significativas (médias consideradas diferentes) em um nível de p<.001, com exceção de NS (não significativo), * (p<.01) e ** (p<.05).

Os valores absolutos de *t* na tabela 5.5 são uma estimativa da "distintividade" de dois falantes quanto a F0. Levando em conta a velocidade normal, alguns falantes, nas comparações par-a-par, são significativamente diferentes de todos os demais (AG, ZR, R2); na velocidade rápida, apenas o falante AG, nas comparações par-a-par, é estatisticamente distinto do resto do grupo. As duas produções não contemporâneas do falante R1/R2 são significativamente diferentes entre si, nas duas condições de velocidade.

De um modo geral, há um maior número de distinções do que de confusões, mas as variações significativas em função da velocidade, assim como a alteração significativa entre as produções com e sem gripe do falante R1/R2, sugerem que esse tipo de teste deve ser usado, na aplicação forense, com bastante cautela, servindo

apenas como um indicador genérico de identidade. Deve-se ressaltar, contudo, que a condição "velocidade rápida de emissão", tal como proposta no presente estudo, estabelece condições de produção um tanto extremas, que não encontraremos, certamente, na maior parte das situações cotidianas, onde a variação intra-falante na taxa de articulação deve ficar bem abaixo das diferenças aqui registradas. A experiência forense tem demonstrado que, quando é possível ter acesso a padrões espontâneos do suspeito, a distribuição de F0 pode fornecer informação relevante - embora nunca conclusiva - para a determinação da identidade da amostra questionada [12]. Seria necessário estudar mais detalhadamente a relação velocidade de emissão/F0, mas é razoável supor que seja possível estabelecer um fator individual de ajuste de F0 em função da velocidade de emissão para cada falante.

5.3.1.1) *F0 Intrínseco*

Vários estudos já verificaram que há uma relação sistemática entre a qualidade vocálica e o F0 local, fenômeno geralmente denominado *F0 intrínseco*. A constatação dessa relação fez inclusive com que alguns autores propusessem modelos perceptuais incluindo o F0 local como um fator tão importante quanto os formantes para a caracterização de vogais (v. Syrdal e Gopal 1986; Traunmüller 1988).

Em geral associa-se o F0 intrínseco a certas restrições articulatórias; assim, nas vogais altas, a posição levantada da língua afetaria a tensão laríngea, tendo como conseqüecia um aumento local de F0 (Delgado Martins 1986) [13]. Ternström *et al.*(1988) verificam que o fenômeno, fora das expectativas, ocorre também no canto; os autores estudam um grupo de cantores profissionais que devem cantar uma nota, em uma altura confortável, mudando de uma vogal para outra, sem empregar *glissando*. Ternström *et al.* verificam que há um acréscimo de F0 da ordem de 10 -

20 % nas passagens /ɛ, e, a, y/ → /i/ e /e, o, ɸ/ → /y/ (ou seja, dentro das expectativas, vogais mais altas têm F0 intrínseco mais alto). A existência do efeito no canto é um tanto surpreendente, já que há uma preocupação especial em manter a altura melódica; os autores argumentam que o fato de haver alterações mesmo quando, conscientemente, se tenta exercer controle, indica que o fenômeno é um efeito involuntário, provavelmente de natureza articulatória.

Do ponto de vista da Identificação de Falantes, o fenômeno do F0 intrínseco tem relevância, na medida em que insere um fator de variação intra-falante, cuja magnitude será importante conhecer. Além disso, já se observou que as alterações do F0 local em função da qualidade vocálica pode variar consideravelmente de falante para falante (Ladd e Silverman 1984).

A figura 5.1 mostra as médias de F0 para cada uma das categorias vocálicas, nas duas condições de velocidade de produção separadamente e reunidas, agrupando todos os falantes do grupo principal (n=9; 7 + R1/R2). Podemos observar que, de acordo com as expectativas, há uma tendência a valores mais altos de F0 nas vogais mais altas. As diferenças relativas são aproximadamente as mesmas nas duas condições de velocidade, havendo apenas uma leve tendência a uma menor diferenciação entre as vogais mais baixas na velocidade rápida. A diferença entre as vogais nos extremos da escala chega a cerca de 18 % (/ã/ *versus* /u/). A vogal nasalizada /ã/ apresenta o valor mais baixo de F0; é provável que a abertura do *velum*, criando um maior volume supra-glotal, tenha o efeito de reduzir a pressão trans-glotal e, conseqüentemente, o F0 local. É interessante observar, nesse sentido, que a diferença entre /ã/ e sua equivalente oral /a/ diminui consideravelmente na velocidade rápida; se considerarmos que a abertura total do *velum* exige cerca de 130 ms (Laver 1980:77), é razoável supor que, em taxas muito rápidas de emissão, a nasalização ocorra de forma incompleta, explicando a menor diferenciação entre /a/ e /ã/.

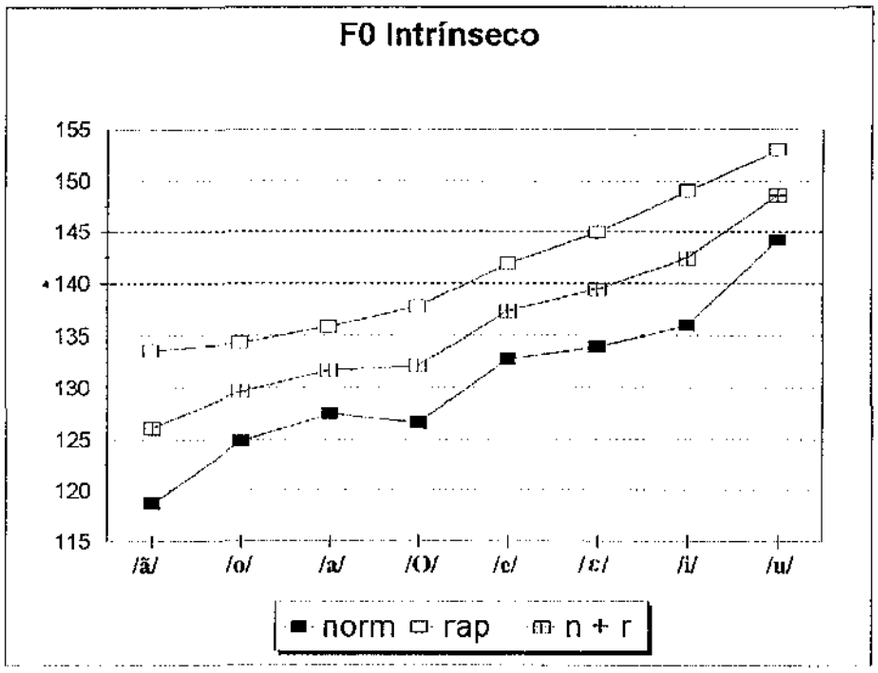


FIGURA 5.1: Médias de F0 em função de cada categoria vocálica (F0 intrínseco), nas duas condições de velocidade de produção, separadamente e reunidas. As médias foram calculadas considerando o grupo total de falantes (n=9; 7 + R1/R2)

De modo a avaliar a força relativa da qualidade vocálica em F0, isoladamente e em interação com as demais variáveis descontínuas, realizamos várias análises de variância (BMDP-7D para os efeitos isolados e BMDP-2V para os modelos incluindo interações), cujos resultados encontram-se na tabela 5.6.

MODELO	VEL.	F =	p <
VOGAL	normal	5.12	.0001
VOGAL	rápida	4.41	.0001
VOGAL	norm.+rap.	8.98	.0001
FALANTE	normal	71.31	.0001
VOGAL		9.37	.0001
FALANTE X VOGAL		.40	NS
FALANTE	rápida	80.31	.0001
VOGAL		9.32	.0001
FALANTE X VOGAL		.42	NS
FALANTE	norm.+rap.	130.4	.0001
VOGAL		16.58	.0001
FALANTE X VOGAL		.59	NS
VOGAL	-----	9.25	.0001
VELOCIDADE		45.8	.0001
VOGAL X VELOCIDADE		.24	NS

TABELA 5.6: Resultados de diversas análises de variância estudando a influência de VOGAL isoladamente (BMDP-7D) e em modelos que incluem interações com as demais variáveis descontínuas (FALANTE e VELOCIDADE)

Analisando os resultados da tabela 5.6, podemos constatar que os efeitos de VOGAL no F0 local são da mesma magnitude nas duas condições de velocidade como se confirma pelos efeitos isolados e pela ausência de interação VOGAL X VELOCIDADE. No modelo que considera o efeito FALANTE percebemos que também não há interação significativa FALANTE X VOGAL, indicando que a influência de VOGAL em F0 é homogênea entre os falantes do grupo analisado; esse resultado discorda das observações de Ladd e Silverman (1984), que relatam algumas diferenças inter-falante, sem, contudo, apresentar resultados estatísticos mais detalhados.

Embora os resultados aqui obtidos sejam, de forma geral, consistentes com os relatados em outros estudos, indicando uma tendência a um F0 intrínseco maior nas vogais altas, a posição relativa de algumas vogais parece não corresponder totalmente às expectativas. A vogal /o/, por exemplo, deveria apresentar um valor maior do que efetivamente foi observado, e a vogal /e/ deveria ter uma média mais elevada do que a mais baixa /ɛ/. Essas distorções podem estar relacionadas ao fato de os dados não estarem balanceados, como já comentamos anteriormente, ou seja, não houve controle quanto aos contextos segmentais e prosódicos onde as vogais observadas ocorrem, fatores que, certamente, interagem com o F0 intrínseco (Cf. Steele 1986).

Um dos fatores fonéticos que podem influir no F0 local é o contexto consonantal precedente. Esse aspecto já foi verificado em House e Fairbanks (1953): o padrão de F0 no *onset* vocálico após consoantes não-sonoras em seqüências CV exibe um contorno descendente, enquanto o oposto ocorre para a consoante sonora equivalente. O efeito, a princípio, seria pista perceptualmente inportante para a sonoridade da consoante, mas esse ponto permanece sob discussão, já que alguns estudos não verificam um padrão consistente (Cf. Haggard *et al.*1970; Umeda 1981). O fenômeno, do ponto de vista articulatório, é normalmente associado à posição mais alta da laringe nos *stops* não sonoros (Hombert *et al.* 1979; Riordan 1980), de acordo com a regra geral que associa um levantamento da laringe a um aumento de F0 (Ohala 1972); assim, articulatoriamente, as alterações de F0 no *onset* vocálico pós-consonantal teriam a mesma causa das alterações em função da qualidade vocálica. É possível que o efeito acima descrito afete de algum modo o F0 intrínseco medido no núcleo vocálico, já que o movimento ascendente ou descendente de F0 a partir do *onset* vocálico pode se estender por cerca de 100 ms (Haggard *et al.* 1970); como no presente estudo não controlamos o contexto fonético, os pequenos desvios

comentados no parágrafo anterior podem estar relacionados a efeitos de interação entre a consoante precedente e a vogal.

É importante ressaltar, entretanto, que o efeito da qualidade vocálica no F0 local manteve-se nos nossos dados, mesmo sem qualquer tipo de controle quanto aos ambientes segmental e prosódico. A influência de VOGAL é sempre estatisticamente significativa, mesmo nos modelos que incluem as variâncias (bem mais expressivas) de FALANTE e VELOCIDADE (v. tabela 5.6). Por outro lado. a atribuição de uma relação estável entre a qualidade vocálica e o F0 intrínseco seria prematura, já que esse tipo de alteração micro-prosódica talvez seja neutralizada, ao menos parcialmente, em outros contextos prosódicos. Ladd e Silverman (1984) observam que as diferenças entre o F0 de vogais altas *versus* baixas é maior em sentenças isoladas do que na leitura fluente. Algo semelhante ocorre nas variações de F0 no *onset* vocálico em função do traço [± sonoro] da consoante precedente (Umeda 1981; Silverman 1986). O que se depreende dessas observações é que contextos entoacionais mais "ricos" tendem a inserir uma variação de tal ordem que neutraliza esses efeitos mais locais.

A questão do F0 intrínseco no contexto da Identificação de Falantes mereceria um estudo mais amplo e detalhado, que fugiria ao escopo do presente trabalho, já que a magnitude dos efeitos está longe de ser desprezível, como vimos na tabela 5.6 e na figura 5.1. A importância das variações micro-prosódicas tem sido ressaltada em diversos sitemas de síntese de fala, onde já se observou que a inclusão desse tipo de informação é fundamental para uma maior naturalidade da saída acústica (Thorsen 1980; Silverman 1986; Di Cristo e Hirst 1986).

5.3.2) *Medidas de F0 obtidas através de LPC*

Na seção anterior examinamos a variação de F0 inter- e intra-falante a partir de medidas obtidas nos núcleos vocálicos de sílabas tônicas. Como já comentamos anteriormente, na situação forense, em virtude da presença de condições adversas de captação/transmissão, é bastante freqüente que medidas confiáveis - de F0 ou de formantes - só sejam possíveis nessa posição. Uma questão importante é verificar em que medida a distribuição de F0 assim obtida apresentaria algum tipo de distorção; uma possibilidade, por exemplo, é que o F0 médio obtido apenas em medidas nos núcleos vocálicos de vogais lexicalmente tônicas seja um pouco diferente do que a média obtida em pontos igualmente espaçados, já que a marca de acento pode estar associada à variação de F0.

Nesta seção estudaremos a distribuição de F0 a partir de medidas obtidas automaticamente pelo programa LPC (Linear Predictive Coding) da KAY Elemetrics (mod. 5635). O método utilizado foi o de Covariância, com *frames* assíncronos, sendo cada medida tomada em *frames* de 20 milisegundos. O programa LPC deriva diversos parâmetros da fala (F0, formantes, BW) e os expõe numericamente diretamente na tela de um micro-computador tipo PC; um arquivo texto pode então ser aberto, de modo a registrar os resultados numéricos desejados.

Através da extração automática via LPC, é possível obter, rapidamente, um grande número de medidas isoladas de F0. Em sinais com boa relação sinal/ruído, o algoritmo utilizado tem um desempenho bastante eficiente. Alguns erros, entretanto, nunca deixam de ocorrer. Em geral esses erros são facilmente detectáveis, já que aparecem, quase sempre, nas fronteiras entre regiões vozeadas e regiões não vozeadas; isso ocorre porque, em alguns casos, o *frame* ajustado para um passo fixo de 20 milisegundos abrange exatamente a região limítrofe, dificultando a aplicação

do algoritmo de extração de F0. De modo a eliminar esse erro, não foram considerados, nos cálculos a seguir, os *frames* situados em regiões de transição vozeado/não-vozeado (ou vice-versa). Alguns erros podem também ocorrer mesmo fora das regiões limítrofes, embora mais raramente; esses erros esporádicos são, na maior parte dos casos, salientes, já que fazem o valor medido divergir fortemente dos valores vizinhos. De modo a eliminar esse tipo de erro desenvolvemos um algoritmo corretor que, após detectar o valor anômalo, reescreve-o com base em uma interpolação baseada nos valores corretos dos *frames* imediatamente vizinhos.

Submetemos os dados obtidos via LPC a uma análise de variância (BMDP-2V), de modo a verificar os efeitos de cada variável (FALANTE e VELOCIDADE), assim como sua interação. Os resultados dessa análise estão na tabela 5.7.

efeito	F=	p<
FALANTE	1071.2	.0001
VELOCIDADE	516.9	.0001
FAL X VEL	56.6	.0001

TABELA 5.7: Resultados de análise de variância (BMDP-2V)

Na tabela 5.7, observamos que todos os efeitos são altamente significativos, embora a variação relacionada ao falante seja bem maior do que os demais efeitos; esse resultado é semelhante ao observado quando realizamos análise de variância apenas com as medidas extraídas de núcleos vocálicos tônicos (v. seção 5.3.1). A existência de efeito de interação indica que a alteração de F0 em função da velocidade de emissão não ocorre de forma homogênea, o que pode ser verificado na tabela 5.8.

A tabela 5.8 apresenta a média e o desvio-padrão de F0 para cada falante isoladamente, extraídos pelos dois métodos: VOG (medidas apenas em núcleos

vocálicos lexicalmente tônicos) e LPC (medidas automáticas em *frames* de 20 milisegundos, para o enunciado inteiro - leitura do texto I; v. anexo), para as duas condições de velocidade de produção, assim como a diferença entre as médias e desvios-padrão entre os dois métodos.

Método ↓	Fal. →	ZR	EN	R1	ZP	AG	WA	MS	R2	DO
VOG	N	112.0	132.5	134.2	143.4	89.5	134.9	146.7	158.9	127.6
		17.2	15.6	15.7	21.2	10.1	22.2	16.0	23.7	24.5
	R	134.5	143.0	137.7	150.3	90.4	140.8	171.9	157.3	147.3
		15.7	14.8	15.0	19.4	8.9	20.6	23.4	21.4	22.9
LPC	N	110.0	132.2	131.8	140.4	94.2	134.0	148.9	153.2	131.3
		14.9	14.0	15.7	16.9	9.7	19.5	15.6	18.4	17.0
	R	132.7	140.1	134.5	146.1	96.0	138.7	156.3	155.3	144.2
		15.9	14.8	16.3	17.2	7.5	17.4	17.9	16.2	18.5
DIF. →	N	+ 2.0	-.03	+ 2.4	+ 3.0	- 4.7	+ 0.9	- 2.2	+ 5.7	- 3.7
		+ 2.3	+ 1.4	+ 0.0	+ 4.1	+ 0.4	+ 2.7	+ 0.4	+ 5.3	+ 7.5
	R	+1.8	+ 2.9	+ 3.2	+ 4.2	- 5.6	+ 2.1	+ 15.6	+ 2.0	+ 3.1
		- 0.2	+ 0.4	- 1.3	+ 2.2	+ 1.4	+ 3.2	+ 5.5	+ 5.2	+ 4.4

TABELA 5.8: Médias e desvios-padrão aferidos por dois métodos diferentes: VOG (apenas medidas em núcleos vocálicos) e LPC (medidas automáticas em *frames* de 20 ms); a última linha dá a diferença entre os resultados (VOG - LPC)

De acordo com a tabela 5.8, a diferença entre as médias das medidas de F0 extraídas apenas nos núcleos vocálicos (método VOG) e as medidas automáticas via LPC é, na maior parte dos casos, inferior a 4 %; apenas o falante MS apresenta um aumento de cerca de 10 % na média de F0 pelo método VOG, velocidade rápida. De modo geral, portanto, os resultados indicam que a estimativa da média de F0 baseada em medidas nos núcleos vocálicos tônicos representa satisfatoriamente a média global de F0 de um falante.

A tabela 5.8 não revela um desvio consistente nas médias obtidas pelos dois métodos na velocidade normal; quatro falantes (ZR,R1/R2,ZP e WA) apresentam médias superiores no método VOG, enquanto os outros quatro (EN,AG,MS e DO)

têm médias menores nesse método. Na velocidade rápida, por outro lado, apenas o falante AG mostra um valor menor na média de F0 obtida pelo método VOG. Essa tendência mais consistente a uma média mais elevada de F0 no método VOG na velocidade rápida pode estar associada ao fato de, nessa condição de produção, serem reduzidas as possibilidades de contrastes de tonicidade baseados na duração, em função do efeito do *princípio de incompressibilidade* (Klatt 1973); assim, o falante na velocidade rápida, marcaria a tonicidade com a variação de F0, em geral com um aumento local nos núcleos vocálicos tônicos ; esse processo levaria, obviamente, a um maior valor médio de F0 pelo método VOG, que considera apenas os núcleos vocálicos tônicos (a tendência divergente observada no falante AG pode estar relacionada ao valor atipicamente baixo de seu F0 médio; o algoritmo de extração de F0 utilizado pelo programa LPC tem alguma dificuldade em tratar valores muito baixos de F0, e pode ter havido alguma imprecisão nos valores obtidos [14]; outro aspecto a ser considerado é o fato de o falante AG ter apresentado a menor alteração de velocidade entre as duas condições de produção; assim, qualquer efeito relacionado à variação pronunciada de velocidade de produção não deveria se manifestar nesse falante.

A redução dos contrastes duracionais pode estar relacionada à maior complexidade articulatória da fala rápida; nesse sentido é interessante observar que o controle de aspectos finos de *timing*, incluindo os contrastes prosódicos baseados na duração, são adquiridos bem mais tarde que o controle de F0 e intensidade (Kent 1976; Pollock et al. 1993).

Uma tendência bastante consistente observada na tabela 5.8 é o maior valor do desvio padrão nas medidas obtidas pelo método VOG, independentemente da velocidade de produção. Essa diferença é um tanto surpreendente, especialmente no caso da velocidade rápida, contrariando, aparentemente, nossa hipótese, acima

formulada, que sugeria ser a marca de tonicidade, na fala hiper-acelerada, menos dependente de contrastes duracionais do que de contrastes de F0; assim, em uma primeira análise, a expectativa seria de um **maior** desvio padrão na condição velocidade rápida de produção. Há, entretanto, uma outra forma de interpretar esses resultados. O que parece ocorrer na fala rápida é uma normalização dos contrastes de F0 como marca de tonicidade, de tal forma que os picos locais de F0 mantêm-se num mesmo patamar, em virtude de uma redução do efeito do fenômeno conhecido como *declinação de F0* [15] no âmbito da sentença, fazendo com que a linha entoacional de base permaneça quase constante. Na fala rápida, portanto, embora haja um uso mais consistente de F0 para marcar acento, a variação de F0 se dá num âmbito menor do que na fala normal. Esse aspecto pode ser verificado na figura 5.2, contendo os histogramas de F0 para as duas velocidades de produção; observamos a distribuição menos achatada da velocidade rápida, concentrando-se em torno de valores centrais.

A redução do efeito da declinação de F0 na fala rápida deve-se, em parte, ao fato de a programação respiratória, nessa condição de produção, abranger trechos mais extensos, quase sempre compostos de várias sentenças, enquanto na leitura normal os *breath-groups* (v. Lieberman 1980) tendem a coincidir com estruturas lógicas menores.

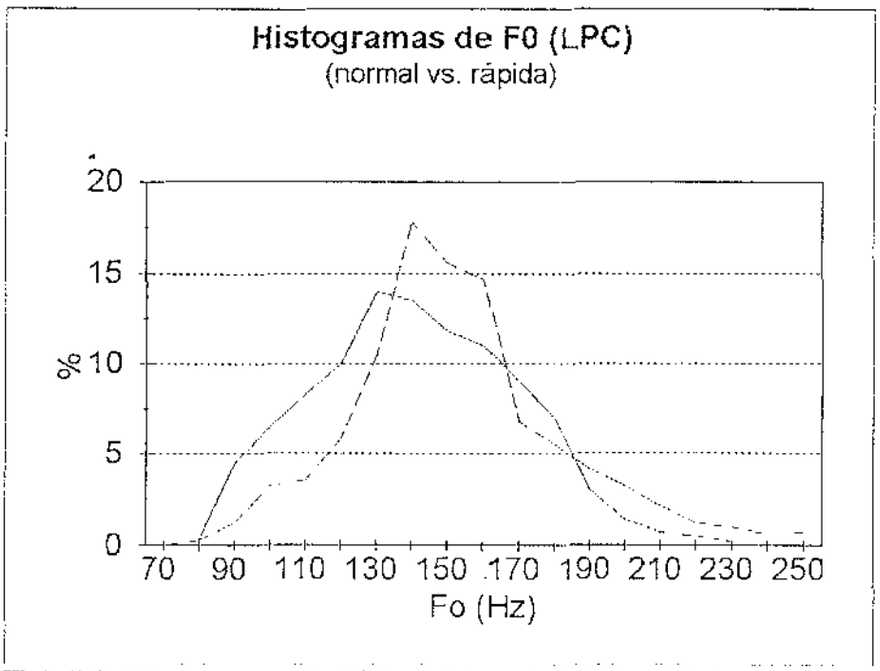


FIGURA 5.2: Histogramas de F0 (medidas via LPC). A linha contínua refere-se à velocidade normal e a tracejada à velocidade rápida. Observa-se uma configuração mais leptocúrtica na distribuição dos dados na velocidade rápida

De modo a verificar a significância estatística das diferenças entre as médias individuais de F0, foi realizada uma série de testes-*t* comparando cada par de falantes (programa BMDP-3D) para cada velocidade separadamente. Os resultados desses testes estão na tabela 5.9.

FAL.\	EN	R1	ZP	AG	WA	MS	R2	DO	n=
ZR	21.1 9.11	21.0 4.9	30.3 12.2	11.1 22.1	22.2 4.5	40.7 33.3	41.0 18.9	21.2 11.0	931 736
EN		NS 3.7 *	11.4 3.8	32.2 34.1	2.8 * 4.3	22.3 30.1	23.2 13.1	2.7 ** 4.0	895 897
R1			11.8 7.0	32.3 27.9	3.1 ** NS	23.3 31.2	24.4 15.3	3.1 ** 6.7	1304 986
ZP				40.2 36.8	7.9 7.7	10.5 27.2	11.9 9.9	7.6 NS	1114 949
AG					32.3 27.9	50.2 52.0	50.0 39.9	30.8 31.9	799 640
WA						18.4 31.8	19.5 15.9	NS 7.3	1053 920
MS							NS 16.1	17.7 22.7	1310 1072
R2								18.8 7.8	1332 1004
DO									1118 741

TABELA 5.9: Resultados de testes-*t* (BMDP-3D) comparando médias de F0 (dados extraídos por programa LPC) entre falantes (n=9). O número superior refere-se à velocidade normal, e o inferior à rápida. Todas as comparações são significativas (médias consideradas diferentes) a um nível de p<.0001, com exceção de NS (não significativo), * (p<.001) e ** (p<.01). A última coluna dá o número de *frames* usados para cada falante.

Os testes-*t* resumidos na tabela 5.9 indicam que quase todos os pares de falantes têm médias significativamente diferentes, sendo o número de distinções entre pares de falantes maior do que o observado anteriormente na tabela 5.5, com medidas extraídas apenas em núcleos vocálicos lexicamente tônicos (v. seção 5.3.1). As duas produções não contemporâneas do falante R1/R2, no entanto, permanecem significativamente diferentes, apesar do maior número de dados utilizado.

A comparação de medidas via LPC com as obtidas apenas nos núcleos vocálicos mostrou que, em termos de F0 médio e desvio padrão, as diferenças são

pequenas, uma indicação de que o método descrito em 5.3.1 reflete razoavelmente bem essas características. Há algum ganho de informação nas distribuições baseadas nas medidas LPC, já que um maior número de pares de falantes é estatisticamente distinto nessa condição. O emprego de medidas automáticas na prática forense, como já comentamos acima, é muitas vezes prejudicado pela presença do ruído de fundo. Há muitos casos, no entanto, onde é possível sua utilização; a limitação de banda imposta pelo canal telefônico, por exemplo, não insere erros significativos na extração de parâmetros LPC, desde que a transmissão seja de boa qualidade (Furui 1981). A própria influência do ruído de fundo depende bastante de suas características. As dificuldades impostas ao algoritmo serão maiores se o ruído interferente for de baixa freqüência; já se verificou que a adição de ruído branco e/ou ruído de quantização ao sinal não alteram significativamente medidas de F0 baseadas em LPC (Sambur e Jayant 1976) e ruídos de alta freqüência podem ser neutralizados através de uma filtragem prévia *low-pass* (Johnson *et al.* 1990).

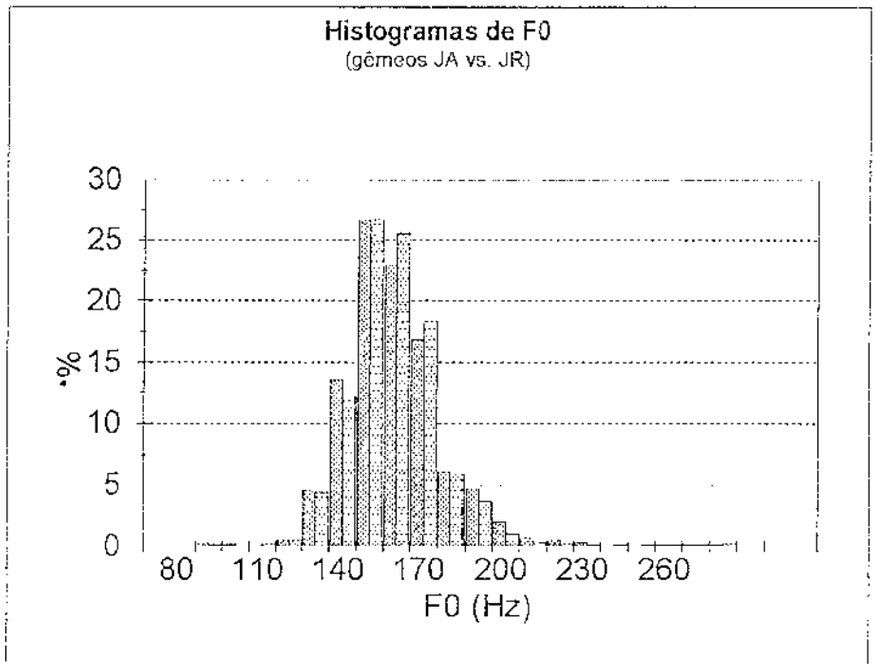
5.3.3) *Análise de F0 em Gêmeos Idênticos*

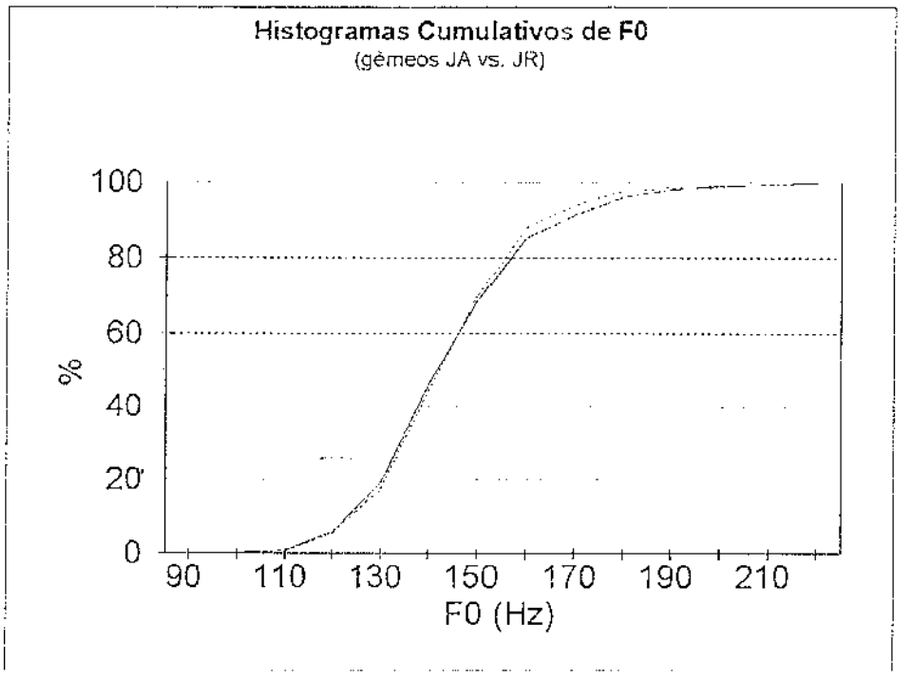
Comparamos as distribuições de F0 das medidas LPC para os dois gêmeos monozigóticos JR e JA. O teste-*t* de comparação de médias indicou que as médias não são significativamente diferentes (ver tabela 5.10).

FALANTE \	JR	JA
média	144.7	144.3
desvio padrão	18.1	17.8
n=	2558	2821
pooled t = .83 p< .40		

TABELA 5.10: Resultados do teste-*t* (BMDP-3D) comparando as médias de F0 (medidas LPC) dos gêmeos idênticos JR e JA (n refere-se ao número de *frames* utilizados no cálculo).

A coincidência entre os valores das médias e desvios-padrão dos gêmeos JR e JA não é um artefato estatístico baseado em medidas genéricas; as figuras 5.3 e 5.4 revelam que as distribuições são praticamente idênticas, o que se destaca especialmente no histograma cumulativo. Atkinson (1976) sugere que o F0 médio está mais fortemente vinculado a características anatômicas, enquanto a variabilidade de F0 (desvio-padrão) representaria diferenças estilísticas e/ou emocionais. As coincidências nos dois parâmetros, para os gêmeos JA e JR indicam que, provavelmente, o que está em jogo aqui não são apenas similaridades fisiológicas mas também "hábitos prosódicos" semelhantes que se refletem em padrões rítmico-entoacionais semelhantes. Nesse sentido é interessante observar que um outro parâmetro, a razão tempo vozeado/tempo não vozeado (RTV; ver seção 7), também relacionado com a dimensão prosódica, apresentou valores praticamente iguais para os gêmeos JA e JR.





FIGURAS 5.3 e 5.4: Histogramas de barras e cumulativo das distribuições de F0 dos gêmeos JA e JR

5.3.4) *Influência do Tamanho da Amostra (número de frames) nas medidas de F0 via LPC*

A extração de médias de F0 para um falante passa necessariamente pela questão do tamanho da amostra de fala utilizada. É importante verificar em que medida pode haver uma variação condicionada à extensão do trecho de fala. De modo a examinar esse aspecto, foram calculadas as médias e os desvios-padrão para diferentes números de *frames* contíguos da análise LPC. Para cada falante foram selecionados, arbitrariamente, diferentes trechos com 15, 30, 60, 120, 250 e 500 *frames* vozeados, a partir dos quais foram calculados média e desvio padrão. As figura 5.5a,b apresentam as médias calculadas a partir das diversas quantidades de *frames*, para cada falante, em cada condição de velocidade de produção.

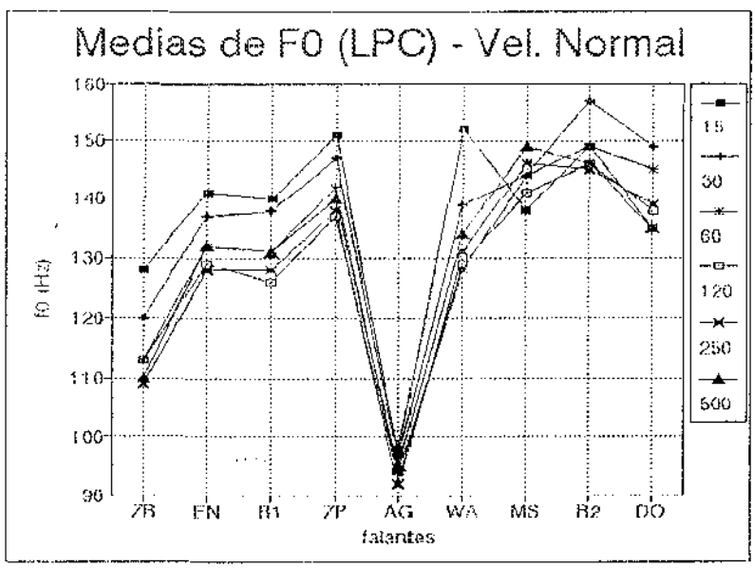


FIGURA 5.5a: Médias de F0 a partir de diferentes números de *frames*

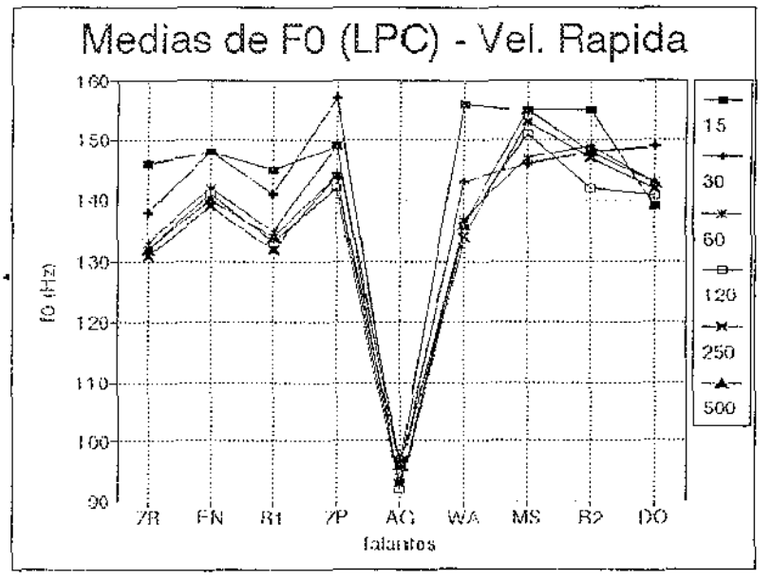


FIGURA 5.5b: Médias de F0 a partir de diferentes números de *frames*.

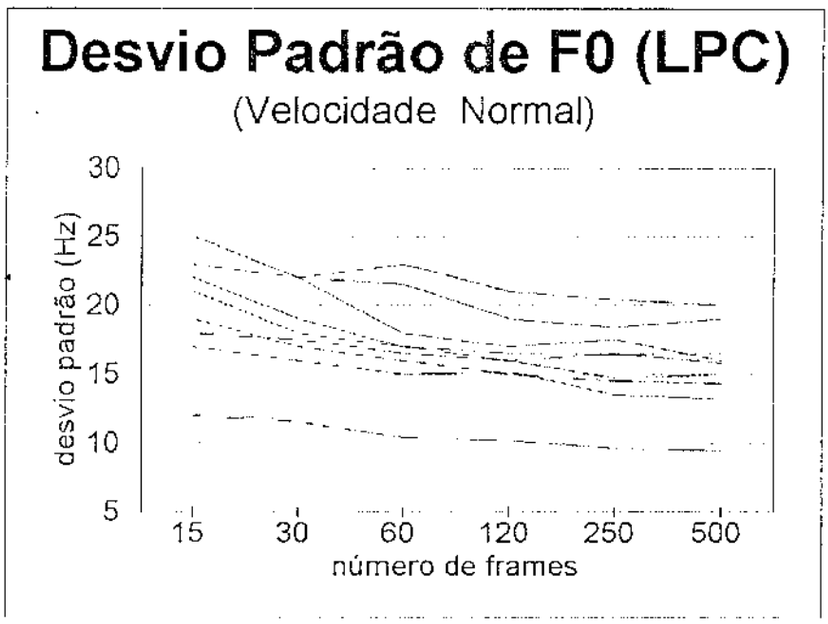
Observamos, na figura 5.5, que existe uma variação considerável nas médias de F0 em função do número de *frames*, com uma tendência à estabilização a partir de 120 *frames*. Os valores obtidos para amostras de 500 *frames* são praticamente iguais às médias obtidas com base no enunciado total de cada falante, indicando que esse tamanho de amostra é suficiente para expressar valores médios individuais.

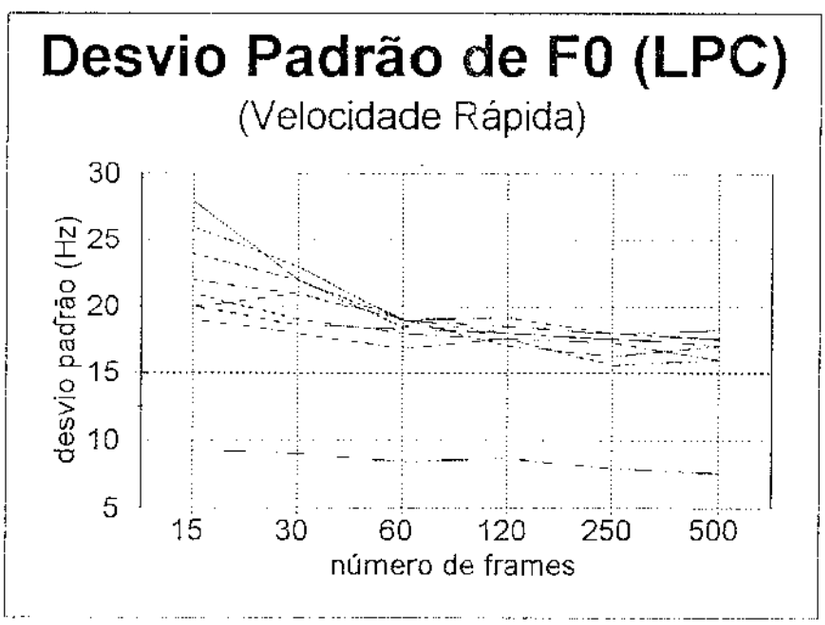
A figura 5.5 enfatiza o fato de que a utilização do F0 médio como indicador da identidade do falante deve ser acompanhada de considerações a respeito do valor absoluto desse parâmetro; observamos que apenas o falante AG destaca-se fortemente do grupo em qualquer tamanho de amostra e nas duas velocidades de produção.

É interessante observar que as diferenças individuais tendem a diminuir na velocidade rápida (com exceção do falante AG). Esse fenômeno já foi observado quando medimos o F0 apenas nas vogais tônicas, e está provavelmente relacionado a um princípio genérico que reduz a variação de um parâmetro qualquer à medida em

que o falante se aproxima de uma condição limite de produção. No domínio das durações, por exemplo, essa tendência se manifesta naquilo que Klatt (1973) chama de *princípio de incompressibilidade*, um limite natural que impediria a redução de um segmento além de um certo ponto. Assim, existe uma expectativa de que a variação inter-falante de F0 (do mesmo modo que a duração segmental e, provavelmente, a amplitude) tenda a diminuir na fala acelerada. A ausência de alteração na média do falante AG em função da velocidade não pode ser considerada, já que esse falante, como já comentamos anteriormente, praticamente não modificou sua taxa de emissão, apesar das instruções específicas do experimento.

As figuras 5.6a,b apresentam a variação do desvio padrão em função do número de *frames* utilizado na análise. Observamos que o desvio padrão tende a cair à medida em que aumenta o tamanho da amostra. A partir de 500 *frames* as medidas se estabilizam e não se alteram significativamente em relação à amostra englobando o enunciado integral. Horii (1975) também observa uma relação inversamente proporcional entre o desvio padrão e o tamanho da amostra, embora relate uma estabilização apenas a partir de 70 segundos de fala. Não podemos estabelecer uma comparação direta com os números de Horii (1975), pois utilizamos apenas *frames* vozeados; como aproximação, podemos avaliar que 500 *frames* vozeados correspondem a cerca de 25-30 segundos de fala contínua - de qualquer modo uma amostra bem menor do que a relatada por Horii (1975).





FIGURAS 5.6a,b: Desvio-padrão de F0 em função do número de *frames*

Os valores do desvio padrão, assim como já observamos para as próprias médias de F0, têm uma menor variação inter-falante na velocidade rápida. Esse resultado é consistente com a hipótese de que, na fala rápida, haja uma normalização nos contrastes de F0.

A diferença pouco significativa inter-falante entre os valores do desvio padrão sugere que esse parâmetro, isoladamente, não é um indicador eficiente da identidade do falante. Apenas o falante AG apresenta um valor nitidamente divergente dos demais falantes; é possível que haja mesmo uma relação sistemática entre a média de F0 e o desvio padrão (média de F0 baixa, desvio padrão também baixo).

5.3.5) *Relação F0/Amplitude*

Ao examinarmos mais de perto as medidas aferidas via LPC, verificamos a existência, para a maioria dos falantes, de uma relação mais ou menos sistemática entre F0 e a amplitude, em dB, medidos em cada *frame* de análise. Observou-se, na maior parte dos casos, uma correlação positiva entre esses dois parâmetros. Essa correlação justifica-se em parte pelo fato de o controle de F0 estar condicionado ao aumento da pressão sub-glotal (Flanagan 1958); assim, existe uma expectativa de que as curvas de amplitude e F0 tenham um certo paralelismo. Examinando, entretanto, o comportamento de cada falante isoladamente, observamos que a correlação F0/AMPLITUDE varia consideravelmente entre os falantes analisados no presente estudo. A tabela 5.11 mostra os coeficientes de correlação (*Pearson - r*) entre F0 e a amplitude (em dB) para cada falante, em cada condição de velocidade de produção. De modo a verificar a consistência desse tipo de medida, foram

extraídos coeficientes de correlação para cada metade do enunciado total de cada falante (leitura do texto I - ver anexo).

Podemos verificar, na tabela 5.11, que alguns falantes tendem a apresentar um índice mais alto da correlação F0/AMPLITUDE (especialmente o falante ZR), enquanto outros apresentam uma correlação baixa (WA, por exemplo) ou sequer uma correlação estatisticamente significativa (DO e MS). Essa diferença de comportamento pode estar relacionada a diferentes estratégias individuais de controle de F0, ou seja, alguns falantes, provavelmente, controlam a entonação apenas - ou prioritariamente - através da variação da tensão das cordas vocais, enquanto outros associam as variações de F0 às variações da pressão sub-glotal. É interessante observar que os valores para as produções não contemporâneas de R1/R2 são da mesma ordem de grandeza.

Outro aspecto relevante observado na tabela 5.11 é o fato de existir uma tendência bastante consistente a uma maior correlação F0/AMPLITUDE nas medidas extraídas das amostras de velocidade rápida: todos os falantes, sem exceção, revelam essa tendência, incluindo aqueles falantes que, na velocidade normal, não apresentam correlação significativa entre as duas medidas. É possível que, na velocidade rápida, em função de uma postura muscular geral mais tensa, o controle de F0 dependa mais fortemente das variações da pressão sub-glotal do que das variações da tensão local das cordas vocais. Dito de outro modo: na velocidade rápida, para manter o controle das variações de F0 - prioritárias para a clareza da mensagem prosódica - o falante recorre, alternativamente (já que as cordas vocais encontram-se em estado de tensão quase máxima), a variações da pressão sub-glotal, estratégia esta que terá como conseqüência uma aumento da correlação F0/AMPLITUDE, tal como observamos na tabela 5.11.

FALANTE	VEL.	Amostra	*Pearson - r*	p<
ZR	N	1	.247	.001
		2	.255	.001
	R	1	.325	.001
		2	.356	.001
EN	N	1	.231	.001
		2	.199	.01
	R	1	.255	.001
		2	.265	.001
R1	N	1	.307	.001
		2	.315	.001
	R	1	.344	.001
		2	.361	.001
ZP	N	1	.402	.001
		2	.398	.001
	R	1	.423	.001
		2	.457	.001
AG	N	1	.235	.01
		2	.221	.01
	R	1	.243	.01
		2	.277	.001
WA	N	1	.151	.01
		2	.134	.05
	R	1	.178	.01
		2	.169	.01
MS	N	1	.009	ns
		2	-.015	ns
	R	1	.126.	.05
		2	.119	ns
R2	N	1	.289	.001
		2	.301	.001
	R	1	.342	.001
		2	.328	.001
DO	N	1	-.021	ns
		2	-.046	ns
	R	1	.141	.05
		2	.164	.01

TABELA 5.11: Índices de correlação (*Pearson - r*) entre F0 e Amplitude (dB) a partir de medidas extraídas por LPC. As amostras **1** e **2** referem-se à primeira e segunda metades do texto I (v. anexo)

O emprego da correlação F0/AMPLITUDE como indicador da identidade do falante coloca-se como uma possibilidade, embora não tenhamos notícia de esse parâmetro ter sido anteriormente utilizado. A medida é razoavelmente consistente, mantendo a mesma ordem de grandeza ao longo de todo o enunciado, como se pode verificar comparando as correlações na primeira e na segunda metades da leitura na tabela 5.11. A variação desse índice aqui verificada, em função das diferentes condições de velocidade de emissão, não deve, na prática, influir muito no desempenho desse parâmetro, já que as condições do experimento aqui realizado exigiram uma performance dos falantes certamente afastada das situações normais de produção. É preciso ressaltar, contudo, que as medidas de amplitude obtidas referem-se ao **pico** de energia em cada *frame* da análise LPC; um outro algoritmo, que extraísse a energia **média** em cada *frame*, certamente indicaria relações F0/AMPLITUDE mais consistentes do que as aqui observadas; ao extrair o pico de energia, a curva obtida tende a ter variações locais mais abruptas, interferindo de algum modo no índice de correlação F0/Amplitude.

5.4) *Comentário Final*

Ao longo da presente seção tivemos a oportunidade de discutir diversos aspectos relacionados às variações inter- e intra-falante de parâmetros derivados de F0. Uma série de fatores exercem uma influência considerável no comportamento de F0, especialmente no que diz respeito ao valor médio.

Resumidamente, listamos abaixo os principais fatores que podem contribuir para variações intra-falante em F0:

- expressividade afetivo/emocional
- presença de *stress* psicológico
- velocidade de emissão
- intoxicação alcoólica
- inflamação laríngea

Essas fontes de variação devem ser consideradas em todas as aplicações relacionadas com a Identificação de Falantes, especialmente no paradigma forense, onde a possibilidade de controlar certas condições é bastante limitada. Por outro lado, é importante destacar que a variabilidade inter-falante do F0 médio tem um âmbito considerável; excetuando-se os casos onde há tentativa de disfarce (raros e, além disso, quase sempre perceptualmente salientes), a distribuição de F0 é informação fundamental, situando o falante em uma **faixa** característica, especialmente se a amostra é suficientemente extensa. Na verdade, na prática forense, uma decisão só é tomada a partir de um conjunto de fatores; nenhum aspecto isolado (indicando ou não uma coincidência) deve ser considerado conclusivo.

SEÇÃO 6 : ESPECTRO DE LONGO TERMO

6.1) *Introdução*

De uma forma geral, os parâmetros acústicos da fala podem ser classificados em duas categorias básicas: (a) traços de curto termo e (b) traços de longo termo. Os traços de curto termo exigem o isolamento de trechos do sinal de fala com uma duração limitada, que correspondem a unidades abstratas tais como fonemas, sílabas, núcleos entoacionais, etc. A utilização de aspectos de curto termo na Identificação de Falantes tem como maior dificuldade a própria definição do evento a ser observado, em função do grande número de fatores contextuais (fonéticos, entoacionais, discursivos, etc) que podem interagir com esse evento; os parâmetros discriminadores de falante definidos no curto termo são fortemente dependentes do material de fala específico, e sua efetividade depende em grande parte do controle de uma série de condições, incluindo contexto fonético, velocidade de emissão, padrão entoacional, etc.

Por outro lado, os aspectos acústicos definidos no longo termo têm a vantagem de ser essencialmente independentes do conteúdo da mensagem falada; idealmente, poder-se-ia dizer que esses são traços invariantes no tempo, refletindo traços estáveis do falante. Existem características da voz de uma pessoa que perpassam a saída acústica como um todo e não podem ser associadas diretamente a realizações de um elemento isolado. O rótulo genérico "Qualidade de Voz" é geralmente aplicado na descrição dessas características quase-permanentes (ver Laver 1980, para uma descrição - mais em termos articulatórios que acústicos - de diferentes "tipos" básicos de voz).

Um modo bastante eficiente de acessar características invariantes de uma voz é o espectro de longo termo (ELT); esse tipo de análise envolve o cálculo seqüencial de uma série de espectros de freqüência/amplitude ao longo da duração de um enunciado; esses espectros de curto termo independentes são tomados como base para a formação de um valor médio final, de tal forma que o espectro freqüência/amplitude resultante represente um espectro composto único. Ao contrário de um espectro individual isolado, o ELT não reflete diretamente características de um evento temporal particular; na verdade, espera-se que, para enunciados acima de uma determinada duração, o ELT independa totalmente do conteúdo segmental (Li *et al.*1969; Gelfer *et al.* 1989).

O emprego de espectros médios tem uma origem longínqua. Já em 1917, Crandall, por meio de cálculos manuais, construía espectros sintéticos médios, representando a composição de energia de diversas vogais. Sivian (1929; *apud* Pittam 1987:2), no entanto, foi o primeiro a realizar uma análise de longo termo a partir de um trecho de fala contínua. As contingências da Segunda Grande Guerra levantaram a possibilidade de empregar análises acústicas de longo termo para verificar o efeito das grandes altitudes na fala e a distorção na voz provocada pelo uso de máscaras de gás (Stevens *et al.* 1947). Nas últimas quatro décadas análises acústicas baseadas no ELT têm sido utilizadas em um grande número de áreas de pesquisa, desde a acústica ambiental até o aprimoramento de aparelhos para deficientes auditivos (Cf. Pittam 1987). Como instrumento para a análise e monitoramento de patologias laríngeas, o ELT tem tido um largo emprego, embora a eficiência desse tipo de informação seja alvo de intensa controvérsia (v. p.ex. Hurme e Sonninen 1986; Sundberg 1986; Wendler *et al.* 1986).

O ELT tem sido empregado com bastante freqüência, e já há algum tempo, na pesquisa focalizando a Identificação de Falantes (ver, entre outros: Hargraves e Starkweather 1963; Wolf 1972; Majewski e Hollien 1974; Zalewski *et al.* 1975; Doherty 1975; Hollien e Majewski 1977; Hollien *et al.* 1978; Doherty e Hollien 1978; Gelfer *et al.* 1989). Em geral, os estudos relatam um bom desempenho do ELT como indicador da identidade do falante, especialmente em condições de laboratório. Existem, entretanto, uma série de condições que podem alterar em alguma medida a eficiência do ELT. A próxima sessão abordará mais de perto alguns desses aspectos.

6.2) *Eficiência do ELT na Identificação de Falantes*

A primeira questão que intuitivamente nos ocorre, ao pensarmos em ELT, refere-se à definição de uma amostra representativa. Sendo o objetivo principal capturar características de longo termo, é fundamental conhecer a extensão mínima da amostra de fala necessária para produzir ELTs estáveis. Gelfer *et al.* (1989) abordam esse problema, testando a eficiência de ELTs produzidos a partir de amostras de 5, 10 e 20 segundos, cobrindo uma faixa de 100 - 10000 Hz (40 dimensões : filtros fixos de 1/6 oitava); através de um critério estatístico de decisão, os autores verificam que para amostras de 10 e 20 segundos o índice de acertos praticamente independe do trecho utilizado para extrair os ELTs, ou seja, diferentes trechos do mesmo falante podem ser usados como amostras teste e referência sem que haja queda na performance. Para amostras de 5 segundos, entretanto, o sistema torna-se fortemente dependente do texto e índices de acerto elevados só são obtidos quando se comparam trechos idênticos. Li *et al.*(1969) sugerem um intervalo de tempo consideravelmente maior para estabilizar o ELT, pelo menos 30 segundos de fala contínua. No experimento descrito mais abaixo (seção 6.3) utilizamos cerca de

25 segundos para cada ELT; testes preliminares, no entanto, indicaram que um trecho de 10-15 segundos, dependendo do falante, é suficiente para produzir ELTs estáveis.

Outro aspecto que pode influir no desempenho do ELT é a faixa útil de freqüência. A questão é extremamente importante para avaliar a possibilidade de emprego desse tipo de informação acústica através de linhas telefônicas. Hollien e Majewski (1977) comparam o desempenho de ELTs com faixas de 80 - 10000 Hz e 315 - 3150 Hz, a segunda simulando a faixa do canal telefônico. Com base em uma métrica de Distância Euclidiana, Hollien e Majewski verificam uma redução de 14 % no índice de acertos para o ELT de banda limitada (embora o índice final ainda seja consideravelmente alto, cerca de 76 %) para um grupo de 100 falantes (v. também Doherty e Hollien 1978).

Em condições ideais de laboratório, o ELT é um dos mais eficientes indicadores da identidade do falante; nessas condições, vários estudos comparando o ELT com outros parâmetros acústicos (F0, *tracking* de Formantes, curvas de amplitude, *speech rate*, etc) apontam o ELT como a informação mais eficiente (Zalewski *et al.* 1975; Doherty 1975; Hollien *et al.* 1978). Por outro lado, o ELT pode ser sensível a tentativas de disfarce (Doherty 1975; Hollien e Majewski 1977; Hollien 1991), presença de *stress* psicológico induzido (Doherty 1975; Doherty e Hollien 1978), intoxicação alcoólica (Klingholz *et al.* 1988) e condições afetivo/emocionais (Pittam 1987). A maior restrição para o emprego do ELT na Identificação de Falantes não vem, entretanto, dessas fontes de variação (fatores que, de uma forma ou de outra, também modificam outros parâmetros acústicos da fala), mas sim da extrema sensibilidade do ELT às condições do meio de transmissão/captação do sinal. Fatores como o nível de absorção acústica ambiental, padrão do ruído de fundo, tipo de gravador, microfone e fita magnética, etc, influem consideravelmente na configuração do ELT (Cf. Doddington 1985; Juang 1991).

Sendo um somatório de espectros no tempo, o ELT captura o efeito combinado desses diversos fatores, integrando-os indissoluvelmente na configuração espectral final, de tal forma que é impossível separar os efeitos do falante e os efeitos do meio; assim, o emprego do ELT fica, de certa forma, restrito às aplicações onde as condições do meio não são variáveis, como é o caso de alguns sistemas automáticos de Verificação de Falantes. No paradigma forense o uso do ELT é mais crítico, já que, apenas em casos excepcionais, é possível reproduzir exatamente as condições originais de captação [1].

6.3) *ELT: Um Experimento*

Nesta seção abordaremos alguns aspectos relacionados à performance do ELT como indicador de identidade do falante ainda não examinados em outros trabalhos (pelo menos na literatura ao nosso alcance). Um deles é a variabilidade do ELT em função de diferentes condições de velocidade de produção. Como se sabe, o aumento de velocidade de fala - mais especificamente da taxa de articulação [2] - pode ter como conseqüência uma redução da qualidade vocálica, fenômeno conhecido como *target undershoot* (Cf. Lindblom 1963; Lindblom e Studdert-Kennedy 1967); a questão que se coloca aqui é verificar se o *target undershoot* , ou outras alterações devidas à maior velocidade, se refletirão na configuração final do ELT.

Outro aspecto a ser aqui examinado diz respeito à eficiência do ELT em diferenciar gêmeos idênticos. Gêmeos idênticos possuem, geralmente, vozes quase indistinguíveis em uma avaliação apenas auditiva; os gêmeos monozigóticos empregados no presente estudo possuem efetivamente essa característica (de acordo com o que nos foi relatado, mesmo familiares próximos teriam dificuldades em identificá-los corretamente apenas pela voz).

6.3.1) *Material e Métodos*

Para o presente experimento utilizamos, além dos falantes do grupo principal (n=9; 7 + R1/R2), também os gêmeos univitelinos JA e JR. As amostras de fala dos falantes do grupo principal foram obtidas através da leitura do texto I (v. anexo) em duas condições de velocidade (normal *vs.* rápida). As amostras dos gêmeos JA e JR foram extraídas da leitura do texto II (v. anexo) (para maiores detalhes sobre a metodologia básica ver seção 3).

Duas amostras foram criadas para cada falante: uma amostra referência e uma amostra teste. Para os falantes do grupo principal (7 + R1/R2) chamamos de amostra teste aquela obtida com base na leitura do texto I na condição *velocidade normal*, e de amostra referência a obtida na condição *velocidade rápida*. No caso dos gêmeos, que não realizaram a leitura do texto II na velocidade rápida, foram utilizadas como referência e teste amostras obtidas a partir da primeira e da segunda metades do texto II, respectivamente.

6.3.2) *Procedimento de Análise (ELT)*

Para obter os ELTs calculou-se um espectro médio (*average*) na faixa 0-8000 Hz, com filtro de análise fixo de 300 Hz, para um intervalo de tempo de cerca de 25 segundos de fala contínua. As análises acústicas foram realizadas através do DSP-500 da KAY Elemetrics.

Embora alguns experimentos utilizem intervalos de tempo um pouco maiores do que o aqui empregado (v. p.ex. Hollien e Majewski 1977; Doherty e Hollien 1978), aceita-se, em geral, que um intervalo de 10-15 segundos de fala produzirá um

ELT representativo, neutralizando quase totalmente o efeito do conteúdo segmental (Gelfer *et al.* 1989) (v. discussão na seção 6.2).

O ELT foi calculado considerando todos os *frames*, incluindo os trechos não vozeados. Embora a inclusão de sons fricativos não sonoros possa enfatizar os componentes de mais alta freqüência (acima de 3-4 KHz), a forma geral do ELT nas regiões mais informativas não parece se alterar significativamente (Cf. Nolan 1983:144ff; Wendler *et al.* 1986).

Para cada falante foram obtidos dois ELTs, um a partir de um trecho lido na velocidade normal e outro a partir de um trecho, de mesma duração (cerca de 25 segundos), lido na velocidade rápida. Observe-se que, em função das diferentes velocidades de emissão, os ELTs para cada falante não se referem exatamente ao mesmo trecho, já que, na velocidade rápida de emissão, o mesmo intervalo de tempo corresponde a uma maior quantidade de conteúdo segmental.

Para os procedimentos estatísticos descritos a seguir foram também incluídos os gêmeos JA e JR, formando assim um total de 11 falantes, considerando as duas amostras não contemporâneas de R1/R2 (11= 7 + R1/R2 + JA + JR). Os ELTs teste e referência dos gêmeos JA e JR também baseiam-se em diferentes conteúdos segmentais (primeira e segunda metades do texto II).

A figura 6.1 mostra ELTs de dois falantes do grupo, extraídos segundo os critérios acima expostos.

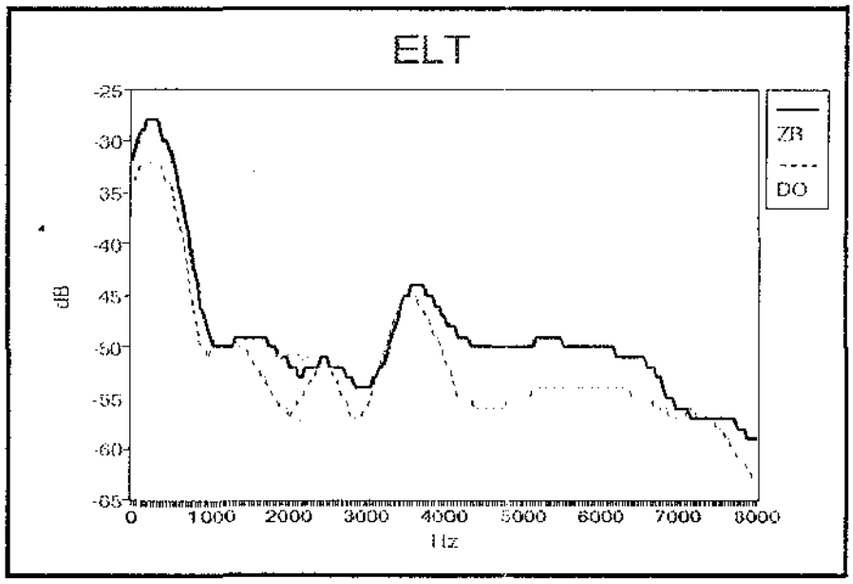


FIGURA 6.1: ELTs de dois falantes (ZR e DO), incluindo trechos não-vozeados (faixa 0-8000 Hz)

6.3.3) *Análises Estatísticas*

6.3.3.1) *Análise Cluster: 200 pontos do ELT*

Cada ELT foi expresso quantitativamente por meio da amplitude em dB em cada um dos 200 pontos no eixo de freqüência (de 40 a 8000 Hz em passos de 40 Hz). Cada um dos pontos no eixo de freqüência foi tratado como uma observação, enquanto cada falante (ou melhor, cada uma das duas amostras de cada falante) representa uma variável isolada. Essa tabulação serviu de entrada para o programa BMDP-1M, que realiza uma análise *cluster* de variáveis.

O objetivo da análise *cluster* é detectar inter-relações entre um conjunto de variáveis em uma matriz de dados. O programa BMDP-1M inicialmente considera

cada variável (no caso em questão, cada produção de cada falante) como um *cluster* independente; as duas variáveis mais semelhantes são então agrupadas para formar um novo *cluster*. O processo continua, passo a passo (reunindo variáveis ou *clusters* de variáveis) até que um único *cluster* seja formado, contendo todas as variáveis. As medidas de similaridade são estabelecidas a partir de uma matriz de correlações entre as variáveis. Para o processo de amalgamação dos *clusters*, BMDP-1M oferece três critérios diferentes: similaridade máxima (SINGLE), similaridade mínima (COMPLETE) e similaridade média (AVERAGE).

Todo o processo pode ser graficamente sintetizado na forma de uma árvore (dendrograma) cujos nódulos representam a junção de uma variável a outra variável, de uma variável a um *cluster* já formado, ou de um *cluster* a outro *cluster*. No caso de o programa reunir pares do mesmo falante **antes** de reunir um dos itens do par a qualquer outro falante ou *cluster*, podemos considerar que houve uma identificação **correta**. De um modo geral, o dendrograma oferece uma avaliação das similaridades entre os falantes, em função do ELT.

A figura 6.2 mostra dendrogramas obtidos através de BMDP-1M, por três métodos diferentes (SINGLE, AVERAGE e COMPLETE), a partir de 200 pontos do ELT na faixa 0 - 8000 Hz (ZR_N representa "falante ZR, velocidade normal", ZR_R representa "falante ZR, velocidade rápida", etc; para os gêmeos JA e JR, os índices 1 e 2 representam as duas diferentes amostras de cada um). Podemos observar que dos onze pares corretos possíveis, nove foram encontrados pelo programa, independentemente do método empregado; a probabilidade de se chegar a esse resultado é bem pequena, da ordem de 10^{-6}.

Apenas as variáveis correspondendo às produções dos falantes ZR e R2 não foram agrupadas antes de serem reunidas a outro *cluster*. R2-N e R2-R, no entanto, foram reunidos ao *cluster* já formado por [R1-N + R1-R], formando assim um novo *cluster* que agrupa corretamente **todas** as produções desse falante (métodos SINGLE

e COMPLETE). Com relação ao falante ZR, observamos que, no método COMPLETE, ZR-R foi reunido incorretamente ao *cluster* formado por [DO-N + DO-R]; o próximo passo do programa, entretanto, foi reunir ZR-N ao *cluster* [ZR-R + DO-N + DO-R], reaproximando assim ZR-N e ZR-R.

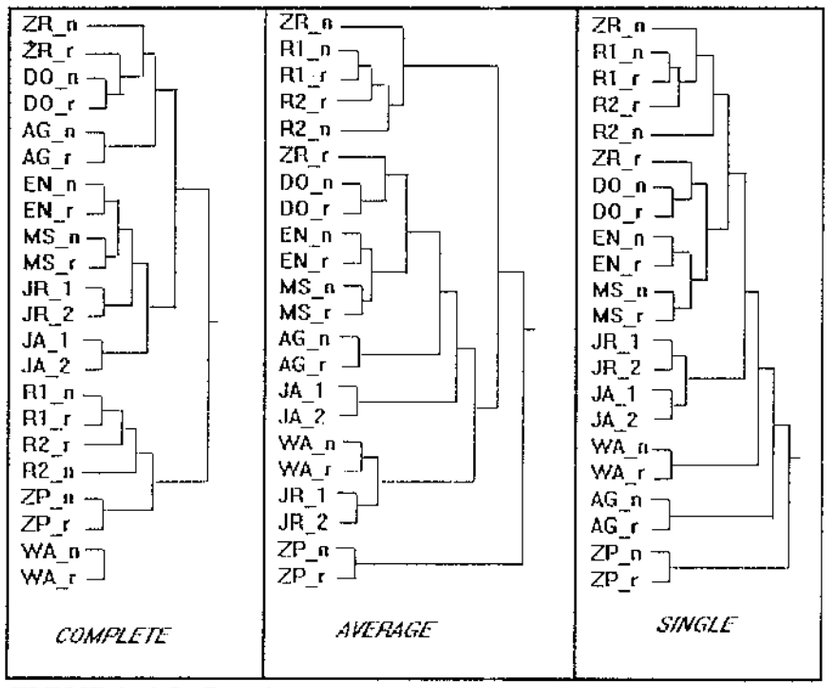


FIGURA 6.2: Dendrogramas resultantes de análise *cluster;* a estrutura do dendrograma reflete o grau de similaridade entre as diferentes amostras, a partir de medidas extraídas do ELT.

Ainda com relação à figura 6.2, podemos observar que os gêmeos JR e JA são corretamente separados em *clusters* individuais [JR_1 + JR_2] e [JA_1 + JA_2], independentemente do método utilizado. Esse acerto é bastante interessante, já que, impressionisticamente, as vozes dos gêmeos JA e JR são quase indistinguíveis. A informação contida no ELT parece, pois, acessar características não diretamente salientes à percepção.

Podemos observar que as estruturas dos dendrogramas gerados apresentam algumas diferenças em função do método de amalgamação empregado. O método SINGLE captura melhor a similaridade entre os gêmeos JA e JR, reunindo-os em um *cluster* único antes de juntá-los a outro ramo do dendrograma; por outro lado, o par ZR-N/ZR-R fica mais afastado com esse método. No método AVERAGE, tanto o falante ZR quanto os gêmeos JA e JR são reunidos em *clusters* iniciais diferentes. Está fora do escopo do presente trabalho interpretar as diferentes estruturas do *cluster* final em função do método empregado. Pareceu importante, entretanto, registrar o fato, na medida em que se coloca aqui o problema mais geral da sensibilidade das decisões em função do procedimento estatístico utilizado. Outros estudos baseados no ELT já observaram que o número de identificações corretas pode variar dependendo do método estatístico utilizado para definir a métrica de distância (v. Doherty 1975; Doherty e Hollien 1978; Zalewski *et al.* 1975).

6.3.3.2) *Análise Cluster: Faixas selecionadas do ELT*

Na seção anterior utilizou-se informação do ELT considerando a faixa 0 - 8000 Hz. Sabemos, entretanto, que, perceptualmente, é possível identificar falantes com razoável precisão a partir de sinais com banda mais limitada. Através do canal telefônico, por exemplo, a faixa de freqüência está, em geral, restrita a 350 - 3500 Hz. A informação contida no ELT talvez não se distribua homogeneamente ao longo de todo o espectro. Para testar essa possibilidade, realizamos mais uma série de análises *cluster* usando como entrada faixas selecionadas do ELT.

Foram definidos arbitrariamente pontos no eixo de freqüência de cada ELT em 0, 500, 1000, 2000, 3500, 5000 e 8000 Hz. Essas marcas serviram como limites inferior e superior para a definição da faixas de freqüência a serem usadas como entrada para o programa BMDP-1M. A tabela 6.1 resume os resultados dos testes

estatísticos apresentando o número de acertos [3] (*clusters* iniciais agrupando pares corretos de falantes) para os três métodos de amalgamação fornecidos em BMDP-1M.

Faixa (Hz)↓	SINGLE	AVERAGE	COMPLETE
0 - 500	1	3	2
0 - 1000	2	4	4
0 - 2000	2	3	4
0 - 3500	7	7	7
0 - 5000	9	9	10
0 - 8000	9	9	9
500 - 1000	1	1	2
500 - 2000	2	2	2
500 - 3500	7	7	8
500 - 5000	9	9	10
500 - 8000	9	9	9
1000 - 2000	0	0	2
1000 - 3500	6	7	7
1000 - 5000	9	10	10
1000 - 8000	9	10	10
2000 - 3500	6	6	7
2000 - 5000	7	8	8
2000 - 8000	10	10	10
3500 - 5000	5	5	5
3500 - 8000	6	7	8
5000 - 8000	5	5	5

TABELA 6.1: Número de identificações corretas em três diferentes métricas, a partir de faixas selecionadas do ELT

Observamos na tabela 6.1 que o método de amalgamação influi consideravelmente no desempenho do programa. Com exceção da faixa 0 - 500 Hz, o método COMPLETE atinge sempre o maior número de acertos. O método SINGLE obtém o pior desempenho em todas as faixas.

Fica evidente, ao examinarmos a tabela 6.1, que algumas faixas do ELT contêm mais informação discriminadora de falante do que outras. A faixa 0 - 3500 Hz, por exemplo, consegue identificar corretamente 7 falantes, enquanto a faixa 3500 - 8000 Hz identifica apenas 5. Dividindo o ELT em 3 faixas, verificamos que a faixa 2000 - 5000 obtém o dobro de acertos, se comparada com as faixas 0 - 2000 e 5000 - 8000.

A figura 6.3 mostra, graficamente, o número de acertos em função de faixas selecionadas do ELT. O gráfico pode ser interpretado como se simulasse o efeito de filtros passa baixa e passa alta, com cada um dos pontos no eixo horizontal representando a freqüência de corte. No ponto assinalado como 1000 Hz, por exemplo, teríamos o número de acertos para dados do ELT utilizando as faixas seletivas 0 -1000 Hz (passa baixa) e 1000 - 8000 Hz (passa alta).

Verificamos, através da figura 6.3, que a faixa mais informativa é a de 2000 - 5000 Hz: ao incluirmos dados extraídos dessa faixa o número de acertos cresce rapidamente. A faixa 1000 - 2000, por outro lado, parece contribuir pouco para a separação correta dos falantes: ao incluirmos dados extraídos dessa faixa o número de acertos não se altera.

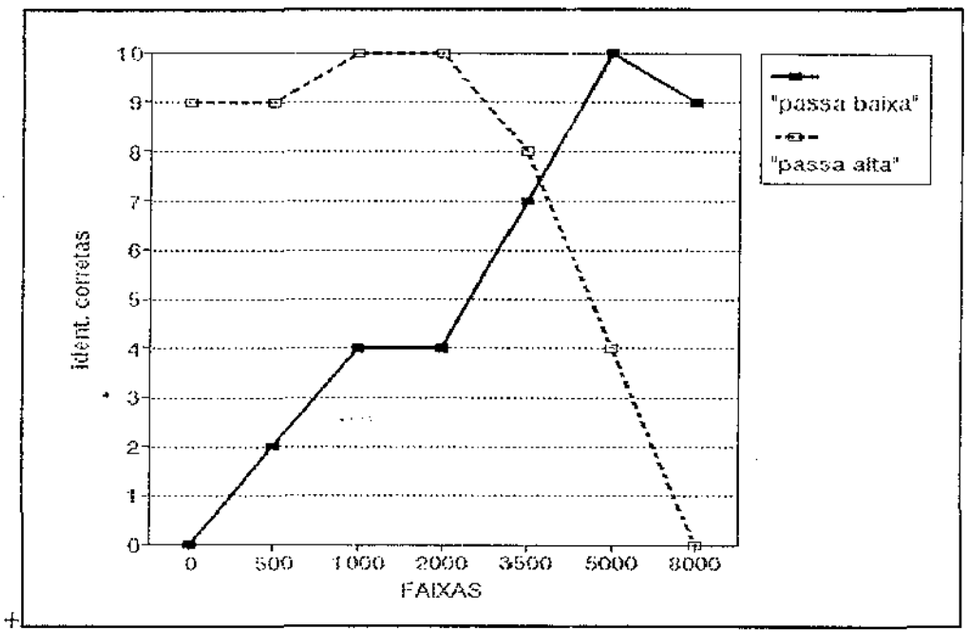


FIGURA 6.3: Identificações corretas a partir de faixas selecionadas do ELT.

A figura 6.4 mostra os acertos a partir de faixas intermediárias do ELT. A linha contínua refere-se às faixas definidas por marcos contíguos (0-500, 500-1000, 1000-2000 Hz, etc), a linha tracejada fina às faixas definidas por 3 marcos contíguos (0-1000, 500-2000 Hz, etc) e a linha tracejada espessa às faixas definidas por 4 marcos contíguos (0-2000, 500-3500 Hz, etc). O gráfico simula o efeito de um filtro passa-banda.

As figuras 6.3 e 6.4 indicam que faixas mais amplas tendem a obter maior número de separações corretas entre falantes, mas isso não ocorre em todos os casos; a faixa 2000 - 8000 Hz, por exemplo, é mais informativa do que a faixa completa 0 - 8000 Hz. Do mesmo modo, a faixa 0 - 5000 Hz obtém mais acertos do que o ELT total. Mesmo a faixa bem mais estreita 1000 - 5000 separa melhor os falantes - pela análise *cluster* - do que o ELT integral 0 - 8000 Hz.

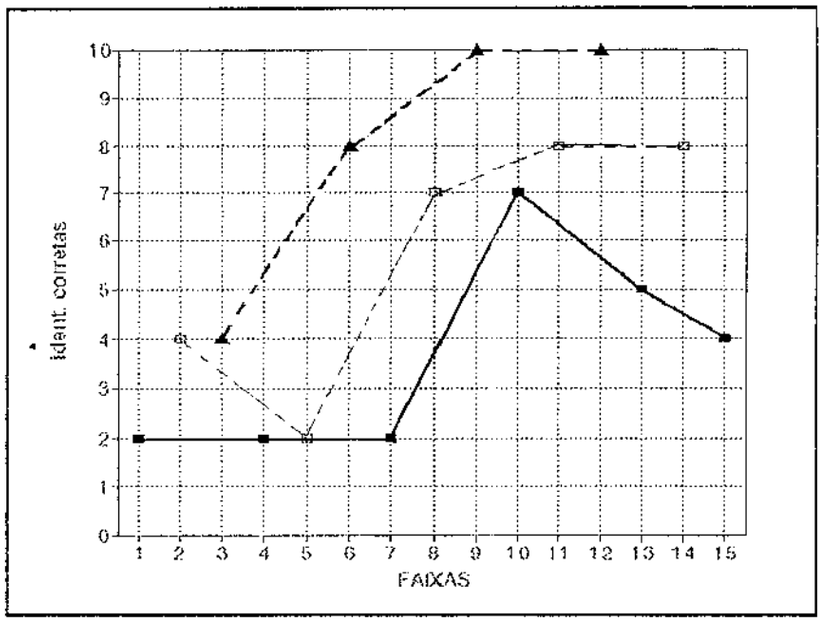


FIGURA 6.4: Identificações corretas a partir de faixas intermediárias selecionadas do ELT (**1**= 0-500; **2**=0-1000; 3=0-2000; **4**=500-1000; **5**=500-2000; **6**=500-3500; **7**=1000-2000; **8**=1000-3500; **9**= 1000-5000; **10**=2000-3500; **11**=2000-5000; **12**= 2000-8000; **13**=3500-5000; **14**=3500-8000; **15**=5000-8000).

A faixa intermediária de 500 - 2000 Hz é pouco informativa. A região do ELT correspondente aos dois primeiros formantes - ou seja, aproximadamente a faixa 500 - 2000 Hz - tem sua configuração determinada quase exclusivamente por F1 e F2, cuja variabilidade é fortemente condicionada pela qualidade vocálica. A informação contida nessa região do ELT, portanto, seria predominantemente lingüística, tornando-se indiferenciada quanto às características do falante. Por outro lado, a faixa 2000 - 3500 Hz, que se mostrou, dentre as faixas estreitas, a mais eficiente para distinguir falantes no nosso teste, contém grande parte da informação referente a F3 e F4, parâmetros mais dependentes do falante do que da qualidade vocálica.

A baixa informação da faixa 5000 - 8000 Hz pode ser atribuída ao fato de haver nessa região do ELT forte influência do ruído fricativo de alta freqüência, o que torna o envelope quase plano e, portanto, indiferenciado para efeito de

correlação (que é a métrica básica utilizada pelo programa BMDP-1M). Essa faixa do ELT, entretanto, pode se tornar mais informativa ao considerarmos o espectro total, já que a amplitude **relativa** dessa faixa reflete algumas características da fonte. Vozes com qualidade *breathy*, por exemplo, podem apresentar um ganho de energia nessa região, em função da presença de ruído turbulento produzido na região glotal (Hammarberg *et al.* 1986; Klatt e Klatt 1990; Childers e Lee 1991).

6.3.3.3) *Análise cluster: Slopes + resíduos + amplitude média na faixa*

Vários métodos já foram aplicados para quantificar diferenças entre ELTs. Algumas dessas tentativas estão relacionadas com a possibilidade de acessar modificações pós-tratamento em diversos tipos de patologia da voz (Cf. Hurme e Sonninen 1986; Löfqvist 1986; Sundberg 1986). Uma das medidas utilizadas freqüentemente é a razão entre os níveis de amplitude de diferentes faixas do espectro. Outra medida é o *slope* , ou seja, a inclinação da reta ajustada a uma determinada faixa do ELT, expressa em dB/oitava.

Essas medidas representam, é claro, uma drástica redução de informação, com perda de qualquer detalhe local do ELT. Até que ponto, porém, serão preservadas distinções entre falantes, com base apenas nesse tipo de medida espectral genérica? Para testar essa possibilidade, criamos algumas novas medidas espectrais a partir dos mesmos ELTs utilizados anteriormente. Em primeiro lugar, foram extraídos os *slopes* das retas ajustadas às faixas já definidas (método dos mínimos quadrados); esses *slopes* são expressos como a tangente do ângulo que o prolongamento da reta ajustada faz com o eixo das freqüências em escala logarítmica (LOG_2 [f_{Hz}]). Dessa forma, obtemos diretamente o decaimento em dB/oitava. Cada ELT fica assim reduzido a apenas 6 *slopes*, como ilustra a figura 6.5.

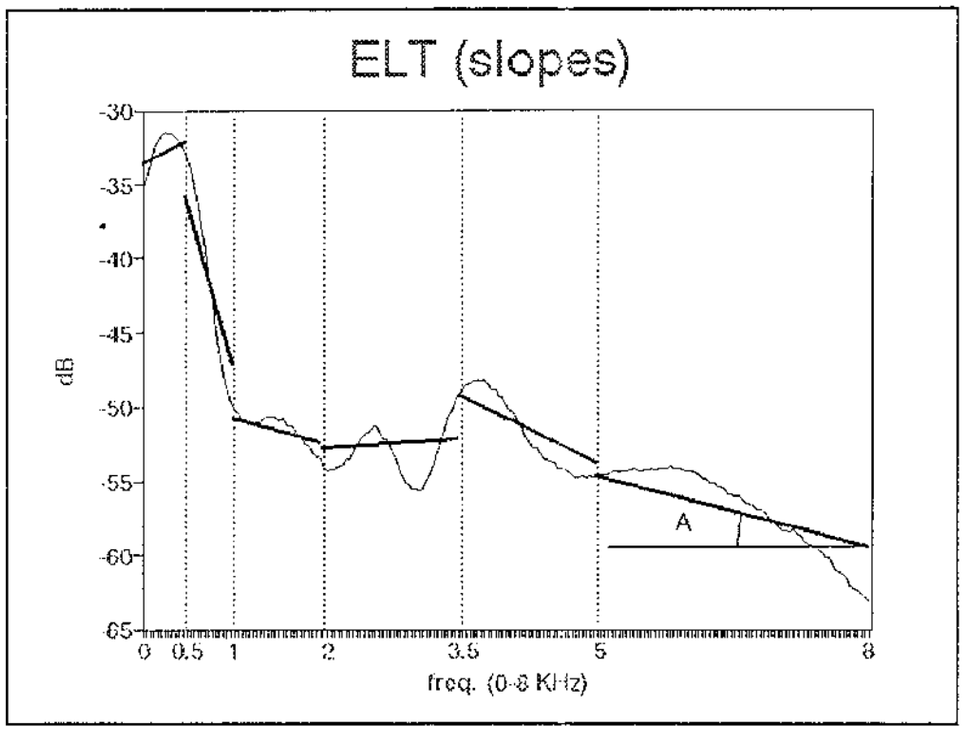


FIGURA 6.5: *Slopes* extraídos de faixas selecionadas do ELT; cada *slope* é expresso pela tangente do ângulo que a reta ajustada faz com o eixo das freqüências expresso em escala LOG ; o ângulo A, por exemplo, na faixa 5-8 KHz (as abscissas não estão representadas em LOG, mas o cálculo das tangentes foi feito com base nessa escala, de modo a obter o *slope* em dB/oitava).

Essas 6 novas variáveis serviram de entrada para o programa BMDP-2M, que executa também uma análise *cluster*, mas de forma um pouco diferente de BMDP-1M. BMDP-2M agrupa observações, e não variáveis. Cada produção de cada falante foi então tabulada como uma observação única, e os *slopes* entraram sob a forma de variáveis. Como métrica de distância para formar clusters, BMDP-2M oferece várias alternativas; optou-se aqui pela medida mais clássica: a distância Euclidiana.

Como primeira tentativa foram usadas apenas as 6 variáveis correspondentes aos *slopes* de cada uma das faixas previamente selecionadas: 0-500, 500-1000,

1000-2000, 2000-3500, 3500-5000 e 5000-8000 Hz (chamaremos essas variáveis de SL0_5, SL5_10, SL10_20, SL20-35, SL35_50 e SL50_80, respectivamente). Com base nesses *slopes*, BMDP-2M separou corretamente 6 pares de falantes. Esse resultado é um tanto surpreendente, já que temos aqui uma drástica redução de informação: em vez dos 200 pontos do ELT usados anteriormente, temos apenas 6 quantidades para definir todo o espectro.

Mais duas tentativas foram realizadas, uma excluindo SL0_5 e outra excluindo SL50_80, duas faixas que nos testes anteriores pareciam veicular pouca informação. A exclusão de SL50_80 provocou um **aumento** nos acertos, totalizando 7 separações corretas. A exclusão de SL0_5, por outro lado, fez o número de pares corretos cair para apenas 4. É interessante observar que a faixa 0-500 Hz, que nos testes anteriores usando 200 pontos do ELT parecia não ser relevante, mostrou-se aqui - quando expressa como *slope* - mais informativa.

Ao ajustar uma reta a uma determinada faixa do ELT resta sempre um resíduo, isto é, o erro quadrático médio, que é a soma das distâncias de cada ponto do espectro à reta ajustada, ao quadrado, dividida pelo número de pontos na faixa. Esse erro residual é importante, pois é possível ter duas retas com o mesmo *slope* ajustadas a configurações espectrais distintas. Assim, incluimos esses valores residuais (mas mantendo a exclusão de SL50-80). Nessa nova tentativa, obtivemos um total de 9 separações corretas, apenas uma a menos do que obtivéramos com todos os pontos do ELT na faixa 0-5000 Hz (ver tabela 6.1).

Mesmo representando cada faixa do ELT como *slope* + resíduo, estamos ainda deixando escapar uma informação provavelmente relevante, que é a amplitude média, em dB, da faixa em questão. Para normalizar os níveis médios de cada faixa, dividiu-se o nível absoluto dessa faixa pelo nível médio do ELT total (0-8000 Hz; L0_80). Assim, o nível normalizado da faixa 0-500 Hz, por exemplo, será:

$L0\_5_{normal} = L0\_5 \div L0\_80$

Incluímos essa nova informação no conjunto de variáveis, mantendo a exclusão de SL50_80. O número de acertos, no entanto, permaneceu igual ao já obtido sem essas novas variáveis (9 acertos).

A tabela 6.2 resume os resultados dos testes realizados através de BMDP-2M, apresentando os conjuntos de variáveis utilizados e as identificações corretas correspondentes.

Variáveis	**I.C.**
Todos os *slopes*	6
Todos os *slopes*, menos SL0_5	4
Todos os *slopes*, menos SL50_80	7
Todos os *slopes*, menos SL50_80 + resíduos	9
Todos os *slopes*, menos SL50_80 + resíduos + amplitudes normalizadas	9

TABELA 6.2: Índice de acertos (I.C.) a partir da combinação das diversas variáveis.

Os resultados obtidos indicam que apenas a informação fornecida pelos *slopes* de faixas selecionadas do ELT (0-5000 Hz) e pelos erros residuais em cada faixa é praticamente igual à informação extraída a partir de 200 pontos do ELT (0-8000 Hz em intervalos de 40 Hz). Embora não se possa garantir que o procedimento tenha a mesma eficácia para um número maior de falantes, é evidente que a redução dos *bytes* alocados para a codificação de cada falante é conveniente para o tratamento de conjuntos extensos, especialmente no paradigma de verificação automática de falante.

No presente experimento, as faixas do ELT foram selecionadas de modo mais ou menos arbitrário. Estudos posteriores com base em bancos de dados extensos

poderiam otimizar essa seleção estabelecendo maiores pesos para faixas mais informativas do ELT.

6.3.3.4) *Efeito da Velocidade de Emissão no ELT*

O sinal de fala representa o produto entre as características da fonte e a função de transferência do trato (Fant 1960). A transformação efetuada pelas configurações articulatórias depende das propriedades segmentais, mas essas variações de curto termo relacionadas à estrutura fonética são neutralizadas no processo de extração do ELT; assim, o ELT resultante pode oferecer informação relevante sobre as características da fonte, permitindo a detecção de certas alterações no comportamento das cordas vocais. As modificações na forma de onda glotal refletem-se no ELT principalmente no que diz respeito à inclinação global do espectro; nas vozes classificadas como *breathy*, por exemplo, a onda glotal tende a uma senóide, privilegiando os componentes de baixa freqüência, e em especial o primeiro harmônico; no ELT, essa característica da fonte se reflete na forma de um *slope* abrupto (Cf. Klatt e Klatt 1990). A produção de voz com maior tensionamento das cordas vocais cria uma onda glotal quase triangular, cujo efeito no ELT é um *slope* menos abrupto (Kitzing 1986; Klatt e Klatt 1990).

É provável que a produção de fala em velocidade muito rápida esteja associada a modificações significativas na onda glotal. Exames preliminares das amostras dos falantes aqui estudados revelaram a existência de um aumento sistemático de F0 na condição velocidade rápida de emissão (ver discussão na seção 5). Já se verificou que a produção de voz *strained* está associada a um aumento do F0 médio (Kitzing 1986). Esse aumento de F0 pode estar associado a um maior tensionamento das cordas vocais e/ou a um aumento da pressão sub-glotal (Flanagan

1958); no primeiro caso deveríamos esperar uma alteração mais significativa na forma de onda glotal, com conseqüências na configuração do ELT.

O uso prolongado da voz pode produzir também alterações no *slope* do ELT (Löfqvist 1986); esse é um aspecto a ser considerado no presente experimento já que os falantes produziram as amostras de fala rápida ao final da sessão, após uma série de leituras de diversos textos.

De modo a examinar alterações no ELT em função da velocidade de emissão e/ou da fadiga vocal, comparamos os *slopes* de diferentes faixas do ELT. A tabela 6.3 mostra os *slopes* para todas as faixas selecionadas do ELT, para todos os falantes nas duas condições de velocidade de produção.

Verificamos, na tabela 6.3, que o *slope* médio para a faixa integral 0 - 8000 Hz fica em torno de -4.5 dB/oitava, uma inclinação menos abrupta do que a de -6 dB/oitava, prevista na teoria acústica de produção de Fant (1960). A inclusão de fricativas não sonoras no cômputo do ELT pode ter acrescentado um ganho de energia na faixa acima de 5000 Hz, tornando assim o ELT menos abrupto.

Na faixa 0-2000 Hz, os *slopes* aproximam-se da previsão de -6 dB/oitava. Nolan (1983:151-153), estudando a variação dos *slopes* em função de diferentes qualidades de voz, também observa uma inclinação de -6 dB/oitava para a reta ajustada à faixa 0-2500 Hz, na voz modal.

Faixas	V	ZR	EN	R1	ZP	AG	WA	MS	R2	DO
0 - 8000	N	-3.9	-4.1	-4.6	-4.4	-4.6	-4.6	-4.8	-3.9	-3.9
	R	-4.6	-4.3	-4.1	-4.7	-4.5	-4.3	-4.8	-3.9	-4.0
0 - 5000	N	-4.0	-3.3	-4.5	-5.2	-5.0	-3.9	-3.9	-3.8	-4.2
	R	-4.3	-3.7	-4.4	-5.5	-4.9	-3.6	-4.3	-4.2	-4.1
0 - 3500	N	-5.5	-4.1	-6.1	-5.8	-5.8	-3.2	-4.8	-5.5	-5.0
	R	-5.1	-4.7	-6.0	-5.7	-5.5	-3.0	-5.6	-6.0	-4.9
0 - 2000	N	-6.1	-4.9	-6.2	-5.3	-7.2	-4.7	-6.3	-4.9	-5.9
	R	-5.1	-5.4	-6.1	-4.9	-6.5	-4.4	-6.8	-5.6	-5.6
0 - 1000	N	-4.3	-3.5	-3.4	-2.7	-4.4	-3.0	-4.8	-3.0	-4.3
	R	-2.9	-3.8	-3.3	-2.5	-3.8	-3.0	-5.1	-3.1	-4.0
0 - 500	N	+.22	+.66	+1.1	+1.1	+.36	+.88	-.40	+.77	+.06
	R	+1.2	+.41	+1.5	+1.5	+.73	+1.1	-.32	+.77	+.27
500 - 1000	N	-19	-17	-23	-17	-21	-16	-19	-19	-18
	R	-18	-19	-22	-18	-22	-17	-19	-19	-18
1000 - 2000	N	-.86	-3.7	+.51	-6.4	-8.9	-4.4	-.58	+3.2	-7.8
	R	-3.7	-5.8	+.89	-4.4	-6.9	-2.5	-4.5	-3.2	-6.9
2000 - 3500	N	+3.9	+2.1	-.44	-8.0	+6.1	+13	+4.8	-10	+7.5
	R	+.98	+.71	-3.9	-7.7	+5.7	+13	+2.5	-.34	+5.5
3500 - 5000	N	-15	-25	-15	+4.7	-10	-32	-23	-6.0	-25
	R	-28	-24	-23	-.75	-15	-38	-26	-5.2	-27
5000 - 8000	N	-16	-11	-17	-12	-12	-6.6	-17	-17	-9.8
	R	-16	-12	-13	-10	-10	-1.5	-13	-13	-9.1

TABELA 6.3: *slopes* de faixas selecionadas do ELT para as duas condições de velocidade de emissão (N= normal; R= rápida)

A velocidade de produção não altera substancialmente o *slope* da faixa integral 0-8000 Hz. Apenas os falantes ZR e R1 têm uma diferença maior ou igual a 0.5 dB/oitava nessa faixa. A maioria das faixas analisadas não apresentou diferença significativa quanto aos *slopes* em função da velocidade de produção. Apenas o *slope* da faixa 1000-2000 Hz parece variar bastante inter-falantes e em função da velocidade. Essa faixa, porém, é praticamente dominada pela presença de um pico espectral correspondendo ao âmbito de variação de F2; os *slopes* daí extraídos talvez

sejam pouco relevantes, já que a inversão de direção do contorno do ELT torna sensível o ajuste da reta de regressão nessa região. Algo semelhante parece também ocorrer na faixa 2000-3500 Hz, região dominada pelo pico de F3 e pelo começo do pico de F4. A extração de *slopes* a partir de faixas muito estreitas (da ordem de 1000 Hz) parece pouco adequada, podendo gerar indicadores espúrios.

Tomando como base uma faixa ampla do ELT, a velocidade de produção não parece afetar substancialmente a configuração espectral. A figura 6.6 mostra os ELTs médios (todos os falantes) para a velocidade normal (linha contínua) e velocidade rápida (linha tracejada). Observamos que o envelope espectral permanece praticamente o mesmo, assim como o *tilt* espectral, que é de -4.318 dB/oitava na velocidade normal, e -4.372 dB/oitava na velocidade rápida. O ganho de amplitude no ELT correspondente à velocidade rápida é praticamente constante ao longo de todo o espectro, não existindo, aparentemente, regiões mais afetadas pela variação de velocidade de produção. É importante considerar, entretanto, que a figura 6.6 representa os valores **médios** reunindo o conjunto de falantes. Portanto, eventuais diferenças individuais podem ter sido neutralizadas. Mais adiante estudaremos as amplitudes médias de faixas selecionadas do ELT, para cada falante isoladamente, em cada condição de velocidade.

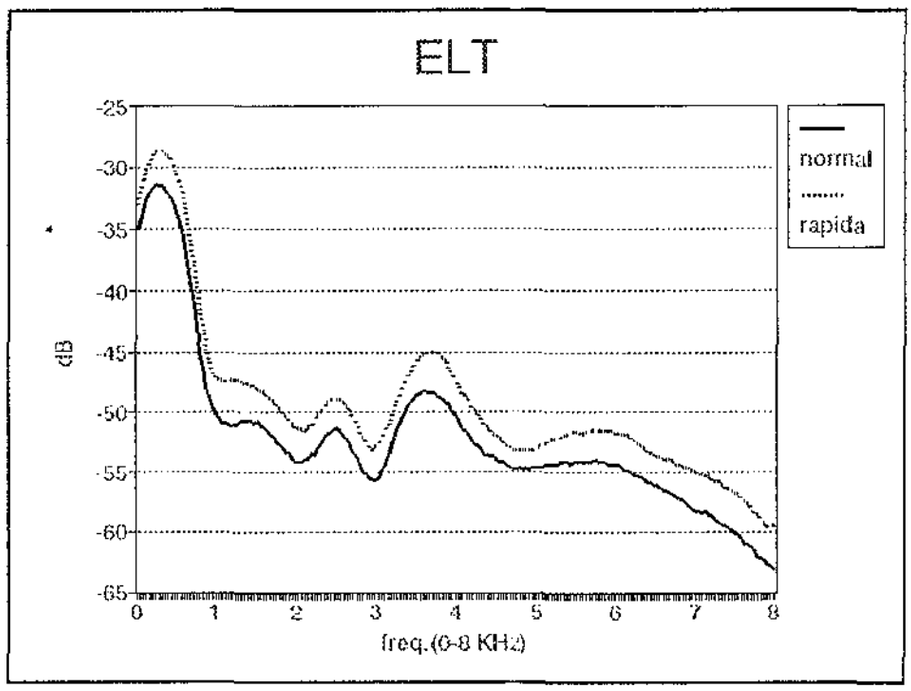


FIGURA 6.6: ELTs médios, reunindo todos os falantes, nas duas velocidades de emissão (linha contínua=velocidade normal; linha pontilhada=velocidade rápida)

A maior amplitude média do ELT correspondendo à velocidade rápida de produção está provavelmente associada a um maior esforço vocal utilizado nessa condição. Como já comentamos anteriormente, embora as condições de gravação não tenham sido estritamente controladas no experimento, procurou-se manter o mesmo nível de entrada para as gravações com o mesmo falante. Assim, parece ser razoável inferir que, efetivamente, a fala rápida foi produzida com maior amplitude.

Na seção 6.3.3.2 observamos que algumas faixas do ELT eram mais informativas do que outras para a identificação dos falantes. Uma maneira de verificar mais localmente a informação contida no ELT é aferir, para cada um dos

200 pontos do ELT, a variabilidade das medidas de amplitude entre os falantes. A figura 6.7 mostra o desvio-padrão inter-falantes para cada um dos 200 pontos do ELT (0-8000 Hz), nas duas velocidades de produção (normal=linha contínua; rápida=linha tracejada). Observamos, na figura 6.7, que a curva do desvio-padrão (i.e. a variabilidade) tem um padrão mais ou menos regular - mais evidente na velocidade normal - com picos de maior variabilidade (i.e maior informação) afastados em intervalos de aproximadamente 1000 Hz. Esses picos de variabilidade parecem corresponder ao espaçamento médio previsto para os formantes em adultos do sexo masculino (Fant 1960), ou seja, as regiões do ELT contendo inflexões correspondentes a formantes médios seriam, potencialmente, mais informativas. Nessa mesma direção aponta o fato de que o máximo da curva do desvio-padrão encontra-se na região de 3500-4000 Hz, exatamente o ponto onde, invariavelmente, ocorre a mais pronunciada inflexão no ELT, correspondente a F4.

Na figura 6.7 podemos observar também que, na velocidade rápida a variabilidade inter-falantes - em relação à velocidade normal - decresce progressivamente a partir de 3500 Hz, ou seja, nessa condição de velocidade, a faixa superior do ELT assume uma configuração parecida para todos os falantes.

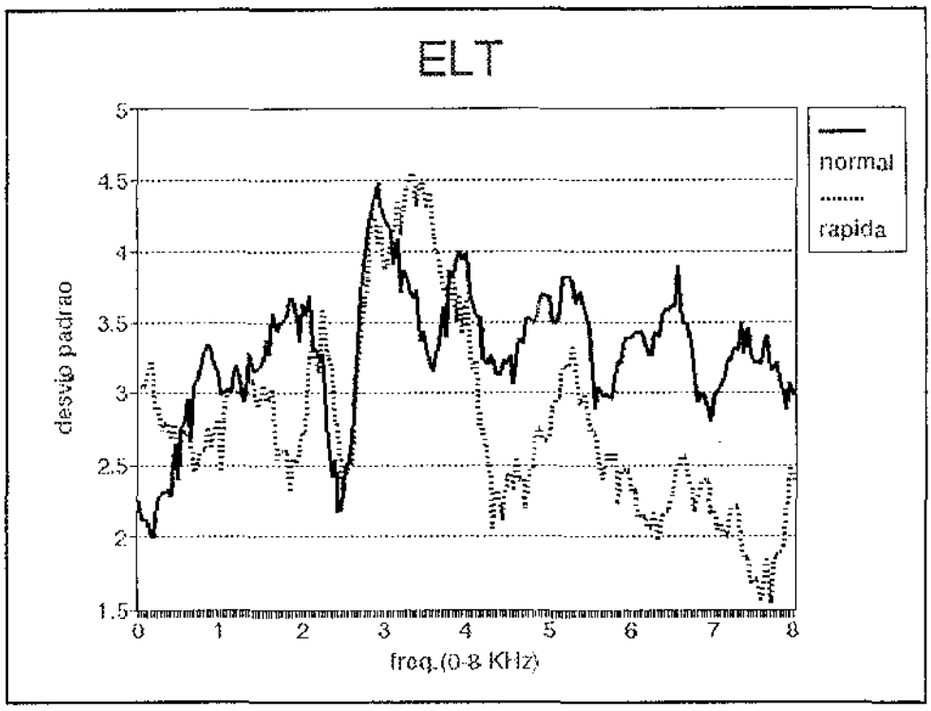


FIGURA 6.7: Desvio-padrão inter-falante para cada ponto no eixo de freqüências do ELT (n=200)

6.3.3.5) *Razões de Amplitude entre Faixas do ELT*

Além dos *slopes* das retas ajustadas ao espectro, outro método freqüentemente utilizado para quantificar diferenças entre ELTs é a razão entre os níveis médios de amplitude de faixas selecionadas do ELT. A delimitação dessas faixas varia entre os autores. Frøkjaer-Jensen e Prytz (1976), usando a medida para aferir mudanças pós-tratamento em pacientes foniátricos, empregam a razão entre a amplitude média da faixa acima de 1000 Hz e a amplitude média da faixa abaixo de 1000 Hz (*apud* Nolan 1983:151); Löfqvist (1986) utiliza como um dos parâmetros para acessar efeitos da fadiga vocal a razão 0-1/1-5 KHz; Kitzing (1986), estudando as características acústicas do ELT em 4 diferentes qualidades de voz (*normal, leaky, strained* e *soft*), emprega, entre outros parâmetros, as razões 0-1/1-5, 0-1/1-2, 0.3-0.8/1.5-3.0 e 0.3/1.5-2.0 KHz; Nolan (1983), verificando alterações no ELT para

diferentes *settings* articulatórios produzidos por ele próprio e pelo foneticista John Laver, testa diversas combinações, observando que a medida mais efetiva, ou seja, aquela que melhor separou os diferentes *settings*, foi a razão entre as faixas 1500-3000 Hz e 0-1500 Hz.

A escolha dessa ou daquela delimitação do ELT tem sempre, nos trabalhos que examinamos, uma base empírica. No presente trabalho optamos, então, pelo teste de diversas faixas. Calculamos assim a razão em dB entre as seguintes faixas:

(1) 0-500/500-1000, (2) 0-500/500-1500, (3) 500-1500/1500-2500, (4) 0-1000/1000-2000, (5) 1000-2000/2000-3500, (6) 1000-2500/2500-3500, (7) 1000-2500/2500-5000, (8) 1000-3500/3500-5000, (9) 0-1500/1500-2500, (10) 0-2000/2000-3500, (11) 0-2500/2500-3500, (12) 0-2500/2500-5000 e (13) 0-3500/3500-5000 Hz.

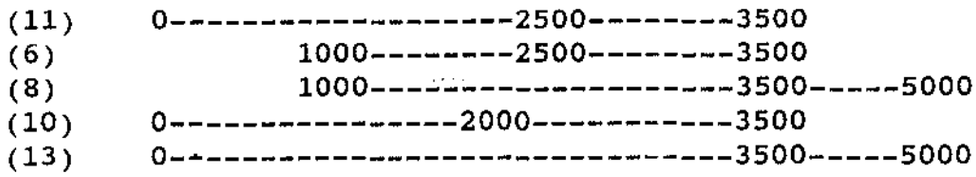
A tabela 6.4 mostra as razões entre as faixas do ELT nas duas condições de velocidade, as médias e o desvio-padrão inter-falantes para cada tipo de medida. As razões expressam a relação de amplitude da faixa de freqüência mais baixa sobre a mais alta; assim, um valor positivo significa que a faixa mais baixa tem uma amplitude maior. O desvio-padrão serve aqui para aferir aproximadamente a eficácia de cada uma das razões de amplitude para separar os falantes (Cf. Nolan 1983:151).

fal.	V.	1	2	3	4	5	6	7	8	9	10	11	12	13
ZR	N	11.6	15.8	6.9	14.5	1.9	1.1	-1.1	-2.9	11.0	9.2	7.3	5.1	1.6
	R	9.8	12.6	6.2	13.0	3.0	3.2	1.9	-.40	10.3	9.5	8.7	7.4	3.8
EN	N	10.0	12.8	5.2	11.5	.40	-1.0	-1.3	-1.1	9.3	6.1	3.9	3.6	2.2
	R	10.4	13.2	6.3	12.5	.90	-.10	-.80	-1.1	10.6	7.2	5.2	4.5	2.6
R1	N	9.6	15.3	6.8	16.3	3.0	3.0	-.20	-3.6	11.8	11.2	9.7	6.5	1.6
	R	9.6	15.4	6.0	16.3	2.7	3.6	-.10	-4.1	11.0	10.8	10.3	6.6	1.0
ZP	N	7.8	11.8	7.6	13.6	4.8	5.6	5.0	2.3	11.5	11.6	11.3	10.7	7.1
	R	7.6	11.4	7.0	12.7	5.3	5.9	6.3	4.2	10.7	11.7	11.4	11.8	8.8
AG	N	11.8	16.3	9.1	17.3	-.30	-1.6	-.80	.30	14.4	8.4	5.6	6.4	5.3
	R	10.1	14.8	8.8	16.3	.10	-1.7	-.30	1.2	13.6	8.2	5.1	6.5	6.0
WA	N	8.8	11.6	5.7	11.3	-2.2	-4.6	.10	5.1	9.4	3.5	.30	5.0	8.0
	R	8.6	11.4	4.7	10.6	-2.2	-4.5	.00	4.8	8.4	3.1	.00	4.5	7.5
MS	N	12.6	16.6	4.1	14.6	-1.4	-1.3	-1.4	-.90	9.5	5.9	4.5	4.4	3.0
	R	13.4	17.2	5.1	15.5	.20	.20	-1.2	-2.3	11.4	8.0	6.5	5.1	2.2
R2	N	8.3	13.0	3.7	12.7	4.5	6.4	1.9	-3.7	7.9	10.9	11.6	7.1	.80
	R	8.5	13.1	8.0	14.4	4.3	3.2	-.20	-3.8	12.2	11.5	9.5	6.1	1.1
DO	N	11.4	14.3	6.9	13.6	.70	-.10	-.20	-.20	11.6	7.5	5.6	5.5	3.8
	R	10.7	13.4	6.6	12.6	1.6	.70	.30	-.20	10.9	7.9	6.1	5.7	3.8
Med.	N	10.2	14.1	6.1	13.9	1.26	.83	.22	-.52	10.7	8.2	6.6	6.0	3.7
	R	9.7	13.6	6.6	13.8	1.76	1.16	.65	-.19	11.0	7.5	6.9	6.4	4.1
Dv. Pad.	N	1.71	1.93	1.70	2.01	2.47	3.58	2.05	2.87	1.91	2.76	3.71	2.06	2.56
	R	1.70	1.89	1.21	1.96	2.32	3.16	2.28	3.16	1.41	2.66	3.47	2.22	2.78

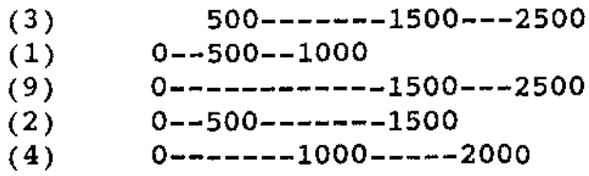
TABELA 6.4: Razões de amplitude entre diversas faixas do ELT; as duas últimas linhas dão as médias e desvios-padrão inter-falante

Os maiores desvios-padrão, para as duas condições de velocidade foram para as razões de amplitude (11), (6), (8), (10) e (13). Os menores, ou seja, aqueles correspondentes às razões menos eficazes para separar falantes, foram (3), (1), (9), (2) e (4). Esquematicamente temos:

Maiores Desvios:



Menores Desvios:



Observamos que as razões que utilizam faixas mais amplas do ELT tendem a ser mais informativas. As razões 0-2500/2500-3500 e 1000-2500/2500-3500 apresentam a maior extensão de variação inter-falante, superior a 10 dB. É interessante observar que as duas razões com maior variabilidade (11 e 6) não têm informação acima de 3500 Hz.

SETTINGS SUPRA-LARÍNGEOS				
	upper/lower 1.5-3/0-1.5 KHz		*slope approximation (0 - 2.5 KHz)*	
SETTING	*Speaker JL*	*Speaker FN*	*Speaker JL*	*Speaker FN*
neutral	-22	-19	-6	-6
raised larynx	-19	-17	-4	-3
low. larynx	-19	-32	-5	-11
spread lips	-20	-20	-5	-11
open round.	-19	-22	-6	-6
close round.	-21	-22	-5	-7
retroflex	-16	-20	-4	-5
lar.-pharyng.	-16	-20	-2	-5
pharyngalised	-20	-21	-4	-6
uvularised	-20	-23	-4	-6
velarised	-19	-21	-4	-6
palatalised	-18	-17	-4	-6
pal.-alveolar.	-20	-18	-4	-6
alveolarised	-19	-22	-5	-6
dentalised	-19	-21	-5	-6
nasalised	-19	-18	-4	-6
denasalised	-20	-19	-4	-4
close jaw	-20	-22	-5	-6
open jaw	-21	-22	-5	-6
SETTINGS LARÍNGEOS				
modal	-20	-20	-4	-5
falsetto	-27	-26	-9	-9
creak	-20	-17	-3	-4
whispery	-17	-18	-5	-5
whisp. falsetto	-24	-27	-8	-8
whisp. creak	-15	-19	-3	-6
creaky voice	-18	-18	-3	-4
creaky fals.	-18	-17	-3	-4
breathy voice	-23	-24	-5	-8
harsh ventric.	-10	-4	-3	-0
harsh.ventric. whisp. falsetto	-16	-11	-6	-2

TABELA 6.5: Razões de amplitude (dB) entre as faixas 1.5 - 3.0 KHz e *slopes* (dB/oitava) da faixa 0 - 2.5 KHz para diferentes qualidades de voz (*settings* articulatórios); reproduzido de Nolan 1983:153, tabela 4.1)

A razão 0-1500/1500-2500 Hz tem um dos mais baixos desvios-padrão, contrastando com a constatação de Nolan (1983:151), que encontra na razão 1500-3000/0-1500 a maior variabilidade. Nolan relata uma diferença de cerca de 20 dB entre a faixa superior (1500-3000 Hz) e a inferior (0-1500 Hz) e um *slope* de -6 dB/oitava no *setting* "neutro" para a faixa 0-2500 Hz (ver tabela 6.5). Esse *slope* é próximo ao que encontramos para a faixa de 0-2000 Hz, com média de -5.72 dB/oitava (ver tabela 6.4). A razão de amplitude entre as faixas 0-1500/1500-2500 Hz que encontramos em nossos falantes está, porém, bem abaixo da observada por Nolan (cerca de 10.5 dB *versus* os 20 dB de Nolan). Essa diferença pode estar relacionada, em parte, (a) com a filtragem *low-pass* com corte em 5000 Hz nos dados de Nolan; na nossa análise o sinal foi filtrado a 8000 Hz, o que pode ter produzido um pequeno ganho na faixa superior, e em parte (b) com diferentes condicões na captação do sinal (características de gravador, fita, transdutores, etc.). Parte da diferença nos resultados, porém, deve estar também relacionada com características individuais dos falantes; para a razão de amplitude em questão, observamos um valor mínimo de 7.9 dB (falante R2, velocidade normal) e um máximo de 14.4 dB (falante AG, velocidade normal). Esse âmbito de 6.5 dB é superior à variação entre a maior parte dos *settings* produzidos no experimento de Nolan (excluindo apenas alguns *setttings* laríngeos que se afastam fortemente da média; ver tabela 6.5).

É preciso considerar que o paradigma do experimento de Nolan (1983) é diferente do aqui realizado. Nolan avalia diferentes *settings* articulatórios produzidos pelo mesmo falante, enquanto aqui trata-se, efetivamente, de diferentes falantes. A simulação de diferentes qualidades de voz por um mesmo falante determina fontes de variabilidade distintas daquelas encontradas em um conjunto de diferentes falantes, já que, no primeiro caso, serão forçosamente mantidas algumas

características não passíveis de manipulação. Nesse sentido é interessante observar, que, mesmo nas medidas espectrais muito genéricas utilizadas por Nolan (*slopes* e razões de amplitude) a variação entre os dois falantes para o mesmo *setting* é freqüentemente maior do que a variação entre os *settings* para o mesmo falante (ver tabela 6.5).

Uma das questões que se coloca é saber até que ponto ELTs do mesmo falante simulando diferentes qualidades de voz não preservariam os picos referentes aos formantes altos, especialmente nos *settings* onde não há alteração significativa na posição da laringe. Outro aspecto, ressaltado pelo próprio Nolan (1983:155), é a presença de vales no ELT decorrentes, provavelmente, do acoplamento acústico com a traquéia. Essas depressões no ELT permaneceriam inalteradas mesmo com grandes alterações no tipo de fonação, já que dependem diretamente das dimensões subglotais.

A influência de zeros e formantes nasais também poderia gerar características mais ou menos constantes para o mesmo falante, no experimento relatado por Nolan (1983). Embora se possa variar o grau de nasalização, alguns aspectos acústicos diretamente relacionados com a cavidade nasal - que é fixa - não devem se alterar substancialmente em função de diferentes formas de fonação ou de diferentes *settings* supra-laríngeos.

Os resultados de Nolan (1983) indicam que o ELT sofre alterações maiores em função de modificações no mecanismo fonatório do que em função de *settings* afetando a tensão muscular e a postura global do trato como um todo. Mas como esses resultados poderiam ser extrapolados para o paradigma da identificação de falantes? As próprias medidas obtidas por Nolan parecem um tanto contraditórias, como podemos observar na tabela 6.5.

Podemos verificar na tabela 6.5 que as alterações nas razões de amplitude (1500-3000/0-1500 Hz) e nos *slopes* (0-2500 Hz) em função dos diversos *settings*

diferem entre os dois falantes do experimento (John Laver=JL, e o próprio Francis Nolan=FN) não só quanto à magnitude mas também - e esse é o aspecto de difícil interpretação - quanto à **direção** da mudança. Assim, para o falante JL, o abaixamento da laringe provoca um **ganho** de 3 dB na faixa 1500-3000 Hz, enquanto para o falante FN, o mesmo *setting* provoca um **decréscimo** de amplitude de 13 dB na mesma faixa (o mesmo fato se reflete no *slope* mais abrupto de FN para a faixa 0-2500 Hz). Efeitos opostos para os dois falantes ocorrem em uma série de outros *settings*, especialmente os supra-laríngeos. Nos *settings* laríngeos há uma maior coerência nos resultados, com a única exceção do *slope* de FN para o *setting whispery creak*, que para essa qualidade de voz torna-se mais abrupto do que na voz modal (ao contrário do que ocorre com JL).

Nolan tenta explicar algumas dessas discrepâncias argumentando que alguns *settings* não teriam sido implementados corretamente (por ele, Nolan), segundo crítica auditiva de John Laver. Assim, de acordo com Laver, na performance de Nolan:

> *phonation type tended towards creaky voice rather than modal; (...) most settings were accompanied by a slightly (...) degree of nasalisation; (...) modal voice with high fundamental frequency replaced falsetto in creaky falsetto; (...) harsh ventricular whispery falsetto approximated more to harsh ventricular creaky voice with high fundamental frequency.* (Nolan 1983:147).

Na verdade, a dificuldade em controlar adequadamente a produção de cada *setting* é admitida pelo próprio Laver; ao verificar que na sua própria performance com levantamento da laringe o comportamento dos formantes afastara-se da previsão

(deveria haver um aumento de freqüência, em lugar da queda efetivamente observada), Laver sugere que, durante a produção do *setting*,

> *...the sustained muscular effort to keep the larynx high may very weell have unwittingly resulted in a severely constricted pharynx* (Laver 1980:27);

a faringalização teria, portanto, suplantado o efeito da alteração na laringe, provocando a redução de freqüência nos formantes [3].

A dificuldade é maior nos *settings* supralaríngeos: em 13 desses *settings* (de um total de 19) podemos observar efeitos opostos na razão de amplitude 1500-3000/0-1500 Hz para os falantes JL e FN. É possível que o controle da produção dos *settings* supralaríngeos seja mais crítico, já que uma alteração na postura global envolve um reajuste constante da programação motora, especialmente em função da maior ou menor suscetibilidade de diferentes segmentos (para a noção de *suscetibilidade* ver Laver 1980:20ff). Por outro lado, nas qualidades de voz envolvendo alterações na fonte, a postura pode ser mantida sem grandes dificuldades ao longo de toda a cadeia segmental; de fato, podemos verificar na tabela 6.5 que as medidas dos dois parâmetros (razão de amplitude e *slope*) nos *settings* laríngeos são mais consistentes para os falantes JL e FN.

Parte do problema pode estar na natureza muito genérica das medidas utilizadas por Nolan (1983) para caracterizar os diferentes *settings*. É possível que alguns *settings* provoquem alterações demasiado locais no ELT para serem detectadas por razões de amplitude e *slopes*, parâmetros que podem permanecer invariantes para *n* configurações espectrais diferentes. Nesse sentido é interessante observar, na tabela 6.5, que os *slopes* - especialmente nos *settings* supralaríngeos e

no falante FN - são invariantes para uma série de *settings*, embora haja alteração considerável nas razões de amplitude (v. p.ex. *palatalizado vs. alveolarizado*, falante FN).

A conclusão de que o ELT é menos suscetível a alterações supralaríngeas do que aos diferentes modos de fonação deve ficar, portanto, restrita a alguns aspectos globais do ELT, tais como aqueles expressos pelas medidas usadas por Nolan (1983). A utilização de outros métodos de quantificação do ELT, mais sensíveis a perturbações locais, poderia levar a um quadro diferente; lembramos que, ao realizarmos a análise *cluster* com métrica baseada na correlação de 200 pontos do ELT (v. seção 6.3.3.1), foi possível separar corretamente os gêmeos incluídos no experimento, apesar da aparente semelhança dos ELTs. Técnicas baseadas em correlação tornar-se-iam, provavelmente, ainda mais eficientes se fosse utilizado um maior número de pontos e/ou fosse empregado um filtro de análise mais estreito, de modo a ressaltar a estrutura fina do ELT [5].

6.4) *Comentário Final*

No presente trabalho estudamos diversos parâmetros extraídos a partir do Espectro de Longo Termo (ELT). De uma forma geral, o ELT é um dos indicadores mais seguros da identidade do falante. Para o nosso grupo de falantes (n=10; 7 + R1/R2 + JA + JR) foi possível expressar o ELT através de medidas mais globais (*slopes* + resíduos) sem que a perda de informação local reduzisse drasticamente a distinção entre os falantes. Conjuntos maiores de falantes, entretanto, devem ser analisados de modo a avaliar mais seguramente a eficácia dessas medidas globais; é provável que a informação local torne-se muito mais importante para grupos maiores.

Os resultados que obtivemos sugerem que a informação contida no ELT não se distribui homogeneamente ao longo de todo o espectro. Transformações que privilegiassem as regiões mais informativas, enfatizando, por exemplo, pontos de inflexão mais acentuada (picos e vales), poderiam criar ELTs ajustados mais eficientes para a identificação de falantes.

Embora não seja certo que a utilização de filtros de análise mais estreitos que o aqui empregado (300 Hz) aumentasse necessariamente a distinção entre os ELTs, é provável que alguns detalhes da estrutura fina que assim surgissem fossem relevantes para o tratamento de conjuntos maiores de falantes.

Uma outra possibilidade seria o emprego de espectros transformados de acordo com um modelo perceptual, baseado em dimensões mais diretamente relacionadas com processos do sistema auditivo: resolução em bandas críticas (BARK : v. Zwicker 1961), filtros assimétricos e ajustáveis segundo a faixa de freqüência, não linearidade da escala de *loudness*, relação não linear freqüência/*loudness*, etc (Cf. Lindblom 1986).

Embora aparentemente atrativa, a adoção de um modelo perceptual para o ELT não é de fácil justificação teórica, pois sabemos que o processamento humano não exige os 10-15 segundos necessários para que o ELT se estabilize; vários experimentos têm demonstrado que ouvintes sem treinamento conseguem identificar falantes (conhecidos e desconhecidos), acima do acaso, a partir de trechos de fala muito curtos, da ordem de alguns centisegundos (Pollack *et al.* 1954; Compton 1963). É claro que testes de laboratório não refletem necessariamente o que humanos **fazem**, mas antes o que são **capazes de fazer.** Assim, é provável que a informação espectral de longo termo seja utilizada complementarmente no processamento humano.

Embora se possa afirmar que o ELT é um indicador efetivo da qualidade de voz e que, como tal, é uma excelente pista para a identidade do falante, é preciso

avaliar também algumas limitações que, em condições menos controladas que as dos testes de laboratório (como a situação forense, por exemplo), podem dificultar, ou mesmo inviabilizar seu uso. O ELT pode ser bastante sensível a certas características do meio e do canal de transmissão tais como: presença de ruído ambiental, distorção harmônica do canal telefônico, características do microfone e fita magnética, etc. Além desses aspectos, algumas condições do falante também podem alterar consideravelmente a configuração do ELT, tais como disfarce, presença de *stress* psicológico, rouquidão, etc. O experimento de Nolan (1983), acima comentado (seção 6.3.3.5), permite visualizar a magnitude da variação do ELT para um mesmo indivíduo, evidenciando o tipo de dificuldade que pode surgir na situação forense típica, onde a possibilidade de disfarce deve ser sempre considerada.

No paradigma da Verificação Automática de Falante (VAF), as dificuldades são menores, já que é possível normalizar uma série de condições do meio e dos canais de transmissão. A presença de disfarce é, nessa situação, bastante improvável, já que o falante em geral é cooperativo. A possibilidade de imitação, por outro lado, deve ser considerada mais cuidadosamente nesse modelo. Embora não existam estudos específicos sobre essa questão, a imitação não deve trazer problemas para o emprego do ELT na VAF; como vimos na seção 6.3.3.1, medidas baseadas no ELT permitiram separar corretamente um par de gêmeos monozigóticos, apesar da extrema semelhança auditiva entre as vozes. A limitação maior do uso do ELT, no caso da VAF, é de natureza mais prática do que teórica, na medida em que, para a maioria das aplicações, parece pouco razoável colher uma amostra de cerca de 15 segundos de fala. É possível, no entanto, que em sistemas dependentes de texto intervalos de tempo menores possam ser empregados.

SEÇÃO 7: ASPECTOS RÍTMICO-TEMPORAIS

7.1) *Introdução*

Há uma série de questões não resolvidas relacionadas à qualificação e à quantificação dos aspectos rítmico-temporais da fala. Nolan (1983:127) observa, com pertinência, que embora as durações segmentais contribuam para a especificação dos fenômenos rítmicos, estes são tradicionalmente considerados como propriedades prosódicas ou suprasegmentais. A própria noção de Ritmo é por si só problemática, na medida em que implica uma natureza dual, envolvendo não só uma **base temporal,** onde deverá ocorrer algum evento periódico, como também a definição de uma **estrutura**, ou seja, a definição do tipo de evento (ou unidade) que está sendo periodicamente repetido ao longo dessa base temporal. Essa estruturação rítmica envolve, a princípio, uma organização hierarquizada em vários níveis (sílaba, pé, palavra, sentença, pausas, etc), embora não se entenda completamente como esses diferentes níveis interagem entre si. A principal dificuldade é estabelecer um mapeamento entre os parâmetros acústicos do sinal de fala e seus correlatos perceptuais; de modo geral essa relação não é claramente biunívoca e a própria noção de periodicidade (ou regularidade rítmica) não pode ser definida acusticamente de forma inequívoca.

A dificuldade em estabelecer uma correspondência entre as dimensões acústica e auditivo-perceptual reflete-se na clássica discussão em torno da categorização das línguas como de ritmo "acentual" *versus* "silábico". Essa distinção, consagrada por Pike (1945) e mais tarde reformulada por Abercrombie (1967:97ff), tem sido, entretanto, alvo de intensa controvérsia. Segundo a definição de Abercrombie, em uma língua de ritmo "silábico":

> *the periodic recurrence of movement is supplied by the syllable-producing process: the chest-pulses, and hence the syllables, recur at equal intervals of time - they are isochronous* (Abercrombie 1967:97, grifo do autor),

enquanto nas línguas "acentuais"

> *the periodic recurrence of movement is supplied by the stress-producing process: the stress-pulses, and hence the stressed syllables, are isochronous* (Abercrombie 1967:97).

As definições de Abercrombie são problemáticas em dois sentidos. Em primeiro lugar, a associação entre a produção de sílabas e a atividade respiratória, inspirada na teoria de Stetson (1951); segundo Stetson, cada sílaba seria iniciada por uma contração muscular da caixa toráxica, fazendo aumentar a quantidade de ar expelida pelos pulmões. A investigação direta da atividade muscular, entretanto, nunca confirmou a hipótese de Stetson, e sua teoria é hoje de difícil aceitação (Ladefoged 1975:221).

O segundo ponto controverso na formulação de Abercrombie diz respeito à noção de "isocronia", explicitamente colocada nas duas definições. O fato é que estudos instrumentais examinando uma possível regularidade nas línguas classificadas como de ritmo silábico (isocronia ao nível da sílaba) e ritmo acentual (isocronia ao nível do pé métrico) não têm observado evidências acústicas suportando tal classificação (Nakatani *et al.*; Roach 1982; Dauer 1983; Fletcher 1991; Fant *et al.* 1991a).

A falta de evidências acústicas suportando a hipótese de isocronia tem levado alguns pesquisadores a encarar o fenômeno da regularidade rítmica como estritamente perceptual; segundo esse ponto de vista, em virtude da relação complexa entre o sinal acústico de fala e o percepto, não seria surpresa se a regularidade **percebida** não encontrasse paralelo direto na regularidade **objetiva**. Lehiste (1973) verifica que, embora os ouvintes avaliem mais objetivamente diferenças duracionais em material não-lingüístico, as mesmas diferenças não são igualmente perceptíveis na escuta de fala natural; Lehiste conclui que, se os ouvintes não são capazes de detectar as diferenças de duração em unidades rítmicas de fala, seria razoável assumir que eles ouvem essas unidades como sendo, em algum sentido, de duração igual, provocando assim uma impressão de isocronia ao nível perceptual (v. também Benguerel e D'Arcy 1986).

A falta de uma correlação direta entre os domínios acústico e perceptual não parece ocorrer apenas no âmbito das durações; Ladefoged e Broadabend (1957), em um experimento bastante conhecido, verificam que a percepção da qualidade vocálica depende do contexto fonético da sentença onde a vogal alvo está embutida. Da mesma forma, sabe-se que um /a/, com a mesma intensidade **física** de um /i/ será avaliado como perceptualmente menos intenso (i.e , mais baixo na escala de *loudness*; Cf. Benguerel e D'Arcy 1986).

Para as dimensões de intensidade e freqüência são freqüentemente empregadas unidades perceptualmente relevantes, obtidas através de experimentos de laboratório envolvendo técnicas de mascaramento e avaliação subjetiva de intervalos escalares; assim, expressa-se a intensidade perceptual em *sones* e a freqüência em *mels* ou *barks*, por exemplo. Falta, entretanto, para a dimensão temporal, unidades que expressem convenientemente a avaliação perceptual dos fenômenos periódicos da fala fluente. Uma sugestão interessante nesse sentido é o conceito de *P-Center*, introduzido por Marcus (1981). O *P-Center* seria o "instante

perceptualmente relevante", isto é, o instante em que, para o ouvinte, se estabeleça um referencial rítmico/temporal no interior da cadeia de fala; esse ponto não seria, necessariamente, "objetivamente" definido no interior do estímulo em função da inflexão de alguma variável acústica, dependendo antes da apreensão *gestáltica* do ouvinte.

O cônceito de *P-Center*, embora teoricamente interessante, não foi suficientemente examinado no nível experimental. Um dos poucos estudos existentes (Nord 1991), baseia-se em estímulos de fala sintética, onde os ouvintes ajustam mecanicamente a duração da pausa P em uma seqüência do tipo

/a:/...P.../Ca:/...P.../a:/,

(onde /C/ pode ser {/#, m, b, br, k, s, sp/}), de modo a fazer a seqüência de sílabas ritmicamente estável. Nord (1991) verifica que o evento de tempo definido pela distância entre os *onsets* vocálicos ocorre perto da fronteira /Ca:/ se a posição consonantal está vazia (C = #), mas tende a se deslocar para dentro do domínio consonantal se uma ou mais consoantes precedem a vogal. Particularmente relevante nos resultados de Nord (1991) é a observação de que, para certas combinações de sons envolvendo *clusters* consonantais (/br, sp/) na posição C, parece ser possível "escutar" de dois modos diferentes, fazendo o pulso rítmico perceptualmente ajustado concordar com a seção consonantal ou, alternativamente, com a seção vocálica. Nord (1991) atribui, a princípio, essa variação a características estilísticas do falante, embora seja razoável supor que fatores relacionados ao ouvinte também estejam em jogo. de qualquer forma, os resultados de Nord (1991) sugerem que mais testes experimentais devem ser realizados antes de se poder empregar operacionalmente a noção de *P-Center*.

A falta de evidência experimental indicando a existência de regularidade no domínio das durações tem levado alguns pesquisadores a estabelecer hipóteses alternativas relacionando a percepção do ritmo da fala a fatores estruturais mais gerais. Fant (1991a), examinando aspectos duracionais de três línguas diferentes (Sueco, Francês e Inglês), verifica que a percepção do Francês como uma língua "silábica" depende menos da duração física observada do que de fatores relacionados à estrutura silábica (predominância de sílabas abertas, especialmente CV) e ao grau de redução vocálica. Assim, a diferença entre ritmo "silábico"e "acentual" pode ser colocada nos seguintes termos:

> *Stress timing is not a matter of physical isochrony of inter-stress intervals but a perceptual dominance of heavy syllables, the sucession of which is sensed as quasiperiodical. A language is sensed as syllable timed when the differences between stressed and unstressed syllables are reduced. This involves both a reduction os stress cues and a relatively greater precision and uniformity of unstressed syllables* (Fant *et al.* 1991a: 363-4).

A formulação de Fant *et al.* (1991a) coloca a possibilidade de uma versão "fraca" da diferença entre ritmo acentual e silábico, expressa apenas em termos de uma tendência a esse ou aquele padrão rítmico de base. Assim, embora não se espere encontrar padrões absolutamente isócronos, é possível observar alguns processos, mesmo na dimensão da duraçãc, que aproximem uma língua de um determinado padrão. Hoequist (1983), por exemplo, verifica que, no Espanhol (uma língua tradicionalmente classificada como "silábica") o alongamento de sílaba em final de palavra (um fenômeno geralmente observado nas línguas "acentuais") é pequeno ou inexistente , assim como é menor o aumento duracional nas sílabas [+ tônica].

A versão fraca da distinção ritmo silábico/acentual permite supor uma escala gradual onde as diferenças seriam apenas relativas. Essas diferenças poderiam ser estabelecidas entre diferentes línguas, diferentes dialetos de uma mesma língua, ou mesmo entre diferentes estilos de fala dentro do mesmo dialeto. Major (1981), examinando propriedades rítmicas do Português Brasileiro, verifica a existência de alguns processos na fala casual que parecem favorecer um padrão rítmico acentual (levantamento de algumas vogais, monotongação de ditongos, apagamento de sílabas [- tônica], etc); por outro lado, na forma de citação há poucos processos de encurtamento, sugerindo que um enunciado nesse estilo de fala tende a possuir um padrão rítmico mais próximo do silábico.

Uma questão importante, no que diz respeito à Identificação de Falantes, é saber se a gradação ritmo silábico/acentual encontraria paralelo nas diferenças entre indivíduos falantes da mesma língua. No caso do Português Brasileiro, por exemplo, parece haver uma tendência ao ritmo silábico em alguns dialetos do Sul do país, e é provável que, mesmo entre falantes do mesmo dialeto, existam diferenças que se possam referir a uma distinção genérica ritmo silábico *versus* acentual. Uma dificuldade aqui é determinar os limites da variabilidade intra-falante em função do estilo de fala, que, como vimos acima no trabalho de Major (1981), pode alterar consideravelmente o padrão rítmico de base (pelo menos no Português Brasileiro). A maior dificuldade, entretanto, consiste na já comentada falta de definição de correlatos acústicos mensuráveis que pudessem expressar inequivocamente a complexa percepção do contraste ritmo silábico/acentual na fala fluente; não surpreende, portanto, o fato de não encontrarmos estudos empregando esse contraste na Identificação ou Verificação de falantes.

7.2) *Eficiência de Aspectos Rítmico-Temporais na Identificação de Falantes*

Vários estudos têm tentado acessar características individuais rítmico-temporais através de diferentes medidas acústicas. Alguns experimentos baseiam-se em avaliações subjetivas de estímulos manipulados de modo a isolar pistas exclusivamente temporais; Abberton e Fourcin (1978) verificam que, mesmo com F0 e características espectrais sinteticamente normalizados para todos os falantes, os ouvintes conseguem identificar corretamente cerca de 60% dos falantes, com base apenas em pistas rítmico-temporais (v. também van Dommelen e Win 1987 para resultado semelhante).

Outra abordagem consiste em testar estatisticamente parâmetros temporais extraídos do sinal de fala. Vários tipos de medidas rítmico-temporais já foram estudados quanto à sua eficácia no reconhecimento de falantes: velocidade de emissão em sílabas por segundo com ou sem a inclusão das pausas ("taxa de fala" ou *speech rate* e "taxa de articulação" ou *articulation rate*, respectivamente); padrão de pausas (respiratórias e/ou de hesitação); razão entre tempo de fala vozeado e não vozeado; razão entre duração consonantal e duração vocálica; distribuição tempo/energia (isto é, tempo total acumulado em níveis quantizados de energia pré-determinados) (Cf. Clarke e Becker 1969; Doherty 1975; Reich *et al.* 1976; Doherty e Hollien 1978; Johnson *et al.* 1984; Hollien 1990:242-3; Nolan 1983:127ff).

Em geral tem-se verificado que medidas baseadas na dimensão temporal são menos eficazes para o reconhecimento de falantes do que outras medidas de longo termo - tais como distribuição de F0 e espectro de longo termo -, embora o desempenho possa ser melhorado quando são empregados vetores múltiplos consistindo na combinação de diferentes parâmetros temporais (Johnson *et al.* 1984; Hollien 1990). Por outro lado, é preciso considerar que os fatores prosódicos da fala

são, em geral, bastante resistentes a diversos tipos de degradação do sinal. Sambur e Jayant (1976) verificam que medidas de F0 e energia extraídas via LPC não são perturbadas se a razão sinal/ruído permanece maior que 12 dB; esse é um aspecto importante para a situação forense típica, onde a má qualidade das gravações é quase uma constante.

Uma limitação mais séria ao emprego de alguns parâmetros temporais diz respeito à possibilidade de um efeito degradante na robustez dos vetores de identificação em função de variações condicionadas ao estado afetivo do falante e/ou tentativa de disfarce (Johnson *et al.* 1984). Outro aspecto importante é a variação dos padrões prosódicos em função do estilo discursivo; Fónagy (1978) emprega estímulos baseados exclusivamente no sinal laringográfico (e, portanto, sem informação espectral) veiculando diferentes estilos de fala (leituras de poesia, tragédia clássica e conto de fadas, conversação, noticiário, etc.) e verifica que as pistas prosódicas são suficientes para que a maioria dos ouvintes reconheça a intenção comunicativa original do falante. A possibilidade de variação intra-subjetiva em função do estilo discursivo ou do estado afetivo pode ser um fator complicador para algumas aplicações forenses, já que a situação de coleta da amostra de confronto ("voz-padrão") envolve, geralmente, alteração no estado emocional do falante (medo, *stress* psicológico, etc) e/ou o estabelecimento de uma condição muito formal, com possíveis conseqüências no estilo de fala, especialmente na dimensão prosódica.

A potencialidade dos aspectos rítmico temporais para o reconhecimento de falantes ainda não foi exaustivamente analisada e seria prematura qualquer opinião conclusiva a respeito da validade desse tipo de parâmetro. A própria complexidade do fenômeno RITMO permite supor que outras dimensões analíticas possam ser exploradas. Crystal (1969), em sua discussão sobre a percepção da *ritmicidade*, sugere três aspectos: *rhythmic/arhythmic*, *spiky/glissando*, e *staccato/legato*, o

primeiro definido como o grau de *loudness* relativo nas sílabas tônicas do enunciado (maior grau ≡ *rhythmic*), o segundo como a menor ou maior suavidade dos saltos de *pitch* entre sílabas (ataques mais abruptos ≡ *spiky*), e o terceiro como a maior ou menor diferença duracional entre as sílabas fortes e fracas (menor diferença ≡ *legato*). As definições de Crystal são um tanto problemáticas, na medida em que recorrem a dimensões perceptuais de difícil confirmação experimental (*loudness* e *pitch*), enquanto o contraste *staccato/legato* aproxima-se, aparentemente, da já discutida - e também problemática - distinção ritmo silábico/acentual. A noção mais produtiva parece ser o contraste *spiky/glissando*, já que pode ser expresso em função dos movimentos de F0 - para todos os efeitos o principal correlato acústico da dimensão perceptual *pitch* (Cf. Barry 1981).

É provável, que correlatos acústicos de longo termo possam ser empregados para expressar o contraste *spiky/glissando* sugerido por Crystal (1969), tais como as durações médias dos trechos alternados vozeado/não vozeado e/ou consonantal/vocálico (nos dois casos excluindo pausas); essas medidas não dependem diretamente da detecção de micro variações locais de F0 e podem ser obtidas com relativa facilidade através de Predição Linear (LPC).

Mais adiante examinaremos alguns parâmetros relacionados à noção genérica de RITMO. A seção 7.3 analisará diversas medidas que podem expressar a velocidade de fala, considerando o âmbito da variação em função das diferentes condições experimentais de velocidade (normal *versus* rápida) e a covariação imposta por diversos fatores (tamanho da palavra, unidade tonal, posição relativa no enunciado, etc). Na seção 7.4 analisaremos a variabilidade inter- e intra-falante de medidas baseadas na razão vozeado/não vozeado e em 7.5 as distribuições individuais a partir de níveis quantizados de amplitude.

7.3) *Velocidade de Fala*

Vários estudos têm examinado a velocidade de fala e outras medidas baseadas em aspectos temporais como um fator potencialmente importante para individualizar a fala de um determinado indivíduo (Reich *et al.* 1976; Doherty 1975; Clarke e Becker 1969; van Lancker *et al.* 1985).

A eficiência relatada de medidas de fluência para a identificação do falante varia bastante de um trabalho para outro, e os resultados geralmente não são diretamente comparáveis, em função da diversidade metodológica. A principal diferença reside na inclusão ou não inclusão das pausas no cômputo global da velocidade de fala; no primeiro caso, fala-se de "taxa de fala" (quase sempre denominada *speech rate*, ou a duração total do enunciado dividida pelo número de unidades de base: fonemas, sílabas ou palavras), e no segundo emprega-se habitualmente o termo "taxa de articulação" (*articulation rate* , ou o tempo de emissão de sons de fala- excluindo pausas - dividido pela unidade de base). Embora o padrão de pausas possa ser associado a características psicológicas idiossincráticas, e , portanto, constituir um indíce potencialmente importante para a identificação do falante (v. Scherer 1979:160 ff. para uma resenha de diversos trabalhos), esse é um aspecto de fluência fortemente influenciado pelo estilo discursivo (Goldman-Eisler 1968; Levin *et al.* 1982) e por alterações psicológicas transitórias, tais como ansiedade, medo, etc (Scherer 1979).

No presente trabalho definiremos velocidade de fala como "taxa de articulação", desprezando pausas respiratórias e de hesitação. A taxa de articulação tem, em relação à taxa de fala, a vantagem de possuir um âmbito de variação intra-falante certamente menor e ser menos facilmente manipulável pelo falante; como veremos abaixo, ao longo do exame dos dados, diferentes falantes parecem ter limites superiores de velocidade naturais, que podem ser marcas razoavelmente

eficientes de individualidade (um limite inferior seria difícil de definir, embora se possa supor que deva haver um limiar - eventualmente distinto para diferentes falantes - antes que a fala se torne "arrastada" e pouco natural).

A quantificação da velocidade de emissão tem como primeira dificuldade a determinação de uma unidade de medição coerente. A maior parte dos estudos utiliza a sílaba como unidade básica, embora o fonema e a palavra também sejam empregados com certa freqüência (Cf. Fant *et al.* 1991a; Goldman-Eisler 1968; McCroskey 1984). A sílaba parece ser uma unidade mais conveniente (Cf. Abercrombie 1967:96), embora não seja possível estabelecer um critério objetivo universalmente válido para definir o número de sílabas em uma palavra ou enunciado. Ladefoged (1975:218 ff) discute várias teorias da sílaba, concluindo que nenhuma delas é totalmente satisfatória; o fato é que não há propriedades acústicas ou fisiológicas marcando direta e inequivocamente cada sílaba. Apesar da dificuldade teórica, parece ser relativamente simples para o falante nativo intuir - na maior parte dos casos - o número de sílabas em uma palavra ou frase ; assim, assumiremos - como o faz Ladefoged (1971:81) -que

> *a neuro physiological definition is possible, even if we cannot at the moment state it in any way.*

Para a grande maioria das palavras do *corpus* aqui empregado (texto I; v. anexo) não houve dificuldade em determinar o número de sílabas em cada palavra segundo as regras de divisão silábica definidas no *Novo Dicionário da Língua Portuguesa* (Aurélio B. de Holanda Ferreira, 1ª edição, item XV, pgs. XI-XVII).

O emprego de uma divisão silábica baseada na ortografia é o procedimento mais freqüente nos estudos de velocidade de emissão. Talvez parecesse mais

razoável utilizar sílabas fonéticas, ou seja, sílabas efetivamente produzidas pelo falante. Não é certo, entretanto, que exista um critério totalmente objetivo para definir o que seria "sílaba fonética", especialmente pelo fato de - como já comentado acima - não existir uma teoria **fonética** satisfatória da sílaba (Cf. Ladefoged 1975:218ff) que permitisse uma divisão sem recurso a uma organização de nível mais alto (fonológica ou mesmo ortográfica). Na verdade, com exceção de eventuais fenômenos epentéticos, é razoável supor que a divisão ortográfica não se afaste consideravelmente da realização efetiva, especialmente no *corpus* aqui estudado, baseado em leitura com manutenção de inteligibilidade plena.

Há alguma evidência experimental indicando que a sílaba ortográfica é uma unidade adequada para a aferição da taxa de articulação; den Os (1985) verifica que medidas de taxa de articulação baseadas na "sílaba lingüística" (= sílaba ortográfica) correlacionam-se melhor com avaliações subjetivas da velocidade de emissão do que medidas expressas em "sílabas fonéticas" e "segmentos fonéticos". É possível que os resultados em den Os tenham relação com o fato de os ouvintes (assim como os falantes) terem a divisão silábica como um dos componentes de sua representação lexical. Nesse sentido é interessante observar que crianças não alfabetizadas e, portanto, sem conhecimento das regras ortográficas de divisão silábica (ou da própria noção de sílaba) conseguem decompor corretamente a maior parte dos vocábulos em unidades silábicas, tais como as reconhecemos tradicionalmente.

A utilização da sílaba ortográfica como unidade básica para aferir a taxa de velocidade de emissão torna-se menos problemática se é empregada uma amostra mais extensa de fala, já que eventuais distorções locais tendem a desaparecer no longo termo.

Como unidade temporal optou-se pelo segundo. Embora não haja diferença essencial em usar segundos ou minutos, uma unidade temporal mais curta, mais próxima da ordem de grandeza de uma palavra, parece mais conveniente e

intuitivamente mais acessível. Assim, a medida básica de velocidade de emissão nas análises que se seguem será o número de sílabas por segundo (sil/seg).

Além da definição da unidade segmental a ser empregada nas medidas de velocidade, é preciso definir também o domínio no qual será aferida cada uma das medidas. Esse aspecto é importante, pois já se verificou que a média e - mais pronunciadamente - o desvio-padrão da taxa de articulação apresentam valores diferentes em função do tamanho do enunciado utilizado para cada aferição da velocidade de emissão (Miller *et al.* 1984).

Na primeira fase de nossa análise as medidas foram tomadas com base na palavra, ou seja, as médias apresentadas são o somatório das medidas da taxa de articulação medida em em cada palavra isoladamente dividido pelo número de palavras medidas. Esse procedimento tem a vantagem de captar mais adequadamente as variações locais da taxa de articulação; essas variações são previsíveis na leitura (Fant *et al.* 1991a) e podem ocorrer mesmo no interior de unidades tonais e sentenças, em função, por exemplo, da posição da palavra na frase (Lehiste 1972; Klatt 1975; Berkovits 1991) e da estrutura tema/rema (Lieberman 1963; Hunnicutt 1985; Fowler e Housum 1987). Mais adiante serão analisados os resultados obtidos a partir de medidas baseadas em trechos ininterruptos de fala, separados por pausas.

A tabela 7.1 mostra as médias para cada falante da velocidade de emissão expressa em sílabas por segundo, para cada uma das condições de velocidade de emissão, além dos valores de F e respectivos níveis de significância correspondendo a uma análise de variância considerando apenas o efeito FALANTE isoladamente. Como na velocidade rápida só foram examinadas as palavras com mais de quatro sílabas, foram também incluídos na tabela 7.1 os resultados na velocidade normal para palavras de quatro ou mais sílabas, de modo a permitir uma comparação mais direta entre as duas condições de velocidade.. Foram desprezados todos os casos

onde não foi possível determinar com precisão a duração exata de uma palavra, em decorrência de efeitos de articulação truncada ou de coarticulação em fronteira de palavras; assim, de modo a evitar desbalanceamento na comparação inter-falante, foram computados apenas casos com observações para no mínimo sete falantes.

As médias por falante na tabela 7.1 foram obtidas a partir de medidas de velocidade extraídas de cada palavra isoladamente [1]; assim, cada média *m* é

$$\mathbf{m} = [\sum (d_i \div ns_i)] \div N$$

(onde d_i é a duração em segundos de cada palavra *i*, ns_i é o número de sílabas dessa palavra, e *N* é o número total de palavras -cerca de 90 na velocidade normal, e 20 na velocidade rápida).

Velocidade Normal					F=5.03 p<.0001				
ZR	**EN**	**R1**	**ZP**	**AG**	**WA**	**MS**	**R2**	**DO**	**Tot.**
6.4	6.2	5.8	6.2	7.5	6.7	6.6	5.9	6.5	6.4
1.9	1.7	1.9	1.9	2.9	2.4	2.1	2.0	1.8	2.1
97	88	94	95	87	92	93	97	97	843
Velocidade Normal (só pal. > = 4 sil.)					F=4.80 p<.0001				
ZR	**EN**	**R1**	**ZP**	**AG**	**WA**	**MS**	**R2**	**DO**	**Tot.**
7.5	6.8	6.2	6.7	7.6	6.9	7.0	6.2	7.3	6.9
1.9	1.6	1.3	1.6	1.3	1.5	1.5	1.5	1.9	1.6
52	46	51	45	40	51	48	50	44	427
Velocidade Rápida					F=4.02 p<.0003				
ZR	**EN**	**R1**	**ZP**	**AG**	**WA**	**MS**	**R2**	**DO**	**Tot.**
9.2	7.5	7.1	7.5	8.2	7.6	7.8	7.1	8.8	7.9
1.8	1.7	1.4	1.4	1.4	1.6	1.6	1.5	1.8	1.7
20	18	20	19	18	20	18	20	17	170

TABELA 7.1: Médias (sil/seg), desvios-padrão e número de ocorrências de cada falante em cada condição de velocidade. Os valores de F e respectivos níveis de significância indicam que o fator FALANTE é significativo.

Os resultados da análise de variância expressos na tabela 7.1 indicam que a categoria FALANTE, isoladamente, é estatisticamente significativa, embora os valores não muito altos de F sugiram um efeito fraco.

Observa-se que a exclusão de palavras com menos de quatro sílabas (velocidade normal) leva invariavelmente a um aumento na média de velocidade de emissão, embora não com a mesma magnitude para todos os falantes. Esse resultado indica que o número de sílabas da palavra pode ser um fator importante na velocidade de emissão. Vários estudos já verificaram uma relação inversamente proporcional entre o número de sílabas da palavra e a velocidade de emissão dessa palavra (Crystal e House 1982; Lehiste 1972; Schwartz 1972; Fant *et al.* 1991b); esse fenômeno tem sido geralmente associado a ajustes rítmicos nas línguas ditas acentuais (Cf. Fant *et al.* 1991b), mas há indícios de que outros fatores, tais como a previsibilidade da palavra (Lieberman 1963; Hunnicutt 1985) a estrutura tema/rema (Fowler e Housum 1987), e a posição da palavra no enunciado (Crystal e House 1988a,b; Cooper e Danly 1981; Fletcher 1991; Shen 1992) possam também influir na velocidade de emissão local . Mais adiante examinaremos mais de perto os efeitos do número de sílabas da palavra e do grupo tonal na velocidade de emissão local, assim como alguns efeitos sintáticos e semânticos.

Observa-se na tabela 7.1 que a diferença entre médias de dois falantes na mesma condição pode ser menor do que a diferença entre as médias do mesmo falante; a maior diferença na velocidade normal (palavras com quatro ou mais sílabas), por exemplo, é 1.4 sil/seg (falantes ZR e R1/R2), enquanto o falante ZR aumenta 1.7 sil/seg na velocidade rápida. O aumento da taxa de emissão na condição velocidade rápida, entretanto, não é uniforme; os aumentos percentuais variam de 22.6 % (falante ZR) a 7.9 % (falante AG).

É interessante observar que, apesar dos diferentes aumentos percentuais, a ordenação dos falantes em função da taxa de emissão silábica é praticamente a mesma nas duas condições (com a única exceção de AG, que é o mais rápido na velocidade normal e o terceiro mais rápido na velocidade rápida). Observe-se também os valores idênticos nas duas produções não-contemporâneas do falante R1/R2 (nas duas condições de velocidade).

A coincidência nas taxas de emissão de R1/R2 e a manutenção da ordem da velocidade dos falantes nas duas condições sugere que as duas tarefas ("leitura normal fluente em ritmo confortável" e "leitura maximamente rápida sem perda de inteligibilidade") refletem alguma espécie de limite ou referência natural das programações motoras individuais para a tarefa genérica de leitura. No caso da velocidade maximamente rápida, pode estar em jogo aquilo que Klatt (1973) denomina *princípio de incompressibilidade*; segundo esse princípio, os segmentos fonéticos resistiriam à compressão além de um determinado limite, cartacterístico de cada tipo de segmento.

A hipótese de um limiar superior de velocidade de emissão é intuitivamente mais aceitável; mas seria razoável supor um limite inferior? É claro que, dentro dos limites da capacidade respiratória, seria possível alongar consideravelmente alguns segmentos da fala (particularmente vogais e fricativas); mas até que ponto isso poderia ser realizado sem que houvesse ruptura na estrutura prosódica e, conseqüentemente, na naturalidade e inteligibilidade?

Nesse sentido são relevantes os resultados obtidos em alguns experimentos focalizando os efeitos duracionais em função de diferentes condições de velocidade. Em Fant *et al.* (1991b), além das condições velocidade "normal" e "rápida", solicitou-se ao falante que lesse o mesmo texto em velocidade "lenta"; os resultados indicaram que na leitura "lenta" o aumento da duração total está mais fortemente relacionado ao aumento de duração das pausas entre e dentro de sentenças do que a

uma expansão significativa na duração média do fonema (75 ms na velocidade normal contra 78 ms na lenta). ou seja, a taxa de articulação propriamente dita altera-se pouco de uma condição para a outra. Esses resultados são consistentes com os relatados em Goldman-Eisler (1968), Lane e Grosjean (1973) e Grosjean (1979). Também a esse respeito é pertinente a observação de Cooper *et al.* (1985) de que a marca de foco semântico nas palavras em posição final de sentença acarreta em um aumento duracional de apenas 15%, contra 40% nas posições inicial e medial; segundo os autores, essa diferença está relacionada ao fato de as palavras em posição final já estarem submetidas a um processo de alongamento. Cooper *et al.* (1985) propõem um princípio geral de restrição à expansibilidade de um segmento (*expandibility constraint*), que impediria um segmento de ser alongado além de um determinado limite. Tal restrição funcionaria de modo análogo ao já comentado *princípio de incompressibilidade* proposto em Klatt (1973), mas na direção inversa.

É como se o falante tivesse dificuldade em reduzir a taxa de articulação além de um limiar naturalmente confortável para ele; nesse sentido é interessante a observação de Lenneberg (1967), de que a taxa "normal" de emissão silábica (≅ 6 sil/seg) pode estar associada a freqüências da mesma ordem de grandeza produzidas nas regiões do cérebro envolvidas com a fala. Uma questão importante aqui seria estudar uma possível relação entre a taxa de articulação "normal" e padrões temporais de natureza neurológica para cada indivíduo. Embora o escopo do presente estudo permita apenas um tratamento meramente especulativo dessa questão, a exata coincidência nas taxas articulatórias não contemporâneas do falante R1/R2 (espaçadas em mais de três meses) sugere que pode estar em jogo algum tipo de controle não periférico de grande estabilidade, tal como proposto por Lenneberg (1967).

7.3.1) *Efeito do Número de Sílabas da Palavra na Velocidade de Emissão*

Já observamos acima que o número de sílabas da palavra parece ter alguma influência sobre a velocidade de emissão local. Antes de examinarmos mais de perto a magnitude desse efeito, é necessário verificar os efeitos isolados e em interação em um modelo que inclua também a variável FALANTE. A tabela 7.2 apresenta, para as duas condições de velocidade de produção, os efeitos de FALANTE, NUMSIL (número de sílabas da palavra sobre a qual uma determinada aferição foi feita) e a interação FALANTE X NUMSIL.

Velocidade Normal			
efeito ↓	**F**	**p<**	**G.L.**
FALANTE	.06	.73	8
NUMSIL	11.5	.0001	6
FAL. X NUMSIL	.53	.9961	48
Velocidade Rápida			
efeito ↓	**F**	**p<**	**G.L.**
FALANTE	1.41	.195	8
NUMSIL	25.93	.0001	3
FAL. X NUMSIL	.26	.9999	24

TABELA 7.2: Resultados de ANOVA (BMDP-7D)

Podemos observar, na tabela 7.2, que, no modelo incluindo FALANTE e NUMSIL, não há influência significativa de FALANTE para nenhuma das duas condições de velocidade de emissão (velocidade normal e rápida) sobre a velocidade de emissão efetivamente medida. Não se observa também um efeito de interação significativo (FALANTE X NUMSIL), indicando que é razoável reunir os dados de todos os falantes para estudar o efeito isolado do Número de Sílabas na palavra (NUMSIL). Os resultados na tabela 7.2 aparentemente contradizem o observado

anteriormente na tabela 7.1, onde verificamos a existência de um efeito isolado significativo de FALANTE. A diferença aqui deve-se ao fato de termos incluído a variável NUMSIL no modelo; o fato é que o efeito de NUMSIL praticamente anula o efeito de FALANTE, sugerindo que as medidas de velocidade individuais não devem ser consideradas indiscriminadamente, mas sempre em função de NUMSIL.

A tabela 7.3 mostra as médias (em sílabas por segundo) e desvios- padrão separadamente para cada um dos níveis de NUMSIL, nas duas condições de velocidade de emissão (normal e rápida), assim como os valores de F para o efeito isolado de NUMSIL. Lembramos que na condição Velocidade Rápida só foram medidas as palavras com mais de quatro sílabas.

Velocidade Normal								
NUMSIL →	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**TOT.**
Média	7.1	6.1	6.4	5.8	7.5	7.7	7.3	6.4
D.P.	2.9	2.1	1.1	1.1	1.2	.42	.60	2.1
n=	196	264	161	168	36	9	9	843
Velocidade Rápida								
NUMSIL →	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**TOT.**
Média	-	-	-	7.2	8.7	10.7	8.9	7.7
D.P.	-	-	-	1.5	1.4	.81	.55	1.7
n=	-	-	-	135	33	9	9	186

TABELA 7.3: Médias e desvios-padrão para diferentes níveis de NUMSIL (número de sílabas na palavra), nas duas condições de velocidade. Na velocidade rápida sóforam medidas as palavras de 4 ou mais sílabas.

O exame da tabela 7.3 mostra que há uma tendência a uma taxa de emissão mais rápida em palavras com maior número de sílabas. Na velocidade normal, observa-se também que os monossílabos tendem a ser produzidos mais rapidamente que palavras de duas, três e quatro sílabas. De modo a verificar a significância estatística dessas diferenças, foram realizados testes *post-hoc* (teste *Student-*

Newman-Keuls; BMDP-7D) comparando os níveis de NUMSIL segundo diferentes agrupamentos desses níveis; as distinções assim obtidas, que apresentaram os valores máximos de F estão na tabela 7.4.

Velocidade Normal				
GRUPOS (níveis reunidos de NUMSIL)	**Média**	**D.P.**	**F=**	**p<**
[2,3,4]	6.31	1.1		
X			21.8	.0001
[1,5,6,7]	7.58	1.7		
Velocidade Rápida				
GRUPOS (níveis reunidos de NUMSIL)	**Média**	**D.P.**	**F=**	**p<**
[4]	7.2	1.5		
X			55.8	.0001
[5,6,7]	9.1	1.4		

TABELA 7.4: Resultados de teste de comparação de médias (*Student-Newman-Keuls*; BMDP-7D) reunindo em cada grupo vários níveis de NUMSIL. Todas as combinações possíveis foram testadas; as distinções apresentadas na tabela foram aquelas com maior valor de F

A tabela 7.4 resume resultados de diversos testes de comparação de médias, incluindo em cada grupo todas as possíveis combinações de níveis de NUMSIL, apresentando apenas as combinações mais significativas. Assim, na condição velocidade normal, as palavras de duas, três e quatro sílabas são significativamente diferentes das de uma, cinco, seis e sete sílabas, quanto à média de velocidade (medida em cada palavra separadamente). Na condição velocidade rápida, palavras de quatro sílabas são significativamente diferentes de palavras com cinco ou mais sílabas.

Vários estudos já verificaram uma relação sistemática entre a velocidade de emissão e o tamanho do enunciado. Lehiste (1972) observa que a duração do radical é consideravelmente reduzida se um sufixo derivacional é acrescentado; assim, em

palavras como *steady, skiddy e skitty*, a seção correspondente ao radical (*stead, skid, skit*) pode ser tão reduzida que a duração total, mesmo com o sufixo, não é muito mais longa do que na produção dos radicais isolados. Crystal e House (1982; 1988a; 1988b), tratando uma extensa base de dados do Inglês Americano, verificam uma tendência a uma relação inversamente proporcional entre as durações segmentais e o número de sílabas da palavra onde se encontra o segmento.

Embora se possa, de forma genérica, estabelecer uma relação entre o número de sílabas da palavra e as durações silábicas dessa palavra, é provável que existam fatores atuando em um nível de estruturação rítmico-temporal superior ao domínio da palavra. Fant *et al.* (1991a; 1991b) verificam que há um padrão alternado de aceleração/desaceleração da velocidade de emissão ocorrendo no domínio da sentença, em função da densidade acentual, isto é, um pequeno números de fonemas por Pé Métrico, condiciona um aumento da duração média do fonema no interior desse Pé Métrico, diminuindo assim a taxa de articulação. A redução proporcional da duração dos monossílabos observada no *corpus* aqui estudado (velocidade normal) deve-se, provavelmente, a restrições rítmicas dessa natureza, em função da posição fraca onde ocorrem esses monossílabos, em sua grande maioria itens funcionais (artigos, preposições, conjunções, etc.); assim, parece haver uma tendência a um quase apagamento dessas partículas funcionais, equivalente à redução duracional observada normalmente em sílabas não tônicas no interior de palavras polissílábicas. Outro aspecto que pode estar em jogo é o fato de esses monossílabos funcionais serem semanticamente pouco informativos, já que há evidências que, independentemente do número de sílabas da palavra, há uma tendência genérica a reduzir a duração de palavras altamente previsíveis (Lieberman 1963; Hunnicutt 1985; Fowler e Housum 1987; Eefting 1992).

É interessante observar, no entanto, que a redução duracional dos monossílabos não é um fenômeno consistente entre os diversos falantes analisados

no presente trabalho, como se pode observar, em parte, pelo desvio-padrão mais alto nas medidas de velocidade dos monossílabos (v. tabela 7.3, velocidade normal). Examinando mais de perto os resultados de cada falante separadamente verificamos que a redução duracional dos monossílabos é muito mais pronunciada em alguns falantes do que em outros. A figura 7.1 ilustra esse aspecto, mostrando que os falantes MS, ZR e, especialmente, o falante DO tendem a reduzir mais a duração dos monossílabos, enquanto os falantes ZP, EN e WA reduzem pouco (em comparação com os demais tipos de palavra). Esse comportamento diferenciado inter-falante quanto à redução duracional dos monossílabos sugere que possa estar em jogo uma característica pessoal potencialmente interessante para a diferenciação de falantes no plano rítmico-temporal.

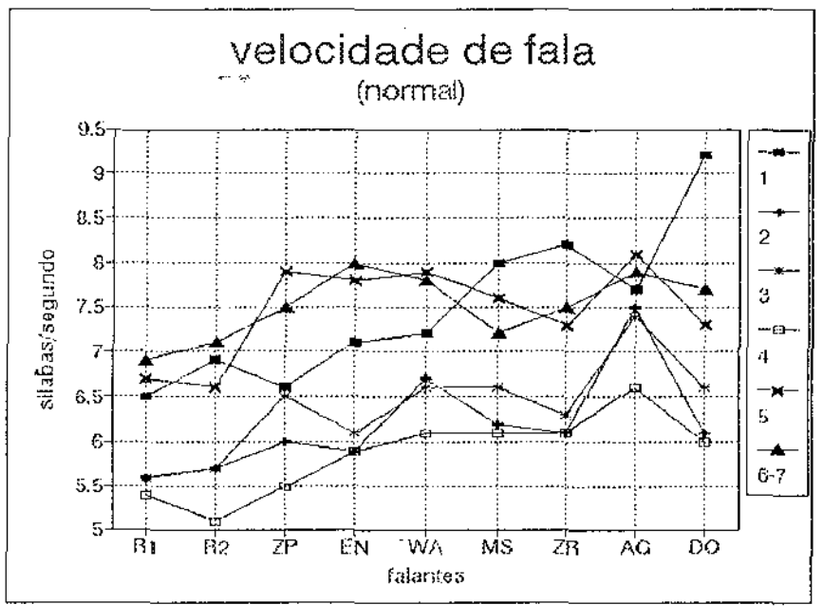


FIGURA 7.1: Médias (sil/seg) para cada falante, em função do número de sílabas da palavra. Observar como os falantes MS, ZR e DO encurtam relativamente mais os monossílabos

7.3.2) *Efeito do Número de Sílabas em Blocos de Palavras na Velocidade de Emissão*

Verificamos em 7.3.1 que o número de sílabas de uma palavra influi significativamente na velocidade de emissão local. Discutimos também a possibilidade de existir uma estruturação temporal de nível superior à palavra, que possa ter também influência sobre a velocidade de fala. Há evidências indicando uma relação dessa natureza; Lehiste (1972) verifica experimentalmente que a duração de um dado segmento tende a diminuir à medida que aumenta o tamanho do enunciado no qual esse segmento ocorre. Schwartz (1972) interpreta esse comportamento como evidência de um mecanismo de *exploração* (*scanning*) realizado pelo falante antes de produzir efetivamente o enunciado; assim, segundo Schwartz (1972), o falante avalia com antecedência o comprimento do enunciado e usa essa informação para determinar a quantidade de tempo que pode dispender na articulação de sons individuais.

A questão que se coloca aqui é definir um domínio coerente, mais amplo que a palavra, ao qual se possa referir as medidas locais de velocidade. A fala corrente é normalmente estruturada em blocos descontínuos, separados por pausas. Algumas dessas pausas são condicionadas pelas limitações inerentes do sistema respiratório. Nesse sentido é bem conhecida a noção de *breath-group* de Lieberman; assim, o *breath-group* seria

> *a condition of minimum departure from the constraints of vegetative breathing* (Lieberman 1977:170).

Ainda segundo Lieberman (1977:169), a organização prosódica em termos de *breath-groups* reflete uma organização sintática em sentenças (ou pelo menos tende a ser coincidente com essa organização sintática).

Há, no entanto, unidades menores que o *breath-group*, também separadas por pausas e não necessariamente ligadas à delimitação de sentenças. Mattoso Câmara Jr. (1980:79) define três causas principais para as interrupções da fonação:

> *1) satisfazer as exigências da respiração; 2) facilitar a elaboração mental e a compreensão; 3) realizar uma impressão rítmica, mais ou menos determinada para cada língua*

Assim, as pausas e interrupções da cadeia de fala não se devem apenas a restrições fisiológicas. Mattoso Câmara chama de *Grupo Melódico* a cada uma das unidades entoacionais que constituam um *contínuo sonoro mínimo*. A essas unidades seria mais apropriado atribuir uma motivação lógica (antes que fisiológica) dentro da estruturação do discurso falado.

Examinando as produções dos falantes aqui estudados, observamos que as pausas "lógicas" ocorrem invariavelmente nas mesmas posições para todos os falantes (no anexo, estão assinaladas, no texto I, essas posições). Os trechos assim delimitados são equivalentes ao que Mattoso Câmara (1980) chama de *Grupo Melódico* (v. também as noções equivalentes de *Unidade tonal* em Mateus *et al.* 1982:518 e de *Groupe Phonétique* em Dubois *et al.* 1973:241). Assim, adotamos essas divisões "naturais" da cadeia de fala como base para a determinação de uma unidade mais ampla que a palavra, à qual se pudesse relacionar, de alguma forma, a variação local de velocidade de emissão; a cada um desses grupos isolados por pausas "lógicas" chamaremos de "bloco".

A figura 7.2 mostra, no eixo vertical, as velocidades de emissão, para as duas condições de velocidade de produção (normal e rápida) em função do tamanho do enunciado, expresso em número de sílabas.

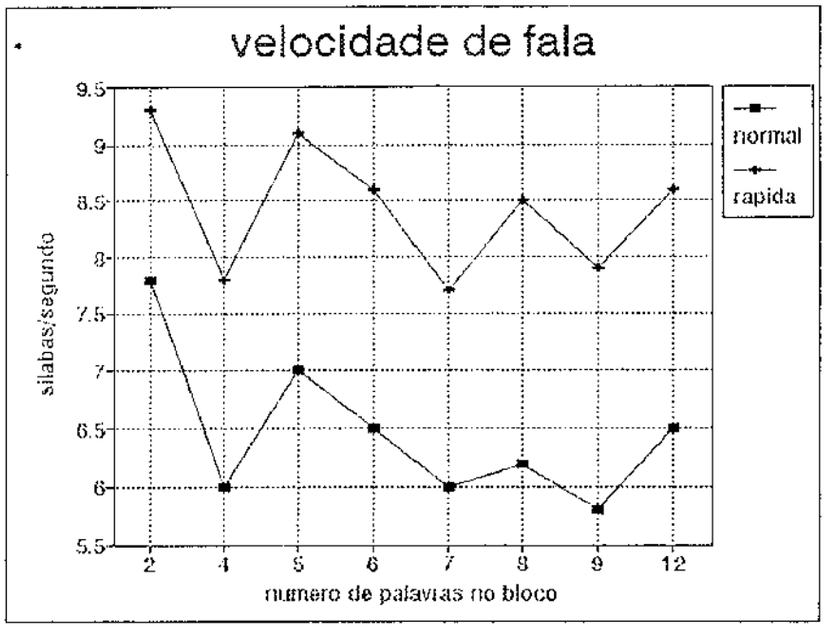


FIGURA 7.2: Velocidade média de emissão em função do número de sílabas no bloco

Podemos observar, na figura 7.2, que não existe, aparentemente, qualquer padrão que pudesse associar o tamanho do enunciado à velocidade de emissão média dentro desse enunciado, o que contradiz os resultados obtidos em Lehiste (1972), já comentados acima. É provável que o efeito de compressão das durações, à medida que aumenta o tamanho do enunciado seja apenas observado no tipo de paradigma experimental usado em Lehiste (1972); nesse estudo, o contexto sintático de base é mantido e são acrescentados mais itens, ou itens mais extensos, de forma a permitir que os diferentes enunciados sejam passíveis de uma comparação mais direta (*The*

stick fell, The stick is broken, The stick was discarded, por exemplo). Esse tipo de comparação é impossível no contexto experimental do presente estudo, já que cada um dos blocos de fala é diferente dos demais quanto à estrutura sintática. De qualquer modo, parece claro que o tamanho do enunciado (definido aqui como um "bloco" de fala) não afeta diretamente a velocidade de emissão média no trecho. Na verdade, alguns estudos apresentam resultados opostos aos de Lehiste (1972), verificando uma diminuição na taxa de articulação média, à medida que aumenta o tamanho do enunciado (Goldman-Eisler 1954; Sternberg *et al.* 1988).

Por outro lado, já verificamos acima que o tamanho da palavra (expresso em número de sílabas) afeta fortemente a velocidade de emissão local; assim, se existe realmente o mecanismo de *scanning* proposto pro Schwartz (1972), seu domínio deve ser a palavra, e não o enunciado (ou o *bloco de fala*, como o denominamos aqui). Variações locais de velocidade de emissão podem estar também relacionadas a ajustes rítmicos no domínio de intervalos inter-acentuais - isto é, dos "pés métricos" - , mas esse ajuste pode ser referido, da mesma forma, ao tamanho (em sílabas) da palavra.

Nas medidas de velocidade realizadas até agora usamos como base cada uma das palavras, isto é, as médias obtidas são o somatório de todas as medidas isoladas dividido pelo número total de palavras. A média assim obtida não será necessariamente igual à média obtida a partir de trechos mais longos de fala. Miller *et al.* (1984) observam que o valor médio da velocidade de emissão depende do tamanho dos blocos de fala usados como base para as medidas; Miller *et al.* (1984) comparam medidas obtidas a partir de blocos de cerca de 12 sílabas com medidas a partir de blocos de 30 sílabas, para o mesmo material de fala, e constatam que as médias finais obtidas nos dois procedimentos são consideravelmente diferentes.

A figura 7.3 apresenta, para a velocidade normal de produção, as médias de cada falante, a partir de três diferentes métodos: (1) medidas tomadas em cada palavra separadamente, (2) medidas tomadas em cada bloco de fala separadamente e (3) uma medida única englobando todo o enunciado. Podemos verificar, através do exame da figura 7.3, que as médias individuais obtidas por cada um dos métodos não são iguais; as médias obtidas a partir de medidas palavra-a-palavra difere consideravelmente das médias bloco-a-bloco, enquanto estas são praticamente iguais à média global (considerando o texto lido inteiro como um todo) [2]. É interessante observar que a ordem relativa dos falantes altera-se consideravelmente dependendo do método utilizado (ver por exemplo o falante DO). Observa-se também que a diferença entre as médias é maior para alguns falantes; é possível que esses falantes tenham uma maior flutuação na velocidade de emissão local, fazendo com que se torne mais nítida a distinção entre os dois métodos de medição. Uma diferença considerável nas médias obtidas pelos dois métodos (palavra-a-palavra *versus* bloco-a-bloco) poderia estar refletindo alguma tendência a um ritmo de base mais "acentual" para esse falante, já que um ritmo silábico "puro" igualaria necessariamente as médias dos dois métodos.

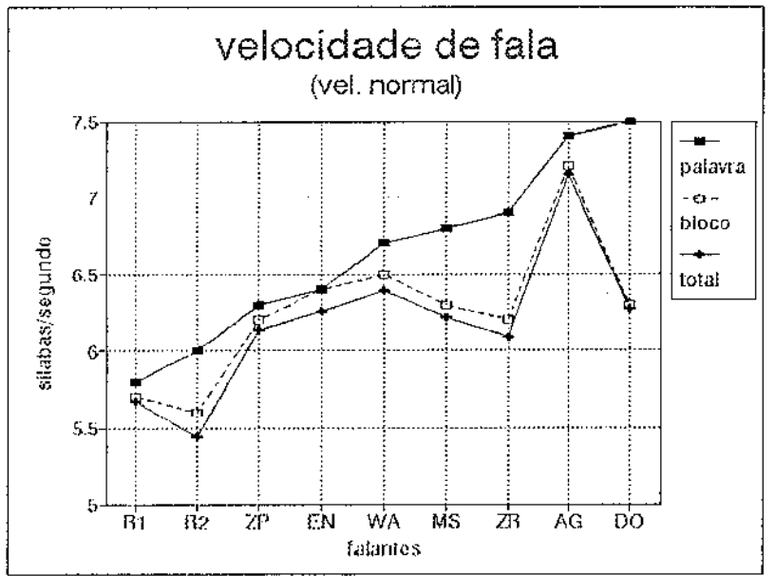


FIGURA 7.3: Médias de velocidade por falante por 3 diferentes métodos de medição

7.3.3) *Efeito da Posição da Palavra no Enunciado na Velocidade de Emissão*

Um outro fator que pode afetar a velocidade local de emissão é a posição relativa da palavra no interior do enunciado. É sabido que finais de enunciados sofrem geralmente uma desaceleração (Crystal e House 1988a, b; Klatt 1975, 1976; Lehiste 1972). O fenômeno aponta para uma relação sistemática entre prosódia e sintaxe, sendo geralmente interpretado como uma marca de fronteira sintática (Price *et al.* 1991; Klatt 1975; Shen 1992), embora possa ter também um componente relacionado à desaceleração esperada em sequências motoras de qualquer natureza (Klatt 1976). Lyberg (1979) sugere que o alongamento de sílabas em final de enunciado pode ser um efeito secundário do controle da freqüência fundamental.

O alongamento em final de enunciado parece ser um fenômeno universal, embora possam existir diferenças de magnitude do efeito em diferentes línguas (Berkovits 1991). O material fonético envolvido também pode influir na magnitude do alongamento: fricativas finais são em geral mais alongadas que outros fonemas (Cooper e Danly 1981; Oller 1973) e, entre as vogais, /i/ sofre um maior alongamento que /a/ (Fletcher 1991).

Há alguma controvérsia quanto à extensão do efeito do alongamento em final de enunciado; alguns resultados indicam um alongamento apenas na palavra ou na sílaba final (Oller 1973), enquanto outros relatam um efeito que se estende ao longo de várias sílabas (Lehiste 1972; Klatt 1976). Kohler (1983) observa que essa diferença pode estar relacionada a estratégias individuais, com alguns falantes desacelerando mais cedo que outros, uma possibilidade potencialmente interessante para a Identificação de Falantes.

De modo a verificar o efeito do alongamento em final de enunciado em nossos dados, classificamos as palavras do *corpus* em duas categorias: (1) palavras em final de bloco (isto é, precedendo imediatamente pausas) e (2) as demais palavras. A tabela 7.5 mostra os resultados de uma análise de variância considerando FALANTE, POSIÇÃO e o efeito de interação.

efeito	G.L.	F =	p <
FALANTE	8	2.68	.0005
POSIÇÃO	1	90.87	.0001
FAL X POS.	8	.37	NS

TABELA 7.5: Resultados de ANOVA (BMDP-7D) testando os efeitops de FALANTE e POSIÇÃO (final *versus* não-final)

Podemos observar que não há qualquer efeito de interação, confirmando que o fenômeno ocorre de modo semelhante para todos os falantes do grupo estudado. Observa-se também , pelo alto valor de F para a variável POSIÇÃO, que a variabilidade devida a essa condição é muito maior do que a atribuída aos diferentes falantes.

A figura 7.4 mostra as médias de velocidade, para cada falante, em cada um dos níveis de POSIÇÃO. Podemos observar que existe, efetivamente, uma redução na velocidade de emissão nas palavras em fronteira de pausas; o efeito é bastante homogêneo no grupo de falantes estudado, com praticamente o mesmo grau de desaceleração para todos os falantes, independentemente da velocidade de produção (normal ou rápida).

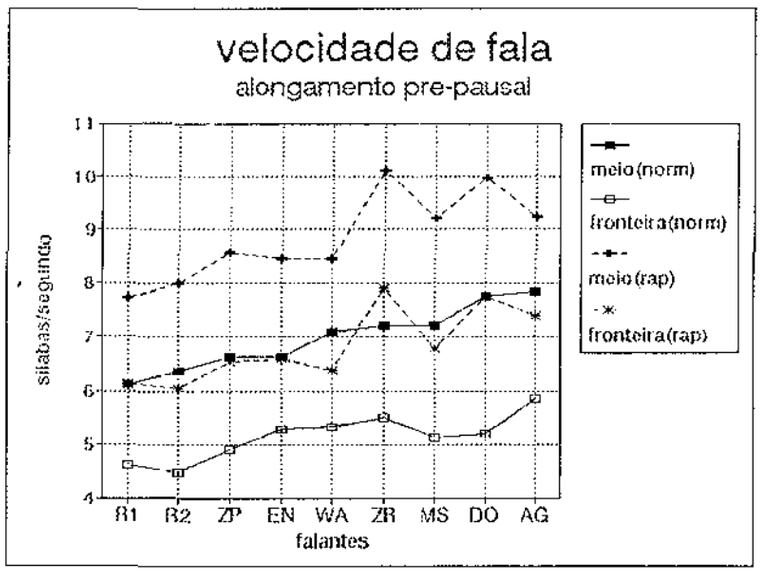


FIGURA 7.4: Médias de velocidade por falante, nas duas condições de velocidade, em função da POSIÇÃO da palavra no enunciado (final *versus não-final)*

Na classificação dentro da variável POSIÇÃO usamos até aqui apenas dois níveis: palavra em posição final de enunciado (bloco) *versus* todas as demais palavras nesse bloco. Alguns estudos, entretanto, indicam que há um efeito de desaceleração que se estende ao longo de várias sílabas. Lehiste (1972) observa que uma determinada palavra se torna mais curta à medida em que sua posição relativa afasta-se mais do final do enunciado; assim, a palavra *stick*, nos contextos *The stick* (a) *fell;* (b) *is broken;* (c) *was discarded*, tem as durações respectivas de (a) 311, (b) 283 e (c) 268 milisegundos (v. também Klatt 1976).

De modo a verificar essa hipótese, dividimos a variável POSIÇÃO em três níveis, de acordo com a posição relativa da palavra no bloco de fala; assim, cada bloco foi dividido aproximadamente em três partes (que chamamos: *início, meio e fim*). A figura 7.5 mostra os intervalos de confiança (95 %) para cada um dos níveis de POSIÇÃO para o conjunto total de dados. Podemos verificar que é evidente o efeito observado em Lehiste (1972); de fato, parece existir uma desaceleração em direção ao final do enunciado, e que se estende por várias sílabas (ou palavras).

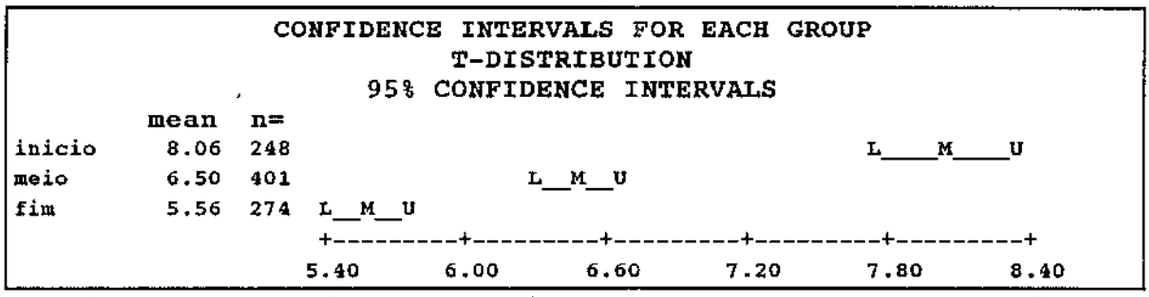


FIGURA 7.5: Variação entre as médias de velocidade de emissão em função da posição da palavra no enunciado (início, meio e fim)

A tabela 7.6 e a figura 7.6 mostram as médias individuais de todos os falantes (velocidade normal de produção) para as três posições relativas dentro do bloco. A análise de variância com esses dados apresentou os seguintes resultados: FALANTE

(F= 5.44; p<.0001; G.L.= 8), POSIÇÃO NO ENUNCIADO (F= 82.09; p< .0001; G.L.= 2), FALANTE X POSIÇÃO (F= 1.01; p<.4399; G.L.= 16). Novamente podemos verificar que a variância relacionada a POSIÇÃO é bem maior do que a FALANTE e não há efeito significativo de interação FALANTE X POSIÇÃO.

Apesar de não haver interação significativa FALANTE X POSIÇÃO, podemos observar na figura 7.6 que a desaceleração não tem a mesma proporcionalidade para todos os falantes; o falante DO, por exemplo, parece concentrar o efeito mais para o fim do enunciado do que os demais falantes. Essas diferenças individuais já foram observadas em outra parte (Kohler 1983) e são potencialmente interessantes para a Identificação de Falantes. É possível que uma melhor exploração desse tipo de tendência rítmica exija um maior detalhamento da variação da velocidade de emissão ao longo do enunciado do que uma divisão em apenas três partes, como a que foi feita aqui; uma possibilidade seria definir perfis de velocidade nos blocos de fala e procurar um padrão individual nessas curvas.

	ZR	EN	R1	ZP	AG	WA	MS	R2	DO
início	8.1	7.5	7.2	7.2	9.1	8.5	8.6	7.8	8.3
	2.6	1.7	2.2	2.2	4.2	3.1	2.8	2.6	2.7
	31	24	26	26	25	26	29	29	32
meio	6.8	6.3	5.6	6.2	7.1	6.3	6.5	5.6	7.7
	2.6	2.6	1.4	1.8	2.0	1.8	1.6	1.2	3.9
	55	39	43	42	37	42	46	44	53
fim	5.9	5.6	5.0	5.5	6.4	5.8	5.6	4.9	5.6
	2.1	1.1	1.0	1.4	1.4	1.3	1.2	1.0	1.3
	33	32	31	29	27	28	30	31	33

TABELA 7.6: Médias individuais (velocidade normal) em função da posição da palavra no bloco

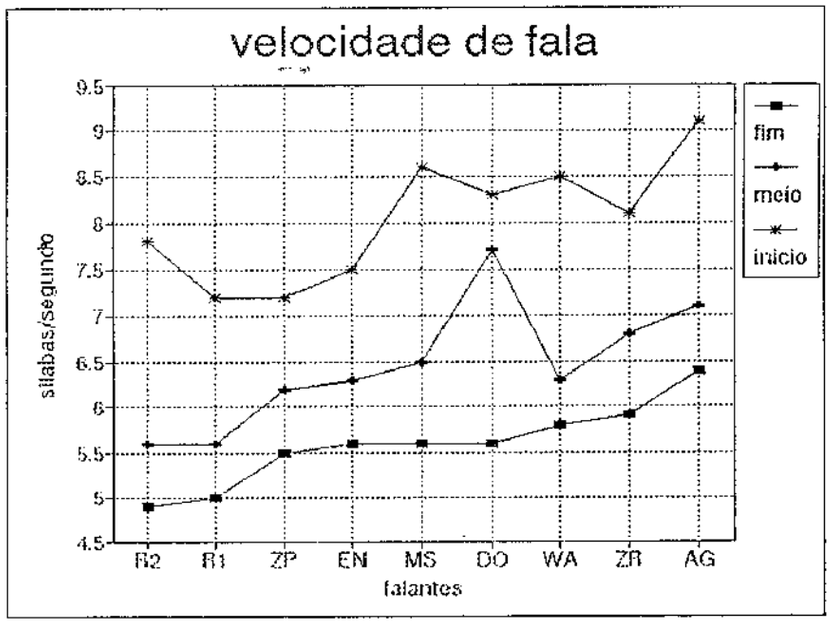


FIGURA 7.6: Médias individuais em função da posição da palavra no bloco

Os experimentos sobre os efeitos do alongamento em final de enunciado baseiam-se geralmente em contextos controlados onde cada fator varia isoladamente (posição da palavra, número de palavras, número de sílabas das palavras, etc), fazendo com que os resultados sejam mais facilmente interpretados. No tipo de material aqui utilizado, cada bloco é, a princípio, diferente dos demais quanto à estrutura gramatical, conteúdo, etc. É possível, portanto, que algumas tendências sejam mascaradas pela covariância de outros fatores. Já verificamos acima que o número de sílabas na palavra influi fortemente na velocidade local, refletindo, provavelmente, algum tipo de ajuste métrico inter-acentual. De modo a verificar a influência do tamanho da palavra, em função da posição da palavra no bloco, realizamos uma análise de variância incluindo a variável NUMSIL (número de sílabas na palavra sobre a qual se mediu um valor de velocidade de emissão). Os resultados dessa análise são: NUMSIL (F= 8.65; p<.0001; G.L.= 4), POSIÇÃO (F=

18.38; p<.0001; G.L.= 2), NUMSIL X POSIÇÃO (F= 11.23; p<.0001; G.L.= 8). Observamos que POSIÇÃO ainda é o fator mais importante, mas com um efeito também significativo da interação NUMSIL-X POSIÇÃO, indicando que o efeito de desaceleração não é homogêneo para todos os tipos de palavra. A figura 7.7 mostra as médias para palavras de uma a quatro sílabas, velocidade normal de produção, em função da posição no bloco. Podemos observar que as palavras com menor número de sílabas variam mais quanto à velocidade de emissão do que as palavras mais longas, em função da posição no bloco; a influência do número de sílabas tende a desaparecer nas posições mais ao final do bloco de fala, sugerindo que há uma alteração do padrão rítmico de base, com uma tendência a um padrão mais "silábico" no fim do bloco de fala.

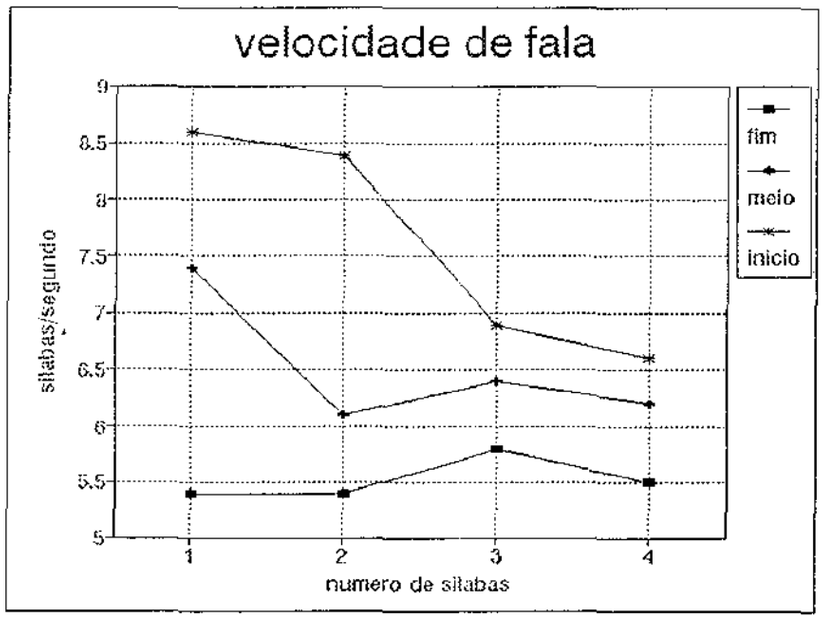


FIGURA 7.7: Médias de velocidade de emissão em função do número de sílabas e da posição da palavra no enunciado.

Ao examinarmos as velocidades médias em cada bloco de fala observamos que havia uma tendência a valores um pouco menores nos blocos finais do enunciado total (texto I; ver anexo). A figura 7.8 mostra as médias de cada bloco, em cada uma das condições de velocidade de produção (normal e rápida) com as respectivas retas de regressão ajustadas a todos os valores medidos, para todos os falantes. Observa-se que o efeito de desaceleração é estatisticamente significativo (os coeficientes das retas são: r= -.348, p<.001 para a velocidade normal e r= -.195, p<.05 para a velocidade rápida). Aparentemente, a desaceleração observada isoladamente em cada bloco reflete-se também na dimensão "macro" do enunciado total.

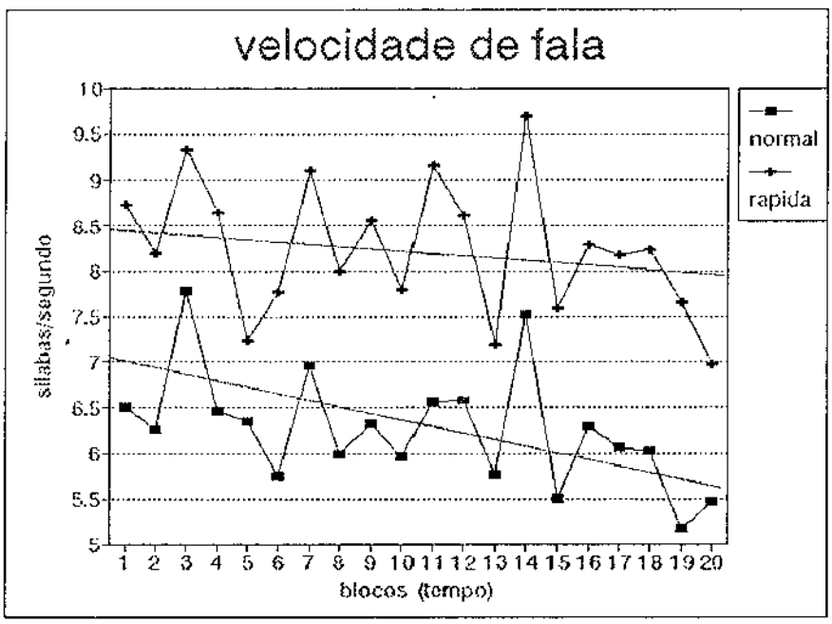


FIGURA 7.8: Velocidades médias em cada bloco de fala, nas duas condições de velocidade (normal *vs.* rápida), para todos os falantes agrupados. As retas de regressão ajustadas à médias de cada bloco mostram que há uma tendência à desaceleração do princípio para o fim do enunciado como um todo.

7.4) *Razão Entre Trechos Vozeados e Não-Vozeados (RTV)*

Dentro da dimensão prosódica há uma série de aspectos de longo termo que ainda não foram sistematicamente estudados com o objetivo de identificar falantes. Um desses aspectos é a proporção de tempo em que ocorre fonação, ou seja, a razão entre a duração total de vozeamento e a duração total do enunciado [3]. Há poucas referências a esse tipo de parâmetro na literatura (uma exceção é Johnson *et al.* 1984). A medida é potencialmente interessante e parece correlacionar-se com aquela dimensão perceptual que Crystal (1969:164) define em termos de um contraste *spiky/glissando.* Nolan (1983:129) sugere os termos - mais corriqueiros - *clipped/drawled*, que poderíamos traduzir livremente por fala *entrecortada/arrastada*. A fala *entrecortada* se caracterizaria por trechos vozeados (principalmente vocálicos) relativamente curtos e trechos não-vozeados (principalmente consonantais) relativamente longos, enquanto a fala *arrastada* seria percebida com a existência da relação oposta.

As medidas necessárias para estabelecer a razão vozeado/não-vozeado podem ser obtidas através de Predição Linear (LPC) com baixa taxa de erro e com razoável rapidez, mesmo em computadores do tipo PC. Para extrair as medidas utilizadas abaixo, empregamos o programa LPC da KAY Elemetrics, modelo 5635, com *frames* de análise de 20 milisegundos. O enunciado inteiro foi assim analisado (leitura do texto I), totalizando cerca de um minuto de fala (exclusive pausas), ou seja, cerca de 3000 *frames* de análise para cada falante.

A tabela 7.7 reúne as medidas do tempo total de fala, tempo total vozeado e a porcentagem de tempo vozeado. As medidas foram tomadas separadamente para as duas condições de velocidade. Os gêmeos JA e JR só leram o texto na velocidade

normal (texto II), e só essa condição está representada na tabela. A média total não inclui os gêmeos JA e JR.

Foram retiradas todas as pausas de respiração de hesitação; dessa forma, a razão *tempo vozeado/ tempo total* refere-se apenas a trechos contendo efetivamente algum segmento de fala, isto é enunciados contínuos. Chamaremos essa medida de *Razão de Tempo Vozeado* (RTV).

Fal.	V	Tempo Total (seg.)	Tempo Vozeado (seg.)	RTV (%)
ZR	**N**	46.3	18.6	40.2
	R	31.4	14.7	46.8
EN	**N**	45.6	19.7	43.2
	R	37.3	17.9	48.1
R1	**N**	49.8	25.5	51.2
	R	39.9	21.2	53.1
ZP	**N**	46.3	22.3	48.1
	R	37.4	19.0	50.7
AG	**N**	41.9	16.0	38.1
	R	35.6	12.8	35.9
WA	**N**	46.5	21.1	45.3
	R	38.7	18.4	47.5
MS	**N**	47.4	26.2	55.3
	R	35.2	21.4	60.9
R2	**N**	53.2	26.6	50.1
	R	38.9	20.1	51.6
DO	**N**	47.3	22.4	47.3
	R	30.7	14.8	48.3
JA	**N**	89.5	56.4	63.0
JR	**N**	82.3	51.2	62.2
Média	**N**	424.3	199.0	46.6
	R	325.1	158.8	49.9

TABELA 7.7: Tempo total de fala (pausas excluídas), tempo total vozeado e a razão vozeado/tempo total (RTV) expressa em termos percentuais, para cada falante separadamente, nas duas condições de velocidade de produção

Observamos, na tabela 7.7, que a velocidade de fala tem um efeito considerável sobre RTV. Há uma tendência a aumentar a proporção de segmentos vozeados na velocidade rápida (com a única exceção do falante AG que apresentou a tendência oposta). Esse resultado sugere que a compressão dos segmentos na fala rápida não se dá de forma homogênea. Se considerarmos que os segmentos vocálicos respondem pela maior parte do tempo vozeado, as medidas parecem indicar que as vogais sofrem menor compressão na fala rápida do que os segmentos consonantais.

A fala rápida provavelmente está associada a uma reorganização motora complexa, que não envolve apenas uma reparametrização linear dos gestos articulatórios (Löfqvist 1991; Adams e Kent 1993; Tuller *et al.* 1982); assim, a fala rápida resulta em uma compressão não-linear das durações segmentais, com reduções não diretamente proporcionais nos segmentos vocálicos e consonantais (Gay 1981). Gay (1978a); verifica, através de medidas eletro-miográficas (EMG), que, com o aumento da taxa de fala (*speech rate*), a atividade da língua em vogais decresce, mas cresce em consoantes (especialmente em *stops* alveolares e consonantais), provocando uma redução duracional de maior magnitude nas vogais em fala rápida (em comparação com consoantes).

Esse resultado, entretanto, não se confirma nos dados aqui estudados, onde o parâmetro RTV tende a ser maior na fala rápida, indicando uma alteração no sentido oposto ao observado em Gay (1978a). Nossos resultados, no entanto, são consistentes com os relatados por van Son e Pols (1990); nesse estudo são comparadas as durações das vogais em duas condições de velocidade de produção (normal *versus* rápida), observando-se que, enquanto a redução global de duração na fala rápida é de cerca de 25%, a redução média das vogais é menor que 15%. É possível que a divergência entre os resultados esteja relacionada ao tipo de material usado em cada estudo; van Son e Pols (1990:1692) sugerem que a leitura de textos

longos, mesmo na condição velocidade "normal", é produzida com taxas mais rápidas do que a produção de frases isoladas em forma de citação, o que faria com que as diferenças duracionais das realizações vocálicas nas duas velocidades fosse menor nesse tipo de material (leitura de texto longo) - semelhante ao empregado no presente trabalho.

É possível que também esteja em jogo aqui o fato de nossos falantes terem sido explicitamente instruídos a produzir, mesmo na velocidade rápida, fala com inteligibilidade plena, evitando assim, intencionalmente ou não, maiores efeitos de redução vocálica (*target undershoot*; v. Lindblom 1963). Alguns estudos já verificaram que o *target undeshoot* vocálico não é uma conseqüência necessária do aumento de velocidade, se os falantes são solicitados a produzir fala "clara" (*clear speech*) (Lindblom e Moon 1988; Moon 1991). É provável que a manutenção da qualidade vocálica na fala rápida "clara" exija uma menor redução duracional nas vogais do que quando não existe essa limitação; assim, a redução global na duração nesse tipo de tarefa pode estar mais fortemente associada a segmentos não-vocálicos, como as fricativas, por exemplo, o que explicaria - pelo menos em parte - o aumento do parâmetro RTV na velocidade rápida, tal como aqui observado.

A faixa de variação interfalante para o parâmetro RTV é considerável, especialmente na velocidade normal de produção, variando de um mínimo de 38.1% (falante AG) até um máximo de 63.0% (falante JA). Uma questão importante aqui é verificar a estabilidade da medida para um diferente número de *frames*. Na seção onde discutimos o Espectro de Longo Termo (ELT) vimos que essa medida só se estabiliza a partir de amostras maiores que 10-15 segundos de fala contínua; é possível que algo semelhante ocorra com o parâmetro RTV. As porcentagens listadas na tabela 7.7 foram obtidas a partir do número total de *frames* para cada

falante - cerca de 2500 na fala normal e 2000 na fala rápida; de modo a examinar a variação de RTV em função do tamanho da amostra, extraímos a medida a partir de amostras de 500, 1000 e 1500 *frames* (é importante ressaltar que as amostras de 500, 1000 e 1500 *frames* têm o menor grau possível de superposição entre si). Os resultados assim obtidos estão reproduzidos nas figuras 7.9a,b, para as duas condições de velocidade de produção.

Podemos observar nas figura 7.9a,b, especialmente na velocidade normal, que as medidas de RTV tendem a convergir a partir de amostras de 1000 *frames*; as medidas obtidas a partir de 500 *frames* são, em alguns casos, bastante diferentes da medida com base no número total de *frames* (falantes WA, DO e R1). É interessante observar que, para alguns falantes, a medida é mais estável, independentemente do tamanho da amostra (particularmente os falantes AG e ZP). Na velocidade rápida, as medidas obtidas a partir de 1000 *frames* ainda são consideravelmente diferentes da medida a partir do número total de frames, e só a partir de 1500 *frames* os resultados parecem convergir mais fortemente. Um aspecto interessante nas medidas de RTV na velocidade rápida é o fato de haver um certo achatamento nas diferenças inter-falantes; apenas os falantes AG e MS permanecem bastante distintos dos demais [4].

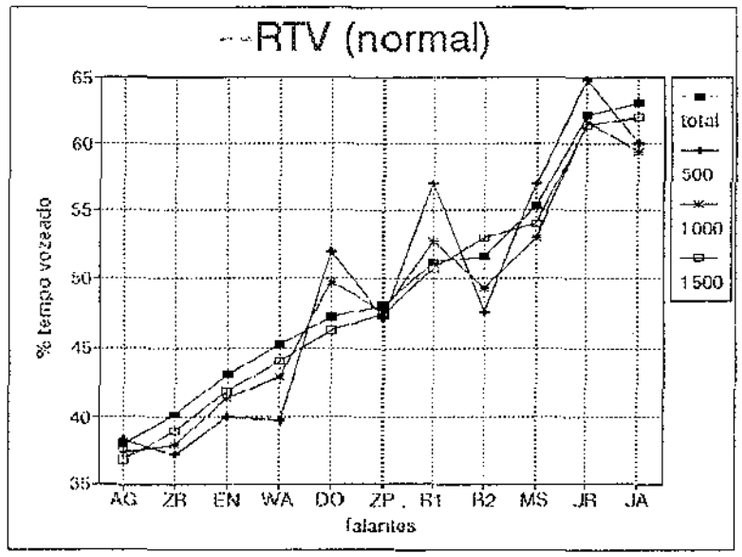


FIGURA 7.9a: Médias por falante de RTV, em função do número de *frames* (velocidade normal)

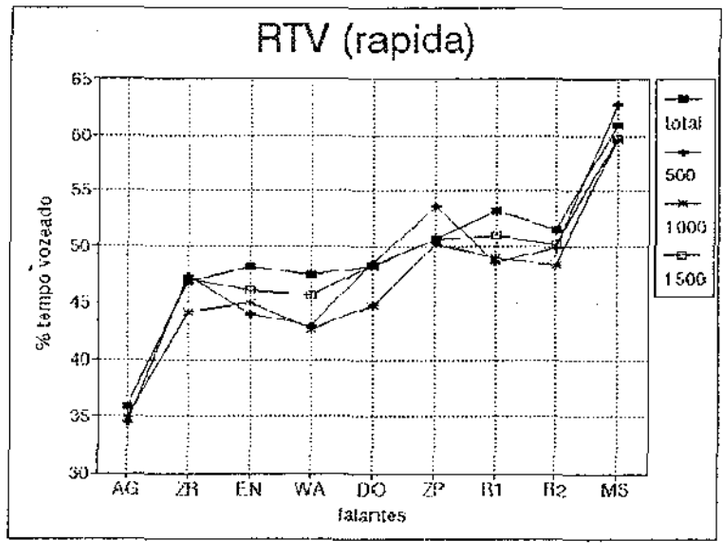


FIGURA 7.9b: Médias por falante de RTV, em função do número de *frames* (velocidade rápida)

É interessante observar, na tabela 7.7, que os gêmeos JA e JR têm RTV praticamente igual; a diferença entre JA e JR é bem menor do que diferenças entre quaisquer dois falantes do Grupo I, ou diferenças entre as duas condições de velocidade para qualquer falante. Esses valores muito próximos de RTV para os gêmeos JA e JR coloca uma questão: em que medida a grande semelhança auditivo-perceptual entre as vozes de JA e JR dependeria - pelo menos em parte -dessa dimensão? É importante lembrar que as análises estatísticas com base em medidas do Espectro de Longo Termo (seção 6) indicaram que os ELTs de JA e JR, a partir da leitura do mesmo texto, eram suficientemente distintos para permitir uma classificação automática correta.

7.5) *Níveis Quantizados de Amplitude*

Medidas derivadas da intensidade de fala têm sido pouco empregadas para a Identificação de Falantes, embora essa seja uma dimensão prosódica importante e, portanto, potencialmente relevante para acessar características individuais (Pruzansky 1963). A princípio, a própria intensidade média da fala parece ser uma característica pessoal mais ou menos estável (Healey 1987), embora possam existir variações intra-falante em função da expressão de diferentes modos emocionais (Lieberman e Michaels 1962) ou da diminuição da capacidade vital (Helfrich 1979).

Uma das dificuldades aqui, é estabelecer um referencial que permita comparar medidas de intensidade, já que esse é um parâmetro acústico sensível às condições do meio (nível de entrada de gravação, distância falante/microfone, tipo de fita e microfone utilizados, largura de banda do canal de transmissão, presença de ruído ambiental, nível de absorção acústica do ambiente, etc.).

Uma possível solução para esse problema é encontrar um conjunto de medidas de caráter **relativo**, definidas a partir do sinal já gravado. Johnson *et al.* (1984) criam um artifício engenhoso para tratar essa questão, estabelecendo níveis quantizados de energia ajustados pelo pico de amplitude de cada falante; Johnson *et al.* definem dez níveis equidistantes entre o máximo de amplitude individual e um mínimo correspondente ao ruído de fundo isolado. Cada um dos níveis gera uma distribuição energia/tempo representando o número e a extensão (tempo) dos *bursts* de fala em cada nível quantizado; as médias e desvios-padrão de cada um dos níveis são então utilizadas para a composição de um vetor multidimensional associado ao falante específico.

Uma dificuldade relacionada ao método de tratamento de amplitude proposto por Johnson et. al. (1984) diz respeito à comparação de amostras com diferentes níveis de ruído de fundo. Embora no paradigma de Verificação Automática de Falante seja possível controlar esse aspecto, na situação forense típica, a expectativa é termos gravações com condições do meio bastante diferentes.

A existência de ruído de fundo com nível estável ao longo da gravação não deve alterar significativamente as medidas extraídas a partir de níveis quantizados; Sambur e Jayant (1976) verificam que medidas de amplitude **relativas** extraídas por LPC são extremamente resistentes à presença de ruído com padrão constante, desde que a razão sinal/ruído permaneça acima de 12 dB. Por outro lado, em se tratando de ruído com padrão de amplitude variável, é evidente que haverá maiores dificuldades em utilizar o procedimento de Johnson *et al.* (1984); nesse caso seria preciso fazer normalizações por trecho isolado, nas regiões onde o ruído se estabilizasse.

No caso específico do conjunto de falantes aqui estudado, não há problema em comparar diretamente as medidas de amplitude normalizadas pelos picos individuais, já que o ruído de fundo é praticamente desprezível. Para obter os dados de amplitude tomamos como base as medidas de energia extraídas por LPC, em

frames de 20 milisegundos, então transformadas para dB; essas medidas em dB foram então normalizadas para cada falante de acordo com sua extensão de amplitude, igualando-se a 10 o máximo individual. Cada medida **normalizada** de amplitude será então

$$\mathbf{Amp_{normal}} = [10 \times (\mathbf{Amp} - \mathbf{MIN})] \div \mathbf{EXT}$$

onde **Amp** é a amplitude em dB medida em um determinado *frame,*
MIN é a amplitude mínima (nível do ruído de fundo) e
EXT é a extensão de amplitude para um determinado falante.

A amplitude assim normalizada (na faixa de 0 a 10 para todos os falantes) foi então quantizada em dez níveis equidistantes (0-1, 1-2, 2-3,..., 9-10). Todos os procedimentos estatísticos abaixo tomaram como base essas medidas assim normalizadas e quantizadas.

A figura 7.10 mostra as distribuições de amplitude para seis falantes (de modo a não sobrecarregar o gráfico); cada ponto no eixo horizontal representa um nível de amplitude normalizado e quantizado, e o eixo vertical apresenta o percentual de tempo em cada um desses níveis. Podemos observar que há uma considerável variabilidade inter-falante quanto às distribuições, com as diferenças individuais concentrando-se principalmente nas proximidades dos níveis mais altos de amplitude (8-9), ou seja, alguns falantes permanecem um maior tempo em níveis mais altos de amplitude (JR e MS, por exemplo), enquanto outros utilizam uma faixa mais ampla ao longo do enunciado (EN e ZR, por exemplo).

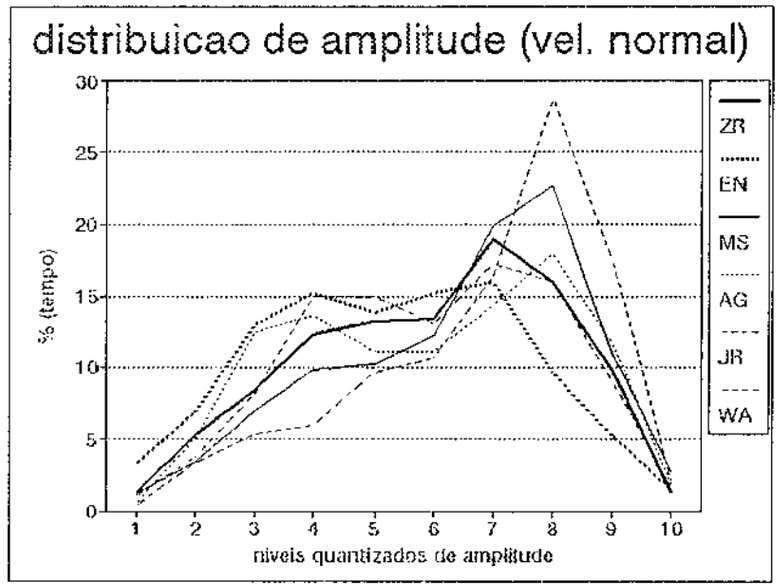


FIGURA 7.10: Distribuições de níveis quantizados de amplitude para 6 falantes

A figura 7.11 mostra as distribuições a partir das médias para cada uma das condições de velocidade de produção (normal *versus* rápida). Podemos observar que as configurações são bastante semelhantes, embora se verifique na velocidade rápida uma tendência a aumentar a proporção de tempo nos níveis mais altos de amplitude. Examinando, entretanto, a variabilidade inter-falante para cada um dos níveis separadamente, pudemos observar que, na velocidade rápida, há um aumento acentuado do desvio-padrão concentrado no nível 9, como demonstra a figura 7.12, indicando que os resultados médios apresentados na figura 7.11 podem ser enganosos, escondendo variações intra-falante importantes (em função da velocidade de produção). De fato, ao examinar o comportamento isolado de cada falante, pudemos constatar que alguns falantes alteram bastante o padrão na velocidade

rápida, enquanto outros mantêm o mesmo padrão. A figura 7.13 mostra as distribuições, nas duas condições de velocidade, para dois falantes: o falante MS altera consideravelmente seu padrão na velocidade rápida, aumentando a proporção de tempo em níveis mais altos de amplitude, enquanto o falante R1 mantém praticamente a mesma distribuição, independentemente da velocidade.

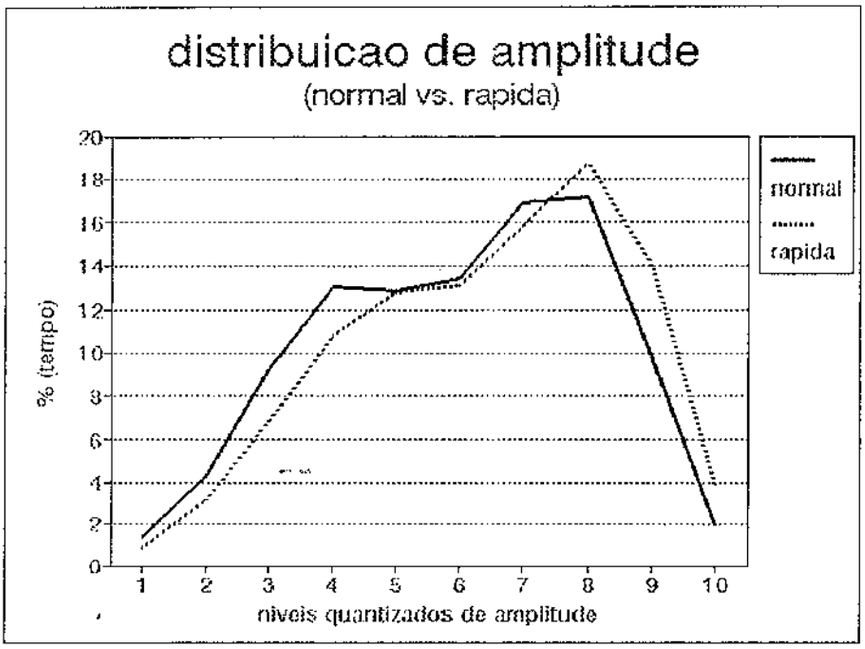


FIGURA 7.11: Distribuições de níveis quantizados de amplitude nas duas condições de velocidade de produção (agrupando todos os falantes.

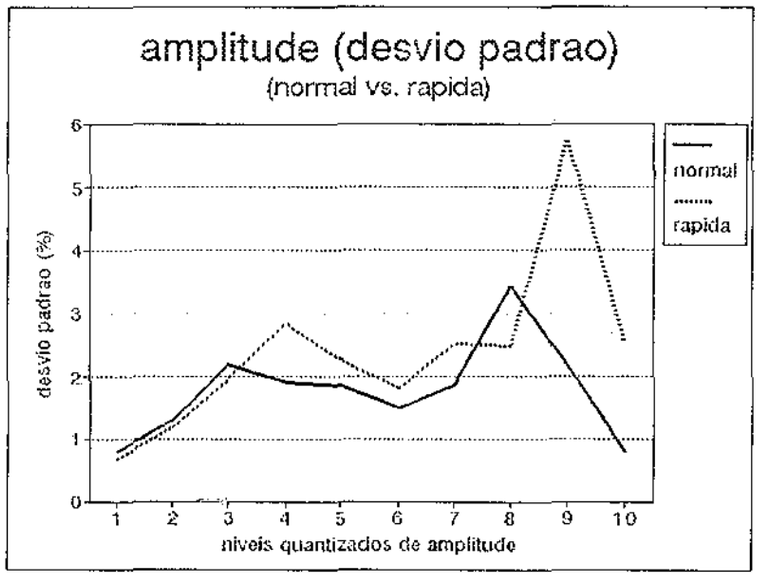


FIGURA 7.12: Desvios-padrão inter-falante em cada nível quantizado de amplitude, nas duas condições de velocidade de produção

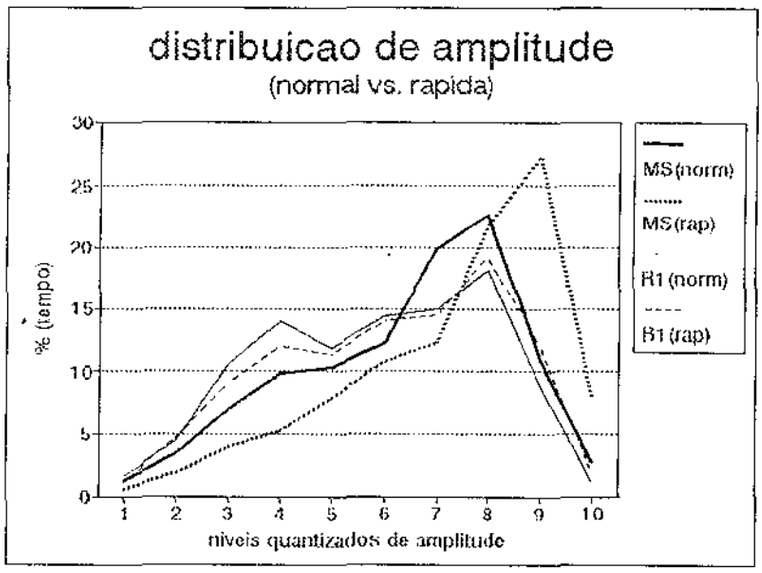


FIGURA 7.13: Distribuições de níveis quantizados de amplitude, comparando os resultados nas duas condições de velocidade para dois falantes do grupo estudado

Utilizamos as medidas quantizadas de amplitude como entrada para uma análise estatística *cluster*, de modo a verificar a eficácia desse tipo de parâmetro para a separação automática dos falantes aqui estudados. O procedimento é semelhante ao já descrito na seção 6, quando examinamos o Espectro de Longo Termo.

Para obter mais de uma amostra para cada falante, dividiu-se o enunciado total pela metade; assim, foram obtidas quatro amostras por falante, duas na velocidade normal e duas na velocidade rápida (primeira e segunda metades, em cada condição de velocidade). Os gêmeos JR e JA foram incluídos no grupo, mas, como não produziram amostra em velocidade rápida, não aparecem no dendrograma para essa condição.

Os dendrogramas da figura 7.14 mostram os melhores resultados dessa análise *cluster* (BMDP-1M) para cada condição de velocidade isoladamente e para as duas condições reunidas. Na velocidade normal, os *clusters* iniciais formados agrupam corretamente seis pares de falantes, inclusive separando corretamente os gêmeos JR e JA. As amostras dos falantes MS e WA permanecem próximas, sendo reunidas no segundo nível de similaridade do dendrograma, enquanto as amostras dos falantes ZR e DO estão bastante afastadas, indicando que esses falantes não mantiveram o mesmo padrão ao longo de todo o enunciado. As amostras não contemporâneas do falante R1/R2 também não foram aproximadas.

Na velocidade rápida apenas três falantes foram corretamente identificados (MS, ZP e AG). O último dendrograma, reunindo as duas condições de velocidade aponta um quadro pouco favorável: poucos falantes foram corretamente identificados e, em nenhum caso formou-se um *cluster* inicial aproximando amostras do mesmo falante em diferentes velocidades.

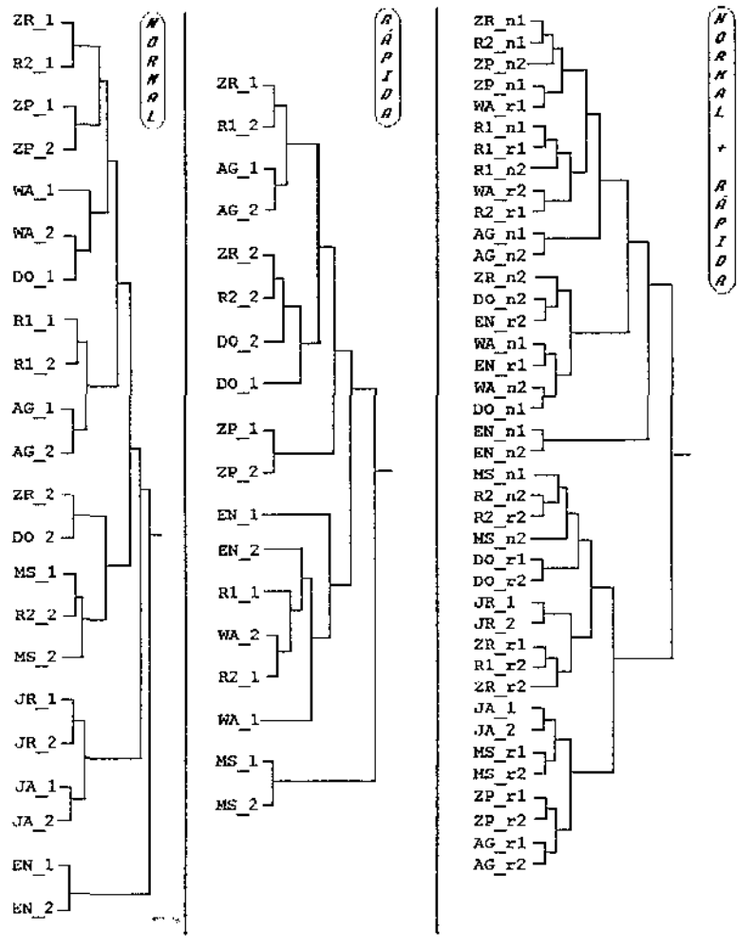


FIGURA 7.14: Dendrogramas criados por análise *cluster* (BMDP-1M) com base em níveis quantizados de amplitude

Os resultados aqui observados sugerem que parâmetros derivados da distribuição amplitude/tempo são pouco robustos (pelo menos na forma como esses parâmetros foram aqui definidos). Johnson *et al.* (1984), com base em um conjunto de 40 falantes, relatam um índice de 60% de identificações corretas (análise discriminante) utilizando um vetor de 40 dimensões composto de medidas obtidas da distribuição energia/tempo. Mesmo empregando um conjunto mais complexo de

medidas temporais, o índice de acerto obtido por Johnson *et al.*(1984) não é muito promissor; a performance cai ainda mais na presença de *stress* psicológico e disfarce, condições que, no mesmo experimento de Johnson *et al.* (1984), obtêm apenas 40 e 30% de identificações corretas, respectivamente. Por outro lado, é preciso considerar que, apesar dos baixos índices de acerto, o vetor energia/tempo utililizado por Johnson *et al.* (1984) é provavelmente ainda mais eficiente, nas condições de *stress* e disfarce do que outros parâmetros como F0 e Espectro de Longo Termo (Cf. Doherty 1975; Hollien e Majewski 1977).

7.6) *Comentário Final*

Ao longo da presente seção examinamos alguns aspectos relacionados à dimensão rítmico/temporal: velocidade de fala, razão vozeado/tempo total de fonação (RTV) e medidas quantizadas de amplitude. Observamos que esses parâmetros, em maior ou menor medida, refletem caraterísticas idiossincráticas, embora sofram eventualmente restrições em função da covariância de outros fatores.

Os aspectos aqui estudados podem servir de base para a composição de um vetor visando quantificar o fenômeno complexo que denominamos genericamente de RITMO, um conceito cuja validade perceptual é inegável, mas que, até o momento, permanece sem uma definição fonética não ambígua. Uma possibilidade, não explorada no presente estudo, mas aparentemente promissora, seria tentar capturar regularidades rítmicas através de uma análise espectral de longo termo das curvas de amplitude. Esse tipo de análise poderia revelar diferentes padrões temporais, de modo equivalente a uma análise que, no domínio das freqüências, evidencia picos espectrais em determinadas regiões. Warner e Mooney (1988) utilizam um procedimento dessa natureza, criando assim o que chamam de *periodogramas*, usados para descrever aspectos cíclicos presentes na conversação entre duas pessoas.

Embora o objetivo de Warner e Mooney seja estudar ciclos longos (tomadas de turno) o método em si é bastante interessante e, a princípio, deve ser capaz de capturar ciclos de várias ordens, refletindo assim a estruturação rítmica de um determinado falante [5].

SEÇÃO 8: CONSOANTES NASAIS

8.1) *Eficiência de Sons Nasais na Identificação de Falantes*

Parâmetros acústicos relacionados a características fisiológicas estáveis do trato vocal são naturalmente interessantes para a identificação de um falante, na medida em que, a princípio, variam menos em função da não-contemporaneidade das amostras e são menos suscetíveis ao controle do falante. Nesse sentido, os sons nasais - em especial as consoantes - são particularmente relevantes, já que as cavidades nasais são uma parte do trato que não podem ser alteradas voluntariamente e, além disso, existe uma variação inter-subjetiva considerável quanto a seu tamanho e estrutura, características que produzirão saídas acústicas particulares. Outra vantagem das consoantes nasais reside no fato de essa classe de sons ser produzida com os articuladores e as cavidades vocais mantidos relativamente fixos (especialmente no caso de /n/; Cf. Glenn e Kleiner 1968), embora se deva esperar, como veremos mais adiante, uma certa variação em função do contexto fonético (o grau de variação relacionado a efeitos de co-articulação, no entanto, pode também refletir características idiossincráticas; Cf. Su *et al.* 1974).

Para os estudos baseados no Inglês, há uma motivação adicional para o estudo das nasais com a finalidade de identificar falantes: a freqüência relativa com a qual esses sons ocorrem no Inglês falado. De acordo com Tobias (1959), as nasais constituem 11 % do conteúdo fonêmico do Inglês Americano falado. No Português Brasileiro é menor o número de ocorrências das nasais, embora esses fonemas se situem ainda entre as consoantes mais freqüentes, totalizando cerca de 6.52 % do total de fones (vogais + consoantes) (Alcaim *et al.* 1993).

Alguns experimentos têm indicado que medidas extraídas do espectro de consoantes nasais são bastante eficientes para a determinação da identidade de um falante. Höfker (1977; *apud* Nolan 1983:75) relata um experimento onde 24 fonemas (Alemão) são avaliados quanto à eficiência na discriminação de 12 falantes; as nasais /n,ŋ,m/ apresentaram o melhor desempenho entre os fonemas estudados. Wolf (1972), testando uma série de parâmetros acústicos, verifica que duas medidas espectrais, uma de /m/ e outra de /n/, encontram-se entre os índices mais eficientes, dentre os parâmetros segmentais analisados. Reich *et al.* (1976), em um experimento testando a eficiência relativa de diferentes palavras para a Identificação de Falantes através de leitura espectrográfica sem auxílio auditivo, verificam que aquelas palavras contendo fonemas nasais (*me, on e and*) tiveram desempenho superior às demais.

A superioridade das nasais, entretanto, não é observada em todas as pesquisas; Paul *et al.* (1975), desenvolvendo um sistema semi-automático de Identificação, baseado no isolamento visual e auditivo de segmentos fonéticos, relatam os seguintes 13 fonemas (Inglês Americano) como os mais eficientes, na ordem decrescente de performance:

/U,u,i,m,I,a,O,n,ŋ,ʌ,ð,A,ɛ/.

Independentemente do resultado relatado, os estudos citados no parágrafo anterior negligenciam um aspecto importante: o efeito do ambiente fonético na realização de um dado segmento. Paul *et al.* (1975) utilizam segmentos com o mesmo contexto fonético circundante, Wolf (1972) baseia suas medidas em uma mesma sentença lida por diferentes falantes e Höfker (1977) emprega apenas segmentos produzidos em isolamento. Esses procedimentos podem ter neutralizado em grande parte a variabilidade intra-subjetiva, tornando difícil uma avaliação

objetiva dos resultados, especialmente se considerarmos uma possível aplicação baseada na fala fluente, tal como ocorre no modelo forense (por outro lado, é importante ressaltar que, em se tratando do paradigma de Verificação Automática de Falante, a manutenção do mesmo ambiente é perfeitamente viável). Na verdade, o espectro nasal varia consideravelmente em função do contexto fonético (Cf. Kurowski e Blumstein 1987:1920).

Glenn e Kleiner (1968) descrevem um experimento de Identificação, onde são utilizadas amostras de /n/ obtidas em diferentes ambientes. Nesse estudo, 30 falantes (Inglês Americano) lêem duas listas de palavras onde o fonema /n/ ocorre em diferentes contextos fonéticos e posições na palavra (inicial, medial e final). Cada ocorrência foi representada por um espectro quantizado (média de 3 medidas extraídas da região central de cada /n/), constituído de 25 faixas de 100 Hz, abrangendo a faixa total 1000-3500 Hz, sendo a amplitude de cada faixa descrita em 6 bits. Assim, cada seção espectral é convertida em um vetor de 25 × 64 dimensões (25 faixas de freqüência × 2^6 níveis de amplitude). Um vetor de referência é formado através da média de 10 amostras de cada falante extraídas da primeira lista de palavras e é comparado, após uma transformação para normalização das amplitudes, com um segundo vetor médio obtido da mesma forma a partir da segunda lista de palavras. Como critério de decisão foi utilizado um método de correlação angular; nas tentativas de identificação, um vetor teste é julgado como pertencendo a um falante se a correlação entre o vetor referência desse falante e o vetor teste for maior do que a correlação entre o vetor referência de qualquer outro dos falantes e o vetor teste.

Glenn e Kleiner (1968) relatam para seu sistema um índice de acerto de 43 %, se é usada apenas uma única amostra referência qualquer de cada falante. Esse número aumenta para 93 % quando é utilizada a média das 10 amostras de cada falante, e para 97 % se a população é reduzida para apenas 10 falantes. Esses

resultados indicam que o uso de nasais, /n/ em particular, é um meio relativamente eficaz para a Identificação de Falantes. Um aspecto importante a ressaltar nesse experimento é o fato de a informação espectral ter sido drasticamente reduzida quanto à estrutura mais fina; os vários procedimentos de quantização, normalização e média sobre média fazem com que ao final apenas os aspectos mais contrastantes do espectro permaneçam, notadamente os picos correspondentes a F3, F4 e F5 de /n/ (em torno de 1500, 2200 e 2600 Hz, respectivamente). A utilização de espectros médios, apesar de neutralizar variações espectrais relacionadas à diversidade dos contextos fonéticos, foi um procedimento relativamente robusto, pelo que se pode depreender dos resultados apresentados. Quanto a isso, os autores comentam que a assunção de unimodalidade parece razoável, visto que as variações resultantes das condições do ambiente fonético são geralmente menores do que as variações devidas à identidade do falante (Glenn e Kleiner 1968:372). Essa afirmação, entretanto, não se confirma totalmente, já que, como vimos acima, há uma redução drástica na performance do sistema quando a amostra teste consiste apenas em um único vetor aleatoriamente escolhido para um falante qualquer do grupo.

Su *et al.* (1974), em outro estudo baseado em nasais, tentam empregar, de forma mais positiva, a informação oriunda dos efeitos de co-articulação entre a consoante nasal e o ambiente fonético, mais especificamente a vogal imediatamente seguinte. Su *et al.* partem da seguinte premissa:

> *if the nasal consonant is followed by a vowel, the tongue anticipates the following vowel segments and moves to the vowel position during the nasal phonation.* (Su *et al* 1974:1877).

A magnitude dessa co-articulação, segundos os autores, deverá ser maior para /m/, pois nessa nasal a língua não tem função específica, enquanto em /n/, há uma maior limitação dos movimentos desse articulador. Se houver co-articulação em função da vogal - afirmam os autores - o espectro médio das nasais seguidas de vogais anteriores deve ser diferente do espectro médio das nasais seguidas de vogais posteriores. Nas consoantes nasais, os formantes (polos da função de transferência) estão relacionados diretamente com a configuração da faringe e da cavidade nasal, estruturas que praticamente não se alteram durante a produção desses sons. Os zeros nasais (anti-formantes) estão relacionados à influência da cavidade oral, que funciona como um ressoador *side-branch* acoplado ao sistema ressoador principal (nasal). Cada nasal possui seu anti-formante em uma região determinada de freqüência: /m/ entre 750-1250 Hz, /n/ entre 1450-2200 Hz e /ŋ/ acima de 3000 Hz (Fujimura 1962). Os efeitos de co-articulação examinados por Su *et al.* (1974), na medida em que alteram a configuração da cavidade oral, devem modificar a posição do zero nasal, cuja conseqüência acústica deve ser o deslocamento do *cluster* formante/anti-formante no espectro nasal. De acordo com Fujimura (1962:1872), a posição do anti-formante, tanto para /m/ quanto para /n/, deve ser mais alta em freqüência na vizinhança de vogais anteriores, em virtude da configuração antecipatória do trato oral (estreitamento da extremidade anterior).

Su *et al.* utilizam em seu experimento palavras *nonsense* da forma /hə'CVd/, onde /ə/ é a vogal central neutra, C é uma das nasais /m/ ou /n/, e V uma das vogais anteriores /i,e,æ/ ou posteriores /a,o,u/. Cada um dos falantes (n=4) fala cada palavra 3 vezes, totalizando 36 enunciados. O sinal foi analisado por um banco de filtros e, após a digitalização (10 bits), as saídas dos primeiros 25 filtros, abrangendo a faixa de 250-3681 Hz, foram transformadas em um vetor. A faixa útil de freqüência adotada por Su *et al.* é suficiente para estudar o fenômeno em questão, já que, como vimos acima, os zeros de /m/ e /n/ encontram-se abaixo de 2500 Hz.

A análise estatística realizada em Su *et al.* (1974) revelou que as amostras espectrais de /m/ antecedendo vogais [+posterior] formam um *cluster* distinto das amostras de /m/ antecedendo vogais [+anterior]. O mesmo, entretanto, não foi observado no caso de /n/, onde se forma um único *cluster*, refletindo o fato de a língua, na produção de /n/, ter seus movimentos limitados pelo contato com a região alveolar.

O aspecto mais relevante do experimento de Su *et al.* reside na constatação de que a estrutura do espaço onde se formam os *clusters* baseados nos espectros de /m/ é diferente para cada falante. Um experimento suplementar de identificação automática através de matriz de correlações atingiu 100 % de acertos para um universo de 10 falantes, com base apenas nas informações derivadas das diferenças dos espectros de /m/ em diferentes contextos fonéticos (precedendo vogal [+posterior] ou [+anterior]). Dito de outro modo, o experimento de Su *et al.* indica que um indivíduo pode ser caracterizado pela extensão de sua co-articulação e que a informação extraída da co-articulação nasal é mais eficiente do que a informação dada pelo espectro nasal isolado.

8.2) *Limitações ao Emprego de Nasais*

Apesar dos resultados promissores de alguns estudos envolvendo medidas derivadas dos espectros de consoantes nasais, existe uma série de dificuldades relacionadas ao tema. A primeira delas refere-se à própria obtenção das medidas; a presença de zeros na função de transferência das nasais e o efeito de *damping* característico dificultam consideravelmente a determinação dos picos correspondentes aos polos da função de transferência (formantes). Além disso, espera-se também uma certa variação na posição do zero ao longo do murmúrio nasal (Fujimura 1962), colocando problemas para a determinação do ponto ótimo

para realizar a medida. Aliada a essas dificuldades, há uma limitação adicional, vinda da redução da energia acústica durante as consoantes nasais; esses sons são, geralmente, de baixa intensidade (excetuando o primeiro formante nasal, pouco interessante como indicador de identidade), especialmente na fala fluente, um aspecto que pode ser bastante perturbador em gravações de baixa qualidade, tais como as que encontramos freqüentemente na aplicação forense. Mesmo considerando apenas o domínio perceptual, há dificuldades para o emprego de nasais. Esses sons tendem a ser os mais afetados pela presença de ruído e/ou reverberação ambiental (Helfer 1992).

Embora seja verdadeiro que a estrutura do trato nasal constitua um fator de variação orgânica inter-subjetiva - o que representa um aspecto positivo para a finalidade de Identificação -, não é menos verdadeiro que as condições de ressonância do trato nasal de um falante podem sofrer algumas alterações vinculadas, por exemplo, à consistência do muco (House 1957). Como possibilidade adicional de variação intra-subjetiva não podemos esquecer que o falante, com algum treino, é capaz de controlar a abertura do *velum*, podendo, no limite, produzir voz desnasalizada, evitando assim fornecer informação sobre a conformação de suas cavidades nasais [1].

8.3) *Um experimento com a nasal /n/*

De modo a testar a variabilidade de características espectrais da consoante nasal /n/, em função do falante e do ambiente fonético, realizamos um pequeno experimento. Cada um dos falantes do grupo principal (n=9: 7 + R1/R2) leu uma lista de palavras contendo seqüências do tipo /V_1n'V_2/, isto é, VOGAL + /n/ + VOGAL TÔNICA, onde V_1 pode ser /a,e,i,o,ã/ e V_2 pode ser /a,E,e,i,O,o,u/. Não existem, entretanto, todas as combinações possíveis entre V_1 e V_2, ou seja, os dados

não estão balanceados quanto a esse aspecto [2]. As palavras utilizadas foram as seguintes:

ANÁPOLIS
ONOFRE
MONÔMERO
INÍCIO
ANÍSIO
INEPTO
CINEMA
CONECTA
INATO
INOVA
ENEMA
ANU
AFINA
HUMANO
MENOS

A tabela 8.1 apresenta as médias e desvios-padrão do quinto e sexto formantes de /n/, para cada falante, medidos na região central da nasal [3]. De modo a verificar uma possível distorção provocada por valores espúrios, incluiu-se também as médias e desvios após *trimming* de 15%; esse procedimento exclui dos dados valores extremos, cortando 15% dos valores nos limites superior e inferior da distribuição. A diferença entre as medidas baseadas na distribuição integral e na distribuição após o *trimming* não diferem substancialmente, indicando que há poucos valores anômalos.

F5 /n/									
fal.→	**ZR**	**EN**	**R1**	**ZP**	**AG**	**WA**	**MS**	**R2**	**DO**
med.	2728.0	2912.0	2670.7	2642.8	2728.6	2738.2	2546.7	2673.3	2486.7
m.15	2738.1	2900.0	2670.9	2656.7	2712.2	2730.4	2558.1	2680.0	2489.5
dp	99.6	236.6	76.8	111.6	153.9	198.1	66.9	134.9	91.1
dp.15	98.1	216.6	72.2	152.3	145.6	165.2	47.5	130.9	94.7
F6 /n/									
fal.→	**ZR**	**EN**	**R1**	**ZP**	**AG**	**WA**	**MS**	**R2**	**DO**
med.	3736.0	3902.8	3810.7	3686.7	3640.8	3490.9	3540.0	3960.0	3662.9
m.15	3731.4	3892.2	3825.7	3680.9	3654.1	3495.6	3556.2	3968.6	3664.1
dp	150.8	93.4	144.4	114.9	215.7	127.7	174.5	160.5	185.8
dp.15	148.8	86.9	110.9	126.9	183.9	98.3	145.5	151.1	188.5

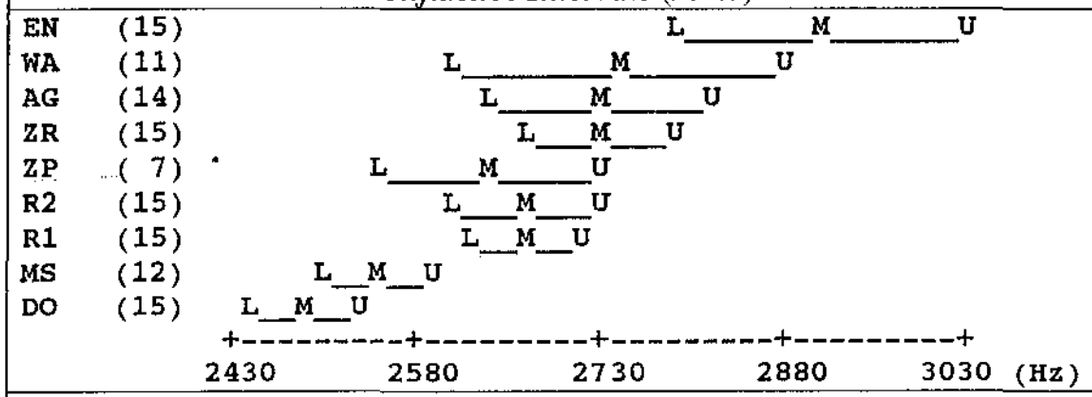
TABELA 8.1: Média e desvio-padrão (Hz) dos formantes nasais para cada falante. Foram também incluídos a média e desvio-padrão depois de *trimming* de 15%.

A tabela 8.1 mostra que existe, em termos de média, uma diferença considerável entre os falantes, tanto para F5 quanto para F6. A significância estatística dessa diferença foi testada através de uma análise de variância, acompanhada do teste de comparação de médias individuais *Student-Newman-Keuls* (programa BMDP-7D). A tabela 8.2 apresenta os resultados da análise de variância para as variáveis F5/n/ e F6/n/ (grupo=FALANTE).

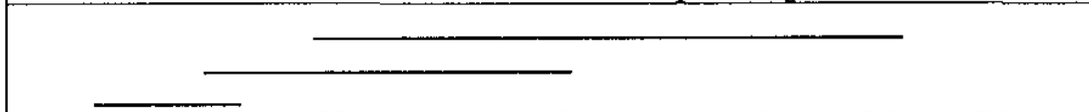
ANOVA Grupo=FALANTE Variável: F5/n/

F = 11.57 G.L. = 8 p< .0001

Confidence Intervals (95 %)



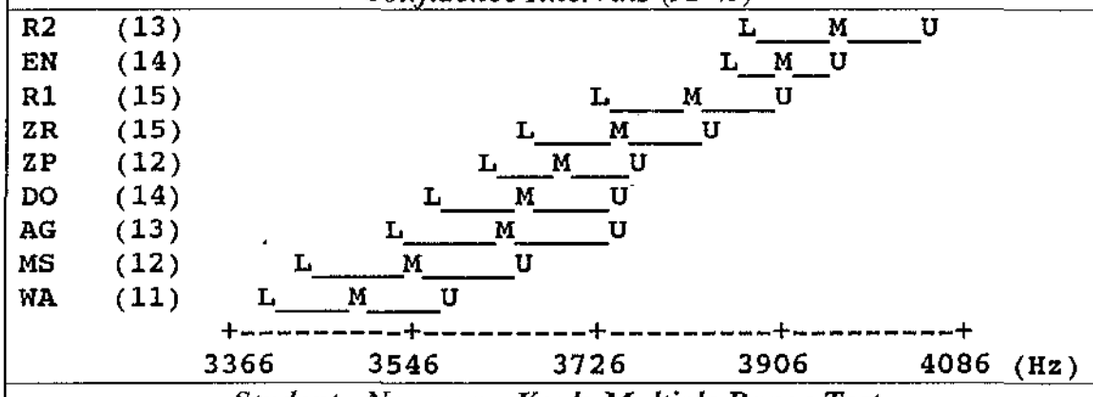
Student - Newman - Keuls Multiple Range Test



ANOVA Grupo=FALANTE Variável: F6/n/

F = 14.09 G.L. = 8 p< .0001

Confidence Intervals (95 %)



Student - Newman - Keuls Multiple Range Test

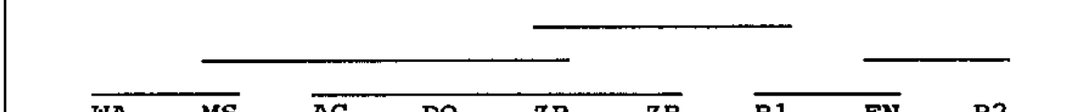


TABELA 8.2: Resultados de ANOVA (BMDP-7D) para as variáveis F5/n/ e F6/n/, agrupadas por FALANTE. O teste *Student-Newman-Keul* indica quais níveis de FALANTE são significativamente diferentes (p<.05), ligando através de linhas horizontais os falantes cujas médias **não** diferem estatisticamente.

Os resultados de ANOVA, na tabela 8.2, mostram que a variável FALANTE influi significativamente no comportamento de F5/n/ e F6/n/ (F= 11.57 e F= 14.09, p<.0001, para F5/n/ e F6/n/, respectivamente). O teste de comparação de médias *Student-Newman-Keuls* une, através de linhas horizontais, os falantes que **não** diferem significativamente quanto à variável estudada; assim, por exemplo, o falante EN é significativamente diferente de todos os demais na distribuição de F5/n/, enquanto o falante WA difere apenas de DO e MS. Há mais pares de falantes significativamente diferentes para a variável F6/n/ do que para F5/n/, o que pode estar relacionado à existência de uma fonte adicional de variação influindo no comportamento de F5/n/; uma possibilidade aqui é a possível maior suscetibilidade de F5/n/ aos efeitos de co-articulação com o ambiente fonético (V_1 e V_2). Examinando os intervalos de confiança a 95 %, verificamos que o âmbito da variação de F5/n/ difere bastante de falante para falante, indicando que, se existe um fator adicional influindo nessa variação (efeito de ambiente fonético, por exemplo), ele não se manifesta igualmente para todos os falantes.

A comparação dos resultados relativos às duas produções não contemporâneas do falante R1/R2 mostra que F5/n/ praticamente não sofreu alteração significativa; por outro lado, F6/n/ sobe consideravelmente na segunda amostra (R2), sem a presença do resfriado (as médias de R1 e R2 para F6/n/ são estatisticamente diferentes). A alteração de F6/n/ em R1/R2 pode estar relacionada às diferentes condições de ressonância devidas à presença de muco nas cavidades nasais, aspecto que pode influir na saída acústica (House 1957; Sambur 1975; Rosenberg 1976:479) [4].

8.3.1) *Influência do Contexto Fonético nos Formantes de /n/*

À luz dos que foi examinado acima, podemos supor que, além da variação relacionada a FALANTE, deve haver uma outra fonte de variação influindo no comportamento de F5/n/ e F6/n/. Vários estudos já verificaram que a qualidade da vogal subseqüente à consoante nasal interfere na sua configuração espectral, em função da co-articulação antecipatória (Fujimura 1962; Glenn e Kleiner 1968; Su *et al.* 1974; Kurowski e Blumstein 1987). A figura 8.1 apresenta as distribuições de F5 e F6 da nasal /n/ (linhas mais espessas), juntamente com F3 e F4 das vogais que precedem e sucedem a nasal (linhas pontilhadas); podemos observar que há uma sobreposição de F5/n/ com F3/V_1,V_2/ e de F6/n/ com F4/V_1,V_2/, sugerindo a possibilidade de efeito co-articulatório.

A tabela 8.3 apresenta as médias e desvios-padrão de F5/n/ e F6/n/, com e sem *trimming* de 15 % dos dados, em função de V_1 e V_2. A tabela 8.4 resume resultados de ANOVA (BMDP-7D), com alguns testes de comparação de médias entre os níveis de V_1 e V_2.

V_1 (vogal imediatamente antes de /n/)											
F5/n/						F6/n/					
V_1	n=	med.	m.15	dp.	dp.15	V_1	n=	med.	m.15	dp.	dp.15
/a/	32	2703.1	2695.2	170.9	130.5	/a/	32	3796.2	3798.2	186.8	190.7
/e/	14	2641.4	2643.7	165.4	158.6	/e/	14	3711.4	3724.9	196.9	195.0
/i/	41	2640.9	2630.0	155.0	128.2	/i/	43	3694.9	3710.6	232.3	204.8
/o/	25	2713.6	2695.4	158.7	136.7	/o/	26	3695.8	3693.2	192.2	168.9
/ã/	7	2834.3	2813.9	298.4	365.5	/ã/	4	3660.0	3668.6	202.7	256.9
Tot.	119	2684.4	--	177.3	--	Tot.	119	3723.1	--	202.5	--
V_2 (vogal imediatamente após /n/)											
F5/n/						F6/n/					
V_2	n=	med.	m.15	dp.	dp.15	V_2	n=	med.	m.15	dp.	dp.15
/a/	17	2756.5	2722.5	185.6	172.9	/a/	16	3757.5	3777.8	223.7	184.9
/ɛ/	17	2629.4	2628.2	105.7	81.5	/ɛ/	17	3764.7	3782.0	206.3	174.6
/e/	15	2576.0	2575.2	107.9	105.6	/e/	18	3755.6	3763.2	166.2	135.5
/i/	22	2618.2	2607.0	146.7	128.1	/i/	25	3732.8	3750.3	221.9	180.3
/O/	17	2762.3	2748.7	173.4	127.9	/O/	17	3691.7	3676.6	204.6	178.6
/o/	23	2735.6	2716.7	192.5	162.4	/o/	17	3612.3	3615.8	203.8	194.5
/u/	8	2720.0	2748.6	147.4	131.3	/u/	9	3760.0	3745.1	236.3	276.7
Tot.	119	2684.4	--	177.3	--	Tot.	119	3723.1	--	202.5	--

TABELA 8.3: Médias de desvios-padrão com e sem *trimming* de 15 % de F5 e F6 da nasal /n/, em função da qualidade das vogais precedendo (V_1) e subseguindo (V_2) a consoante.

ANOVA Grupo=V_1 Variável: F5/n/

F = 2.44 G.L. = 4 p< .06

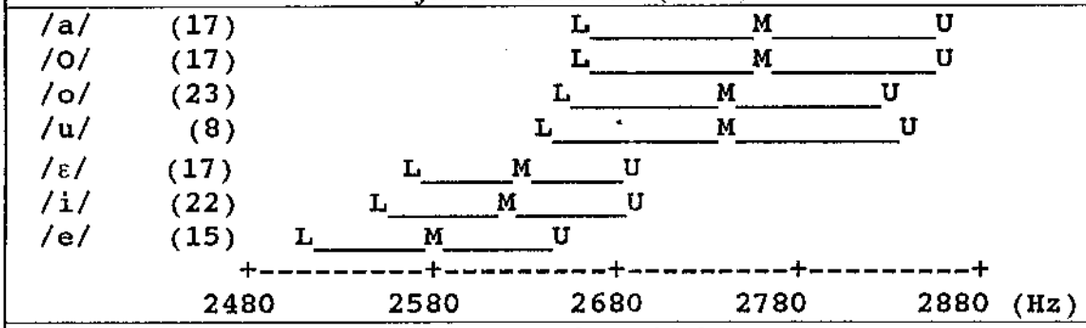
ANOVA Grupo=V_1 Variável: F6/n/

F = 1.50 G.L. = 4 p< .20

ANOVA Grupo=V_2 Variável: F5/n/

F = 3.50 G.L. = 6 p< .003

Confidence Intervals (95 %)



Student - Newman - Keuls Multiple Range Test

/e/ /i/ /ɛ/ /u/ /o/ /a/ /O/

ANOVA Grupo=V_2 Variável: F6/n/

F = 1.26 G.L. = 4 p< .28

TABELA 8.4: Resultados resumidos de ANOVA, testando influência do ambiente fonético (qualidade da vogal antes e após /n/) na variação de F5/n/ e F6/n/. Os intervalos de confiança e o teste *Student - Newman - Keuls* foram mostrados apenas para F5/n/, Grupo= V_2, sendo esta a única combinação com um valor de *F* significativo; todas as demais não indicam qualquer padrão quanto à uma possível influência da qualidade do contexto vocálico.

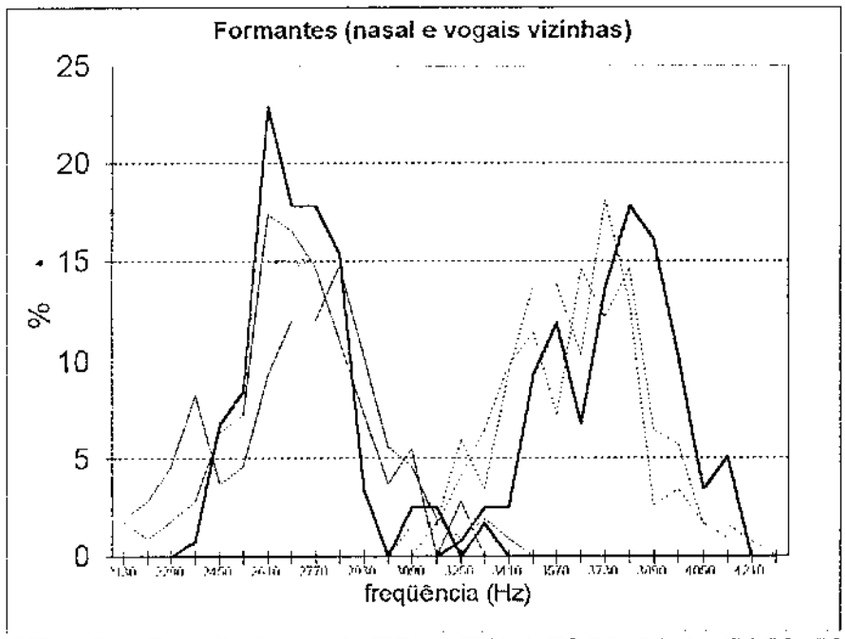


FIGURA 8.1: Distribuições de F5/n/, F6/n/, F3/V_1,V_2/ e F4/V_1,V_2/. As distribuições da nasal estão representadas pelas linhas mais espessas. O eixo vertical mostra o percentual de ocorrências em cada freqüência no eixo horizontal (Hz).

Os resultados das tabelas 8.3 e 8.4 indicam que a influência do contexto vocálico nos formantes de /n/ restringe-se a um efeito co-articulatório de natureza antecipatória, que faz com que haja uma queda de freqüência em F5/n/, quando V_2 é uma vogal anterior. Nenhum efeito significativo relacionado à co-articulação retentiva (influência de V_1) foi observado. A análise dos intervalos de confiança na tabela 8.4 indica que podemos, sem qualquer prejuízo, agrupar as 3 vogais anteriores /i,e,ɛ/ em um único grupo, contrapondo-as a um segundo grupo formado pelas vogais posteriores /a,O,o,u/ (o mesmo procedimento foi utilizado *a priori* no experimento de Su *et al.* 1974, já discutido acima). Após esse agrupamento, realizamos mais uma análise com apenas dois níveis de V_2 (VOGAL ANTERIOR *vs.* VOGAL POSTERIOR), obtendo os resultados apresentados na tabela 8.5.

Variável: F5/n/ VOGAL ANTERIOR *versus* ANTERIOR F= 20.22 G.L. = 4 p<.0001					
VOGAL	n=	média	med. 15	d.padrão	d.padrão.15
anterior	54	2610.0	2606.5	124.2	104.7
posterior	65	2746.1	2730.2	176.0	130.8
Total	119	2684.4	--	177.3	--

TABELA 8.5: Resultados de ANOVA (BMDP-7D), para a variável contínua F5/n/, testando a influência do contexto vocálico, integrando as vogais /i,e,ɛ/ e /a,O,o,u/ em dois grupos distintos (ANTERIOR e POSTERIOR, respectivamente). Estão incluídos na tabela as médias e desvios-padrão após *trimming* de 15 %.

Fica claro, pela tabela 8.5, que a divisão dos dados em apenas dois grupos de vogais aumenta consideravelmente a significância estatística do efeito do contexto fonético sobre F5/n/. Após a integração das vogais V_2 em dois únicos grupos e o conseqüente aumento do número de casos em cada nível (ANTERIOR *vs.* POSTERIOR), parece razoável testar o efeito de interação FALANTE X VOGAL, considerando, entretanto, que o número de casos em cada célula será ainda relativamente pequeno. A tabela 8.6 mostra os resultados de uma nova análise de variância (BMDP-7D), incluindo no modelo os efeitos isolados de FALANTE, VOGAL e a interação FALANTE X VOGAL.

ANOVA	Variável: F5/n/	FALANTE (9) X VOGAL (2)	
Fonte ↓	F =	p<	G.L.
FALANTE	17.23	.0001	8
VOGAL	39.90	.0001	1
FAL. X VOG.	3.05	.005	8

TABELA 8.6: Resultados de ANOVA (BMDP-7D), incluindo no modelo efeito de interação FALANTE X VOGAL.

Podemos verificar através do exame da tabela 8.6 que existe um efeito significativo da interação FALANTE X VOGAL, embora os efeitos de FALANTE e VOGAL, isoladamente, sejam bem maiores, dentro das expectativas. A existência de interação indica que o grupo de falantes não é homogêneo quanto ao comportamento de F5/n/ em função da qualidade da vogal V_2, ou seja, é possível que, a exemplo dos resultados obtidos em Su *et al.* (1974), o fenômeno da co-articulação antecipatória ocorra mais intensamente em alguns sujeitos, uma informação potencialmente útil para a determinação da identidade do falante.

A tabela 8.7 apresenta as médias e desvios-padrão de cada célula de FALANTE X VOGAL, assim como os resultados de um teste *t* (BMDP-3D) comparando as distribuições dos dois níveis de VOGAL (ANTERIOR *vs.* POSTERIOR). Um valor estatisticamente significativo de *t* indica que, para esse falante específico, há uma diferença consistente no espectro nasal em função da qualidade da vogal imediatamente após /n/, isto é, esse falante poderia ser classificado genericamente como um "bom co-articulador".

Fal.→	ZR	EN	R1	ZP	AG	WA	MS	R2	DO
m.A.	2685.7	2762.8	2625.7	2600.0	2620.0	2576.0	2566.7	2597.1	2434.3
dp.A	62.9	155.9	80.6	169.7	115.2	92.1	30.1	109.8	74.6
m.P	2765.0	3042.5	2710.0	2660.0	2820.0	2873.3	2526.7	2740.0	2532.5
dp.P	107.8	184.7	59.5	87.2	150.8	132.5	92.7	88.2	51.2
n=(A)	7	7	7	2	6	5	6	7	7
n=(P)	8	8	8	5	8	6	6	8	8
t =	1.70	3.14	2.33	.660	2.57	4.22	-1.01	2.80	3.01
p<	.112	.008	.036	.539	.025	.002	.338	.02	.01

TABELA 8.7: Médias e desvios-padrão para cada falante em cada nível de VOGAL (ANTERIOR (A) *vs.* POSTERIOR (P)). A última linha mostra os resultados de um teste *t* (BMDP-3D) para verificar a significância do efeito de V_2 em F5/n/ em cada falante separadamente.

Pela tabela 8.7, verificamos, que os falantes ZR, ZP e MS não têm médias de F5/n/ significativamente diferentes em diferentes contextos fonéticos (qualidade de V_2), ou seja, esses falantes co-articulam pouco a nasal com a vogal subseqüente. Já os falantes EN, AG, WA, R1/R2 e DO apresentam um índice maior de co-articulação.

8.3.2) *Discussão*

O *corpus* limitado sobre o qual se baseia o experimento descrito nesta seção não permite, obviamente, uma avaliação segura da eficiência de medidas derivadas de fenômenos co-articulatórios. No entanto, a constatação de padrões individuais mesmo com o pequeno número de dados aqui utilizado sugere que esse é um campo potencialmente fértil para futuras explorações envolvendo um *corpus* mais amplo e balanceado.

A utilização de fenômenos co-articulatórios com a finalidade de determinar a identidade de falantes já foi explorada com sucesso em Nolan (1983:77ff),

focalizando a magnitude da variação dos formantes de laterais em função da qualidade da vogal seguinte. Nolan, quantificando o grau de co-articulação através de coeficientes de correlação não paramétricos (*Spearman* - r), constata que esses coeficientes têm um âmbito extenso de variação para um grupo de 15 falantes do mesmo dialeto.

Outras combinações CV poderiam ser também exploradas, especialmente aquelas onde a configuração articulatória favorece efeitos de co-articulação. Uma possibilidade aqui, por exemplo, são as consoantes apicais. O volume da cavidade à frente da língua na produção de /d/ em seqüências /dV/ varia consideravelmente em função da vogal, devido à co-articulação antecipatória (Sundberg e Lindblom 1990), um aspecto cujas conseqüências acústicas podem ser relevantes para a Identificação de Falantes.

Efeitos de co-articulação atendem, a princípio, algumas das condições estabelecidas em Wolf (1972) para a seleção de parâmetros ideais na Identificação de Falantes: (a) ocorrem freqüentemente na fala normal, (b) a variação intra-falante é menor do que a inter-falante e (c) não são facilmente passíveis de modificação consciente por parte do falante (resistência a tentativas de disfarce). Em contrapartida, o emprego desse tipo de parâmetro requer um número razoável de amostras, já que a avaliação do grau de co-articulação só pode ser estabelecida por meio de um critério estatístico.

Outra dificuldade relaciona-se com a própria mensuração das quantidades básicas. Medidas dependentes da detecção de formantes possuem um certo grau de incerteza, especialmente em laterais e nasais, onde o aparecimento de picos espectrais espúrios não é raro. Uma forma de contornar esse problema seria representar os espectros de amplitude como vetores *n*-dimensionais baseados em saídas de bancos de filtros, cuja média neutralizaria eventuais variações locais aleatórias. É evidente, no entanto, que essa é uma solução que envolve um certo

compromisso, pois à medida que se perde informação espectral fina podem também ser perdidas, a princípio, fontes de variação inter-subjetiva potencialmente relevantes.

8.4) *Aspectos do Espectro Nasal em Gêmeos Idênticos*

Um dos argumentos a favor do emprego de consoantes nasais para a Identificação de Falantes vem da presumida associação direta entre a configuração espectral desses sons e a estrutura individual das cavidades nasais. Sendo a forma dessas cavidades não mutável (excetuando estados transientes anormais, como resfriados), espera-se uma certa estabilidade do espectro, desde que mantidas as mesmas condições quanto ao ambiente fonético.

Ao estudar a variação espectral de consoantes nasais em vários falantes, é praticamente impossível separar as influências da anatomia individual das cavidades dos efeitos de co-articulação, efeitos esses que, como já vimos acima, variam em magnitude de falante para falante. Por outro lado, no caso de gêmeos monozigóticos, espera-se que as características anatômicas do trato nasal e oral sejam praticamente idênticas. Assim, qualquer diferença acústica entre os gêmeos, na produção de uma consoante nasal, deve estar relacionada a condições articulatórias e não à conformação física do trato.

A figura 8.2 mostra o espectrograma no tempo e o espectro de seção (BW=300 Hz) extraído na região central da nasal /n/ demarcada pelos cursores, para cada um dos gêmeos idênticos (JA e JR) durante a produção do enunciado (leitura fluente) *O Santos entra em campo amanhã no Maracanã preparado para enfrentar uma verdadeira guerra*. O trecho mostrado na figura 8.2 são as duas últimas sílabas de "Maracanã".

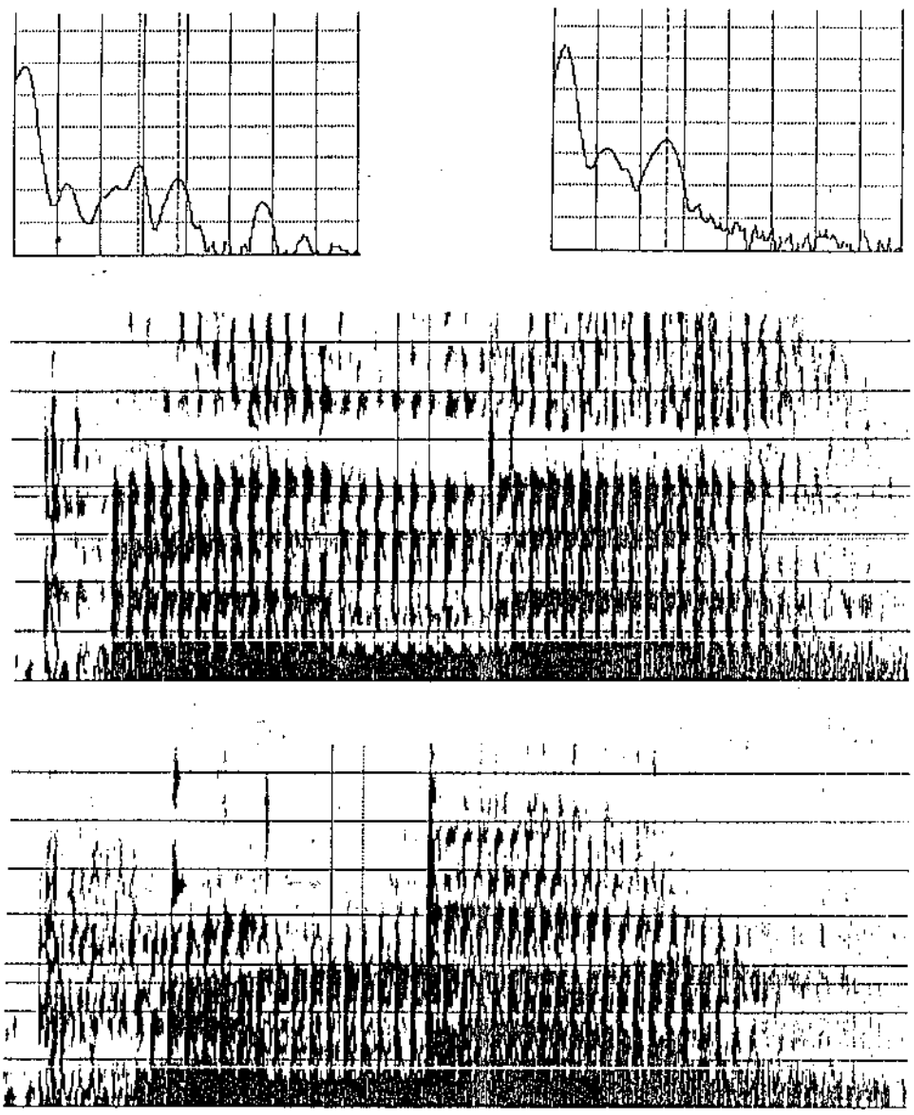


FIGURA 8.2: Comparação dos espectros nasais dos gêmeos JA (espectrograma inferior, espectro de seção à direita) e JR.

Apesar da presumível identidade entre as anatomias particulares podemos observar, na figura 8.2, que há diferenças bastante evidentes entre os espectros dos gêmeos JA e JR. No espectro de JR aparecem 2 picos nítidos em 2880 e 3800 Hz, assinalados pelos cursores, enquanto para JA existe apenas um pico espectral nessa região, situado em 2600 Hz. A região acima de 3000 Hz do espectro de JA apresenta um forte efeito de *damping* que achatou completamente o espectro nessa faixa, enquanto no espectro de JR há dois picos bem visíveis em freqüências mais altas, um em 5800 Hz e outro em 6720 Hz.

Os picos em 2880 e 3800 Hz no espectro de JR podem ser interpretados como F5 e F6 de /n/, já que não há expectativa de um zero nessa região que pudesse estar associado ao vale que se observa entre esses dois picos (Cf. Fujimura 1980:1871). A existência de um único pico, em 2600 Hz, no espectro de JA, não é de fácil interpretação De um modo geral não há expectativa de variação em F5 /n/, já que este formante está relativamente distante do âmbito de influência do anti-formante, sendo mais diretamente relacionado a características do falante (Fujimura 1980:1874). Outra possibilidade é que o pico em 2600 Hz no espectro de JA esteja relacionado com F4 e não com F5. Comparando os espectros de JA e JR podemos verificar que o anti-formante mais importante parece mais alto em JA (observar o vale em ≅ 2000 Hz), em comparação com JR (observar o vale em ≅ 1700 Hz); a presença de um zero mais alto "empurraria" F5 /n/ para freqüências mais altas do que o esperado. O efeito isolado do zero, no entanto, não deveria causar um deslocamento tão grande em F4 /n/, já que em consoantes nasais antecendo vogais [+ posterior] o zero em questão tende a ser de freqüência relativamente mais baixa (Fujimura 1980:1871). Assim, é possível que o pico em 2600 Hz no espectro de JA, seja um efeito combinado de F4 e F5, este último tendo sido abaixado em virtude do forte *damping* em freqüências acima de 4000 Hz observado nesse falante; a largura

de banda consideravelmente larga do pico em questão reforça a hipótese de que possa ter havido uma integração de F4 e F5.

Tanto em JA quanto em JR, observamos um pico espectral na região de ≅ 1200 Hz, certamente correspondendo a F2 /n/. Esse formante está suficientemente distante do anti-formante para sofrer sua influência e pode ser diretamente associado à ressonância natural do trato nasal, sendo primariamente determinado pelas suas dimensões (Fujimura 1980:1874). A mesma posição de F2/n/ para JA e JR está dentro das expectativas, em função das configurações anatômicas semelhantes dos gêmeos idênticos.

Se as anatomias dos gêmeos JA e JR são semelhantes [5], as diferenças observadas entre os espectros nasais só podem estar relacionadas a posturas articulatórias distintas. A configuração espectral de consoantes nasais é bastante sensível ao *timing* do movimento do *velum* e à posição específica da língua, que podem alterar significativamente as razões entre as aberturas dos subsistemas ressoadores envolvidos (faringe, cavidade oral e trato nasal), sendo essas razões o principal determinante das posições relativas de polos e zeros na função de transferência (Pickett 1980; Fujimura 1980; Laver 1980). Nesse sentido é interessante observar, no falante JA, que o efeito de *damping* em freqüências altas inicia-se já na vogal que antecede a consoante nasal /n/ (ver espectrograma no tempo, figura 8.2); é possível que, para esse falante, a abertura do *velum* seja mais antecipada do que em JR, tendo como conseqüência um pórtico para a cavidade nasal mais amplo, explicando algumas das diferenças espectrais em relação a seu irmão gêmeo.

O exemplo aqui estudado, embora limitado, indica que a configuração do espectro nasal tem uma relação fraca com as características anatômicas particulares, sugerindo maior cautela quanto ao emprego de parâmetros derivados do espectro de

consoantes nasais. A interação entre polos e zeros na produção de nasais é influenciada por uma série de variáveis articulatórias, fazendo com que a relação entre o espectro observado e a função de transferência seja complexa, dificultando assim uma interpretação que pudesse remeter a saída acústica, sem ambigüidade, a características anatômicas estáveis do falante.

SEÇÃO 9: VOT (Voice Onset Time)

9.1) *Introdução*

Ao definir os principais requisitos na escolha de parâmetros acústicos eficientes para identificar falantes, Wolf (1972) enfatiza aquelas medidas relacionadas a aspectos da fala cujo mecanismo articulatório subjacente seja transparente para o falante, ou seja, ações articulatórias sobre as quais o falante não possa exercer controle consciente. Parâmetros dessa natureza, observa Wolf, seriam mais resistentes a tentativas de disfarce e imitação. Como exemplo de uma medida que atende a essa condição, Wolf cita a duração do pré-vozeamento em uma plosiva sonora precedida por um segmento não vozeado.

O *timing* relativo entre a soltura de uma plosiva e o momento em que se inicia o movimento periódico das cordas vocais é normalmente chamado de *voice onset time* (VOT), e é uma das pistas mais importantes, embora não exclusiva, para a distinção de consoantes sonoras e não-sonoras (Lisker e Abramson 1964; Klatt 1975; Ladefoged 1975:124ff). Atribuindo tempo zero ao momento da soltura, observa-se que, nas plosivas sonoras precedidas de segmentos não vozeados, o VOT geralmente é negativo, ou seja, há um pré-vozeamento antes da soltura da oclusão da consoante. Nas plosivas não sonoras, em qualquer posição, há um período de tempo sem atividade das cordas vocais, que só começam a vibrar após a soltura; assim, o VOT é positivo.

A distinção do traço de sonoridade, colocada em termos do sinal positivo ou negativo do VOT, só é válida, entretanto, para a posição medial, já que em posição inicial tanto plosivas sonoras quanto não-sonoras são geralmente produzidas com intervalos silenciosos de oclusão (Lisker e Abramson 1964:384). Assim, o emprego

do VOT de plosivas sonoras, tal como sugerido por Wolf (1972), tem como primeira limitação a variação da medida em função da posição da consoante. Não encontramos na literatura especializada estudos focalizando diretamente a eficiência do VOT para a Identificação de Falantes, onde pudéssemos avaliar a variação intra-falante da medida. Outros estudos, no entanto, podem servir de base para uma primeira avaliação. Lisker e Abramson (1964) apresentam um conjunto de medidas de VOT de plosivas sonoras e não-sonoras a partir de 11 línguas diferentes; os resultados desse estudo indicam que, independentemente da língua, para uma determinada oposição sonora/não-sonora, a faixa de variação intra-falante do VOT é, em termos absolutos, geralmente maior para a plosiva sonora, mesmo sendo mantidas as mesmas condições de ambiente fonético. Esses resultados sugerem que as plosivas não-sonoras podem fornecer informação mais consistente do que as sonoras no que diz respeito a identidade do falante, ao contrário do que sugerira Wolf (1972).

9.2) *Material e Métodos*

De modo a obter uma estimativa da variação do VOT de plosivas não-sonoras para o grupo de falantes aqui estudado, realizamos uma série de medidas baseadas na sílaba /tO/ na palavra "laboratório". Cada falante do grupo principal (n=9: 7 + R1/R2) leu, em velocidade normal, as frases abaixo, onde a palavra "laboratório" aparece em diferentes posições:

1) **Laboratório** de Fonética Acústica
2) O **Laboratório** provavelmente chegou a uma parametrização
3) Os condicionadores do anteprojeto previam o **Laboratório**
4) As interações no **Laboratório** criavam condicionadores

A frase (1) foi lida 6 vezes e as demais 2 vezes cada, totalizando 12 medidas de VOT para cada falante. As produções de cada uma das frases (1-4) foi intercalada pela leitura de outros textos de modo a evitar que qualquer uma das frases fosse produzida em seqüência (consigo mesma ou com qualquer uma das demais).

Lembramos que, no caso do falante R1/R2, duas amostras foram colhidas com defasagem de alguns meses, sendo que na primeira (R1) o falante encontrava-se sob forte estado gripal, provavelmente com inflamação no nível da laringe. Assim, os resultados para esse falante poderão servir como indicadores para a estabilidade do VOT em amostras não contemporâneas e, acumulativamente, em função da presença de alterações orgânicas transientes.

De acordo com a definição de VOT dada por Lisker e Abramson (1964:390), segundo a qual o VOT é o

> *interval between the release of the stop and the onset*
> *of the glottal vibration,*

realizamos as medidas, diretamente na forma de onda, tomando como ponto inicial a descontinuidade característica da soltura da plosiva e como ponto final o primeiro pulso periódico nítido, correspondendo ao início da vibração glotal. A figura 9.1 mostra 3 exemplos do procedimento, em amostras dos falantes ZR, R1 e AG.

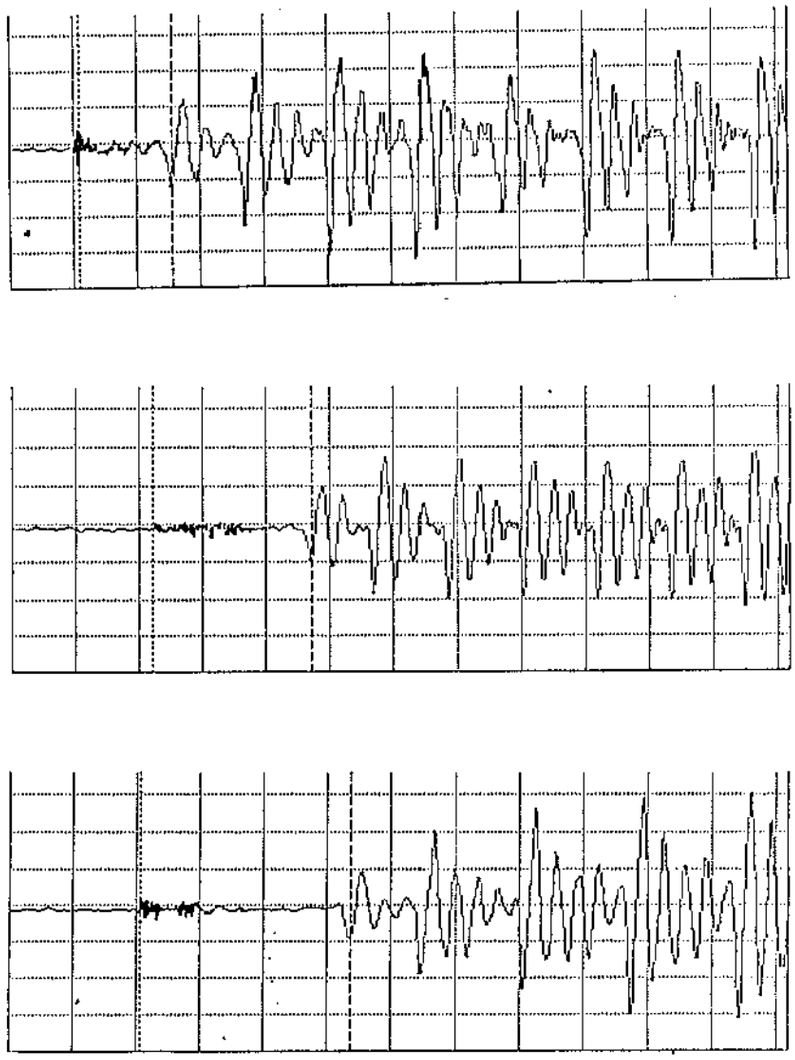


FIGURA 9.1: Medidas de VOT obtidas diretamente na tela do DSP-5500 da KAY Elemetrics. O primeiro cursor (linha pontilhada) marca o momento da soltura da oclusão do /t/, e o segundo cursor (linha tracejada) assinala o início do movimento periódico das cordas vocais, ou seja, o *onset* vocálico. Os gráficos referem-se a amostras dos falantes ZR (superior), R1 (meio) e AG (inferior).

9.3) *Resultados*

A tabela 9.1 apresenta as médias e desvios-padrão dos VOT para cada falante separadamente e para o total de falantes, além dos valores máximos e mínimos observados e a extensão da variação. A tabela mostra também o resultado do teste de comparação de médias *Student-Newman-Keuls* (BMDP-7D), unindo através de linhas horizontais os falantes cujas médias de VOT **não** são significativamente diferentes entre si. A figura 9.2 complementa a tabela 9.1, apresentando, graficamente, as faixas de variação individuais.

Fal.→	ZR	EN	R1	ZP	AG	WA	MS	R2	DO	Tot.
med.	10.87	18.03	14.70	14.99	20.41	16.42	18.16	14.21	17.33	15.95
D.P.	1.11	.89	1.30	.33	.97	1.34	1.25	.80	.80	2.96
min.	9.18	16.80	13.48	14.65	19.14	15.43	16.99	13.67	16.60	9.18
max.	11,72	19.14	16.02	15.43	21.48	18.36	19.34	14.65	18.36	21.48
var.	2.54	2.34	2.54	.78	2.34	2.93	2.35	.98	1.76	12.30
Teste *Student-Newman-Keuls*										
ZR	**R2**	**R1**	**ZP**	**WA**	**DO**	**EN**	**MS**	**AG**		

TABELA 9.1: Média, desvio-padrão, valores mínimo e máximo e faixa de variação das medidas de VOT, para cada falante separadamente e para o total de falantes. A linha inferior mostra o resultado de um teste de comparação de médias; os falantes unidos por linhas horizontais **não** são estatisticamente diferentes entre si, para o conjunto de medidas em questão.

Podemos observar na tabela 9.1 que a faixa total de variação inter-falante é de 12.30 ms. A ordem de grandeza dessa variação recomenda algumas considerações quanto à possibilidade de eventuais erros de medida. Durante a realização das medidas de VOT, verificamos que o momento exato da descontinuidade relacionado com a soltura da plosiva estava sempre bem definido na forma de onda e podia ser determinado com maior precisão dessa forma do que através do espectrograma no tempo. A determinação do primeiro pulso glotal, por outro lado, é um pouco mais

crítica, já que não existe aí uma descontinuidade abrupta, e o primeiro pulso pode ser bastante débil em amplitude, um aspecto já observado também em Lisker e Abramson (1964:416). Para um F0 de cerca de 125 Hz, a diferença de um pulso acarretaria um erro de 8 ms no VOT medido, uma distorção grande demais para a ordem de grandeza do parâmetro em questão. Para o conjunto de medidas que tomamos não existem, aparentemente, valores espúrios devidos a erros dessa natureza, já que o desvio-padrão intra-falante é sempre menor que 1.5 ms (ver tabela 9.1) [1].

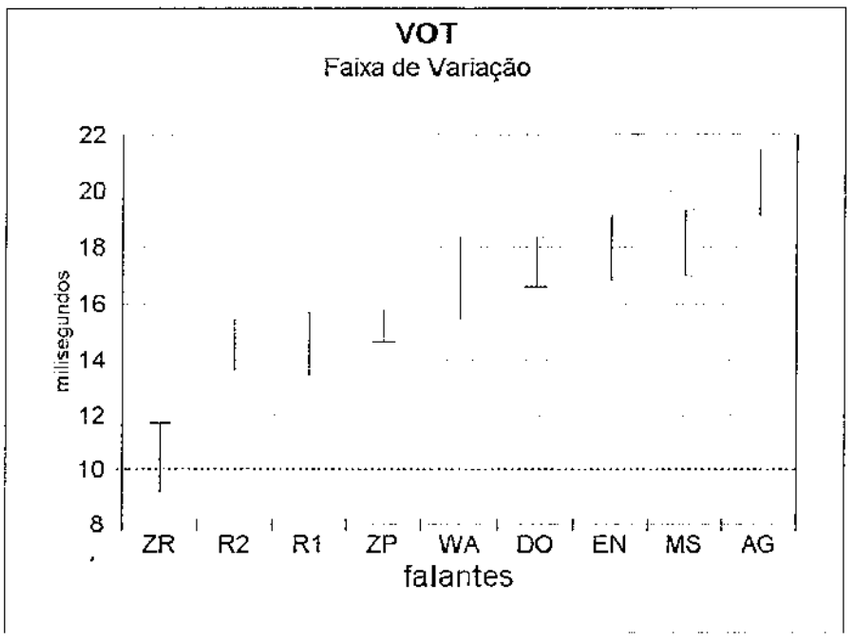


FIGURA 9.2: Faixas de variação de VOT para cada falante.

O teste *Student-Newman-Keuls* (tabela 9.1) indica que apenas os falantes ZR e AG, com as médias mínima e máxima do grupo, respectivamente, são significativamente diferentes de todos os demais falantes, para o conjunto de medidas de VOT obtido. Os grupos [R1/R2, ZP] e [WA, DO, EN, MS] são distintos

dos demais, mas cada um dos elementos do grupo não difere dos outros elementos do mesmo grupo.

Os valores semelhantes para as duas amostras não contemporâneas do falante R1/R2 sugerem que a medida é relativamente estável, ao menos se mantidas as mesmas condições de contexto fonético. O forte estado gripal que acometia o falante durante a primeira coleta (R1) não influenciou as medidas de VOT, ao contrário do que aparentemente ocorreu com alguns parâmetros examinados anteriormente (F0, F4, espectro da nasal /n/; ver seções 4, 5 e 8). É possível, como já discutimos anteriormente, que essas alterações tenham relação com o estado inflamatório no nível da laringe, especialmente no que diz respeito a F0, que apresentou uma média significativamente menor na amostra onde existia o estado gripal (R1), uma conseqüência, provavelmente, da variação de massa e consistência dos tecidos na região da glote e/ou de uma diminuição da pressão trans-glotal em função da diminuição da capacidade pulmonar. Esse conjunto de fatores não é compatível com a afirmação de Löfqvist (1992), que atribui a alguns aspectos relacionados à glote (área glotal, pressão trans-glotal e tensão) uma função importante no comportamento do VOT. A ausência de variação significativa nas amostras do falante R1/R2 sugere que o VOT é determinado principalmente pelo *timing* inter-articuladores, pelo menos em línguas sem aspiração.

Sendo um fenômeno diretamente relacionado ao controle do *timing* articulatório, espera-se que o VOT seja sensível a variações na velocidade de fala. Com efeito, esse aspecto já foi comentado em alguns estudos. Klatt (1975), embora não tenha controlado a taxa de articulação em seu experimento, verifica que os menores valores de VOT correspondiam ao falante de fala mais rápida. Também Lisker e Abramson (1964), comparando os VOTs extraídos de palavras isoladas com medidas extraídas de fala fluente, observam uma diminuição dos valores absolutos

no segundo caso (embora os valores relativos correspondentes à distinção sonora/não-sonora tenham se mantido).

No nosso experimento não houve controle da velocidade de fala na produção das frases (1-4), mas é possível estabelecer uma relação a partir das médias individuais obtidas na leitura do texto I (ver anexo). A figura 9.3 confronta o valor médio do VOT com a taxa de articulação média de cada falante (condição velocidade normal de produção). Podemos verificar que não há um padrão bem definido relacionando VOT e taxa de articulação média. Na verdade, parece haver para alguns falantes uma tendência oposta à relatada em Klatt (1975), ou seja, o falante AG, que é o mais rápido na condição velocidade normal, apresenta o maior valor médio de VOT, enquanto o falante ZR, com uma taxa de articulação bem mais baixa do que AG, tem o VOT médio mais curto.

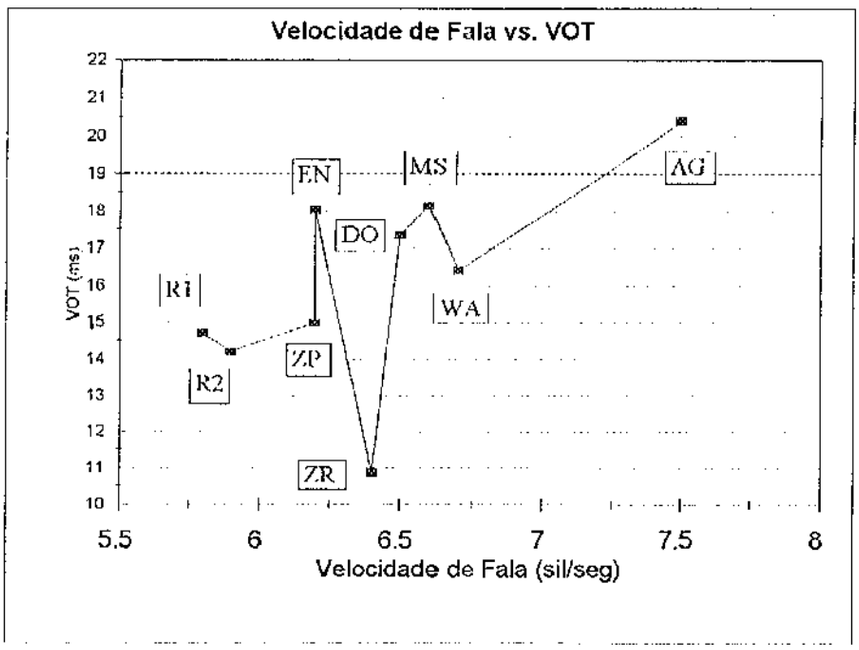


FIGURA 9.3: Velocidade de emissão média (normal) *versus* VOT, para cada falante.

É evidente que as tendências observadas na figura 9.3 nada dizem a respeito de uma possível variação do VOT em função de alterações da velocidade de produção **intra-falante**, já que esse aspecto não foi contemplado em nosso experimento. Os resultados, no entanto, sugerem que não existe uma relação genérica simples, inversamente proporcional e independente de falante, entre o VOT e a velocidade de fala, ou seja, as modificações no *timing* inter-articuladores em função da variação da velocidade de emissão não se dão de forma homogênea para todos os eventos articulatórios e/ou para todos os falantes. Assim, diferentes falantes alterariam sua velocidade de emissão implementando diferentes estratégias articulatórias (maior alteração no VOT do que nas durações vocálicas, por exemplo). No plano da variação intra-falante, há evidências de que as alterações de velocidade de emissão não implicam uma reparametrização linear de um programa motor genérico, atuando igualmente em todos os tipos de segmento, isto é, as razões entre intervalos inter-articulatórios não permanecem constantes, independentemente de condições de velocidade (Gay 1981; Sternberg *et al.* 1988; Löfqvist 1991; Adams e Kent 1993). No que diz respeito à variação inter-falante, já se verificou, através de cine-radiografia, que a relação entre os padrões de deslocamento articulatório e velocidade de emissão é dependente do falante (Kuehn e Moll 1976). Esse aspecto já foi experimentalmente verificado no fenômeno do *target undershoot*, onde se observa que o grau de redução vocálica depende do falante (Lindblom e Moon 1988).

9.4) *Comentário Final*

Os resultados aqui obtidos indicam que o VOT é um parâmetro potencialmente útil para a Identificação de Falantes. No entanto, assim como alguns aspectos acústicos já abordados, é mais um indicador genérico do que um

determinante de identidade. Informações dessa natureza não deixam de ser válidas, especialmente na aplicação forense, onde a evidência de identificação negativa (exclusão de suspeitos) também cumpre um papel relevante. Assim, diferenças inter-individuais acima de um determinado limiar, sem qualquer sobreposição das distribuições - tal como ocorre com os falantes ZR e AG - devem ser consideradas, ao menos como evidência complementar.

Uma avaliação mais detalhada da eficiência do VOT para a Identificação de Falantes escapa ao escopo do presente experimento, e passa por um estudo das variações do VOT em função de várias condições, como contexto fonético, velocidade de fala, duração do enunciado, etc, fatores que, sabemos, influenciam o parâmetro em questão (Cf. Klatt 1975). Outros fatores que possam alterar os processos neuro-musculares e, conseqüentemente, o controle fino da coordenação inter-articuladores, tais como a ingestão de bebidas alcoólicas, terão, provavelmente, efeitos sensíveis nas medidas de VOT (Cf. Johnson *et al.* 1990:217).

SEÇÃO 10: ABORDAGEM ESPECTROGRÁFICA

10.1) *Introdução*

Vimos na seção 2.1 que Hecker (1971) admite três métodos básicos para a Identificação de Falantes, que denominamos abordagens *perceptual, automática* e *espectrográfica*. Discutiremos na presente seção a eficiência da inspeção visual de espectrogramas no tempo, examinando os resultados de diversos experimentos de laboratório focalizando a questão.

Há uma série de mal-entendidos envolvidos na avaliação do espectrograma como evidência em casos forenses, sendo o mais difundido a analogia feita entre esse tipo de padrão e as impressões digitais (v. p. ex. Gocke e Oleniewski 1973; Anghelescu 1974; Cunha Lima 1976; Smrkovski 1981). Essa perspectiva é equivocada, na medida em que o sinal de fala não reflete diretamente características anatômicas, como já discutimos mais extensamente na seção 1 (v. também Bolt *et al.* 1970:606ff). Parte do mal-entendido vem do entusiasmo exagerado da mídia com relação à evidência espectrográfica, quase sempre apresentada como prova "concreta" e "inequívoca" da identidade de um falante, uma informação que, tal como é veiculada, esconde a verdadeira natureza das técnicas efetivamente utilizadas em um exame competente de identificação de falante.

O espectrograma tradicional é, de fato, um instrumento poderoso para a determinação da identidade de um falante, como aliás também tem sido, há longo tempo, para a análise fonético/fonológica de diversos aspectos da fala. O engano é considerar um espectrograma como um mero padrão visual, ignorando as

especificidades dos complexos processos articulatórios subjacentes. Embora em situações controladas de laboratório se deva esperar uma certa estabilidade nos padrões espectrográficos em enunciados do mesmo falante, o mesmo pode não ocorrer em casos forenses reais, onde a presença de ruído e outras distorções podem interferir substancialmente no espectrograma. Nessas situações, se o examinador não possuir um bom conhecimento dos mecanismos articulatórios, e basear sua decisão apenas na configuração gráfica, é muito provável que cometa erros grosseiros. Hollien (1990:215) relata um desses casos, onde um "perito" sem conhecimentos lingüísticos baseia seu parecer em um aspecto espectrográfico que nada mais era que um simples *click*, provavelmente produzido pelo sistema de gravação (o "perito" avaliou as vozes como não-idênticas em função da inexistência da barra vertical associada ao *click* na gravação de confronto, enquanto a mesma aparecia no enunciado questionado).

Embora possa parecer estranho que se proponha a mera comparação visual de espectrogramas, sem referência às dimensões lingüísticas, como um método válido de identificação, esse foi o paradigma utilizado em diversos experimentos. Os altos índices de acerto relatados inicialmente por alguns desses estudos (Kersta 1962, Tosi *et al.* 1972, entre outros; v. seção 10.2) levaram, perigosamente, à generalização de que técnicas semelhantes poderiam ser empregadas também nos casos forenses, desconsiderando as diferenças significativas que podem existir entre a situação laboratorial e os casos da "vida real".

É importante ressaltar que não consideramos a evidência espectrográfica destituída de valor para o modelo forense. Na verdade, as objeções ao uso de técnicas de inspeção visual de espectrogramas concernem mais aos procedimentos de decisão e à natureza da informação sobre a qual a decisão se baseia do que ao potencial informativo do espectrograma convencional. Como já frisamos acima, o

espectrograma é altamente eficiente, na medida em que concentra várias fontes de informação acústica em um só gráfico. Atualmente, a flexibilidade dos modernos equipamentos digitais, permitindo o ajuste rápido de várias características do *display*, facilita a observação de aspectos "microscópicos", que podem ser de grande relevância para a determinação da identidade de um falante.

Não acreditamos que qualquer foneticista encarregado da análise de um caso forense de identificação prescinda da informação espectrográfica, embora não precise (e nem deva) basear sua decisão apenas nesse tipo de evidência. Nossa experiência tem demonstrado que o exame comparativo de espectrogramas freqüentemente revela similaridades bastante significativas, especialmente no que diz respeito ao *timing* articulatório (v. seção 10.3). A avaliação dessas similaridades deve, no entanto, ser feita à luz de uma interpretação fonético/fonológica, sempre em contraponto com a monitoração auditiva e um profundo conhecimento do âmbito de variação do aspecto focalizado.

10.2) *A Eficiência da Inspeção Visual de Espectrogramas na Identificação de Falantes*

Durante a Segunda Guerra Mundial surgiu um grande interesse estratégico na pesquisa envolvendo a Identificação de Falantes, com o objetivo de monitorar transmissões militares de rádio. O termo *voiceprint* (em analogia a *fingerprint*) foi cunhado nesse contexto por Gray e Kopp (1944; *apud* Tosi *et al.* 1972:2031), dois pesquisadores dos Laboratórios Bell; afirmava-se então que a inspeção visual de espectrogramas seria potencialmente eficaz para determinar a identidade de um falante. Após o término da Segunda Guerra, sem que a pesquisa atingisse resultados efetivos, a questão foi praticamente esquecida, até que, no final dos anos cinqüenta,

o Departamento de Polícia de Nova Iorque se viu às voltas com uma série de casos envolvendo ameaças de bomba em aviões, feitas por via telefônica. Em função dessa nova demanda, foi solicitado aos Laboratórios Bell que desenvolvesse um método para identificar falantes através de análise espectrográfica. A tarefa ficou a cargo de Lawrence G. Kersta, um físico com alguns anos de experiência na análise de fala por espectrogramas. Kersta conduziu um programa de pesquisa de dois anos, usando uma população de 123 falantes (Inglês Americano) do sexo masculino. Um grupo de 9 estudantes do segundo grau foi treinado para realizar a tarefa de identificação, com base apenas na inspeção visual de espectrogramas, sem qualquer tipo de monitoração auditiva. Ao longo desses dois anos, cerca de 50000 tentativas de identificação foram realizadas, a partir de 16000 diferentes espectrogramas de 10 monossílabos freqüentes em Inglês. Testes fechados, onde os juízes deviam reunir em pilhas espectrogramas teste e referência dos mesmos indivíduos, atingiram índices de 99.6%, 99.2% e 99.0%, para sub-conjuntos de 5, 9 e 12 falantes, respectivamente. O emprego de palavras extraídas do contexto de uma sentença chave, ao invés de palavras produzidas em isolamento, diminuiu em apenas 0.2% o índice de acertos (Kersta 1962).

As altas taxas de acerto relatadas em Kersta (1962) não foram totalmente aceitas como evidência da eficiência do método de identificação por inspeção visual de espectrogramas. Na verdade, surgiram algumas dúvidas quanto a aspectos metodológicos, já que Kersta não explicitou procedimentos experimentais importantes para uma avaliação objetiva dos resultados. A pequena diferença entre o índice de acertos para palavras isoladas e palavras em contexto era particularmente discutível, já que outros estudos indicavam que, na identificação auditiva, essa diferença era bem maior, sinalizando uma variação significativa em função do ambiente fonético (Cf. Pollack *et al.* 1954; Bricker e Pruzansky 1966). Young e

Campbell (1967) reavaliaram o método originalmente proposto por Kersta, conduzindo um experimento mais controlado, com material semelhante ao utilizado em Kersta (1962). Young e Campbell utilizam 10 examinadores para identificar 5 falantes, com base em palavras produzidas em isolamento (monossílabos) e as mesmas palavras extraídas de diferentes contextos fonéticos. Os resultados indicaram que os examinadores tiveram uma dificuldade muito maior em identificar os falantes por meio de palavras produzidas em contexto do que por meio de palavras produzidas isoladamente, atingindo índices de acerto de 37.3% e 78.4%, respectivamente. Os autores atribuem essa diferença a dois fatores:

> *(1) ...the shorter duration [of the word spoken in context] results in the transmission of less acoustic information; (2) ... the phonetic environment of the test word interact with the word's acoustic representation*;

a conjunção desses dois efeitos, segundo os autores,

> *...outweigh any intratalker consistency for the repeated production of these short monosyllabic words* (Young e Campbell 1967:1253).

Os resultados desse estudo contrastam consideravelmente com os de Kersta (1962), não só quanto à diferença de desempenho entre as duas condições (0.2% em Kersta *versus* 41.1% em Young e Campbell), mas também quanto ao índice absoluto de acertos na identificação por palavras produzidas em isolamento para a mesma população (5 falantes), onde Kersta (1962) obtém 99.6%, contra os mais modestos 78.4% observados em Young e Campbell (1967). Os autores comentam que o menor índice de acertos em seu estudo pode estar relacionado a uma maior homogeneidade

do grupo de falantes, com respeito a sexo, dialeto, idade e educação; essa afirmação, entretanto, não pode ser devidamente avaliada, já que Kersta não oferece qualquer informação quanto à constituição do grupo de falantes que utilizou.

O paradigma experimental de Kersta (1962) não reproduz a situação forense. Em primeiro lugar, a comparação de palavras produzidas em isolamento nunca ocorrerá; teremos, em geral, a comparação de palavras produzidas em diferentes contextos (com sorte, algumas repetições de palavras e - com muito mais sorte - de alguns contextos parciais). Além disso, os testes de Kersta (e também os de Young e Campbell 1967) são fechados, isto é, o examinador **sabe** que cada amostra teste corresponde **necessariamente** a alguma amostra referência. No modelo forense, ao contrário, os exames devem pressupor que **qualquer pessoa,** a princípio, pode ser o "culpado", ou seja, trata-se de um teste aberto [1].

As imperfeições metodológicas dos experimentos envolvendo identificação através do exame visual de espectrogramas levou vários pesquisadores a duvidar da possibilidade de extrapolar os resultados assim obtidos para a situação forense real (Cf. Henessy e Romig 1971a;b; Bolt *et al.* 1970). Na esteira da polêmica assim criada, Tosi *et al.* (1972) realizaram um estudo mais amplo sobre o tema, considerando aspectos não observados anteriormente, de modo a aproximar as condições experimentais à situação forense real. Foram testados os efeitos de 7 variáveis na taxa de identificações corretas, cuja descrição segue abaixo:

1) *Número de Palavras-Chave*; dois níveis: 6 e 9 palavras, extraídas do conjunto {*it, is, on, you, and, the, I, to, me*}, escolhidas pela alta porcentagem de ocorrência no Inglês.

2) *Número de Enunciados*; três níveis: 1,2 ou 3 diferentes enunciados da mesma palavra chave.

3) *Condições de Gravação*; três níveis: (a) diretamente no gravador em ambiente sem ruído, (b) através de linha telefônica, sem ruído e (c) através de linha telefônica com adição de ruído branco (50 dB).

4) *Contexto das Palavras-Chave*; três níveis: (a) palavras produzidas em isolamento, (b) palavras em contexto fixo (são comparadas as mesmas sentenças produzidas pelo falante conhecido e o desconhecido) e (c) palavras em contexto aleatório (sentenças diferentes são comparadas).

5) *Número de Falantes*; três níveis: 10, 20 e 40 falantes.

6) *Variação Intra-Falante*; dois níveis: (a) amostras contemporâneas (os espectrogramas teste e referência são obtidos na mesma sessão de gravação) e (b) amostras não-contemporâneas (os espectroramas teste são obtidos em uma segunda sessão de gravação, no mínimo um mês após a primeira).

7) *Tipo de Teste*; dois níveis: (a) teste fechado (o examinador sabe que o falante "desconhecido" está entre os "conhecidos" e (b) teste aberto (o examinador não sabe se o falante "desconhecido" está ou não entre os "conhecidos").

Para o experimento foram utilizados 250 falantes aleatoriamente selecionados de uma população de 25000 falantes de Inglês Americano, sem marcas dialetais acentuadas. Um total de 29 examinadores participou do experimento; esse grupo foi treinado durante um mês, recebendo nocões gerais de fonética e treinamento específico para a tarefa proposta. Os principais resultados obtidos em Tosi *et al.* estão resumidos na tabela 10.1.

Examinando a tabela 10.1, verificamos que os índices de acerto são consideravelmente menores do que os relatados no estudo de Kersta (1962), mas ainda maiores do que os de Young e Campbell (1967); o treinamento mais extenso realizado por Tosi *et al.* pode ter influenciado, com relação à discrepância com Young e Campbell (1967). Algumas condições parecem ter um maior efeito no desempenho; o tipo de contexto, tamanho da população, tipo de teste e contemporaneidade das amostras apresentaram distinções estatisticamente significativas entre os diferentes níveis.

Condição ↓	9 palavras	6 palavras
Número de Enunciados:		
a) 1	91.29	89.71
b) 2	90.26	91.62
c) 3	92.49	92.39
Condições de Gravação:		
a) direto no gravador	92.42	91.41
b) telefone sem ruído	91.31	91.20
c) telefone com ruído	91.02	91.10
Tipo de Contexto:		
a) isolamento	95.77 *	93.83 *
b) contexto fixo	92.39 *	91.68 *
c) contexto aleatório	86.59	88.20
Número de Falantes:		
a) 10	93.30	93.78 *
b) 20	91.87 *	90.40
c) 40	89.58	89.52
Variação Intra-falante:		
a) amostras contemporâneas	92.51 *	95.13 *
b) amostras não-Contemporâneas	87.95	87.35
Tipo de Teste:		
a) fechado	94.48 *	94.31 *
b) aberto	90.14	89.71

TABELA 10.1: Resultados de Tosi *et al.* (1972). Os números referem-se aos percentuais de identificações corretas através da inspeção visual de espectrogramas, nas diferentes condições. Os asteriscos indicam que o nível especificado difere estatisticamente do(s) seguinte(s) (adaptado da tabela I, em Tosi *et al.* 1972:2037).

No que diz respeito à simulação do modelo forense, interessam os resultados que cruzam as condições compatíveis com esse modelo, ou seja: testes abertos, espectrogramas não contemporâneos e palavras em contexto (fixo ou aleatório). Acumulando essas condições, o índice de erros observado em Tosi *et al.* (1972) cresce consideravelmente, atingindo aproximadamente 19 %, sendo 6% erros de falsa identificação e 13% erros de falsa rejeição. Tosi *et al.* argumentam que, na situação forense, os índices reais seriam menores, já que, ao contrário do experimento, os examinadores não seriam forçados a tomar uma decisão, sem que estivessem seguros em sua avaliação. Considerando apenas os casos onde os examinadores auto-avaliaram positivamente sua decisão (74% das tentativas compatíveis com o modelo forense), Tosi *et al.* observam que o índice de erros diminui para apenas 7% (2% falsa identificação e 5% falsa rejeição).

Embora os resultados relatados em Tosi *et al.* (1972) tenham sido considerados pelos proponentes do método "*voiceprint*" como evidência sustentando a validade da técnica em casos forenses, a questão continuou controversa. O fato é que o estudo de Tosi *et al.*(1972) ainda ignora algumas variáveis que podem estar presentes no modelo forense, tais como tentativas de disfarce e imitação, alterações no estado psicológico do falante, diferentes tipos de distorção do sinal, etc (Cf. Bolt *et al.* 1973:532-3). Tosi *et al.* (1972) incluem, na verdade, diferentes condições de gravação em seu experimento (ver acima), mas os espectrogramas comparados são obtidos sempre nas **mesmas** condições. Ora, na situação forense real isso só ocorre quando é possível reproduzir exatamente as condições originais da gravação questionada - o que nem sempre é viável; a situação mais realista, não contemplada no experimento de Tosi *et al.*, seria comparar amostras produzidas em **diferentes** condições de gravação/transmissão.

Outro aspecto deficiente no experimento de Tosi *et al.* diz respeito às condições de contextualização da palavra chave. Tanto na condição contexto "fixo", quanto na condição contexto "aleatório", as palavras chave foram extraídas de sentenças *nonsense* lidas pelos falantes; resta saber se esse material guarda uma relação próxima com a fala espontânea. Com efeito, já se verificou que o emprego de amostras efetivamente extraídas de conversação espontânea (mesmas palavras chave de Tosi *et al.* 1972) aumenta drasticamente a taxa de identificações incorretas (Hazen 1973).

Bolt *et al.* (1973:533), levando em conta que a situação forense deve introduzir outras fontes de variabilidade intra-falante não observadas em Tosi *et al.* (1972), consideram os resultados (taxas de acerto) aí obtidos como

> *...artificial minima which are likely to increase when conditions depart from the laboratory situation in which the voice samples were recorded.*

As observações de Bolt *et al.*(1973) são pertinentes, mas é preciso considerar também a existência de alguns fatores que, fora do laboratório, podem contribuir para uma maior confiabilidade das identificações feitas com base na inspeção visual de espectrogramas. Na situação forense real o perito não sofre, a princípio, qualquer pressão quanto ao tempo que deve empregar para tomar sua decisão, e pode ter à sua disposição, para confronto, tanto material de fala quanto necessite (a limitação existe, obviamente, apenas quanto ao material questionado), enquanto no laboratório as decisões, por evidentes razões logísticas, devem ser tomadas em um tempo pré-determinado (apenas 15 minutos em Tosi *et al.*1972, independentemente do número de falantes) e com base em um número restrito de itens (apenas 6 ou 9 monossílabos em Tosi *et al.* 1972). A possibilidade de ter acesso a trechos de fala mais longos

pode ser fundamental no exame de espectrogramas, já que alguns aspectos relacionados ao *timing* só se manifestarão mais claramente em enunciados de uma certa duração (maiores, com certeza, que os monossílabos empregados nos experimentos aqui discutidos) (Cf. Stevens *et al.* 1968).

O grau de treinamento do examinador também é um fator relevante na presente discussão. Em geral, observa-se uma tendência a um maior número de identificações corretas em grupos de examinadores que tiveram a oportunidade de um treinamento mais extensivo (Cf. Black *et al.* 1973:536). Espera-se que no tratamento de casos forenses estejam envolvidos profissionais com conhecimentos mais sólidos e mais experiência do que os examinadores normalmente utilizados nos experimentos de laboratório (na maior parte dos casos estudantes secundários sem formação lingüística específica). Embora seja um aspecto não mensurável, a própria responsabilidade do perito profissional diante da tarefa deverá impor um grau de prudência que normalmente não existe na situação mais descompromissada do laboratório; o perito profissional não é **obrigado** a tomar uma decisão (identificação ou rejeição) se não estiver totalmente seguro em sua análise, lhe sendo facultado o direito de, sempre que houver dúvidas, eximir-se de emitir opinião conclusiva. Com efeito, o próprio experimento de Tosi *et al.*, como já vimos acima, revela que, mesmo com a limitação de tempo para a decisão (e a obrigatoriedade de tomar essa decisão), os índices de erro caem consideravelmente se são computados apenas os casos onde os examinadores avaliam positivamente suas decisões.

Por último, mas não menos importante, é fundamental ressaltar que os experimentos acima descritos não fornecem informação sonora para os examinadores. Essa é uma diferença crucial em relação à situação forense, onde o perito, obviamente, terá acesso a pistas auditivas durante o decorrer da análise,

auxiliando-o a localizar aspectos potencialmente relevantes para posterior exame espectrográfico.

10.3) *Espectrogramas: Alguns Exemplos*

Examinaremos a seguir exemplos espectrográficos extraídos do nosso *corpus*, assim como alguns espectrogramas utilizados em casos forenses reais nos quais tivemos oportunidade de participar. Nosso objetivo aqui não é fazer qualquer tipo de apologia ao método "*voiceprint*", mas apenas ressaltar a importância do exame de espectrogramas como um elemento **acessório** em um processo de Identificação de Falante.

A figura 10.1 mostra quatro espectrogramas extraídos de produções de dois falantes do grupo utilizado em nossos experimentos. Os espectrogramas representam a produção da palavra "Laboratório", no contexto "Laboratório de Fonética Acústica". As análises foram realizadas no DSP-500 da KAY Elemetrics, em uma faixa de 0 - 8000 Hz, com pré-ênfase (*high-shape*) e filtragem "banda larga" (300 Hz). Os dois espectrogramas à esquerda da figura 10.1 referem-se a duas produções não contemporâneas do mesmo falante (R1/R2; inferior:R1, superior:R2); lembramos que essas produções foram gravadas com um intervalo de cerca de 4 meses e que o falante encontrava-se sob forte estado gripal na primeira coleta (R1). À direita da figura 10.1 vemos dois espectrogramas contemporeâneos do falante MS, obtidos na mesma sessão de gravação com intervalo de alguns minutos, durante os quais MS realizou outras tarefas de leitura (textos, palavras isoladas, dígitos, etc), ou seja, embora tenham sido registradas na mesma sessão, as produções não são imediatamente consecutivas. Ressaltamos que os espectrogramas de R1/R2 e MS foram selecionados em função de sua **semelhança**, exatamente para destacar os aspectos que, apesar da aparente similaridade visual, são distintos entre os dois

falantes, em um exame mais detalhado. Para maior clareza, cada par de espectrogramas está alinhado com referência à soltura da plosiva /t/.

Um dos aspectos coincidentes nos espectrogramas de R1/R2 e MS é o *timing* (na verdade, esse foi o motivo de escolher esses falantes); as produções têm aproximadamente a mesma duração total e os eventos acústicos ocorrem aproximadamente nos mesmos pontos ao longo do enunciado. Podemos verificar, entretanto, que há uma série de diferenças mais locais entre os dois falantes, entre elas: (1) o falante MS apresenta um VOT mais longo na plosiva /t/; (2) o fonema /r/ em posição intervocálica na seqüência /ora/ é realizado como *tap* por R1/R2, envolvendo um momento de oclusão real, enquanto a produção de MS parece mais uma aproximante; (3) as posições e movimentos transicionais de F3 e F4 na sílaba inicial /la/ são bastante diferentes entre os dois falantes (mas semelhantes para cada falante); (4) as transições de F2 e F3 na sílaba /tO/ são diferentes para os dois falantes; (5) o falante MS apresenta "vazios" de energia nas regiões de 5000 Hz e 6000 Hz nos núcleos vocálicos das sílabas /bo/ e /tO/, respectivamente, inexistentes nos espectrogramas de R1/R2; (6) de um modo geral, o falante MS tem uma menor concentração de energia nas freqüências altas do que o falante R1/R2.

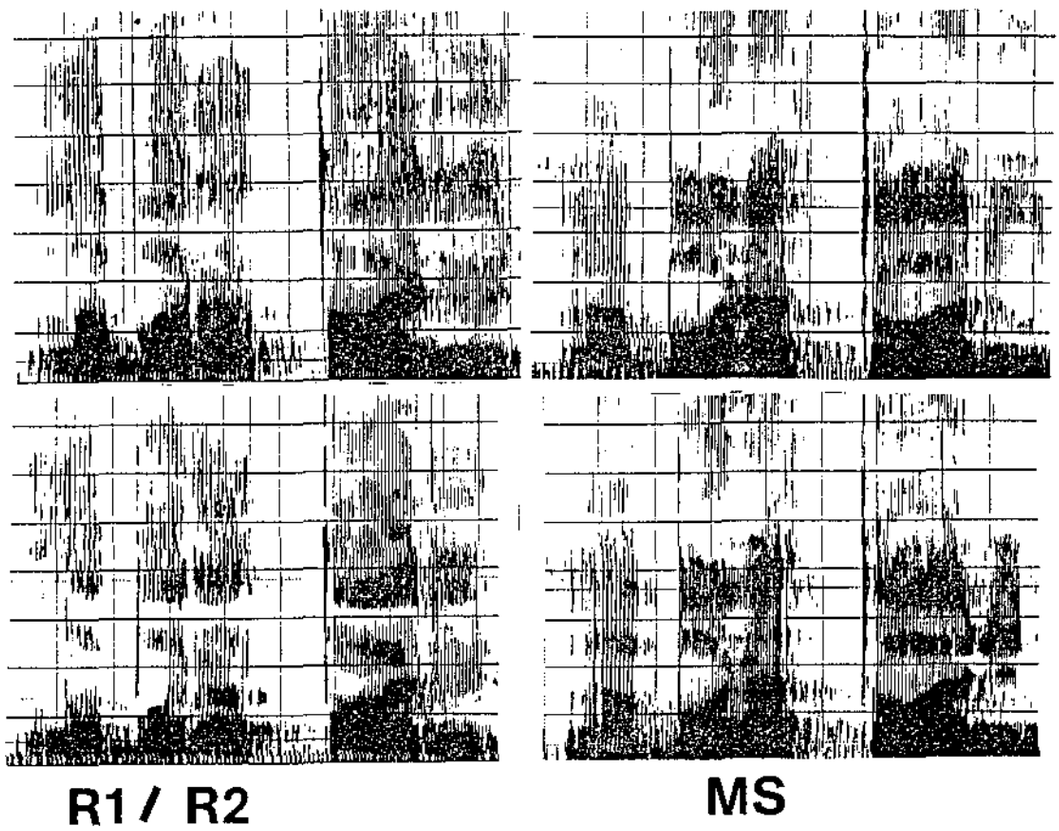


FIGURA 10.1: Espectrogramas dos falantes R1/R2 (esquerda) e MS referentes à produção da palavra "Laboratório". Os espectrogramas foram realizados no DSP-5500 da KAY Elemetrics, na faixa de 0 -8000 Hz, com pré-ênfase (*high-shape*) e filtragem "banda larga" (300 Hz).

A título de comparação, a figura 10.2 mostra os espectrogramas de outros falantes do grupo (ZP, AG e EN, de baixo para cima), produzindo a mesma palavra "Laboratório", com os mesmos parâmetros de análise no DSP-5500 utilizados para produzir os espectrogramas da figura 10.1. Os espectrogramas estão alinhados pelo ponto de soltura da plosiva /t/. Observa-se aqui uma maior variabilidade inter-falante quanto aos padrões espectrográficos, tanto no eixo de tempo quanto no da freqüência.

Observe-se, ainda com relação à figura 10.1, que apesar da diferença de cerca de quatro meses entre as amostras R1 e R2, as características básicas dos espectrogramas permanecem semelhantes. É evidente que, nesse caso, a igualdade de contextos (lingüístico e situacional) contribui para a diminuição da variabilidade intra-falante. Na situação forense nem sempre é possível reproduzir exatamente o contexto original, especialmente no que diz respeito aos componentes afetivo/emocionais. O sucesso na coleta de material para confronto, em casos forenses, depende em grande parte da experiência e habilidade do profissional responsável pela gravação, que deve ser capaz, na medida do possível, de criar uma ambiência menos formal durante a gravação com o(s) suspeito(s) (v. seção 1.2). Outro aspecto importante é a construção do texto que será apresentado ao suspeito; esse texto deverá conter trechos relevantes (palavras ou frases selecionadas através de uma análise prévia do material questionado), em contextos semelhantes aos da gravação questionada. Preferencialmente, não deve ficar evidente para o suspeito quais são os aspectos relevantes para o perito; uma estratégia eficiente, que temos utilizado em casos onde participamos, é elaborar textos que simulem notícias de periódicos ou trechos de ficção, envolvendo temas totalmente distintos daqueles presentes na gravação questionada, mas contendo no seu interior, dissimuladamente,

os trechos de interesse pericial (procedimento análogo é freqüentemente utilizado também na perícia grafotécnica) [2].

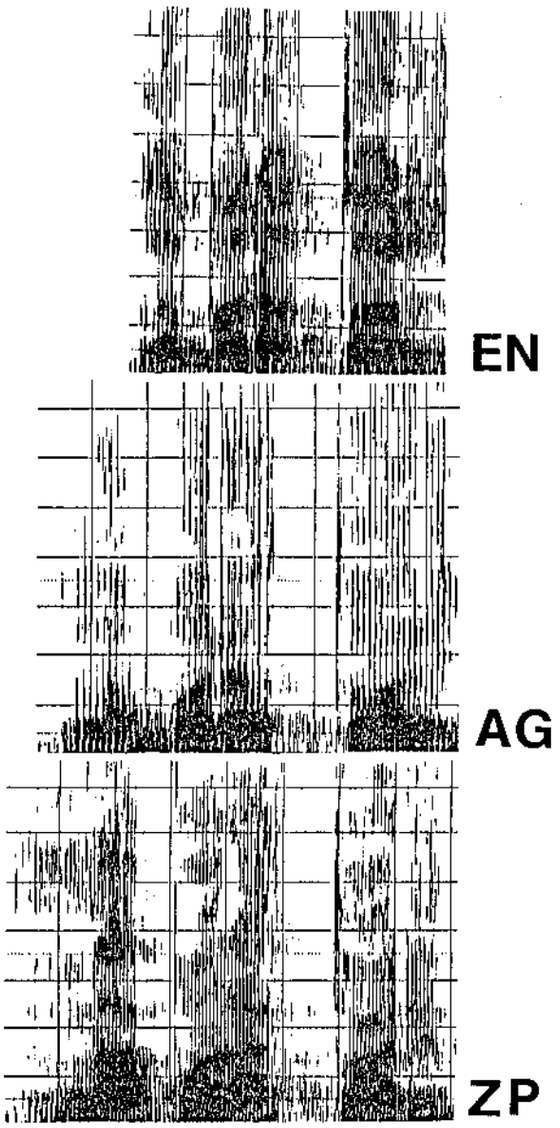


FIGURA 10.2: Espectrogramas da palavra "Laboratório", produzidas pelos falantes ZP, AG e EN (de baixo para cima). O *setup* do DSP-5500 da KAY Elemetrics é o mesmo do utilizado para a confecção dos espectrogramas da figura 10.1

As figuras 10.3 e 10.4 mostram espectrogramas comparativos extraídos de um caso forense (os *setups* do DSP-500 estão descritos nas legendas das figuras). As gravações foram realizadas através de um "grampo" telefônico, antes de passar pela sub-estação retransmissora; assim, o sinal não foi cortado em 3400 Hz (embora os espectrogramas utilizem apenas a faixa de 0-4000 Hz, havia energia acústica até cerca de 6500 Hz). As amostras questionadas encontram-se na parte inferior das figuras e as amostras de confronto (as duas do mesmo suspeito) na parte superior. A duas figuras revelam uma grande coincidência em vários aspectos, apesar de as amostras de confronto e questionada terem sido gravadas com um intervalo de mais de cinco meses. Observe-se, principalmente, a similaridade no *timing* e nos movimentos transicionais mais extensos. No caso em questão, tivemos oportunidade de verificar coincidências semelhantes em mais de vinte comparações espectrográficas, o que foi um elemento fundamental (mas não a única evidência) para chegarmos à conclusão de que se tratava efetivamente da mesma pessoa.

A evidência espectrográfica pode, eventualmente, se basear em sons que extrapolam o domínio da Linguagem, avançando no terreno daqueles aspectos que Trager (1958) chamou de *vocal characterizers* (choro, riso, muxoxo, etc). Sons dessa natureza podem apresentar características acústicas tão idiossincráticas quanto os sons da fala normal. Na verdade, já tivemos a oportunidade de empregar esse tipo de evidência em um caso forense. No referido caso, pudemos observar, na escuta da

gravação questionada, que um dos interlocutores produzia esporadicamente uma espécie de risada com tom sarcástico. Ao analisarmos o material de confronto ficou evidente que o suspeito possuía o mesmo tipo de "cacoete" vocal, que, tal como no diálogo questionado, aparecia recorrentemente ao longo da gravação. A análise espectrográfica revelou que os padrões acústicos eram semelhantes, como ilustra a figura 10.5. O som em questão era produzido com um golpe de glote inicial e a língua, aparentemente, em posição de repouso, sendo possível visualizar claramente as ressonâncias decorrentes dessa configuração articulatória. Os espectros de seção, na parte superior da figura 10.5, revelam a coincidência dos picos espectrais nas amostras questionada e de confronto. A semelhança é relevante, na medida em que, nessas condições articulatórias, o padrão de formantes estará relacionado com as dimensões totais do trato individual, aproximando-se de uma vogal "neutra" para aquele falante (observar o igual espaçamento entre os formantes). Um possível imitador precisaria ter o mesmo comprimento do trato para produzir padrão semelhante - o que seria extremamente improvável.

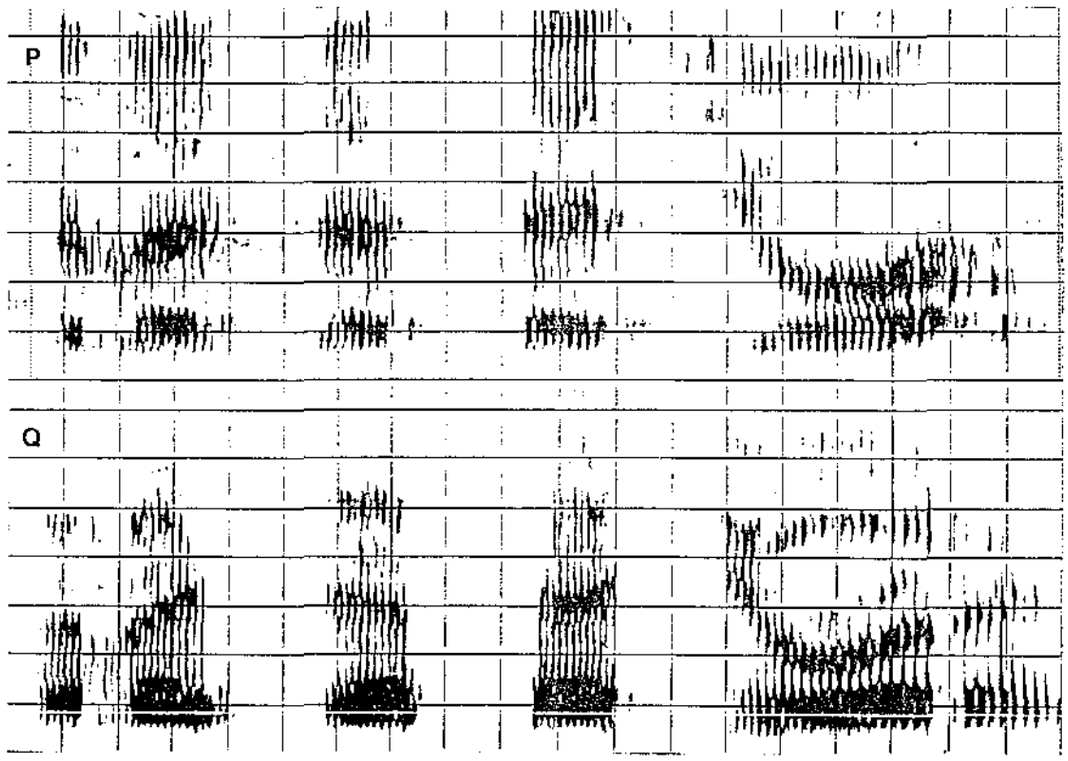


FIGURA 10.3: Espectrogramas comparativos de um caso forense real. A expressão produzida foi "pelas dezessete horas"; a amostra inferior refere-se à voz questionada e a superior à voz do suspeito. A análise no DSP-500 da KAY Elemetrics utiliza a faixa de 0-4000 Hz, com pré-ênfase (*high shape*), com filtro de 200 Hz; cada divisão vertical corresponde a 50 milisegundos.

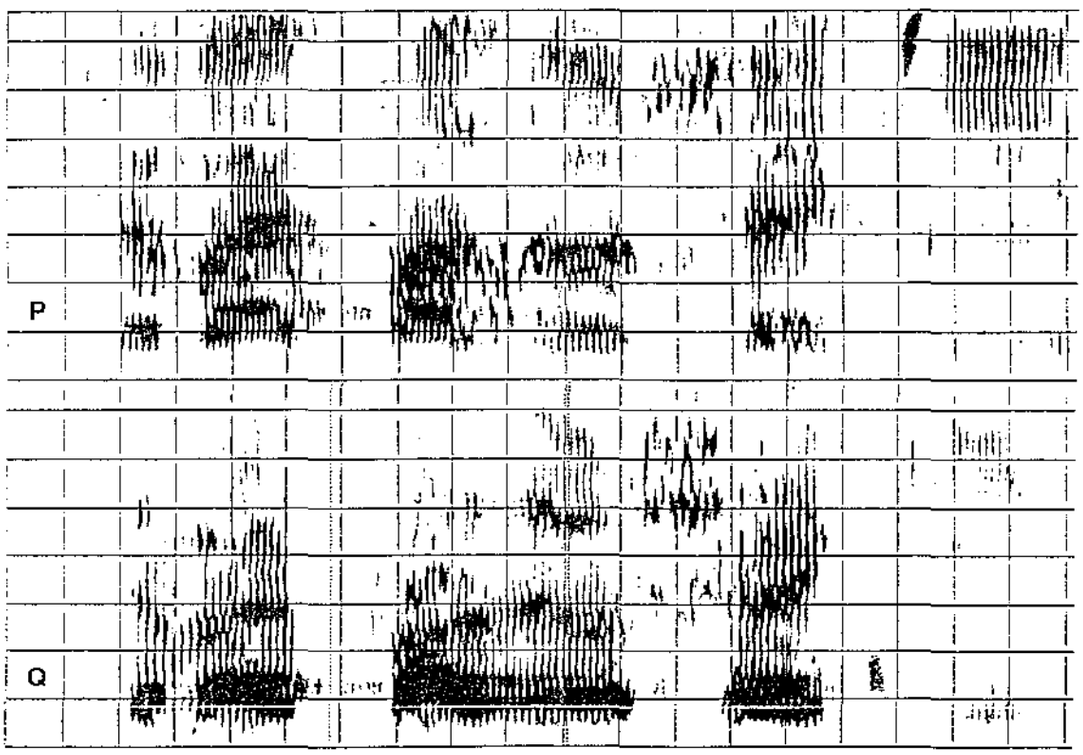


FIGURA 10.4: Espectrogramas comparativos de um caso forense real. A expressão produzida foi " 'cê' vai fazer o *check-in*"; a amostra inferior refere-se à voz questionada e a superior à voz do suspeito. A análise no DSP-500 da KAY Elemetrics utiliza a faixa de 0-4000 Hz, com pré-ênfase (*high shape*), com filtro de 200 Hz; cada divisão vertical corresponde a 50 milisegundos.

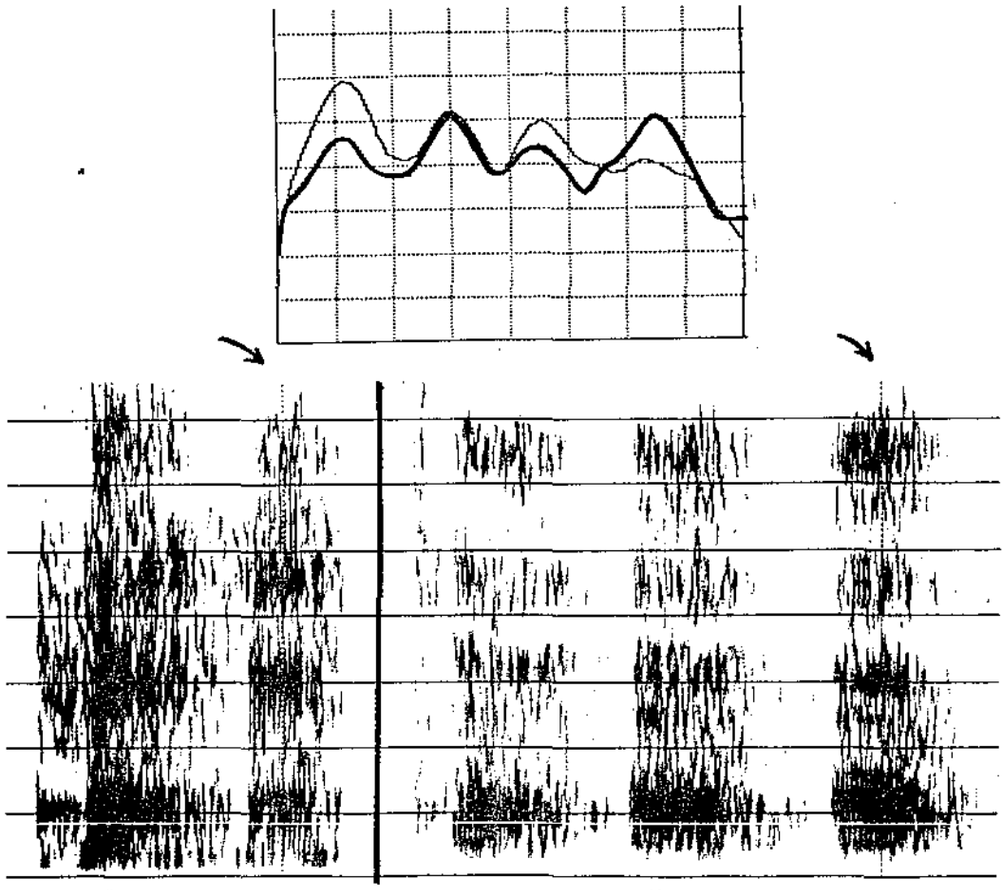


FIGURA 10.5: Espectrogramas de um caso forense. O som produzido é uma espécie de riso curto, consistindo em um golpe de glote inicial seguido de uma vogal neutra. À esquerda estão representadas duas amostras extraídas da gravação de confronto, e à direita três amostras da gravação questionada. A faixa utilizada é de 0-4000 Hz, com pré-ênfase (*high shape*) e filtro "banda larga" (300 Hz). Na parte superior da figura são mostrados dois espectros de seção (faixa 0-4000 Hz, filtro de 300 Hz) extraídos nos pontos assinalados por setas; a linha mais espessa refere-se à amostra questionada. Pode-se observar a coincidência no padrão de formantes.

Além da determinação da identidade do falante, a análise espectrográfica, em alguns casos, pode oferecer auxílio para a transcrição de trechos duvidosos da gravação questionada, onde, auditivamente, não é possível entender exatamente a palavra ou expressão produzida. Durante o exame de um dos casos forenses em nosso laboratório, surgiram algumas dúvidas quanto a uma expressão dita por um dos interlocutores. Por motivos decorrentes da situação, os interlocutores preferiam permanecer incógnitos (mesmo não sabendo que estavam sendo gravados) e não se trataram pelos nomes. Em um certo momento, entretanto, um dos envolvidos chamou seu parceiro usando, aparentemente, a expressão "baixinho". A transcrição do trecho era importante, pois tratava-se da alcunha pela qual um dos suspeitos era conhecido. O ruído de fundo interferente no trecho, no entanto, prejudicava a compreensão auditiva da expressão questionada, não permitindo uma transcrição segura . A análise espectrográfica, apesar do ruído de fundo, revelou que a palavra produzida era mesmo "baixinho", como se constatou ao compararmos uma produção da mesma expressão extraída do material de confronto com a amostra questionada, tal como ilustra a figura 10.6 [3].

Uma questão que surge com uma certa freqüência a respeito da inspeção visual de espectrogramas é a do disfarce/imitação. Já discutimos antes o tema, ressaltando que disfarce e imitação são eventos raros na situação forense, especialmente na amostra de confronto. Admitamos, no entanto, que na gravação questionada o falante disfarçou sua voz ou, intencionalmente, tentou imitar uma determinada voz, de modo a incriminar alguém. Há vários desdobramentos possíveis para essas situações. Se o disfarce na gravação questionada for realmente eficiente, é muito provável que não seja possível estabelecer com segurança se a voz pertence

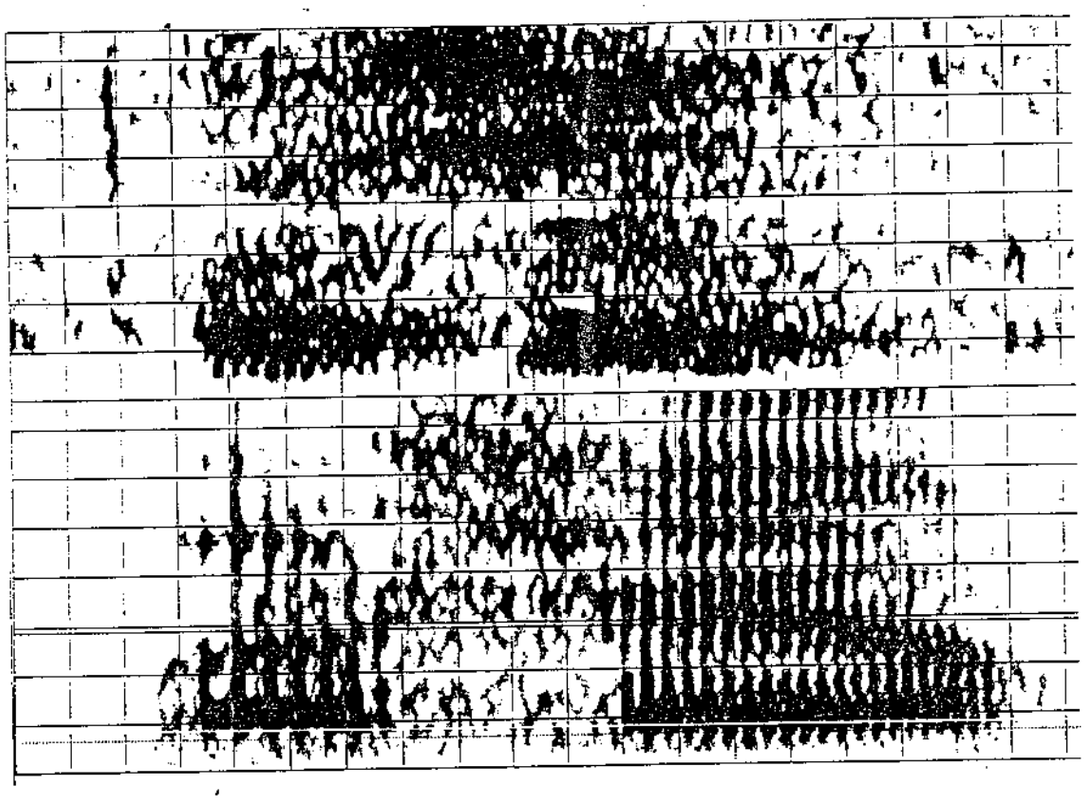


FIGURA 10.6: Espectrogramas de um caso forense. A análise espectrográfica foi usada aqui para confirmar a transcrição de um termo duvidoso para a decodificação auditiva. A palavra representada nos dois espectrogramas é "baixinho"; na parte inferior está a amostra dita pelo suspeito e na superior a amostra questionada. O *setup* do DSP-500 foi ajustado para a faixa de 0-4000 Hz, com pré-ênfase (*high shape*), filtro de 150 Hz; cada divisão vertical corresponde a 25 milisegundos.

ou não ao suspeito; alguns "disfarces", entretanto, serão facilmente detectáveis, não ocultando características idiossincráticas relevantes. Poder-se-ia alegar que, eventualmente, uma voz disfarçada possa se aproximar da voz de um suspeito, que poderia ser erroneamente incriminado. Isso talvez pudesse ocorrer, mas é estatisticamente extremamente improvável que a voz alterada de um criminoso acidentalmente se torne, **de todos os modos**, semelhante à de alguém que seja suspeito do crime.

Uma possibilidade mais realista seria a da imitação: em uma gravação forjada, alguém tenta deliberadamente imitar a voz de uma determinada pessoa, de modo a incriminá-la mais tarde. Já discutimos anteriormente (v. seção 1.2) a questão da imitação, observando que, embora o impostor possa eventualmente ter algum sucesso em julgamentos auditivos, não consegue se aproximar efetivamente da voz alvo, especialmente se são considerados aspectos espectrais . Nesse sentido, a análise espectrográfica pode ser de grande valia; em geral, padrões espectrográficos de imitadores são facilmente distinguíveis dos extraídos das vozes imitadas (Endres *et al.* 1971; Hall e Tosi 1975).

Há algum tempo atrás nos foi solicitado, que avaliássemos a performance de um imitador, de modo a verificar espectrograficamente o grau de similaridade entre sua imitação e a voz de um conhecido cantor. O imitador teve acesso à gravação de um trecho de fala do cantor (cerca de 50 palavras), e dispôs do tempo que achou necessário para realizar a tarefa proposta. O referido cantor possui certos traços de fala bastante característicos, incluindo uma qualidade de voz bastante nasalizada, mais evidente em vogais tônicas, e uma tendência a *creaky voice* em alguns finais de frase. Esses traços foram evidentemente explorados pelo imitador; além disso, observou-se que o artista procurou reproduzir os padrões prosódicos da voz alvo, ajustando ritmo e contorno entoacional. Através desses recursos o imitador consegue, para qualquer um que escute o trecho, definir a identidade do imitado.

Com efeito, muitos ouvintes serão iludidos pela sua performance, especialmente se não tiverem acesso à amostra da voz original. Acreditamos, entretanto, que o ouvinte (mesmo não treinado) conseguirá reconhecer a fraude se puder comparar diretamente as duas amostras; na verdade, alguns experimentos têm demonstrado que, em testes de discriminação, a maioria dos ouvintes consegue identificar a voz imitada (Rosenberg 1971; Hall e Tosi 1975).

O imitador de vozes age de forma análoga ao caricaturista gráfico, na medida em que ressalta alguns traços específicos, perceptualmente mais salientes. O ouvinte, reconhecendo esses aspectos estereotipados estabelece imediatamente uma conexão com a personalidade representada, independentemente do fato de ser ou não iludido pela imitação. Uma análise mais detalhada revelará, certamente, a existência de uma série de diferenças entre as duas vozes. A figura 10.7 apresenta dois espectrogramas relacionados com a tentativa de imitação acima discutida; na parte inferior da figura encontra-se uma amostra da fala do imitado produzindo a palavra "determinadas" (no contexto "gosto de determinadas cores...") e na parte superior uma amostra do imitador produzindo a mesma palavra no mesmo contexto (no DSP-5500 foi usada a faixa de 0 -4000 Hz, com pré-ênfase e filtro de 200 Hz). Pelo espaçamento dos pulsos glotais podemos verificar que o F0 médio do imitado é consideravelmente mais baixo do que o do imitador. Já se verificou que os imitadores, em geral, procuram ajustar seu F0 médio, de modo a aproximá-lo do valor da voz alvo, embora não consiga uma coincidência exata (Endres *et al.*1971; Hall e Tosi 1975). É provável que algo semelhante tenha ocorrido aqui, visto que o F0 médio do imitador parece mais baixo do que na sua fala normal (essa é uma impressão baseada apenas na memória perceptual, pois não tivemos acesso a amostras da fala normal do imitador). Os espectrogramas deixam transparecer também uma série de diferenças no nível espectral e no *timing* relativo entre os eventos acústicos (embora a duração total da palavra em questão seja semelhante). Comentaremos apenas as realizações

relativo entre os eventos acústicos (embora a duração total da palavra em questão seja semelhante). Comentaremos apenas as realizações da fricativa final /s/; observe-se que, na produção do imitado, o centro de energia desse som encontra-se em uma região de freqüência consideravelmente mais baixa do que na fala do imitador, indicando uma posição articulatória mais posteriorizada, provavelmente relacionada a aspectos dialetais. Esse traço, aparentemente, passou desapercebido para o imitador, embora seja facilmente detectável na análise espectrográfica.

Do ponto de vista da prática forense, tentativas de disfarce e imitação, efetivamente introduzem algumas dificuldades, mas nunca no sentido de incriminar a pessoa errada. Tanto no caso da imitação quanto do disfarce, pode ser difícil, em alguns casos, chegar ao impostor, mas, com certeza, a análise espectrográfica comparativa revelará a existência da fraude. Em suma: embora algum culpado possa continuar impune, nenhum inocente será injustamente responsabilizado.

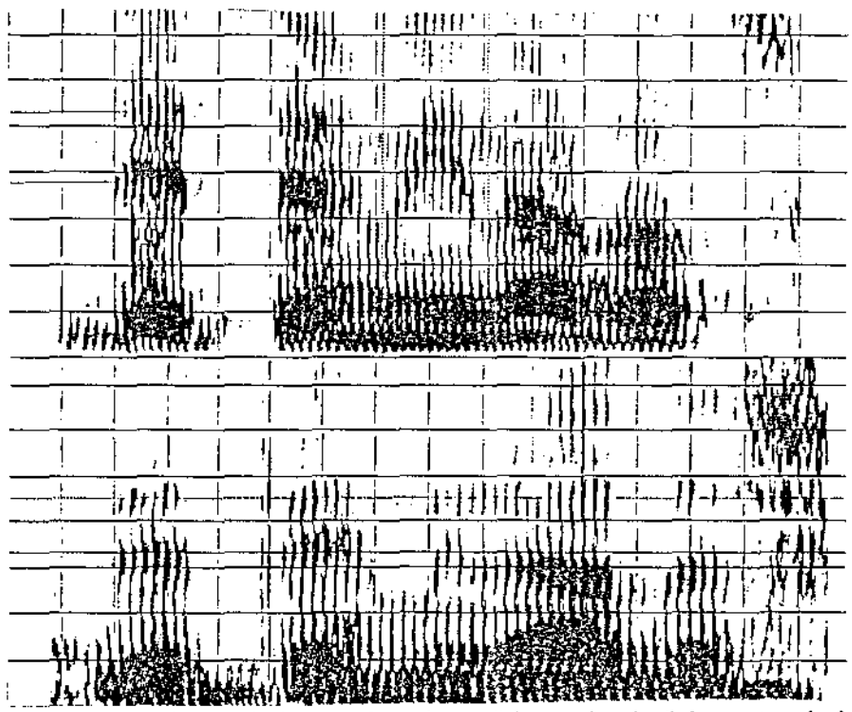


FIGURA 10.7: Tentativa de imitação da palavra "determinadas" (amostra do imitador no espectrograma superior, voz original no inferior)

SEÇÃO 11: COMENTÁRIO FINAL

11.1) *Eficiência Relativa de Alguns Parâmetros*

Os experimentos descritos ao longo do presente trabalho destacaram uma série de aspectos acústicos que podem ser explorados com o objetivo de identificar falantes, tanto no paradigma de Verificação Automática, quanto no - mais complexo - modelo forense. Observamos que alguns parâmetros são potencialmente mais eficazes. A tabela 11.1 apresenta todos os pares possíveis de falantes do grupo principal (n=9; 7 + R1/R2). Cada asterisco na tabela 11.1 representa, dependendo de sua posição, um determinado parâmetro ou conjunto de parâmetros; a presença do asterisco indica uma distinção estatisticamente significativa para aquele par de falantes. F3 e F4 vocálicos foram agrupados; assim, um asterisco nessa posição indica que os dois falantes são diferentes em F3 e F4. O mesmo ocorre com F5 e F6 da nasal /n/, que também foram agrupados. A última coluna à direita apresenta, para cada falante, o número de distinções corretas para cada aspecto acústico (para o par R1/R2, considerou-se uma **não-distinção** como acerto); o quadro separado logo abaixo dessa coluna mostra os totais globais, isto é, o número de separações corretas considerando todos os falantes.

Podemos verificar na tabela 11.1 que o ELT é o parâmetro mais eficiente e os formantes nasais o menos eficaz. F0 separa a maioria dos falantes; R1 e R2, no entanto, são, erroneamente, considerados distintos, indicando que a variação intra-falante de F0 pode ser significativa. É importante observar, contudo, que R1 e R2 não diferem em todos os demais aspectos (agrupando-se os formantes altos como um só parâmetro); observe-se também que os demais pares são diferentes em pelo

menos três aspectos (com exceção de WA/DO, que diferem em apenas dois aspectos, [F3 + F4] e ELT).

	EN	R1	ZP	AG	WA	MS	R2	DO	TOTAL
ZR	* *	*	* *	* *	* *	*	*	* *	8 0 5
	* *	* *	* *	* *	* *	* *	* *	*	7 0 8
EN		*	* * *	* *	*	* *	* *	*	5 2 7
		* *	* * *	* *	*	* *	* *	* *	8 3 5
R1			* *	* * *	*		*	*	3 2 6
			*	* *	* *	* * *		* *	8 2 7
ZP				* * *	* *	* *	* *	* * *	7 4 8
				* *	* *	* *	*	* *	8 1 5
AG					*	* *	* * *	*	8 3 6
					* *	* *	* *	* *	8 0 8
WA						* *	* *	*	5 0 7
						*	* *	*	8 0 5
MS							* *	* *	6 1 5
							* * *	*	8 3 5
R2								* * *	7 5 5
								* * *	8 3 7

↓

49	17	49
63	12	50

[F0]	[F1]	[F3,4 V]
[ELT]	[F5,6/n/]	[VOT]

TABELA 11.1: Resumo das distinções estatisticamente significativas para cada par possível de falantes do grupo principal (n=9: 7 + R1/R2). A presença de um asterisco indica que os dois falantes cruzados são significativamente distintos no parâmetro correspondente à posição do asterisco; as posições de cada parâmetro são definidas no quadro com moldura em linha dupla (F0 ≡ superior esquerda; VOT ≡ inferior direita, F1 ≡ superior centro, etc). Os formantes vocálicos altos foram reunidos como um só parâmetro; assim, um asterisco nessa posição indica que os dois falantes são distintos quanto a F3 e F4 (o mesmo procedimento foi usado para F5 e F6 de /n/).

De acordo com o número total de distinções corretas, os parâmetros mais eficientes foram, em ordem decrescente:

ELT > VOT > [F3+F4] = F0 > F1 > [F5 + F6 /n/].

Essa ordenação, no entanto, deve ser avaliada com uma certa cautela. Uma comparação direta entre os parâmetros é difícil, visto que os conjuntos de medidas não são totalmente compatíveis (há poucas observações de VOT, em comparação com Formantes e F0, por exemplo). Outro fator a considerar é a relação de cada parâmetro com cada falante. Certos aspectos parecem mais relevantes para certos falantes; observamos, por exemplo, que apenas o falante ZP distingue-se de todos os demais quanto aos formantes vocálicos altos. De modo geral, um determinado parâmetro será proporcionalmente mais importante, à medida em que, para um determinado falante, esse parâmetro afasta-se da média global. Especialmente no paradigma forense, onde cada caso é examinado separadamente, uma avaliação "objetiva" da eficiência de um aspecto acústico dependerá do valor absoluto para o falante em questão.

Wolf (1972) sugere alguns fatores que devem ser levados em conta na avaliação de uma característica de fala quanto ao seu potencial para a Identificação de Falantes; as características mais úteis seriam as que

1) ocorrem natural e freqüentemente na fala normal.

2) são facilmente mensuraveis.

3) variam mais inter- do que intra-falante.

4) variam pouco em amostras não-contemporâneas e/ou em função do estado de saúde do falante.

5) são pouco afetadas pelo ruído de fundo e não dependem de características específicas do meio de transmissão.

6) são mais resistentes a tentativas de disfarce.

Aparentemente, nenhum parâmetro preenche **todos** os requisitos listados por Wolf (1972). A importância de cada item dependerá, evidentemente, do tipo de aplicação; para um sistema de Verificação Automática de Falante, os itens (1), (5) e (6) são pouco relevantes, enquanto no modelo forense os mesmos aspectos adquirem um peso maior.

11.2) *Novas Perspectivas*

Embora tenhamos concentrado nossa atenção nos aspectos acústicos que podem servir de pistas para determinar a identidade de um falante, há uma série de dimensões mais abstratas que podem ser potencialmente relevantes para o mesmo fim. Os falantes não se diferenciam apenas no que diz respeito ao som de sua fala, mas também quanto ao modo como estruturam as unidades lingüísticas em vários níveis. Determinadas escolhas lexicais, construções sintáticas características, e outros fatores, podem ser tão idiossincráticos quanto a qualidade de voz. Na verdade, a participação de lingüistas em processos legais não está limitada aos foneticistas; podemos falar mais genericamente de uma *Lingüística Forense*, que abrangeria, não só o aspecto fonético, mas também questões sintáticas, semânticas, discursivas, etc. (Cf. Di Paolo e Green 1990; Dumas 1993; Shuy 1993; Tiersma 1993).

NOTAS

Seção 2: Abordagem Perceptual

1) Kretschmer definiu três tipos físicos básicos: o *pícnico* caracteriza-se pela baixa altura, formas arredondadas e membros curtos; o *leptossômico* é alto e magro, e o *atlético*, como diz o próprio nome, tem porte vigoroso e constituição geral forte e musculosa. Kretschmer associava esses tipos básicos também a traços de caráter, uma extrapolação bastante discutível (o tipo *pícnico*, por exemplo, corresponderia ao caráter *ciclotímico*), cuja referência aqui passa apenas pelo caráter pitoresco do fato.

2) As alterações de parâmetros relacionados a F0, em função da idade, serão abordadas mais detalhadamente na seção 5.

3) Essa questão será discutida mais detalhadamente quando examinarmos a eficiência dos formantes vocálicos na identificação de falantes (seção 4).

4) O caso em questão envolve o seqüestro do filho do aviador Charles Lindbergh, o primeiro homem a sobrevoar o Atlântico em vôo *solo*. O incidente teve conseqüências trágicas, com o filho do aviador tendo sido encontrado morto algum tempo depois do seqüestro. Lindbergh teria ouvido a voz do seqüestrador duas vezes: uma por telefone (de modelo antigo e baixa fidelidade) e outra pessoalmente, mas por breve espaço de tempo e à noite, sem a possibilidade de visualizar claramente o criminoso. Dois anos depois, durante o julgamento, Lindbergh testemunhou, afirmando reconhecer a voz de um certo Bruno

Hauptmann como sendo a mesma voz do seqüestrador. O testemunho foi aceito e não surgiram dúvidas quanto à sua validade, exceto para a psicóloga Frances McGehee, que não se mostrou convencida que a identificação auditiva, passado tal intervalo de tempo, fosse confiável. Os experimentos de McGehee comentados no texto queriam demonstrar que intervalos de tempo muito grandes entre as apresentações de amostras de fala para confronto têm um efeito fortemente negativo na performance, o que colocaria em cheque o testemunho de Lindbergh.

5) Também seria possível uma tarefa de *discriminação* com uma defasagem entre as amostras comparadas, mas não temos notícia de um estudo empregando esse modelo experimental.

Seção 4: Formantes (Vogais)

1) O F1 /ã/ muito baixo observado para o falante DO poderia ser interpretado como um formante nasal (e não o F1 "verdadeiro"). Interações complexas de pólos e zeros na região de F1 dificultam a interpretação dos picos espectrais que aí surgem, e não é raro que o espectro de vogais nasalizadas se torne pouco definido nesa faixa (Cf. Hawkins e Stevens 1985). No caso do falante DO, não se observou, em nenhum espectro de /ã/, um pico espectral nítido, na faixa de 500-700 Hz, que pudesse ser diretamente associado a F1; o aparecimento sistemático de uma ressonância na faixa de 300 HZ (e apenas essa ressonância aparece nos espectros de DO, antes de F2) não deixou outra alternativa a não ser rotular esse pico como F1. Para os objetivos da Identificação de Falantes, é menos importante a explicação da "origem" desse formante do que o fato de sua regularidade, por si só um elemento importante para a determinação da identidade do falante DO.

2) A existência de alterações voluntárias da voz na situação forense, tais como disfarce e imitação, é muito mais rara do que geralmente se imagina (Cf. Ladefoged 1984). Na maior parte dos casos, o falante, na gravação questionada, **não sabe** que está sendo gravado. Na coleta de voz padrão, tentativas de disfarce seriam por demais evidentes e raramente são utilizadas pelos suspeitos. Além disso, é preciso considerar que o falante, em geral, não tem controle dos parâmetros relevantes para a análise (a não ser que seja um imitador profissional, ou um foneticista treinado).

Seção 5: Freqüência Fundamental

1) O mesmo efeito pode também ser causado pelo natural constrangimento imposto pela situação - bastante desagradável para o suspeito. Com alguma habilidade, entretanto, é quase sempre possível estabelecer uma ambiência mais informal durante a coleta de voz padrão, de modo a obter amostras confiáveis.

2) Os momentos de ordem *n* de uma distribuição são definidos como

$[\Sigma x^n] \div N$

onde x = X - (média), sendo X o valor de uma ocorrência qualquer e N = número total de ocorrências.

(v. Lutz 1983)

3) *Jitter* e *Shimmer* são os termos geralmente usados para expressar as flutuações aleatórias período-a-período, no domínio da freqüência (*jitter*) e da amplitude (*shimmer*) (Horii 1979, 1980). Algum grau de perturbação sempre existirá, mesmo

em fonação sustentada; na verdade, a ausência total dessa pequena variação local faz a fala soar pouco natural, sendo esse o motivo pelo qual se tem procurado integrar efeitos de *jitter* e *shimmer* em sistemas de síntese de voz (Cf. Klatt e Klatt 1990). Embora seja uma perturbação de F0, essas aperiodicidades não são percebidas dentro do domínio entoacional, mas sim como um componente de longo termo, relacionado à qualidade de voz; assim, índices mais elevados de *jitter* e/ou *shimmer* emprestam à voz uma qualidade *harsh* (Cf. Laver 1980:127).

4) Em uma área relacionada, entretanto, medidas de microperturbação de F0 têm mostrado alguma eficiência. Referimo-nos à detecção de *stress* psicológico, um componente do famigerado "detetor de mentiras" (v. Disner 1982 para uma resenha de diversos trabalhos nessa área).

5) Embora as línguas tonais também sinalizem aspectos afetivos através de variações entoacionais, o uso lingüístico das inflexões de F0 restringe consideravelmente a manipulação de parâmetros relacionados a F0 (Cf. Ross *et al.* 1986).

6) A expressão de diferentes estados emocionais através da entonação, a partir de frases curtas e semanticamente neutras, era um dos exercícios preferidos pelo diretor russo Stanislawski para testar os atores com quem trabalhava. Antes dele, porém, uma certa Madame Modjeska, por volta da virada do século, utilizava material ainda mais neutro, incluindo em suas performances *solo* a recitação do alfabeto polonês expressando diferentes estados emocionais (Cf. ben Avram 1969). Mais perto de nós, para quem se recorda, o grande ator Procópio Ferreira também fazia o mesmo exercício de Stanislawski em algumas de suas apresentações.

7) A seção 5.2.3 pode ser entendida como uma extensão da seção 5.2.2, já que continuamos tratando de variações motivadas por estados emocionais. A separação aqui deve-se ao fato de haver uma linha de pesquisa específica voltada ao estudo dos efeitos vocais de vários tipos de *stress* psicológico.

8) Uma terceira abordagem, que não será enfocada aqui, examina os efeitos de *stress* provocado por causas fisiológicas. A aplicação desse tipo de metodologia experimental deixa algumas dúvidas quanto ao aspecto ético, submetendo o sujeito a experiências bastante desagradáveis, tais como o isolamento sensorial em câmaras escuras à prova de som (Rubenstein 1966), exposição a odores desagradáveis, como o emanado pelo cloridrato de amônia (Ostwald 1963), ou a aplicação aleatória de descargas elétricas durante a leitura de um texto (McGlone 1975) (todos *apud* Disner 1982).

9) Uma variável importante nesse sentido é a própria capacidade do indivíduo em suportar situações envolvendo pressão psicológica. Treinamento, por exemplo, pode fazer uma grande diferença. Alguns agentes de organismos de segurança são treinados para não produzir padrões coerentes no polígrafo, através do controle do estado fisiológico de órgãos específicos, de forma a produzir respostas equivalentes para todos os itens ou, alternativamente, produzindo respostas mais amplas para questões irrelevantes (Honts 1987). Sabe-se também que certas personalidades patológicas não apresentam qualquer alteração fisiológica (relacionada à voz ou não) detectável no polígrafo, quando mentindo (Hollien *et al.* 1987). Embora esses exemplos representem casos extremos, uma variação

considerável na resistência a pressões psicológicas é esperada em qualquer grupo de indivíduos normais.

10) Optou-se aqui pela medição de F0 diretamente na forma de onda por ser esse método mais rápido, em relação ao exame das distâncias entre harmônicos em espectrogramas de banda estreita (um fator importante, tendo em vista o grande número de dados no presente trabalho). A verificação visual na forma de onda é também muitas vezes impossível em gravações degradadas, e poder-se-ia argumentar que os resultados obtidos pela medição dos períodos na forma de onda podem ser diferente do obtido nos espectrogramas de banda estreita. Com efeito, medidas extraídas na forma de onda tendem a ser mais exatas, mas a diferença em relação ao método dos harmônicos é irrelevante se, nesse último caso, toma-se a distância entre 3 ou mais harmônicos. Um teste preliminar, comparando as medidas nos dois métodos (forma de onda e distância inter-harmônicos), indicou que a diferença média, para um conjunto de 100 medidas, fica em torno de 2 %, um desvio desprezível para a variável em questão.

11) É importante ressaltar que os intervalos de confiança não representam o *range* total ou parcial, mas sim uma estimativa da faixa onde a média de cada grupo, com base em uma distribuição normal, deve ser encontrada com uma determinada probabilidade (95%, no caso em questão). Dito de outro modo: se o experimento fosse repetido, aceitando-se a hipótese de uma distribuição normal, haveria apenas 5% de probabilidade de a média encontrar-se fora dos limites estabelecidos pelo intervalo de confiança.

12) Em muitos casos onde estão envolvidas personalidades da vida pública, é possível obter, para confronto, entrevistas em meios de rádio difusão, discursos, etc. Esse tipo de material fornece, em geral, um ótimo quadro de referência para estabelecer comparações baseadas em F0. Em um rumoroso caso ocorrido há poucos anos atrás, envolvendo um ex-ministro de Estado acusado de corrupção passiva, tivemos a oportunidade de empregar trechos de entrevistas e discursos do acusado, de onde se extrairam diversas medidas de F0 (apenas em núcleos vocálicos); a comparação dessas medidas com as obtidas na fita questionada (pelo mesmo método) resultou em histogramas quase coincidentes, uma informação que foi um elemento importante para a definição do laudo pericial (v. Figueiredo *et al.* 1993).

13) Zawadzki (1989) prefere associar o F0 intrínseco à posição da mandíbula, antes que ao dorso da língua, argumentando que a dimensão *altura vocálica* não tem uma relação clara com a altura do ponto mais alto da língua. Através de exames cine-radiográficos, Zawadski observa que a vogal /I/, por exemplo, tem a posição da língua **mais baixa** do que a vogal /e/, mas se for considerada apenas a posição da mandíbula, observa-se o oposto.

14) A dificuldade do algoritmo para extrair valores exatos quando o F0 médio é muito baixo deve-se à relação entre o tamanho do *frame* utilizado (20 milisegundos) e o período da forma de onda. A extração é facilitada se pelo menos um período completo está contido no interior do *frame*, o que só ocorre para um

F0 local maior que 100 Hz. O programa LPC utilizado permite estabelecer *frames* mais longos (até 30 milisegundos). Esse procedimento pode ser empregado quando a média de F0 é muito baixa, tal como ocorre com o falante AG; não o fizemos porque isso faria com que a distribuição de F0 para esse falante tivesse um número de pontos muito menor do que as dos demais falantes, desbalanceando uma série de comparações baseadas em medidas obtidas via LPC.

15) O termo *declinação de F0* refere-se ao declínio gradual do contorno entoacional, freqüentemente observado ao longo de frases e enunciados. O fenômeno é considerado como universal por alguns autores; segundo esse ponto de vista, o F0 de todos os enunciados tem uma linha de base descendente subjacente (que está virtualmente presente, mesmo que o contorno não seja de fato descendente) (para uma resenha da literatura sobre o tema, v. Ladd 1984). Há alguma evidência experimental indicando que o mesmo valor de F0 é percebido como mais alto se ocorre no fim de uma sentença *nonsense* do que se ocorre perto do começo (Pierrehumbert 1979), sugerindo que há uma expectativa do ouvinte em relação a uma linha entoacional de base descendente (o F0 no final é avaliado como mais alto porque o ouvinte refere-o a essa linha virtual declinada). Umeda (1982) observa que a declinação de F0 raramente ocorrerá na fala espontânea, mas será freqüente em leituras "mecânicas", quando o falante "is too concerned about not making mistakes and is not able to digest and express the idea".

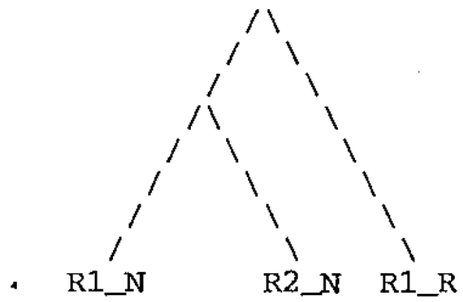
Seção 6: Espectro de Longo Termo

1) A possibilidade de reproduzir as condições originais em casos forenses reais depende muito do modo como foi realizada a gravação questionada. Em "grampos" telefônicos (judicialmente autorizados!) feitos diretamente na linha, sem passar por

sub-estações, não há dificuldades em repetir as condições originais. Sempre que há acesso ao local da gravação questionada, aparelho gravador e fita originais, e tem-se certeza das posições relativas dos interlocutores durante a gravação questionada, é viável empregar o ELT como indicador da identidade do(s) falante(s). É preciso considerar também que o ELT pode ser um instrumento útil também para verificar a autenticidade de uma gravação, na medida em que permite, através de comparações de diversas amostras do mesmo falante ao longo da gravação questionada, detectar tentativas de fraude por meio de edição, ou seja, a criação de um diálogo fictício a partir de trechos colhidos em diferentes gravações do mesmo falante. Sendo o ELT um espectro composto, as configurações finais de trechos suficientemente longos de um mesmo falante só serão semelhantes se todos os trechos foram gravados no mesmo ambiente.

2) Falamos aqui de "taxa de articulação", medida diferente da "taxa de fala" (*speech rate*), essa última incluindo pausas de respiração e/ou hesitação, representando assim mais diretamente aspectos de fluência. A medida relevante aqui é a taxa de articulação, que reflete efetivamente as possíveis alterações ao nível segmental (para uma definição das duas medidas, ver Scherer 1979:161).

3) Surgiram algumas dificuldades para avaliar o que seria uma identificação correta no caso das amostras do falante R1-R2. Em algumas análises um *cluster* inicial foi gerado agrupando as produções não contemporâneas desse falante ([R1-N + R2-N], por exemplo). Nesses casos o agrupamento não foi considerado como "acerto", a não ser quando o próximo item a ser conectado ao *cluster* já formado fosse também uma produção desse mesmo falante (R1-R ou R2-R, para o exemplo dado acima). Assim, para uma configuração inicial como



pareceu razoável considerar a decisão como um acerto.

4) A principal dificuldade da classificação dos *settings* em Laver (1980) é a falta de uma descrição acústica objetiva. Ladefoged (1984:86), em uma resenha de Nolan (1983), observa que

> *the articulatory settings* [baseados em Laver 1980] *are described simply in impressionistic auditory terms, so there is no way in which anybody not familiar with these terms (i.e. the majority of phoneticians) can pursue this line of research.*

Essa observação assume um sentido mais incisivo se considerarmos *settings* complexos como *harsh ventricular whispery falsetto* (!).

5) É importante observar, no entanto, que ao empregar filtros mais estreitos seria necessário representar o ELT através de um maior número de pontos no eixo de freqüência. O sonógrafo KAY 5500 tem uma definição fixa de apenas 200 pontos no eixo de freqüência, independentemente do filtro de análise empregado. Nesse caso, um ELT obtido com filtro mais estreito não traria uma vantagem muito grande.

Seção 7: Aspectos Rítmico-Temporais

1) A duração de cada palavra foi medida diretamente na forma de onda, através dos cursores do DSP 5500, da KAY Elemetrics. Evidentemente, a definição exata de fronteiras entre palavras na fala fluente depende de decisões mais ou menos arbitrárias. No entanto, a possibilidade de monitoramento auditivo constante, através de *loops* na saída de áudio, oferecida pelo DSP 5500, é um apoio fundamental para as decisões a respeito das fronteiras de palavras. Além disso, as durações das palavras foram medidas seqüencialmente, ou seja, após definir uma duração, o cursor mais à direita era mantido fixo, tornando-se o ponto inicial da medida para a próxima palavra contígua. Com esses cuidados, acreditamos que eventuais distorções locais tenham um peso mínimo na composição das médias.

2) As médias extraídas a partir dos blocos e a média a partir do enunciado são ligeiramente diferentes porque, no segundo caso, não foram removidas eventuais pausas entre os blocos independentes. Assim, as médias considerando o total do enunciado são um pouco menores do que no primeiro método.

3) Embora o traço de vozeamento seja normalmente associado a propriedades segmentais, a medida aqui estudada, por suas características de longo termo, pode ser melhor classificada como uma dimensão prosódica, especialmente em função dos seus prováveis correlatos perceptuais, como ficará claro ao longo do texto.

4) Parece haver uma tendência geral a uma perda de diferenciação inter-falante na velocidade rápida para diversos parâmetros (alguns formantes, F0, durações, RTV, etc). Essa tendência pode estar associada a um princípio genérico que poderíamos

chamar de *efeito de limite*, ou seja, à medida que a velocidade de fala aumenta, reduz-se o "espaço" para a variação de alguns parâmetros acústicos, em função de serem atingidas certas condições articulatórias críticas. O já comentado *princípio de incompressibilidade* , proposto por Klatt (1973), seria um caso especial, no domínio das durações.

5) Alternativamente, poderiam ser empregadas outras curvas que não a de amplitude. Nolan (1983:129), por exemplo, sugere que as inflexões no contorno de F1 têm uma relação direta com a taxa de produção silábica, na medida em que as variações de F1 refletem em grande parte os movimentos de abertura/fechamento do trato.

Seção 8: Consoantes Nasais

1) Vale ressaltar aqui, mais uma vez, um aspecto que temos comentado ao longo deste trabalho. Tentativas de disfarce em casos forenses reais são muito mais raras do que habitualmente se imagina, por vários motivos: (1) quanto à gravação questionada, o falante, em geral, **não sabe** que está sendo gravado (excetuando, talvez - mas nem sempre - casos de seqüestro, ameaças de bomba, etc); (2) durante a coleta de voz padrão, tentativas de disfarce seriam **perceptualmente salientes** (não esquecer que o suspeito, antes de ser submetido à análise de voz, já passou por diversos tipos de interrogatórios e entrevistas; qualquer alteração radical no momento da gravação soaria falsa para os que acompanharam o caso) e, finalmente, mas não menos importante, (3) o falante não sabe **como** nem **o que** exatamente deve alterar para produzir disfarce eficiente. A possibilidade de o suspeito utilizar voz denasalizada (manutenção do *velum* levantado durante toda a

produção), de modo a evitar pistas acústicas relacionadas às cavidades nasais, é remota, já que se trata de controle difícil de ser exercido conscientemente (o leitor pode, como um exercício, tentar produzir essa qualidade).

2) A lista de palavras utilizada não foi especialmente desenhada para testes específicos com a nasal /n/, pois faz parte de uma lista mais extensa destinada a verificar também outros aspectos da fala (em experimentos não descritos no presente trabalho).

3) Entenda-se aqui "formante" no sentido de pólo da função de transferência. Obviamente, em consequência da interação pólos/zeros, o espectro de nasais é complexo e não apresenta uma estrutura onde os picos espectrais visíveis possam ser referidos de forma direta aos pólos; o que rotulamos aqui como F5 e F6 são, portanto, os pólos "teóricos" equivalentes na função de transferência (Cf. Fujimura 1962).

4) É interessante observar que, na seção 4, verificamos que F4 das vogais (não nasais) de R2, situado na mesma região de freqüência de F5/n/, também sofre alguma alteração em relação a R1, mas na direção oposta (queda de 3.5 % em relação a R1). É possível que os dois efeitos estejam de alguma forma relacionados, embora não seja evidente de que forma.

5) Presumivelmente também são semelhantes os índices de absorção acústica relacionados à consistência dos tecidos. Ambos os gêmeos não apresentavam qualquer sinal de resfriado ou qualquer sinal de obstrução das cavidades nasais.

Seção 9: Voice Onset Time

1) A influência de erros aleatórios relacionados à eventual imprecisão de medidas pode ser eliminada, ou ao menos drasticamente reduzida, adotando-se um ajuste que exclui da distribuição um percentual pré-definido de valores que se afastam demais da média (*trimming*). Esse procedimento, no entanto, pressupõe um conjunto significativo de dados.

Seção 10: Abordagem Espectrográfica

1) Alguns casos forenses, raros, podem envolver situações semelhantes às simuladas em testes fechados. Isso só ocorre quando se sabe que, dentre um grupo de suspeitos, um é, com certeza, o culpado. Situações desse tipo podem ocorrer quando se localiza rapidamente a origem de um telefonema (dentro de uma empresa, por exemplo), limitando assim o número de suspeitos aos presentes em um determinado recinto.

2) Em alguns casos é impossível "disfarçar" o trecho relevante, em virtude da natureza muito peculiar do que foi dito na gravação questionada, e a expressão de interesse pericial deve ser repetida isoladamente pelo suspeito. Essa situação impõe uma maior dificuldade, já que o suspeito (se for efetivamente a mesma pessoa que aparece na gravação questionada) pode tentar algum tipo de disfarce. Já comentamos anteriormente (v. seção 1) que a possibilidade de disfarce durante a coleta de material de confronto é remota, já que qualquer alteração seria saliente para os que acompanham o processo. Além disso, o suspeito, em geral, não sabe como disfarçar eficientemente sua voz; na maior parte dos casos o "disfarce"

consiste em uma mera modificação de padrões rítmico/entoacionais (embora, eventualmente, essa modificação não seja intencional, mas antes um efeito do natural constrangimento da situação). Se perceber que a amostra soa artificial, o perito responsável pela gravação deve solicitar ao suspeito que produza o enunciado em outras condições, dando instruções do tipo: "fale a expressão X mais rapidâmente/lentamente", "fale a expressão X como se estivesse na situação Y", etc. Em geral, esse simples procedimento é suficiente para obter amostras representativas, mesmo para expressões isoladas.

3) A alta definição da tela colorida do DSP-5500 da KAY Elemetrics permite distinguir mais claramente os níveis de amplitude nos espectrogramas, diminuindo assim a interferência gráfica do ruído de fundo. Em espectrogramas com níveis de cinza, tais como os que serviram de base para as figuras deste trabalho, pode ser mais difícil separar visualmente o ruído de fundo do sinal de fala relevante. As reproduções xerográficas dos espectrogramas aqui utilizadas degradam ainda mais a definição gráfica, na medida em que "achatam" os níveis originais de cinza, transformando o espectrograma em uma figura monocromática. Esperamos que ainda seja possível ao leitor visualizar os aspectos espectrográficos importantes, especialmente na figura 10.6, onde, com a redução monocromática, o ruído de fundo aparecerá com a mesma intensidade do sinal de fala.

BIBLIOGRAFIA

Abberton, E. e A.J. Fourcin, 1978 "Intonation and speaker identification", Lang.Speech 21, 305-18

Abe, I., 1980 "How vocal pitch works", in L.R.Waugh e C.H. von Schooneveld 1980:1-24

Abercrombie, D., 1967 Elements of General Phonetics, Aldine.Atherton, Chicago, NY

Adams, S.G. e G.Weisner e R.D. Kent, 1993 "Speaking rate and speech movement velocity profiles", JSHR 36, 41-54

Alcaim, A., J.A. Solewicz e J.A. Moraes, 1993 "Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no Português falado no Rio de Janeiro", manuscrito

Anghelescu, I., 1974 "Método de Identificación de las personas por la voz y la manera de hablar en rumeno", Revista Internacional de Polícia Criminal, 274, 2-9

Atal, B.S., 1972 "Automatic speaker recognition based on pitch contours", JASA 52, 1687-97

Atal, B.S., 1974 "Effectiveness of linear prediction characteristics os the speech wave for automatic speaker identification and verification", JASA 55, 1304-12

Atal, B.S., 1976 "Automatic recognition of speakers from their voices", Proc. IEEE 64, 4, 460-75

Atkinson, J.E., 1976 "Inter- and intraspeaker variability in fundamental voice frequency, JASA 60, 440-5

Baken, R.J. e R.G. Daniloff, 1991 Readings in Clinical Spectrography of Speech, Singular Pub. Group e Kay Elemetrics Corp.

Barry, W.J., 1981 "Prosodic functions revisited again!", Phonetica 38, 320-40

Behlau, M.S., 1984 Uma Análise das Vogais do Protuguês Brasileiro falado em São Paulo: Perceptual, Espectrográfica de Formantes e Computadorizada de Freqüência Fundamental, Diss. Mestrado, Escola Paulista de Medicina

ben Avram, R., 1969 The Act and the Image, Odyssey Press, NY

Benguerel, A-P. e J. D'Arcy, 1986 "Time-warping and the perception of rhythm in speech", J.Phon 14, 231-46

Berkovits,R., 1991 "The effect of speaking rate on evidence for utterance-final lengthening", Phonetica 48, 57-66

Black, J.W., W. Lashbrook, E. Nash, H.J. Oyer, C. Pedrey, O.I. Tosi e H. Truby, 1973 "Reply to 'Speaker identification by speech spectrograms: some further observations'", JASA 54, 535-7

Bladon, R.A. e A. Al-Bamerni, 1976 "Coarticulation resistance in English /l/, J.Phon. 4, 137-50

Bogner, R.E., 1981 "On talker verification via orthogonal parameters", IEEE Trans. ASSP-29, 1, 1-12

Bolt, R.H., F.S. Cooper, E.E. David Jr., P.B. Denes, J.M. Pickett, K.N. Stevens, 1973 "Speaker identification by speech spectrograms: some further observations", JASA 54, 531-34

Bolt, R.H., F.S. Cooper, E.E. David Jr., P.B. Denes, J.M. Pickett e K.N. Stevens, 1970 "Speaker identification by speech spectrograms: a scientists' view of its reliability for legal purposes", JASA 47, 597-612

Brend, R., 1971 "Male-female intonation patterns in American English", Proc. 7th Int. Cong. Phon. Sci., 866-9

Bricker, P.D. e S. Pruzansky, 1976 "Speaker recognition", in N.J. Lass (ed), 1976, 295-326

Bricker, P.D. e S. Pruzansky, 1966 "Effects of stimulus content and duration on talker identification", JASA 40, 1441-9

Broad, D.J. e R. Fertig, 1970 "Formant trajectories in selected CVC-syllable nuclei", JASA 47, 1572-82

Brokx, J.P.L. e S.G. Nooteboom, 1982 "Intonation and the perceptual separation of simultaneous voices", J.Phon. 10, 23-36

Brown Jr., W.S. e S.H. Feinstein, 1975 "Speaker sex identification utilizing a constant laryngeal source", JASA 58, S107,ZZ16

Brown, Murry e Hugues, 1976 "Comfortable effort level: an experimental variable", JASA 60, 696-9

Brown, R., 1981 "An experimental study of the relative importance of acoustic parameters for auditory speaker recognition", Lang.Speech 24, 295-310

Brown, R., 1983 "Parameters in auditory speaker recognition", in Cohen e v.d. Broecke, 1983 (eds), 380

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

Butters, R. R., 1990 "Forensic Linguistics comes of age", American Speech 68, 109-12

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

Chalikia, M.H. e A.S. Bregman, 1989 "The perceptual segregation of simultaneous auditory signals: pulse train segregation and vowel segregation", Perc.Psychoph. 46, 487-96

Cheung, R:S., 1978 "Feature selection using adaptive learning networks for text-independent speaker verification", JASA 64, S183, NNN27

Chevrie-Muller, C., N. Seguier, A. Spira e M. Dordain, 1978 "Recognition of psychiatric disorders from voice quality", Lang.Speech 21, 87-111

Chevrie-Muller, C., P. Sevestre e N. Seguier, 1985 "Speech and Psychopathology", Lang.Speech 28, 57-79

Childers, D.G. e C.K. Lee, 1991 "Vocal quality factors: analysis, synthesis, and perception", JASA 90, 2394-410

Childers, D.G. e K. Wu, 1991 "Gender recognition from speech. Part II: fine analysis", JASA 90, 1841-56

Clarke, F.R. e R.W. Becker, 1969 "Comparison of techniques for discriminating among talkers", JSHR 12, 747-61

Cohen, A., R. Collier e J. t'Hart, 1982 "Declination: construct or intrinsic feature of speech pitch?", Phonetica 39, 254-73

Cohen, J.R., T.H. Crystal, A.S. House e E. P. Neuburg, 1980 "Weighty voices and shaky evidence: a critique", JASA 68, 1884-6

Cohen, A. e M.P.R. van der Broecke (eds) 1983, Abstracts Xth Int. Cong. Phon. Sci., Foris Pub., USA

Coleman, R.O., 1973 "Speaker identification in the absence of inter-subject differences in glottal source characteristics", JASA 53, 1741-3

Coleman, R.O., 1971 "Male and female voice quality and its relationship to vowel formant frequencies", JSHR 14, 565-77

Compton, A., 1963 "Effects of filtering and vocal duration upon the identification of speakers, aurally", JASA 35, 1748-52

Cooper, W.E. e M. Danly, 1981 "Segmental and temporal aspects of utterance-final lengthning", Phonetica 38, 106-15

Cooper, W.E., N. Tye-Murray, E. Tye-Murray e J. Stephen, 1985 "Acoustical cues to the reconstruction of missing words in speech perception", Perc.Psychoph. 38, 30-40

Cox, R.M., G.C. Alexander e C. Gilmore, 1987 "Intelligibility of average talkers in typical listening environments", JASA 81, 1598-608

Crandall, I.B., 1917 "The composition of speech", Physical Review 10, 74-6

Crystal, D., 1969 Prosodic Systems and Intonation in English, Cambridge U. Press

Crystal, T.H. e A.S. House, 1988 "Segmental durations in connected speech signals: syllabic stress", JASA 83, 1574-85

Crystal, T.H. e A.S. House, 1988 "The duration of American English vowels: an overview", J.Phon. 16, 263-84

Crystal, T.H. e A.S. House, 1982 "Segmental durations in connected speech signals: preliminary results", JASA 72, 705-16

Cunha Lima, J.M., 1976 "A perícia da voz", Arquivos da Polícia Civil, Vol. XXVII, 193-5

Das, S.K e W.S. Mohn, 1971 "A scheme for speech processing in automatic speaker verification", IEEE, Trans. Audio Electroacoust., AU-19, 1, 32-43

Dauer, R.M., 1983 "Stress-timing and syllable-timing reanalyzed", J.Phonetics 11, 51-62

David, D. e F. Denes (eds), 1972 Human Communication: A Unified View, McGraw-Hill, Ny

Delgado Martins, M.R., 1986 Sept Études sur la Perception, I.N.I.C., Lisboa

den Os, E., 1985 "Perception of speech rate of Dutch and Italian utterances", Phonetica 42, 124-34

Di Cristo, A. e D.J. Hirst, 1986 "Modelling French micromelody: analysis and synthesis", Phonetica 43, 11-30

Di Paolo, M. e G. Green, 1990 "Juror's beliefs about the interpretation of speaking style", American Speech 65, 304-22

Disner, S.F., 1982 "Stress evaluation and voice lie detection: a review", UCLA Working Papers in Phonetics, 54, 78-92

Dixon, W.J., P. Sampson e P. Mundle, 1990 "One- and two-way analysis of variance with data screening", BMDP manual, programa 7D, Vol. I, Univ. Calif. Press, 189-212

Doddington, G.R., 1985 "Speaker recognition- Identifying people by their voices", Proc. IEEE 73, 11, 1651-64

Doherty, E.T., 1975 "Evaluation of selected acoustic parameters for use in speaker identification", JASA 58, S107

Doherty, T. e H. Hollien, 1978 "Multiple-factor speaker identification of normal and distorted speech", J.Phon. 6, 1-8

Dubois, J., M. Giacomo, L. Guespin, C. Marcellesi, J-B. Marcellesi e J-P. Mével, 1973 Dictionnaire de Linguistique, Larousse, Paris

Dumas, B. K., 1993 "Forensic Linguistics: A brief anthology", American Speech 68, 106-9

Dumas, B.K., 1990 "Voice identification in a criminal law context", American Speech 65, 341-8

Eefting, W., 1992 "The effect of accentuation and word duration on the naturalness of speech", JASA 91, 411-20

Ekman,P., W.V. Friesen e K.R. Scherer, 1976 "Body movement and voice pitch in deceptive interaction", Semiotica 16, 23-7

Emmorey, K., D. van Lancker e J. Kreiman, 1984 "Recognition of famous voices given vowels, words, and two-second texts", UCLA Working Papers 59, 120-4

Endres, W., W. Bambach e G. Flössler, 1971 "Voice spectrograms as a function of age, voice disguise and voice imitation", JASA 49, 1842-48

Engelman, L., 1990 "2D: Detailed data description including frequencies", BMDP Statistical Software Manual V.I, Univ. California Press, 135-44

Engstrand, O., 1988 "Articulatory correlates of stress ans speaking rate in Swedish VCV utterances", JASA 83, 1863-75

Fant, G., 1962 "Sound spectrography", in R.J. Baken e R.G. Daniloff (eds), 1991, 47-66

Fant, G., 1973 Speech Sounds and Features, MIT, Cambridge

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

Fant, G., 1960 Acoustic Theory of Speech Production, Mouton, The Hague

Fant, G., A. Kruckenberg e L. Nord, 1991 "Tempo and stress", Perilus XIII, 31-4

Fant, G., A. Kruckenberg e L. Nord, 1991 "Durational correlates of stress in Swedish, French and English", J.Phon., 19, 351-65

Fay, P.J. e W.C. Middleton, 1940 "Judgement of Kretschmerian body types from the voice as transmitted over a public address system", J.Social Psychology 12, 151-62

Figueiredo, R. M., 1993 "Variabilidade inter- e intra-falante da freqüência fundamental em função da velocidade de emissão", Anais do XLI Seminário do G.E.L., Ribeirão Preto, maio, 1993

Figueiredo, R.M., 1990 Identificação de Vogais: Aspectos Acústicos, Articulatórios e Perceptuais, Dissertação de Mestrado, IEL-UNICAMP

Figueiredo, R.M., F.A.B. Palhares e E.J. Nagle, 1993 "Description of a real case involving the identification of a speaker and the authentication of a magnetic tape recording", Proc. 13th Meeting, IAFS, Düsseldorf, Alemanha, Agosto, 1993, Abst. pg. A44

Firth, J.R., 1950 "Personality and Language in society", The Sociological Review, xlii, 3, rep. em Papers in Linguistics:1934-1951, Oxford U.P., 177-89, 1969

Flanagan, J.L., 1958 "Some properties of the glotal sound source", in D. Fry, 1976, 31-51

Flanagan, J.L., 1972 Speech Analysis, Synthesis and Perception, Springer

Fletcher, F., 1991 "Rhythm and final lengthening in French", J.Phon., 19, 193-212

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

Fowler, C.A. e J. Housum, 1987 "Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction", J.Mem.Lang. 26, 489-504

Fritzell, B. e G. Fant (eds), 1986 Voice Acoustics and Dysphonia, Theme Issue, J.Phon., 1986,14

Fromkin, V.A. (ed), 1985 Phonetic Linguistics: Essays in Honor of Peter Ladefoged, Acad.Press

Fujimura, O., 1962 "Analysis of nasal consonants", JASA 34, 1865-75

Furui, S., 1978 "Effects of long-term spectral variability on speaker recognition", JASA 64, S183, NNN28

Furui, S., 1981 "Cepstral analysis technique for automatic speaker verification", IEEE Trans. ASSP-29, 2, 254-72

Fónagy, I., 1978 "A new method of investigating the perception of prosodic features", Lang.Speech 21, 34-49

Fónagy, I., 1981 "Emotions, Voice and Music", in Sundberg,J. (ed), Research Aspects on Singing, Royal Swedish Academy of Music, Stockolm, 51-79

Fónagy, I. e E. Bérard, 1972 "Il est huit heures: contribution à l'analyse sémantique de la vive voix", Phonetica 26, 157-92

Fónagy, I. e G. Magdics, 1963 "Emotional patterns in intonation and music"

Garrett, K.L. e C.H. Healey, 1987 "An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day", JASA 82, 58-62

Garvin, P.L. e P. Ladefoged, 1963 "Speaker identification and message identification in speech recognition", Phonetica 9, 193-99

Gay, T., 1978 "Effect of speaking rate on vowel formant movements", JASA 63, 223-30

Gay, T., 1981 "Mechanisms in the control of speech rate", Phonetica 38, 148-58

Gelfand, S.A. e S. Silman, 1979 "Effects of small room reverberation upon the recognition of some consonant features", JASA 66, 22-9

Gelfer, M.P., K.P. Massey e H. Hollien, 1989 "The effects of sample duration and timing on speaker identification accuracy by means of long-term spectra", J.Phon. 17, 327-38

Gerstman, L.J., 1968 "Classification of self-normalized vowels", IEEE Trans. Audio Electroacoustics AU-16, 78-80

Glenn, J.W. e N. Kleiner, 1968 "Speaker identification based on nasal phonation", JASA 43, 368-72

Gocke, J.W. e W.A. Oleniewski, 1973 "Voiceprint identification in the courtroom", J.Forensic Sciences 18, 232-6

Goldman-Eisler, F., 1968 Psycholinguistics: Experiments in Spontaneous Speech, Acad. Press, London-NY

Goldman-Eisler, F., 1956 "The determinants of the rate of speech output and their mutual relations", J. Psychosom. Res. 1, 137-43

Goldman-Eisler, F., 1954 "On the variability of the speed of talking and on its relation to the length of utterances in conversations", Br. J. Psychol. 45, 94-107

Goldstein, U.G., 1976 "Speaker-identifying features based on formant tracks", JASA 59, 176-82

Gomes, G. M. P., 1993 "Reconhecimento de locutor, com palavras isoladas, pelo método do alinhamento temporal dinâmico", Anais do XI SBT, Natal, Vol. II, 520-7

Graddol e Swann, 1983 "Speaking fundamental frequency: some physical and social correlates", Lang.Speech 26, 351-66

Grosjean, F., 1979 "A study of timing in a manual and a spoken language: American sign language and English", J. Psycholing. Res. 8, 379-405

Gunter, C. e W. Manning, 1982 "Listener estimations of speaker height and weight in unfiltered and filtered conditions", J.Phon. 10, 251-57

Günzburger, D., A. Bresser e M. Ter. Keurs, 1987 "Voice identification of prepubertal boys and girls by normally sighted and visually handicapped subjects", Lang. Speech 30, 47-58

Haggard, M., S. Ambler e M. Callow, 1970 "Pitch as a voicing cue", JASA 47, 613-17

Hair, G.D. e T.W. Rekieta, 1972 "Automatic speaker verification using phoneme spectra", JASA 51, 131(A)

Hall, M. e O. Tosi, 1975 "Spectrographic and aural examination of professionally mimicked voices", JASA 58, S107

Hammarberg, B., B. Fritzell, J. Gauffin e J. Sundberg, 1986 "Acoustic and perceptual analysis of vocal dysfunction", J.Phon. 14, 533-48

Hargraves, W. e J. Starkweather, 1963 "Recognition of speaker identity", L.Speech 6, 63-7

Hawkins, S. e K.N. Stevens, 1985 "Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels", JASA 77, 1560-75

Hazen, B., 1973 "Effects of differing phonetic contexts on spectrographic speaker identification", JASA 54, 650-60

Hecker, M.H.L., 1971 "Speaker recognition: basic considerations and methodology", JASA 49, 138(A)

Hecker, M.H.L., K.N. Stevens, G. von Bismarck e C.E. Williams, 1968 "Manifestations of task-induced stress in the acoustic speech signal", JASA 44, 993-1001

Helfer, K.S., 1992 "Aging and the binaural advantage in reverberation and noise", JSHR 35, 1394-401

Helfer,K.S e R.A. Huntley, 1991 "Aging and consonant errors in reverberation and noise", JASA 90, 1786-96

Helfrich, H., 1979 "Age markers in Speech", in Scherer, K.R. e H. Giles (eds), 1979, 63-107

Hennessy, J.J. e H.A. Romig, 1971 "Sound, speech, phonetics, and voiceprint identification", J. Forensic Sciences 16, 438-54

Hennessy, J.J. e H.A. Romig, 1971 "A review of the experiments involving voiceprint identification", J. Forensic Sciences 16, 183-98

Hockett, C.F., 1987 Refurbishing our foundations, John Benjamins, Amsterdam-Philadelphia

Hockett,C.F., 1958 A Course in Modern Linguistics, The MacMillan Company

Hoequist Jr., C.,' 1983 "Durational correlates of linguistic rhythm categories", Phonetica 40, 19-31

Hollien, H., 1974 "Peculiar case of 'voiceprints', JASA 56, 210-3

Hollien, H., 1981 Resenha se Tosi, O., 1979: Voice Identification: Theory and Legal Applications, Baltimore, JASA 70, 263-5

Hollien, H., 1990 The Acoustics of Crime: The New Science of Forensic Phonetics, Plenum Press, NY-London

Hollien, H. e T. Shipp, 1972 "Speaking fundamental frequency and chronologic age in males", JSHR 15, 155-9

Hollien, H. e W. Majewski, 1977 "Speaker identification by long-term spectra under normal and distorted speech conditions", JASA 62, 975-80

Hollien, H., C.C. Johnson e E.T. Doherty, 1978 "Speaker identification: new vectors for SAUSI", JASA 64, S182, NNN25

Hollien, H., G. Bennett e M.P. Gelfer, 1983 "Criminal identification comparison: aural versus visual identifications resulting from a simulated crime", J. Forensic Sciences 28, 208-21

Hollien, H., L. Geison e J.W. Hicks, 1987 "Voice stress evaluators and lie detection", J. Forensic Sciences 32, 405-18

Hollien, H., W. Majewski e E.T. Doherty, 1982 "Perceptual identification of voices under normal, stress and disguise speaking conditions", J.Phon.

Holmberg, E.B., R.E. Hillman e J.S. Perkell, 1988 "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice", JASA 84, 511-29

Hombert, J.M., J.J. Ohala, e W.G. Ewan, 1979 "Phonetic explanations for the development of tones", Language 55, 37-58

Honts, C.R., 1987 "Interpreting research on polygraph countermeasures", J. of Police Science and Administration 15, 204-9

Horii, Y., 1979 "Fundamental frequency perturbation observed in sustained phonation", JSHR 22, 5-19

Horii, Y., 1975 "Some statistical characteristics of voice fundamental frequency", JSHR 18, 192-201

Horii, Y., 1980 "Vocal shimmer in sustained phonation", JSHR 23, 202-9

House, A.S., 1957 "Analog studies of nasal consonants", JSHD 22, 190-204

House, A.S. e G. Fairbanks, 1953 "The influence of consonant environment upon the secondary acoustical characteristics of vowels", JASA 25, 105-13

Hudson, A.I. e A. Holbrook, 1981 "A study of the reading fundamental vocal frequency of young black adults", JSHR 24, 197-201

Hunnicutt, S., 1985 "Intelligibility versus redundancy-conditions of dependency", Lang.Speech 28, 48-56

Hurme, P. e A. Sonninen, 1986 "Acoustic, perceptual and clinical studies of normal and dysphonic voice", J.Phon. 14, 489-92

Hydrick, B.M. e G.R. Doddington, 1978 "Performance evaluation of speaker verification in entry control", JASA 64, S182, NNN24

Ichikawa, A., A. Nakajima e K. Nakata, 1978 "Speaker verification from actual telephone voice", JASA 64, S182, NNN26

Ingemann, F., 1968 "Identification of the speaker's sex from voiceless fricatives"< JASA 44, 1142-44

Jakobson, R., 1976 Seis Lições sobre o Som e o Sentido, Moraes Ed., Lisboa-SP

Jassem, W. e K. Kudela-Dobrogowska, 1980 "Speaker independent intonation curves", in Waugh, L. C.H. van Schooneveld (eds), 1980, 135-48

Jetzt, J.J., 1979 "Critical distance measurement of rooms from the sound energy spectral response", JASA

Johnson, C.C., H. Hollien e J.W. Hicks, 1984 "Speaker identification utilizing selected temporal speech features", J.Phon. 12, 319-26

Johnson, K., 1990 "The role of perceived speaker identity in F0 normalization of vowels", JASA 88, 642-54

Johnson, K., D.B. Pisoni e R.H. Bernacki, 1990 "Do voice recordings reveal whether a person is intoxicated?: a case study", Phonetica 47, 215-37

Johnson, N.F., 1987 "A tutorial symposium on dinamic conceptions of vowel perception: an introduction", J.Mem.Lang. 26, 539-41

Joos, M., 1948 Acoustic Phonetics, Language (sup.), 24

Juang, B.H., 1991 "Speech recognition in adverse environments", Computer, Speech and Language 5, 275-94

Karlsson, I., 1986 "Glottal wave froms for normal female speakers", J.Phon. 14, 409-13

Karlsson, I., 1992 "Evaluations of acoustic differences between male and female voices; a pilot study", STL - QPSR 1/1992, 19-31

Kashyap, R.L., 1976 "Speaker recognition from an unknown utterance and speaker-speech interaction", IEEE Trans. ASSP-24, 481-8

Kent, R.D., 1976 "Anatomical and neuromuscular maturation of the speech mechanism", JSHR, 19, 421-47

Kent, R.D. e C. Read, 1992 The Acoustic Analysis of Speech, Singular Pub. Group, S. Diego-California

Kersta, L.G., 1962 "Voiceprint Identification", Nature, 196, 1253-7

Kitzing, P., 1986 "LTAS criteria pertinent to the measurement of voice quality", J.Phon. 14, 477-82

Klatt, D., 1973 "Interaction between two factors thet influence vowel duration", JASA 54, 1102-4

Klatt, D., 1975 "Vowel lengthning is syntactically determined in a connected discourse", J. Phon., 129-40

Klatt, D. e L.C. Klatt, 1990 "Analysis, synthesis, and perception of voice quality variations among female and male talkers", JASA 87, 820-57

Klatt, D.H., 1976 "Linguistic uses of segmental duration in English: acoustic and perceptual evidence", JASA 59, 1208-221

Klatt, D.H., 1975 "Voice onset time, frication, and aspiration in word-initial consonant clusters", JSHR 18, 686-706

Klingholz, F., R. Penning e E. Liebhardt, 1988 "Recognition of low-level alcohol intoxication from speech signal", JASA 84, 929-35

Kohler, K.J., 1983 "Prosodic boundary signals in German", Phonetica 40, 89-134

Kretschmer, E., 1925 Physique and Character, New York

Kuehn, D. e K. Moll, 1976 "A cineradiographic study of VC and CV articulatory velocities". J. Phon. 4, 303-20

Kurowski, K. e S. Blumstein, 1987 "Acoustic properties for place of articulation in nasal consonants", JASA 81, 1917-27

Kuwabara, H. e T. Takagi, 1991 "Acoustic parameters of voice indivituality and voice quality control by analysis-synthesis method", Speech Communication 10, 491-5

Künzel, H.J., 1989 "How well does average fundamental frequency correlate with speaker height and weight?", Phonetica 46, 117-25

Künzel, H.J., 1990 Phonetische Untersuchungen zur Sprecher-Erkennung durch Linguistisch Naive Personen, Franz Steiner Verlag, Stuttgart

La Rivière, C., 1974 "Speaker identification from turbulent portions of fricatives", Phonetica 29, 246-52

La Rivière, C., 1975 "Contributions of fundamental frequency and formant frequencies to speaker identification", Phonetica 31, 185-97

Labov, W., 1972 Sociolinguistic Patterns, Philadelphia Univ. Pennsylvania Press

Ladd, D.R., 1984 "Declination: a review and some hypotheses", Phonology Yearbook 1, 53-74

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 Univ. Press

Ladefoged, P., 1971 Preliminaires to Linguistic Phonetics, Univ. Chicago Press

Ladefoged, P., 1975 A Course in Phonetics, Harcourt Brace Jovanovich Inc.

Ladefoged, P., 1978 "Expectation affects identification by listening", Lang. Speech 21, 373-4

Ladefoged, P., 1984 "Review: The Phonetic Bases of Speaker Recognition by F.J. Nolan", J.Phon. 12, 85-9

Lane, H. e F. Grosjean, 1973 "Perception of reading rate by speakers and listeners", J. Exp. Psychol. 97, 141-7

Lass N.J., P.M. Mertz e K.L. Kimmel, 1978 "The effect of temporal speech alterations on speaker race and sex identifications", Lang. Speech 21, 279-90

Lass, N.J, G.A. Dicola, A.S. Beverly, C. Barbera, K.G. Henry e M.K. Badali, 1979 "The effect of phonetic complexity on speaker height and weight identification", Lang. Speech 22, 297-309

Lass, N.J. (ed), 1984 Speech and Language: Advances in Basic research and Practice, Acad. Press

Lass, N.J. e E.G. Colt, 1980 "A comparative study of the effect of visual and auditory cues on speaker height and weight identification", J.Phonetics 8, 277-85

Lass, N.J. e L.A. Harvey, 1976 "An investigation of speaker photograph identification", JASA 59, 1232-6

Lass, N.J. e M. Davis, 1976 "An investigation of speaker height and weight identification", JASA 59, 700-3

Lass, N.J., A.S. Beverly, D.K. Nicosia e L.A. Simpson, 1978 "An investigation of speaker height and weight identification by means of direct estimations", J.Phonetics 6, 69-76

Lass, N.J., C.A. Hendricks e M.A Iturriaga, 1980 "The consistency of listener judgements in speaker height and weight identification", J.Phon. 8, 439-48

Lass, N.J., D.T. Kelley, C.M. Cunningham e K.J. Sheridan, 1980 "A comparative study of speaker height and weight identification from voiced and whispered speech", J.Phonetics 8, 195-204

Lass, N.J., G.W. Ciccolella, S.C. Walters e E.L. Maxwell, 1980 "An investigation of speaker height and weight discriminations by means of paired comparison judgements", J.Phonetics 8, 205-12

Lass, N.J., J.K. Phillips e C.A. Bruchey, 1980 "The effect of filtered speech on speaker height and weight identification", J. Phonetics 8, 91-100

Lass, N.J., P.J. Barry, R.A. Reed, J.M. Walsh e T.A. Amuso, 1979 "The effect of temporal speech alterations on speaker height and weight identification", Lang.Speech 22, 163-71

Laver, J., 1980 The Phonetic Description of Voice Quality, Cambridge Univ. Press

Laver, J. e P. Trudgill, 1979 "Phonetic and linguistic markers in speech", in Scherer, K.R. e H. Giles (eds), 1979, 1-32

Lehiste, I., 1973 "Rhythmic units and syntactic units in production and perception", JASA 54, 1228-34

Lehiste, I., 1972 "The timing of utterances and linguistic boundaries", JASA 51, 2018-24

Lenneberg, E.H., 1967 "The problem of the organizing principle: rhythm", capítulo de Biological Foundations of Language, John Wiley & Sons, NY, 107-24

Levin, H., C.A. Schaffer e C. Snow, 1982 "The prosodic and paralinguistic features of reading and telling stories", Lang, Speech 25, 43-54

Levine, S., 1971 "Stress and behavior", Scientific American 224, 26-31

Li, K.P., G.W. Hugues e A.S. House, 1969 "Correlation characteristics and dimensionality of speech spectra", JASA 46, 1019-25

Li, K.P., J.E. Dammann e W.D. Chapman, 1966 "Experimental studies in speaker verification, using an adaptive system", JASA 40, 966-78

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

Lieberman, P., 1977 Speech Physiology and Acoustic Phonetics: An Introduction, MacMillan Pub. Co. Inc., NY

Lieberman, P., 1980 "The innate, central aspect of intonation", in Waugh, L. e C.H. van Schooneveld (eds), 1980, 187-99

Lieberman, P. e S.B. Michaels, 1962 "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech", JASA 34, 922-7

Lieberman, Ph., 1963 "Some effects of semantic and grammatical context on the production and perception of speech", Lang.Speech 6, 172-87

Lin, W.C. e S.K. Pillay, 1975 "Selection of features and speech segments for speaker verification", JASA 58, S107,ZZ12

Lindblom, B., 1962 "Accuracy and limitations of Sona-Graph measurements", Proc. IVth Int. Cong. hon. Sci., 188-202, in Baken, R.J. e R.G. Daniloff (eds), 1991, 34-46

Lindblom, B., 1963 "Spectrographic study of vowel reduction", JASA 35, 1773-81

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

Lindblom, B. e J. Lubker, 1985 "The speech homunculus and a problem of Phonetic Linguistics", in Fromkin, V.A. (ed), 1985, 169-92

Lindblom, B. e M. Studdert-Kennedy, 1967 "On the role of formant transitions in vowel recognition", JASA 42, 830-43

Lindblom, B. e S-J. Moon, 1988 "Formant undershoot in clear and citation-form", Perilus VIII, 20-33

Linville, S.E., 1988 "Intraspeaker variability in fundamental frequency stability: an age related phenomenon?", JASA 83, 741-5

Lisker, L. e A.S. Abramson, 1964 "A cross-language study of voicing in initial stops: Acoustical measurements", Word 20, 384-422

Luck, J.E., 1969 "Automatic Speaker Verification using cepstral measurements", JASA 46, 1026-32

Lummis, R.C., 1973 "Speaker verification by computer using speech intensity for temporal registration", IEEE Trans. Audio Electroacoust. AU-21, 80-9

Lummis, R.C. e A.E. Rosenberg, 1972 "Test of an automatic speaker verification method with intensively trained professional mimics", JASA 51 (abst.)

Lutz, G.M., 1983 Understanding Social Statistics, MacMillan Pub. Co. Inc., NY

Lyberg, B., 1979 "Final lengthening - partly a consequence of restrictions on the speed of fundamental frequency change?", J.Phon. 7, 187-96

Löfqvist, A., 1986 "The long-time-average spectrum as a tool in voice research", J.Phon. 14, 471-6

Löfqvist, A., 1992 "Acoustic and aerodynamic effects of interarticulator timing in voiceless consonants", Lang. Speech 35, 15-28

Majewski, W. e H. Hollien, 1974 "Euclidean distance between long-term speech spectra as a criterion for speaker identification", Speech Communication Seminar, Stockolm, aug., 1-3

Major, R.J., 1981 "Stress-timing in Brazilian Portuguese", J.Phon. 9, 343-51

Marcus, P., 1981 "Acoustic determinants of perceptual center (p-center) location", Perc.Psychoph. 30, 247-56

Markel, J.D., 1972 "Digital inverse filtering - a new tool for formant trajectory estimation", IEEE Trans. Audio Electroacoust., AU-20, 129-37

Markel, J.D. e A.H. Gray Jr., 1976 Linear Prediction of Speech, Springer Verlag, NY-Berlin

Markel, J.D. e S.B. Davis, 1979 "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base", IEEE Trans., ASSP-27, 1, 74-82

Markel, J.D., B.T. Oshika e A.H. Gray Jr., 1977 "Lomg term feature averaging for speaker recognition", IEEE Trans., ASSP, 25, 4, 330-7

Matsumoto, H., S. Hiki, T. Sone e T. Nimura, 1973 "Multidimensional representation of personal quality and its acoustical correlates", IEEE Trans. Audio Electro-Acoust., AU-21, 428-36

Mattoso Câmara Jr., J., 1980 Princípios de Lingüística Geral, Padrão-Liv. Editora Ltda., RJ

McCandless, S.S., 1974 "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-22, 135-41

McCroskey, R.L., 1984 "Auditory timing: its role in speech-language pathology", in Lass, N.J. (ed), 1984,141-84

McGehee, F., 1944 "An experimental study in voice recognition", J.Gen.Psychol. 31, 53-65

McGehee, F., 1937 "The reliability of the identification of the human voice", J.Gen.Psychology 17, 249-71

Miller, J.D., 1989 "Auditory-perceptual interpretation of the vowel", JASA 85, 2114-34

Miller, J.L., F. Grosjean e C. Lomanto, 1984 "Articulation rate and its variability in spontaneous speech: a reanalysis and some implications", Phonetica 41, 215-25

Miller, R.L., 1953 "Auditory tests with synthetic vowels", JASA 25, 114-21

Monsen, R.B. e A.M. Engebretson, 1983 "The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction", JSHR 26, 89-97

Moon, S-J., 1991 "An acoustic and perceptual study of undershoot in clear and citation-form speech", Perilus, 153-6

Moses, P.J., 1941 "Theories regarding the relation of constitution and character through the voice", Psychological Bulletin 38, 746

Nakatani, L.H., K.D. O'Connor e C.H. Aston, 1981 "Prosodic aspects of American English speech rhythm", Phonetica 38, 84-106

Nardini, W., 1987 "The polygraph technique: an overview", J. of Police Science and Administration 15, 239-49

Nearey, T., 1989 "Static, dynamic and relational properties in vowel perception", JASA 85, 2088-113

Niederjohn, R.J. e M. Lahat, 1985 "A zero-crossing consistency method for formant tracking of voiced speech in high noise levels", IEEE Trans. ASSP-33, 2, 349-55

Nilsonne et al., 1988 "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression", JASA 83, 716-28

Nolan, F., 1983 The Phonetic Bases of Speaker Recognition, Cambridge

Nord, L., 1991 "Rhythmical- In what sense? Some preliminary considerations", Perilus XIV, 107-11

Nushikyan, E., 1987 "The typological analysis of emotional speech prosody", Proc. XIth Cong, Phon. Sci., Estonia, 1987

Ohala, J., 1972 "How is pitch lowered?", JASA 52, 124

Ohala, J. e J.J Jaeger (eds), 1986 Experimental Phonology, Acad. Press, NY

Öhman, S.E.G., 1966 "Corticulation in VCV utterances: Spectrographic measurements", JASA 39, 151-68

Olive, J., 1971 "Automatic formant tracking by a Newton-Raphson technique", JASA 50, 661-70

Oller, D.K., 1973 "The effect of position in utterance on speech segment duration in English", JASA 54, 1235-47

Osgood et al., 1957 The Measurement of Meaning, Univ. of Illinois Press, 1971

Paliwal, K.K., 1984 "Effectiveness of different vowel sounds in automatic speaker identification", J.Phon 12, 17-21

Papamichalis, P.E. e G.R. Doddington, 1984 "A speaker recognizability test", Proc. ICASSP-84, 18B.6

Papçun, G. e P. Ladefoged, 1974 "Two 'voiceprint' cases", JASA 55, 463,LL15

Paul, J.E., A.S. Rabinowitz, J.P. Riganati e J.M. Richardson, 1975 "Development of analytical methods for a semi-automatic speaker identification system", Proc. Carnaham Conf. on Crime Countermeasures, 52-64, Univ. Kentucky

Peterson, G.E. e H.L. Barney, 1952 "Control methods in a study of the vowels", JASA 24, 175-84

Pickett, J.M., 1980 The Sounds of Speech Communication, PRO-ED Inc., Austin

Pierrehumbert, J., 1979 "The perception of fundamental frequency declination", JASA 66, 363-9

Pike, K., 1945 Intonation of American English, Ann Arbor, Univ. Michigan Press

Pittam, J., 1987 "The long term spectral measurement of voice quality as a social and personality marker: a review", Lang. Speech 30, 1-12

Pollack, I., J.M. Pickett e W.H Sumby, 1954 "On the identification of speakers by voice", JASA 26, 403-6

Pollock, K.E, D.M. Brammer e C.F. Hageman, 1993 "An acoustic analysis of young children's productions of word stress", J.Phon. 21, 183-203

Potter, R., G. Kopp e H. Green, 1947 Visible Speech, Dover, NY, 1966 (re-edição do original de, 1947)

Price, P.J., M. Ostendorf, S. Shattuch-Hufnagel e C. Fong, 1991 "The use of prosody in synthatic disambiguation", JASA 90, 2956-70

Pruzansky, S., 1963 "Pattern-matching procedure for automatic talker recognition", JASA 35, 354-8

Ptacek, P.H. e E.K. Sander, 1966 "Age recognition from voice", JSHR 9, 273-7

Ramig, L.A. e R.L. Ringel, 1983 "Effects of physiological aging on selected acoustic characteristics of voice", JSHR 26, 22-30

Reich, A.R., 1981 "Detecting the presence of vocal disguise in the male voice", JASA 69, 1458-61

Reich, A.R. e J.E Duke, 1979 "Effects of selected vocal disguises upon speaker identification by listening", JASA 66, 1025-8

Reich, A.R., K.L. Moll e J.F. Curtis, 1976 "Effects of selected vocal disguises upon spectrographic speaker identification", JASA 60, 919-25

Riordan, C.J., 1980 "Larynx heigth during english stop consonants", J.Phon. 8, 353-60

Roach, P., 1982 "On the distinction between stress-timed and syllable-timed languages", in Crystal, D. (ed.), 1982, Linguistic Controversies, Edward Arnold, London, 73-9

Rosenberg, A., 1973 "Listener performance in speaker verification tasks", IEEE, Trans. Audio Electroacoustics, AU-21, 221-5

Rosenberg, A.E., 1975 "Evaluation of an automatic speaker verification system over telephone lines", JASA 57, S23, L1

Rosenberg, A.E., 1976 "Automatic speaker verification: a review", Proc. IEEE 64, 4, 475-87

Rosenberg, A.E., 1973 "Listener performance in speaker verification tasks", IEEE Trans. Audio Electroacoust., AU-21, 3, 221-5

Rosenberg, A.E., 1971 "Listener performance in a speaker verification task", JASA 50, 106A

Rosenberg, A.E. e M.R. Sambur, 1975 "New techniques for automatic speaker verification", IEEE Trans. Acoust. , Speech Signal Processing, ASSP-23, 169-76

Ross, E.D., J.A. Edmondson e G. Burton Seibert, 1986 "The effect of affect on various acoustic measures of prosody in tone and non-tone languages: a comparison based on computer analysis of voice", J.Phon. 14, 283-302

Rothman, H.B., 1975 "Perceptual (aural) and spectrographic investigation of speaker homogeneity", JASA 58, S107, ZZ13

Salmon, V., 1986 "Security by masking", JASA 79, 2077-8

Sambur, M.R., 1975 "Speaker recognition using orthogonal linear prediction", JASA 58, S107

Sambur, M.R., 1975 "Selection of acoustic features for speaker identification", IEEE, Trans. ASSP-23, 176-82

Sambur, M.R. e N.S. Jayant, 1976 "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise", IEEE Trans. ASSP-24, 6, 488-94

Sapir, E., 1927 "A fala como traço de personalidade", in Lingüística como Ciência, Liv. Acadêmica, 1969, RJ, 63-78

Schaefe, R.W. e L.R. Rabiner, 1970 "System for automatic formant analysis of voiced speech", JASA 47, 634-48

Schenker, S., 1982 "Effects of alcohol on the brain", in Lieber, C.S. (ed), Medical Disorders,of Alcholism, Pathogenesis and Treatment, Saunders-California

Scherer, K.R., 1978 "Inference rules in personality attribution from voice quality: the loud voice of extroversion", European Journal of Social Psychology 8, 467-87

Scherer, K.R., 1979 "Personality markers in speech", in Scherer, K.R. e H. Giles, 1979 (eds), 147-209

Scherer, K.R. e H. Giles (eds), 1979 Social Markers in Speech, Cambridge

Scherer, K.R., S. Feldstein, R.N. Bond e R. Rosenthal, 1985 "Vocal cues to deception: a comparative channel approach", J.Psycholing.Res. 14, 409-25

Schmidt-Nielen, A. e R. Stern, 1985 "Identification of known voices as a function of familiarity and narrow-band coding", JASA 77, 658-63

Schulman, R., 1990 "Articulatory dynamics of loud and normal speech", JASA 85, 295-312

Schwartz, M., 1968 "Identification of speaker sex from isolated, voiceless fricatives", 43, 1178-9

Schwartz, M.F., 1972 "Influence of utterance length upon bilabial closure duration for /p/", JASA 51, 666

Schwartz, M.F. e H.E. Rine, 1968 "Identification of speaker sex from isolated, whispered vowels", JASA 44, 1736-7

Shen, X.S., 1992 "A pilot study on the relation between the temporal and synthatic structures in Mandarim", J. IPA 22, 35-43

Shipp, T. e H. Hollien, 1969 "Perception of the aging male voice", JSHR 12, 703-10

Shoup, J.E. e L.L. Pfeifer, 1976 "Acoustic characteristics of speech sounds", in Lass, N., 1976 (ed), 172-224

Shuy, R. W., 1993 "Risk, deception, confidentiality, and informed consent", American Speech 68, 103-6

Shuy, R. W., 1993 "Using language evidence in money-laundering trials", American Speech 68, 3-19

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

Smith, J.E.K., 1962 "Decision-theoretic speaker recognizer", JASA 34,1988

Smith, P.M., 1979 "Sex markers in speech", in Scherer, K.R. e H. Giles, 1979 (eds), 104-46

Smrkovski, L.L., 1981 "Forensic Voice Identification", Michigan Department of State Police

Sonoda, Y., 1987 "Effect of speaking rate on articulatory dynamics and motor event", J.Phon. 15, 145-56

Steele, S., 1986 "Interaction of vowel f0 and prosody", Phonetica 43, 92-105

Sternberg, S., R.L. Knoll, S. Monsell e C.E. Wright, 1988 "Motor programs ans hierarquical organization in the control of rapid speech", Phonetica 45, 175-97

Stetson, R.H., 1951 Motor Phonetics, The Hague-Amsterdam

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

Stevens, K.N., 1989 "On the quantal nature of speech", J.Phon. 17, 3-45

Stevens, K.N., C.E. Williams, J.R. Carbonell e B. Woods, 1968 "Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material", JASA 44, 1596-607

Stevens, S.S., J.P. Egan e G.A. Miller, 1947 "Methods of measuring speech spectra", JASA, 19, 771-80

Streeter, L., R. Krauss, V. Geller, C. Olson e W. Apple, 1977 "Pitch changes during attempted deception", J. Personality and Social Psychology 35, 345-50

Su, L-S., K.P. Li e K.S. Fu, 1974 "Identification of speakers by use of nasal coarticulation", JASA 56, 1876-82

Sundberg, J., 1986 "Session 5. Long Term Average Spectrum Analysis: Chairman's Summary", J.Phon. 14, 493-4

Sundberg, J. e B. Lindblom, 1990 "Acoustic estimations of the front cavity in apical stops", JASA 88, 1313-7

Sundberg, J., C. Johansson, H. Wilbrand e C. Ytterberg, 1987 "From sagital distance to area: a study of transverse, vocal tract cross-sectional area", Phonetica 44, 76-90

Sundberg, J., S. Ternström, W.H. Perkins e P. Gramming, 1988 "Long-term average spectrum analysis of phonatory effects of noise and filtered auditory feedback", J.Phon. 16, 203-19

Syrdal, A.K. e H.S. Gopal, 1986 "A perceptual model of vowel recognition based on the auditory representation of American English vowels", JASA 79, 1086-100

Tartter, V.C., 1991 "Identifiability of vowels and speakers from whispered syllables", Perc.Psychoph. 49, 365-72

Tate, D.A., 1979 "Preliminary data on dialect in speech disguise", in Hollien, H. e P. Hollien (eds), Current Issues in the Phonetic Sciences, John Benjamins, Amsterdam

Ternström, S., J. Sundberg e A. Colldén, 1988 "Articulatory f0 perturbations and auditory feedback", JSHR 31, 187-92

Thorsen, N.G., 1980 "A study of the perception of sentence intonation: evidence from Danish", JASA 63, 1014-30

Tiersma, P. M., 1993 "Linguistic issues in the law", Language 69, 1, 113-37

Titze, I.R., 1987 "Physiology of the female larynx", JASA, Sup. 1, 82, S90

Titze, I.R., Y. Horii e R.C. Scherer, 1987 "Some technical considerations in voice perturbation measurements", JRHR 30, 252-60

Tobias, J.V., 1959 "Relative ocurrence of phonemes in American English", JASA 31, 631(L)

Tosi, O., H. Oyer, W. Lashbrook, C. Pedrey, J. Nicol e E. Nash, 1972 "Experiment on volce identification", JASA 51, 2030-43

Trager, G.L., 1958 "Paralanguage: a first approximation", in Hymes, D., 1958 (ed), 274-88

Traunmüller, H., 1988 "Paralinguistic variation and invariance in the characteristics frequencies of vowels", Phonetica 45, 1-29

Trojan, F. e K. Kryspin-Exner, 1968 "The decay of articulation under the influence of alcohol and paraldehyde", Folia Phoniatrica 20, 217-39

Tuller, B., K.S. Harris e J.A. Scott Kelso, 1982 "Stress and rate: differential transformations of articulation", JASA 71, 1534-43

Uldall, E., 1961 "Dimensions of meaning in intonation", in Intonation, Penguim Books, 250-9

Umeda, N., 1982 "'F0 declination' is situation dependent", J.Phon. 10, 279-90

Umeda, N., 1981 "Influence of segmental factors on fundamental frequency in fluent speech", JASA 70, 350-5

Vaissière, J., 1988 "Prediction of velum movement from phonological specifications", Phonetica 45, 122-39

van Dommelen, W.A., 1993 "Speaker height and weight identification: a re-evaluation of some old data", J.Phon. 21, 337-41

van Dommelen, W.A. e A. Win, 1987 "The contribution of speech rhythm and pitch to speaker recognition", Lang.Speech 30, 325-38

van Lancker, D., J. Kreiman e T.D. Wickens, 1985 "Familiar voice recognition: patterns and parameters- part II: recognition of rate-altered voices", J.Phon 13, 39-52

van Lancker, D., J. Kreiman e K. Emmorey, 1985 "Familiar voice recognition: patterns and parameters. PArt I: recognition of backward voices", J. Phonetics 13, 19-38

van Lancker, d. e J. Kreiman, 1985 "Unfamiliar voice discrimination and familiar voice recognition are independent and unordered abilities", UCLA Working Papers 62, 50-60

van Riper, C. e J.V. Irwin, 1958 Voice and Articulation, Prentice-Hall, Englewood Cliffs

van Son, R.J.J.H. e L.C.W. Pols, 1990 "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", JASA 88, 1683-93

van Summers, M. et al., 1988 "Effects of noise on speech production: acoustic and perceptual analyses", JASA 84, 917-28

Vaz, N. M., 1983 "Idéias para uma nova imunologia", Ciência Hoje II, 7,32-8

Voiers, W.D., 1964 "Perceptual bases of speaker identity", JASA 36, 1065-73

Wakita, H., 1975 "On the use of linear prediction error energy for speech and speaker recognition", JASA 57, S23, L5

Walden, B.E., A.A. Montgomery, G.J. Gibeily, R.A. Prosek e D.M. Schwartz, 1978 "Correlates of psychological dimensions in talker similarity", JSHR 21, 265-75

Warner, R. e K. Mooney, 1988 "Individual differences in vocal activity rhythm: Fourier analysis of ciclicity in amount of talk", J.Psicholing.Res. 17, 99-111

Warner, R.M., 1979 "Periodic rhythms in conversational speech", Lang.Speech 22, 381-96

Waugh, L.R. e C.H. van Schooneveld (eds), 1980 The Melody of Language, Univ. Park Press, Baltimore

Weinstein, C.J., 1991 "Opportunities for advanced speech processing in military computer-based systems", Proc. IEEE 79, 11, 1626-41

Wendler, J., A. Rauhut e H. Krüger, 1986 "Classification of voice qualities", J.Phon. 14, 483-8

Williams, C.E. e K.N. Stevens, 1972 "Emotions and speech: some acoustical correlates", JASA 52, 1238-50

Wolf, J.J., 1972 "Efficient acoustic parameters for speaker recognition", JASA 51, 2044-56

Wood, C.A., 1978 "Speaker identification by analysis of sound islands", JASA 64, S183, NNN29

Wright, J.T., 1986 "The behavior of nasalized vowels in the perceptual vowel space", in Ohalâ, J.J. e J.J. Jaeger, 1986 (eds), 45-67

Wu, K. e D.G. Childers, 1991 "Gender recognition from speech. Part I: coarse analysis", JASA 90, 1828-40

Young, M.A. e R.A. Campbell, 1967 "Effects of context on talker identification", JASA 42, 1250-4

Zalewski, J., W. Majewski e H. Hollien, 1975 "Cross correlation of long-term speech spectra as a speaker identification technique", Acustica 34, 20-4

Zawadzki, P.A. e H.R. Gilbert, 1989 "Vowel fundamental frequency and articulator position", J.Phon. 17, 159-66

Zlatoustova, L.V. e G.Y. Kedrova, 1987 "Perceptive and acoustic characteristics of emotions", Proc. XIth Cong. Phon Sci., Estonia, 1987, 218-21

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

ANEXO

TEXTO I (extraído de VAZ 1983 (pg. 33, item 2)

A reatividade dos linfócitos, //as células do sangue que fabricam anticorpos, //, é individualizada.// Em cada organismo, as células do fígado são provavelmente iguais entre si,// as da pele também, //mas os linfócitos são diferentes uns dos outros.// Cada um difere do seguinte por possuir na membrana diferentes receptores,// moléculas que garantem a aderência a certas estruturas (ou a capacidade de fixar certas substâncias) [1].// Assim, o linfócito seguinte adere a estruturas diferentes. //Para ser mais exato, as diferenças existem entre clones de linfócitos.// Quando um determinado linfócito se multiplica e gera duas, quatro, oito milhares de cópias idênticas [2],// este conjunto constitui um clone linfocitário. //Dentro de um mesmo clone, os linfócitos são iguais:// têm os mesmos receptores de membrana,// aderem às mesmas coisas,// participam das mesmas interações.

Obs:

- até [1] : trecho lido utilizado para cálculo do ELT na velocidade normal
- até [2] : trecho lido utilizado para cálculo do ELT na velocidade rápida

- as barras duplas {//} marcam os limites dos "blocos de fala" utilizados na seção 7

-*TEXTO II* (extraído do Jornal *Correio Popular*, Campinas, 6/6/92, pg. 21)

O Santos entra em campo amanhã no Maracanã preparado para enfrentar uma verdadeira guerra por parte de diretores, jogadores e torcedores vascaínos. A estratégia utilizada durante toda a semana pelo Vasco, com o objetivo de caracterizar o Santos como uma equipe violenta e com isso pré-condicionar a arbitragem, chegou em alguns momentos a criar um clima tenso na Vila Belmiro. No entanto, depois que os jogadores e o técnico Geninho conversaram e decidiram denunciar a manobra, a tranqüilidade voltou a tomar conta da equipe [1]. A indicação do juiz Márcio Resende de Freitas contribuiu para isso. Para Geninho, ele é um árbitro equilibrado e certamente saberá distinguir o que é um jogo duro e o que é violência. Márcio Resende é o árbitro que será o representante brasileiro nas Olimpíadas de Barcelona.

A preocupação de Geninho agora é só montar a equipe. Ele conta com a volta do garoto Axel, expulso contra o Bahia e que retorna contra o Vasco para jogar como cabeça de área. A única dúvida continua em relação a quem substituirá o meio campista Sérgio Manoel, punido com três cartões amarelos [2]

até [1] : trecho lido utilizado para cálculo do ELT (primeira metade)
de [1] a [2] : trecho lido utilizado para cálculo do ELT (segunda metade)