Cad.Est.Ling., Campinas, (25):129/160, Jul./Dez. 1993

A EFICÁCIA DE MEDIDAS EXTRAÍDAS DO ESPECTRO DE LONGO TERMO PARA A IDENTIFICAÇÃO DE FALANTES

RICARDO MOLINA DE FIGUEIREDO
IEL/UNICAMP

1) INTRODUÇÃO

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, ou seja, 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 esses procedimentos precisa, no entanto, ser estabelecida. O fato é que 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 através 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**.

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 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).

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 fonético 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.

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).

Vários fatores que podem, eventualmente, alterar a performance do ELT como discriminador de falantes já foram abordados: o efeito da limitação de banda (Doherty e Hollien 1978; Hollien e Majewski 1977), da duração do enunciado (Gelfer et al. 1989), da intensidade da voz (Doddington 1985:1659), da não-contemporaneidade das amostras (Gelfer et al. 1989), da presença de *stress* psicológico artificialmente induzido (Doherty e Hollien 1978; Hollien e Majewski 1977), da tentativa de disfarce (Hollien e Majewski 1977; Doherty 1975), da língua específica (Hollien e Majewski 1977; Pittam 1987) e do tipo de microfone utilizado na captação do sinal (Doddington 1985:1659). O presente estudo pretende abordar 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[1] - 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).

---

[1] 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)

2) METODOLOGIA

2.1) *Sujeitos e Amostras de Fala*

As amostras de fala utilizadas no presente estudo foram produzidas por 10 indivíduos do sexo masculino, com idades entre 22 e 45 anos, todos gozando de bom estado geral de saúde e sem exibir distúrbios aparentes de voz. Apenas um dos falantes (R1) apresentava um forte resfriado com quadro de inflamação laríngea e provável presença de muco nas cavidades nasais; esse mesmo falante fez uma segunda gravação, com um intervalo de quatro meses da primeira, já sem qualquer sinal de resfriado (nessa segunda gravação esse falante foi identificado como R2). A inclusão das produções do falante R1/R2 acrescenta, portanto, duas condições adicionais ao estudo: a existência do resfriado (com a possível alteração de ressonâncias nasais, e a eventual alteração do ELT) e, acumulativamente, a comparação de amostras não contemporâneas do mesmo falante.

Oito falantes do grupo total leram um texto com 126 palavras extraído de artigo científico (Vaz 1983; texto I, no apêndice). Os dois gêmeos, JA e JR, leram texto de 241 palavras, extraído de noticiário esportivo em jornal diário (texto II, no apêndice). Todos os falantes, com exceção dos gêmeos JA e JR, leram o texto em duas condições de velocidade: normal e rápida. As instruções para a leitura em cada velocidade foram assim explicitadas:

Normal: "leia o seguinte texto (texto I) da forma mais confortável possível para você"
Rápida: "leia o seguinte texto (texto I) o mais rapidamente possível, evitando, no entanto, perda de inteligibilidade"
(obs: os gêmeos JA e JR leram o texto II apenas na velocidade normal.)

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, mantendo 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.

2.2) *Procedimento de Análise (ELT)*

Para obter os ELTs calculou-se um espectro médio 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 sonógrafo da KAY Elemetrics, modelo 5500, empregando o *setup* #04.

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).

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. Os trechos utilizados para a extração dos ELTs estão assinalados no apêndice. 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 (10 + R2). Os gêmeos não produziram amostras na condição velocidade rápida de produção; para obter os dois ELTs foram utilizados, então, dois diferentes trechos da mesma leitura do Texto II (esses trechos estão assinalados no apêndice) com cerca de 25 segundos cada.

A figura 1 mostra ELTs de dois falantes do grupo, extraídos segundo os critérios acima expostos.

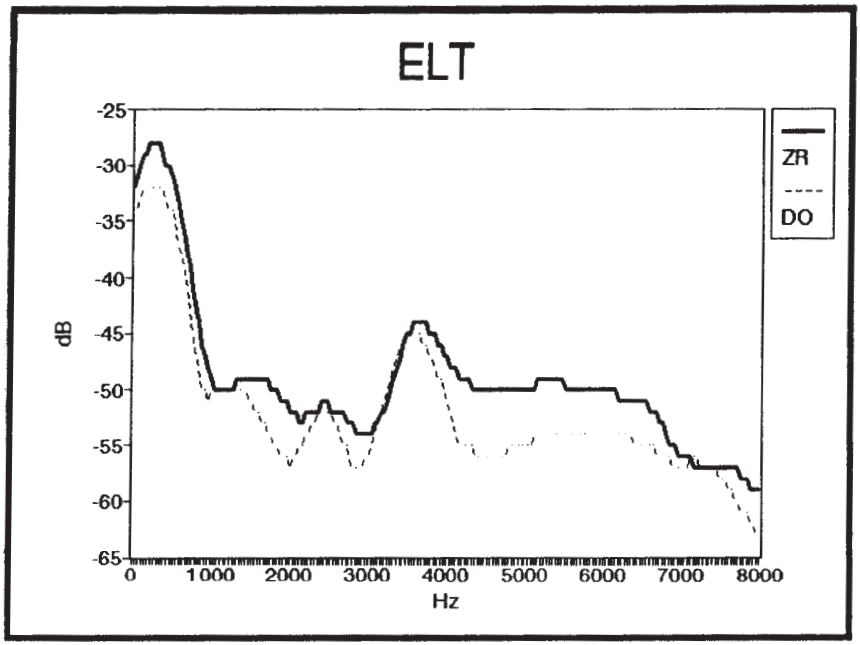

figura 1: ELTs de dois falantes (ZR e DO), incluindo trechos não-vozeados (faixa 0-8000 Hz)

3) ANÁLISE ESTATÍSTICA

3.1) *Análise Cluster a partir de 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 (Univ. da California), 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 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} (Olivier, Com. Pess.).

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.

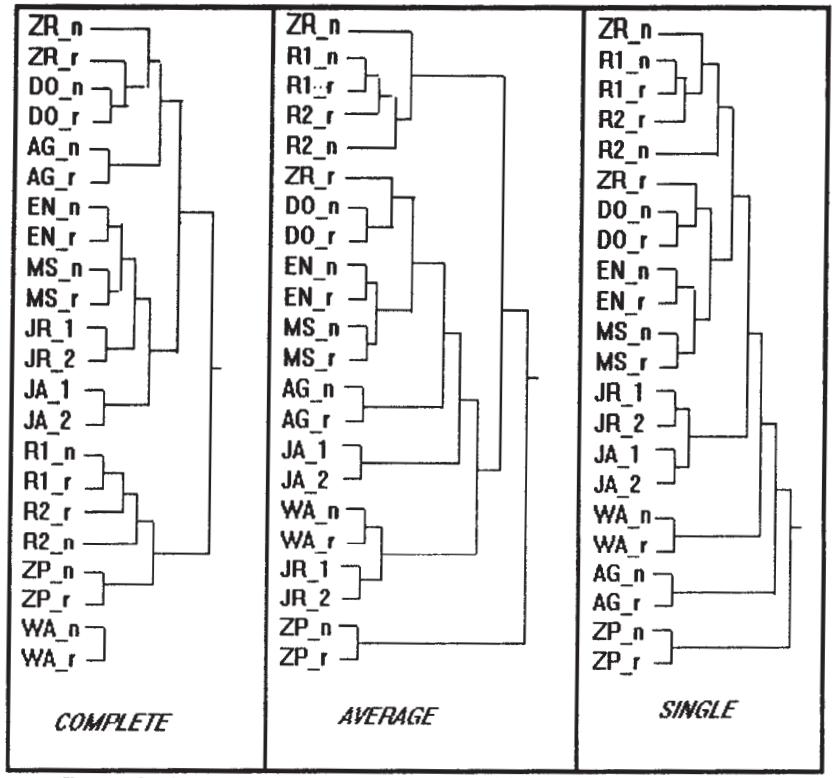

figura 2: dendrogramas resultantes de análise *cluster;* a estrutura do dendrograma reflete o grau de similaridade entre as diferentes amostras, a partir de medidas extraídas do ELT.

Ainda com relação à figura 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 indistingüíveis. 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).

3.2) *Análise Cluster a partir de faixas selecionadas do ELT*

Na seção anterior (3.1) utilizou-se informação do ELT considerando a faixa 0 - 8000Hz. 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, 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 1 resume os resultados dos testes estatísticos apresentando o número de acertos[2] (*clusters* iniciais agrupando pares corretos de falantes) para os três métodos de amalgamação fornecidos em BMDP-1M.

---

[2] 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.

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 1: número de identificações corretas em três diferentes métricas, a partir de faixas selecionadas do ELT

Observamos na tabela 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 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 3 mostra, graficamente, o numero 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 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.

+

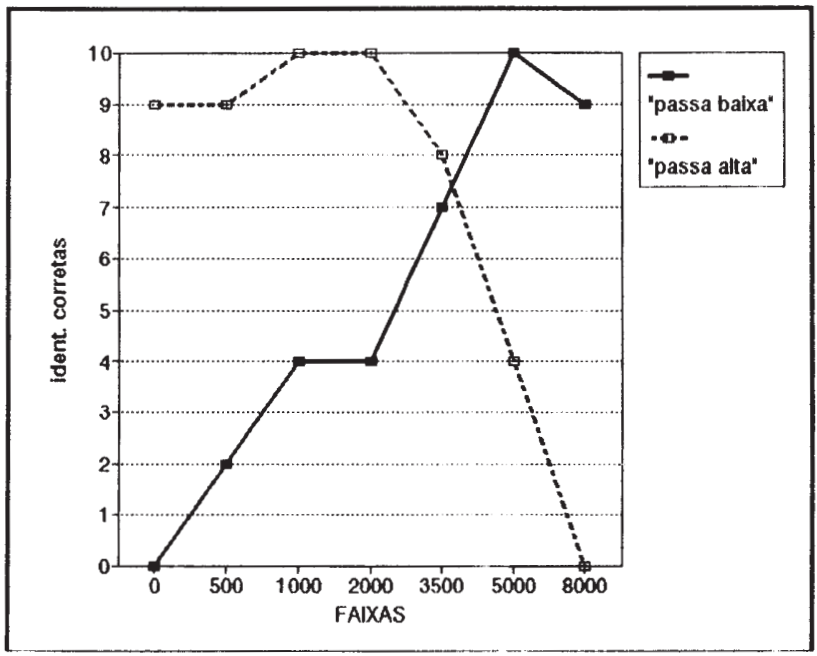

figura 3: identificações corretas a partir de faixas selecionadas do ELT

A figura 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.

O gráfico indica 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.

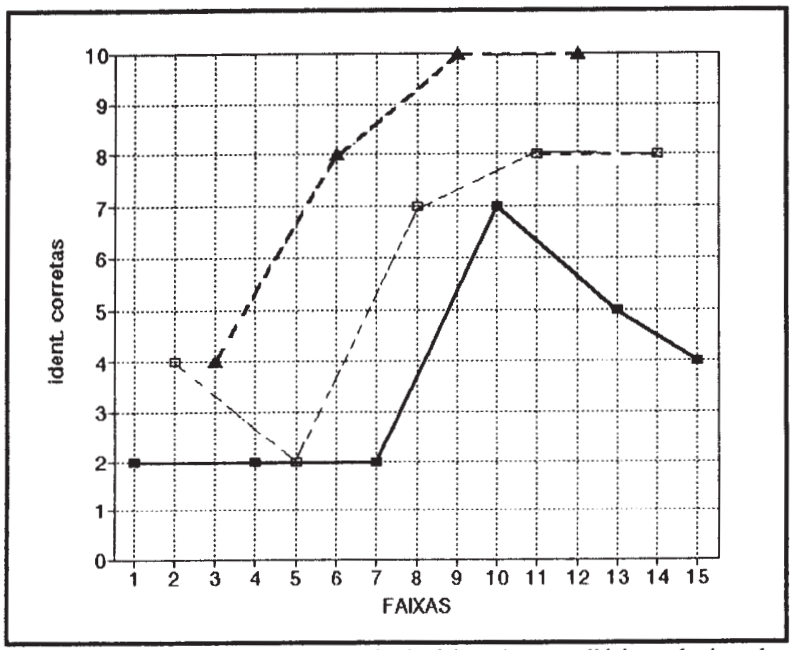

figura 4: identificações corretas a partir de faixas intermediárias selecionadas do ELT (1=0-500; 2=0-1000; 3=0-2000; 4=500-1000; 5=500-2000; 6=500-3500; 7=1000-2000; 8=1000-3500; 9=1000-5000; 10=2000-3500; 11=2000-5000; 12=2000-8000; 13=3500-5000; 14=3500-8000; 15=5000-8000)

As faixas menos informativas são as de 500 - 2000 Hz e 5000 - 8000 Hz. 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 mais eficiente para distingüir 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).

3.3) *Análise cluster com 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). 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. Dessa forma, obtemos diretamente o decaimento em dB/oitava. Cada ELT fica assim reduzido a apenas 6 *slopes*, como ilustra a figura 5.

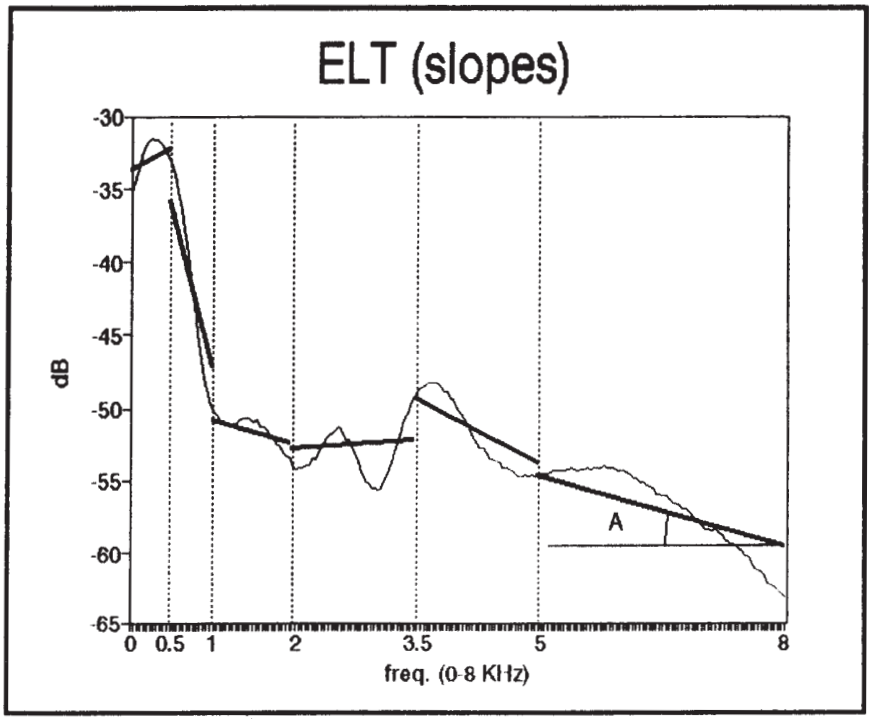

figura 5: *slopes* extraídos de faixas selecionadas do ELT; cada *slope* é expresso pela tangente do ângulo que a reta ajustada faz com o eixo das freqüências (o ângulo A, por exemplo, na faixa 5-8 KHz)

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 SL_0_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 1 acima).

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). Assim, o nível normalizado da faixa 0-500 Hz, por exemplo, será:

$$L0\text{-}5_{normal} = L0\_5 \div L0\_80$$

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 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 2

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.

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 (*tilt*) 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).

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 (Figueiredo 1993). 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 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 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 -6dB/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.

SLOPES **(dB/oitava)**

Falantes

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 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 faixa1000-2000 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) não é adequada, podendo gerar índicadores espúrios.

Tomando como base uma faixa ampla do ELT, a velocidade de produção não parece afetar substancialmente a configuração espectral. A figura 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 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.

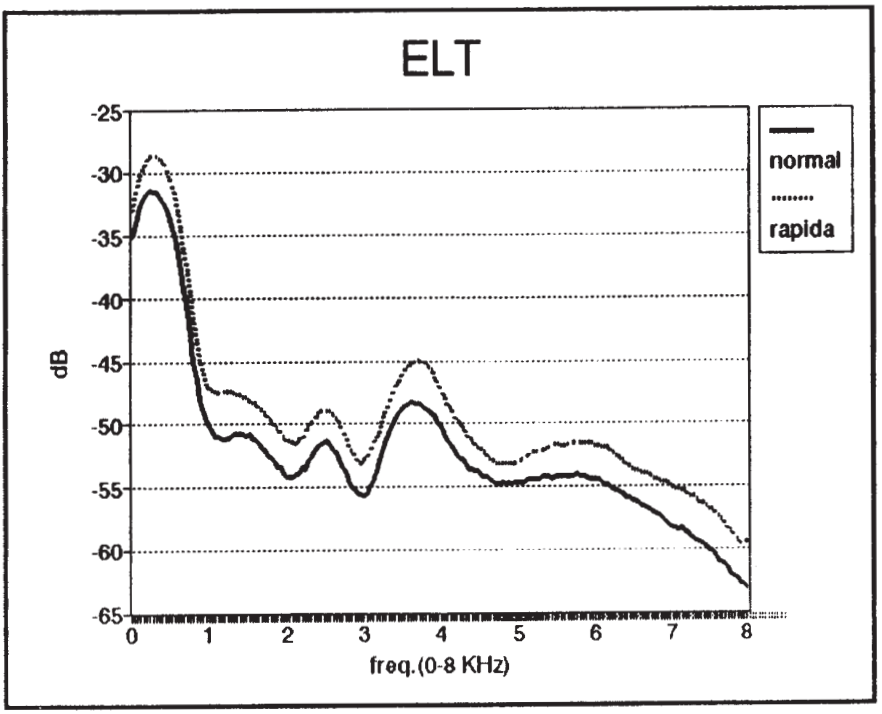

figura 6: ELTs médios, reunindo todos os falantes, nas duas velocidades de emissão (linha contínua=velocidade normal; linha pontilhada=velocidade rápida)

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 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 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 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 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.

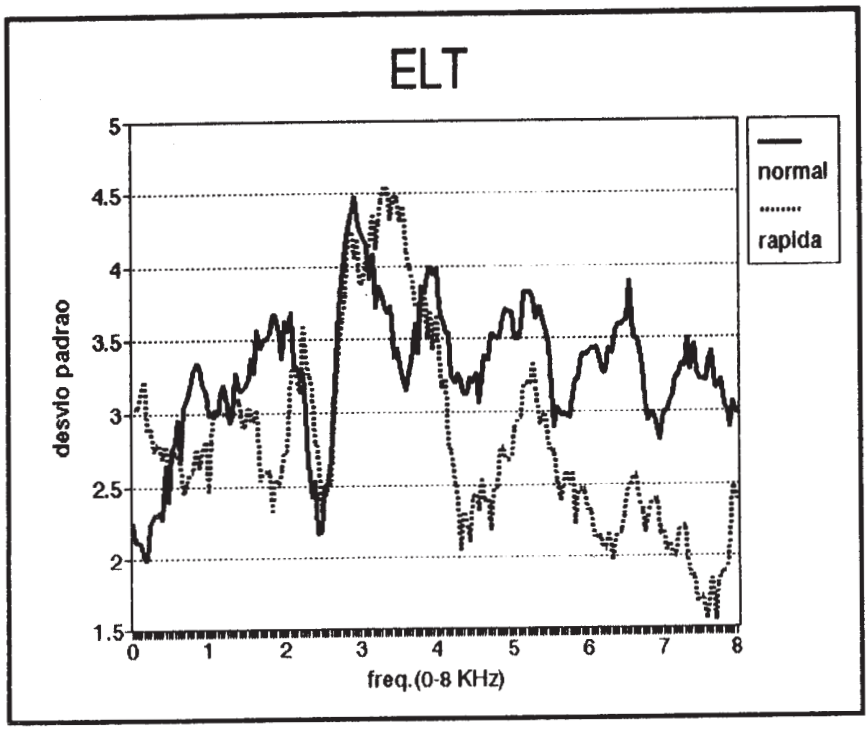

figura 7: desvio padrão inter-falante para cada ponto no eixo de freqüências do ELT (n=200)

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 usam 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 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).

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:

```
(11)    0-------------------2500-------3500
(6)             1000--------2500--------3500
(8)             1000--------------------3500-    5000
(10)    0---------------2000----------3500
(13)    0-------------------------------3500-   -5000
```

Menores Desvios:

```
(3)        500-------1500---2500
(1)     0--500--1000
(9)     0------------1500---2500
(2)     0--500-------1500
(4)     0-------1000-----2000
```

fal.	V.	1	2	3	4	5	6	7	8	9	10	11	12	13
ZR	N	11.6	15.8	5.9	14.5	1.9	1.1	-1.1	-2.9	11.0	9.2	7.3	5.1	1.6
	R	8.8	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	2.7
	R	10.4	13.2	6.3	12.5	.90	-.10	-.80	-1.1	10.6	7.2	5.2	4.5	2.6
RI	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.	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
Pad.	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 4: razões de amplitude entre diversas faixas do ELT; as duas últimas linhas dão as médias e desvios padrão inter-falante

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.

A razão 0-1500/1500-2500 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 5, abaixo, reproduzida de Nolan 1983:153, tabela 4.1). Esse *slope* é próximo ao que

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.wh isp. falsetto	-16	-11	-6	-2

tabela 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)

encontramos para a faixa de 0-2000 Hz, com média de -5.72 dB/oitava (ver tabela 4, acima). A razão de amplitude entre as faixas 0-1500/1500-2500 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 5, acima).

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 frequentemente maior do que a variação entre os *settings* para o mesmo falante (ver tabela 5, acima).

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 5.

Podemos verificar na tabela 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 uma **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].

---

[3] 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.*

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 5 que as medidas dos dois parâmetros (razão de amplitude e *slope*) são consistentes para ambos 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 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 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.

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) 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

---

Essa observação assume um sentido mais incisivo se considerarmos *settings* complexos como *harsh ventricular whispery falsetto* (!).

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[4].

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 adoçao 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 3.5), permite visualizar a magnitude da variação do ELT para um mesmo indivíduo,

---

[4] 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.

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 paradigma. 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 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.

BIBLIOGRAFIA

CHILDERS, D.G. e C.K. Lee 1991 "Vocal quality factors: analysis, synthesis, and perception", *JASA* 90, 2394-410

COMPTON, A. 1963 "Effects of filtering and vocal duration upon the identification of speakers, aurally", *JASA* 35, 1748-52

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

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

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

FLANAGAN, J.L. 1958 "Some properties of the glotal sound source", in D. Fry (ed.) 1976, *Acoustic Phonetics*, Cambridge, 31-51

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

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

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

HURME, P. e A. Sonninen 1986 "Acoustic, perceptual and clinical studies of normal and dysphonic voice", *J.Phon.* 14, 489-92

KITZING, P. 1986 "LTAS criteria pertinent to the measurement of voice quality", *J.Phon.* 14, 477-82

KLATT, D. e L.C. Klatt 1990 "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *JASA* 87, 820-57

LADEFOGED, P. 1984 "Review: The Phonetic Bases of Speaker Recognition by F.J. Nolan", *J.Phon.* 12, 85-9

LAVER, J. e P. Trudgill 1979 "Phonetic and linguistic markers in speech", in Scherer, K.R. e H. Giles (eds) 1979, *Social Markers in Speech* , Cambridge, 1-32

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

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, *Experimental Phonology*, Acad. Press, 13-44

LÖFQVIST, A. 1986 "The long-time-average spectrum as a tool in voice research", *J.Phon.* 14, 471-6

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

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

PITTAM, J. 1987 "The long term spectral measurement of voice quality as a social and personality marker: a review", *L.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

SCHERER, K.R. e H. Giles (eds) 1979 *Social Markers in Speech*, Cambridge

VAZ, N. M. 1983 "Idéias para uma nova imunologia", *Ciência Hoje* II, 7,32-8

WENDLER, J., A. Rauhut e H. Krüger 1986 "Classification of voice qualities", *J.Phon.* 14, 483-8

WOLF, J.J. 1972 "Efficient acoustic parameters for speaker recognition", *JASA* 51, 2044-56

ZALEWSKI, J., W. Majewski e H. Hollien 1975 "Cross correlation of long-term speech spectra as a speaker identification technique", *Acustica* 34, 20-4

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

APÊNDICE

TEXTO I (extraído de VAZ 1983 (pg. 33, item 2)

A reatividade dos linfócitos, as células do sangue que fabricam anticorpos, , 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.

até [1] : trecho lido utilizado para cálculo do ELT na velocidade normal
até [2] : trecho lido utilizado para cálculo do ELT na velocidade rápida

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)