DETERMINAÇÃO DE PONTOS TERMINAIS (*END-POINTS*) BASEADA NO BANCO DE FILTROS DE COEFICIENTES *PLP*

M. A. R. ANDRADE E S. C. B. SANTOS

Departamento de Engenharia Elétrica – Instituto Militar de Engenharia – IME
Pça General Tibúrcio, 80 – Praia Vermelha, Rio de Janeiro RJ – Brasil, CEP.:22290-000
Tel: (21) 546 70 30, Fax: (21) 546 70 39, rocca@cds.eb.mil.br, sidney@aquarius.ime.eb.br

SUMÁRIO

A determinação dos pontos terminais ('end-points') de uma locução é uma das etapas iniciais na atividade de reconhecimento da fala. Existem vários métodos para a determinação destes pontos. É apresentado neste artigo um método utilizando os valores obtidos com o banco de filtros empregado na determinação dos coeficientes PLP. O método proposto apresentou bons resultados em ambientes de gravação ruidosos empregando os mesmos atributos ('features') intermediários usados no processamento do sinal de voz após a determinação dos pontos terminais.

1. INTRODUÇÃO

A diferenciação entre o que é ruído e o que é voz, pelo computador, é uma das tarefas iniciais no reconhecimento automático da fala. Uma diferenciação incorreta pode prejudicar, logo nas primeiras etapas, o processo de reconhecimento. Gravada uma locução, a determinação do ponto onde ocorre a transição entre sinais pertencentes ao ruído de fundo e os sinais relevantes da voz tem por objetivo orientar a pesquisa para o intervalo em que o sistema de reconhecimento deverá ser aplicado. Conhecidos os pontos terminais, pode-se concentrar o esforço computacional no intervalo de interesse e economizar tempo de processamento.

Um método simples para a determinação dos pontos terminais da locução é a observação direta da voz digitalizada em um dos diversos programas disponíveis hoje no mercado. Com o auxílio de gráficos e repetição de locução, o operador pode estimar os pontos que limitam o sinal relevante.

Um método mais usado para tornar automático o processo é o da energia e da taxa de cruzamento por zero [1,2]. Neste processo, o sinal é janelado tipicamente em intervalos de dez milisegundos e as comparações são feitas entre janelas. O início da gravação, um intervalo em torno de 100 milisegundos, é considerado como contendo o ruído de fundo ambiente. Deste início são extraídos a energia e a taxa de cruzamento por zeros. De posse destes valores iniciais são determinados valores que serão usados como limiares de decisão na busca do início e do fim da locução. Após a comparação das características de uma janela do sinal com os limiares e a aceitação desta como um sinal relevante, as janelas vizinhas são reexaminadas para testar a continuidade da locução e impedir que um pico espúrio seja erroneamente admitido como relevante. Há formas de determinação dos limiares de comparação e meios de exclusão de picos espúrios dentro desta linha de determinação de pontos terminais [2]. Os dois métodos citados apresentam alguns inconvenientes. O primeiro é demorado, tedioso e não pode ser aplicado em tempo real. O segundo é relativamente vulnerável às condições de ruído ambiente.

No XVII SBT foi apresentado um método alternativo utilizando algoritmo de agrupamento k-médias modificado [3].

O presente artigo propõe um método para a determinação dos limites de uma locução relevante. Este método possui o mérito de realizar a delimitação utilizando os mesmos atributos intermediários que serão usados no processo de reconhecimento. Isto, com menor custo computacional do que o uso de um algoritmo de agrupamento.

Na seção 2 são apresentadas algumas considerações sobre o ruído ambiente. A seção 3 cita sumariamente a obtenção de coeficientes PLP. A seção 4 apresenta as adaptações feitas no processo de extração de coeficientes PLP. Na seção 5 é apresentado o método para estimar o limiar de decisão. O processo de determinação dos pontos terminais é descrito na seção 6. A seção 7 apresenta o equipamento e o ambiente de gravação de teste. Os resultado obtidos e a comparação entre os três métodos utilizados é apresentada na seção 8. E a seção 9 expõe as conclusões do artigo.

2. CONSIDERAÇÕES SOBRE RUÍDO AMBIENTE

Em ambientes com tratamento acústico, o ruído de fundo tem características estatísticas bem definidas. Estas características facilitam a discriminação dos limites de uma locução sobreposta a este ruído, e os limiares de comparação podem ser facilmente determinados no caso do método da energia e taxa de cruzamento por zero.

Quando o ambiente não possui um tratamento acústico ótimo, as características estatísticas do ruído passam a ser mais complexas, ou de ordens superiores, exigindo um maior cuidado na forma de se encontrar os limiares de decisão sobre o que é voz e o que não é. Neste ponto, o método da energia e taxa de cruzamento por zero começa a tornar-se mais complexo de modo a não perder eficiência.

Em ambientes normais de trabalho, o ruído de fundo já não pode ser considerado como tendo uma distribuição gaussiana. Em vez disso, passa a apresentar características mais específicas, tais como: freqüências fundamentais e seus harmônicos, picos ritmados, picos esporádicos, etc.

Consequentemente, uma forma mais complexa de representação (através do emprego de um número maior atributos) dos sons do ambiente de gravação deve proporcionar uma melhoria na decisão quanto aos limites de uma locução válida pronunciada neste ambiente.

3. ESCOLHA DE ATRIBUTOS

No sistema de reconhecimento, usado após a rotina de determinação dos pontos terminais descrita neste trabalho, utilizou-se os seguintes atributos: Log-energia de curto período e os cinco primeiros coeficientes *PLP-Ceps* (*Perceptual predictive analysis*) [4] com suas derivadas de primeira e segunda ordens. O sinal de voz é agrupado por janelas de Hamming de vinte milisegundos (quadros de dez milisegundos com 50% de superposição com os quadros vizinhos no tempo). O vetor de atributos de cada janela possui então 18 atributos.

Para poupar esforço computacional, optou-se, para a determinação dos pontos terminais, por empregar os mesmos valores usados em uma etapa intermediária (banco de filtros) do processo de extração de coeficientes *PLP* [4]. Sendo assim, estes atributos intermediários são determinados no início do processamento do sinal gravado, e não precisam mais ser alterados ou abandonados posteriormente, como seria o caso da energia e da taxa de cruzamento por zero extraídas em janelas de dez milisegundos.

Com o objetivo de levar em consideração as características do sistema auditivo, o método *PLP* [4] altera o espectro do sinal de voz antes da análise de predição linear. Faz-se uso de um banco de filtros assimétricos de banda não muito estreita espaçados pela escala Bark [1]. Faz-se uso também de pré-ênfase e compressões com o objetivo de simular determinadas características da audição humana. Os passos envolvidos na transformação para cada janela do sinal de voz são citados a seguir:

- **Análise espectral**: realizada com o cálculo da transformada rápida de Fourier e do espectro de potência.

- **Análise de banda crítica**: pondera-se o espectro de freqüência, de acordo com a escala Bark, utilizando a seguinte equação [4]:

$$\Omega(\omega) = 6\ln\left(\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right) \qquad (1)$$

onde ω é a freqüência angular em *rad/s*. Em seguida, faz-se a convolução do espectro ponderado com o espectro de potência de uma banda crítica simulada, onde a forma do filtro é dada pela seguinte equação [4]:

$$\Psi(\Omega) = \begin{cases} 0 & \Omega < -1{,}3 \\ 10^{2{,}5(\Omega+0{,}5)} & -1{,}3 \le \Omega \le -0{,}5 \\ 1 & -0{,}5 \le \Omega \le 0{,}5 \\ 10^{-\Omega+0{,}5} & 0{,}5 \le \Omega \le 2{,}5 \\ 0 & \Omega \ge 2{,}5 \end{cases} \qquad (2)$$

e a convolução por [4]:

$$\Theta(\Omega_i) = \sum_{\Omega=-1{,}3}^{2{,}5} S(\Omega - \Omega_i)\Psi(\Omega) \qquad (3)$$

A função $\Theta(\Omega)$ é amostrada em intervalos de aproximadamente 1 Bark. O valor exato do intervalo de amostragem é escolhido de forma que um número inteiro de amostras espectrais cubra todo o intervalo de análise. Como exemplo, para cobrir uma faixa de 0 a 5 kHz, equivalente a 0 até 16,9 Barks, pode-se utilizar 18 amostras espectrais espaçadas de 0,994 Bark cada uma.

- **Pré-ênfase**: o espectro amostrado $\Theta(\Omega)$ é pré-enfatizado por uma curva que simula as diferenças de sensibilidade do ouvido humano às diversas freqüências. O sinal resultante é representado por [4]

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \qquad (4)$$

onde $E(\omega)$ é dado por [4]

$$E(\omega) = \frac{(\omega^2 + 56{,}8\text{x}10^6)\omega^4}{(\omega^2 + 6{,}3\text{x}10^6)^2(\omega^2 + 0{,}38\text{x}10^9)(\omega^6 + 9{,}58\text{x}10^{26})} \qquad (5)$$

para sons em níveis moderados e em freqüências de 0 até superiores a 5.000 Hz. Esta curva de simulação de sensibilidade não é muito rígida, podendo sofrer ligeiras alterações sem que ocorram mudanças significativas. O fato de ainda não serem conhecidas todas as características da audição humana média deixa em aberto a determinação de $E(\omega)$.

- **Compressão cúbica da amplitude**: Para simular a relação não linear entre a intensidade do som real e a percebida, faz-se uma compressão cúbica conforme a equação [4]

$$\Theta(\Omega) = \Xi(\Omega)^{0{,}33} \qquad (6)$$

- **Modelo só de pólos**: A partir da função $\Theta(\Omega)$ é calculada a transformada inversa de Fourier e obtém-se um modelo só de pólos utilizando o método da autocorrelação.

- **Transformações adicionais**: na utilização do PLP no reconhecimento de voz é comum a realização de transformações adicionais sendo a mais comum a passagem de PLP para PLP-Cepstro e suas primeiras e segundas derivadas [4].

Para o método de extração de pontos terminais proposto, são usados os valores obtidos até a etapa de de pré-ênfase. O restante da análise *PLP* só será efetuado no trecho delimitado pelo método.

O número de coeficientes a serem gerados e utilizados depende da aplicação. Para tarefas de reconhecimento de *locutor* são utilizados mais de 12 coeficientes (modelos de ordens elevadas), e para tarefas de reconhecimento de voz *independente do locutor* são utilizados modelos de ordens inferiores com 5 coeficientes apenas [4].

4. ADAPTAÇÕES NA EXTRAÇÃO DE COEFICIENTES *PLP*

O esforço computacional para o cálculo do PLP é comparável ao gasto para a análise preditiva linear tradicional [4], sendo que as etapas de análise de banda crítica e pré-ênfase podem ser unidas e seus coeficientes calculados a priori. A operação mais custosa é a transformada rápida de Fourier. A passagem para coeficientes auto regressivos se realiza com poucos pontos oriundos da análise de banda crítica.

Para o presente trabalho, optou-se por realizar a FFT apenas com número inteiros (16 *bits*), em vez de números de ponto flutuante (até 80 bits) para acelerar o cálculo computacional. Esta redução de precisão não causou variações muito acentuadas nos espectrogramas gerados a partir dos resultados com os dois níveis de precisão, e evitou a normalização dos valores inteiros obtidos da placa de aquisição de sinal. Os efeitos foram mais atuantes na transformada inversa. Os coeficientes de ordem mais elevada passaram a apresentar valores diferentes dos obtidos com precisão estendida. Porém, como os coeficientes elevados não possuem utilidade para o reconhecimento independente do locutor [4], esta divergência foi considerada como sem relevância para o trabalho.

O programa para cálculo de FFT direta e inversa com números inteiros em linguagem *C*, foi adquirido via Internet através do endereço http://www.jjj.de/fxt/, de autoria de Tom Roberts, sofrendo pequenas correções em função de teste realizados antes da inclusão na rotina de extração de coeficientes PLP, utilizada neste artigo. O autor do método PLP forneceu, posteriormente à publicação do seu artigo, os valores dos bancos de filtros das bandas críticas já ponderados por uma curva de pré-ênfase. Porém, os dados eram para as freqüências de amostragem de 8.000 Hz ou para 16.000 Hz, com uma curva de pré-ênfase diferente da utilizada no artigo[4]. Para compatibilizar o banco de filtros com a freqüência de amostragem de fs=11.025 Hz, de uso comum no Instituto, realizou-se a interpolação da envoltória do banco de filtros correspondente a fs=16.000 Hz para se obter a nova curva de pré-ênfase. O número de filtros também foi modificado de 19 para 15, com espaçamento de 1,08 Bark. Manteve-se o número de 512 pontos para a transformada. O gráfico superior da Figura 1 (apresentada no final do artigo) contém as curvas dos 19 filtros fornecidos para cálculo de coeficientes PLP para uma f_s = 16 kHz. O gráfico inferior contém as curvas dos 15 filtros propostos para a f_s = 11.025 Hz, seguindo as equações do item 2.3.1., porém equalizado por uma curva interpolada do gráfico superior. Em ambos os gráficos na Figura 1, a primeira curva, correspondente a banda de freqüências mais baixas, possui amplitude e comprimento muito reduzidos, ficando imperceptível neste gráfico.

5. DETERMINAÇÃO DE LIMIARES DE DECISÃO

A determinação de limiares de decisão foi orientada pela idéia de que o sinal de voz apresenta uma variação entre atributos de janelas adjacentes maior do que as variações do sinal de ruído ambiente. A simples observação de espectrogramas de elocuções realizadas em ambientes com baixo ruído permite explorar a determinação de pontos terminais sob essa ótica. A Figura 2 apresenta o espectrograma da elocução da frase *bom dia*, ladeada por amostras de ruído ambiente. Observa-se a relativa uniformidade do ruído em comparação com o trecho da elocução relevante (amostras de 3.000 a 7.000 aproximadamente na Figura 2). Assim, distâncias euclidianas entre janelas que contenham atributos de ruído ambiente estariam abaixo de um limiar, indicando pouca variação e sinal irrelevante. Janelas adjacentes distantes entre si acima de um limiar apontam para uma seqüência de transições que provavelmente seria causada pela seqüência de fonemas, indicando um sinal relevante para identificação. Supõe-se, neste ponto, que o ruído ambiente possui a menor variação de atributos entre janelas, ou, de outro modo, é o trecho mais uniforme da gravação. A Figura 3 apresenta a mesma elocução *bom dia* após ter passado pelo banco de filtros do item 4, onde a faixa de freqüências mais altas corresponde ao banco de filtro de maior ordem.

O cálculo do limiar de decisão foi feito com base nas cinqüenta primeiras janelas do intervalo gravado, admitiu-se que neste início havia somente atributos representativos do ruído ambiente, com uma distribuição não necessariamente guassiana. Este intervalo possui duração e função similares ao intervalo inicial do método da energia e taxa de cruzamento por zero. Calculados as médias e os desvios padrão de cada atributo ao longo das cinqüenta janelas iniciais, os valores são consolidados em um único limiar determinado pela distância euclidiana entre um vetor *M+ (M mais*, formado pelas médias mais os desvios), e um vetor *M- (M menos*, formado pelas médias menos os respectivos desvios padrão). Isto equivale ao comprimento de um vetor composto por 2 vezes os desvios padrão. Caso a distância entre duas janelas adjacentes do restante da gravação seja maior que este limiar, estas janelas serão candidatas a representar sinais relevantes de voz.

6. DETERMINAÇÃO DE PONTOS TERMINAIS

A determinação do ponto inicial da locução dentro da gravação é feita varrendo-se as janelas ao longo do tempo, iniciando-se a varredura logo após a última janela usada na caracterização do desvio-padrão do ruído de fundo. O primeiro vetor das janelas a serem testadas que estiver afastado (distância euclidiana) do vetor da janela seguinte de um valor maior que duas vezes o desvio-padrão determinado no item anterior é considerado como forte candidato a representar uma janela de locução válida. Para se ter a confirmação de que a janela corresponde realmente ao início de uma locução válida, são examinadas as nove janelas seguintes; se, dentre elas, mais de quatro também forem válidas

(distantes entre si por mais de duas vezes o desvio-padrão), a primeira janela deste conjunto é aceita como início de locução relevante. Caso contrário, prossegui-se na busca. Este procedimento visa evitar o falso reconhecimento de picos espúrios.

A determinação do ponto final da locução segue o processo inverso ao da determinação do ponto inicial. Procura-se por um conjunto de dez janelas em que pelo menos 4 possuam distâncias inferiores ao limiar de decisão. A última deste grupo é o final da locução.

Para evitar que uma pequena pausa entre sílabas seja erroneamente identificada como final da locução, janelas a frente são examinadas. Caso ainda haja sinal relevante após esse falso silêncio, ele é acrescido a locução e a busca por um final reinicia.

7. GRAVAÇÃO DE LOCUÇÕES PARA TESTE

Para testar o método proposto, foram feitas dez gravações, cada uma contendo um dos dez dígitos.

O ambiente de gravação usado continha como ruído de fundo sons comuns em um ambiente doméstico. As gravações foram feitas em um cômodo situado no terceiro andar, com janela aberta para uma rua de pouco movimento, próximo a um aeroporto regional, e com um aparelho de televisão ligado em volume moderado no cômodo vizinho.

A aquisição do sinal foi realizada por um microfone de eletreto comum, conectado a uma placa 'Sound Blaster 16' em um computador pessoal.

8. RESULTADOS

Para avaliar os três métodos citados anteriormente, cada gravação teve seus pontos iniciais e finais determinados por cada método. A Tabela 1 apresenta os resultados. O método A corresponde à edição manual da gravação. O método B é o que emprega somente os valores de energia e taxa de cruzamento por zero. E o método C é o proposto, utilizando o banco de filtros do método *PLP*.

O método B realiza uma determinação mais rápida que o método C. A demora do terceiro método é devida à maior quantidade de cálculos necessários para montar um vetor de 15 características por janela, maior que as duas usadas pelo segundo método.

9. CONCLUSÃO

Pelo estudo dos resultados da Tabela 1, comparando-se especificamente as colunas A e C, nota-se que o método proposto aproxima-se muito dos valores determinados pela edição manual das gravações, em que o ouvido humano é a principal ferramenta.

Nos pontos limites, onde o fonema adjacente é sibilante [5], o método proposto apresentou um retardo na identificação do início no caso dos dígitos cinco, seis e sete.

	Número de ordem da janela inicial			**Número de ordem da janela final**		
Método / Dígito	**A**	**B**	**C**	**A**	**B**	**C**
Um	124	122	123	164	167	164
Dois	120	67	122	188	207	189
Três	226	1	226	268	472	267
Quatro	212	23	212	184	301	187
Cinco	178	4	175	264	336	262
Seis	188	159	187	268	322	269
Sete	200	142	198	260	276	260
Oito	224	10	224	284	492	284
Nove	198	198	199	262	276	265
Zero	192	46	191	260	438	263

Tabela 1. Comparação dos resultados de cada método

No ambiente ruidoso das gravações, o método B realizou vários erros de estimação, quanto ao ponto inicial e ao ponto final das locuções. O segundo método apresentou resultados próximo aos demais nas gravações dos dígitos um e nove, em que o ruído de fundo foi casualmente mais baixo.

O método proposto, apesar de mais complexo, de exigir maior esforço computacional e maior período inicial de silêncio, mostra-se robusto em ambientes ruidosos. Melhorias no processo podem ser obtidas com um estudo sobre os atributos mais adequados e formas alternativas de determinação dos limiares de decisão.

Uma melhor caracterização do ruído ambiente permite uma determinação de pontos terminais mais eficiente.

O número de atributos usados para a determinação dos pontos terminais tem implicação no esforço computacional realizado. É necessário um estudo das características das locuções que serão processadas para determinar se o aumento de atributos compensará o acréscimo no tempo de processamento. Se o silêncio abranger a maior parte das gravações a serem analisadas, um menor número de atributos é sugerido para que não se perca a maior parte do tempo analisando o que é irrelevante no processo.

A mudança de atributos usados pode modificar em muito o desempenho da tarefa e determinação dos pontos iniciais e finais. Dependendo do sistema de reconhecimento a ser usado, subconjuntos de atributos podem ser empregados para a determinação dos pontos terminais de forma mais eficiente.

10. REFERÊNCIAS

[1] Rabiner L.R. e Juang B.H., "Fundamental of Speech Recognition", Prentice Hall, USA, 1993.

[2] Lamel L., Rabiner L. R., Rosemberg A., & Wilpon J.; *Improved end-point detector for isolated word recognition*, IEEE Trans. On ASSP, Vol. 29 pp. 777-785, 1981.

[3] Andrade M.A.R, Santos S.C.B. e Miscow R. F., "Determinação de Pontos Terminais ('End-Points') Baseada no Método de Médias-k Modificado", XVII Simpósio Brasileiro de Telecomunicações, 1999.

[4] Hermansky H., "Perceptual predictive (PLP) analysis of speech", J. Acoust. Soc. Am. 87 (4), April 1990.

[5] Silveira R.C.P., "Estudo de Fonologia Portuguesa", Cortez Editora, São Paulo, 1986.

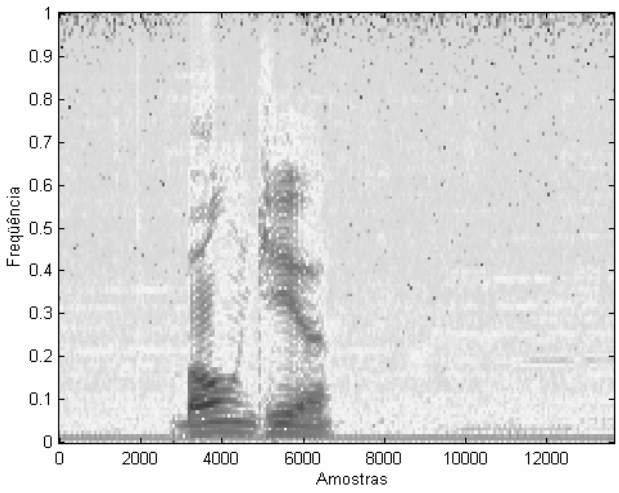

Figura 2. Espectrograma da elocução *bom dia*.

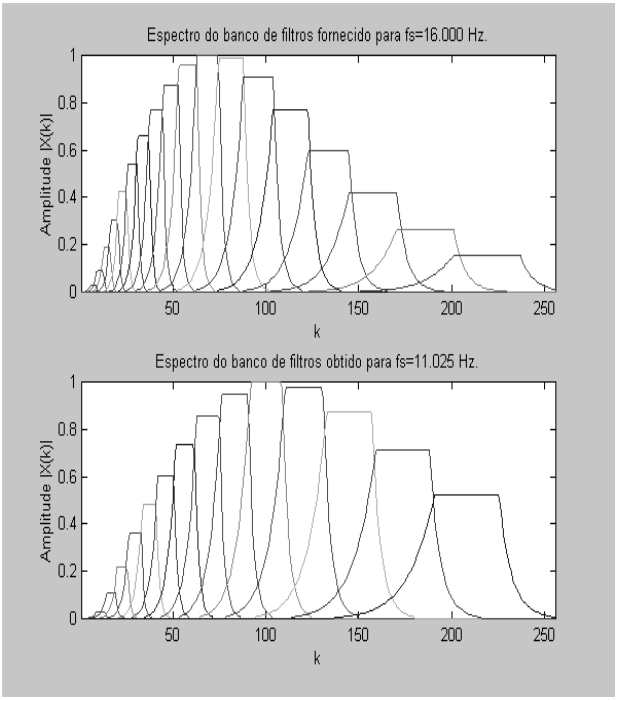

Figura 1. Bancos de filtros.

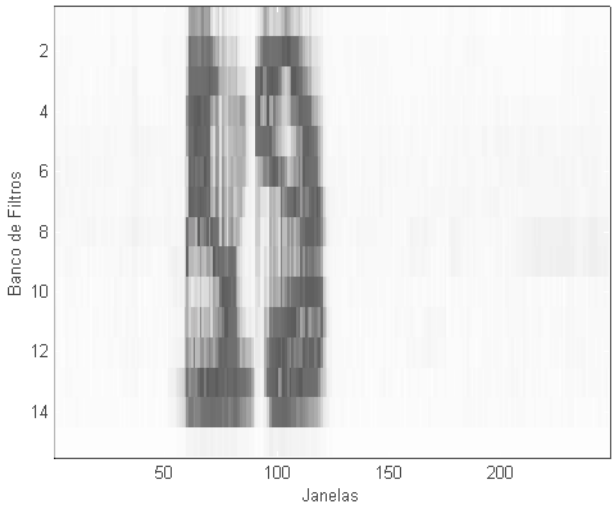

Figura 3. Saída do banco de filtros.