ESTUDO DA DETERMINAÇÃO DE PONTOS TERMINAIS (*END-POINTS*) EM ELOCUÇÕES

*Marco Antônio Rocca de Andrade**
*Sidney Cerqueira Bispo dos Santos**

SUMÁRIO

A determinação dos pontos terminais (*end-points*) de uma elocução é uma das etapas iniciais na atividade de reconhecimento da fala. Existem alguns métodos para a determinação destes pontos. Neste artigo é apresentado um estudo de métodos alternativos baseados no agrupamento de amostras de ruído ambiente pelo processo de médias-k modificado e no desvio padrão dos atributos de amostras de ruído. Os métodos propostos, empregando os mesmos atributos (*features*) usados no processamento do sinal após a determinação dos pontos terminais, apresentaram bons resultados em ambientes de gravação ruidosos.

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.

* Ambos do Departamento de Engenharia Elétrica, Instituto Militar de Engenharia.

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 muito usado para tornar automático o processo é o da energia e da taxa de cruzamentos por zero.[1,2] Neste processo, o sinal é janelado tipicamente em intervalos de dez milissegundos e as comparações são feitas entre janelas. O início da gravação, um intervalo em torno de 100 milissegundos, é considerado como sendo constituído por ruído de fundo ambiente. Desse início são extraídas a energia e a taxa de cruzamento por zeros. A partir 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 admitindo que ela contém 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 várias formas de determinação dos limiares de comparação e meios de exclusão de picos espúrios dentro desta linha de determinação de pontos terminais.[4]

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.

O presente artigo propõe dois outros métodos para a determinação dos limites de uma locução relevante.

CONSIDERAÇÕES SOBRE RUÍDO AMBIENTE

Em ambientes com tratamento acústico, o ruído de fundo tem características estatísticas bem definidas. Essas 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 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 para que não se perca a eficiência.

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

Conseqüentemente, uma forma mais complexa de representação dos sons do ambiente de gravação, isto é, sem a presença de sinais de voz, deve proporcionar uma melhoria na decisão quanto aos limites de uma locução válida pronunciada nesse ambiente.

ESCOLHA DOS ATRIBUTOS (*FEATURES*)

Para poupar esforço computacional, optou-se, na determinação dos pontos terminais, por usar os mesmos atributos usados nos processos

de reconhecimento. Sendo assim, esses atributos 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 milissegundos.[4]

Em um sistema de reconhecimento a ser usado após a rotina de determinação dos pontos terminais, os atributos usados são a energia de curto período e os 12 primeiros coeficientes mel-cepstrais. Nesse sistema, o sinal de voz é agrupado em janelas de Hamming de 20 milissegundos (quadros de dez milissegundos com 50% de superposição com os quadros vizinhos no tempo). O vetor de atributos que caracterizam cada janela possui 39 atributos: os 13 da janela atual *t*, 13 da janela *t*-2, e 13 atributos da janela *t*+2. Os atributos deslocados no tempo destinam-se à composição de derivadas espectrais.[1]

Em um outro sistema de reconhecimento, utilizaram-se valores intermediários do processo de extração de 15 coeficientes de predição linear preceptiva (PLP),[6] com janelas de 20 milissegundos.

Para sistemas de reconhecimento diferentes, os atributos escolhidos podem ser outros.

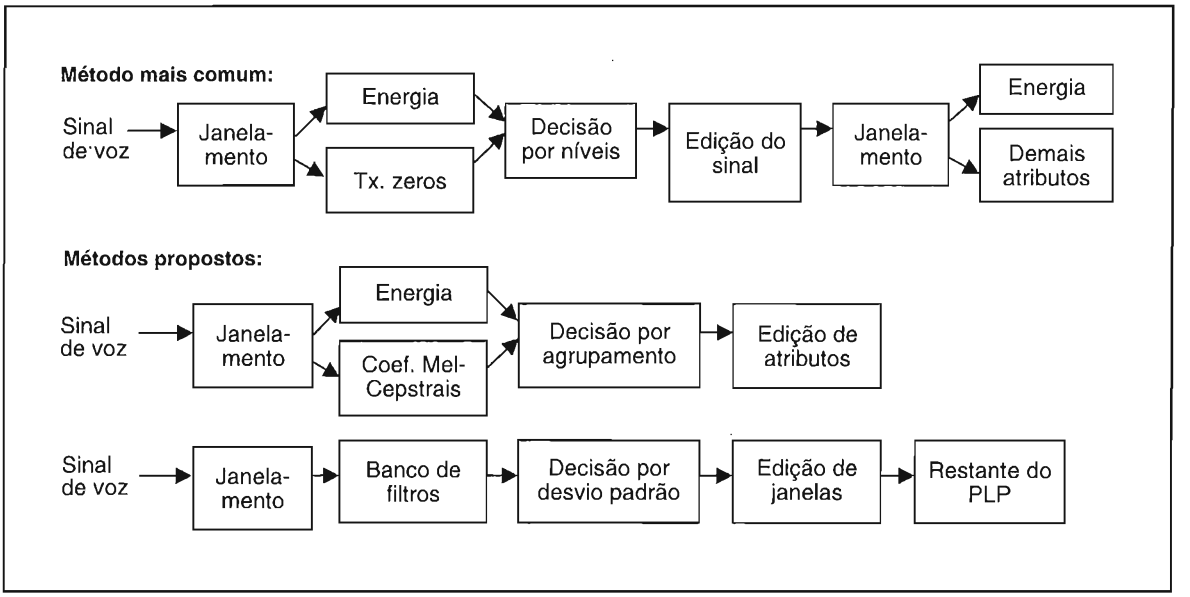

Figura 1: Resumo dos métodos

MÉTODO UTILIZANDO ALGORITMO DE AGRUPAMENTO POR MÉDIAS-K MODIFICADO

O primeiro processo proposto consiste em formar um conjunto de vetores representantes de janelas selecionadas no conjunto das janelas

do início da gravação que contenham amostras do ruído ambiente. Experimentalmente, um período em torno de 90 quadros de dez milissegundos, totalizando 900 milissegundos, mostrou-se adequado para o processo de seleção. As janelas que melhor representarem as diferentes nuances do ruído de fundo são adotadas como padrões para a comparação com as demais janelas da gravação. Esta comparação é feita por meio de distâncias euclidianas entre o vetor de atributos da janela padrão e o vetor da janela a ser testada. Caso a distância seja elevada, a janela testada será rejeitada como ruído de fundo e, então, poderá ser considerada como candidata a ponto limite de uma locução.

Para agrupar as janelas de ruído do início da gravação é usado o algoritmo de *k* médias modificado.[3]

No método proposto, as janelas são divididas em três grupos, e as janelas centróides dos grupos são tomadas como padrões de ruído. Um número diferente de padrões pode ser adotado, porém, em experimentos preliminares, três grupos apresentaram um satisfatório compromisso entre resultado e esforço computacional.

Durante os trabalhos, houve dificuldade em determinar qual seria o melhor limite de distância euclidiana para a qual um dado vetor de janela de teste estaria próximo ou não de um vetor de janela padrão de ruído. A imprecisão desse limite tornaria inviável o processo.

Durante os cálculos do algoritmo *k*-médias modificado, são calculados valores como: número de elementos por grupo, elemento central, distância intragrupo, diâmetro do grupo, entre outros.[3]

Através de experimentos chegou-se à relação:

$$Rn = 1{,}5\ Dmax - DIn / 2 \qquad (1)$$

onde Rn é um valor apropriado para o *raio* em torno do centróide do enésimo grupo padrão de ruído. Dmax é o diâmetro do maior grupo de ruído, e DIn é a distância intragrupo do enésimo grupo de ruído. Qualquer vetor de teste fora destes círculos é candidato a sinal de voz.

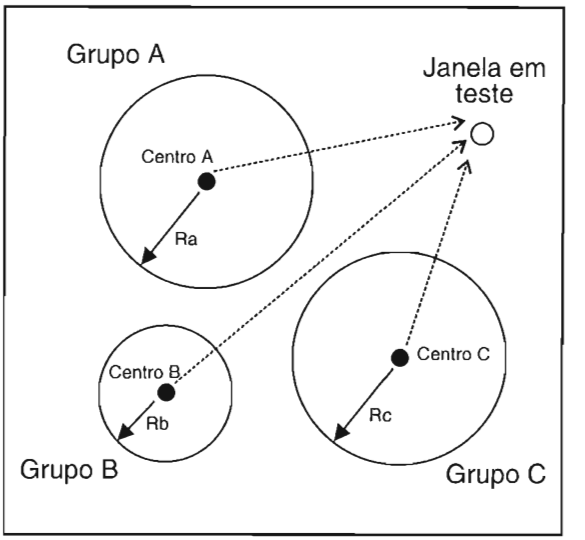

Figura 2: Teste de uma janela

A figura 2 apresenta, no plano, a disposição de três grupos de amostras de silêncio e uma janela a ser testada. Os centros dos grupos, os *raios* de decisão e os vetores que ligam os centros à amostra em teste estão ali representados. Por essa configuração, a janela em teste é candidata a ponto limite de locução, por estar fora dos *círculos de ruído*.

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 ruído de fundo. O primeiro vetor da janela de teste que estiver afastado das janelas padrão de ruído além dos limites determinados 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, a primeira janela deste conjunto é aceita como início de locução relevante. Caso contrário, prossegue-se na busca. Este procedimento visa a 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 quatro sejam de ruído. A última deste grupo é o final da locução.

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

MÉTODO UTILIZANDO DESVIO PADRÃO

O segundo método proposto é regido 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. Assim, distâncias euclidianas entre janelas que contenham atributos de ruído ambiente estariam abaixo de um limiar, indicando pouca variação e sinal irrelevante, e janelas que contenham atributos de sinal de voz possuiriam distâncias acima de um limiar, indicando 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 possua a menor variação de atributos entre janelas, ou, de outro modo, seja o trecho mais uniforme da gravação, e que um fonema muito prolongado não traga informação importante para a tarefa de reconhecimento de voz, podendo assim ser confundido com ruído quando apresentar poucas variações entre janelas vizinhas.

No sistema de reconhecimento usado após a determinação dos pontos terminais por esse método são utilizados os 15 primeiros coeficientes de predição linear perceptiva (PLP),[6] porém para uso na determinação de pontos terminais são utilizados valores intermediários no processo de extração de coeficientes PLP. Esses valores são as amplitudes de 15 bancos de filtros da etapa inicial do processamento PLP. Assim, o sistema teria a vantagem de só prosseguir no processamento das janelas que sejam relevantes para a identificação. Cada janela passa a ser representada por um vetor de 15 atributos.

O cálculo do limiar é feito com base nas 50 primeiras janelas do intervalo gravado, supondo-se que neste início haja somente amostras do ruído ambiente com uma distribuição próxima da gaussiana. São calculados as médias e os desvios padrão de cada atributo (cada faixa do banco de filtros) ao longo destas 50 janelas iniciais. Os valores são consolidados em um único limiar determinado pela distância euclidiana entre um vetor M+, formado pelas médias mais os desvios, e um vetor M-, formado pelas médias menos os respectivos desvios padrão. Caso a distância entre duas janelas do restante da gravação seja maior que este limiar, estas janelas serão candidatas a representar sinais relevantes de voz.

O método adotado para evitar falsos reconhecimentos devido a picos espúrios foi o mesmo mecanismo de verificação citado no método anterior.

GRAVAÇÃO DE LOCUÇÕES PARA TESTE

Para testar os métodos propostos, 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, com taxa de amostragem de 11.025Hz.

CRITÉRIO DE AVALIAÇÃO E RESULTADOS OBTIDOS

Para avaliar os quatro 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. O método C, ora proposto, utiliza agrupamento por k-médias (utilizando energia

	Número de ordem da janela inicial				Número de ordem da janela final			
Método / Dígito	A	B	C	D	A	B	C	D
Um	124	122	124	123	164	167	164	164
Dois	120	67	124	122	188	207	196	189
Três	226	1	226	226	268	472	286	267
Quatro	212	23	210	212	184	301	292	187
Cinco	178	4	192	175	264	336	274	262
Seis	188	159	216	187	268	322	254	269
Sete	200	142	212	198	260	276	282	260
Oito	224	10	224	224	284	492	296	284
Nove	198	198	200	199	262	276	282	265
Zero	192	46	196	191	260	438	274	263

Tabela 1: Comparação dos resultados de cada método

e coeficientes mel-cepstrais). E o método D, também proposto, utiliza desvio padrão de amostras de ruído.

O método B apresenta o resultado em até um oitavo do tempo gasto pelo método C. A demora do terceiro método é devida à maior quantidade de cálculos necessários para montar um vetor de 39 características por janela, bem maior que as duas usadas pelo segundo método. O tempo gasto pelo método D fica entre os gastos pelos métodos B e C.

CONCLUSÃO

Pelo estudo dos resultados da tabela 1, comparando-se especificamente as colunas A, C e D, nota-se que os métodos propostos aproximam-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 C apresentou um retardo na identificação do início no caso dos dígitos cinco, seis e sete. O mesmo problema não foi observado no método D.

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

O método proposto C, 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 bem como sobre o número de grupos e as formas alternativas de agrupamentos e estimação dos limiares de decisão.

O método proposto D é mais simples que o método C, mais rápido em termos computacionais e com a melhor proximidade dos valores encontrados pelo método A. Nenhuma correção seria necessária para os pontos terminais determinados dessa forma.

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

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 identificar 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 calculados previamente para a determinação dos pontos terminais de forma mais eficiente.

REFERÊNCIAS

[1] RABINER, L. R. e JUANG, B. H. *Fundamental of Speech Recognition*, Prentice Hall, USA, 1993.

[2] ANDRADE, M. A. R. e VICENTE, J. V. M. *Reconhecimento de Comandos Isolados à Voz*, Projeto de Fim de Curso, IME, 1997.

[3] WILPON, J. G. e RABINER, L. R. *A Modified K-Means Clustering Algorithm for Use in Isolated Work Recognition,* IEEE Transactions on Acoustics, Speech, and Processing, vol. ASSP-33, nº 3, junho de 1985.

[4] LAMEL, L., RABINER, L. R., ROSEMBERG, A. e WILPON, J. *Improved end-point detector for isolated word recognition,* IEEE Trans. On ASSP, vol. 29, pp. 777-785, 1981.

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

[6] HERMANSKY, H. *Perceptual predictive (PLP) analysis of speech*, J. Acoust. Soc. Am. 87 (4), abril de 1990.