SÍLABAS COMO UNIDADES FONÉTICAS PARA O RECONHECIMENTO AUTOMÁTICO DE VOZ CONTÍNUA EM PORTUGUÊS

Sidney Cerqueira Bispo dos Santos
Dep Eng Elétrica - IME - DE/3
Pça Gen Tibúrcio, 80 - Praia Vermelha
22290 - 000 Rio de Janeiro - RJ
Tel/Fax: (021) 546 7030
E-mail: sidney@aquarius.ime.eb.br

Abraham Alcaim
CETUC - PUC – Rio
Rua Marquês de São Vicente, 225 – Gávea
22453-900 Rio de Janeiro – RJ
Tel (021) 529 92 54, 529 93 84
E-mail: alcaim@cetuc.puc-rio.br

Resumo: Este artigo examina o papel das sílabas como unidades fonéticas (**UF**) em Sistemas de Reconhecimento de Voz Contínua (RVC) para o português. Essas unidades possuem um desempenho muito pobre em reconhecedores baseados em língua inglesa e uma possível razão para isso é que o inglês não possui uma divisão silábica trivial. O português, por outro lado, é uma língua silábica por natureza onde a sílaba é o núcleo com que se formam as palavras. Essas unidades se tornam atraentes pelo seu número reduzido quando comparado ao número necessário de trifones para a mesma tarefa, além do seu grau de consistência, quando comparadas aos fones independentes do contexto. Foram realizados dois testes. Os resultados obtidos, 98,81% no modo dependente do locutor e 95,01% no modo independente do locutor, permitem concluir que as sílabas são UFs bastante atraentes para utilização no RVC quando o número de modelos a serem treinados é pequeno. Entretanto, para dicionários grandes o número de modelos torna o treinamento inviável, quando então, outras unidades passam a ser mais atraentes. Com base nesses resultados e nos trabalhos de reconhecimento de voz para a língua inglesa, conclui-se que as sílabas possuem um desempenho melhor para o idioma português do que para o idioma inglês.

Palavras chave: Reconhecimento de voz contínua; unidades fonéticas.

Abstract: This paper examines the role of syllables as phonetic units (PU) in Portuguese-based continuous speech recognition (CSR) systems. These units have not shown promising results for the English language. This is probably due to the fact that English does not have a trivial syllabic splitting. However, in the Portuguese language these structures are simple and constitute the nucleus on which words are formed. They are attractive due to the extremely reduced number, as compared to triphones, and because of their consistency, as compared to context-independent units. The test results (98.81% in the speaker-dependent mode and 95.01% in the speaker-independent mode) allow us to conclude that syllables are attractive PUs for small-sized vocabularies CSR schemes. However, for large vocabularies the inventories may be excessively large and other units may be more appealing. Based on these results, we conclude that syllables offer a better performance for Portuguese than for the English language.

Keywords: Continuous speech recognition; phonetic units.

Artigo Submetido em 08/09/99
1a. Revisão em 18/02/00; 2a. Revisão em 07/11/00.
Aceito sob recomendação do Ed. Consultor Prof. Dr. Liu Hsu

1 INTRODUÇÃO

Os sistemas para reconhecimento de voz contínua, devido às exigências de armazenamento e quantidade de processamento, utilizam normalmente como padrões, unidades fonéticas (UF) menores que a palavra. Essas unidades são utilizadas em conjunto com o dicionário de pronúncias para formar o vocabulário do sistema e são as principais responsáveis pelo baixo desempenho em reconhecedores com grandes vocabulários devido à enorme dificuldade em se encontrar um conjunto de unidades que seja treinável e consistente (Young, 1996). Ao longo dos anos, várias unidades ou conjunto de unidades têm sido propostos e utilizados por diversos pesquisadores. Cada um desses conjuntos possuem vantagens e desvantagens e são capazes de representar qualquer palavra. Entretanto, existem sempre problemas de sensibilidade ao contexto ou ao treinamento (Kai-Fu Lee, 1990).

As unidades mais utilizadas e que apresentam o melhor desempenho são os *trifones* - que são de difícil treinabilidade (Young, 1996; Kai-Fu Lee, 1990; Chin-Hui, Rabiner e Pieracini). A carga computacional envolvida é tão grande que só recentemente, com o avanço na tecnologia de armazenamento de dados e com o aumento na velocidade dos processadores foi que os reconhecedores modernos passaram a utilizar HMMs (*Hidden Markov Models*) contínuos com trifones. No treinamento desses modelos são necessárias as estimativas das médias, covariâncias e coeficientes de cada componente da mistura de Gaussianas de cada estado. Para um sistema na língua inglesa que utilize um grande vocabulário, usualmente é necessário o treinamento de aproximadamente

60.000 modelos de trifones. Na língua portuguesa, onde existem aproximadamente 50 fones, existiriam 50^3 = 125.000 possibilidades (embora nem todos os trifones ocorram devido a restrições fonéticas da linguagem).

Como na prática, a utilização de 5 componentes nas misturas produz um bom desempenho (Picone, 1990), e admitindo que as matrizes de covariância sejam diagonais, um reconhecedor com vetores acústicos compostos de 26 atributos necessitará estimar aproximadamente 265 parâmetros por estado, ou seja, 26x5 médias mais 26x5 variâncias mais 5 coeficientes da mistura. Utilizando-se HMMs com 3 estados para modelar cada um dos trifones teremos 99.375.000 parâmetros (3x265x125.000) para serem estimados e armazenados. Vale ressaltar entretanto, que menos da metade das 125.000 possibilidades são realizáveis na língua portuguesa, totalizando aproximadamente 48 milhões de parâmetros. É óbvio que a utilização de um número maior de atributos ou de componentes por mistura na Função de Probabilidade de Saída (FPS) ocasionará um aumento significativo nesse número de parâmetros.

Esse número excessivo e a quantidade necessária de dados para o treinamento, que nunca será suficiente o bastante para permitir uma boa estimativa de todos os contextos, são cruciais no projeto de um reconhecedor de voz. A escolha correta da *Unidade Fonética* influenciará não só o número de parâmetros a serem treinados como também a precisão da modelagem acústica.

Baseado nas características da língua portuguesa, foram descritos em (Santos e Alcaim, 2000) dois inventários reduzidos de unidades fonéticas. Em particular, o Inventário 1, com 149 unidades, é composto de 7 vogais orais, 5 vogais nasais, 4 consoantes silábicas finais e 133 (19x7) combinações CV. Esse Inventário considera que sílabas C_1C_2V (p. ex. BRA) podem ser formadas pela concatenação de unidades C_1V (Ba) e C_2V (RA). Desse modo, é possível formar padrões silábicos de forma apropriada pela concatenação de unidades fonéticas (sub-palavras). Um outro exemplo é a sílaba GRÃOS que pode ser formada pelas unidades: /GA/ + /RA/ + /Ã/ +/U/ + /S/.

As unidades fonéticas apresentadas nos inventários 1 e 2 descrito em (Santos e Alcaim, 2000) são atraentes para utilização no RVC devido ao número extremamente reduzido de modelos, quando comparado ao número de trifones, e ao grau de consistência, quando comparado aos fones independentes do contexto. Entretanto, seu treinamento requer um esforço computacional maior do que o usado para trifones.

Em sistemas com dicionários grandes, que exijam unidades fonéticas com alto grau de treinabilidade e consistência, as unidades dos inventários 1 e 2 são uma excelente escolha. Contudo, para dicionários pequenos e médios, as sílabas podem ser mais atraentes como unidades básicas de reconhecimento.

O primeiro artigo sugerindo sílabas como unidades para o RVC foi publicado em 1975 (Fujimura, 1975) mas não apresentou nenhum teste ou resultado prático. Nos anos seguintes, foram realizados exprimentos com sílabas e semi-sílabas (Hunt, Lemmig e Mormelstein, 1980; Ruske, 1982; Rosemberg *et alii*, 1983), mas os resultados não foram promissores e essas unidades foram praticamente abandonadascomo unidades fonéticas aplicáveis ao RVC. Porém, é importante observar que esses experimentos foram realizados utilizando a língua inglesa como base, que não possui uma divisão silábica trivial. A divisão de palavras dessa língua em sílabas é algo muito complexo e que as pessoas não estão familiarizadas. Tanto, que é prática dos cidadãos de língua inglesa, quando querem tornar clara uma palavra que não foi bem entendida, fazê-lo através dos fonemas que a compõem. Já na língua portuguesa a separação de uma palavra em sílabas é uma conseqüência natural. As pessoas que a falam, quando querem explicitar o que disseram, falam devagar e silabicamente.

No português, a sílaba é o núcleo com que se formam as palavras, é o segmento de voz que dá ao nosso ouvido a impressão de unidade de som. Toda sílaba possui uma vogal, com alta energia, cercada por consoantes ou semivogais que possuem energia mais baixa (Fig. 1). Podemos até dizer que num vocábulo haverá tantas sílabas quanto forem os acentos silábicos, ou seja, as vogais (Nicola e Infante, 1991). É uma UF estável, isto é, leva em consideração a maior parte da coarticulação entre fonemas, esperando-se com isso que proporcione, para o português, um desempenho melhor do que as outras unidades fonéticas menores que a palavra. Além disso, as sílabas possuem uma grande capacidade de generalização, ou seja, a partir delas pode-se gerar qualquer palavra.

Dentro desse contexto, o objetivo do trabalho apresentado neste artigo é examinar o emprego das sílabas como unidades

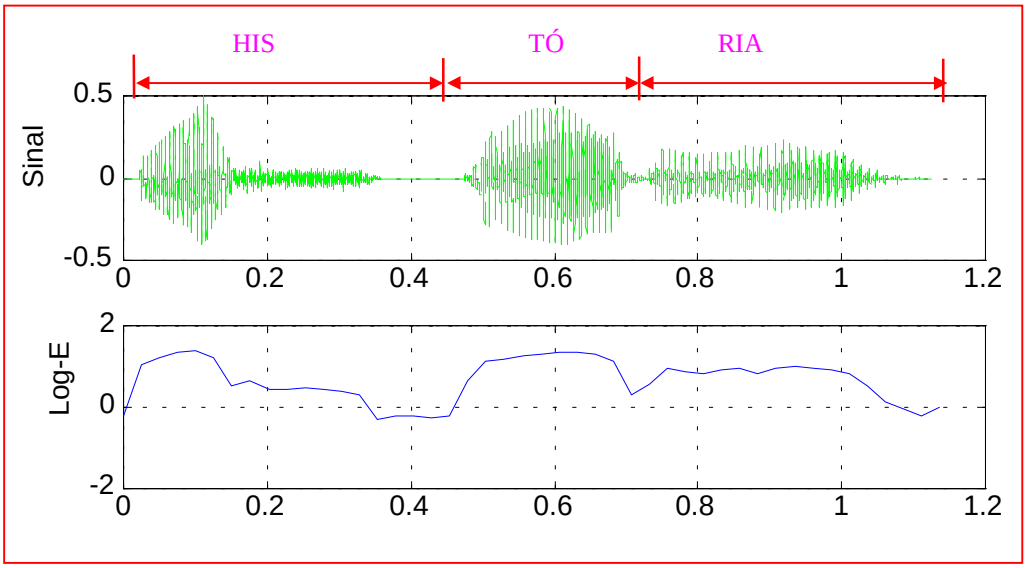

Figura 1 - Sinal acústico e log-energia da palavra *HISTÓRIA*

fonéticas e mostrar que o seu desempenho em reconhecedores de voz contínua, com dicionários pequenos e médios, é compensador. À medida em que o número de palavras no dicionário cresce sua atratividade diminui tendo em vista o grande número de modelos que se tornam necessários treinar.

A seção 2 apresenta o sistema utilizado no treinamento do Reconhecedor Automático de voz contínua, a seção 3 a estrutura do sistema, a seção 4 a base de dados e a seção 5 os resultados dos testes e conclusões.

2 TREINAMENTO DAS SÍLABAS

O treinamento dos modelos das sílabas é mais simples do que o das unidades dos inventários 1 e 2, descritos em (Santos e Alcaim, 2000), porque alguns passos do algoritmo proposto em (Santos e Alcaim, 2000) são desnecessários. Na realidade, o algoritmo para treinamento dos modelos das sílabas é o mesmo utilizado para treinamento de fones, difones e trifones. As sentenças são segmentadas diretamente concatenando-se os modelos das sílabas que as formam, utilizando-se o algoritmo de Viterbi. Para essas unidades não existe a necessidade de gravações das mesmas individualmente, nem para o treinamento inicial do modelo nem para a fase intermediária de treinamento com as palavras do dicionário, devido à sua baixa variabilidade acústica. Isto faz com que todo o processo fique bem mais simples e rápido ao se utilizar dicionários pequenos e médios. No final do treinamento, são obtidas unidades assemelhadas às sílabas. Entretanto, como a utilização será no contexto em que os modelos foram treinados, isso não representa problema.

Definido o contexto em que vai ser empregado, após a definição da estrutura do HMM, ou seja, do número de estados e do número M de componentes da mistura, esse treinamento se traduz na estimação da matriz de probabilidades de transição A e dos parâmetros c_{jm}, μ_{jm} e U_{jm} da Função de Probabilidade de Saída (FPS) de cada estado, para cada unidade, dada por

$$b_j(O_t) = \sum_{m=1}^{M} c_{jm} N(O_t; \mu_{jm}, U_{jm}) \qquad (1)$$

onde O_t é o vetor de observações correspondente ao instante t, j é o j-ésimo estado, m é a m-ésima Gaussiana da mistura e as notações c, μ e U representam, respectivamente, os coeficientes da combinação linear, os vetores média e as matrizes de covariância. O algoritmo para treinamento das sílabas consta do seguinte:

PASSO 1

Represente cada sentença a ser utilizada no treinamento por um HMM composto. Este HMM deve ser formado pela combinação dos HMMs que representam as unidades fonéticas que compõem cada sentença. A Fig. 2 apresenta um exemplo para a frase *Isto é um teste* utilizando sílabas como UF.

PASSO 2

Divida uniformemente os vetores de atributos de cada sentença de treinamento pelos estados do HMM composto assumindo, inicialmente, que só existe silêncio no início e no fim de cada sentença, uma pronúncia de cada palavra e um único modelo para cada UF. Desse modo, inicialmente, todas UFs serão consideradas com a mesma duração.

PASSO 3

Utilize o algoritmo Médias-K Modificado (Wilpon e Rabiner, 1985) para dividir em M agrupamentos os vetores pertencentes a cada estado de cada modelo de cada unidade. Este passo deve ser realizado para todos os estados de todos os modelos.

PASSO 4

Calcule as médias μ_{jm}, as covariâncias U_{jm} e os coeficientes da mistura c_{jm}, para cada agrupamento m de cada estado j. Calcule as probabilidades a_{ij} da matriz A. Este passo é efetuado para todos estados de todas UFs. Os cálculos devem ser efetuados utilizando-se os procedimentos dos PASSOS 3 e 4 do algoritmo *Médias-K Segmentado* (Rabiner et alii, 1985)-(Juang e Rabiner, 1990), descrito a seguir (Fig. 3), para Unidades Fonéticas, que no nosso caso são as sílabas.

PASSO 5

Calcule as verossimilhanças, $P(O|\lambda)$, para um conjunto de sentenças de teste.

PASSO 6

Segmente, novamente, as sentenças de treinamento em UFs utilizando o novo conjunto de parâmetros, obtidos no Passo 4, e o algoritmo de Viterbi. Neste ponto pode-se utilizar vários modelos para a mesma palavra.

PASSO 7

Repita os passos 3 a 6 até que a soma das verossimilhanças das sentenças de teste parem de aumentar.

O algoritmo Médias-K Segmentado, mostrado na Fig. 3, pode ser implementado através dos seguintes passos:

PASSO 1

Execute um dos seguintes procedimentos:

- utilize qualquer modelo já existente da Unidade Fonética (UF);
- divida uniformemente os vetores de atributos da UF pelo número de estados, aplique o algoritmo Médias-K Modificado (Wilpon e Rabiner, 1985), estime a_{ij} e calcule μ_{jm}, U_{jm} e c_{jm} para cada um dos M agrupamentos de cada estado de acordo com os procedimentos do PASSO 3;

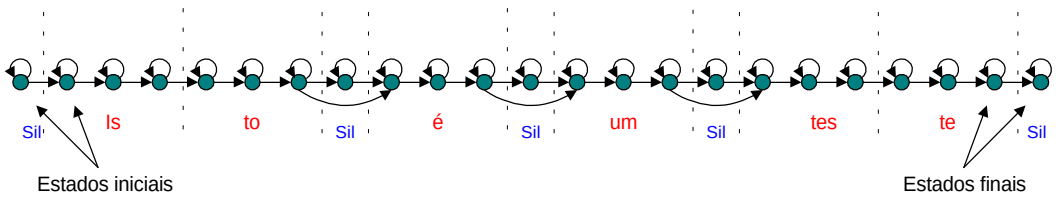

Figura 2 - HMM que representa a frase *Isto é um teste* composta pela combinação dos HMM das unidades fonéticas

- inicialize os parâmetros a_{ij} e $b_j(O_t)$ com valores extraídos de uma seqüência de números aleatórios uniformemente distribuídos entre 0 e 1.

PASSO 2

Utilize o algoritmo de Viterbi para segmentar todas repetições das UFs pelos estados.

PASSO 3

Aplique o algoritmo Médias-K Modificado e divida os vetores de cada estado em M agrupamentos.

Para cada estado, estime novos parâmetros para o modelo estimado λ_{est} utilizando os seguintes procedimentos:

a_{ij} = número de transições do estado i para o estado j dividido pelo número de transições do estado i para qualquer estado;

c_{jm} = número de vetores classificados no agrupamento m do estado j dividido pelo número de vetores no estado j;

μ_{jm} = média amostral dos vetores classificados no agrupamento m do estado j;

U_{jm} = matriz de covariância amostral dos vetores classificados no agrupamento m do estado j.

PASSO 4

Definindo-se a variável *Forward*

$$\alpha_t(j) = P(O_1, O_2, \ldots, O_t, x(t) = S_j / \lambda) \qquad (2)$$

como a probabilidade de ocorrer a seqüência de observações parcial O_1, O_2, ... , O_t e o estado em t ser S_j, dado o modelo λ, e a variável *Backward*

$$\beta_t(j) = P(O_{t+1}, O_{t+2}, \ldots, O_T / x(t) = S_j, \lambda) \qquad (3)$$

realize a estimativa formal dos parâmetros do modelo utilizando as Eq. 4 a 7 para encontrar um novo conjunto de parâmetros λ_{reest} .

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1} \hat{\alpha}_t^{(k)}(i)\, a_{ij}\, b_j(O_{t+1}^{(k)})\, \hat{\beta}_{t+1}^{(k)}(j)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\sum_{j=1}^{N} \hat{\alpha}_t^{(k)}(i)\, a_{ij}\, b_j(O_{t+1}^{(k)})\, \hat{\beta}_{t+1}^{(k)}(j)} \qquad (4)$$

$$\bar{c}_{jm} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{i=1}^{N} \hat{\alpha}_t^{(k)}(j)\, a_{ji}\, b_i(O_{t+1}^{(k)})\, \hat{\beta}_{t+1}^{(k)}(i)\, G_{jm}^{(k)}(t)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{i=1}^{N} \hat{\alpha}_t^{(k)}(j)\, a_{ji}\, b_i(O_{t+1}^{(k)})\, \hat{\beta}_{t+1}^{(k)}(i)} \qquad (5)$$

$$\bar{\mu}_{jm} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{i=1}^{N} \hat{\alpha}_t^{(k)}(j)\, a_{ji}\, b_i(O_{t+1}^{(k)})\, \hat{\beta}_{t+1}^{(k)}(i)\, G_{jm}^{(k)}(t)\, O_t^{(k)}}{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{i=1}^{N} \hat{\alpha}_t^{(k)}(j)\, a_{ji}\, b_i(O_{t+1}^{(k)})\, \hat{\beta}_{t+1}^{(k)}(i)\, G_{jm}^{(k)}(t)} \qquad (6)$$

$$\bar{U}_{jm} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{i=1}^{N} \hat{\alpha}_t^{(k)}(j)\, a_{ji}\, b_i(O_{t+1}^{(k)})\, \hat{\beta}_{t+1}^{(k)}(i)\, G_{jm}^{(k)}(t)\left(O_t^{(k)} - \mu_{jm}\right)\left(O_t^{(k)} - \mu_{jm}\right)^T}{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{i=1}^{N} \hat{\alpha}_t^{(k)}(j)\, a_{ji}\, b_i(O_{t+1}^{(k)})\, \hat{\beta}_{t+1}^{(k)}(i)\, G_{jm}^{(k)}(t)} \qquad (7)$$

onde

$$G_{jm}^{(k)}(t) = \frac{c_{jm}\, N(O_t^{(k)}, \mu_{jm}, U_{jm})}{\sum_{s=1}^{M} c_{js}\, N(O_t^{(k)}, \mu_{js}, U_{js})} \qquad (8)$$

k é a k-ésima elocução da unidade fonética utilizada no treinamento e T_k é o a duração da k-ésima elocução.

PASSO 5

Compare os parâmetros reestimados λ_{reest} com os parâmetros antigos do modelo λ através de uma distância que reflita a similaridade estatística entre os HMMs. Aqui, foi utilizada a

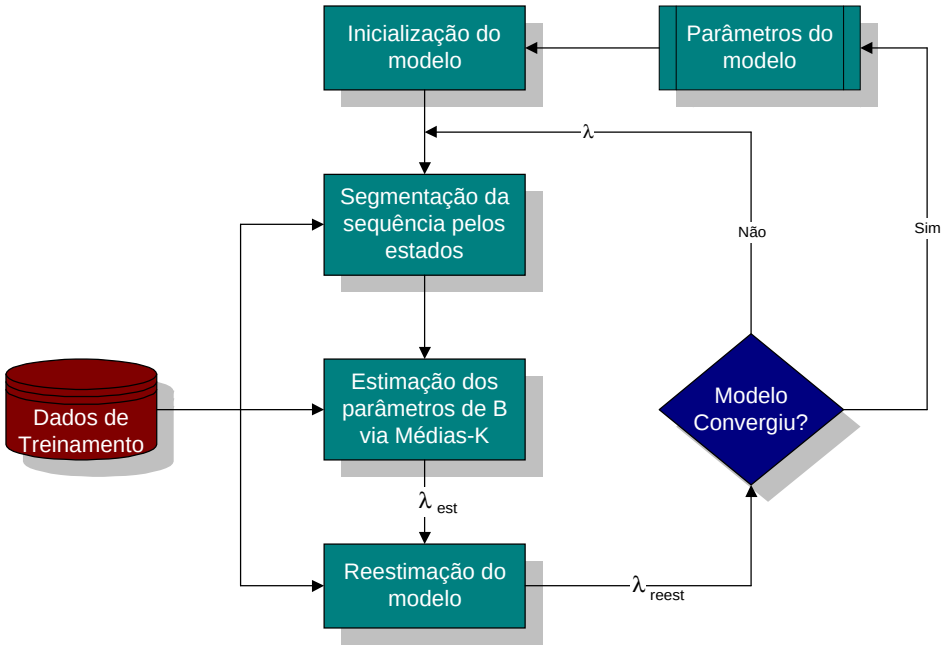

Figura 3 - Algoritmo *Médias-K* Segmentado para unidades fonéticas isoladas

Tabela 1 - Unidades fonéticas do inventário de sílabas

INVENTÁRIO DE SÍLABAS									
1) a	2) ka	3) da	4) fa	5) ga	6) ja	7) ma	8) pa	9) qua	10) ra
11) sa	12) ta	13) é	14) dé	15) lé	16) mé	17) né	18) pré	19) qué	20) rré
21) sé	22) té	23) tré	24) zé	25) ê	26) dê	27) guê	28) lê	29) mê	30) nê
31) prê	32) quê	33) sê	34) tê	35) trê	36) rrê	37) vê	38) zê	39) i	40) di
41) gui	42) li	43) mi	44) ni	45) oi	46) qui	47) ri	48) si	49) ti	50) vi
51) zi	52) kó	53) nó	54) ô	55) dô	56) kô	57) jô	58) rô	59) tô	60) trô
61) zô	62) u	63) du	64) ku	65) ju	66) pru	67) ru	68) tu	69) tru	70) zu
71) eu	72) chã	73) en	74) gem	75) nen	76) nhen	77)quen	78) ren	79) sen	80) ten
81) ven	82) zen	83) quin	84) sin	85) trin	86) vin	87) on	88) fon	89) kon	90) um
91)num	92) brar	93) kar	94) dar	95) gar	96) lar	97) mar	98) nar	99) zar	100) is
101) dis	102) zer	103) faz	104)guós	105) ção	106) três	107) dez	108) tor	109) zôi	110) tus
111)tôs	112) mil	113)guôs							

medida

$$D = \log P(O,X|\lambda_{reest}) - \log P(O,X|\lambda) \quad (9)$$

onde X é a sequência ótima de estados determinada pelo algoritmo de Viterbi. Se a distância exceder um determinado limiar (aqui foi utilizado 0,01 log P(O,X|λ)), substitua os parâmetros antigos λ pelos novos λ_{reest} e retorne ao PASSO 2, senão, encerre o processo.

A aplicação do algoritmo anterior pressupõe a disponibilidade de K repetições de cada UF para que o treinamento seja executado. Além disso, para que os modelos possam ser utilizados no reconhecimento de sentenças é necessário que sejam representativos dos contextos em que as palavras poderão ser encontradas. Por esse motivo devem ser extraídos de sentenças típicas que podem ocorrer no sistema.

A grande vantagem da utilização das sílabas como unidade fonética está na simplicidade e rapidez do treinamento. Para dicionários pequenos, o número de modelos a serem treinados poderá ser menor do que o das unidades do inventário 1 descrito em (Santos e Alcaim, 2000). Entretanto, à medida que o número de palavras do dicionário aumenta, o número de modelos a serem treinados quando se utiliza a sílaba como unidade fonética pode superar (e muito) o número de modelos do inventário 1.

Foram utilizados nos testes efetuados 113 modelos das sílabas, apresentadas na Tabela 1. Este foi o número de modelos necessários para compor o dicionário de pronúncias que podem formar as 83 palavras de um Sistema de RVC para Ligações Telefônicas Automáticas em Língua Portuguesa (Santos, 1997). Esse é o sistema que foi utilizado para avaliação das sílabas como unidade fonéticas. Nesse sistema, em princípio, ao se ouvir um tom de discar ou uma voz sintetizada informando para pronunciar o número do telefone, o usuário o faria de forma natural – sem se preocupar em falar os números na forma de dígitos ou de cardinais (p. ex. zero oitocentos vinte trinta ou zero oito zero zero dois zero três zero ou zero oitocentos dois zero trinta etc).

A Fig. 4 apresenta a visualização da segmentação, utilizando-se o algoritmo de Viterbi, da sentença *Eu gostaria de dar um telefonema pra nove cinco cinco, três quatro, dezesseis* após o treinamento.

3 ESTRUTURA DO SISTEMA

A Fig. 5 mostra os modelos utilizados no sistema. As palavras e as sílabas foram representadas por Modelos de Markov Escondidos Contínuos (**HMMC**). Cada unidade foi modelada utilizando-se três estados e modelos Bakis. As palavras foram representadas como uma combinação de unidades.

O Sistema foi implementado utilizando-se gramática bigrama, gramática regular e conhecimentos dependentes da tarefa em um HMM contínuo. O reconhecimento foi efetuado utilizando-se o algoritmo *One-Pass Modificado* (Santos, 1997).

4 BASE DE DADOS

A base de dados utilizada neste trabalho foi obtida da seguinte

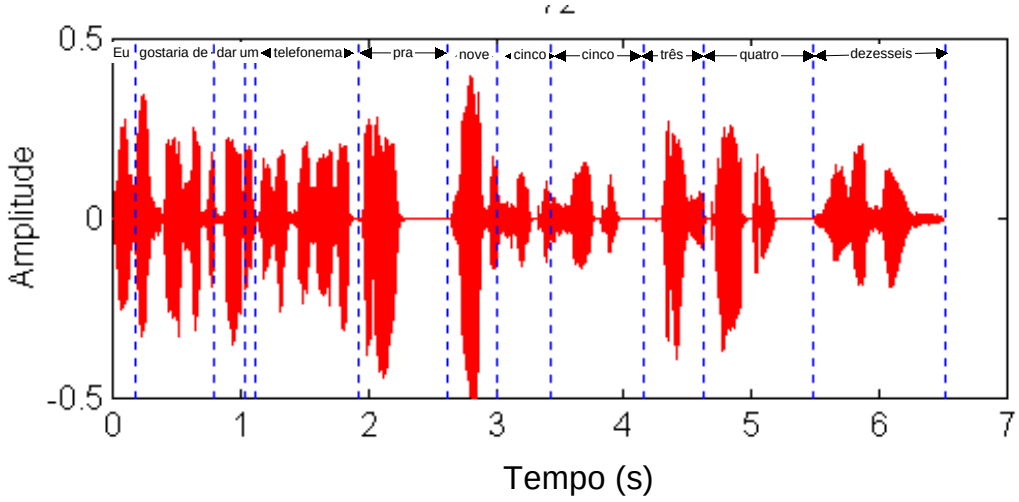

Figura 4 -Segmentação da frase *Eu gostaria de dar um telefonema pra nove cinco cinco, três quatro, dezesseis*

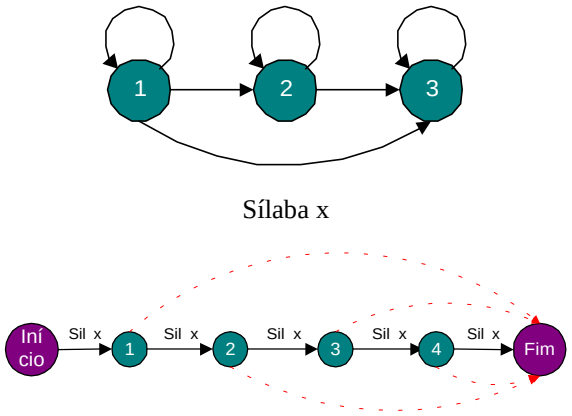

Figura 5 - Estrutura dos modelos

maneira: para o treinamento dependente do locutor, um mesmo locutor pronunciou 20 repetições de cada palavra do dicionário e 154 locuções geradas pelo algoritmo descrito em (Brown, McGee e Rabiner, 1991). Cada gravação foi digitalizada a uma taxa de 11025 Hz, pré-enfatizada por um filtro de primeira ordem e segmentada em janelas de 20 ms. Após o janelamento de Hamming, um vetor de atributos composto de 12 coeficientes Mel-Cepestros, 12 ΔMel-Cepestros, 12 Δ^2Mel-Cepestros, log E, Δlog E e Δ^2log E, totalizando 39 atributos, foi gerado, para cada gravação, com superposição de 50% entre janelas.

Para a determinação dos coeficientes Mel-Cepestros deve-se inicialmente calcular o espectro suavizado do sinal de voz digitalizado, através de coeficientes LPC. Em seguida, este espectro é filtrado através da sua multiplicação por uma série de 20 filtros triangulares espaçados segundo a escala MEL - escala projetada para simular a resposta de freqüência do ouvido humano. Essa escala é linear até 1000 Hz e logarítmica acima dessa freqüência. A energia resultante da filtragem é aplicada a uma função logarítmica para tornar as estatísticas do espectro de potência estimado aproximadamente Gaussianas e, finalmente, é utilizada a Transformada Cosseno Discreta (DCT). Essa transformação possui a propriedade de comprimir a informação espectral nos coeficientes de baixa ordem e também produz uma descorrelação adicional permitindo assim, a utilização com menor perda de precisão, de matrizes de covariância diagonais para o modelamento estatístico da voz (Young. , 1996).

Os parâmetros Δ e Δ^2 são determinados com o objetivo de capturar as variações dinâmicas do espectro dos sinais de voz. O parâmetro ΔMel-Cepestro é calculado de acordo com a seguinte equação:

$$\Delta c_k(n) = c_k(n+N_d) - c_k(n-N_d) \qquad (10)$$

onde $c_k(n)$ é o k-ésimo componente da n-ésima janela e N_d é a distância para a qual se quer calcular a diferença. O valor utilizado para N_d foi 2. Os coeficientes Δ^2 são calculados reaplicando-se os valores de $\Delta c_k(\bullet)$ em substituição a $c_k(\bullet)$ na equação (10).

Para o teste no modo dependente do locutor utilizaram-se 30 sentenças geradas pelo mesmo algoritmo que gerou as sentenças de treinamento mas com bigrama inicial diferente e 11 sentenças escolhidas para levar em consideração contextos ambíguos, totalizando 41 sentenças pronunciadas pelo mesmo locutor com média de 12,3 palavras por sentença.

Para o treinamento independente do locutor, foi utilizado o mesmo processador de sinais anterior, entretanto, os atributos utilizados foram 5 coeficientes PLP-Cepestros, 5 ΔPLP-Cepestros, 5 Δ^2PLP-Cepestros, log E, Δlog E e Δ^2log E, totalizando 18 atributos. Utilizaram-se os coeficientes PLP neste teste devido aos resultados obtidos por Hermansky em (Hermansky, 1990) que mostram que os sistemas PLP de baixa ordem apresentam resultados melhores para o modo independente do locutor.

Os coeficientes PLP são determinados com o objetivo de levar em consideração as características do sistema auditivo humano (Hermansky, 1990). O espectro do sinal de voz é modificado de acordo com características acústicas antes da aproximação pelo modelo autoregressivo. A idéia é semelhante à utilizada no cálculo dos coeficientes Mel-Cepestros, entretanto são utilizados filtros assimétricos e com banda maior que a dos filtros triangulares para simular as bandas críticas e a escala Bark para espaçamento desses filtros. Além disso, o cálculo dos coeficientes PLP incorpora pré-ênfases e compressões com o objetivo de simular determinadas áreas do ouvido humano. Os passos envolvidos nessas modificações são detalhados em (Hermansky, 1990).

O treinamento no modo independente do locutor foi realizado utilizando-se 154 frases pronunciadas por 15 locutores masculinos e 5 femininos, totalizando 3080 frases. O reconhecimento foi realizado utilizando-se as 41 frases pronunciadas por 5 locutores masculinos e 5 femininos, diferentes dos utilizados no treinamento, totalizando 410 frases.

5 RESULTADOS DOS TESTES E CONCLUSÕES

A Tabela 2 mostra os resultados obtidos. Verifica-se que no modo dependente do locutor as sílabas apresentaram uma redução na taxa de erro de 15% em relação às unidades constantes do inventário 1 (Santos e Alcaim, 2000). Entretanto, no modo independente do locutor a taxa de erro aumentou de 6,4%. Uma possível explicação para esse fato é o

Tabela 2 - Resultados das simulações com as sílabas e com as Unidades do Inventário 1 (Santos e Alcaim, 2000)

		I	D	S	Total	Acertos
Sílabas	**DL**	3	1	2	6	98,81%
	IL	46	23	182	251	95,01%
Inventário 1	**DL**	3	0	4	7	98,60%
	IL	5	99	132	236	95,31%

DL - Dependente do Locutor IL - Independente do Locutor
I – Inserções D – Deleções S – Substituições

número de modelos utilizados por unidade. Como as sílabas são unidades com duração temporal grande e englobam vários fones, elas captam melhor a coarticulação existente na fala contínua e indiretamente captam também a prosódia dos locutores. Consequentemente, para que as diversas formas de pronúncias sejam incorporadas pelo reconhecedor é necessária a utilização de vários modelos por sílaba. No modo dependente do locutor, como a pronúncia é mais constante, as sílabas capturam de forma vantajosa as suas características.

Outro aspecto a considerar é que no modo independente do locutor, as unidades do inventário 1 foram treinadas a partir de modelos bastante precisos e já treinados no modo dependente do locutor. O aproveitamento dos modelos treinados no modo dependente para o modo independente do locutor é feito utilizando-se esses modelos como semente. Utiliza-se os modelos já existentes para fazer a segmentação das sentenças (contidas na base de dados do modo independente do locutor) em unidades fonéticas e treina-se os modelos independentes do locutor com essa nova base de dados (Santos e Alcaim, 1999).

Já para o treinamento dos modelos das sílabas o procedimento acima descrito é desnecessário tendo em vista que elas apresentam uma menor variabilidade acústica podendo ser treinadas de forma semelhante aos fones, difones ou trifones. É exatamente este fato que torna as sílabas mais atraentes que as unidades do Inventário 1 sob o ponto de vista de complexidade do treinamento.

Os resultados obtidos permitem confirmar que as sílabas são unidades fonéticas que produzem resultados vantajosos em sistemas de RVC baseados na língua portuguesa com dicionários pequenos e médios. Para esses casos as sílabas apresentam um desempenho comparável ao obtido com o inventário 1 e um treinamento que requer um esforço computacional menor. Entretanto, para dicionários maiores o número de modelos torna o treinamento inviável, quando então, o inventário 1 passa a ser mais atraente.

Finalmente, vale salientar que se fossem utilizados os trifones como unidades de reconhecimento, seria necessário o treinamento de aproximadamente 2000 modelos para a tarefa específica, em contraste com os 113 modelos treinados para as sílabas.

6 REFERÊNCIAS BIBLIOGRÁFICAS

Brown, M. K., M. A. McGee e L. R. Rabiner, Junho 1991 'Trainning Set Design for Connected Speech Recognition', *IEEE Trans. on Signal Processing*, Vol 39, No 6, pp. 1268-1281.

Chin-Hui, L. R. Rabiner e R. Pieracini, '*Speaker Independent Continuous Speech Recognition Using Continuous Density Hidden Markov Models'*, Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ, USA

Fujimura, O. Fevereiro 1975, 'Syllable as a Unit of Speech Recognition', *IEEE Trans. ASSP*, Vol 23, No 1, pp. 82-87.

Hermansky, H., Abril 1990. 'Perceptual Linear Predictive (PLP) Analysis of Speech', *J. Acoustical Society of America*, Vol 87, No 4, pp. 1738-1752.

Hunt, M. J., M. Lennig e P. Mermelstein, 1980, 'Experiments in Syllable-Based Recognition of Continuous Speech', *Proc. ICASSP'80*, pp 880-883.

Juang B. H., and L. R. Rabiner, Setembro 1990, 'A Segmental k-Means Algorithm for Estimating Parameters of Hidden Markov Models', *IEEE Trans. ASSP*, Vol 38, No. 9, pp. 1639-1641.

Kai-Fu Lee, Abril 1990, 'Contex-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition', *IEEE Trans ASSP*, Vol 38, No 4, pp. 599-609.

Nicola, J. de, e U. Infante, 1991, 'Gramática Contemporânea da Língua Portuguesa', Editora Scipione, 6ª edição.

Picone, J., Julho 1990, "Continuous Speech Recognition Using Hidden Markov Models', *IEEE ASSP Mag.*, Vol 7, No. 3, pp. 26-41.

Rabiner, L. R. , B. H. Juang, S. E. Levinson e M. M. Sondhi, Julho-Agosto 1985. 'Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixtures Density', *AT&T Tech J.*, Vol 64, No. 6, pp. 1211-1234.

Rosenberg, A. E., L. R. Rabiner et al. , Junho 1983 'Demi-syllable Based Isolated Word Recognition System', *IEEE Trans. ASSP*, Vol. 31, pp. 713-726.

Ruske, G. 1982, 'Automatic Recognition of Syllabic Speech Segments Using Spectral and Temporal Features', *Proc. ICASSP'82*, pp 550-553.

Santos, S. C. B. dos, e A. Alcaim, Março 2000, Reduced Sets of Subword Units for Continuous Speech Recognition of Portuguese, *Electronics Letters*, Vol. 36, No. 6, pp 586-588.

Santos, S. C. B., Dezembro 1997, *Reconhecimento de Voz Contínua para o Português Utilizando Modelos de Markov Escondidos*, Tese de Doutorado, CETUC, PUC-Rio.

Santos, S. C. B., e A. Alcaim, Setembro 1999. Treinamento de Modelos HMM Para Reconhecedores de Voz Contínua que Utilizam Unidades com Variabilidade Acústica, *Anais do XVII Simpósio Brasileiro de Telecomunicações*, Vila Velha, ES, , pp 444-448.

Wilpon, J. G. e L. R. Rabiner, , Junho 1985. 'A Modified k-Means Clustering Algorithm for Use in Isolated Word Recognition', *IEEE Trans. ASSP*, Vol 33, pp. 587-594.

Young, S., Setembro 1996 "A Review of Large-vocabulary Continuous-Speech Recognition". *IEEE Signal Processing Magazine*, pp. 45-57.