MODELOS DA LÍNGUA BASEADOS EM CLASSES DE PALAVRAS PARA SISTEMA DE RECONHECIMENTO DE FALA CONTÍNUA

Luis A. de Sá Pessoa[1], Fábio Violaro[1] e Plínio A. Barbosa[2]

[1]UNICAMP – FEEC - Departamento de Comunicações
[2]UNICAMP - Instituto de Estudos da Linguagem (IEL)
lpessoa@hotmail.com, fabio@decom.fee.unicamp.br, plinio@iel.unicamp.br

Resumo - Este artigo apresenta o desenvolvimento de dois modelos da língua baseados em classes de palavras: um modelo baseado em classificação gramatical de palavras e outro baseado em classificação automática de palavras utilizando o algoritmo Simulated Annealing. Alguns resultados do reconhecimento de fala contínua utilizando os modelos da língua desenvolvidos também são apresentados.

Abstract - This paper presents the development of two word class language models: one model based on grammatical classification of words and the other one based on automatic word classification using Simulated Annealing algorithm. Some results of continuous speech recognition experiments using the above language models are also presented.

Palavras-chave: Modelo da língua, classes de palavras, modelo bigram, classificação automática de palavras.

1. INTRODUÇÃO

O reconhecimento automático de fala contínua pode ser colocado como a busca pela seqüência de palavras correspondente à elocução de entrada. Os sistemas existentes baseiam-se normalmente em princípios de reconhecimento estatístico de padrões e assumem que a elocução de entrada corresponderá à seqüência de palavras mais provável, segundo os modelos adotados no sistema.

Costuma-se representar a elocução de entrada por uma seqüência $O = o_1,...,o_T$ de vetores de parâmetros extraídos do sinal de fala (cepstrais, mel-cepstrais, PLP, etc.) [10].

Para encontrar a seqüência de palavras $\hat{W} = \hat{w}_1...\hat{w}_N$, correspondente à elocução de entrada, aplica-se o critério da máxima probabilidade a posteriori, conforme a equação (1).

$$\hat{W} = \arg\max_{W} P(W \mid O) \qquad (1)$$

Utilizando a regra de Bayes e excluindo o termo $P(O)$, que não interfere na maximização, chegamos à expressão (2).

$$\hat{W} = \arg\max_{W} \{P(O \mid W).P(W)\} \qquad (2)$$

O termo $P(O \mid W)$ é avaliado pelo **modelo acústico** e representa a probabilidade do modelo da sentença $W = w_1...w_N$ gerar a seqüência observada de vetores $O = o_1,...,o_T$. Neste trabalho, as simulações foram realizadas empregando o modelo acústico desenvolvido por Morais [7] baseado no modelo híbrido HMM/MLP [8], no qual as probabilidades de emissão de símbolos dos modelos de Markov são estimadas por redes neurais Multilayer Perceptron (MLP).

O termo $P(W)$ é avaliado pelo **Modelo da Língua** e consiste na probabilidade a priori de observar a seqüência de palavras W, independente do sinal observado.

Neste trabalho, propomos dois modelos estatísticos da língua nos quais assumimos que a ocorrência de uma palavra depende da palavra anterior (modelo bigram). Os modelos propostos consideram que a relação entre os pares de palavras não será estabelecida diretamente, mas através de *classes de palavras* [11]. Apresentaremos o modelo construído a partir da classificação manual das palavras, utilizando classes gramaticais, e o modelo que emprega classificação automática de palavras através do algoritmo *Simulated Annealing* [4].

2. MODELO DA LÍNGUA BIGRAM DE CLASSES DE PALAVRAS

O problema foi estruturado de tal forma que o procedimento de busca pela seqüência de palavras inclui o cálculo do termo $P(W) = P(w_1...w_N)$.

Tratando a ocorrência das palavras como um processo estocástico, podemos representar a probabilidade $P(w_1...w_N)$ através de (3).

$$P(W) = P(w_1).\prod_{n=2}^{N} P(w_n \mid w_1...w_{n-1}) \qquad (3)$$

O modelo de língua deve ser capaz de estimar as probabilidades condicionais $P(w_n \mid w_1...w_{n-1})$ para qualquer seqüência de palavras $w_1...w_n$. Entretanto, pode-se perceber que o número de seqüências possíveis é

proibitivamente alto. Para $n = 4$ e um vocabulário de 1000 palavras, por exemplo, teríamos de estimar cerca de 10^{12} valores de probabilidade!

Para simplificar o problema, podemos assumir que a escolha da palavra w_n não depende de toda a seqüência passada (história) $h_n = w_1 ... w_{n-1}$ e, desta forma, é razoável propor que agrupemos as várias seqüências passadas em *classes de equivalência* [5]. O problema agora reduz-se a calcular (4) onde $h_n \rightarrow \Phi(h_n)$ é um mapeamento da história h_n em M classes.

$$P(w_n \mid w_1 ... w_{n-1}) = P(w_n \mid \Phi(h_n)) \tag{4}$$

Existem diversas maneiras de efetuar este mapeamento. Uma maneira simples de defini-lo é considerar que ele depende somente das m-1 últimas palavras anteriores. Neste caso, agruparemos todas as seqüências com as mesmas m-1 palavras finais na mesma classe.

Para m=2, temos a expressão (5). Usualmente, o modelo da língua definido por (5) é chamado de *modelo bigram de palavras.*

$$P(w_n \mid w_1 ... w_{n-1}) = P(w_n \mid w_{n-1}) \tag{5}$$

As probabilidades condicionais podem ser estimadas a partir da expressão (6), onde $N\{\cdot\}$ é o número de ocorrências de palavras e pares de palavras no conjunto de frases de treinamento.

$$P(w_n \mid w_{n-1}) \cong \frac{N\{w_{n-1} w_n\}}{N\{w_{n-1}\}} \tag{6}$$

Uma forma de construir modelos da língua estatisticamente mais confiáveis usando pequenas bases de treinamento consiste em reduzir a quantidade de parâmetros a serem estimados através do agrupamento das palavras em classes [11].

Existem várias maneiras de definir e aplicar o mapeamento em classes de palavras, conforme podemos observar em [9]. Assumindo um modelo bigram e definindo um mapeamento determinístico $w \rightarrow G(w)$ das V palavras em um conjunto de K classes, podemos calcular a probabilidade $P(w_n \mid w_{n-1})$, usando o mapeamento em classes para a palavra atual e também para a palavra anterior, $\Phi(w_1 ... w_{n-1}) = G(w_{n-1})$, obtendo a equação (7).

$$P(w_n \mid w_{n-1}) = P(w_n \mid G(w_n)). \\ .P(G(w_n) \mid G(w_{n-1})) \tag{7}$$

A equação (7) pode ser justificada se representarmos a ocorrência de palavras através do processo estatístico representado na Figura 1.

No processo da Figura 1, primeiro há um mapeamento da palavra anterior w_{n-1} numa classe correspondente c_i. Depois, há a transição da classe c_i para a classe c_j da palavra seguinte e, finalmente, a transição da classe c_j para a palavra w_n. As transições são vistas como probabilísticas e as probabilidades estão indicadas ao lado da respectiva transição.

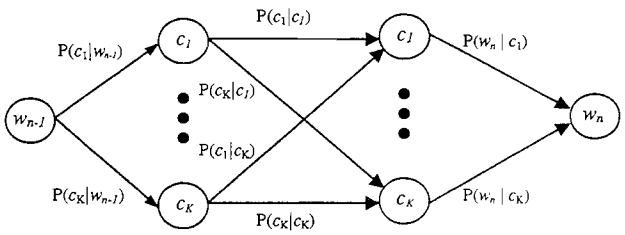

Figura 1. Processo de ocorrência de palavras.

Pode-se observar que se o mapeamento $w \rightarrow G(w)$ for determinístico (uma palavra possui somente uma classe), a transição da palavra w_{n-1} para a classe $c_i = G(w_{n-1})$ terá probabilidade 1. Também observa-se que somente a probabilidade de transição da classe $c_j = G(w_n)$ para a palavra w_n será não-nula, visto que esta palavra não pertence às demais classes. Com estas restrições, chega-se à expressão (7).

Estimar diretamente $P(w_n \mid w_{n-1})$, para todas as combinações de duas palavras, significaria avaliar um total de $V.(V-1)$ parâmetros independentes. Usando (7), teremos de avaliar somente $K.(K-1)$ parâmetros independentes referentes a $P(G(w_n) \mid G(w_{n-1}))$ e $K.(V-1)$ parâmetros independentes referentes a $P(w_n \mid G(w_n))$.

Uma vez que o número de classes, K, é normalmente muito menor que o número de palavras do vocabulário, V, teremos grande redução no número de parâmetros a serem estimados e, consequentemente, a necessidade de uma menor base de treinamento.

Neste trabalho, adotaremos modelos da língua bigram de classes de palavras, mas o mapeamento das palavras em classes não será necessariamente determinístico, podendo uma palavra estar associada a várias classes com diferentes probabilidades.

Considere o vocabulário de palavras $\mathbf{V} = \{v_1, ..., v_V\}$ e o conjunto de classes $\mathbf{C} = \{c_1, ..., c_K\}$. Usaremos o símbolo w_n para representar uma palavra qualquer na n-

ésima posição da frase e o símbolo g_n para representar a classe correspondente.

As probabilidades condicionais de palavra serão estimadas através da expressão (8) (vide [4]).

$$P(w_n \mid w_{n-1}) = \sum_{\forall g_{n-1}} \sum_{\forall g_n} [P(w_n \mid g_n).P(g_n \mid g_{n-1}).P(g_{n-1} \mid w_{n-1})] \quad (8)$$

Um mapeamento determinístico $G(.)$ pode ser obtido considerando que $G(.)$ leva uma palavra w para a classe $G(w)$ com probabilidade um e para as demais classes com probabilidade zero. Aplicando estas restrições à equação (8), obtemos a equação (7).

A classificação das palavras pode ser feita manualmente, segundo algum critério preestabelecido, possivelmente tirando vantagem da existência natural de classes de palavras na língua (verbos, substantivos, adjetivos, pronomes, etc.), ou pode ser usado algum procedimento automático normalmente baseado em conceitos de teoria de informação como em [4].

As fronteiras das frases também podem ser consideradas no cálculo de $P(W)$ através da utilização do marcador "$" como se este fosse mais uma palavra do vocabulário. Observe que o marcador de fronteira de frase define sua própria classe, sendo ele mesmo o único integrante.

Tomando a estrutura $(\$, w_1, w_2, ..., w_N, \$)$, podemos definir a equação (9).

$$P(\$\, w_1 w_2 ... w_N\, \$) = P(\$ \mid w_N) \cdot P(w_N \mid w_{N-1}) \cdot ... \cdot P(w_2 \mid w_1) \cdot P(w_1 \mid \$) \cdot P(\$) \quad (9)$$

Por simplificação, tomamos $P(\$) = 1$. Os termos $P(w_1 \mid \$)$ e $P(\$ \mid w_N)$ podem ser estimados usando as equações (10) e (11).

$$P(w_1 \mid \$) = \sum_{\forall g_1} P(w_1 \mid g_1).P(g_1 \mid \$) \quad (10)$$

$$P(\$ \mid w_N) = \sum_{\forall g_N} P(\$ \mid g_N).P(g_N \mid w_N) \quad (11)$$

Mesmo utilizando modelos bigram baseados em classes de palavras, não é possível evitar o problema de pares de classes não observados. Para garantir que as probabilidades condicionais de classe sejam não-nulas, aplicaremos o *método de interpolação linear* apresentado em [9], através da expressão (12).

$$P(c_j \mid c_i) = \alpha . \frac{N\{c_i, c_j\}}{N\{c_i\}} + p_0 \quad (12)$$

O termo $\alpha = 1 - K.p_0$ é o fator de desconto das probabilidades condicionais. O valor total descontado é redistribuído uniformemente entre os pares de classes (vide [4]). Como desejamos somente evitar que as probabilidades condicionais sejam nulas, adotamos $p_0 = 10^{-30}$.

3. MODELO BASEADO EM CLASSES GRAMATICAIS DE PALAVRAS

A estrutura de uma frase é normalmente representada através de um diagrama em árvore, como mostrado na Figura 2. Esta estrutura em árvore representa as relações hierárquicas existentes entre os diversos constituintes (SN = sintagma nominal, SV = sintagma verbal, Art = artigo, N = nome, V = verbo) que compõem a frase, que é representada pelo símbolo F.

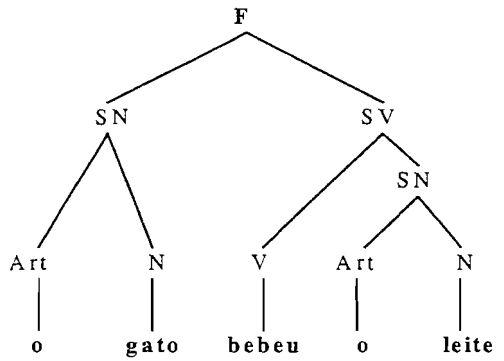

Figura 2. Exemplo de estrutura hierárquica de uma frase.

Por enquanto, estaremos preocupados somente com a *estrutura linear* da frase, ou seja, com a seqüência de categorias lexicais (artigo, nome, verbo, etc.), pois utilizaremos um modelo no qual a ocorrência de uma palavra depende apenas da palavra imediatamente anterior (modelo bigram).

No caso da frase "o gato bebeu o leite", a estrutura linear seria **(Art)(N)(V)(Art)(N)**, conforme verificamos na Figura 2. São as regularidades da estrutura linear que desejamos capturar nos modelos estatísticos que propomos neste trabalho.

A classificação das palavras utilizada no modelo bigram de classes foi baseada na classificação adotada na gramática tradicional da língua portuguesa conforme apresentado em [3]. O conjunto de classes considerado é apresentado na Tabela 1.

Para estimar as probabilidades condicionais de classe, necessárias à construção do modelo da língua, disporemos de um conjunto de 204 frases sendo 89 frases fornecidas

pelo Instituto de Estudos da Linguagem (IEL) e 115 frases retiradas do jornal Folha de São Paulo. Estas 204 frases possuem um total de 1474 palavras sendo 682 palavras distintas.

Estabelecemos que seriam utilizadas principalmente frases constituídas por uma oração ("o saldo é suficiente", "o preço do café aumentou", "a sela foi guardada numa cela nos subterrâneos do castelo", "a TELEBRÁS está investindo em pesquisa"). Frases com orações reduzidas de infinitivo (vide [3], p. 594) ocorrem em menor número ("Brasil tenta colocar satélite em órbita hoje"). Evitamos estruturas mais complexas envolvendo orações subordinadas, coordenadas, apostos, etc.

Símbolos	Significado
Sub	Substantivo
Art	Artigo
Adj	Adjetivo
num	Numeral
adv	Advérbio
prep	Preposição
conj	Conjunção
prep+art	Preposição+Artigo
prep+pron-pess	Preposição+Pronome pessoal
prep+pron-dem	Preposição+Pronome demonstrativo
pron-pess	Pronome pessoal
pron-dem	Pronome demonstrativo
pron-poss	Pronome possessivo
pron-ind	Pronome indefinido
v	Verbo
v-part	Verbo particípio
v-ger	Verbo gerúndio
v-inf	Verbo infinitivo
v-aux	Verbo auxiliar
v-lig	Verbo de ligação

Tabela 1. Classes de palavras usadas no modelo da língua

As probabilidades condicionais de classes de palavras, $P(c_j \mid c_i)$, são estimadas através da equação (13).

$$P(c_j \mid c_i) \cong \frac{N\{c_i, c_j\}}{N\{c_i\}} \qquad (13)$$

Observe que $N\{c_i, c_j\}$ representa o número de ocorrências de pares de classe $c_i\, c_j$, enquanto $N\{c_i\}$ representa o número de ocorrências da classe c_i nas frases de treinamento.

Tomamos inicialmente o arquivo texto das frases de treinamento e executamos uma classificação manual das palavras, gerando um outro arquivo texto conforme indicado na Figura 3.

Depois da classificação, executamos os programas que processam o arquivo texto com as estruturas e calculam as estatísticas desejadas. Os valores estimados de probabilidade condicional são mostrados na forma de imagem através da Figura 4, na qual tons mais escuros representam valores mais altos de probabilidade condicional

Os valores de probabilidade elevados correspondem a "grupos naturais" de classes de palavras. Somente para ilustrar, tomemos o ponto referente a $c_i = 2$ (artigo) e $c_j = 1$ (substantivo). Trata-se de um alto valor de probabilidade que pode ser justificado se observarmos que os *artigos* são normalmente seguidos por *substantivos* nas frases do português do Brasil. Os demais valores de probabilidade condicional podem ser justificados da mesma forma.

ele guarda a sela do cavalo numa prateleira de uma antiga cela

ele guarda a sela numa prateleira de uma cela do palácio

a sela do cavalo é guardada numa prateleira de uma antiga cela

(...)

Arquivo texto com frases

pron-pess v art sub prep+art sub prep+art sub prep art adj sub

pron-pess v art sub prep+art sub prep art sub prep+art sub

art sub prep+art sub v-aux v-part prep+art sub prep art adj sub

(...)

Arquivo texto com estrutura linear das frases

Figura 3. Procedimento manual de classificação das palavras

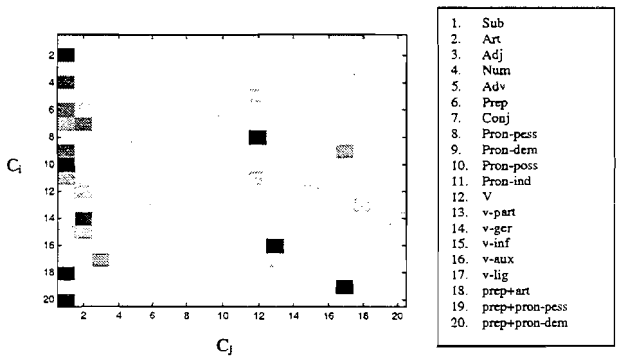

Figura 4. Probabilidades condicionais $P(c_j \mid c_i)$ na classificação manual.

4. MODELO BASEADO EM CLASSIFICAÇÃO AUTOMÁTICA

Nesta seção, apresentaremos um algoritmo que permite a classificação automática de palavras a partir de um conjunto de frases de treinamento. As classes não são previamente definidas segundo critérios lingüísticos e nem as palavras etiquetadas, somente o número de classes é fixado.

Assumiremos que cada palavra pode pertencer somente a uma classe, mas poderíamos permitir também que cada palavra pertencesse a mais de uma classe, obtendo melhores resultados na classificação, como pode ser visto em [4]. Entretanto, isso também aumentaria a complexidade do nosso modelo.

O algoritmo de classificação automática de palavras baseia-se na minimização de um parâmetro denominado **perplexidade** [4], avaliado sobre um conjunto de frases de treinamento.

Formalmente, podemos definir a perplexidade usando conceitos provenientes da teoria de informação. Definindo uma seqüência de palavras de comprimento N como $W = w_1 ... w_N$, podemos definir a perplexidade como função da entropia estimada $\hat{H}(W)$, conforme (14). A entropia $\hat{H}(W)$ é obtida a partir de estimativas da probabilidade $P(w_1 ... w_N)$, segundo a equação (15).

$$PP = 2^{\hat{H}(W)} \tag{14}$$

$$\hat{H}(W) = -\frac{1}{N}.\log_2 \hat{P}(w_1 w_2 ... w_N) \tag{15}$$

Algoritmo Simulated Annealing (SA)

Em problemas como do "caixeiro-viajante" ou em problemas de otimização combinatória de maneira geral, pode-se usar uma técnica denominada *Simulated Annealing* (SA) [1] para se atingir o ótimo global do sistema.

Observando que o problema de classificação automática de palavras visando minimizar a perplexidade é um problema de otimização combinatória, podemos aplicar a técnica de SA como método de otimização.

Considere um problema de otimização combinatória onde $f(.)$ é a função de custo adotada e S, o espaço de soluções (possíveis configurações ou estados) do sistema.

O objetivo é partir de um estado inicial i_{start} e chegar ao ótimo global i_{opt} que define o mínimo global da função de custo $f(.)$.

Dado um estado $i \in S$, podemos gerar (seleção de Monte Carlo) um estado j dentro da vizinhança (S_i) do estado i, onde a probabilidade de aceitação do estado j é dada por (16).

$$P\{\text{aceita } j \mid \text{estado\_atual} = i\} = \begin{cases} 1 & , \text{ se } f(j) \le f(i) \\ \exp\left(\dfrac{f(i) - f(j)}{\varepsilon_k}\right) & , \text{ se } f(j) > f(i) \end{cases} \tag{16}$$

O número de transições efetuadas em cada época é L_k. O parâmetro de controle tem valor inicial $\varepsilon_0 > 0$ e vai diminuindo a cada época completada.

O valor de ε_0 deve ser alto, de forma a permitir que qualquer transição seja aceita. Neste caso, o procedimento adotado foi calcular o custo médio das transições ($f(j) - f(i)$) e definir um valor de ε_0 tal que levasse a uma alta taxa de aceitação (tipicamente da ordem de 95%).

A estratégia de "resfriamento" consiste em adotar $\varepsilon_{k+1} = \alpha.\varepsilon_k$ onde $k = 0,1,2,....$ O valor típico de α usado fica entre 0,8 e 0,99 pois a diminuição do parâmetro de controle deve ser lenta, correspondendo à lenta diminuição da "temperatura do sistema".

As transições são aceitas com probabilidade decrescente e o número de transições deve ser tal que um quase-equilíbrio seja atingido a cada iteração. Dessa forma, o correto seria fazer $L_k \to \infty$ para $\varepsilon_k \to 0$. Na prática, limitamos o número de transições a um valor máximo $\overline{L}$.

A execução do algoritmo pode terminar quando a função de custo permanece inalterada durante algumas épocas ou definindo um número máximo de épocas.

A seguir, apresentamos o algoritmo *Simulated Annealing* proposto em [1].

Algoritmo Simulated Annealing

```
Begin
 INITIALIZE(i_start , ε_0 , L_0 )
 k = 0
 i = i_start
 repeat
  for l = 1 to L_k
  begin
   GENERATE ( j from S_i )
   if f(j) ≤ f(i) then i = j
   else if exp( (f(i) - f(j)) / ε_k ) > random[0,1] then i = j
  end
  k = k + 1
  CALCULATE_LENGTH( L_k )
  CALCULATE_CONTROL( ε_k )
 until stop_criterion
end
```

Para o problema de classificação automática de palavras, desejamos minimizar a perplexidade avaliada sobre o texto de treinamento. A configuração inicial deve ser tal que proporcione uma alta perplexidade. As novas configurações (estados) vão sendo geradas movendo uma palavra de sua classe para uma nova classe. Tanto a palavra quanto a nova classe são escolhidas aleatoriamente (seleção de Monte Carlo).

Um dos pontos decisivos para tornar rápidos e viáveis os programas de classificação automática é a minimização eficiente da perplexidade. Na verdade, não utilizaremos

diretamente a perplexidade como fator a ser minimizado, mas uma função denominada FL (Função-Log), definida pela equação (17), relacionada à log-probabilidade da seqüência completa de palavras nas frases de treinamento (vide [6]). Pode ser percebido que a perplexidade está relacionada a esta função através da equação (18).

$$FL = -\log_2 \hat{P}(w_1 ... w_N) \tag{17}$$

$$PP = 2^{\frac{1}{N}.FL} \tag{18}$$

Visto que se trata de uma função monótona crescente e o fator N permanece constante, podemos minimizar a perplexidade, minimizando o termo FL.

Uma vez calculada a função FL, somente alguns termos precisam ser recalculados a cada mudança de palavra de uma classe-origem para uma classe-destino.

A partir da expressão (17), adotando o modelo bigram de classes, chega-se à expressão (19) que define a função FL.

$$\begin{aligned} FL &= -\log_2\left(\prod_n \hat{P}(w_n \mid w_{n-1})\right) \\ &= -\log_2\left(\prod_n \hat{P}(w_n \mid G(w_n))\hat{P}(G(w_n) \mid G(w_{n-1}))\right) \end{aligned} \tag{19}$$

As probabilidades condicionais de (19) podem ser estimadas através das expressões (20) e (21).

$$\hat{P}(w_n \mid G(w_n)) = \frac{N\{w_n, G(w_n)\}}{N\{G(w_n)\}} \tag{20}$$

$$\hat{P}(G(w_n) \mid G(w_{n-1})) = \frac{N\{G(w_{n-1}), G(w_n)\}}{N\{G(w_{n-1})\}} \tag{21}$$

Substituindo (20) e (21) na expressão (19), e transformando o log do produtório em somatório de log's, teremos a expressão (22).

$$\begin{aligned} FL = &-\sum_n \log_2 N\{w_n, G(w_n)\} \\ &+ \sum_n \log_2 N\{G(w_n)\} \\ &- \sum_n \log_2 N\{G(w_{n-1}), G(w_n)\} \\ &+ \sum_n \log_2 N\{G(w_{n-1})\} \end{aligned} \tag{22}$$

Os termos de (22) podem ser calculados em termos do conjunto de classes e palavras do vocabulário, conforme indicado nas expressões (23) e (24).

$$\sum_n \log_2 N\{w_n, G(w_n)\} = \sum_i N\{v_i\}.\log_2 N\{v_i\} \tag{23}$$

$$\sum_n \log_2 N\{G(w_n)\} = \sum_i N\{c_i\}.\log_2 N\{c_i\} \tag{24}$$

Efetuando estas substituições, teremos finalmente a expressão (25) que corresponde a função-log definida em (17).

$$\begin{aligned} FL = &-\sum_i N\{v_i\}.\log N\{v_i\} \\ &+ 2.\sum_i N\{c_i\}.\log N\{c_i\} \\ &- \sum_{i,j} N\{c_i, c_j\}.\log N\{c_i, c_j\} \end{aligned} \tag{25}$$

Como exemplo de funcionamento do algoritmo de classificação automática, elaboramos um conjunto de 10 frases com um total de 41 palavras sendo 33 palavras distintas (vide Tabela 2). Selecionamos frases declarativas compostas por uma oração e predicado nominal, nas quais indicamos estados ou qualidades dos seres ou objetos.

a bola é redonda
o céu é azul
os sapatos são feios
as meninas estão tristes
sapos são pequenos
a menina está suja
coelhos são bonitos
a casa de maria é grande
as casas das pessoas são caras
dinheiro era importante

Tabela 2. Conjunto de frases de treinamento do exemplo

Aplicando o algoritmo SA ao texto-exemplo, obtivemos a divisão em classes mostrada na Tabela 3. Os números entre parêntesis representam o número de ocorrências de cada palavra no texto-exemplo

classe 0	classe 1	classe 2	classe 3	classe 4	classe 5
Redonda(1), azul(1), feios(1), tristes(1), pequenos(1), suja(1), bonitos(1), grande(1), caras(1), importante(1),	bola(1) céu(1) sapatos(1) meninas(1) menina(1) casa(1) casas(1)	de(1) das(1)	a(3) o(1) os(1) as(2)	sapos(1) coelhos(1) maria(1) pessoas(1) dinheiro(1)	é(3) são(4) estão(1) está(1) era(1)

Tabela 3. Divisão em classes usando *Simulated Annealing* (exemplo).

Podemos observar as seguintes classes de palavras na Tabela 3: Adjetivos (classe 0), substantivos (classe 1 e classe 4), preposições (classe 2), artigos (classe 3) e verbos (classe 5).

Os substantivos que fazem parte da estrutura *artigo+substantivo* estão na classe 0, enquanto os que fazem parte da estrutura *substantivo* (no início da oração: "sapos", "coelhos" e "dinheiro") ou *preposição+substantivo* foram colocados na classe 4.

Deve-se notar que nem sempre é possível associar um significado às diferentes classes obtidas pela classificação automática.

Para o treinamento do modelo da língua, dispomos de um conjunto de 470 frases, sendo 96 frases fornecidas pelo IEL, 115 frases retiradas do jornal Folha de São Paulo, 200 frases provenientes de [2] e 59 frases do jornal Folha da Tarde. Estas 470 frases possuem um total de 3665 palavras sendo 1472 palavras distintas.

Considerando o número total de palavras no conjunto de treinamento (3665 palavras), mas levando em conta a esparsidade natural da matriz de probabilidades condicionais, resolvemos limitar o número máximo de classes em 80.

Na Figura 5, apresentamos os valores de perplexidade final em função do número de classes de palavras no treinamento.

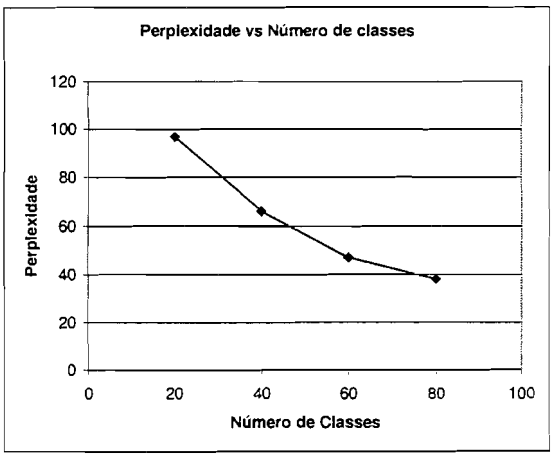

Figura 5. Comportamento da perplexidade em função do número de classes no treinamento.

O treinamento utilizou como critério de parada a condição: $\frac{\Delta FL}{FL} < 0{,}001$. Foram necessárias normalmente cerca de 500 épocas para concluir cada treinamento (cada época com 5000 mudanças de classe e uma taxa de decremento do parâmetro de controle de 0,95). O tempo de treinamento num computador Pentium II 300 MHz ficou em torno de 3h.

Na Figura 6, temos o gráfico correspondente às probabilidades condicionais de classe do modelo da língua obtido com 20 classes. Podemos verificar que mesmo utilizando somente 20 classes de palavras, a matriz de probabilidades condicionais ainda permanece esparsa.

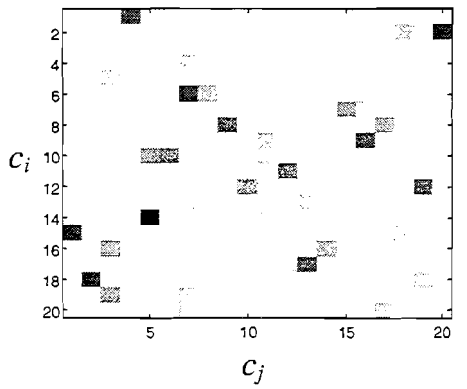

Figura 6. Probabilidades condicionais $P(c_j \mid c_i)$ na classificação automática.

5. RESULTADOS NO RECONHECIMENTO

Utilizando os modelos da língua bigram de classes de palavras, avaliamos o desempenho do sistema de reconhecimento de fala contínua em função da taxa de erro de palavra.

Foi utilizado o sistema de reconhecimento de fala contínua dependente do locutor desenvolvido por Morais [7] que emprega o modelo híbrido HMM/MLP [8] e utiliza o algoritmo de busca *Level Building*. O vocabulário utilizado foi de 312 palavras. O conjunto de teste é formado por 74 frases com um total de 499 palavras.

Comparamos a utilização do modelo bigram de classes gramaticais com a utilização do modelo que assume densidade de probabilidade gaussiana para a duração das palavras (vide [7] e [10]), avaliando o desempenho do reconhecimento nos seguintes casos (vide Tabela 4): sem utilizar o modelo da língua ou o modelo de duração de palavra, utilizando o modelo de duração de palavra (MDUR), utilizando somente o modelo da língua (ML) e empregando ambos os modelos (ML+MDUR).

Modelo Usado	Erro de Palavra (%)
<nenhum>	56,5
MDUR	24,8
ML	22,2
ML+MDUR	19,2

Tabela 4. Reconhecimento usando ML baseado em classes gramaticais.

Utilizando o modelo bigram de classes com classificação automática e o modelo de duração de palavra, obtivemos os resultados apresentados na Tabela 5.

Número de classes	Erro de Palavra (%)	Perplex. no treinamento
20	16,8	97
40	8,4	66
60	7,0	47
80	6,0	38

Tabela 5. Resultados no reconhecimento empregando modelo bigram de classes com classificação automática (SA).

A taxa de erro de palavra não decresce linearmente à medida que aumentamos o número de classes, conforme podemos constatar na Figura 7. De fato, a perplexidade final no treinamento exibe um comportamento semelhante ao comportamento da taxa de erro de palavra no reconhecimento.

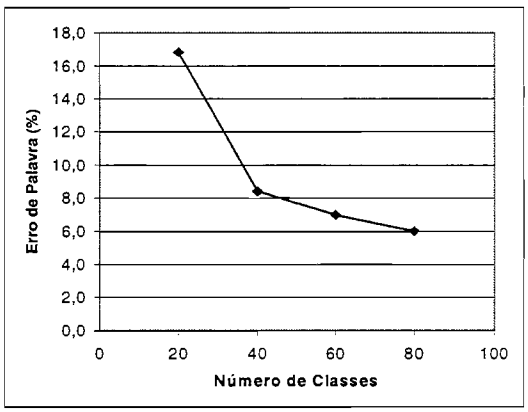

Figura 7. Comportamento do erro de palavra em função do número de classes.

Segundo [4], quando utilizamos poucas classes, aumentamos o poder de generalização do modelo da língua, mas as probabilidades não são suficientemente *restritivas* para proporcionar bons resultados no reconhecimento. Quando o número de classes é grande demais, o modelo refletirá mais as características do texto de treinamento, permitindo pouca generalização. Isso explicaria porque a taxa de erro de palavra cai rapidamente quando passamos de 20 classes para 40 classes, mas varia pouco à medida que aumentamos mais o número de classes.

Observa-se que enquanto a classificação manual em classes gramaticais utiliza somente informações morfológicas das palavras, a classificação automática tende a capturar outras relações entre as palavras (objetivando sempre minimizar a perplexidade avaliada sobre o texto de treinamento).

No modelo que usa classificação manual, o número de classes poderia ser aumentado, diferenciando as palavras quanto ao gênero, número ou mesmo características semânticas, mas isto não constitui uma tarefa simples e exige cada vez mais conhecimento lingüístico. No caso da classificação automática, podemos variar facilmente o número de classes, buscando uma melhor relação entre o número de classes e a taxa de erro de palavra no reconhecimento.

Na classificação automática, as palavras são divididas em classes de forma não-supervisionada, facilitando o processo de construção do Modelo da Língua, pois nenhum conhecimento da língua é necessário e evita-se a preparação manual do conjunto de treinamento, embora as classes obtidas não tenham necessariamente um significado, como no caso da classificação manual.

Todas estas questões apontam para vantagens na utilização da classificação automática de palavras na construção de modelos da língua, seja pela facilidade, seja pelos melhores resultados. Não se deve esquecer, entretanto, que utilizando classes gramaticais temos um controle maior do que está sendo realizado e também poderemos aproveitar uma classificação gramatical mais sofisticada em modelos da língua que utilizem gramática, por exemplo.

<nenhum>	**Modelo de duração de palavra**
, isto o é suficiente , , isto o é suficiente que , , isto o o o é suficiente , , isto o o o é suficiente que , , o saldo o é suficiente , , o saldo o é a suficiente , , os a o o é a suficiente , , os a o do o é a suficiente ,	, isto o é suficiente , , isto é suficiente , , e isto o é suficiente , , e e isto o é suficiente , , o saldo o é suficiente , , o saldo o o é suficiente , , o saldo o o é a suficiente , , o saldo é suficiente ,
Modelo de 20 classes gramaticais	
, isto é suficiente , , isto é suficiente , , , isto é suficiente , , , , isto é suficiente , , , , , o saldo o queda suficiente , , o saldo é suficiente , , , o saldo o queda suficiente , , , o saldo o queda suficiente , ,	
Modelo de 20 classes gramaticais e modelo de duração de palavra	**Modelo de 40 classes automáticas e modelo de duração de palavra**
, isto é suficiente , , isto é suficiente , , , isto é suficiente , , , , isto é suficiente , , , , , o saldo é suficiente , , o saldo o queda suficiente , , o saldo o queda suficiente , , , os plano real suficiente ,	, isto é suficiente , , isto é suficiente , , , isto é suficiente , , , , isto é suficiente , , , , , o saldo com é suficiente , , o saldo é suficiente , , o saldo com é suficiente , , , o saldo com é suficiente , , ,

Tabela 6. Frases reconhecidas.

Tomando as frases de teste "isto é suficiente" e "o saldo é suficiente", apresentamos na Tabela 6 os resultados (frases reconhecidas) obtidos no final do algoritmo de busca referentes aos quatro níveis do *Level Building* com maior

verossimilhança final, organizados em ordem decrescente de verossimilhança, ou seja, a frase reconhecida corresponde à primeira frase de cada lista. Na apresentação das frases, a "vírgula" representa a presença do silêncio.

De forma geral, o reconhecimento utilizando modelo da língua produz frases mais "próximas" das frases corretas, pois existem menos inserções de palavras curtas (como "os", "a", "o", "a" e "do"). Sem utilizar modelo da língua ou modelo de duração de palavra, obtemos seqüências como **"isto o o o é suficiente"** e **"os a o do é a suficiente"**. Utilizando o modelo de duração de palavras, conseguimos melhores resultados, mas ainda ocorrem erros como **"e e isto o é suficiente"**. Aplicando os modelos da língua, conseguimos melhorar sensivelmente as frases obtidas.

6. CONSIDERAÇÕES FINAIS

Neste trabalho, foram apresentados dois modelos da língua para um sistema de reconhecimento de fala contínua baseado em modelo híbrido HMM/MLP: um modelo bigram de classes gramaticais e um modelo bigram de classes com classificação automática utilizando o algoritmo *Simulated Annealing*.

O modelo bigram de classes gramaticais possui suas classes definidas segundo a classificação gramatical de palavras adotada pela gramática tradicional. A classificação das palavras é feita manualmente, tornando a preparação do texto de treinamento custosa e exigindo bom conhecimento da língua. Este modelo possui a vantagem de que a inclusão de uma nova palavra no vocabulário implicaria somente na indicação das classes gramaticais a que ela pertence.

O modelo da língua bigram de classes construído com classificação automática de palavras proporcionou melhores resultados no reconhecimento do que o modelo bigram de classes gramaticais. Utilizando modelos de 20 classes, obtivemos taxa de erro de palavra de 19,2% usando classes gramaticais e 16,8% usando classificação automática.

No modelo que usa classificação manual, o número de classes poderia ser aumentado, diferenciando as palavras quanto ao gênero, número ou mesmo características semânticas, embora isto não seja uma tarefa muito simples. No caso da classificação automática, podemos variar facilmente o número de classes, buscando uma melhor relação entre o número de classes e a taxa de erro de palavra. Nesse sentido, poderíamos tentar definir um mecanismo que encontrasse o número *ótimo* de classes, pois, conforme aumentamos o número de classes, obtemos diminuições cada vez menores da perplexidade e da taxa de erro de palavra. Entretanto, deixaremos essa idéia para ser abordada em outros trabalhos.

Na classificação automática, as palavras são divididas em classes de forma não-supervisionada, facilitando o processo de construção do modelo da língua (nenhum conhecimento da língua é necessário e evita-se a preparação manual do conjunto de treinamento), entretanto as classes obtidas não têm significado definido, como no caso da classificação manual.

REFERÊNCIAS

[1] Aarts, E., Korst, J., *Simulated Annealing and Boltzman Machines*. John Wiley & Sons, 1989.

[2] Alcaim, A., Solewics, J.A., Moraes, J.A., *Freqüência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro.* Revista da Sociedade Brasileira de Telecomunicações, v. 7, n. 1, Dezembro 1992.

[3] Cunha, C., *Nova Gramática do Português Contemporâneo*. Editora Nova Fronteira, 1985.

[4] Jardino, M., Adda, G., *Automatic Determination of a Stochastic Bi-Gram Class Language*. Proceedings of ICASSP, II – 41, 1993.

[5] Jelinek, F., *Language Modeling for Speech Recognition*. Proceedings of the ECAI 96: Workshop on Extended Finete State Models of Language, ed. A. Kornai, 1996.

[6] Martin, S., Liermann, J., Ney, J., *Algorithms for Bigram and Trigram Word Clustering*. EUROSPEECH'95, p. 1253-1256, 1995.

[7] Morais, E. S., *Reconhecimento Automático de Fala Contínua Empregando Modelos Híbridos ANN + HMM*. Tese de Mestrado, UNICAMP, Campinas, 1997.

[8] Morgan, N., Bourlard, H.A., *Continuous Speech Recognition: An Introduction to the Hibrid HMM/Connectionist Approach*. IEEE Signal Processing Magazine, p. 25-42, Maio, 1995.

[9] Ney, H., Essen, U., Kneser, R., *On Structuring Probabilistic Dependences in Stochastic Language Modelling*. Computer Speech and Language, v. 8, p. 1-38, 1994.

[10] Rabiner, L., Juang, B.H., *Fundamentals of Speech Recognition*. PTR Prentice-Hall, 1993.

[11] Suhm, B., Waibel, A., *Towards Better Language Models for Spontaneous Speech*. Proceedings of ICSLP, 1994.

Luis A. S. Pessoa nasceu em Jaboatão, Pernambuco, em 17 de abril de 1973. Graduou-se em Engenharia Elétrica pela Universidade Federal de Pernambuco em setembro de 1996. Recebeu o título de Mestre em Engenharia Elétrica pela Faculdade de Engenharia Elétrica e de Computação da UNICAMP em fevereiro de 1999. Atualmente trabalha na Fundação CPqD como pesquisador em Telecomunicações. Seus principais interesses em pesquisa são Reconhecimento de Fala, Processamento Digital de Sinais e Comunicações Móveis.

Fábio Violaro nasceu em Campinas, São Paulo, em 8 de dezembro de 1950. Graduou-se em Engenharia Elétrica, obteve os títulos de Mestre e Doutor em Engenharia Elétrica, todos pela Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas

(FEEC/UNICAMP), em 1973, 1975 e 1980, respectivamente. Atualmente é professor titular do Departamento de Comunicações da FEEC e coordenador do Laboratório de Processamento Digital de Fala. Suas áreas de interesse se concentram em Processamento Digital de Sinais de Fala: Análise, Codificação, Reconhecimento e Síntese.

Plínio A. Barbosa nasceu em Itabuna, Bahia, em 1966 . Formou-se em Engenharia Eletrônica pelo ITA, em São José dos Campos, em 1988 e recebeu o título de Mestre em Ciência pelo mesmo instituto em 1990. Doutorou-se na área de Ciência da Fala em 1994 pelo Institut de la Communication Parlée, em Grenoble, França. É professor do Departamento de Lingüística do Instituto de Estudos da Linguagem (IEL/UNICAMP) e pesquisador do Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE) onde desenvolve pesquisa sobre a Estrutura Rítmica do Português do Brasil com aplicações em Síntese da Fala. Seus interesses também cobrem as áreas de Fonética e Fonologia do Português do Brasil, Fundamentos Cognitivos do Ritmo, Produção de Fala, Bases Fonéticas para Síntese de Fala e Redes Neurais.