Efeitos da Segmentação em Sistemas Híbridos ANN+HMM

José Antonio Moreira de Rezende

Dissertação apresentada ao Instituto Nacional de Telecomunicações, como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Carlos Alberto Ynoguti

Santa Rita do Sapucaí
2005

Dissertação defendida e aprovada em 09/09/2005, pela comissão julgadora:

Prof. Dr. Carlos Alberto Ynoguti - DCOM - INATEL

Prof. Dr. Luís Geraldo P. Meloni - DECOM - FEEC - UNICAMP

Prof. Dr. Sandro Adriano Fasolo - DTE - INATEL

Coordenador do Curso de Mestrado

Prof. Dr. Adonias Costa da Silveira

Aos meus pais, José e Maria.
Às tias Nilza (in memoriam) e Marília.

Agradecimentos

A Deus por todos os obstáculos que foram colocados em minha caminhada para que pudesse chegar a conclusão desta Dissertação.

Ao Prof. Ynoguti pela valorosa orientação, pelo incentivo, por ter acreditado em meu potencial e pela ajuda imprescindível que culminou neste trabalho.

Aos meus pais, José e Maria, por tudo que fizeram por mim naquilo que fosse possível em todas as etapas da minha vida.

Aos meus familiares, pelo apoio em todos os momentos.

A Eliana, pela total compreensão nos momentos em que precisei me ausentar para poder concluir este trabalho.

A Edmilson Morais, pela grande ajuda nos pontos cruciais do desenvolvimento do sistema.

Aos amigos do LABPG por todas as críticas, sugestões e por também fazerem do laboratório um bom ambiente de trabalho.

Aos funcionários do INATEL pela cordialidade.

A Dilú e César, do Diretório Acadêmico, pela amizade.

A CAPES pelo financiamento parcial desta pesquisa.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ANN	*Artificial Neural Network* – Rede Neural Artificial
CLSP	*Center for Language and Speech Processing* – Centro de Processamento de Fala e Linguagem
CSLU	*Center for Spoken Language Understanding* – Centro para Entendimento da Língua Falada
CU-HTK	*Cambridge University - Hidden Markov Model Toolkit* – Universidade de Cambridge - Pacote Modelo Oculto de Markov
EBP	*Error Backpropagation* – Retropropagação do Erro
FDP	*Função Densidade de Probabilidade*
GMM	*Gaussian Mixture Model* – Modelo de Mistura de Gaussianas
HMM	*Hidden Markov Model* – Modelo Oculto de Markov
HTK	*Hidden Markov Model Toolkit* – Pacote Modelo Oculto de Markov
ICSI	*International Computer Science Institute* – Instituto Internacional de Ciência da Computação
ISIP	*Institute for Signal and Information Processing* – Instituto de Processamento de Sinais e Informação
JHU	*Johns Hopkins University* – Universidade Johns Hopkins
MAP	*Maximum a Posteriori Probability* – Máxima Probabilidade a Posteriori

MFCC *Mel Frequency Cepstrum Coefficients* – Coeficientes de Frequência Mel-Cepstral

ML *Maximum Likelihood* – Máxima Verosimilhança

MLP *Multilayer Perceptrons* – Perceptrons Multicamadas

MSE *Minimum Square Error* – Mínimo Erro Quadrático

NIST *National Institute of Standards and Technology* – Instituto Nacional de Padrões e Tecnologia

REMAP *Recursive Estimation-Maximization of a Posteriori Probability* – Estimação-Maximização Recursiva da Probabilidade a Posteriori

SLP *Single Layer Perceptron* – Percepton de Camada Única

SCLITE *Score Lite* – Ferramenta de avaliação que faz parte do pacote SCTK

SCTK *Speech Recognition Scoring Toolkit* – Pacote de Avaliação de Sistemas de Reconhecimento de Fala

SPRACH *SPeech Recognition Algorithms for Connectionist Hybrids*) – Algoritmos de Reconhecimento de Fala para Redes Conexionistas Híbridas

SVM *Support Vector Machines* – Máquinas de Vetor de Suporte

SRM *Structural Risk Minimization* – Minimização do Risco Estrutural

SWBD *Switchboard conversation* – Base de dados utilizada no sistema de avaliação *Hub-5*

WER *Word Error Rate* – Taxa de Erro de Palavras

Lista de Símbolos

$\mathbf{X}$	Seqüência de vetores acústicos associada a uma locução específica.
N	Comprimento da seqüência de vetores acústicos $\mathbf{X}$.
$\mathbf{X}_e$	Conjunto de vetores acústicos $\{x_1, \ldots, x_{N_e}\}$ usados como exemplos de treinamento da rede neural.
N_e	Número de exemplos de treinamento do conjunto $\mathbf{X}_e$.
$\mathbf{X}_{n-c}^{n+d}$	Subseqüência de vetores acústicos $\mathbf{X}$ com uma informação contextual de c vetores no passado e d vetores no futuro, formando assim uma janela de $c+d+1$ vetores acústicos.
Q	Conjunto de estados do HMM a partir dos quais os modelos de fones, palavras e sentenças serão construídos.
M_i	Modelo da *i-ésima* sentença construído a partir da seqüência de L_i estados estacionários $q_l \in Q$.
W_i	Conjunto de modelos de palavras que formam o modelo M_i.
M_{ω_j}	HMM associado a sequência X_j de vetores acústicos.
Θ	Conjunto de parâmetros que descrevem todos os modelos M_i, que engloba os parâmetros da rede neural e do HMM.
a_{jk}	Probabilidade de ocorrer uma transição entre os estados j e k.
$b_j(x_n)$	Probabilidade de ocorrer um símbolo x_n quando se atingir o estado j.
L	Número de estados do modelo.
L_{M_i}	Número de estados do *i-ésimo* modelo M_i.
π_j	Probabilidade de o processo iniciar-se no estado j.
η	Passo de aprendizagem da rede neural.
Γ_i	Seqüência de estados de comprimento N da *i-ésima* sentença.
s	Sobreposição entre janelas adjacentes.
q_k^n	Estado k visitado no instante n.

K	Número de fones ou classes de saída da rede neural artificial.
G	Número de elementos distintos gerados pelo quantizador vetorial.
q_l	Estado do HMM que pertence a sequência de estados Γ_i.
W_i	Modelo da *i-ésima* palavra.
$P(\cdot)$	Probabilidade de ocorrência de um determinado evento.
$f(\cdot)$	Função densidade de probabilidade.
$Var\{\cdot\}$	Operador variância.
$E\{\cdot\}$	Operador esperança.
$y_k(x_n)$	*k-ésima* classe de saída da ANN que corresponde a resposta a um estímulo representado pelo vetor de entrada x_n.
$d_k(x_n)$	Resposta desejada da *k-ésima* classe de saída da ANN, para o correspondente vetor x_n de entrada.
$\upsilon_j(x_n)$	Potencial de ativação do *j-ésimo* neurônio no instante n.
$\varphi(\upsilon_j)$	Função de ativação do neurônio j.
$\xi_n^{(i)}(j,k)$	Probabilidade de estar no estado q_j, no instante n, e no estado q_k no instante $n+1$ do modelo M_i.
$\alpha_n(k)$	Verossimilhança *forward* do estado k no instante n.
$\beta_n(k)$	Verossimilhança *backward* do estado k no instante n.
c_n	Fator de normalização das verossimilhanças *forward* e *backward*, independente do estado k.
$\hat{\alpha}_n(k)$	Verossimilhança *forward* normalizada.
$\hat{\beta}_n(k)$	Verossimilhança *backward* normalizada.
$\gamma_n(k)$	Probabilidade *a posteriori* global de um estado q_k ser visitado no instante n – alvos suaves.
D_j	Duração média (em milissegundos) do fone associado ao estado q_j.
f_i	*i-ésimo* fone.
t_i	*i-ésima* marca de segmentação do fone f_i.
$dist_n$	Distorção da *n-ésima* época de treinamento.
E_n	Erro quadrático médio da *n-ésima* época de treinamento.
σ_T	Desvio padrão de todas as componentes de todos os vetores acústicos

Resumo

Os modelos ocultos de Markov se tornaram uma ferramenta largamente utilizada na tarefa de reconhecimento de fala pela sua sólida abordagem probabilística. Porém, ao se treinar as cadeias de Markov através do método de máxima verossimilhança esta robustez é prejudicada pela falta de poder discriminativo da técnica. Ao longo dos anos foram desenvolvidos algoritmos cada vez mais eficientes onde podemos notar a evolução crescente do desempenho de um dado sistema. Uma das abordagens usadas para o aumento de tal desempenho é a utilização das Redes Neurais Artificiais em conjunto com os HMM's, conhecida na literatura por modelos híbridos ANN+HMM, onde as Redes Neurais tratam da variabilidade acústica enquanto que os HMM's tratam das variabilidades temporais de um determinado conjunto de locuções. A vantagem de se utilizar Redes Neurais é que o seu algoritmo de treinamento é, por natureza, discriminativo e portanto elimina-se a necessidade de fazer suposições quanto a distribuição estatística das probabilidades de emissão de símbolos.

Nesta Dissertação foi proposta uma investigação dos efeitos das segmentações manual, assim como a simulação dos efeitos de segmentação no desempenho de um sistema híbrido ANN+HMM de reconhecimento de fala contínua, com algoritmo de treinamento chamado REMAP (*Recursive Estimation and Maximization of A Posteriori Probabilities*), no qual consiste na modelagem das transições dos fones, além de utilizar-se de uma técnica de treinamento discriminativo chamada Máxima Probabilidade a Posteriori, onde se maximiza um determinado modelo ao mesmo tempo em que se minimiza a probabilidade a posteriori dos modelos rivais.

Palavras-chave: Modelos ocultos de Markov, redes neurais artificiais, REMAP, modelos híbridos ANN+HMM, efeitos de segmentação.

Abstract

The Hidden Markov Models has become the widely used speech recognition framework due to its solid probabilistic approach. However, this robustness is damaged when the Markov chains are trained under the maximum likelihood method, which is poor in discriminative power. Until now several efficient techniques and algorithms have been developed to increase the performance of a given speech recognition system. For example, the use of an Artificial Neural Network working with a Hidden Markov Models, that yields to a Hybrid Model ANN+HMM, where a Neural Network is in charge of modeling the acoustic variability and the HMM's are in charge of modeling the temporal variability of a given set of training sentences. Because of the Neural Network discriminative learning, there is no need to assume for assumption about the statistical distribution of the emission probability.

In this dissertation we purpose an investigation about the effects of a manual segmentation and a simulation of manual segmentation errors on the hybrid ANN+HMM continuous speech recognition system performance, trained under REMAP (*Recursive Estimation and Maximization of A Posteriori Probabilities*) algorithm, which improves a posterior probability of the correct model while reducing the probabilities of all rival models.

Keywords: Hidden Markov models, Artificial Neural Networks, REMAP, Hybrid Models ANN+HMM, segmentation effects.

Capítulo 1

Introdução

1.1 Considerações iniciais

Para a tarefa de reconhecimento de fala, surgiu no início dos anos 80 os Modelos Ocultos de Markov (*Hidden Markov Models – HMM*), sendo largamente utilizados até então e considerados o estado da arte devido a sua flexibilidade em função da modelagem matemática baseada em cálculos probabilísticos. Entretanto, para que um sistema baseado em HMM's seja matematicamente tratável, é necessário que se façam algumas suposições. Por exemplo, os modelos seguem a hipótese de Markov (HMM de primeira ordem) e é assumido que a função densidade de probabilidade dos símbolos emitidos seja uma mistura de gaussianas.

Outra técnica que conseguiu posição de destaque a partir da metade dos anos 80 foi a aplicação de Redes Neurais Artificiais (*Artificial Neural Networks – ANN*), que vieram como uma alternativa às suposições limitantes que são lançadas ao se utilizar HMM's e que, no caso da abordagem conexionista, não são levadas em consideração. Isto se deve ao fato de seu algoritmo ser inerentemente discriminativo, além de possuir a capacidade de executar uma modelagem não paramétrica dos dados de entrada, o que faz com que não surja a necessidade de se fazer suposições quanto à sua distribuição estatística. Apesar destas vantagens, as Redes Neurais mostraram-se pouco eficazes no tratamento das variabilidades temporais, campo onde os HMM's conseguem boa desenvoltura.

Com o intuito de aproveitar as vantagens das duas abordagens mencionadas nos parágrafos anteriores, surgiram os modelos híbridos ANN+HMM no início dos anos 90, onde os HMM's cuidam das variabilidades temporais, enquanto as ANN's cuidam das variabilidades acústicas. Desta forma, aproveita-se o bom modelamento temporal dos HMM's e a discriminabilidade das ANN's.

1.2 Estado da Arte

Hoje em dia, em função da padronização dos procedimentos de avaliação (taxa de erro de palavras, taxa de acerto de fonemas, etc.), criação de grandes bases de dados para treinamento, teste e avaliação (TIMIT, TIDIGITS, ATIS e WSJ) e, ainda, o aumento na velocidade de processamento dos computadores, é possível comparar o desempenho de várias ferramentas e daí poder apontar aquela que é a melhor para uma determinada aplicação.

Como exemplo disto, o NIST (*National Institute of Standards and Technology*), criou uma série de critérios de avaliação ao longo dos anos em que todos os sistemas de reconhecimento de fala envolvidos fossem treinados com uma mesma base de dados. Atualmente, o critério utilizado é o *Hub-5 Evaluation*.

A Tabela 1.1 mostra os resultados obtidos pelos sistemas envolvidos nesta avaliação, em termos de taxa de erros de palavras, sendo o HTK a ferramenta que obteve o melhor desempenho:

Tabela 1.1: *Desempenho dos sistemas de fala contínua submetidos à avaliação Hub-5 no ano de 2001 [1].*

Participante	*framework*	SWBD-saved	SWBD-2ph3	SWBD-cell
AT&T	HMM	24,0%	27,5%	33,5%
BBN	HMM	20,5%	26,3%	32,7%
CU-HTK	**HMM**	**19,8%**	**24,5%**	**29,2%**
IBM	HMM	22,0%	27,5%	33,5%
ISIP	HMM	35,6%	42,1%	48,2%
JHU	GMM	25,2%	30,2%	35,5%
SRI	HMM	23,5%	29,0%	34,5%

De acordo com a tabela acima, os testes do critério *Hub-5* são feitos em três conjuntos de avaliação distintos designados por SWBD-saved (conjunto original), SWBD-2ph3 (conversações telefônicas da região do sul dos Estados Unidos) e SWBD-cell (conversações em telefones celulares), sendo SWBD a abreviação para *Switchboard conversation*. Estes conjuntos contém 20 conversações telefônicas de 5 minutos cada (exceto em SWBD-cell que é composta de 20 conversações de 6 minutos).

Existem outros sistemas disponíveis desenvolvidos pelas universidades ao redor do mundo e que não estiveram presentes nas avaliações de 2001. Abaixo, estão listadas algumas ferramentas conhecidas:

- *Sphinx* – desenvolvido pela *Carnegie Mellon University*, baseado em modelos ocultos de Markov.

- CSLU (*Center for Spoken Language Understanding*) – pertencente a *Oregon Graduate Institute of Science and Technology*, possui o sistema baseado em modelos híbridos ANN+HMM.
- SPRACH (*SPeech Recognition Algorithms for Connectionist Hybrids*) – desenvolvido pelo ICSI (*International Computer Science Institute*), situado na Universidade da Califórnia, em conjunto com outros institutos de pesquisa. Outra abordagem baseada em modelos híbridos.

1.3 Descrição do Trabalho

Este trabalho visou o estudo, a implementação computacional, a análise e a interpretação dos resultados de um sistema de reconhecimento de fala contínua baseado em Modelos Híbridos ANN+HMM. A implementação do módulo de treinamento foi baseada nas pesquisas de Konig que culminaram na criação do algoritmo REMAP (*Recursive Estimation-Maximization of A Posteriori Probabilities*) [2, 3].

Investigou-se os efeitos causados no desempenho do sistema estudado, através do deslocamento das marcas originais de segmentação manual. Este estudo vem mostrar até que ponto um segmentador automático poderá cometer erros na segmentação das locuções sem afetar drasticamente no resultado final, que é a taxa de acerto de palavras.

1.4 Estrutura da Dissertação

Esta dissertação está estruturada nos seguintes capítulos:

Capítulo 2 – Apresenta os Modelos Híbridos ANN+HMM. Primeiramente serão abordados os Modelos Ocultos de Markov e os elementos que o compõem. Também será analisado o critério ML, mostrando as suposições que limitam o seu poder de discriminabilidade para um sistema de reconhecimento de fala contínua. Uma seção será dedicada às Redes Neurais MLP no contexto de estimador estatístico e as suas vantagens em relação ao HMM serão listadas.

Capítulo 3 – Aborda o Algoritmo REMAP que, de forma iterativa, faz a estimativa dos alvos suaves de uma Rede Neural, em função dos valores de suas saídas, fazendo com que maximize a probabilidade *a posteriori* de um dado modelo, ao mesmo tempo minimizando os modelos rivais.

Capítulo 4 – Lista os blocos funcionais utilizados para a montagem de todo o sistema de reconhecimento, assim como as sub-unidades fonéticas. Será mostrado como foi elaborado o algoritmo para a simulação de erros de segmentação, cujas

novas durações dos fonemas são tiradas a partir do deslocamento das marcas de segmentação originais advindas da segmentação manual.

Capítulo 5 – Mostra os resultados dos experimentos realizados ao longo da pesquisa e o melhor conjunto de parâmetros que levaram a tal objetivo. A seguir são feitos os respectivos comentários.

Capítulo 6 – São feitas as considerações finais e as propostas para trabalhos futuros.

Apêndice A – São apresentadas as frases que compõem a base de dados.

Apêndice B – Apresenta o vocabulário contendo o conjunto de palavras que podem ser reconhecidas pelos modelos híbridos ANN+HMM.

Apêndice C – É feita uma listagem de algumas frases reconhecidas pelo sistema.

Apêndice D – Apresenta a ferramenta de avaliação SCLITE (*Score Lite*), que faz parte do pacote SCTK (*Speech Recognition Scoring Toolkit*) [4], desenvolvido pelo NIST (*National Institute of Standards and Technology*), que foi utilizado nos cálculos de taxa de acerto de palavras das sentenças reconhecidas.

Capítulo 2

Modelos Híbridos ANN+HMM

2.1 Introdução

Este capítulo apresenta uma breve descrição das teorias dos modelos ocultos de Markov e das redes neurais artificiais, mais precisamente da arquitetura *Multi-layer Perceptron – MLP*. A seguir é feita a junção entre as duas abordagens resultando nos modelos híbridos ANN+HMM, sendo os HMM's responsáveis pelas variações temporais da fala e as ANN's pelas suas variações acústicas. É apresentada uma demonstração que justifica o papel da rede neural de estimador de probabilidades. Ao final do capítulo é mostrado o uso de informação contextual.

2.2 Modelos Ocultos de Markov

2.2.1 Definição

Os modelos ocultos de Markov são estruturas que definem dois processos estocásticos: o primeiro processo é a seqüencia L de estados que compõem o modelo e o segundo é a seqüencia dos símbolos de saída. O modelo é chamado de oculto pois existe um fenômeno que não é diretamente observável, mas que influencia diretamente na análise da seqüencia de observação $\mathbf{X} = \{x_1, \ldots, x_n, \ldots, x_N\}$, sendo x_n o *n-ésimo* vetor acústico da sentença parametrizada $\mathbf{X}$ e N o número de vetores. Tal fenômeno é a seqüencia de estados $\Gamma_i = \{q_1, \ldots, q_l, \ldots, q_L\}$ de comprimento N, da *i-ésima* sentença. É também chamado de Markov (de primeira ordem), pois a probabilidade de estar em um determinado estado q_k no instante n depende do estado anteriormente visitado q_j^{n-1}.

A topologia utilizada para representação de uma sentença é a *left-right* (ou modelo de Bakis), como mostra a Figura 2.1.

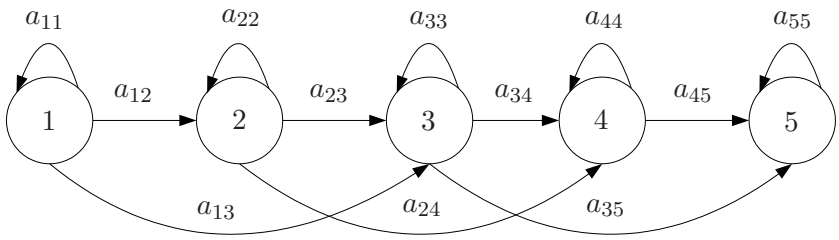

Figura 2.1: *Modelo left-right de um HMM de cinco estados.*

Desta forma, podemos especificar um HMM através dos seguintes elementos:

1. Distribuição de probabilidade de transição de estados $A = \{a_{ij}\}$ dada por $a_{ij} = P(q_j^n | q_i^{n-1})$, $1 \leq (i,j) \leq L_{M_i}$, sendo L_{M_i} o número de estados do *i-ésimo* modelo M_i.

Assume-se que a distribuição das probabilidades de transição é a mesma ao longo de toda seqüencia de observação $\mathbf{X}$. Vale ressaltar que as seguintes condições devem ser satisfeitas:

$$a_{ij} \geq 0\,, \quad 1 \leq (i,j) \leq L_{M_i} \tag{2.1}$$

$$\sum_{j=1}^{L_{M_i}} a_{ij} = 1\,, \quad 1 \leq i \leq L_{M_i} \tag{2.2}$$

2. Uma função de distribuição de probabilidade (*fdp*) de emissão de símbolos $B = \{b_j(x_n) = p(x_n|q_j)\}$: a probabilidade do estado j gerar o símbolo x_n. Quanto à sua natureza, há uma divisão dos tipos de HMM em discretos e contínuos. Nos HMM's discretos, os símbolos são escolhidos a partir de um alfabeto finito de G elementos distintos [5, 6]. Já no caso dos HMM's contínuos, assume-se que o formato da *fdp* é gerado através de uma mistura de um número finito de gaussianas multimensionais.

3. Uma função massa de probabilidade (*fmp*) de estado inicial $\Pi = \{\pi_i\}$:

$$\pi_i = P(q_i^1)\,, \quad 1 \leq i \leq L_{M_i} \tag{2.3}$$

onde q_i^1 é o estado i no instante $n = 1$. Pelo fato do sinal de fala ser seqüencial, o processo sempre se inicia a partir do primeiro estado. Portanto teremos $\pi_1 = 1$ e $\pi_i = 0$, $\forall i > 1$.

Assim, pode-se dizer que um HMM é simplificadamente representado por $M = \{A, B, \Pi\}$.

2.2.2 O critério ML

Dado o modelo M_i da *i-ésima* sentença e uma seqüencia de vetores acústicos (seqüência de observação) $\mathbf{X}$, a probabilidade do modelo gerar uma determinada seqüência de palavras $W_i = \{w_1, \ldots, w_I\}$, tal que W_i é o conjunto de modelos de palavras que formam o modelo M_i é dada por $P(M_i|\mathbf{X})$. Aplicando a regra de Bayes, tem-se:

$$P(M_i|\mathbf{X}) = \frac{P(\mathbf{X}|M_i)P(M_i)}{P(\mathbf{X})} \tag{2.4}$$

onde $P(\mathbf{X}|M_i)$ representa a contribuição do *modelo acústico*, enquanto $P(M_i)$ é a contribuição do *modelo de linguagem* [7, 8, 9]. Ou seja, deve-se encontrar a seqüência de palavras W_i que maximize $P(M_i|\mathbf{X})$ [10].

Durante a etapa de treinamento, $P(\mathbf{X})$ é dependente dos parâmetros Θ de todos os modelos possíveis [11]. Explicitando esta dependência tem-se:

$$P(\mathbf{X}) = P(\mathbf{X}|\Theta) \tag{2.5}$$

onde Θ é o conjunto de parâmetros que descrevem todos os modelos de treinamento.

Como os modelos são independentes entre si, $P(\mathbf{X}|\Theta)$ pode ser escrito em termos de probabilidade marginal:

$$P(\mathbf{X}|\Theta) = \sum_{k=1}^{K} P(\mathbf{X}, M_k|\Theta) = \sum_{k=1}^{K} P(\mathbf{X}|M_k, \Theta)P(M_k, \Theta) \tag{2.6}$$

Substituindo (2.6) em (2.4), vem:

$$P(M_i|\mathbf{X}, \Theta) = \frac{P(\mathbf{X}|M_i, \Theta)P(M_i, \Theta)}{\sum_{k=1}^{K} P(\mathbf{X}|M_k, \Theta)P(M_k, \Theta)} \tag{2.7}$$

reescrevendo o somatório do denominador da Equação (2.7), de tal forma que o termo que corresponde a maximização do modelo M_i fique evidente em relação aos modelos rivais $k \neq i$, temos:

$$P(M_i|\mathbf{X}, \Theta) = \frac{P(\mathbf{X}|M_i, \Theta)P(M_i, \Theta)}{P(\mathbf{X}|M_i, \Theta)P(M_i, \Theta) + \sum_{k \neq i} P(\mathbf{X}|M_k, \Theta)P(M_k, \Theta)} \tag{2.8}$$

Como pode notar na equação acima, para que se possa chegar à maximização de $P(M_i|\mathbf{X}, \Theta)$, deve-se utilizar o conjunto Θ de parâmetros que compreende todos os modelos envolvidos no treinamento, sendo este o chamado critério da

Máxima Probabilidade a Posteriori (*Maximum a Posteriori Probability – MAP*). O critério da Máxima Verossimilhança (*Maximum Likelihood – ML*) é o resultado da simplificação a ser feita na Equação (2.8) a respeito do uso exclusivo do conjunto de parâmetros daquele modelo a ser maximizado (θ_i). Desta forma, o algoritmo *Baum-Welch* fará uma otimização modelo a modelo, ou seja, somente serão utilizados os parâmetros pertencentes ao modelo a ser maximizado. O preço a ser pago por esta abordagem é que não sendo levados em consideração os modelos rivais, não haverá garantia de minimização das verossimilhanças $P(M_k|\mathbf{X}, \Theta)$ em relação a $P(M_i|\mathbf{X}, \Theta)$, com $k \neq i$. Portanto, isto fará com que o algoritmo perca poder discriminativo.

Esta fraca discriminabilidade do critério de treinamento apresentado acima, aliada com a suposição do formato da *fdp* de emissão de símbolos obtida através de uma mistura de N gaussianas e a informação de contexto acústico inserida pelas primeira e segunda derivadas, motivaram as pesquisas envolvendo o uso de redes neurais artificiais em sistemas de reconhecimento de fala, assunto este que será abordado na próxima seção.

2.3 Redes Neurais Artificiais

2.3.1 Introdução

O uso de redes neurais artificiais em reconhecimento de fala é um campo de pesquisa atraente e promissor para inúmeros avanços científicos. Em meados dos anos de 1990 surgiram os modelos híbridos ANN+HMM [7, 8, 11], com o objetivo de eliminar as suposições lançadas quando se usa um sistema baseado puramente em HMM e utilizar as vantagens das duas técnicas onde, de um lado, os modelos ocultos de Markov lidam com as variabilidades temporais (matriz de transição de estados) e de outro as redes neurais, tratando das variabilidades acústicas (distribuição de probabilidades de emissão de símbolos). O estudo de redes neurais nesta dissertação ficou dedicado às redes MLP, que é a arquitetura mais utilizada para este fim.

2.3.2 Vantagens das Redes Neurais

Abaixo estão citadas algumas vantagens que motivaram o uso de redes neurais:

1) Possuem alto poder discriminativo em virtude do algoritmo de treinamento supervisionado e baseado na minimização, por exemplo, do erro médio quadrático.

2) São capazes de aprender através do mapeamento entrada-saída e generalizar conhecimento adquirido.

3) Tratam com facilidade informações contextuais. É possível inserir vetores acústicos à esquerda e à direita do vetor sob análise.

4) Possuem tolerância a ruídos.

5) Possuem grande tolerância a falhas. Isto se explica pelo fato de sua arquitetura permitir múltiplas conexões entre os neurônios.

2.3.3 Redes Multilayer Perceptron - MLP

Uma rede MLP é composta por um conjunto de neurônios amplamente conectados que formarão a camada de entrada (camada pela qual não é feito nenhum cálculo e portanto os neurônios são denominados de passivos), uma ou mais camadas ocultas e uma camada de saída [12]. A sua principal diferença em relação às redes de camada única (*Single Layer Perceptrons – SLP's*) é a capacidade de resolver problemas não-linearmente separáveis, ou seja, problemas onde as classes a serem reconhecidas possuem um certo grau de sobreposição. Em cada camada da rede o cálculo é feito de maneira paralela, de sorte que cada neurônio (a unidade básica de processamento de uma rede neural artificial) é responsável pela propagação do sinal em direção a camada de saída, sendo cada classe q_k representada por um neurônio. O objetivo é treinar a rede de tal forma que esta aprenda a formar relações corretas de entrada e saída e, com isto, adquirir um alto grau de discriminabilidade. Para tal propósito, é utilizado um algoritmo supervisionado *Error Back-Propagation – EBP* [12], que visa o ajuste dos pesos sinápticos, obedecendo um determinado critério de otimização. Este pode ser a minimização do erro quadrático médio (*Mean Square Error – MSE*):

$$E = \sum_{n=1}^{N_e} \sum_{k=1}^{K} [y_k(x_n) - d_k(x_n)]^2 \tag{2.9}$$

ou entropia relativa [11]:

$$E = \sum_{n=1}^{N_e} \sum_{k=1}^{K} d_k(x_n) \ln \frac{d_k(x_n)}{y_k(x_n)} \tag{2.10}$$

onde N_e é o número total de exemplos de treinamento, $y_k(x_n)$ é a *k-ésima* saída da rede neural para um dado conjunto de valores de entrada representado por x_n, $d_k(x_n)$ é o valor de saída desejada para a *k-ésima* saída e K é o número de saídas da rede.

A forma com que o cálculo é feito em cada neurônio está ilustrada na Figura 2.2.

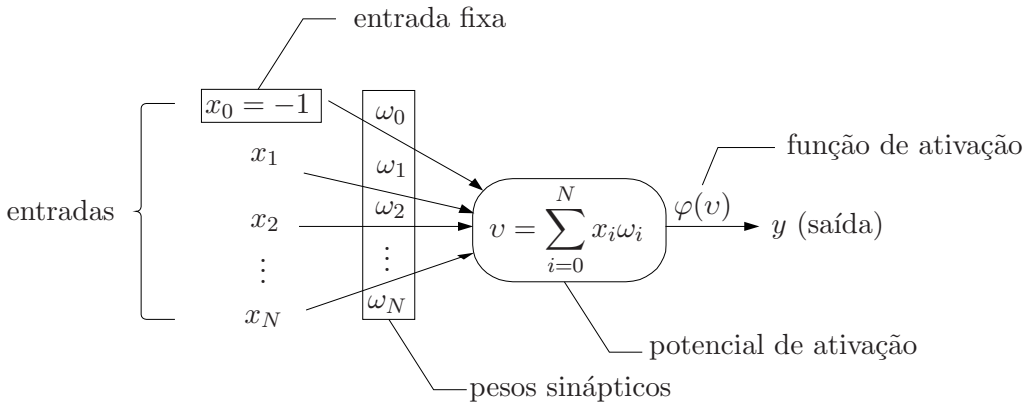

Figura 2.2: *O neurônio artificial e os elementos que o compõe.*

Todas as entradas de um neurônio possuem o seu respectivo peso sináptico ω_i $(i = 1, \ldots, N)$, que pode assumir valores positivos (sinapses excitatórias) ou negativos (sinapses inibitórias) que variam a medida que são atualizados pelo algoritmo de aprendizagem. O valor do potencial de ativação é conhecido através da soma de todas as entradas, multiplicadas pelas suas respectivas sinapses. Para limitar o valor de saída dentro de uma excursão fixa, o valor resultante deste potencial de ativação é passado por uma função que, para o caso de redes MLP, deve ser não-linear, diferenciável e não decrescente. A saída do neurônio será interligada com a entrada de um ou mais neurônios da próxima camada ou será a própria representação da classe de saída q_k. As funções de ativação mais utilizadas são as funções logística:

$$y = \varphi(\upsilon) = \frac{1}{1 + \exp(-a\upsilon)}\,, \quad a > 0 \tag{2.11}$$

e tangente hiperbólica:

$$y = \varphi(\upsilon) = a\tanh(b\upsilon)\,, \quad (a, b) > 0 \tag{2.12}$$

Seus gráficos são apresentados, respectivamente, nas Figuras 2.3 e 2.4, para constantes $a = b = 1$.

Desta forma, para que uma rede neural possua uma arquitetura do tipo MLP, os neurônios são interconectados como mostrado na Figura 2.5.

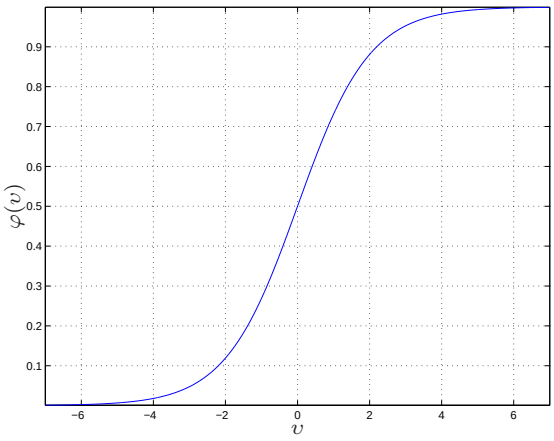

Figura 2.3: *Função logística*

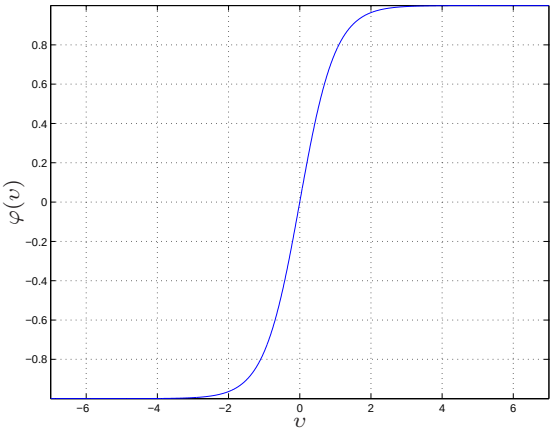

Figura 2.4: *Função tangente hiperbólica*

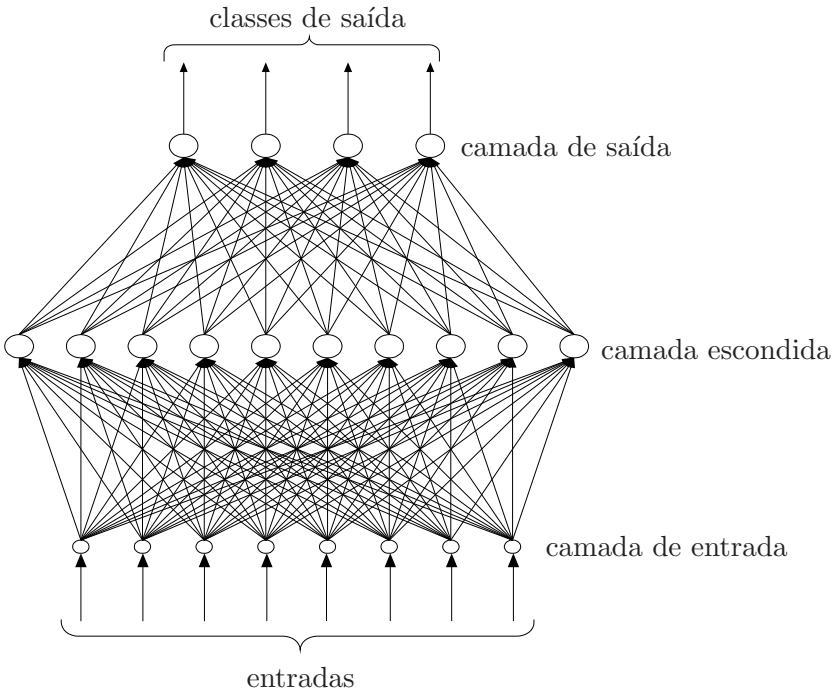

Figura 2.5: *Exemplo de uma rede MLP com oito neurônios na camada de entrada, dez neurônios na camada oculta e quatro neurônios na camada de saída. Notar a completa interconexão entre os neurônios das camadas de entrada e oculta e das camadas oculta e de saída.*

Para o neurônio j de uma determinada camada (oculta ou de saída) e x_n o vetor de entrada da rede no instante n, teremos então para as funções logística e tangente hiperbólica as seguintes notações:

$$y_j(x_n) = \varphi(v_j(x_n)) = \frac{1}{1 + \exp(-a v_j(x_n))}, \quad a > 0 \tag{2.13}$$

$$y_j(x_n) = \varphi(v_j(x_n)) = a\tanh(bv_j(x_n))\,, \quad (a,b) > 0 \tag{2.14}$$

2.3.4 Estimando probabilidades com MLP

A tarefa de uma rede neural artificial, em um sistema de reconhecimento de fala contínua baseado em modelos híbridos, é a de estimar a função densidade de probabilidade de emissão de símbolos $p(x_n|q_k)$. Esta abordagem estatística fornece um ponto de vista interessante no que diz respeito às saídas da rede neural, pois se atribui a cada saída $y_k(x_n)$ da rede a geração de uma probabilidade *a posteriori* da classe q_k dada a entrada x_n:

$$y_k(x_n) = P(q_k|x_n) = \frac{P(x_n|q_k)P(q_k)}{P(x_n)} \tag{2.15}$$

Desta forma, para classificar um vetor acústico x_n em, por exemplo, uma das duas classes q_1 e q_2, a decisão é para aquela classe com maior probabilidade de estar correta [13]. Ou seja, será q_1 a classe vencedora se $P(q_1|x_n) > P(q_2|x_n)$, que resulta no numerador do segundo termo da Equação (2.15) em

$$P(x_n|q_1)P(q_1) > P(x_n|q_2)P(q_2) \tag{2.16}$$

Isolando o termo $P(x_n|q_k)$ da Equação (2.15), chega-se às verossimilhanças de emissão de símbolos:

$$P(x_n|q_k) = \frac{P(q_k|x_n)P(x_n)}{P(q_k)} \tag{2.17}$$

Treinar uma rede neural a partir do critério MSE ou entropia relativa para classificação estatística nada mais é do que otimizar as estimativas das probabilidades *a posteriori* da classe dada a entrada que, neste caso, é a probabilidade da classe fonética q_k dado o vetor acústico de entrada x_n.

A interpretação estatística da saída de uma rede neural está fundamentada na seguinte demonstração, segundo [14]:

Seja uma rede neural treinada a partir do critério MSE. O conjunto de parâmetros livres (matrizes de pesos sinápticos) serão escolhidos para minimizar:

$$E_{MSE} = E\left\{\sum_{k=1}^{K}[y_k(x) - d_k]^2\right\} \tag{2.18}$$

sendo $E\{\cdot\}$ o operador esperança. Como os vetores acústicos assumem valores contínuos, a Equação (2.18) pode ser reescrita na forma:

$$E_{MSE} = \int \left\{ \sum_{k=1}^{K} [y_k(x) - d_k]^2 \right\} f(x)\, dx \tag{2.19}$$

Se escrevermos $f(x)$ em termos de probabilidade marginal: $f(x) = \sum_{j=1}^{K} f(x, q_j)$, daí teremos:

$$E_{MSE} = \int \sum_{j=1}^{K} \left\{ \sum_{k=1}^{K} [y_k(x) - d_k]^2 \right\} f(x, q_j)\, dx \tag{2.20}$$

A Equação (2.20) mostra a soma quadrática dos erros (somatório em k), com K erros aparecendo para cada par entrada–saída (somatório em j).

Mas $f(x, q_j) = P(q_j|x)\, f(x)$, logo:

$$E_{MSE} = \int \left[\sum_{j=1}^{K} \sum_{k=1}^{K} [y_k(x) - d_k]^2 \, P(q_j|x) \right] f(x)\, dx \tag{2.21}$$

e reescrevendo na forma:

$$E_{MSE} = E \left\{ \sum_{j=1}^{K} \sum_{k=1}^{K} [y_k(x) - d_k]^2 \, P(q_j|x) \right\} \tag{2.22}$$

Expandindo o termo entre colchetes de (2.22):

$$E_{MSE} = E \left\{ \sum_{j=1}^{K} \sum_{k=1}^{K} [y_k(x)^2 P(q_j|x) - 2y_k(x) d_k P(q_j|x) + d_k^2 P(q_j|x)] \right\} \tag{2.23}$$

Fazendo o somatório em j dos termos entre colchetes da Equação (2.23), tem-se:

$$\begin{aligned} E_{MSE} &= E \left\{ \sum_{k=1}^{K} \left[\sum_{j=1}^{K} y_k(x)^2 P(q_j|x) - \sum_{j=1}^{K} 2y_k(x) d_k P(q_j|x) + \sum_{j=1}^{K} d_k^2 P(q_j|x) \right] \right\} \\ &= E \left\{ \sum_{k=1}^{K} \left[y_k(x)^2 - 2y_k(x) \sum_{j=1}^{K} d_k P(q_j|x) + \sum_{j=1}^{K} d_k^2 P(q_j|x) \right] \right\} \\ &= E \left\{ \sum_{k=1}^{K} \left[y_k(x)^2 - 2y_k(x) E\{d_k|x\} + E\{d_k^2|x\} \right] \right\} \end{aligned} \tag{2.24}$$

onde $E\{d_k|x\}$ e $E\{d_k^2|x\}$ são as médias condicionais de d_k e d_k^2 respectivamente. Somando e subtraindo a Equação (2.24) por $\sum_{k=1}^{K} E^2\{d_k|x\}$, tem-se:

$$
\begin{aligned}
E_{MSE} &= E\left\{\sum_{k=1}^{K}\left[y_k(x)^2 - 2y_k(x)E\{d_k|x\} + E\{d_k^2|x\}\right]\right\} + \\
&\quad + \sum_{k=1}^{K} E^2\{d_k|x\} - \sum_{k=1}^{K} E^2\{d_k|x\} \\
&= E\left\{\sum_{k=1}^{K}\left[y_k(x)^2 - 2y_k(x)E\{d_k|x\} + E\{d_k^2|x\} + E^2\{d_k|x\} - E^2\{d_k|x\}\right]\right\}
\end{aligned}
\tag{2.25}
$$

Sabe-se que

$$
Var\{d_k|x\} = E\{d_k^2|x\} - E^2\{d_k|x\} \tag{2.26}
$$

e que

$$
[y_k(x) - E\{d_k|x\}]^2 = y_k^2(x) - 2y_k(x)E\{d_k|x\} + E^2\{d_k|x\} \tag{2.27}
$$

Substituindo (2.26) e (2.27) em (2.25):

$$
\begin{aligned}
E_{MSE} &= E\left\{\sum_{k=1}^{K}[y_k(x) - E\{d_k|x\}]^2 + Var\{d_k|x\}\right\} \\
&= E\left\{\sum_{k=1}^{K}[y_k(x) - E\{d_k|x\}]^2\right\} + E\left\{\sum_{k=1}^{M} Var\{d_k|x\}\right\}
\end{aligned}
\tag{2.28}
$$

Note que o primeiro termo da Equação (2.28) é o erro quadrático médio entre a saída da rede $y_k(x)$ e a média condicional das saídas desejadas $E\{d_k|x\}$. Como os parâmetros da rede são estimados para que a função de custo seja minimizada, em termos de erro médio quadrático, as saídas estimam as médias condicionais das saídas desejadas.

Para um problema *1 de M* [14], ou seja, a saída desejada k igual a 1 e as demais iguais a zero, d_k será igual a 1 se a entrada x pertencer a classe q_j e zero para as demais. Esta análise resulta no seguinte:

$$
E\{d_k|x\} = \sum_{j=1}^{K} d_k P(q_j|x) = \begin{cases} P(q_j|x) & \text{se } k = j \\ 0 & \text{se } k \neq j \end{cases}
$$

Portanto, para um problema *1 de M*, as saídas estimam probabilidades a

posteriori da classe dada a entrada e também minimizam o erro quadrático médio. Como esta prova baseia-se em critério de minimização, é válida para qualquer tipo de rede neural que utilize o critério MSE. Para o critério de entropia relativa, a dedução encontra-se em [14].

Assim, pode-se utilizar uma rede neural para estimar as probabilidades de emissão de símbolos e a vantagem desta metodologia é que cada probabilidade gerada pela rede é resultado da manipulação direta dos dados de entrada x_n, fornecendo neste caso uma *fdp* não paramétrica.

Como a rede fornece a probabilidade a posteriori $P(q_k|x_n)$, aplica-se a regra de Bayes como mostrada na Equação (2.17). Para o cálculo da probabilidade a priori da classe $P(q_k)$ será usada a Equação (2.29) pois a base de dados não é balanceada, ou seja, existem fonemas que ocorrem com mais freqüência que outros. Desta forma, a estimação de $P(q_k)$ será feita usando o conceito de probabilidade marginal:

$$P(q_k) = \sum_{j=1}^{N_e} P(q_k, x_j) = \sum_{j=1}^{N_e} P(q_k|x_j)P(x_j)\,,\ 1 \leq k \leq K \tag{2.29}$$

sendo N_e o número total de vetores que constituem os exemplos de treinamento. Assumindo que os vetores acústicos x_j são equiprováveis, temos:

$$P(q_k) \cong \frac{1}{N_e} \sum_{j=1}^{N_e} P(q_k|x_j)\,,\ 1 \leq k \leq K \tag{2.30}$$

Finalmente, as verossimilhanças de emissão de símbolos $P(x_n|q_k)$ são dadas pela equação:

$$p(x_n|q_k) \cong \frac{P(q_k|x_n)}{\frac{1}{N_e}\sum_{j=1}^{N_e} P(q_k|x_j)} P(x_n)\,,\ 1 \leq k \leq K \tag{2.31}$$

Geralmente a Equação (2.31) é escrita na forma:

$$\frac{P(x_n|q_k)}{P(x_n)} = \frac{P(q_k|x_n)}{\frac{1}{N_e}\sum_{j=1}^{N_e} P(q_k|x_j)} \tag{2.32}$$

chamada de verossimilhança de emissão de símbolos normalizada. Durante o reconhecimento, o fator $p(x_n)$ é uma constante para todas as classes e não afetará a classificação.

2.3.5 Utilizando informação contextual

O uso de informação contextual em uma rede MLP é extremamente simples, bastando apenas inserir um determinado número de vetores acústicos c antes e depois do vetor acústico x_n, formando uma janela de $2c + 1$ vetores de entrada. Com isto, o conjunto de entrada é denotado por $\mathbf{X}_{n-c}^{n+c} = \{x_{n-c}, \ldots, x_n, \ldots, x_{n+c}\}$.

Como a rede fornece uma relação entrada-saída e esta afirmação não se altera com o uso de informação contextual, os cálculos das probabilidades *a posteriori* da classe dada a entrada e das verossimilhanças de emissão de símbolos terão o mesmo enfoque dado pelas Equações (2.15) e (2.17), respectivamente:

$$y_k(\mathbf{X}_{n-c}^{n+c}) = P(q_k|\mathbf{X}_{n-c}^{n+c}) = \frac{P(\mathbf{X}_{n-c}^{n+c}|q_k)P(q_k)}{P(\mathbf{X}_{n-c}^{n+c})} \tag{2.33}$$

$$P(\mathbf{X}_{n-c}^{n+c}|q_k) = \frac{P(q_k|\mathbf{X}_{n-c}^{n+c})P(\mathbf{X}_{n-c}^{n+c})}{P(q_k)} \tag{2.34}$$

A Figura 2.6 mostra a rede neural para o uso da informação contextual:

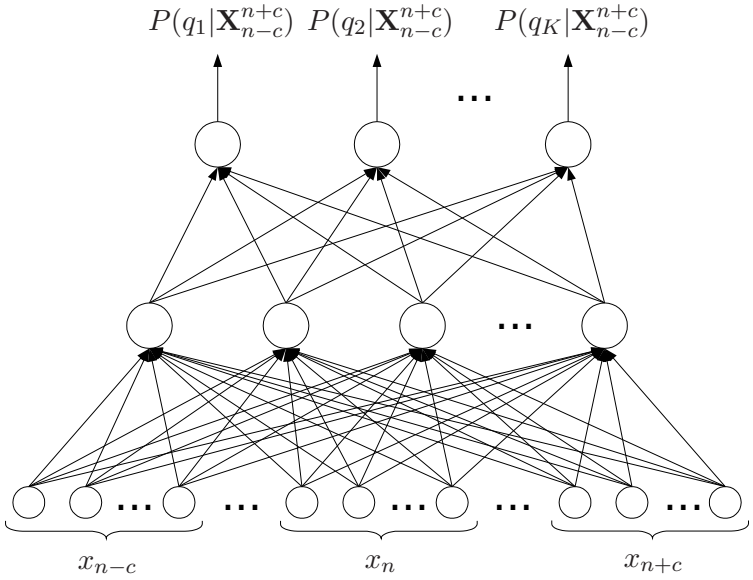

Figura 2.6: *Rede neural com uso de informação contextual* $\boldsymbol{X}_{n-c}^{n+c}$.

Capítulo 3

REMAP

3.1 Introdução

Como mencionado no capítulo anterior, em sistemas híbridos ANN+HMM, as redes neurais são responsáveis pela estimativa das verossimilhanças de emissão de símbolos $P(x_n|q_k)$ para cada sentença. Para isto, é necessário que seja executado um treinamento supervisionado (via *Error Back-Propagation*) para que a estimativa de cada verossimilhança seja a mais precisa possível.

Quando se iniciou o uso em fala contínua de redes neurais em conjunto com HMM's, a estimativa dos alvos (ou saídas desejadas) da rede para cada exemplo de treinamento era feita através do alinhamento forçado de Viterbi. O resultado obtido por este alinhamento constava de decisões abruptas quanto a saída que representa um determinado fonema (problema *1 de M*). Por exemplo, a rede neural ao analisar os quadros correspondentes ao fone "*/s/*", espera-se que a sua saída representativa esteja ativada com valor igual a um e as outras com valor igual a zero. Um exemplo deste alinhamento é mostrado na Figura 3.1.

Como o processo de produção da fala é mecânico, ocorre o efeito de coarticulação, que é o resultado da transição gradual da execução de uma seqüencia de fones, conforme mostrado na Figura 3.2. Pelo fato de ser necessário o treinamento da rede neural em termos das sub-unidades fonéticas, isto faz com que toda base de dados utilizada passe pelos processos de transcrição fonética e de segmentação – que pode ser manual ou automática. Esta incerteza quanto a fronteira de separação pode ser modelada por uma série de transições (suaves) ao longo do tempo, diferente do caso do problema *1 de M* [14], fazendo com que esta abordagem se mostre mais razoável que o alinhamento forçado de Viterbi, onde as decisões abruptas de fato não constroem um cenário apropriado para que as saídas da rede estimem de modo eficaz as verossimilhanças de emissão de estados.

Neste caso, para que os alvos da rede neural ilustrem bem as transições suaves

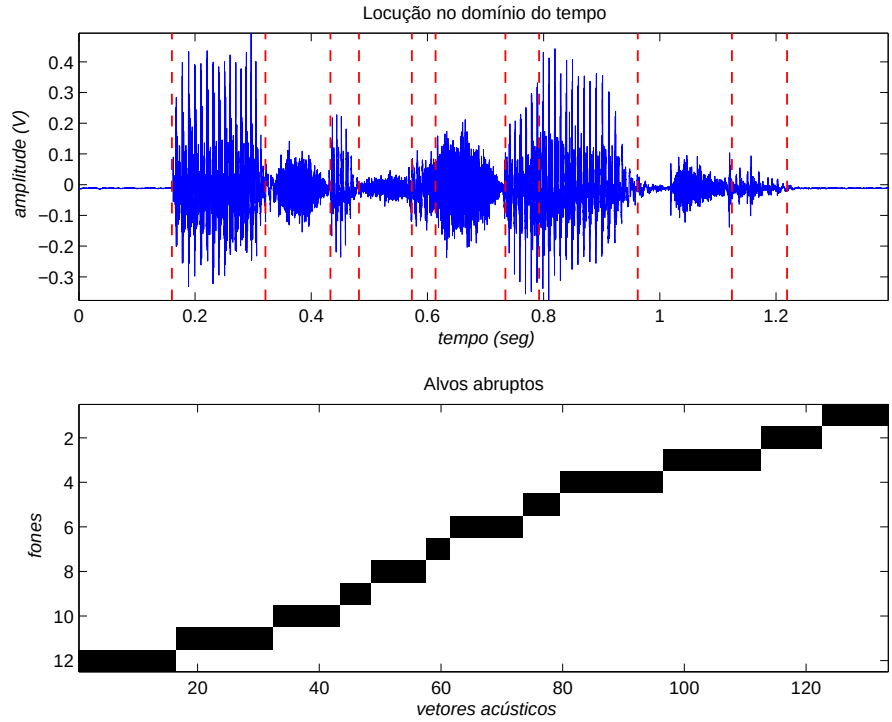

Figura 3.1: *Sentença "É suficiente", com os respectivos alvos abruptos da rede neural.*

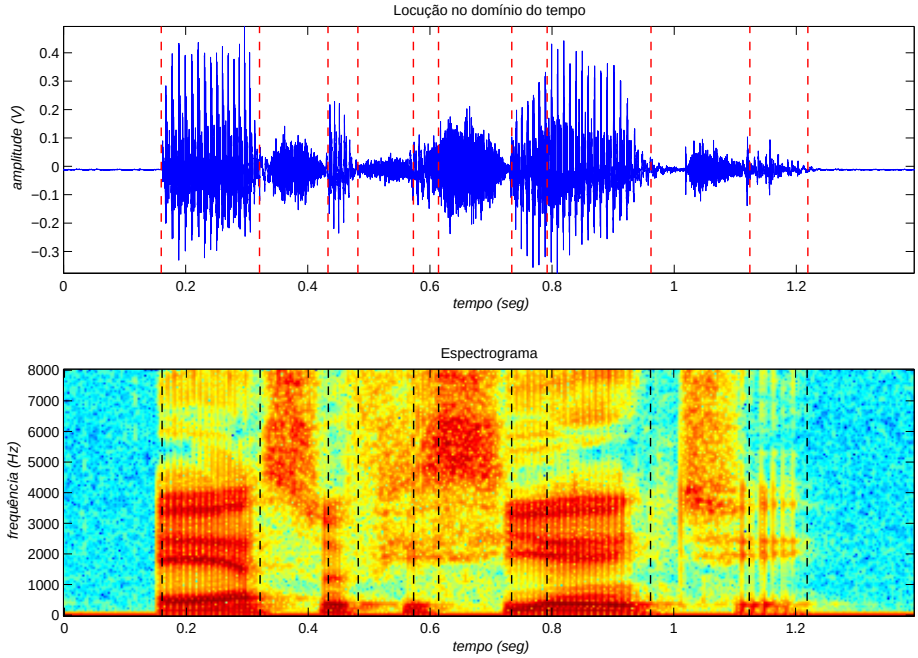

Figura 3.2: *Efeito de coarticulação resultante da inércia do aparelho fonador mostrado através do espectrograma. O treinamento em termos de modelo de fones faz com que cada locução seja previamente segmentada (linha tracejada).*

entre as unidades fonéticas, seus valores deveriam ser contínuos e tendo até mais de uma classe de saída ativa, como apresentado na Figura 3.3. Para que este objetivo seja alcançado, foi desenvolvido por Konig [2] uma metodologia onde o cálculo dos alvos se baseia na modelagem das transições fonéticas (o que foi chamado de reconhecimento de fala baseado em transições). No seu esquema original, a modelagem é feita por intermédio das recursões *forward* e *backward* equivalentes às convencionais do algoritmo *Baum-Welch*, em um sistema híbrido ANN+HMM discriminativo [3]. Como neste trabalho está sendo utilizado um sistema híbrido padrão, as estimativas dos alvos suaves serão feitas a partir da abordagem proposta em [15], que utiliza o algoritmo *Baum-Welch* convencional [6].

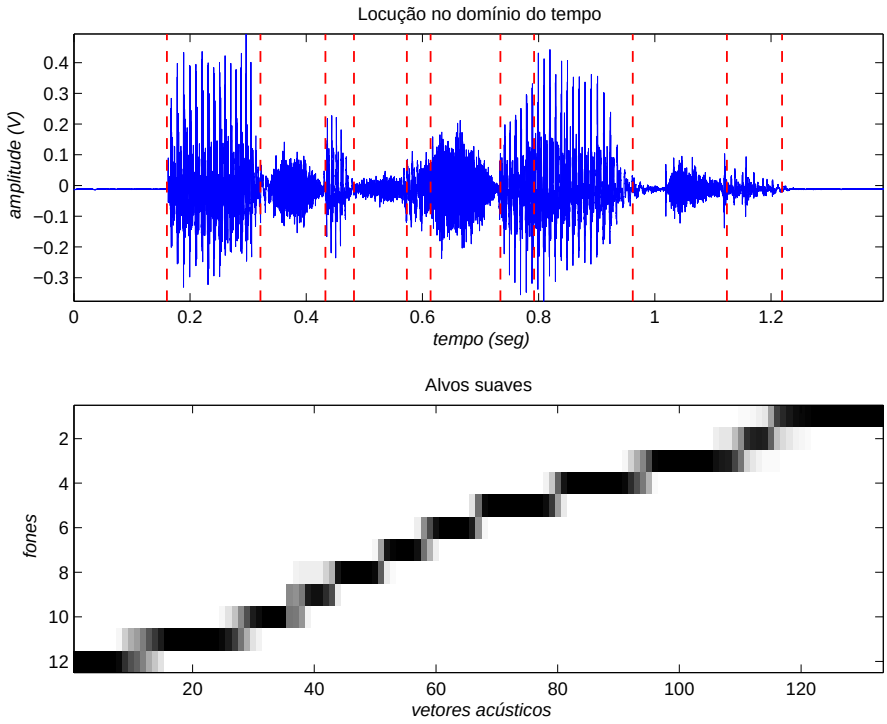

Figura 3.3: *Sentença "É suficiente", agora com alvos suaves. As transições suaves são resultantes de mais de uma classe de saída ativada.*

3.2 Formulações da abordagem do treinamento baseado em transições

O que motivou Konig em suas pesquisas foi a idéia de que a fronteira entre fonemas poderia ser modelada por uma janela onde estariam sendo representadas transições suaves, de tal maneira que o algoritmo de treinamento pudesse

aproveitar estas informações para a estimação de $p(x_n|q_k)$ a partir das probabilidades *a posteriori* $P(q_k|x_n)$ advindas da rede neural. Como estas fronteiras não são bem definidas em fala contínua, durante a estimativa dos alvos da rede neural deveria resultar em mais de uma classe de saída ativa.

3.2.1 Os alvos suaves

Os alvos suaves de uma rede neural são estimativas da probabilidade *a posteriori* global definida por $\gamma_n(k) = P(q_k^n|\mathbf{X}, M, \Theta)$ [9], ou seja, a probabilidade *a posteriori* do estado q_k ser visitado no instante n, dada a seqüencia de vetores acústicos $\mathbf{X}$ que é modelada pela cadeia de Markov M, que por sua vez está representada pelo conjunto Θ de parâmetros. Pode-se representar matematicamente $\gamma_n(k)$ por:

$$\begin{aligned} \gamma_n(k) &= P(q_k^n|\mathbf{X}, M, \Theta) && (3.1)\\ &= \frac{P(q_k^n, \mathbf{X}, M, \Theta)}{P(\mathbf{X}, M, \Theta)} && (3.2)\\ &= \frac{P(q_k^n, \mathbf{X}|M, \Theta)}{P(\mathbf{X}|M, \Theta)} && (3.3) \end{aligned}$$

Sabendo que $\mathbf{X} = \{x_1, \ldots, x_n, x_{n+1}, \ldots, x_N\} = \{\mathbf{X}_1^n, \mathbf{X}_{n+1}^N\}$ e aplicando apenas no numerador da expressão acima (o denominador será tratado mais adiante), tem-se:

$$\begin{aligned} \gamma_n(k) &= \frac{P(\mathbf{X}_1^n, \mathbf{X}_{n+1}^N, q_k^n|M, \Theta)}{P(\mathbf{X}|M, \Theta)} && (3.4)\\ &= \frac{P(\mathbf{X}_1^n, q_k^n|M, \Theta)P(\mathbf{X}_{n+1}^N|\mathbf{X}_1^n, q_k^n, M, \Theta)}{P(\mathbf{X}|M, \Theta)} && (3.5) \end{aligned}$$

Admitindo que os vetores acústicos são independentes:

$$\gamma_n(k) = \frac{P(\mathbf{X}_1^n, q_k^n, |M, \Theta)P(\mathbf{X}_{n+1}^N|q_k^n, M, \Theta)}{P(\mathbf{X}|M, \Theta)} \quad (3.6)$$

Na Equação (3.6), percebe-se que as verossimilhanças do numerador se assemelham às definições das probabilidades *forward* e *backward*, respectivamente:

$$\alpha_n(k) = P(\mathbf{X}_1^n, q_k^n, |M, \Theta) \quad (3.7)$$

$$\beta_n(k) = P(\mathbf{X}_{n+1}^N|q_k^n, M, \Theta) \quad (3.8)$$

Assim, conclui-se que os alvos suaves são estimados em função das recorrências

forward e *backward* convencionais:

$$\gamma_n(k) = \frac{\alpha_n(k)\beta_n(k)}{P(\mathbf{X}|M,\Theta)} \tag{3.9}$$

Extendendo esta dedução para $P(\mathbf{X}|M,\Theta)$, a verossimilhança do vetor acústico $\mathbf{X}$, dado o modelo M e o conjunto Θ de parâmetros:

$$\begin{aligned} p(\mathbf{X}|M,\Theta) &= p(\mathbf{X}_1^N|M,\Theta) && (3.10)\\ &= \sum_{k=1}^{L} p(\mathbf{X}_1^N, q_k^N|M,\Theta) && (3.11)\end{aligned}$$

sendo L o número de estados do modelo M. Pela definição de verossimilhanças *forward* dada em (3.7), vem:

$$P(\mathbf{X}|M,\Theta) = \sum_{k=1}^{L} \alpha_N(k) \tag{3.12}$$

A expressão dos alvos suaves em (3.9), resulta em:

$$\gamma_n(k) = \frac{\alpha_n(k)\beta_n(k)}{\sum_{k=1}^{L} \alpha_N(k)} \tag{3.13}$$

Mas o denominador da Equação (3.9) também pode ser escrito em função de $\alpha_n(k)$ e $\beta_n(k)$:

$$\begin{aligned} P(\mathbf{X}|M,\Theta) &= \sum_{k=1}^{L} P(\mathbf{X}, q_k^n|M,\Theta)\\ &= \sum_{k=1}^{L} P(\mathbf{X}_1^n, \mathbf{X}_{n+1}^N, q_k^n|M,\Theta) && (3.14)\end{aligned}$$

O termo do somatório em k da Equação (3.14) é igual ao numerador da Equação (3.4), desta forma:

$$P(\mathbf{X}|M,\Theta) = \sum_{k=1}^{L} P(\mathbf{X}_1^n, \mathbf{X}_{n+1}^N, q_k^n|M,\Theta) = \sum_{k=1}^{L} \alpha_n(k)\beta_n(k) \tag{3.15}$$

Reescrendo o denominador da Equação (3.13) em termos de $\alpha_n(k)$ e $\beta_n(k)$:

$$\gamma_n(k) = \frac{\alpha_n(k)\beta_n(k)}{\sum_{k=1}^{L} \alpha_n(k)\beta_n(k)} \tag{3.16}$$

Toda a dedução da expressão dos alvos suaves foi feita sem levar em conta a questão de implementação para o cálculo das verossimilhanças *forward* e *backward*. Isto ocorre sempre que o cálculo é executado para longas seqüencias de vetores acústicos ($N > 100$), onde inevitavelmente surgirão problemas de precisão numérica por tratar-se de cálculos de probabilidade, convergindo exponencialmente a zero e ultrapassando a precisão da máquina. A solução seria utilizar a normalização de $\alpha_n(k)$ e $\beta_n(k)$ para evitar o problema de *underflow*.

Neste caso, para $\alpha_n(k)$ e $\beta_n(k)$ normalizados:

$$\hat{\alpha}_n(k) = \frac{\alpha_n(k)}{\sum_{k=1}^{L} \alpha_n(k)} = c_n \cdot \alpha_n(k) \tag{3.17}$$

$$\hat{\beta}_n(k) = \frac{\beta_n(k)}{\sum_{k=1}^{L} \alpha_n(k)} = c_n \cdot \beta_n(k) \tag{3.18}$$

sendo c_n o fator de normalização, independente do estado k, para o instante $n \in [1, N]$. Isolando os termos não normalizados:

$$\alpha_n(k) = \frac{\hat{\alpha}_n(k)}{c_n} \tag{3.19}$$

$$\beta_n(k) = \frac{\hat{\beta}_n(k)}{c_n} \tag{3.20}$$

Substituindo as Equações (3.19) e (3.20) em (3.16), vem:

$$\gamma_n(k) = \frac{\alpha_n(k)\beta_n(k)}{\sum_{k=1}^{L} \alpha_n(k)\beta_n(k)} \tag{3.21}$$

$$= \frac{\frac{\hat{\alpha}_n(k)}{c_n} \cdot \frac{\hat{\beta}_n(k)}{c_n}}{\sum_{k=1}^{L} \frac{\hat{\alpha}_n(k)}{c_n} \cdot \frac{\hat{\beta}_n(k)}{c_n}} \tag{3.22}$$

$$\gamma_n(k) = \frac{\frac{1}{c_n^2} \cdot \hat{\alpha}_n(k)\hat{\beta}_n(k)}{\frac{1}{c_n^2}\sum_{k=1}^{L} \hat{\alpha}_n(k)\hat{\beta}_n(k)} \tag{3.23}$$

$$= \frac{\hat{\alpha}_n(k)\hat{\beta}_n(k)}{\sum_{k=1}^{L} \hat{\alpha}_n(k)\hat{\beta}_n(k)} \tag{3.24}$$

Portanto:

$$\gamma_n(k) = \frac{\hat{\alpha}_n(k)\hat{\beta}_n(k)}{\sum_{k=1}^{L} \hat{\alpha}_n(k)\hat{\beta}_n(k)} \tag{3.25}$$

Na implementação computacional, as estimativas dos alvos suaves de cada locução foram realizadas utilizando a Equação (3.25).

3.3 O algoritmo

De acordo com a formulação desenvolvida na seção anterior, serão listados os passos do algoritmo para a reestimação dos parâmetros de um modelo híbrido ANN+HMM padrão:

1. Inicializar uma rede neural com os exemplos de treinamento centrados nas marcas de segmentação das unidades fonéticas. Em seguida, calcular as probabilidades *a priori* iniciais das classes de saída $P(q_k)$, de acordo com a Equação (2.30) e as probabilidades de transição $P(q_k^n|q_j^{n-1})$ conforme as equações abaixo:

$$a_{jj} = \frac{D_j - s}{D_j} \tag{3.26}$$

$$a_{jk} = 1 - a_{jj} \tag{3.27}$$

sendo D_j a duração média (em milissegundos) do fonema associado ao estado q_j e s é a sobreposição entre janelas adjacentes.

2. Calcular os alvos suaves $\gamma_n(k) = P(q_k^n|\mathbf{X}, M, \Theta)$ de cada uma das sentenças de treinamento. Após estimados todos os alvos, reestimar as probabilidades

de transição para a iteração seguinte conforme Equação (3.36), a ser vista adiante.

3. Treinar a rede neural com os alvos suaves estimados no *passo 2*, a partir do critério MSE ou entropia relativa, reestimando assim parâmetros da rede (matriz de pesos sinápticos). Com os mesmos alvos suaves, a rede será treinada durante n_reest vezes.

4. Reestimar as probabilidades a priori das classes $P(q_k)$ de saída da rede.

5. Se o sistema não convergiu, ou alcançou o número máximo de iterações (n_iter), voltar para o *passo 2*.

Note que os passos do algoritmo se resumem em duas etapas:

a. Uma etapa de maximização, correspondente ao *passo 3*, onde os novos pesos sinápticos da rede da próxima iteração são encontrados para a minimização do critério MSE ou entropia relativa.

b. Uma etapa de estimação, correspondente aos *passos 1* (estimação das probabilidades iniciais $P(q_k)$ e $P(q_k^n|q_j^{n-1})$, *2* e *4*.

Todas as etapas do algoritmo REMAP estão ilustradas no fluxograma da Figura 3.4.

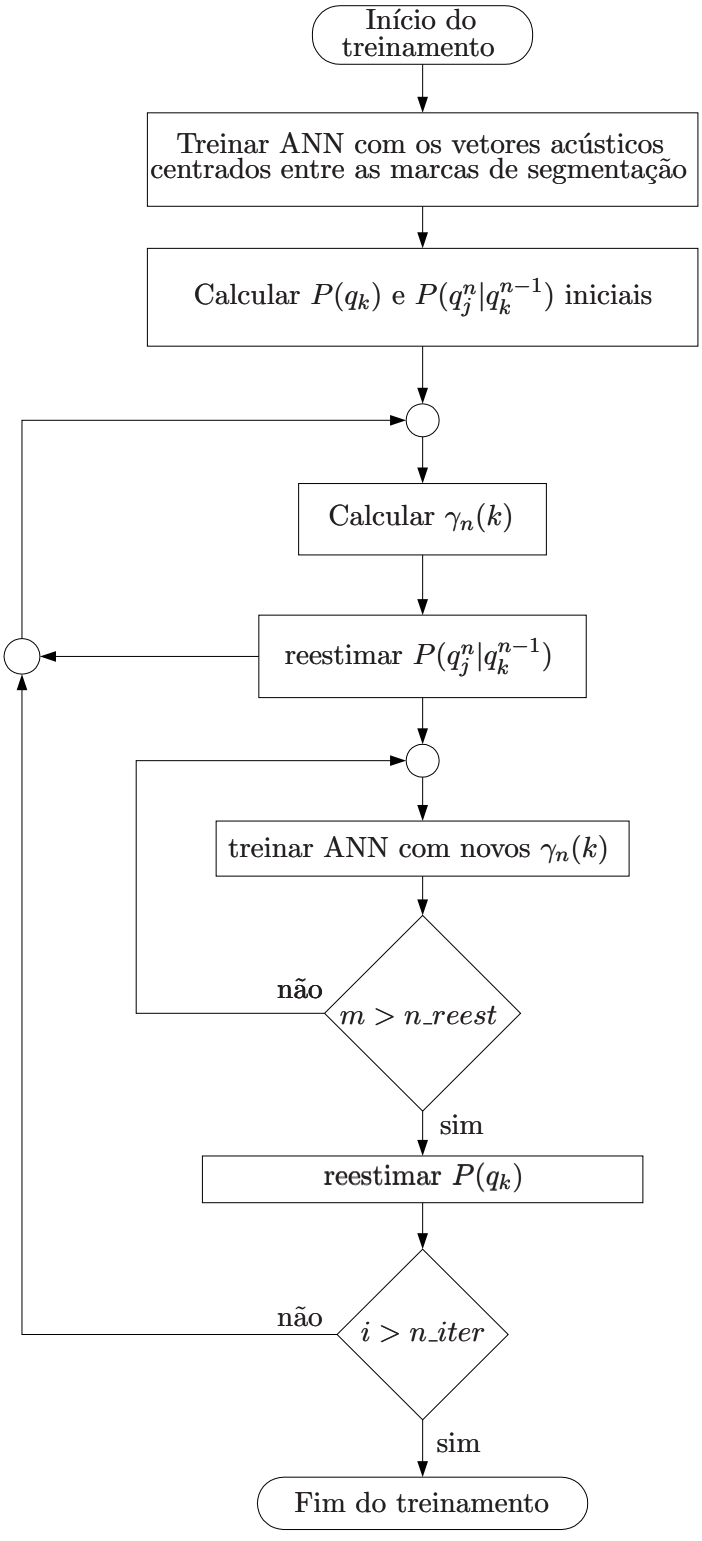

Figura 3.4: *Fluxograma do algoritmo REMAP.*

3.3.1 Reestimação das probabilidades de transição

De acordo com o *passo 2* do algoritmo REMAP, após a estimativa dos alvos suaves da rede ocorre a reestimação das probabilidades de transição para a próxima iteração. Esta corresponde ao terceiro problema básico para os HMM's [6], ou seja, ajustar os parâmetros do modelo para maximizar $P(\mathbf{X}|M,\Theta)$. Como a rede neural cuida do ajuste das verossimilhanças de emissão de cada estado do modelo, cabe ao HMM a reestimação de $P(q_k^n|q_j^{n-1})$. Define-se a seguinte probabilidade:

$$\xi_n^{(i)}(j,k) = P(q_j^n, q_k^{n+1}|\mathbf{X}_{M_i}, M_i, \Theta) \tag{3.28}$$

onde $\xi_n^{(i)}(j,k)$ é a probabilidade de estar no estado q_j, no instante n, e no estado q_k no instante $n+1$, dada a *i-ésima* seqüencia de vetores acústicos correspondente ao modelo oculto de Markov M_i e o conjunto Θ que engloba os parâmetros do HMM M_i e da Rede Neural. Manipulando (3.28), chega-se a seguinte expressão:

$$\xi_n^{(i)}(j,k) = \frac{P(q_j^n, q_k^{n+1}, \mathbf{X}_{M_i}|M_i,\Theta)}{P(\mathbf{X}_{M_i}|M_i,\Theta)} \tag{3.29}$$

$$= \frac{\alpha_n^{(i)}(j)\, a_{jk}\, b_k(x_{n+1}^{(i)})\, \beta_{n+1}^{(i)}(k)}{\sum_{j=1}^{L} \alpha_n^{(i)}(j)\beta_n^{(i)}(j)} \tag{3.30}$$

Tanto para os alvos suaves $\gamma_n^{(i)}(k) = P(q_k^n|\mathbf{X}_{M_i}, M_i, \Theta)$ quanto para a probabilidade $\xi_n^{(i)}(j,k) = P(q_j^n, q_k^{n+1}|\mathbf{X}_{M_i}, M_i, \Theta)$ é feita a seguinte interpretação:

$$\sum_{n=1}^{N-1} \gamma_n^{(i)}(k) = \text{número esperado de transições a partir do estado } q_k \tag{3.31}$$

$$\sum_{n=1}^{N-1} \xi_n^{(i)}(j,k) = \text{número esperado de transições de } q_j \text{ para } q_k \tag{3.32}$$

Usando as Equações (3.31) e (3.32), a reestimação das probabilidades de transição pode ser dada pela seguinte relação:

$$a_{jk} = \frac{\text{número esperado de transições de } q_j \text{ para } q_k}{\text{número esperado de transições a partir do estado } q_k} \tag{3.33}$$

ou seja:

$$a_{jk} = \frac{\sum_{n=1}^{N-1} \xi_n(j,k)}{\sum_{n=1}^{N-1} \gamma_n(k)} = \frac{\sum_{n=1}^{N-1} \alpha_n(j) a_{jk} b_k(x_{n+1}) \beta_{n+1}(k)}{\sum_{n=1}^{N-1} \alpha_n(k) \beta_n(k)} \tag{3.34}$$

Para múltiplas seqüencias de vetores acústicos:

$$a_{jk} = \frac{\sum_{i=1}^{I_M} \sum_{n=1}^{N-1} \xi_n^{(i)}(j,k)}{\sum_{i=1}^{I_M} \sum_{n=1}^{N-1} \gamma_n^{(i)}(k)} = \frac{\sum_{i=1}^{I_M} \sum_{n=1}^{N-1} \alpha_n^{(i)}(j) a_{jk} b_k(x_{n+1}^{(i)}) \beta_{n+1}^{(i)}(k)}{\sum_{i=1}^{I_M} \sum_{n=1}^{N-1} \alpha_n^{(i)}(k) \beta_n^{(i)}(k)} \tag{3.35}$$

sendo I_M o número total de sentenças de treinamento. Para as probabilidades $\alpha_n(k)$ e $\beta_n(k)$ normalizadas:

$$a_{jk} = \frac{\sum_{i=1}^{I_M} \sum_{n=1}^{N-1} \xi_n^{(i)}(j,k)}{\sum_{i=1}^{I_M} \sum_{n=1}^{N-1} \gamma_n^{(i)}(k)} = \frac{\sum_{i=1}^{I_M} \sum_{n=1}^{N-1} \hat{\alpha}_n^{(i)}(j) a_{jk} b_k(x_{n+1}^{(i)}) \hat{\beta}_{n+1}^{(i)}(k)}{\sum_{i=1}^{I_M} \sum_{n=1}^{N-1} \hat{\alpha}_n^{(i)}(k) \hat{\beta}_n^{(i)}(k) / c_n^{(i)}} \tag{3.36}$$

Após apresentado todo embasamento teórico do sistema híbrido ANN+HMM estudado, o próximo capítulo é dedicado a uma breve descrição dos blocos utilizados para montar todo sistema.

Capítulo 4

O Sistema Implementado

Um sistema de reconhecimento de fala contínua é constituído basicamente pelos blocos funcionais mostrados nas Figuras 4.1 e 4.2, que representam os blocos de treinamento e reconhecimento, respectivamente.

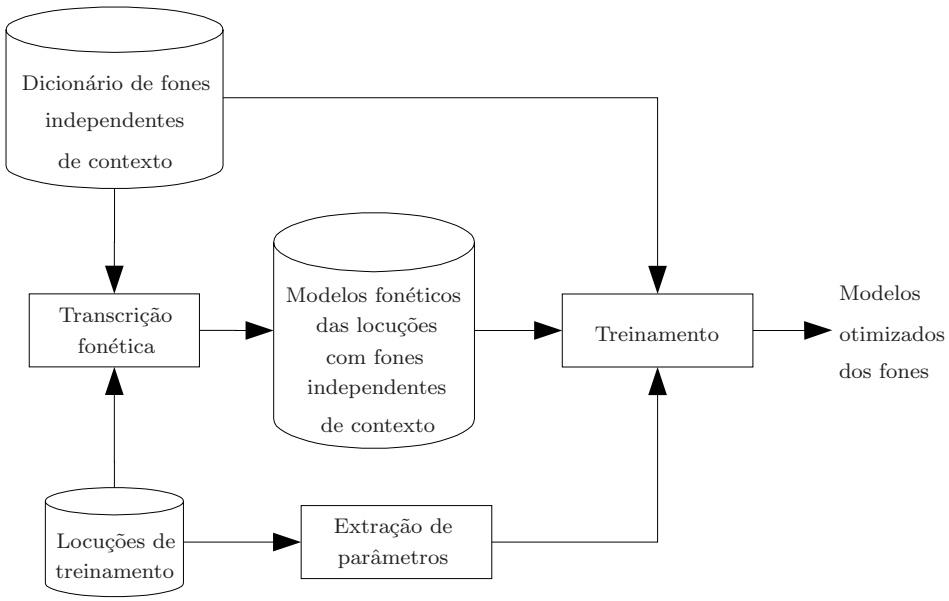

Figura 4.1: *Diagrama em blocos do módulo de treinamento de um sistema de reconhecimento de fala contínua.*

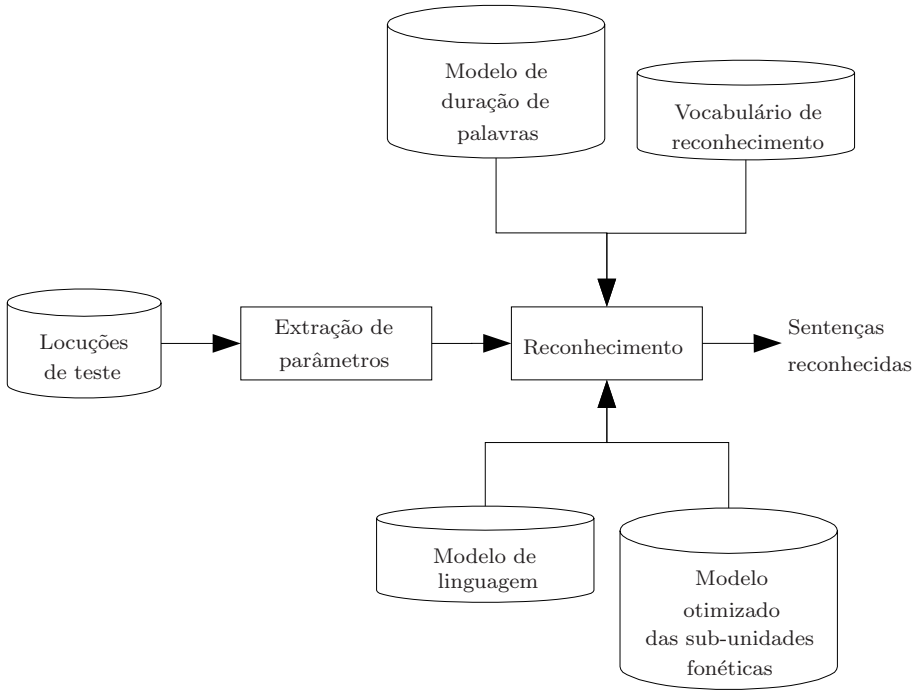

Figura 4.2: *Diagrama em blocos do módulo de reconhecimento de um sistema de reconhecimento de fala contínua.*

Para o sistema híbrido ANN+HMM estudado nesta dissertação, para cada bloco operacional foram usadas as abordagens que serão apresentadas nas Seções a seguir.

4.1 Extrator de parâmetros

Foi utilizado um extrator de coeficientes mel-cepstrais (*Mel Frequency Cepstrum Coefficients – MFCC*) de ordem 12, com quadros de 20 ms, tomados a cada 10 ms. Antes da parametrização, o sinal foi submetido a um filtro de pré-ênfase $H(z) = 1 - 0,95z^{-1}$ e janelado através de uma janela de Hamming.

Como a maior parte destes coeficientes ocorrem em uma faixa de grande amplitude (entre -20 a +20), ao serem inseridos na rede neural podem fazer com que ocorra a saturação dos neurônios de saída. Com isto é necessária a normalização da amplitude destes coeficientes, a fim de reduzir a faixa dinâmica. Para isto, são executados os seguintes procedimentos para tal redução:

1. Retirada da média

2. Normalização da variância

Estes passos também podem ser descritos pela expressão:

$$X_{norm} = \frac{X - \mu}{\sigma_T} \tag{4.1}$$

onde:

- X é o vetor de parâmetros original
- X_{norm} é o vetor acústico normalizado
- μ é o vetor média da locução
- σ_T é o desvio padrão de todas as componentes de todos os vetores acústicos

Com isto, garante-se que aproximadamente 95% destes coeficientes se concentram dentro do intervalo -1 e +1, seguindo uma distribuição de média nula e desvio padrão 0,49. O histograma levantado antes e depois da normalização está ilustrado na Figura 4.3.

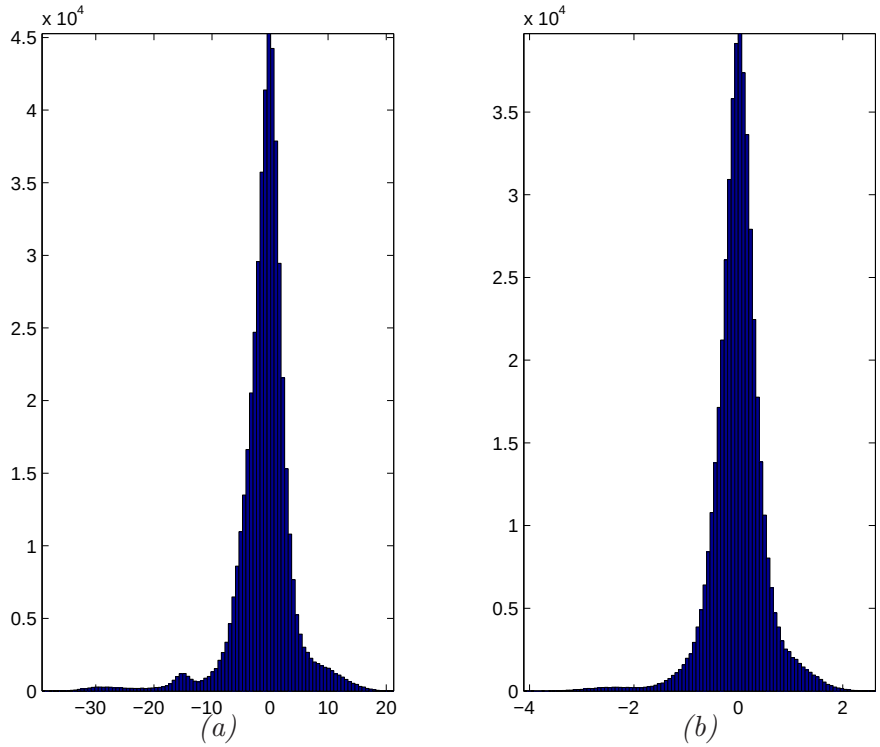

Figura 4.3: *Histograma dos coeficientes mel-cepstrais. (a) antes da normalização. (b) após a normalização, obtendo média nula e $\sigma = 0{,}49$.*

4.2 Treinamento das sub-unidades fonéticas

4.2.1 Modelos das sub-unidades fonéticas

Em sistemas de reconhecimento de fala contínua é comum o treinamento no contexto das sub-unidades fonéticas (p.ex. sílabas, fones, trifones) no intuito de aumentar a treinabilidade destes modelos. Mas para que o sistema possa ser treinado nesta concepção é necessário fazer a transcrição *ortográfico-fonética* de toda base de dados, fazendo com que cada locução seja decodificada em uma concatenação de unidades básicas.

Mas antes de ser feita a transcrição e a segmentação é necessário fazer uma escolha prévia do tipo de unidades fundamentais a serem utilizadas. Esta escolha é diretamente influenciada pelo tamanho do vocabulário.

Para o caso de sistemas de vocabulário pequeno é comum a utilização de palavras como unidade fundamental. Sendo assim, é necessário um grande número de exemplos de cada palavra para que o sistema seja treinado adequadamente. Entretanto, o treinamento se torna inviável quando o vocabulário se torna muito extenso.

Para vocabulários grandes, sub-unidades fonéticas como fones, difones, trifones, sílabas, semissílabas, etc. se tornam úteis pois o número de unidades fundamentais a serem treinadas reduz, em relação à grande quantidade de palavras no treinamento de um sistema de reconhecimento de fala para grandes vocabulários. Desta forma, surge a demanda de uma grande quantidade de exemplos de cada sub-unidade.

Mas para que se faça a escolha do tipo de sub-unidade, deve-se estar atento à estes dois critérios, a saber:

- **Consistência (ou discriminabilidade)**: quando existem exemplos diferentes de uma mesma unidade e que possuem características similares.
- **Treinabilidade**: para que seja criado um modelo robusto, devem existir exemplos de treinamento suficientes para cada unidade.

Sub-unidades tais como difones, trifones, sílabas e semissílabas são consistentes mas difíceis de treinar, enquanto que fones são treináveis, porém inconsistentes.

A unidade fundamental escolhida foi o fone independente de contexto que, para o português brasileiro, são identificados aproximadamente 40 fones [16]. Mas alguns destes foram unificados neste trabalho [5], perfazendo um total de 36 ($K = 36$).

Estas fones serão representadas no HMM por apenas um estado, com auto-transição e uma transição para o fone seguinte (resultando em uma correspondência um para um), em que concatenadas formarão os modelos de palavras e que por sua vez formarão os modelos de sentenças.

A seguir são mostrados na Tabela 4.1 os 36 fones utilizados nesta dissertação com seus respectivos exemplos, número de ocorrência em toda base de dados e duração média. Na Figura 4.4 apresenta uma cadeia de Markov que é resultado da concatenação dos fones da frase *"Diariamente"*:

Tabela 4.1: *Lista dos fones utilizados, com número de ocorrências e duração média.*

Fone	Exemplo	Número de ocorrências	Duração média (ms)
#	pausa	238	315,886
A	efic ***á*** cia	147	63,2517
a	medid ***a***	346	91,3699
$\tilde{\alpha}$	***an*** terior	95	86,6947
e	f ***e*** charam	263	75,4068
ɛ	caf ***é***	59	128,9322
$\tilde{\text{e}}$	sufici ***en*** te	114	117,0964
i	anal ***i*** a	431	57,2761
$\tilde{\text{i}}$n	v ***in*** te	90	105,2111
o	b ***o*** lsa	113	85,1062
ɔ	p ***o*** sso	20	157,9
$\tilde{\text{o}}$	c ***on*** ta	79	101,7215
u	ins ***u*** ficiente	508	56,7264
$\tilde{\text{u}}$	f ***un*** cionário	22	98,9545
b	***B*** arroso	52	64,3654
d	***d*** escontos	180	50,0556
ʤ	cré ***d*** ito	60	56,4833
f	di ***f*** erentes	53	97,6038
g	***g*** overno	32	53,0625
ʒ	***j*** uros	11	73,0901
k	***c*** ondomínio	187	84,0214
l	***l*** ocalizado	58	44,931
ʎ	traba ***lh*** o	6	57,8333
m	***m*** onetário	116	62,569
n	funcio ***n*** ário	120	44,7167
ɲ	ju ***nh*** o	2	91,5
p	o ***p*** comunidade	108	79,6852
r	sofre ***r*** á	208	36,25
r̄	co ***rr*** vigias	19	64,7368
R	cu ***r*** va	50	58,14
s	pa ***ss*** ará	338	113,497
t	***t*** elefônica	226	77,3628
ʧ	se ***t*** e	79	96,2278
v	de ***v*** ido	71	62,2113
ʃ	***ch*** amada	19	104,0526
z	de ***z*** embro	155	62,2322

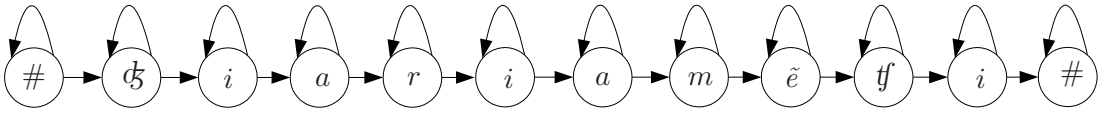

Figura 4.4: *HMM da frase "Diariamente".*

4.2.2 Segmentação fonética

Para que o sistema estudado otimize os modelos fonéticos durante o treinamento, são necessárias as informações de transcrição, junto com suas respectivas informações de duração (em milissegundos). Este último procedimento é o que chamamos de *segmentação*, podendo ser feito através de duas maneiras distintas:

- Segmentação manual – preferencialmente feita por um especialista (*foneticista*) onde este define as fronteiras de cada sub-unidade fonética. Este processo possui grande precisão porém pode ser demorado em função do tamanho da base de dados a ser manipulada.
- Segmentação automática – é executado por um programa de computador cujo algoritmo é baseado, por exemplo, no algoritmo de Viterbi ou em redes neurais. Este processo é mais rápido porém não tem a mesma precisão da segmentação manual.

A Tabela 4.2 apresenta a transcrição ortográfico-fonética da frase "*Diariamente*", que contém o resultado da segmentação fonética obtida manualmente, assim como a sua transcrição em fones:

Tabela 4.2: *Transcrição ortográfico-fonética da frase "Diariamente". O símbolo "#" indica silêncio no início e no final da frase.*

Transcrição:	#	ʤ	i	a	r	i	a	m	ẽ	ʧ	i	#
Duração (ms):	377	79	127	109	20	44	39	66	161	161	50	200

4.2.3 Modelo Híbrido ANN+HMM

Para o treinamento das variabilidades acústicas dos fonemas foi utilizada uma Rede Neural Artificial do tipo *Multilayer Perceptron* com as seguintes características:

- **Número de entradas**: seu valor foi definido em função da média dos valores das durações médias dos fonemas da Tabela 4.1. Assim, chegou-se ao valor de $c = 4$ (informação contextual de 4 vetores acústicos anteriores e

posteriores a $\mathbf{X}_n$), resultando em um total de $2c + 1 = 9$ quadros presentes na entrada. Dado que os vetores acústicos são de dimensão 12, tem-se 108 unidades de entrada. Também foram executados testes para o valor de $c = 3$ (84 entradas).

- **Número de neurônios da camada escondida**: foi testado o grau de generalização que a rede adquire para 100 neurônios.

- **Número de neurônios da camada de saída**: com a correspondência *um para um* entre fonemas e classes de saída da ANN, foi obtido um total de 36 neurônios de saída.

- **Função de ativação**: logística, dada pela Equação (2.13), aplicada às camadas escondida e de saída. Para que os neurônios de saída pudessem ter significado estatístico (somatório das saídas igual a 1), foi feita a normalização $\bar{y_j} = \dfrac{y_j}{\sum_{k=1}^{K} y_k}$.

- **Pesos sinápticos**: os pesos sinápticos da rede neural foram inicializados com valores aleatórios dentro do intervalo $[-0,1; +0,1]$, seguindo uma distribuição uniforme.

- **Algoritmo de aprendizagem**: *Error Back-Propagation* [12], sem uso de momento e com atualização instantânea dos pesos sinápticos. Foi investigado o desempenho do sistema com passos de aprendizagem dentro do intervalo $[0,1; 1,0]$ com intervalos de 0,1.

- **Critérios de parada (Pré-REMAP)**: foram feitas 1000 épocas de treinamento na etapa Pré-REMAP utilizando a verificação do número de erros de treinamento e o cálculo de *distorção* (o mesmo usado no treinamento de sistema de reconhecimento baseados puramente em HMM) como critérios de parada. A expressão da distorção da *n-ésima* época é dada abaixo:

$$dist(n) = \frac{E(n) - E(n-1)}{E(n)} \tag{4.2}$$

onde $E(n)$ é o erro médio quadrático da época n e $E(n-1)$ é o erro médio quadrático da época $n-1$. O treinamento é interrompido quando uma das condições forem satisfeitas:

  - O número máximo de épocas for alcançado;
  - O número de erros de treinamento for nulo;
  - O valor da distorção, calculado pela Equação 4.2, atingir um limiar ε pré-estabelecido.

Na etapa REMAP o treinamento da rede neural, as reestimações das probabilidades de transição de estados e das probabilidades *a priori* das classes seguem os passos apresentados na seção 3.3.

4.3 Reconhecimento

O algoritmo de busca utilizado foi o *Level Building* com 50 níveis de busca e com critério de parada automática proposto em [5].

O número de níveis indica a quantidade máxima de palavras que uma frase reconhecida pode conter. Neste trabalho, a base de dados é composta de frases que variam de 1 a 47 palavras. Portanto, baseado no número de palavras existentes na maior frase e nas suas possíveis pausas, adotou-se os 50 níveis.

Mas para que o algoritmo de busca possa executar o reconhecimento das frases, é necessária a geração de um vocabulário, contendo o universo de palavras que poderão ser reconhecidas. Desta forma, o vocabulário foi definido a partir das frases que compõem a base de dados (100 frases dependentes de locutor). Adotou-se que cada uma das palavras foi pronunciada da mesma maneira, resultando em 319 palavras distintas.

O arquivo de vocabulário é composto por uma lista dos fones utilizados para a transcrição das palavras e pela descrição das palavras, sua transcrição fonética, a média e o desvio padrão de suas durações. Um exemplo de vocabulário utilizado é apresentado abaixo:

Neste arquivo são utilizadas algumas palavras reservadas para a indicação do início e final dos blocos de listagem dos fones e das palavras do vocabulário. Desta forma, o programa de reconhecimento será capaz de extrair as informações necessárias para o seu funcionamento. A saber:

- ***fonemas**: indica o início da listagem dos fones utilizados na transcrição fonética das palavras.
- ***vocab**: indica o início da listagem das palavras do vocabulário.
- ***fim**: indica o final das listagem dos fones e das palavras do vocabulário.

Na definição das palavras do vocabulário, conforme apresentado na Figura 4.5, segue a seguinte estrutura: *palavra / transcrição fonética / média da duração (ms) / desvio padrão da duração.*

```
*fonemas
#
a
A
an
...
v
x
z
*fim
*vocab
, / # / 296.41908 / 144.87533
A / a / 70.8065 / 38.4451
ACEITARÃO / a s e t a r an u / 552.2500 / 53.8973
ACORDO / a k o r d u / 433.6667 / 35.5715
...
NECESSÁRIO / n e s e s A r i u / 647 / 215.6667
NEM / n e i / 170 / 56.6667
NO / n u / 86.5455 / 15.3842
...
VISA / v i s a / 292 / 97.3333
VOCÊ / v o s e / 393 / 131
VÔO / v o / 269 / 89.6667
*fim
```

Figura 4.5: *Exemplo de arquivo do vocabulário.*

4.3.1 Modelo de duração de palavras e modelo de linguagem

Com o intuito de aumentar o desempenho do bloco de reconhecimento foram usados os seguintes métodos:

1) Modelo de duração de palavras

O reconhecedor em si não estabelece nenhuma limitação temporal a uma palavra na qual o mesmo pretende decodificar. Como resultado, alguns modelos podem ser "comprimidos" (por exemplo, ao invés de reconhecer o modelo da palavra "*mas*", pode-se reconhecer o modelo de "*as*") ou "expandidos" (ao invés de "*convênio*", pode-se reconhecer "*com vôo em mil*") ao longo de uma frase. Rabiner [17], propôs um modelo de duração onde cada palavra do vocabulário é modelada através de uma função densidade de probabilidade gaussiana $f_i(d)$:

$$f_i(d) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(d-\bar{d}_i)^2}{2\sigma_i^2}\right] \tag{4.3}$$

onde $\bar{d}_i$ e σ_i são, respectivamente, a média e o desvio padrão da duração da *i-ésima* palavra, sendo seus valores obtidos através da informação de

duração, disponível na segmentação (manual ou automática) e armazenados no arquivo de vocabulário como apresentado na Figura 4.5.

Assim, ao final de cada nível, a cada instante t, do *Level Building* é calculada a duração $d_i(t)$ e a verossimilhança acumulada é então atualizada [5, 9, 17].

2) Modelo de linguagem

Seja uma seqüencia de n palavras $W = \{w_1, \ldots, w_n\}$, a probabilidade do modelo de linguagem é dada por:

$$P(W) = P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_0, \ldots, w_{i-1}) \tag{4.4}$$

Ou seja, a ocorrência de uma dada palavra w_i depende das palavras que foram anteriormente pronunciadas $(w_0, \ldots, w_{i-1})$. O problema nesta abordagem surge quando há um aumento no número de palavras anteriores a w_i a serem analisadas, fazendo portanto com que aumente a complexidade. Uma maneira de diminuir a quantidade de cálculos é fazer com que a análise seja truncada em um número menor de palavras. Isto fez surgir os modelos de gramática *n-gram* sendo que os mais utilizados são os modelos *Bi-gram*:

$$P(W) \approx \prod_{i=1}^{n} P(w_i | w_{i-1}) \tag{4.5}$$

e o modelo *Tri-gram*:

$$P(W) \approx \prod_{i=1}^{n} P(w_i | w_{i-1}, w_{i-2}) \tag{4.6}$$

Neste trabalho foi utilizado um modelo de linguagem que é uma simplificação do modelo *Bi-gram*, chamado *modelo de pares de palavras*, que é representado por:

$$P(W) = \begin{cases} 1 & \text{se o par } w_i, w_{i-1} \text{ é válido} \\ 0 & \text{caso contrário} \end{cases}$$

Capítulo 5

Resultados dos Experimentos

5.1 Introdução

Neste capítulo serão apresentados os resultados dos experimentos computacionais obtidos por intermédio do sistema híbrido estudado, utilizando como material de treinamento fones independentes de contexto listados na Tabela 4.1. A base de dados foi gentilmente cedida pelo LPDF (Laboratório de Processamento Digital da Fala) da UNICAMP, sendo esta constituída de 100 sentenças dependentes de locutor, gravadas em ambiente de estúdio.

A partir destes resultados, as análises do sistema seguiram os seguintes objetivos:

- A verificação do comportamento dos parâmetros responsáveis pelo desempenho dos modelos híbridos ANN+HMM: Através destes parâmetros foi possível acompanhar a evolução do aprendizado do sistema frente ao material de treinamento apresentado. Este comportamento é influenciado pela informação de segmentação fonética.

- A avaliação da influência dos erros de segmentação manual: Verificou-se a robustez dos modelos híbridos em relação ao deslocamento das marcas originais das locuções de treinamento.

- A avaliação da influência da informação contextual: Mostrou como a inserção de c vetores acústicos antes e depois do quadro sob análise pode ser útil na estimação dos parâmetros do sistema.

O sistema foi treinado de acordo com o algoritmo apresentado na Seção 3.3, sendo feitas três reestimações dos alvos suaves $\gamma_n(k)$. Ao final de cada reestimação de $\gamma_n(k)$ a rede neural foi treinada durante n_{reest} vezes ($n_{reest} = 10, 50, 100$). Desta forma, a apresentação dos resultados seguirá a seguinte denominação:

- *Pré-REMAP*: Referente ao treinamento da rede neural com os exemplos de treinamento centrados entre as marcas de segmentação fonética.
- *REMAP-1*: Treinamento do sistema após a primeira reestimação dos alvos suaves.
- *REMAP-2*: Treinamento do sistema após a segunda reestimação dos alvos suaves.
- *REMAP-3*: Treinamento do sistema após a terceira reestimação dos alvos suaves.

5.2 O critério de avaliação

A avaliação é o processo pelo qual a saída de um sistema de reconhecimento de fala contínua é comparado com frases de referência (frases corretas). Os possíveis erros são contabilizados e apresentados em um formato que ajude a compreender as deficiências do sistema.

A métrica utilizada é a *taxa de erros de palavras* (do termo em inglês *Word Error Rate – WER*), que calcula o número e avalia os tipos de erros cometidos pelo reconhecedor segundo a equação:

$$WER = \left(\frac{N_I + N_S + N_D}{N_r} \right) \times 100 \tag{5.1}$$

onde:

WER: taxa de erro de palavras

N_I: número de erros de inserção

N_S: número de erros de substituição

N_D: número de erros de deleção

N_r: número de palavras presentes no conjunto de referência

Para esta tarefa foi utilizada uma ferramenta que faz parte do pacote SCTK desenvolvido pelo NIST chamado SCLITE [4], que será apresentado no Apêndice D desta Dissertação.

5.3 Resultados

5.3.1 Segmentação Manual

Na primeira parte da Tabela 5.1 apresenta os valores de taxa de erros de palavras ao se utilizar uma rede neural de 84 entradas e 100 reestimações das matrizes de pesos sinápticos após cada atualização das matrizes de alvos suaves. O mesmo foi feito para uma rede neural de 108 entradas, que obteve os melhores resultados com o mesmo número de atualizações dos pesos sinápticos da rede anterior, sendo apresentada na parte seguinte. Tais valores foram obtidos ao se utilizar modelo de duração de palavras e gramática *pares de palavras* no reconhecimento das sentenças.

Tabela 5.1: *Desempenho do sistema híbrido ANN+HMM para uma base de dados segmentada manualmente, utilizando modelo de duração de palavras e gramática pares de palavras.*

Número de entradas	Etapas de treinamento	$S(\%)$	$D(\%)$	$I(\%)$	$WER(\%)$
84	Pré-REMAP	14,2	3,2	8,1	25,5
	REMAP-1	5,1	2,8	4,0	11,9
	REMAP-2	4,1	2,4	3,5	10,0
	REMAP-3	3,3	2,3	2,7	8,3
108	Pré-REMAP	9,9	3,3	5,3	18,5
	REMAP-1	4,2	2,2	4,1	10,5
	REMAP-2	4,1	2,3	4,9	11,3
	REMAP-3	3,7	2,3	3,7	9,7

5.3.2 Segmentação Uniforme

Partindo do pressuposto de que os fonemas possuem mesma duração em uma determinada sentença, foram executados os testes seguindo a mesma metodologia abordada na subseção anterior. Desta forma, a Tabela 5.2 apresenta os resultados obtidos utilizando-se modelo de duração de palavras e gramática.

Pôde-se notar nesta seção que a suposição de uniformidade das durações dos fonemas a cada sentença mostrou ser uma aproximação muito ruim, comparada com os resultados obtidos pelo sistema ao reconhecer sentenças submetidas à segmentação manual.

Deve-se estar atento também ao fato de que algumas taxas de erros de palavras possuírem valores maiores que 100%. Isto mostra como o sistema errou em quantidade tal que chegou a superar o número de palavras presentes no conjunto de

Tabela 5.2: *Desempenho para uma base de dados submetida a segmentação uniforme, utilizando modelo de duração de palavras e gramática p-gram.*

Número de entradas	Etapas de treinamento	$S(\%)$	$D(\%)$	$I(\%)$	$WER(\%)$
84	Pré-REMAP	78,5	6,6	22,7	107,8
	REMAP-1	47,1	5,4	32,6	85,1
	REMAP-2	29,6	5,4	26,5	61,5
	REMAP-3	21,0	5,4	19,3	45,7
108	Pré-REMAP	72,0	10,8	14,6	97,4
	REMAP-1	52,0	6,8	33,1	91,9
	REMAP-2	32,2	5,2	22,6	60,0
	REMAP-3	27,5	4,4	17,3	49,2

referência, ou seja, $N_I + N_S + N_D > N_r$, ocasionando em uma brutal queda no desempenho, se comparado com os valores obtidos com segmentação manual.

5.3.3 Simulação dos erros de segmentação manual

O objetivo é analisar a influência na taxa de erros de palavras através do deslocamento das marcas de segmentação manual. Para a simulação destes erros foi implementado um algoritmo, com seus procedimentos apresentados abaixo.

Implementação computacional para simulação dos efeitos de erro de segmentação

Seja uma dada locução genérica de duração T_l composta de N fones, cada um com duração d_{f_n}, como mostrado na Figura 5.1:

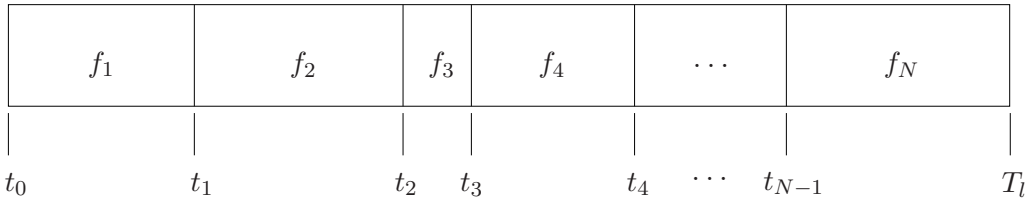

Figura 5.1: *Locução genérica contendo N fones cujas marcas de segmentação foram extraídas manualmente.*

Deslocando as marcas de segmentação em $\pm\tau_i$ milissegundos tem-se:

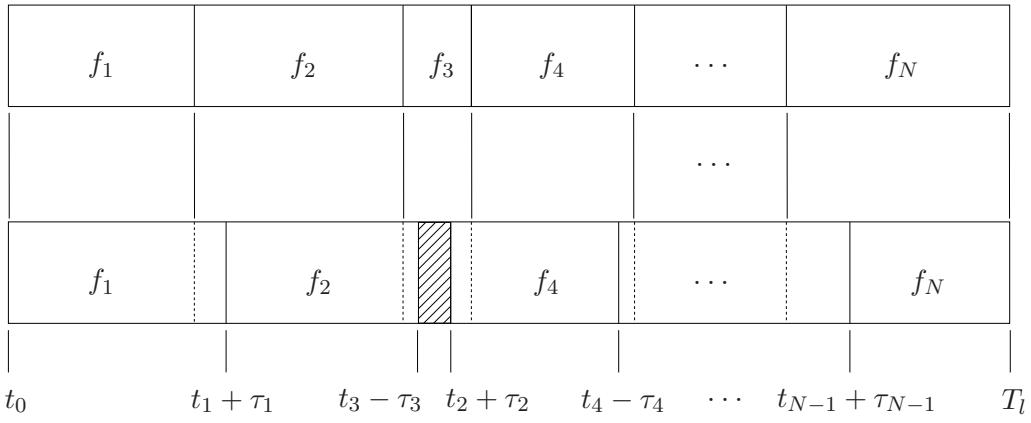

Figura 5.2: *Locução genérica contendo N fones com as marcas de segmentação originais deslocadas de* $\pm\tau_i$.

Os valores de τ_i são números aleatórios de distribuição uniforme que foram gerados dentro do intervalo $[-T; +T]$. Pode-se notar a existência de uma sobreposição, indicada pela área hachurada na Figura 5.2. Para esta situação, considera-se que a referida área seja um fone (neste caso o fone f_3) com marcas inicial e final em $t_3 - \tau_3$ e $t_2 + \tau_2$, respectivamente, e então são calculadas as novas durações a partir das novas marcas encontradas.

Mas apesar de contornado o problema da "*sobreposição de fones*", corre-se o risco de que o valor da duração arbitrada pela abordagem acima seja muito pequeno ou nulo. Para que estas durações sejam penalizadas, deve ser feito um levantamento *a priori* da média μ_{f_i} e do desvio padrão σ_{f_i} do conjunto de fones utilizado e caso a duração arbitrada (d_{f_3}) seja menor que o seu correspondente desvio padrão, é considerado que a duração do "*fone sobreposto*" seja igual a $d'_{f_3} = \sigma_{f_3}$. Porém, a penalização de uma dada duração resulta na diminuição da duração dos fones anterior e posterior em

$$e = \frac{d'_{f_i} - d_{f_i}}{2} \tag{5.2}$$

acarretando no cálculo de suas respectivas novas marcas segundo as equações abaixo:

$$t'_{f_{i-1}} = t_{f_{i-1}} - e \tag{5.3}$$

$$t'_{f_{i+1}} = t_{f_i} + e \tag{5.4}$$

Este ajuste está ilustrado na Figura 5.3.

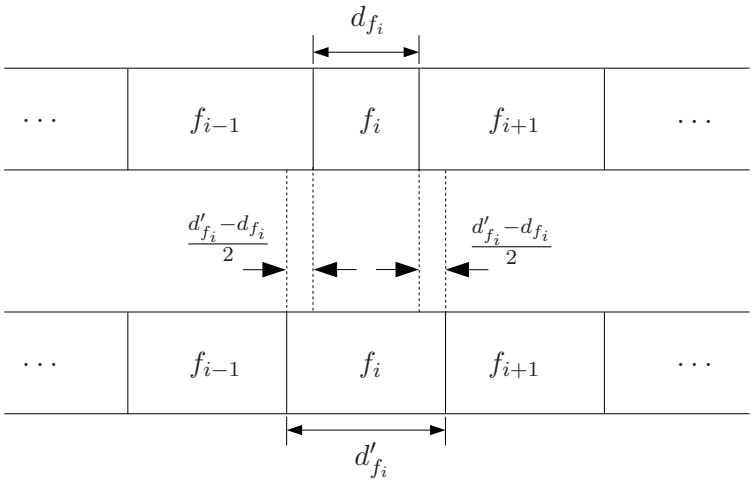

Figura 5.3: *Trecho de locução no qual ocorre a penalização da duração do fone f_i e posterior ajuste das marcas de segmentação dos fones f_{i-1} e f_{i+1}.*

Todo procedimento deste método está ilustrado nos passos a, b, c e d da Figura 5.4.

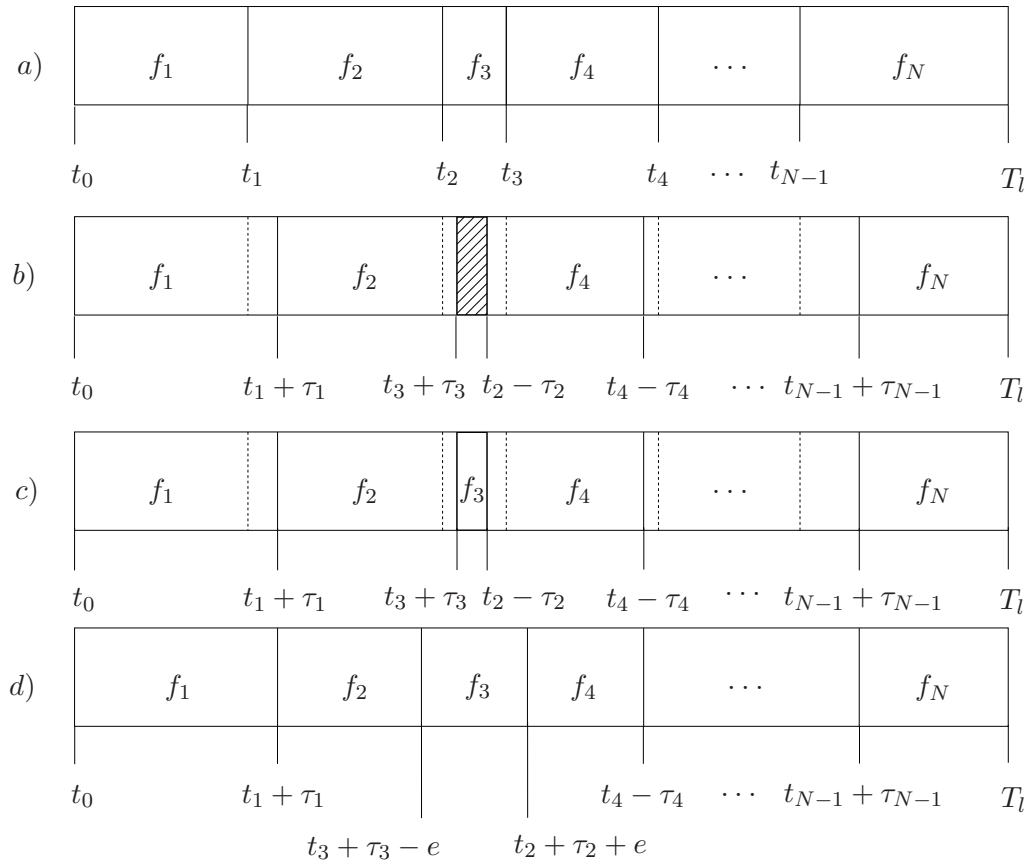

Figura 5.4: *Processo de simulação do erro de segmentação. a) locução com as marcas originais, b) deslocando as marcas em $\pm\tau_i$, c) definindo duração do fone f_3 como sendo σ_{f_3}, d) novas marcas de segmentação, tendo a duração de f_3 penalizada.*

A Figura 5.5 mostra o resultado do deslocamento das marcas de segmentação manual originais para um desvio máximo de $T = \pm 60\ ms$ da locução "*É suficiente*". As linhas tracejadas correspondem às informações originais da segmentação manual, enquanto que as linhas contínuas foram obtidas através do processo de simulação dos erros de segmentação manual descrito acima.

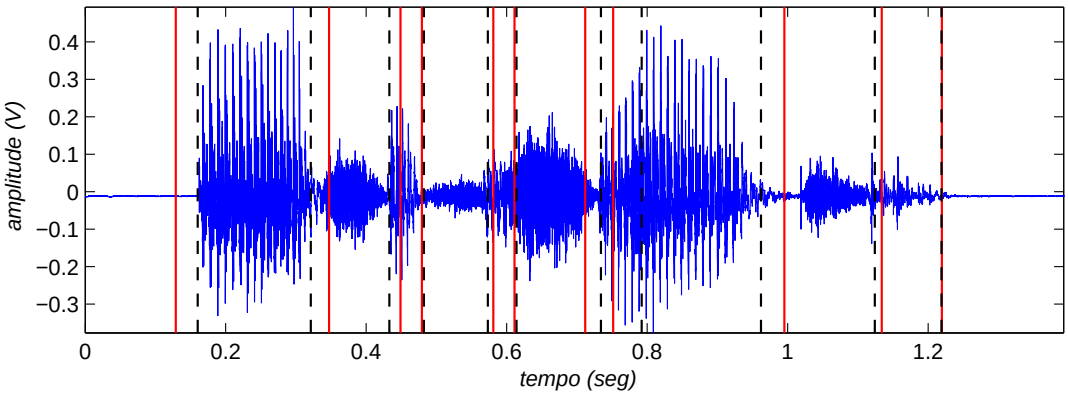

Figura 5.5: *Resultado da mudança das marcas de segmentação manual originais (linhas tracejadas) para um desvio máximo $T = \pm 60$ ms (linhas contínuas) para a sentença "É suficiente".*

Resultados experimentais da abordagem

A partir do procedimento descrito acima, os testes foram executados para valores máximos de T variando de $\pm 10\ ms$ a $\pm 60\ ms$, com passos de 10 ms. Isto equivale a dizer que as marcas originais de segmentação manual sofreram desvios de 1 a 6 vetores acústicos.

As Tabelas 5.3 e 5.4 mostram as taxas de erros de palavras obtidas pelo sistema ao se utilizar o modelo de duração de palavras e gramática *pares de palavras*, com redes neurais de 84 e 108 entradas, respectivamente. A primeira coluna é referente aos valores máximos de desvio das marcas originais, a segunda coluna refere-se ao número de atualizações das matrizes de pesos sinápticos, a cada reestimação dos alvos suaves.

Pode-se notar através das Tabelas 5.3 e 5.4 que, a medida que o desvio das marcas originais aumenta (para valores de T maiores que 20 milissegundos), ocorre o aumento da taxa de erros de palavras. Isto mostra que uma segmentação manual executada de maneira displicente pode resultar em um fraco desempenho do sistema implementado.

Como pôde ser notado nos testes em uma base de dados submetida à segmentação manual, a restrição gramatical contribuiu muito para que as taxas de

erros de palavras diminuíssem em todas as avaliações executadas. Este comportamento era esperado para todos os testes pois a gramática faz com que diminua o campo de possibilidades de composição de pares de palavras, eliminando assim maior parte das combinações que acarretariam em erro de reconhecimento.

Tabela 5.3: *Resultados obtidos em uma rede neural de 84 entradas, utilizando modelo de duração de palavras e gramática p-gram.*

$T(ms)$	N_{reest}	etapas de treinamento	$S(\%)$	$D(\%)$	$I(\%)$	$WER(\%)$
10	100	Pré-REMAP	13,7	3,0	7,4	24,1
		REMAP-1	4,3	2,9	5,5	12,7
		REMAP-2	4,2	2,2	5,7	12,1
		REMAP-3	4,5	2,4	4,9	11,8
20	50	Pré-REMAP	12,9	2,9	7,3	23,1
		REMAP-1	6,0	3,3	5,9	15,2
		REMAP-2	4,6	2,3	4,2	12,8
		REMAP-3	3,5	2,2	4,9	10,6
30	50	Pré-REMAP	14,8	2,9	7,2	24,9
		REMAP-1	5,7	1,7	5,4	12,8
		REMAP-2	4,3	2,0	4,8	11,1
		REMAP-3	4,2	2,1	6,9	13,2
40	10	Pré-REMAP	16,3	4,1	12,3	32,7
		REMAP-1	8,0	5,6	15,6	29,2
		REMAP-2	5,5	2,5	10,8	23,6
		REMAP-3	4,9	2,2	9,7	16,8
50	50	Pré-REMAP	13,0	3,9	14,5	31,4
		REMAP-1	12,9	4,4	12,8	30,1
		REMAP-2	12,5	4,0	12,8	29,3
		REMAP-3	12,5	2,7	7,5	22,7
60	10	Pré-REMAP	16,4	3,0	13,0	33,4
		REMAP-1	13,0	2,7	10,9	26,6
		REMAP-2	12,1	2,8	9,3	24,2
		REMAP-3	14,8	3,0	5,2	23,0

Tabela 5.4: *Resultados obtidos em uma rede neural de 108 entradas, utilizando modelo de duração de palavras e gramática p-gram.*

$T(ms)$	N_{reest}	etapas de treinamento	$S(\%)$	$D(\%)$	$I(\%)$	$WER(\%)$
10	100	Pré-REMAP	12,5	2,7	7,5	22,7
		REMAP-1	5,6	2,3	5,9	13,8
		REMAP-2	3,9	2,2	4,3	10,4
		REMAP-3	3,4	2,7	4,3	10,3
20	50	Pré-REMAP	12,2	2,8	8,1	23,1
		REMAP-1	4,6	2,9	4,4	11,9
		REMAP-2	4,5	2,6	4,4	11,5
		REMAP-3	3,8	2,1	3,8	9,7
30	100	Pré-REMAP	13,8	2,7	9,5	26,0
		REMAP-1	5,3	2,2	4,6	12,1
		REMAP-2	3,3	2,2	4,4	9,9
		REMAP-3	3,6	2,0	5,3	10,9
40	10	Pré-REMAP	14,8	3,2	9,9	27,9
		REMAP-1	6,0	2,7	6,4	15,1
		REMAP-2	5,3	2,3	6,2	13,8
		REMAP-3	4,7	2,6	4,4	11,7
50	50	Pré-REMAP	8,9	2,1	9,3	20,3
		REMAP-1	11,4	3,1	5,3	19,8
		REMAP-2	7,1	3,1	6,4	16,6
		REMAP-3	6,5	2,4	5,2	14,1
60	10	Pré-REMAP	15,4	3,8	10,3	28,5
		REMAP-1	12,8	2,6	8,1	23,5
		REMAP-2	10,1	2,4	7,3	19,8
		REMAP-3	8,0	1,9	5,8	15,7

Tendo em vista os resultados obtidos nesta subseção, concluiu-se que para erros de segmentação de até $T = \pm 30\ ms$, ou seja, para um desvio de 3 vetores acústicos, os modelos híbridos ANN+HMM conseguem resultados que se equiparam com a segmentação manual e após este valor ocorre uma piora considerável na taxa de erros de palavras. Significa também que o sistema é bastante sensível à segmentação das locuções de treinamento.

O sistema apresentou uma diminuição na taxa de erro de palavras com o aumento da informação contextual de 3 para 4 vetores acústicos antes e depois do quadro de análise x_n. Isto mostra que para a simulação dos erros de segmentação manual foi necessária a inserção de um trecho maior de fala ($40\ ms$) na entrada da rede neural, para que sejam compensadas as imperfeições ocorridas na definição das marcas de segmentação das sentenças.

5.3.4 Probabilidades de transição

Esta seção tem como objetivo apresentar alguns fenômenos observados nas estimações das probabilidades de transição de estados. Notou-se que o tipo de segmentação adotado influencia na estimação destes parâmetros e consequentemente no desempenho do sistema estudado.

Todos os gráficos apresentados nesta seção foram retirados dos modelos híbridos ao se utilizar uma rede neural de 108 entradas.

Segmentação manual

Nas Figuras 5.6 e 5.7 são apresentadas, respectivamente, as probabilidades de autotransição e transição estimadas nas etapas Pré-REMAP, REMAP-1, REMAP-2 e REMAP-3 do sistema híbrido ANN+HMM, com uma rede neural composta de 84 entradas na camada escondida, para locuções de treinamento submetidas à segmentação manual. Estas probabilidades são referentes ao desempenho apresentado na Tabela 5.1:

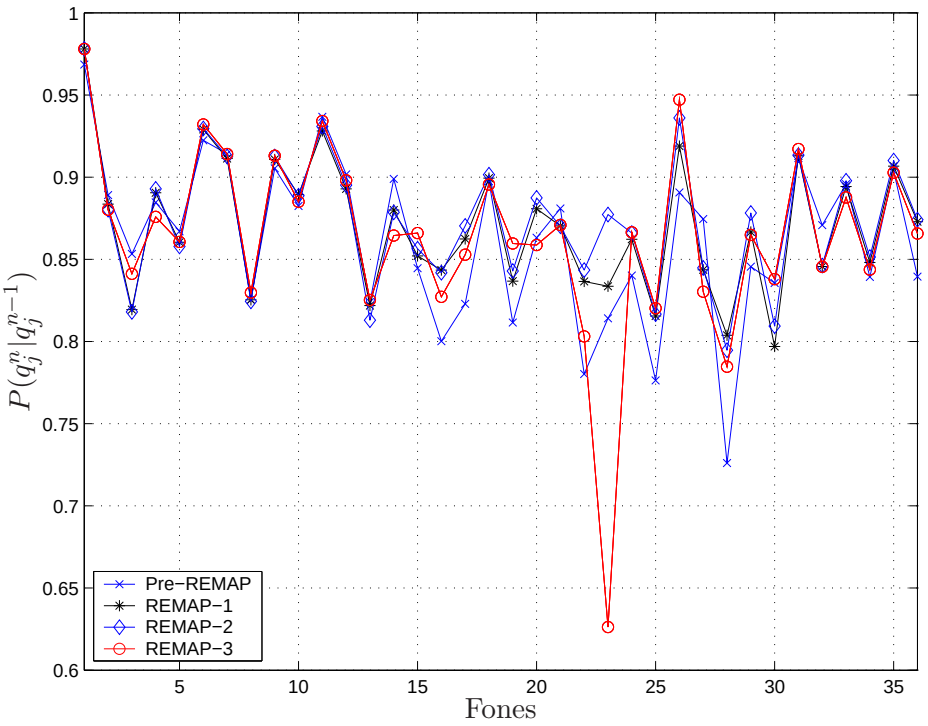

Figura 5.6: *Estimações das probabilidades de autotransição de estados, com as sentenças de treinamento submetidas à segmentação manual.*

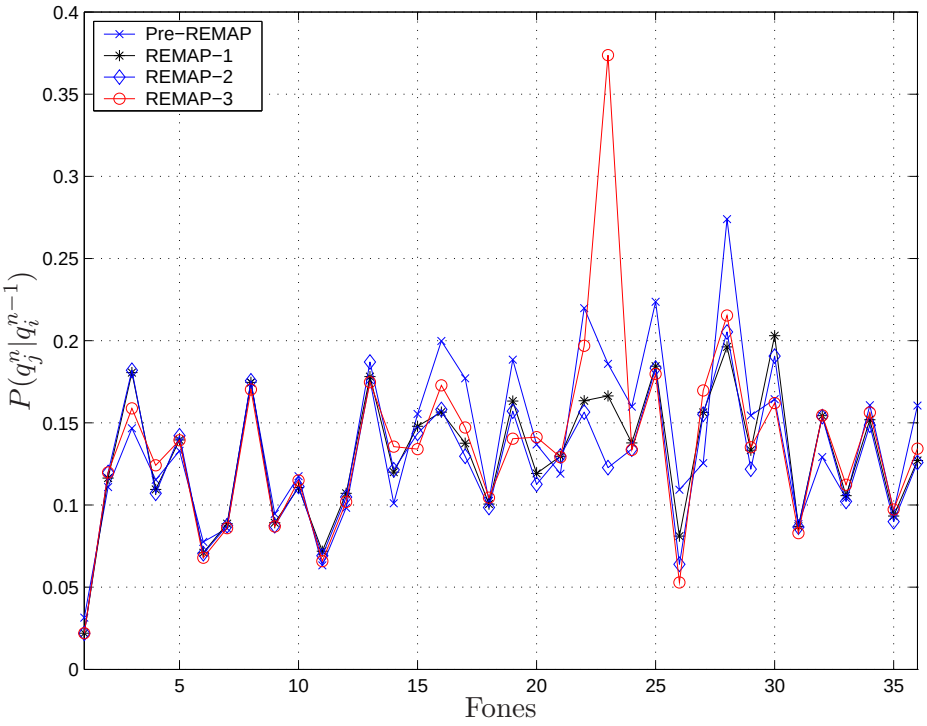

Figura 5.7: *Estimações das probabilidades de transição de estados, com as sentenças de treinamento submetidas à segmentação manual.*

Segmentação uniforme

As Figuras 5.8 e 5.9 apresentam, respectivamente, as estimações das probabilidades de autotransição e probabilidades de transição de estados. Estes valores foram obtidos através do treinamento de sentenças submetidas à segmentação uniforme.

Nestas figuras é possível notar o comportamento da primeira estimação destas probabilidades (Pré-REMAP) que, devido à suposição das durações dos fones, fez com que os valores das probabilidades fossem praticamente os mesmos para todos os estados. Comportamento este que não se estende para as demais etapas de reestimação.

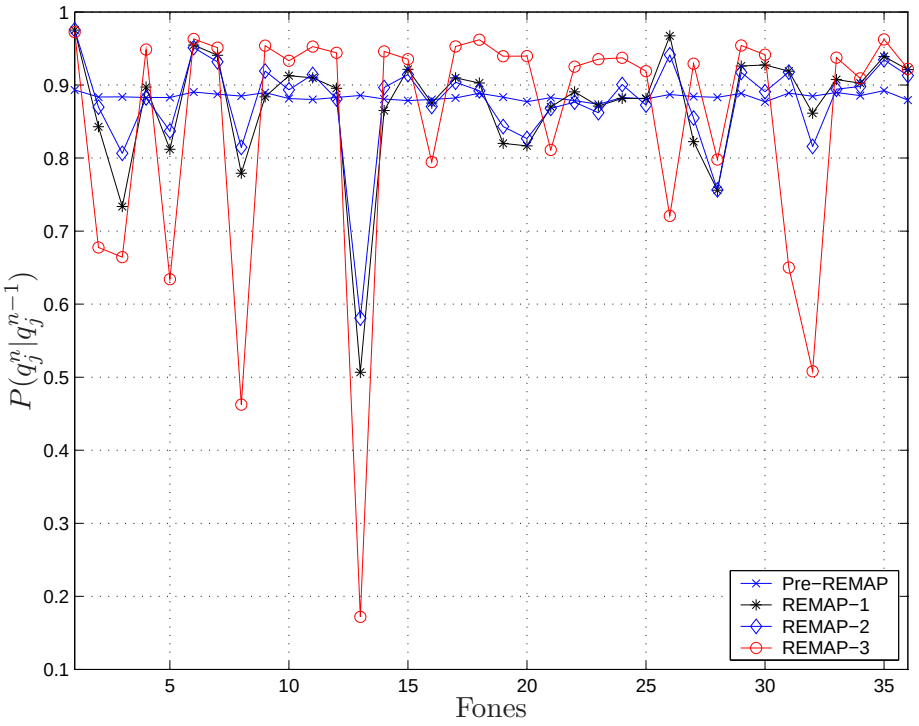

Figura 5.8: *Estimações das probabilidades de autotransição de estados, com as sentenças de treinamento submetidas à segmentação uniforme.*

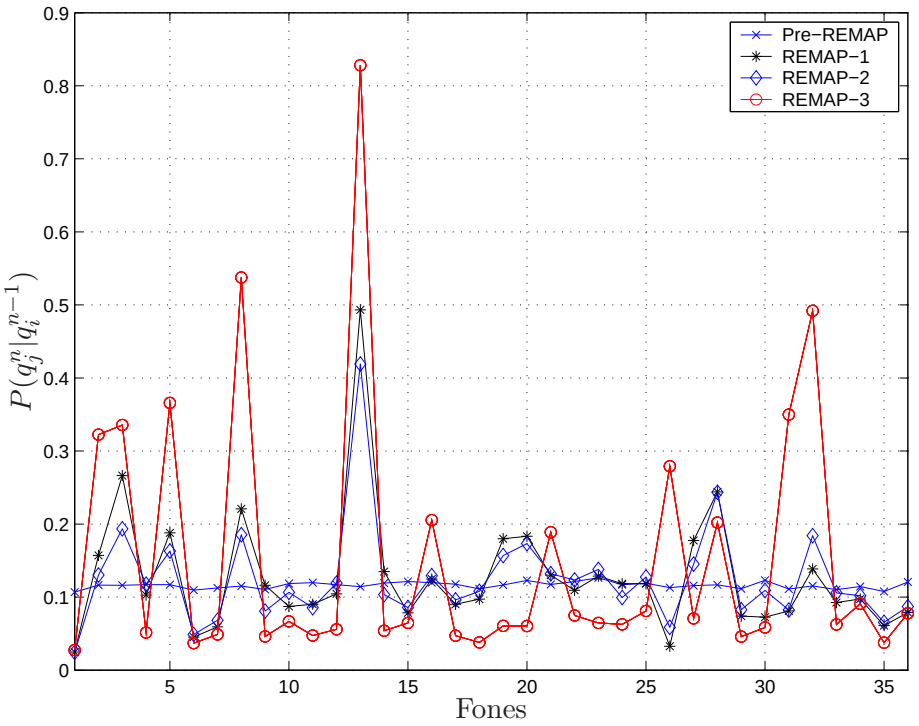

Figura 5.9: *Estimações das probabilidades de transição de estados, com as sentenças de treinamento submetidas à segmentação uniforme.*

Simulação dos erros de segmentação manual

Aqui serão apresentadas as probabilidades de transição de estados estimadas pelo sistema híbrido ANN+HMM, para as sentenças de treinamento submetidas ao procedimento de simulação dos erros de segmentação manual descrito na Subseção 5.3.3.

As Figuras 5.10 e 5.11 apresentam as estimações das probabilidades de autotransição e transição de estados, respectivamente, para a simulação dos erros de segmentação manual, utilizando um deslocamento máximo das marcas originais de ± 10 ms.

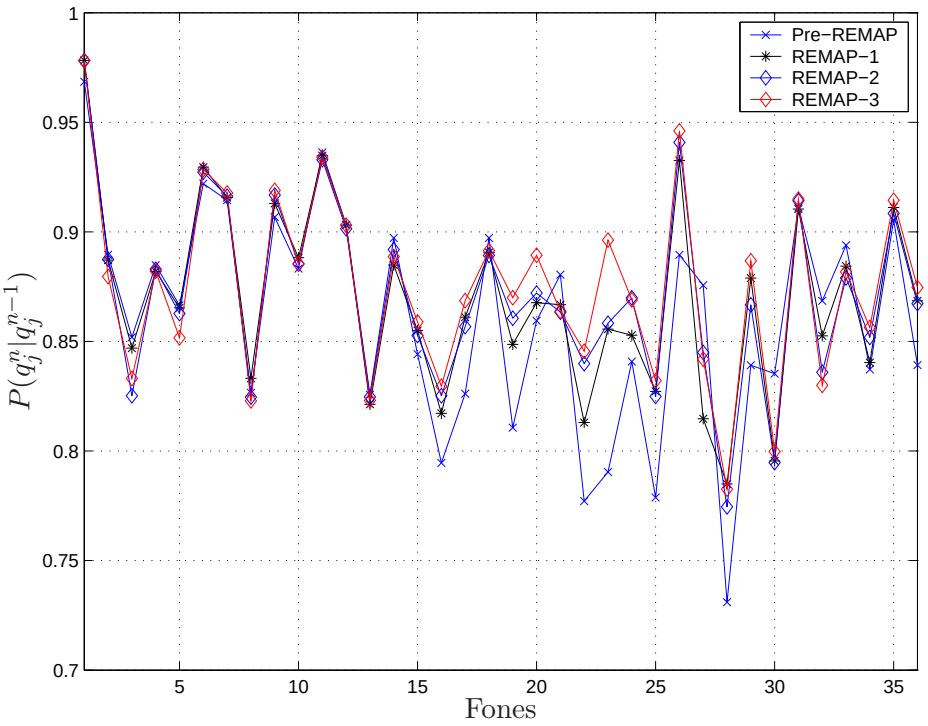

Figura 5.10: *Probabilidades de autotransição de estados, para desvio* $T = \pm 10$ *ms.*

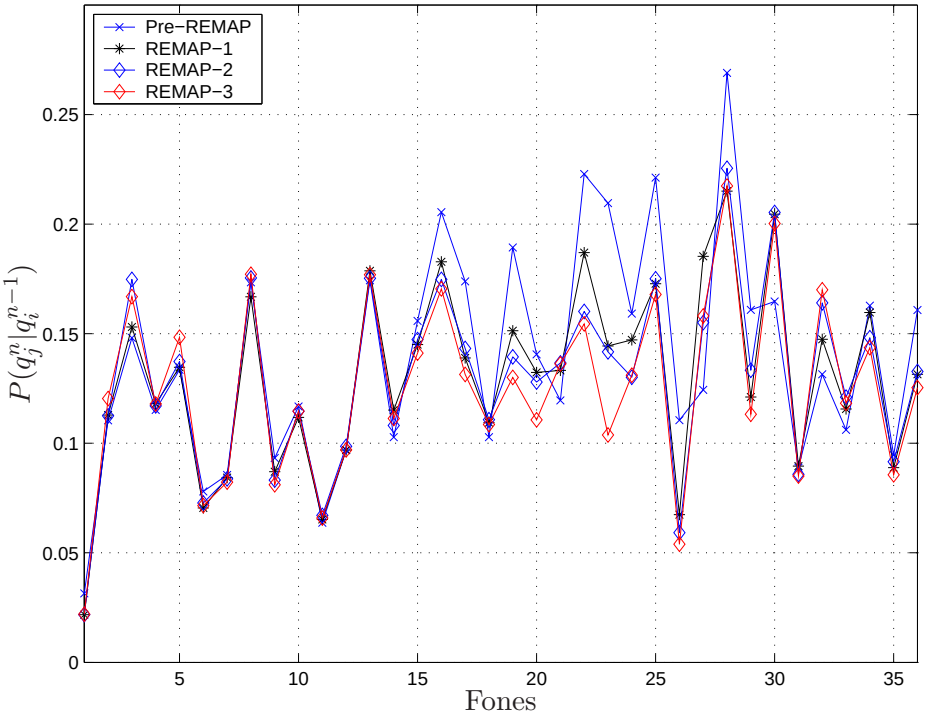

Figura 5.11: *Probabilidades de transição de estados, para desvio* $T = \pm 10$ *ms.*

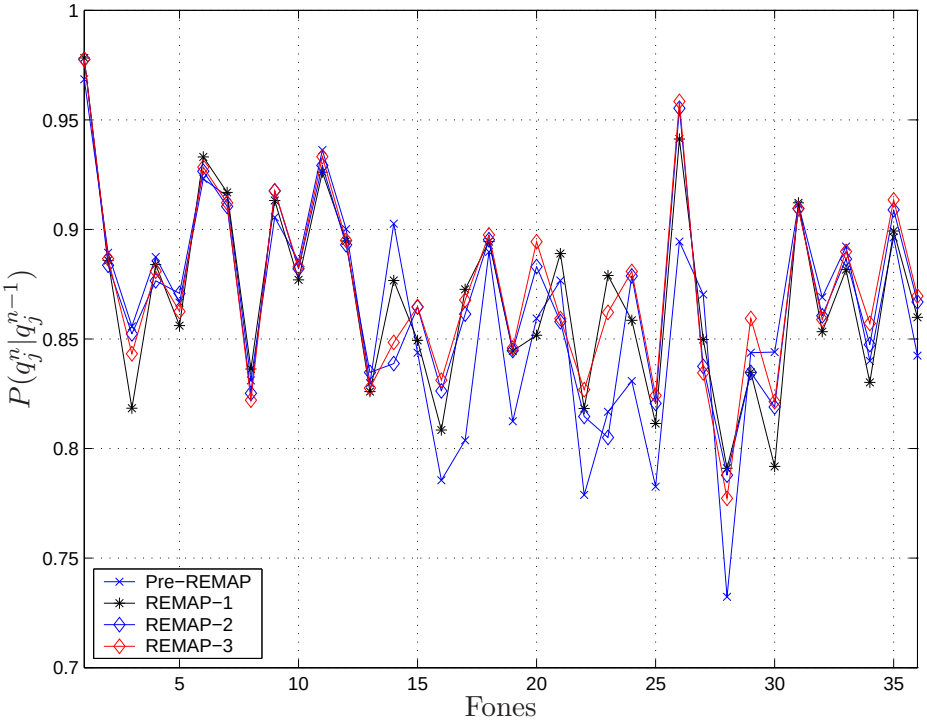

Figura 5.12: *Probabilidades de autotransição de estados, para desvio* $T = \pm 20$ *ms.*

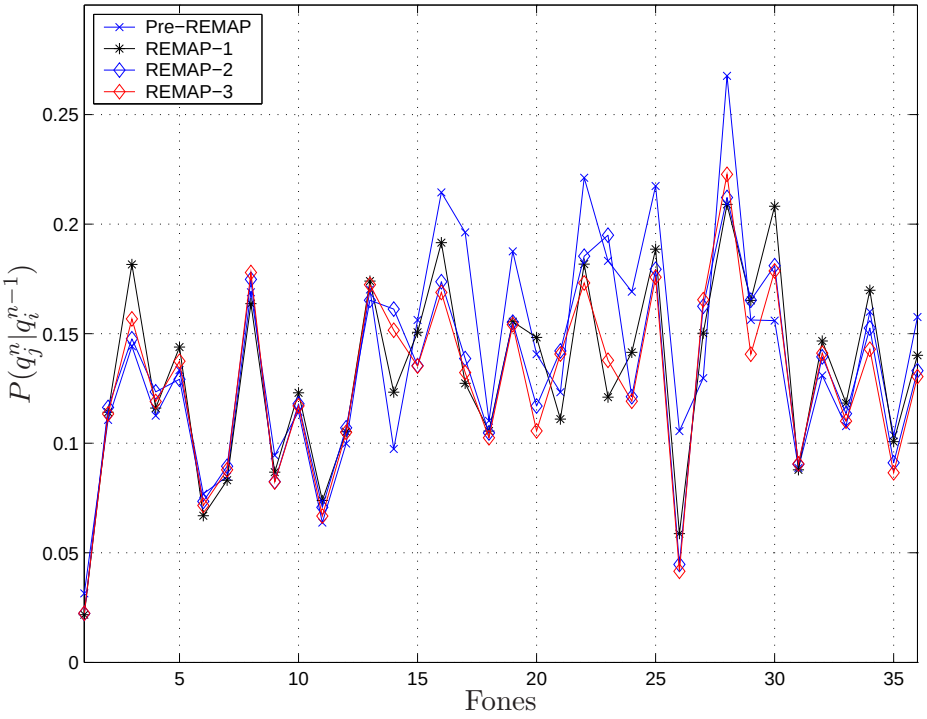

Figura 5.13: *Probabilidades de transição de estados, para desvio* $T = \pm 20$ *ms.*

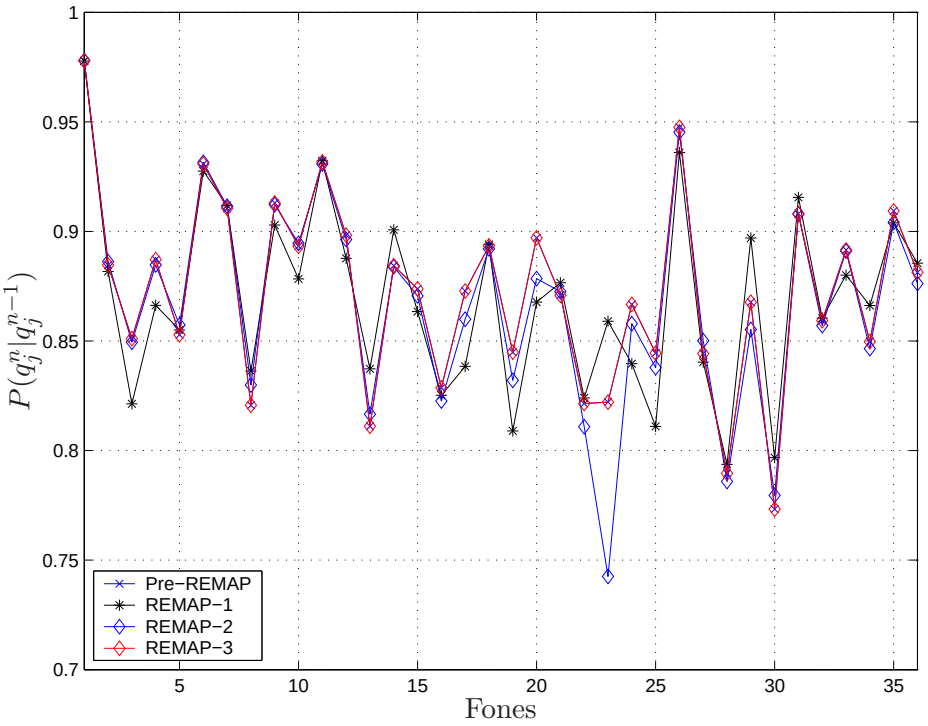

Figura 5.14: *Probabilidades de autotransição de estados, para desvio* $T = \pm 30$ *ms.*

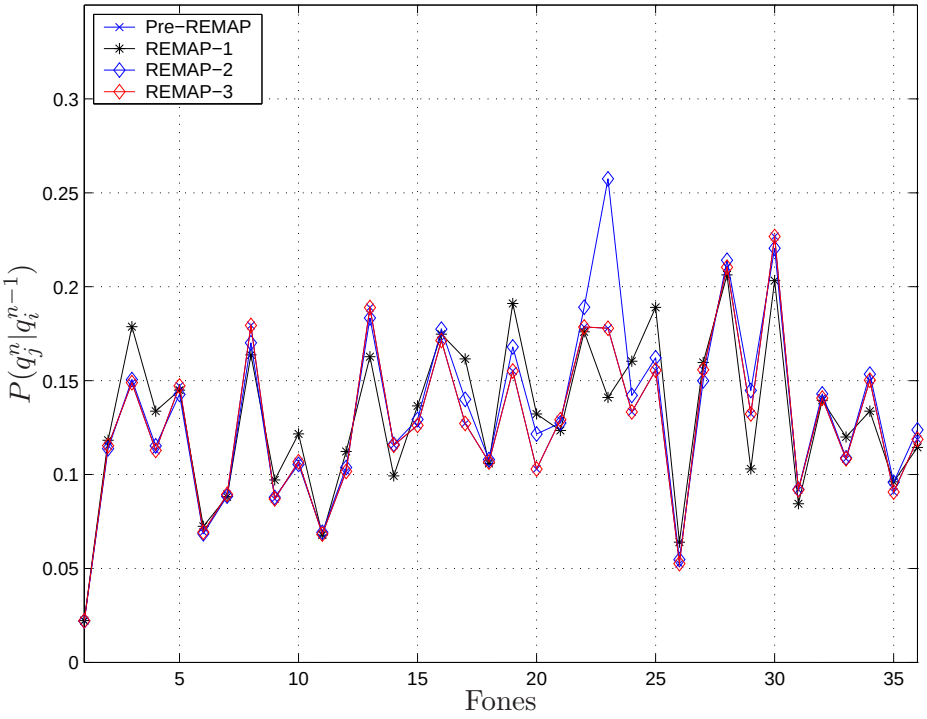

Figura 5.15: *Probabilidades de transição de estados, para desvio* $T = \pm 30$ *ms.*

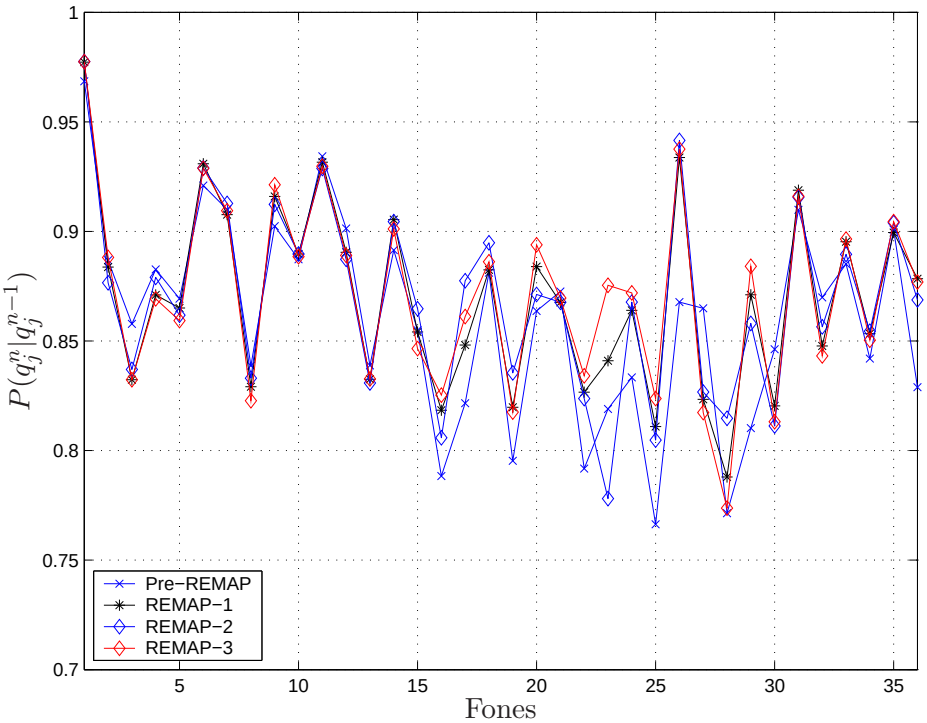

Figura 5.16: *Probabilidades de autotransição de estados, para desvio* $T = \pm 40$ *ms.*

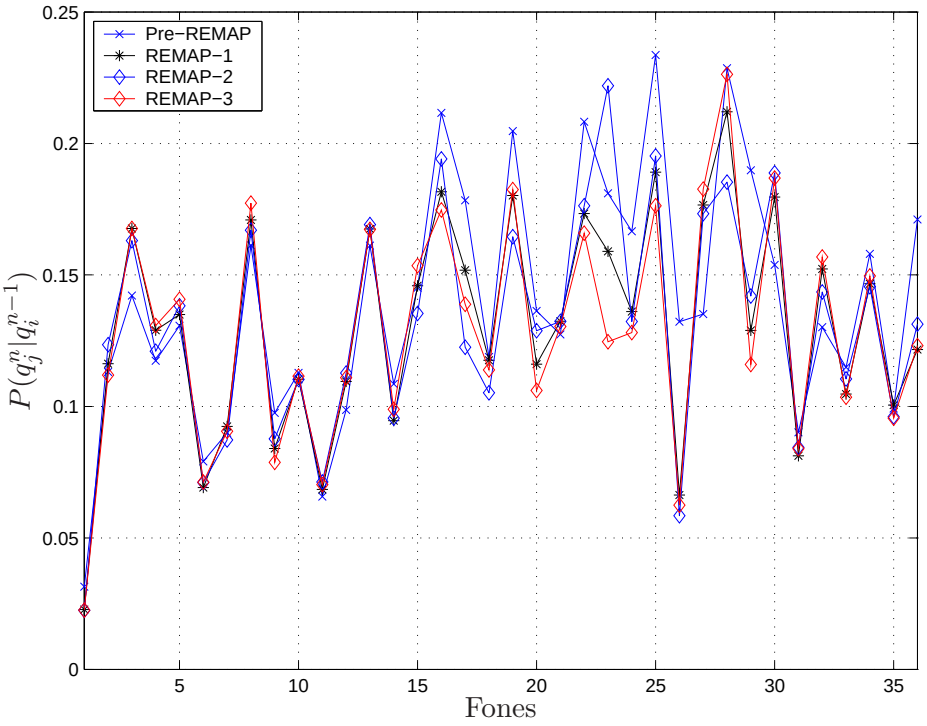

Figura 5.17: *Probabilidades de transição de estados, para desvio* $T = \pm 40$ *ms.*

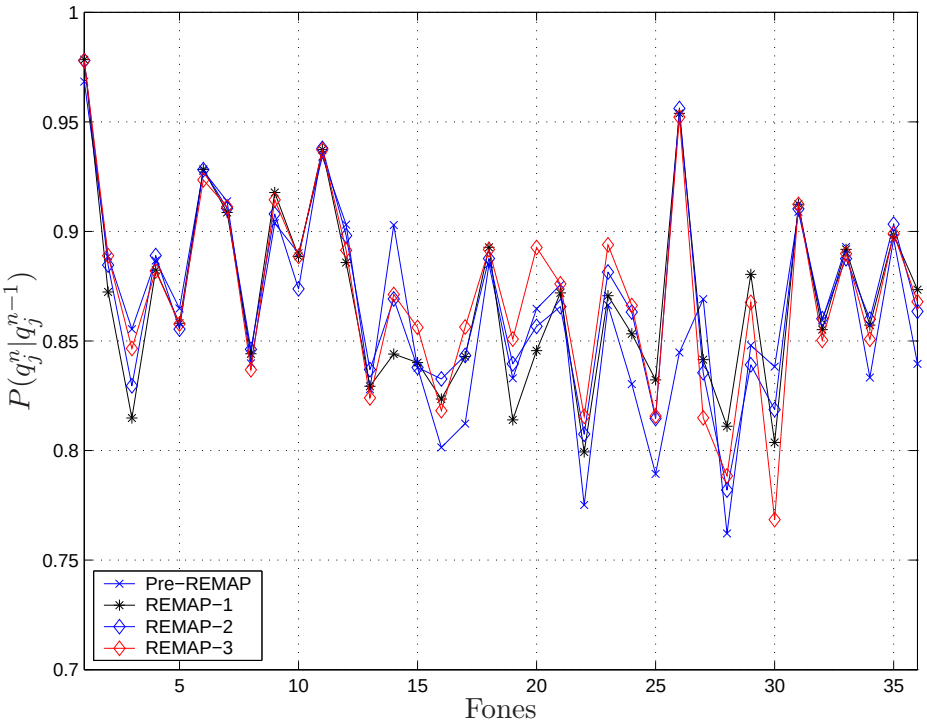

Figura 5.18: *Probabilidades de autotransição de estados, para desvio* $T = \pm 50$ *ms.*

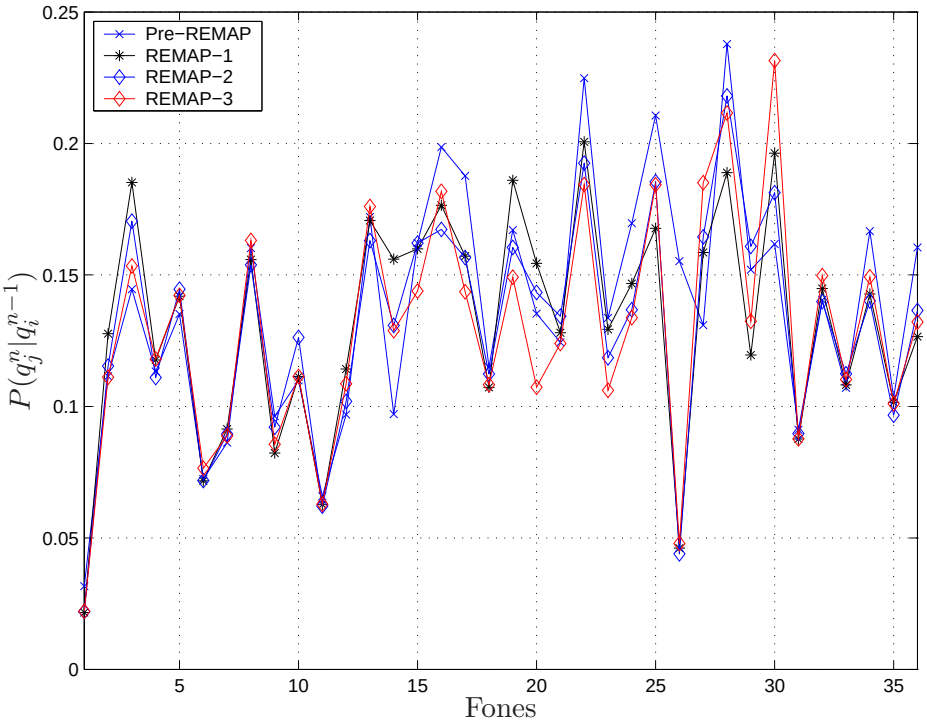

Figura 5.19: *Probabilidades de transição de estados, para desvio* $T = \pm 50$ *ms.*

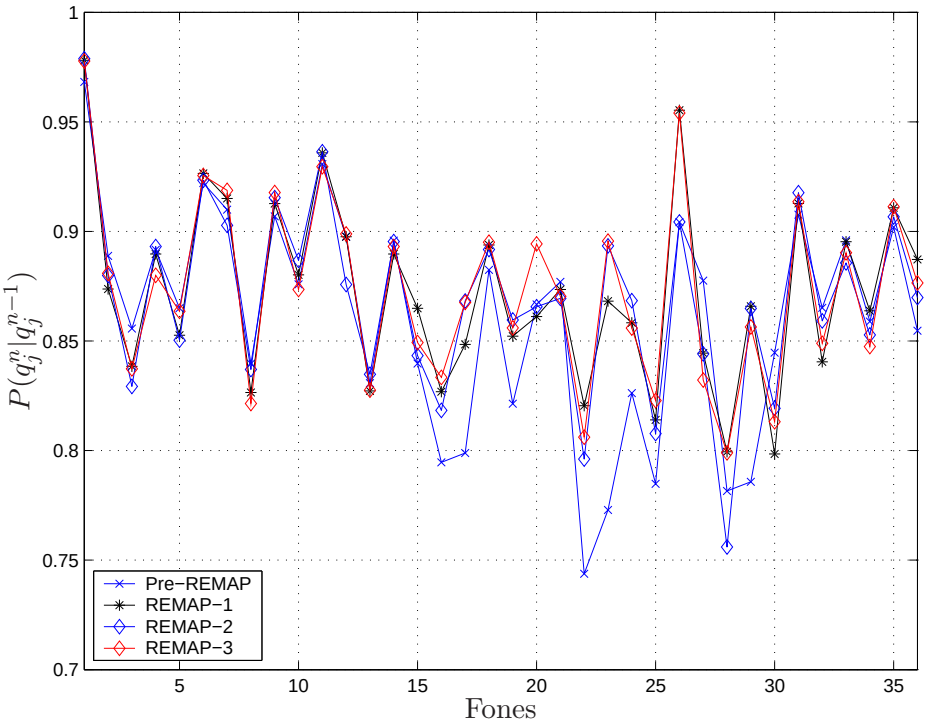

Figura 5.20: *Probabilidades de autotransição de estados, para desvio* $T = \pm 60$ *ms.*

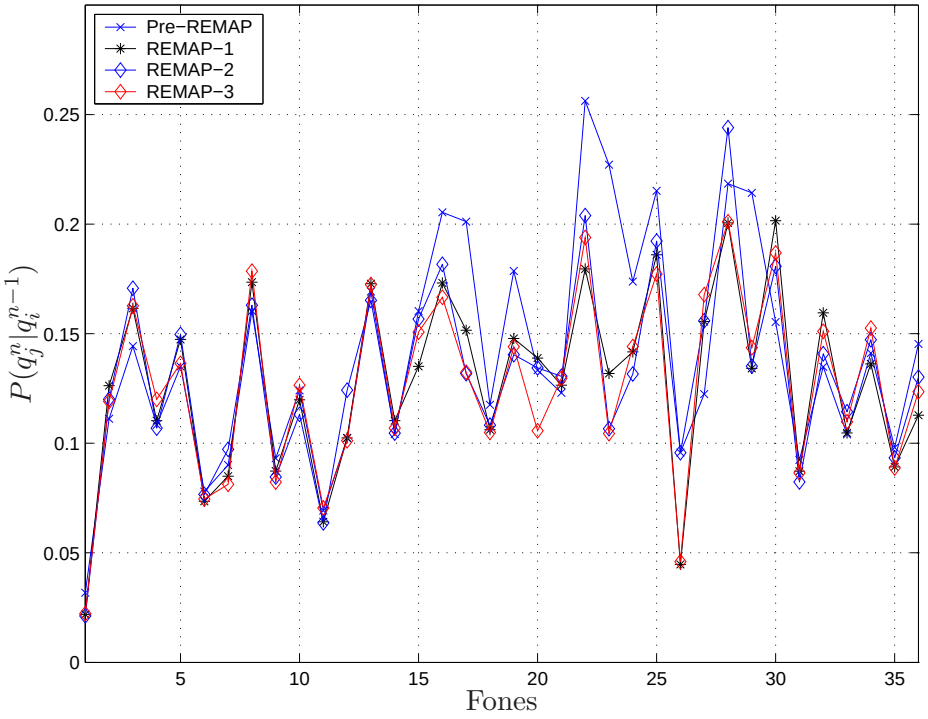

Figura 5.21: *Probabilidades de transição de estados, para desvio* $T = \pm 60$ *ms.*

Discussão

Nesta Subseção foram apresentados os resultados das reestimações das probabilidades de autotransição e de transição de estados calculadas pelos modelos híbridos ANN+HMM.

Pôde ser notado que as probabilidades de transição reestimadas nas etapas REMAP-1, REMAP-2 e REMAP-3 para a base de dados submetida à segmentação manual seguiram o comportamento das probabilidades na etapa Pré-REMAP, calculadas através da Equação (3.26), baseada nas durações médias dos fones. Isto possibilitou a estimação dos parâmetros dos modelos de forma satisfatória, levando às taxas de erros de palavra apresentada na Tabela 5.1. Para a simulação dos erros de segmentação o comportamento das probabilidades $P(q_j^n|q_i^{n-1})$ foi similar à segmentação manual.

Na segmentação uniforme, ao contrário das duas últimas análises, os gráficos das reestimações das probabilidades de transição de estados nas etapas REMAP-1, REMAP-2 e REMAP-3 não seguiram o mesmo padrão obtido pela etapa Pré-REMAP. Este fenômeno mostra como a suposição da uniformidade das durações fonéticas nas sentenças não foi razoável, prejudicando o desempenho do sistema em termos de taxa de erros de palavra.

Portanto, conclui-se que a estimação inicial das probabilidades de transição de estados para o treinamento de um sistema baseado em modelos híbridos

ANN+HMM será boa quando a base de dados for submetida à segmentação manual, podendo ser cometidos erros de segmentação de até ±30 milissegundos.

5.3.5 Probabilidade *a priori* das classes

A Figura 5.22 apresenta o histograma da ocorrência dos fones com os dados da Tabela 4.1.

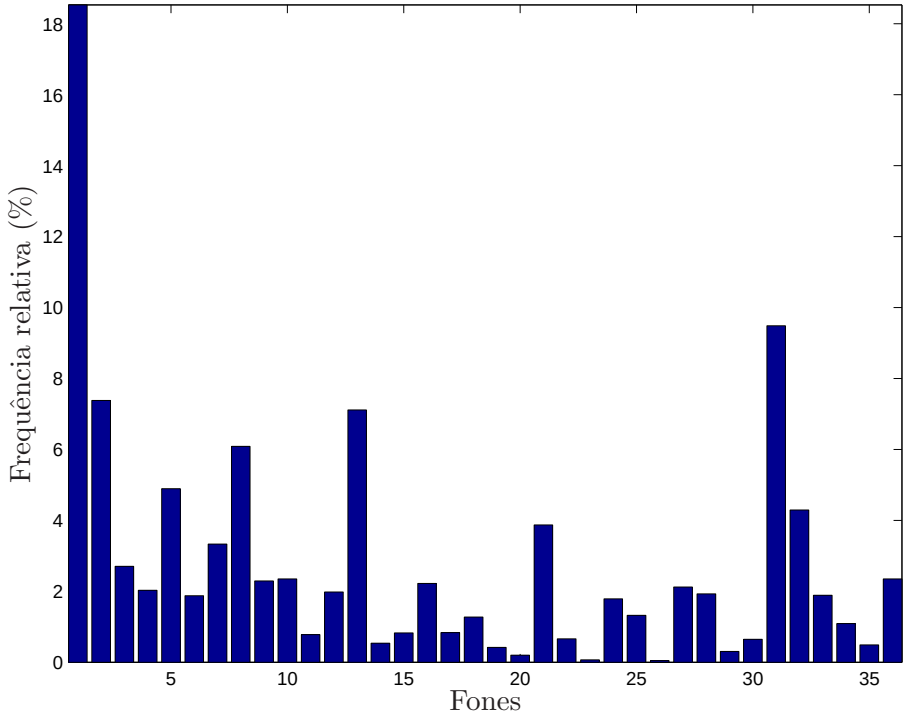

Figura 5.22: *Histograma da ocorrência dos fones para a base de dados submetida à segmentação manual.*

A probabilidade *a priori* das classes $P(q_k)$ foi calculada segundo Equação (2.30) da Subseção 2.3.4. Espera-se que as estimativas destas probabilidades levem a um histograma semelhante ao apresentado acima. Com isto, pode-se concluir que o sistema estimou corretamente as probabilidades de ocorrência de cada fone.

A seguir são apresentadas as diversas estimações de $P(q_k)$ obtidas através do treinamento das sentenças no sistema híbrido ANN+HMM.

Segmentação manual

A Figura 5.23 apresenta as probabilidades das classes estimadas pelo sistema ao se utilizar uma rede neural de 108 entradas.

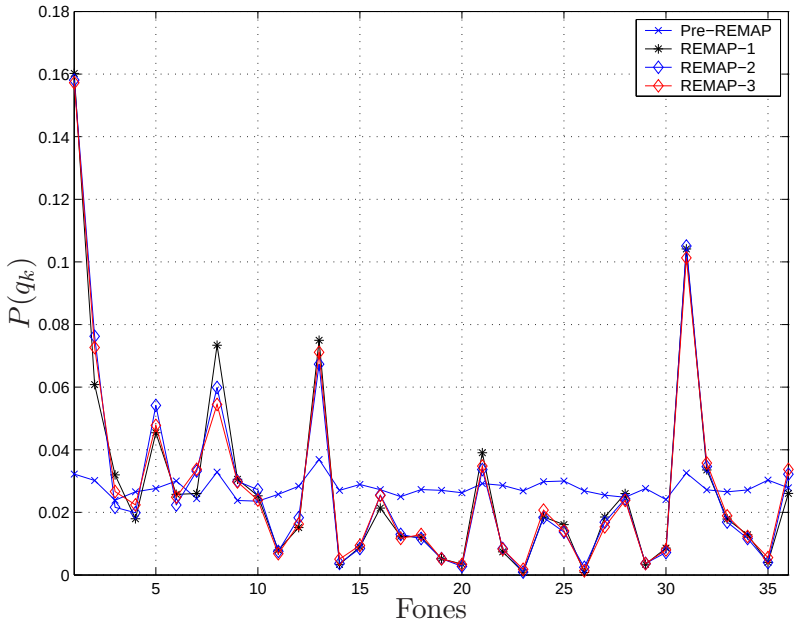

Figura 5.23: *Probabilidades das classes estimadas por uma rede neural de 108 entradas, utilizando sentenças de treinamento submetidas à segmentação manual.*

Segmentação uniforme

Admitindo que os fones das sentenças de treinamento possuam a mesma duração dentro de uma dada frase, foram estimadas as probabilidades $P(q_k)$, que é apresentada pela Figura 5.24.

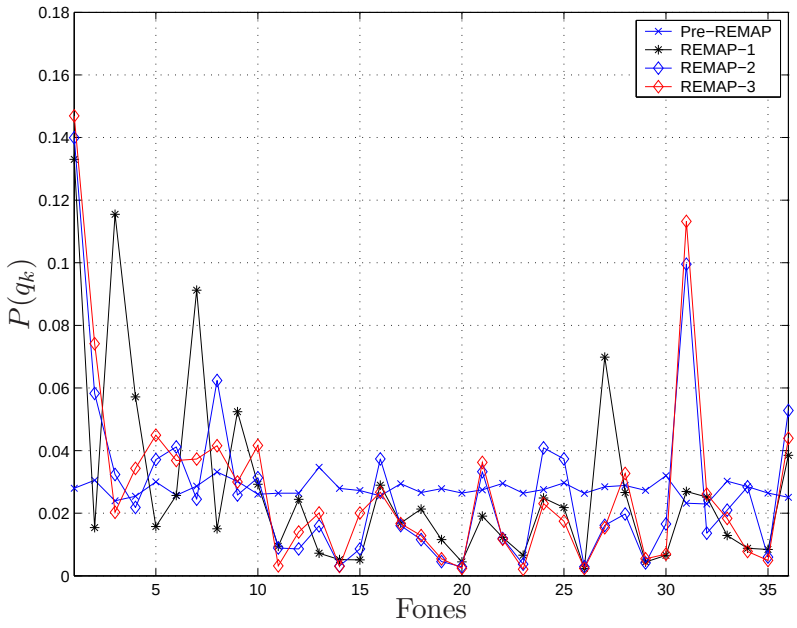

Figura 5.24: *Probabilidades das classes estimadas por uma rede neural de 108 entradas, utilizando sentenças de treinamento submetidas à segmentação uniforme.*

Simulação dos erros de segmentação manual

Do mesmo modo que foram apresentadas as probabilidades de transição de estados, abaixo têm-se os gráficos das probabilidades de classe estimadas para cada valor de desvio das marcas originais de segmentação manual.

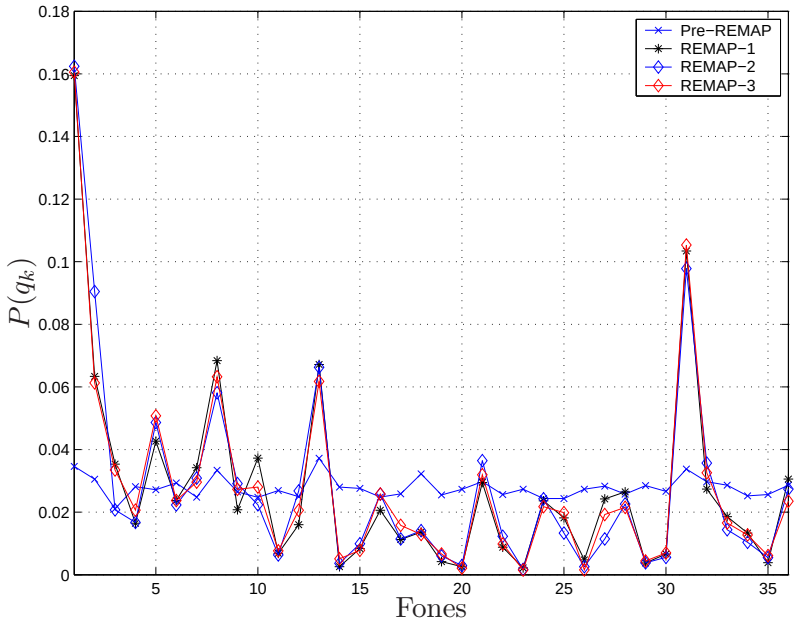

Figura 5.25: *Probabilidades das classes estimadas por uma rede neural de 108 entradas, utilizando desvio máximo de* ± 10 *ms.*

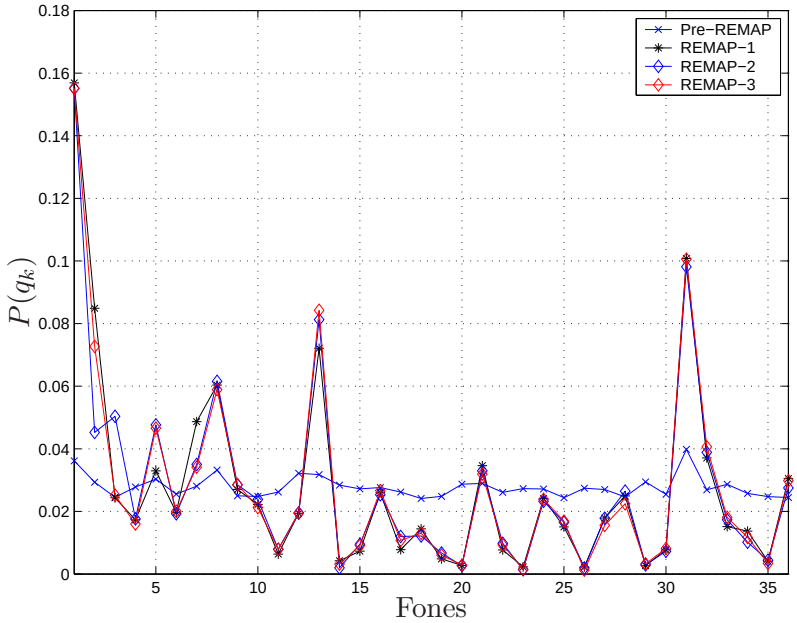

Figura 5.26: *Probabilidades das classes estimadas por uma rede neural de 108 entradas, utilizando desvio máximo de* ± 20 *ms.*

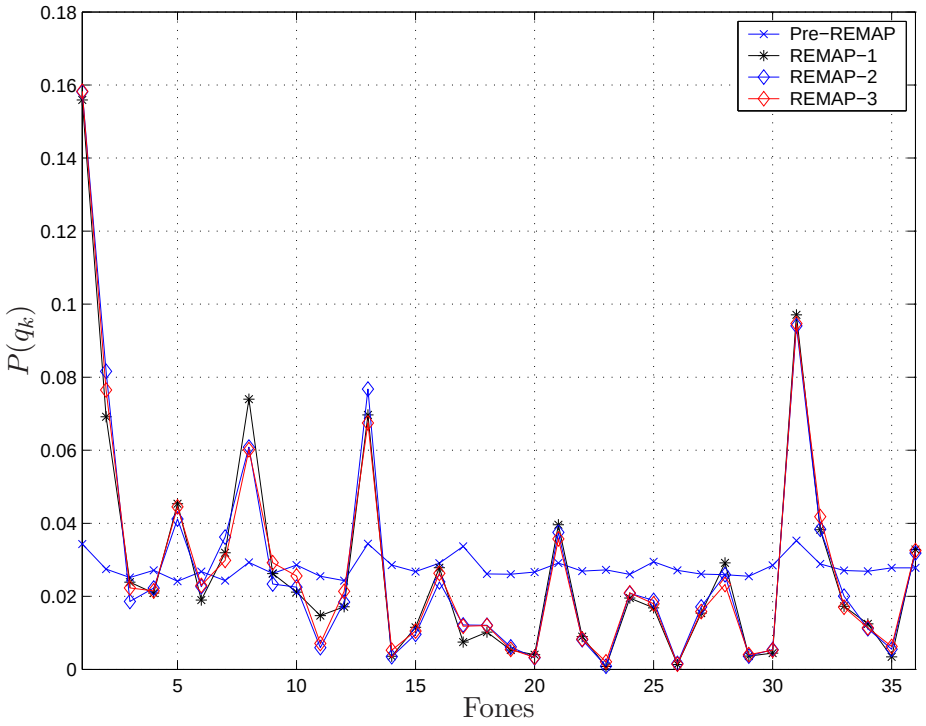

Figura 5.27: *Probabilidades das classes estimadas por uma rede neural de 108 entradas, utilizando desvio máximo de* ± 30 *ms.*

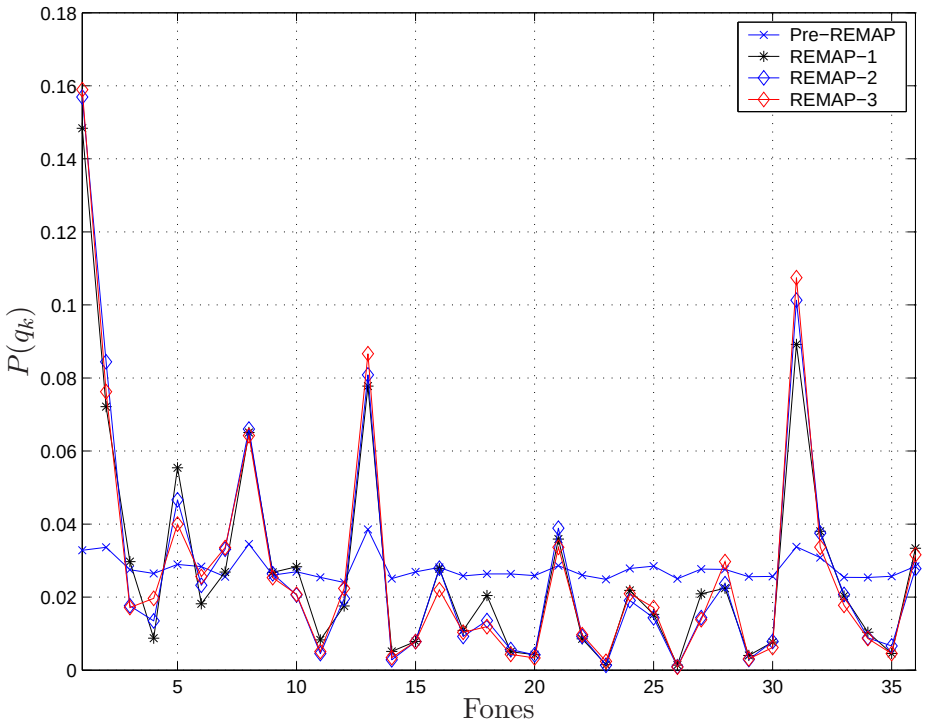

Figura 5.28: *Probabilidades das classes estimadas por uma rede neural de 108 entradas, utilizando desvio máximo de* ± 40 *ms.*

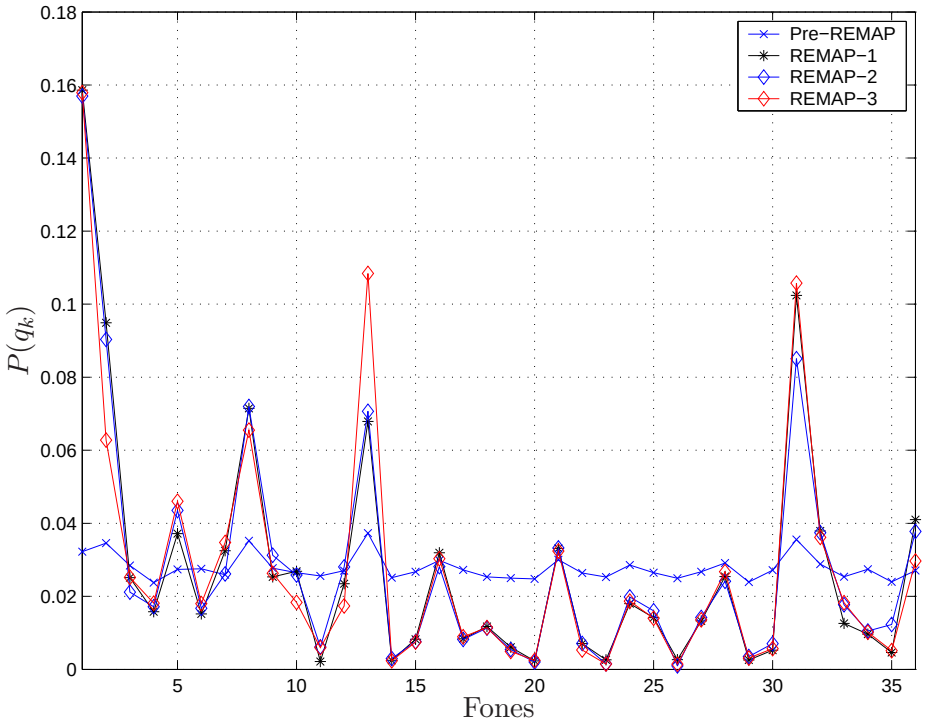

Figura 5.29: *Probabilidades das classes estimadas por uma rede neural de 108 entradas, utilizando desvio máximo de* ± 50 *ms.*

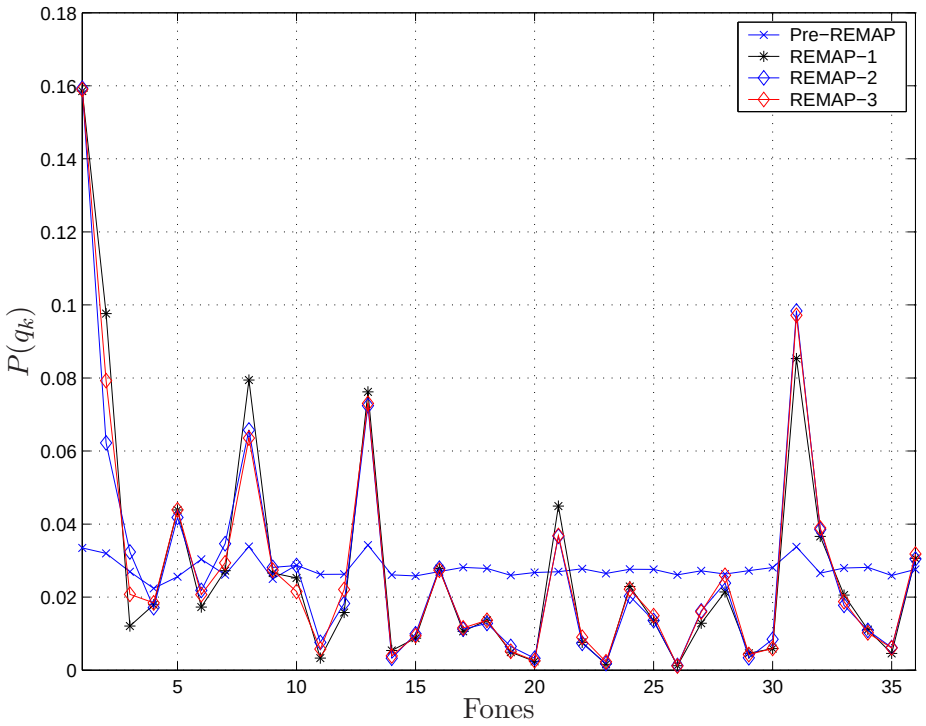

Figura 5.30: *Probabilidades das classes estimadas por uma rede neural de 108 entradas, utilizando desvio máximo de* ± 60 *ms.*

Discussão

Em todas as figuras referente ao avanço da estimação das probabilidades *a priori* das classes, notou-se que as estimativas referentes à etapa Pré-REMAP foram muito ruins. Isto se deve à pequena quantidade de exemplos de treinamento disponíveis nesta etapa, retirados apenas entre as marcas de segmentação.

Outra observação pertinente se refere ao formato aproximado dos gráficos obtidos pela segmentação manual e pelos erros de segmentação, comparados com o histograma da ocorrência dos fones, apresentados na figura 5.22. Isto mostra que as probabilidades de classe foram corretamente estimadas, ao contrário do que ocorreu com as estimações para segmentação uniforme.

Estas comparações também podem ser facilmente observadas através dos histogramas apresentados nas Figuras 5.31, 5.32, 5.33 e 5.34. A legenda *"ref."* indica a probabilidade de ocorrência dos fones, apresentado pela Figura 5.22, extraídos a partir dos dados fornecidos pela Tabela 4.1.

Os histogramas foram extraídos a partir do resultado da estimação de $P(q_k)$ ao final da etapa *REMAP-3*.

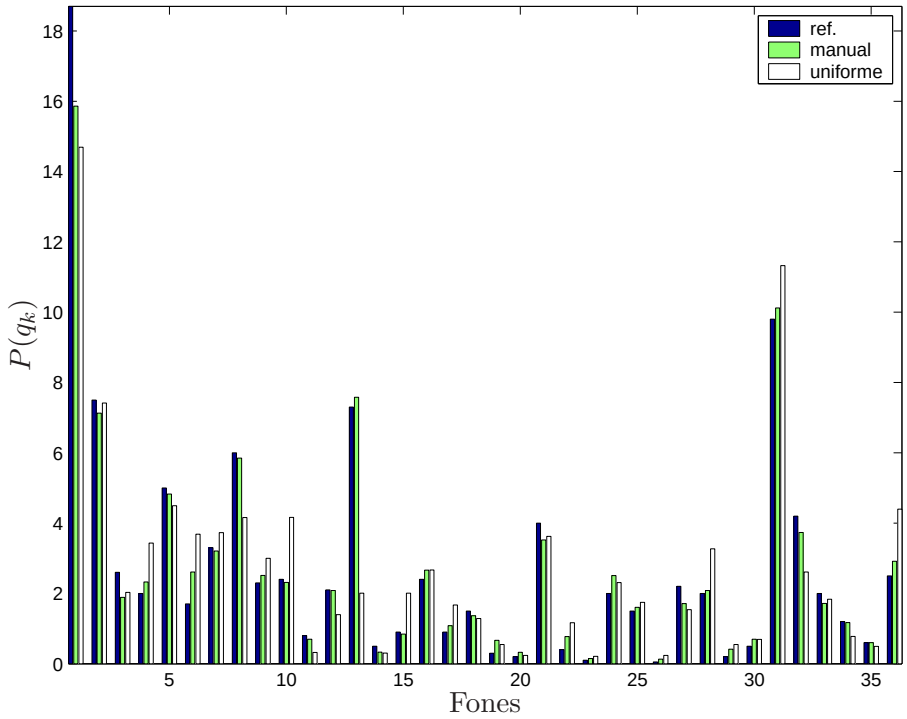

Figura 5.31: *Histograma comparativo das probabilidades de classe (referência, segmentação manual e segmentação uniforme).*

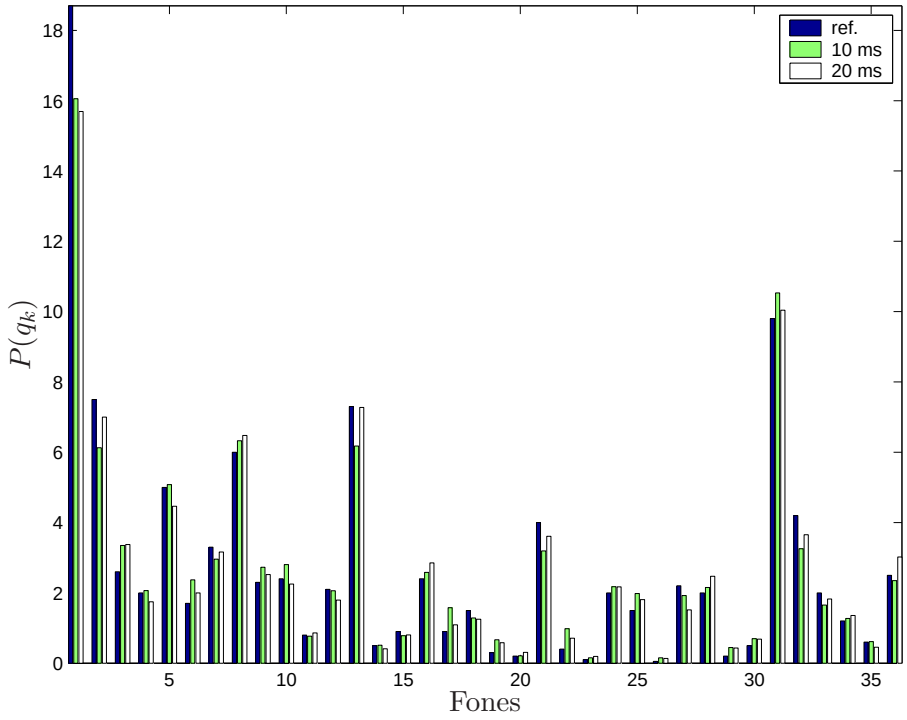

Figura 5.32: *Histograma comparativo das probabilidades de classe (referência, erro de segmentação de 10 ms e erro de segmentação 20 ms).*

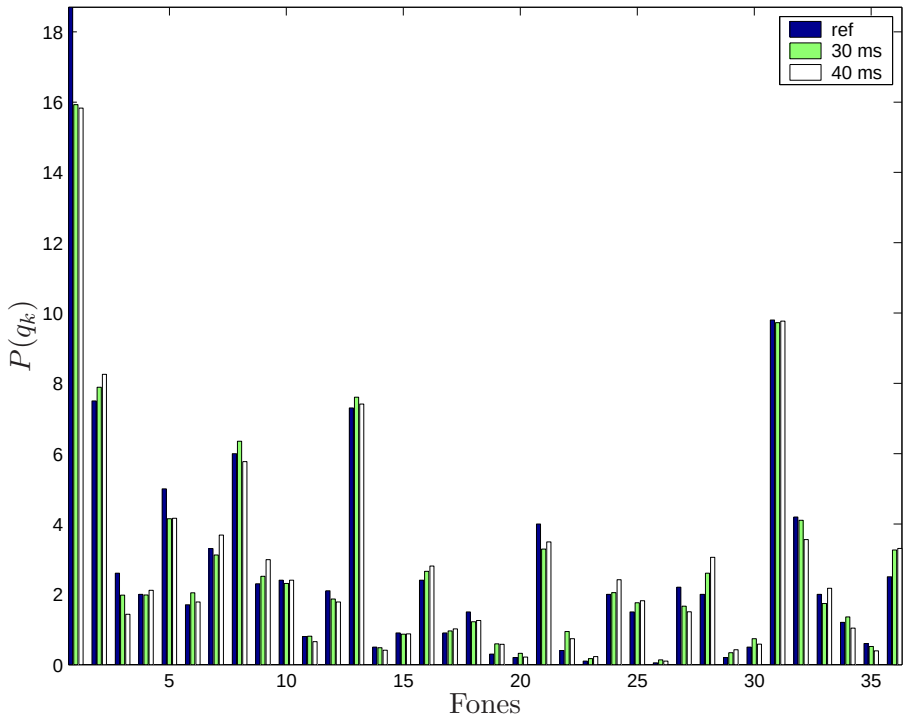

Figura 5.33: *Histograma comparativo das probabilidades de classe (referência, erro de segmentação de 30 ms e erro de segmentação 40 ms).*

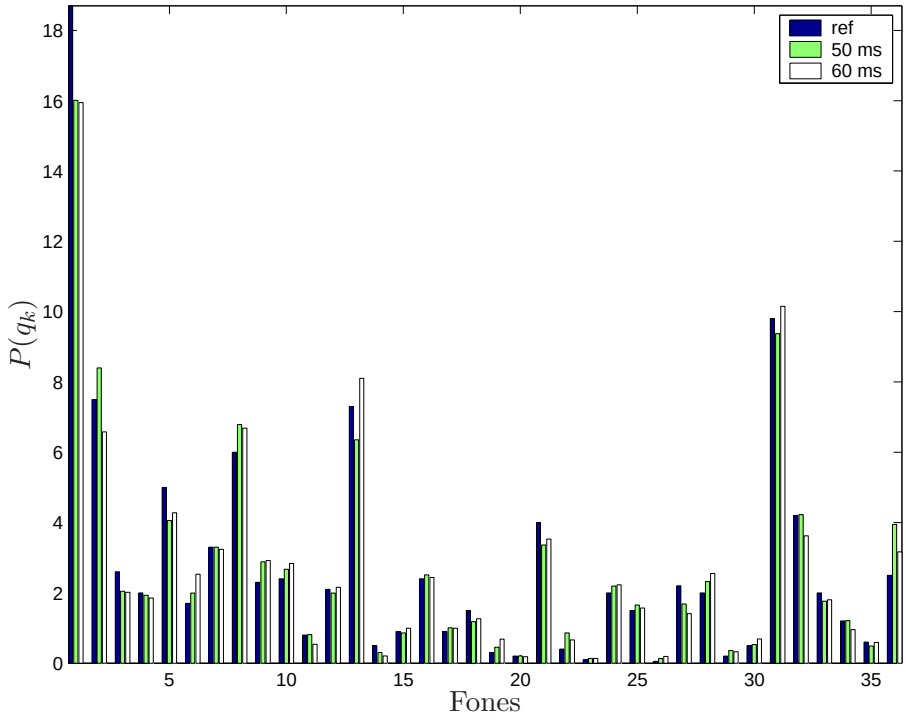

Figura 5.34: *Histograma comparativo das probabilidades de classe (referência, erro de segmentação de 50 ms e erro de segmentação 60 ms).*

Conclui-se que uma segmentação mal executada pode afetar negativamente os valores de $P(q_k)$, contribuindo para um desempenho ruim do sistema de reconhecimento de fala.

5.3.6 Erro médio quadrático da etapa REMAP

Esta subseção apresenta os gráficos de erro médio quadrático da etapa REMAP de treinamento dos modelos híbridos ANN+HMM, no intuito de mostrar a dinâmica do aprendizado adquirido pelo sistema, dentro da abordagem proposta por Konig [3]. Nos experimentos foi possível notar que a modelagem matemática do treinamento baseado em transições fez com que as redes neurais aprendessem corretamente as relações de entrada e saída, permitindo a maximização da probabilidade *a posteriori* do modelo correto, ao mesmo tempo que minimiza as probabilidades dos modelos rivais.

Em cada gráfico de erro médio quadrático apresentado nesta Subseção, será também especificado o valor da taxa de aprendizagem (η) com o qual este foi gerado. O parâmetro η é utilizado na correção dos pesos sinápticos da rede neural, executado de acordo com o algoritmo *Error Back Propagation* [12].

Segmentação manual

A Figura 5.35 apresenta a evolução do erro médio quadrático no treinamento de sentenças submetidas à segmentação manual, nos modelos híbridos ANN+HMM com uma rede neural de 84 entradas (gráfico à esquerda) e 108 entradas (gráfico à direita), respectivamente. Os pontos espaçados uniformemente no gráfico indicam o valor do erro ao final de cada etapa REMAP de treinamento.

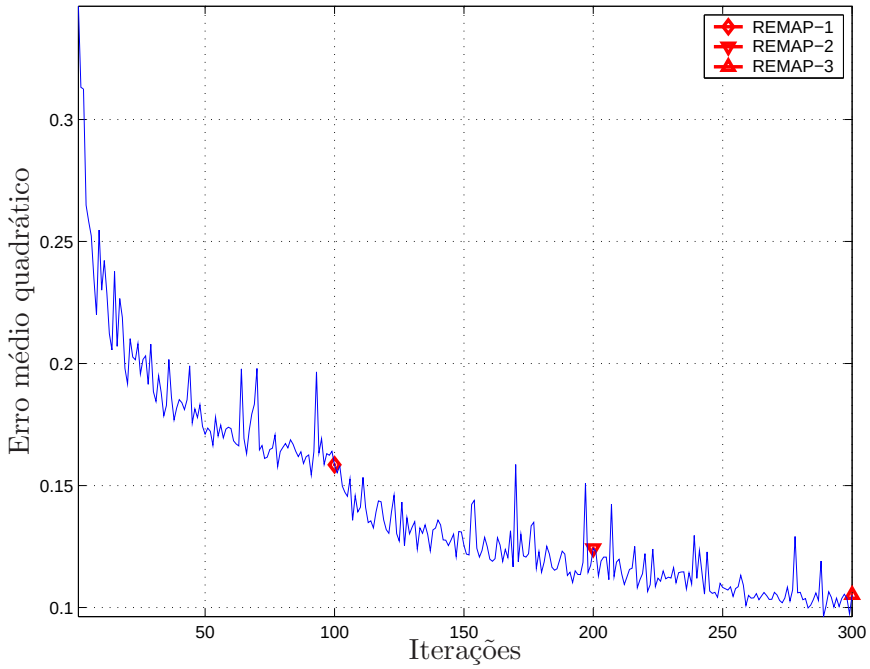

Figura 5.35: *Erro médio quadrático obtido através da rede neural de 108 entradas e passo de aprendizagem de 0,1.*

Como pode ser visto na figura acima, o comportamento da curva indica o aumento do aprendizado do sistema, via atualização das matrizes de pesos sinápticos, ao serem apresentadas as locuções de treinamento ao longo de cada iteração, possibilitando assim a estimativa cada vez mais precisa das probabilidades de transição de estados, probabilidades *a priori* das classes e verossimilhanças de emissão de símbolos.

Segmentação uniforme

Para as sentenças de treinamento submetidas à segmentação uniforme, o decaimento da curva de erro médio quadrático continua presente, conforme apresentado na Figura 5.36. Mas a suposição de uniformidade das durações dos fonemas em uma determinada sentença faz com que o desempenho do sistema caia abruptamente, comparado com os valores de *WER* obtidos com segmentação manual e com erros de segmentação manual.

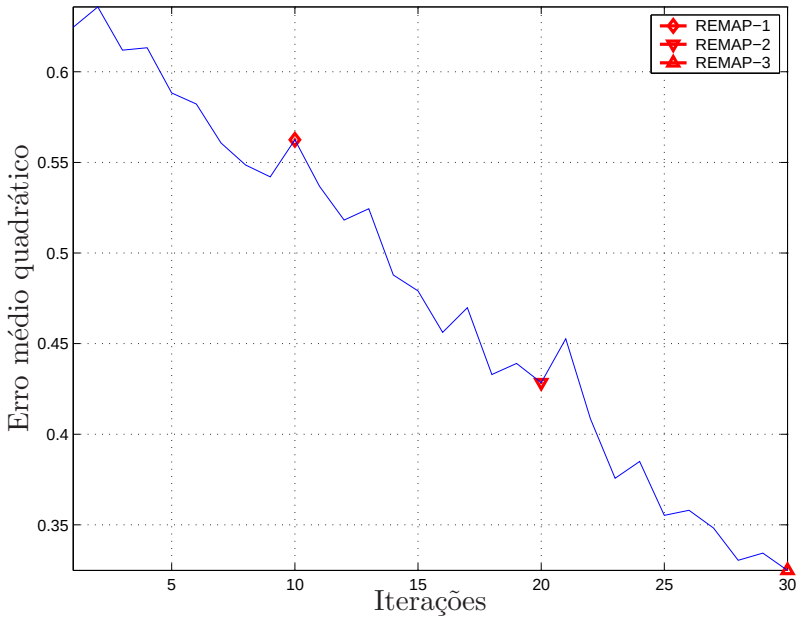

Figura 5.36: *Erro médio quadrático obtido através da rede neural de 108 entradas e passo de aprendizagem de 0,6.*

Simulação dos erros de segmentação manual

As próximas figuras apresentam o desempenho dos modelos híbridos ANN+HMM, em termos dos gráficos de erro médio quadrático, ao treinar sentenças submetidas aos erros de segmentação manual.

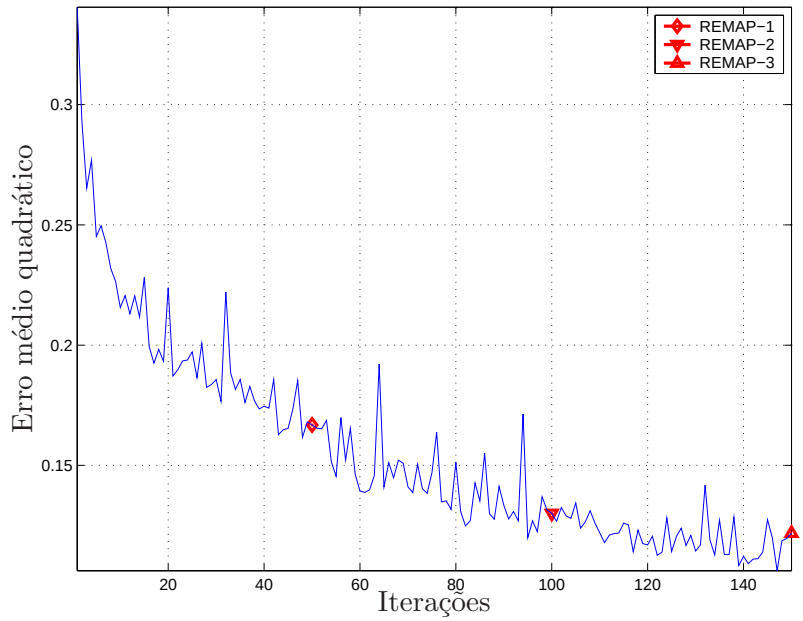

Figura 5.37: *Erro médio quadrático. Rede neural de 108 entradas com* $\eta = 0,1$. *Desvio máximo de 10 ms.*

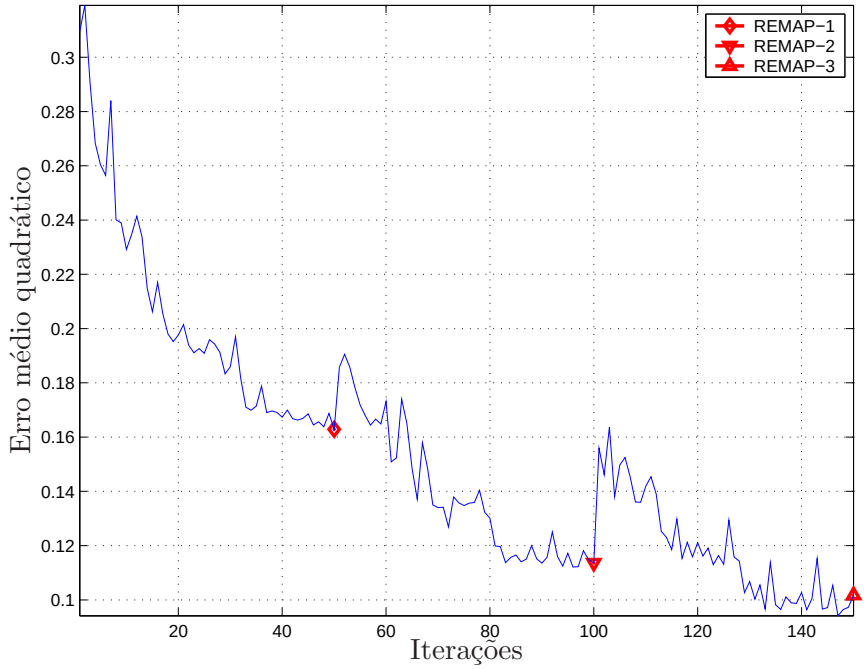

Figura 5.38: *Erro médio quadrático. Rede neural de 108 entradas com* $\eta = 0,1$. *Desvio máximo de 20 ms.*

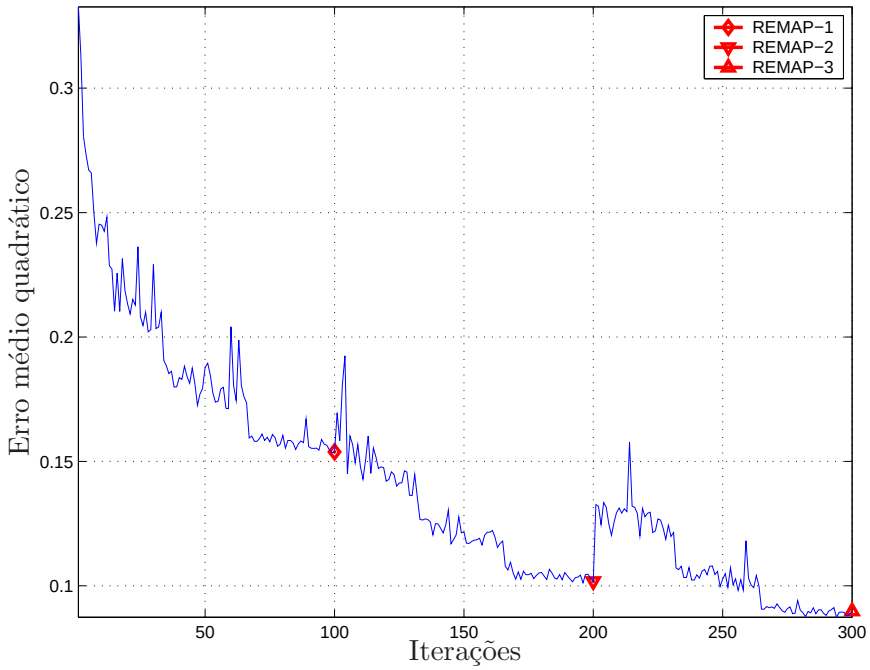

Figura 5.39: *Erro médio quadrático. Rede neural de 108 entradas com* $\eta = 0,3$. *Desvio máximo de 30 ms.*

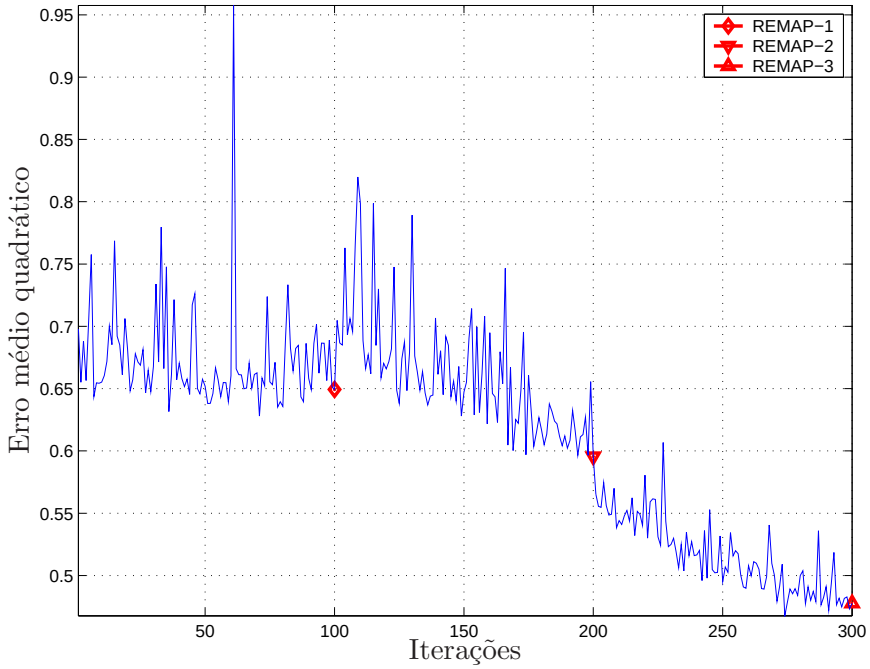

Figura 5.40: *Erro médio quadrático. Rede neural de 108 entradas com* $\eta = 0,1$*. Desvio máximo de 40 ms.*

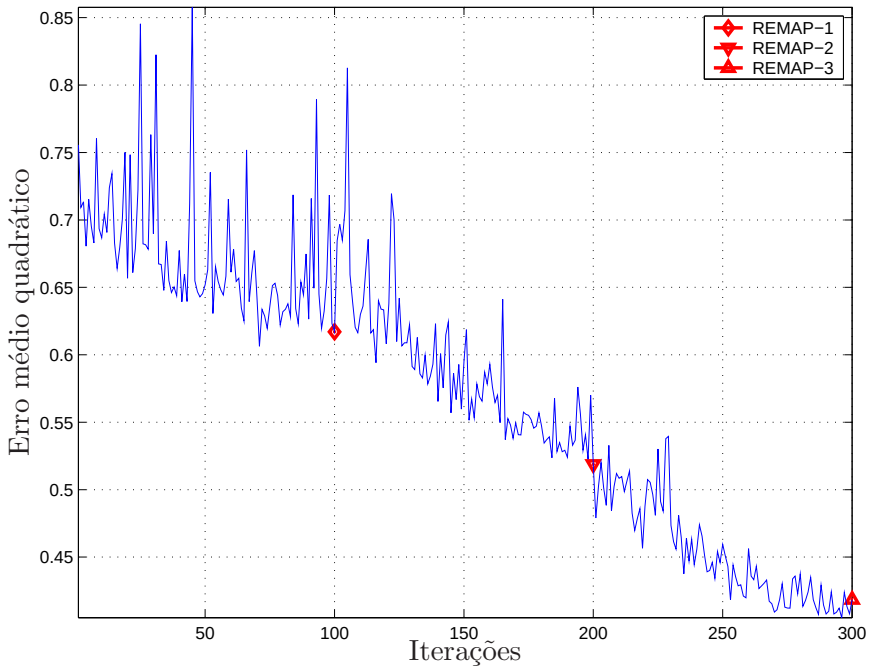

Figura 5.41: *Erro médio quadrático. Rede neural de 108 entradas com* $\eta = 0,1$*. Desvio máximo de 50 ms.*

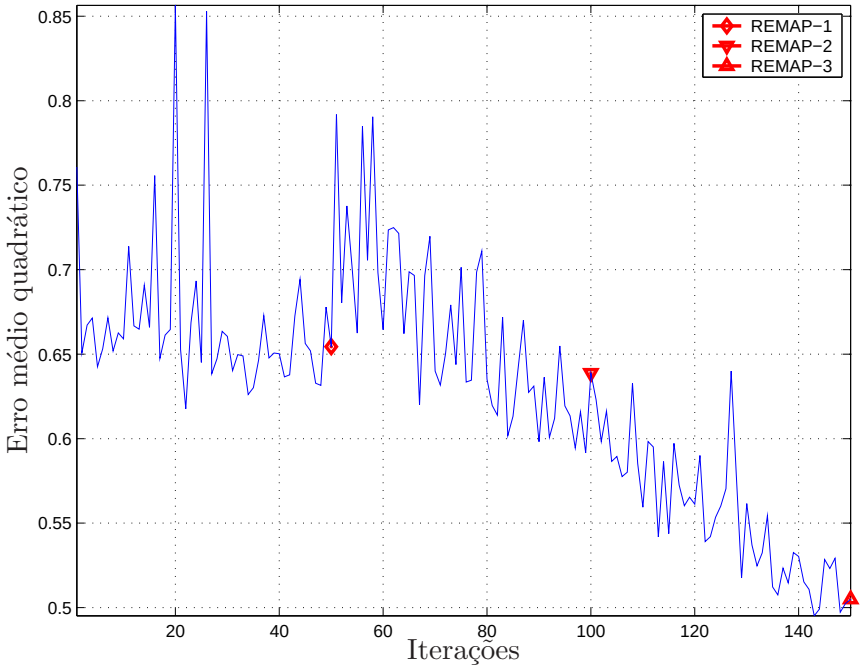

Figura 5.42: *Erro médio quadrático. Rede neural de 108 entradas com* $\eta = 0,1$*. Desvio máximo de 60 ms.*

Discussão

Notou-se um comportamento interessante nos gráficos de erro quadrático médio. Quando se inicia o treinamento dos modelos híbridos com alvos suaves recém estimados, os valores de erro quadrático médio tendem a ser piores que as últimas reestimações da etapa REMAP anterior. A explicação para tal fenômeno é que, com os novos valores de alvos suaves, a rede neural precisa se readaptar (via matrizes de pesos sinápticos), fazendo com que as saídas da rede convirjam para os novos valores de referência. Através do algoritmo de treinamento *EBP*, os valores de erro quadrático médio voltam a descrever uma tendência de queda ao longo das iterações.

5.4 Alvos suaves

Os alvos suaves, conforme apresentado no Capítulo 3, são calculados levando-se em consideração as mudanças graduais quando se passa de um fonema para outro, fazendo assim uma modelagem do efeito de coarticulação.

A Figura 5.43 apresenta a primeira reestimação dos alvos suaves para a sentença *"É suficiente"*, sendo esta submetida à segmentação manual. Pode-se perceber que os modelos híbridos ANN+HMM modelaram de maneira satisfatória as transições para esta primeira reestimação dos parâmetros da rede neural e dos HMM's. No caso das segunda e terceira reestimações, Figuras 5.44 e 5.45, nota-se uma perda das suavidades nas transições, o que poderia ser solucionado com a utilização de fones dependentes de contexto ou então a modelagem de cada fonema treinado através de mais estados.

Devido à suposição quanto às durações dos fones, na segmentação uniforme a reestimação dos alvos suaves apresentou faixas de valores diferente quanto às "durações" de cada sub-unidade. Isto já era esperado, pelo fato de serem atribuídos novos valores a cada fonema. Mas o fato negativo desta mudança é que os modelos híbridos não conseguiram estimar corretamente $P(M_i|\mathbf{X})$, a probabilidade do modelo M_i dado a seqüência $\mathbf{X}$ dos vetores acústicos. Isto é mostrado na Figura 5.46.

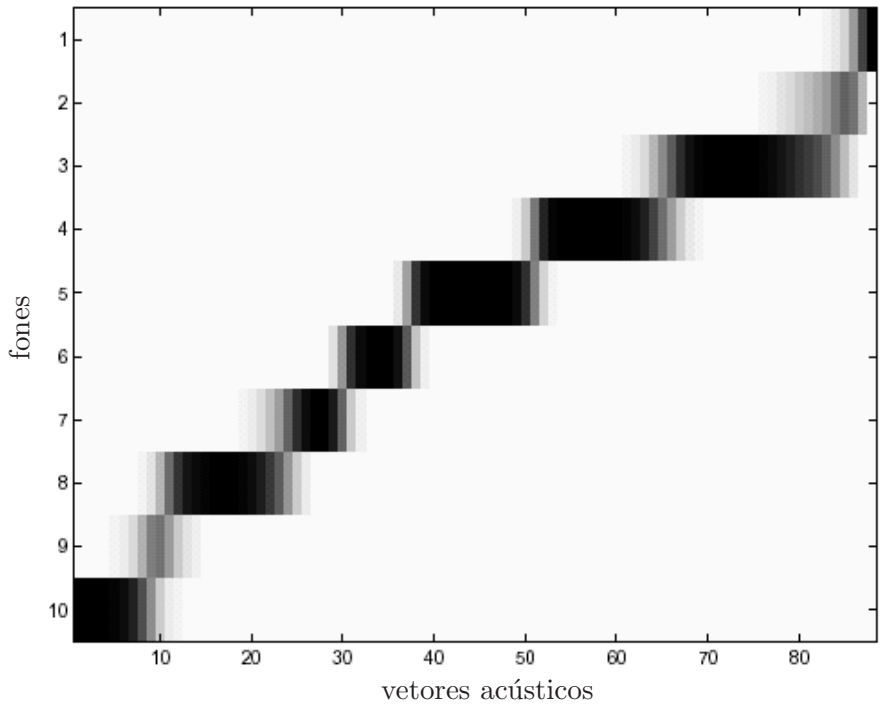

Figura 5.43: *Alvos suaves da sentença "É suficiente", na primeira reestimação dos parâmetros do sistema híbrido ANN+HMM.*

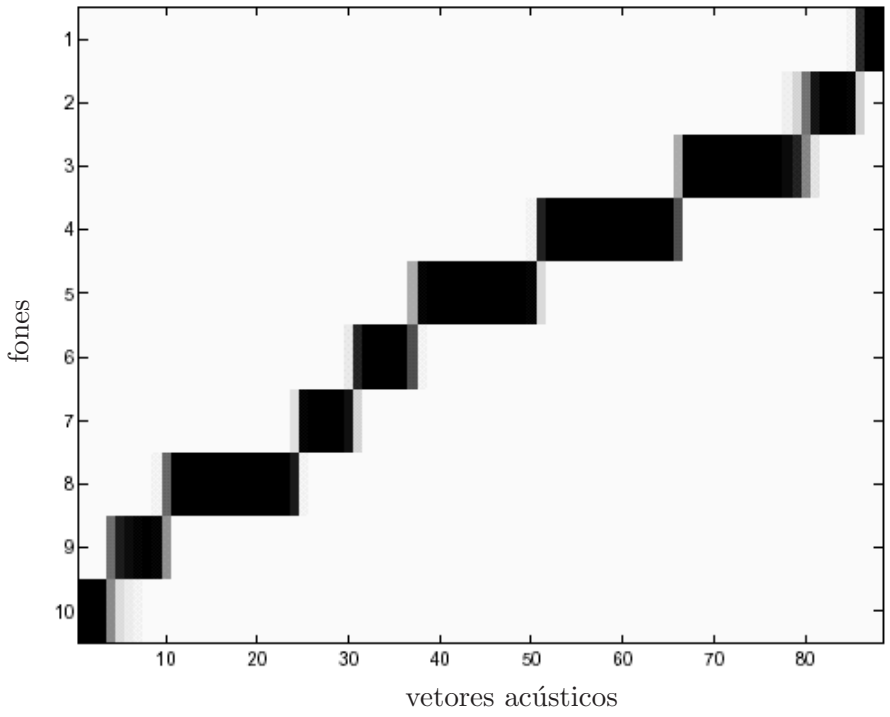

Figura 5.44: *Alvos suaves da sentença "É suficiente", na segunda reestimação dos parâmetros do sistema híbrido ANN+HMM.*

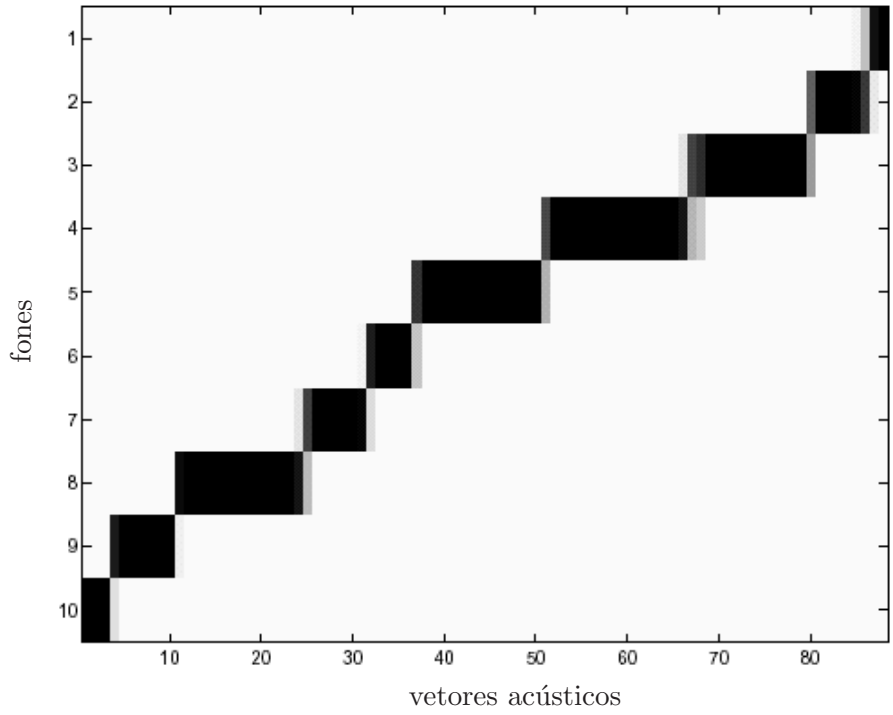

Figura 5.45: *Alvos suaves da sentença "É suficiente", na terceira reestimação dos parâmetros do sistema híbrido ANN+HMM.*

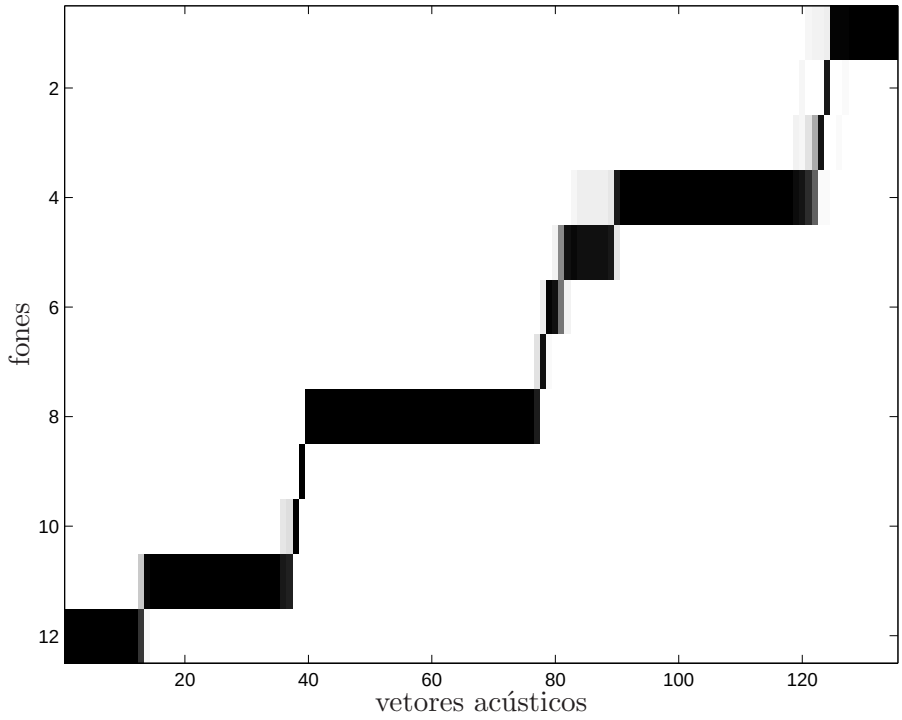

Figura 5.46: *Alvos suaves da sentença "É suficiente", obtido através do treinamento das sentenças submetidas à segmentação uniforme.*

Quanto aos erros de segmentação manual, nos gráficos de alvos suaves as durações dos fonemas se assemelham aos gráficos de segmentação manual. Nas Figuras 5.47, 5.48 e 5.49 estão ilustradas as dinâmicas dos alvos suaves para erro de segmentação de 30 milissegundos. Nestas referidas figuras nota-se que o modelo de um estado por fone não modela satisfatoriamente as transições suaves. Desta forma, alterando para um HMM de mais de um estado por fone e utilizando fones dependentes de contexto poderiam sanar esta dificuldade de modelamento das transições.

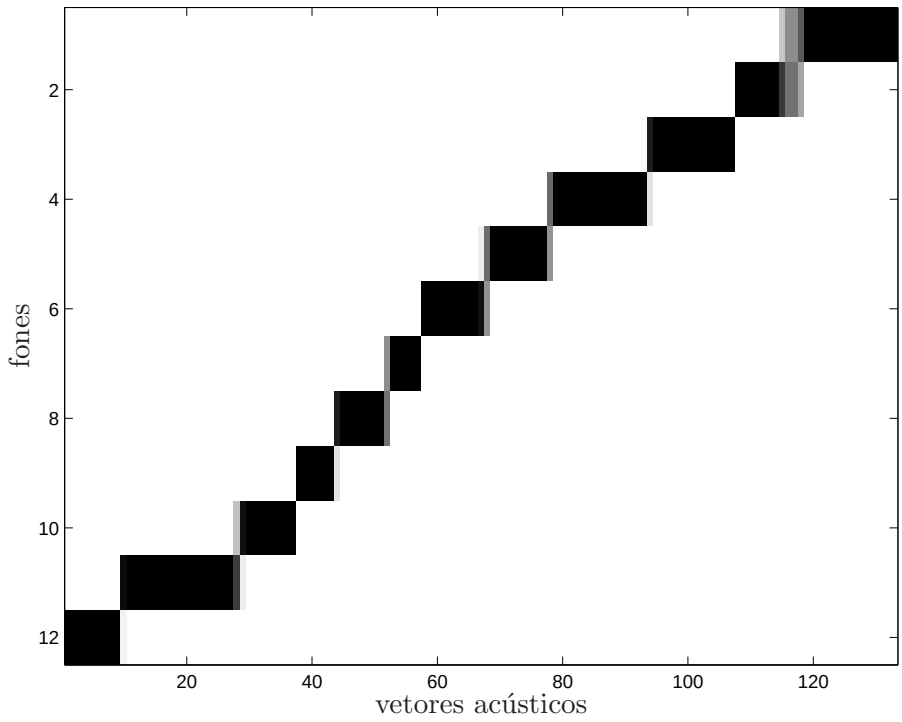

Figura 5.47: *Alvos suaves da sentença "É suficiente", na primeira reestimação dos parâmetros do sistema híbrido ANN+HMM. Para erros de segmentação manual de 30 ms.*

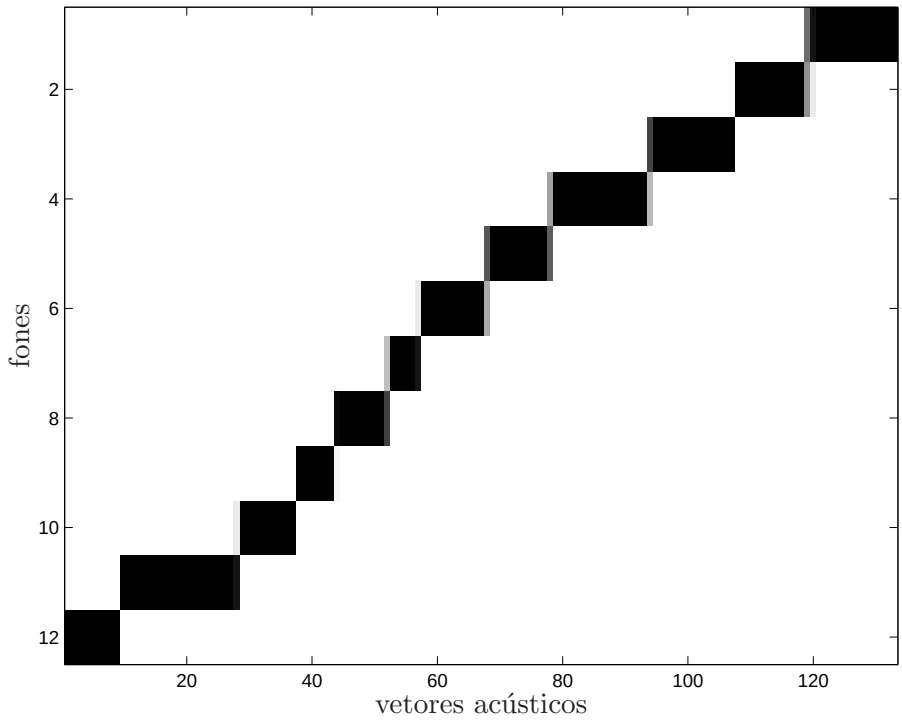

Figura 5.48: *Alvos suaves da sentença "É suficiente", na segunda reestimação dos parâmetros do sistema híbrido ANN+HMM. Para erros de segmentação manual de 30 ms.*

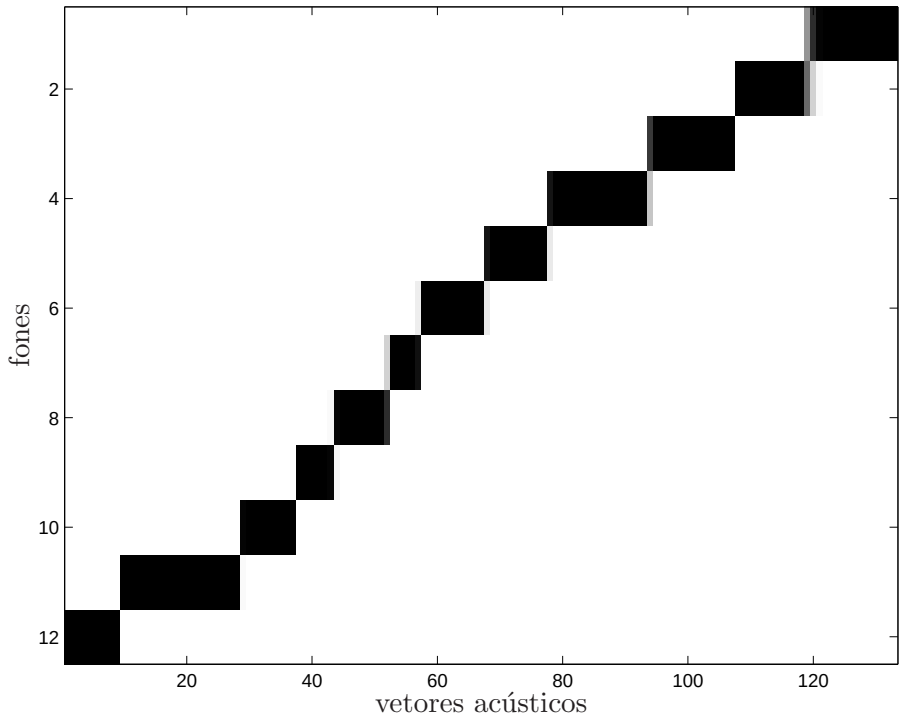

Figura 5.49: *Alvos suaves da sentença "É suficiente", na terceira reestimação dos parâmetros do sistema híbrido ANN+HMM. Para erros de segmentação manual de 30 ms.*

Capítulo 6

Conclusão

6.1 Considerações finais

Neste trabalho foi estudado e implementado um sistema de reconhecimento de fala contínua baseado nos modelos híbridos ANN+HMM com uso da técnica REMAP [2, 3] para reestimação dos alvos suaves segundo [15], no treinamento de sentenças submetidas à transcrição fonética e segmentação. Para o reconhecimento das sentenças utilizou-se o algoritmo *Level Building*, em conjunto com as técnicas de modelo de duração de palavras [17], gramática do tipo pares de palavras e critério de parada automática proposto em [5]. O NIST SCLITE [4, 18] foi a ferramenta que executou a contagem dos erros cometidos no reconhecimento, conseguindo assim as estatísticas de taxa de erros de palavras.

Os testes seguiram os objetivos apresentados na Seção 5.1. Desta forma, verificou-se o desempenho do sistema ao se utilizar no reconhecedor modelo de duração de palavras e restrição gramatical do tipo pares de palavras.

Foram conseguidos desempenhos similares, em relação à segmentação manual, para simulação de erros segmentação de até 30 ms. Conclui-se então que o sistema tolera, no máximo, erros de segmentação da ordem de 3 vetores acústicos. Pode-se também observar que, caso a base de dados seja segmentada automaticamente, deve-se assegurar de que o segmentador erre na atribuição das marcas de segmentação em, no máximo, 30 ms para que não comprometa o desempenho do sistema estudado.

Como esperado, a segmentação uniforme apresentou os piores resultados, fato este atribuído à suposição da duração dos fonemas de uma sentença possuírem mesmo valor. Isto dificultou a estimação correta dos parâmetros da rede neural e dos HMM's, fazendo com que na avaliação do reconhecimento algumas taxas de erros de palavras superassem o patamar de 100%. Isto foi possível devido à grande quantidade de erros de substituição e inserção cometidos ao longo das

sentenças reconhecidas.

No caso da simulação dos erros de segmentação, verificou-se que o desempenho do sistema permanece relativamente inalterado para erros de segmentação de até 30 milissegundos. Isto indica que os modelos híbridos ANN+HMM são bastante sensíveis à segmentação das locuções de treinamento.

6.2 Propostas para trabalhos futuros

Como sugestões para trabalhos futuros, vale citar:

- Utilização de três estados por fone e fones dependentes de contexto no sistema estudado nesta Dissertação. Com três estados abre-se a possibilidade de modelar o início, o meio e o fim de cada fone. Os fones dependentes de contexto fazem com que na modelagem seja levada em consideração a posição deste em relação aos seus fones vizinhos.

- Implementação de modelos híbridos, utilizando Máquinas de Vetor de Suporte (SVM/HMM) ou outras arquiteturas de Redes Neurais Artificiais (por exemplo, redes de Kohonen).

- Implementação de sistema baseado em GMM (*Gaussian Mixture Model*).

- Implementação de modelos híbridos ANN+HMM utilizando algoritmos discriminativos para modelamento temporal (por exemplo, *Segmental GPD*), ao invés do algoritmo *Baum-Welch* utilizado nesta Dissertação.

- Verificação do desempenho do sistema estudado em bases de dados maiores (por exemplo, TIMIT).

- Utilização de gramática *Bi-gram* ou *Tri-gram*.

- Implementação de um segmentador automático, no intuito de acelerar o processo de segmentação de grandes bases de dados.

Apêndice A

Frases utilizadas na base de dados

A seguir são apresentadas as frases que compõem a base de dados.

1. A cotação do dólar aumentou e as bolsas fecharam em baixa.
2. A cotação do dólar aumentou, mas as bolsas fecharam em baixa.
3. A bolsa ficará estável ou sofrerá uma pequena queda.
4. Não haverá ajustes, nem modificações radicais no plano.
5. Foi detectado um problema em seu cartão, ele deve ser substituído.
6. É necessário que o convênio permita o intercâmbio.
7. Posso afirmar-lhes que o convênio permite o intercâmbio.
8. O convênio que foi assinado recentemente permite o intercâmbio.
9. O convênio permite o intercâmbio porque visa a integração entre alunos de culturas diferentes.
10. O convênio permite o intercâmbio quando se trata de universidades vinculadas ao projeto de integração.
11. O convênio, que foi assinado recentemente, permite o intercâmbio.
12. O convênio assinado na última reunião é mais interessante do que o anterior.
13. A medida que o tempo passa, mais nos convencemos da eficácia do convênio.
14. Se o convênio permite o intercâmbio, devemos aproveitar a oportunidade para desenvolver o projeto de integração.

15. Localizado a uma quadra do centro da cidade, o condomínio permite que você una trabalho e conforto.
16. É suficiente.
17. Isto é suficiente.
18. O saldo é suficiente.
19. O saldo de sua conta é suficiente.
20. O saldo disponível é insuficiente.
21. O saldo disponível em sua conta é insuficiente.
22. Isto parece insuficiente.
23. O saldo parece ser insuficiente.
24. O saldo sempre está disponível.
25. O saldo está sempre disponível no início do mês.
26. No início do mês, o saldo está disponível.
27. Esta é a última chamada para o vôo sete três sete da Rio Sul.
28. Isto é uma pesquisa de opinião pública.
29. O valor de sua conta telefônica é baixo.
30. É de trinta mil cruzeiros, o valor de sua conta telefônica.
31. O vencimento de sua prestação será no dia quatro de junho.
32. O preço aumentou.
33. O preço do café aumentou.
34. O preço do café expresso aumentou.
35. O preço do café aumentou consideravelmente.
36. O preço do café aumentou consideravelmente na semana passada.
37. Aumentou o preço do café.
38. As taxas de juros no mercado interno estão subindo bastante.

39. As contas chegaram atrasadas.

40. As contas chegaram muito atrasadas ontem.

41. As contas telefônicas deste mês chegaram muito atrasadas ao banco.

42. Ontem, as contas chegaram aqui muito atrasadas.

43. Chegaram atrasadas.

44. Chegaram atrasadas todas as contas telefônicas deste mês.

45. O governo aumentou o imposto no mês passado.

46. O governo aumentou o imposto sobre importação.

47. O governo entregou os formulários aos contribuintes.

48. O governo entregou aos contribuintes os formulários.

49. O banco colocará a sua disposição o novo cheque.

50. A conta telefônica em nome de Adelaide Barroso, terá vencimento amanhã.

51. A perda da atratividade das aplicações em caderneta de poupança, está provocando um aumento de consumo no país.

52. O mercado foi considerado inadequado.

53. O mercado foi considerado inadequado pelos analistas.

54. O mercado foi considerado inadequado naquele momento.

55. Naquele momento, o mercado foi considerado inadequado pelos analistas das melhores instituições de pesquisa.

56. Diariamente.

57. Curva perigosa.

58. Dia vinte do sete.

59. Sim.

60. Não.

61. Saldo: vinte e cinco reais.

62. Estação Santa Cruz.

63. Passageiros com destino a São Paulo, Recife e Fortaleza, embarque imediato, portão sete, última chamada.

64. Os bancos atrás de mais eficiência.

65. Descontos de até cinquenta porcento.

66. Número incompleto.

67. Vinte e cinco.

68. Vinte cinco reais.

69. Cento e vinte cinco.

70. Cento e vinte e cinco reais.

71. Quatrocentos e quarenta e nove.

72. Dois mil, cento e vinte e cinco.

73. Dezesseis mil e quinhentos.

74. Oitenta milhões, trezentos e sessenta mil e duzentos e setenta e um.

75. A Telebrás, a empresa de telecomunicações brasileira, está investindo em pesquisa.

76. A Telebrás, uma empresa estatal, está investindo em pesquisa.

77. Tivemos recentemente a seguinte notícia, a Telebrás passará a investir mais em pesquisa.

78. Telesp informa, dezenove horas e trinta minutos.

79. Empresário, é preciso antecipar o futuro.

80. Prezado cliente, aguardaremos o seu comparecimento.

81. O código foi registrado pelo funcionário.

82. O convênio, um documento de trinta páginas, tem permitido o intercâmbio.

83. Os caixas eletrônicos não aceitarão depósitos.

84. Os caixas eletrônicos não aceitarão mais depósitos a partir das quinze horas.

85. Os caixas eletrônicos não aceitarão mais depósitos a partir das quinze horas, do próximo dia vinte e nove.

86. Os caixas eletrônicos não aceitarão mais depósitos a partir das quinze horas, e não estarão disponíveis para saques, a partir das dezessete horas do dia trinta.

87. Todos os bancos devem fazer a atualização cadastral até trinta e um de dezembro.

88. Todos os bancos devem fazer a atualização cadastral de seus clientes até trinta e um de dezembro.

89. Todos os bancos devem fazer a atualização cadastral de seus clientes até trinta e um de dezembro, de acordo com a determinação do banco central.

90. Todos os bancos devem fazer a atualização cadastral de seus clientes até trinta e um de dezembro, de acordo com a determinação do banco central, em cumprimento a resolução dois mil e vinte e cinco, do conselho monetário nacional.

91. Todos os bancos instalados no Brasil, devem fazer a atualização cadastral de seus clientes até trinta e um de dezembro, de acordo com a determinação explícita do banco central, em cumprimento a resolução dois mil e vinte e cinco, do conselho monetário nacional.

92. As operações continuam.

93. As operações de crédito continuam.

94. As operações de crédito e financiamento continuam a seguir as regras.

95. As operações de crédito e financiamento continuam a seguir as regras do Banco Central.

96. As operações de crédito e financiamento corrigidas por outros indexadores, continuam a seguir as regras do Banco Central, ainda em vigor.

97. As operações de crédito e financiamento corrigidas por outros indexadores, continuam normalmente a seguir as regras do Banco Central, ainda em vigor, beneficiando assim a maioria.

98. O aumento no consumo devido ao plano real impulsionou os preços.

99. Segundo dados do IBGE, o aumento no consumo devido ao plano real impulsionou os preços.

100. Segundo dados do IBGE, o aumento no consumo devido ao plano real tem impulsionado consideravelmente os preços nas últimas semanas.

Apêndice B

Vocabulário de reconhecimento

A seguir é apresentado o vocabulário de reconhecimento. As informações de média e desvio padrão da duração de cada palavra foram obtidas pela segmentação manual.

***fonemas**

#

a

A

an

e

E

en

i

in

o

O

on

u

un

b

d

D

f

g

j

k

l

L

m

n

N

p

r

rr

R

s

t

T

v

x

z

***fim**

***vocab**

, / # / 296.41908 / 144.87533

A / a / 70.8065 / 38.4451

ACEITARÃO / a s e t a r an u / 552.2500 / 53.8973

ACORDO / a k o r d u / 433.6667 / 35.5715

ADELAIDE / a d e l a i d i / 463 / 154.3333

AFIRMAR / a f i r m a R / 467 / 155.6667

AGUARDAREMOS / a g u a R d a r en m u z / 712 / 237.3333

AINDA / a in d A / 354 / 1.4142

AJUSTES / a j u s t i s / 628 / 209.3333

ALUNOS / a l un n u z / 432 / 144

AMANHÃ / a m an N an / 383 / 127.6667

ANALISTAS / a n a l i s t A s / 706 / 161.2203

ANTECIPAR / an t e s i p a r / 633 / 211

ANTERIOR / an t e r i o R / 579 / 193

AO / a u / 89.8000 / 11.1669

AOS / a u s / 168 / 49.4975

APLICAÇÕES / a p l i k a s on i z / 623 / 207.6667

APROVEITAR / a p r o v e i t a r / 635 / 211.6667

AQUI / a k i / 332 / 110.6667

AS / a s / 152.7778 / 21.7838

ASSIM / a s in / 340 / 113.3333

ASSINADO / a s i n a d u / 454.6667 / 32.5013

ATÉ / a t E / 247.8333 / 22.5337

ATRASADAS / a t r a z a d A s / 773.5000 / 108.2603

ATRÁS / a t r a z / 415 / 138.3333

ATRATIVIDADE / a t r a T i v i d a D i / 933 / 311

ATUALIZAÇÃO / a t u a l i z a s an u / 776 / 56.4447

AUMENTO / a u m en t u / 412.7500 / 23.4858

AUMENTOU / a u m en t o u / 521.3000 / 52.6098

BAIXA / b a i x A / 437.5000 / 9.1924

BAIXO / b a i x u / 499 / 166.3333

BANCO / b an k u / 290.6250 / 57.9604

BANCOS / b an k u z / 407.6667 / 30.3557

BARROSO / b a rr o z u / 524 / 174.6667

BASTANTE / b a s t an T i / 649 / 216.3333

BENEFICIANDO / b e n e f i s i an d u / 642 / 214

BOLSA / b o s A / 381 / 127

BOLSAS / b o s A s / 420.5000 / 0.7071

BRASIL / b r a z i u / 464 / 154.6667

BRASILEIRA / b r a z i l e r A / 596 / 198.6667

CADASTRAL / k a d a s t r a u / 576.6000 / 17.0968

CADERNETA / k a d e r n e t A / 527 / 175.6667

CAFÉ / k a f E / 368.2000 / 49.3579

CAIXAS / k a i x a z / 405.7500 / 29.0904

CARTÃO / k a R t an u / 489 / 163

CENTO / s en t u / 380.6667 / 23.7136

CENTRAL / s en t r a u / 500.8333 / 24.9272

CENTRO / s en t r u / 389 / 129.6667

CHAMADA / x a m a d A / 498.5000 / 28.9914

CHEGARAM / x e g a r an u / 429.3333 / 45.3373

CHEQUE / x E k i / 500 / 166.6667

CIDADE / s i d a D i / 520 / 173.3333

CINCO / s in k u / 404.6250 / 77.2971

CINQUENTA / s in k u en t A / 510 / 170

CLIENTE / k l i en T i / 506 / 168.6667

CLIENTES / k l i en T i z / 514.2500 / 8.4607

CÓDIGO / k O d i g u / 502 / 167.3333

COLOCARÁ / k o l o k a r / 402 / 134

COM / k on / 118.7500 / 14.1038

COMPARECIMENTO / k on p a r e s i m en t u / 954 / 318

CONDOMÍNIO / k on d o m i n i u / 515 / 171.6667

CONFORTO / k on f o R t u / 650 / 216.6667

CONSELHO / k on s e L u / 514.5000 / 28.9914

CONSIDERADO / k on s i d e r a d u / 586 / 37.8506

CONSIDERAVELMENTE / k on s d e r a v e u m en T i / 1009.6667 / 74.7418

CONSUMO / k on s un m u / 525.7500 / 22.4258

CONTA / k on t A / 402.8000 / 59.2132

CONTAS / k on t A s / 407.8000 / 36.8130

CONTINUAM / k on T i n u an u / 590.3333 / 53.9654

CONTRIBUINTES / k on t r i b u in T i s / 824 / 94.7523

CONVENCEMOS / k on v en s en m u z / 712 / 237.3333

CONVÊNIO / k on v e n i u / 501 / 51.7408

CORRIGIDAS / k o rr i j i d A s / 604 / 14.1421

COTAÇÃO / k o t a s an u / 521.5000 / 19.0919

CRÉDITO / k r E D i t u / 535.2000 / 32.7368

CRUZ / k r u s / 495 / 165

CRUZEIROS / k r u z e r u z / 607 / 202.3333

CULTURAS / k u t u r A z / 494 / 164.6667

CUMPRIMENTO / k un p r i m en t u / 618.5000 / 20.5061

CURVA / k u r v A / 325 / 108.3333

DA / d A / 81.2500 / 21.6699

DADOS / d a d u z / 360.5000 / 36.0624

DAS / d A z / 210.5000 / 39.8635

DE / D i / 88.2703 / 18.3449

DEPÓSITOS / d e p O z t u s / 741 / 77.0281

DESCONTOS / d e s k on t u z / 572 / 190.6667

DESENVOLVER / d e s en v o v e r / 704 / 234.6667

DESTE / d e s T i / 289.5000 / 6.3640

DESTINO / d e s T in / 358 / 119.3333

DETECTADO / d e t e k t a d / 566 / 188.6667

DETERMINAÇÃO / d e t e r m i n a s an u / 769.3333 / 5.1316

DEVE / d E v i / 253 / 84.3333

DEVEM / d E v en / 310 / 30.7490

DEVEMOS / d e v en m u z / 449 / 149.6667

DEVIDO / d e v i d / 343.3333 / 42.5480

DEZEMBRO / d e z en b r u / 523.8000 / 37.2317

DEZENOVE / d e z en n O v i / 501 / 167

DEZESSEIS / d e z e s e z / 544 / 181.3333

DEZESSETE / d e z e s E T i / 566 / 188.6667

DIA / D i A / 208.2500 / 42.5784

DIARIAMENTE / D i a r i a m en T i / 856 / 285.3333

DIFERENTES / D i f e r en T i s / 805 / 268.3333

DISPONÍVEIS / D i s p o n i v e s / 702 / 234

DISPONÍVEL / D i s p o n i v i u / 634 / 35.4753

DISPOSIÇÃO / D s p o z i s an u / 728 / 242.6667

DO / d u / 94.6800 / 23.6356

DOCUMENTO / d o k u m en t u / 533 / 177.6667

DOIS / d o i z / 221.3333 / 30.0222

DÓLAR / d O r / 305 / 36.7696

DUZENTOS / d u z en t u z / 641 / 213.6667

E / i / 56.2414 / 19.1883

É / E / 128 / 53.5147

EFICÁCIA / e f i k a s i A / 654 / 218

EFICIÊNCIA / e f i s i en s i A / 743 / 247.6667

ELE / e l i / 158 / 52.6667

ELETRÔNICOS / e l e t r on n i k u s / 745.7500 / 22.0964

EM / en / 89.1538 / 17.2233

EMBARQUE / en b a r k / 474 / 158

EMPRESA / e in p r e z A / 498 / 56.5685

EMPRESÁRIO / en p r e z a r i u / 700 / 233.3333

ENTRE / en t r i / 269 / 89.6667

ENTREGOU / en t r e g o u / 445.5000 / 61.5183

ESTA / E s t / 315 / 105

ESTÁ / e s t A / 282.5000 / 67.3817

ESTAÇÃO / e s t a s an u / 572 / 190.6667

ESTARÃO / i s t a r an u / 396 / 132

ESTATAL / e s t a t a u / 592 / 197.3333

ESTÃO / i s t an u / 300 / 100

ESTÁVEL / e s t A v i u / 553 / 184.3333

EXPLÍCITA / e s p l i s i t A / 647 / 215.6667

EXPRESSO / s p r E s u / 562 / 187.3333

FAZER / f a z e r / 379 / 23.1193

FECHARAM / f e x a r an / 424.5000 / 20.5061

FICARÁ / f i k a r A / 324 / 108

FINANCIAMENTO / f i n an s i a m en t u / 745.7500 / 28.2887

FOI / f o i / 189.1250 / 21.6428

FORMULÁRIOS / f o R m u l A r i u z / 736 / 117.3797

FORTALEZA / f o R t a l e z A / 706 / 235.3333

FUNCIONÁRIO / f un s on n a r i u / 665 / 221.6667

FUTURO / f u t u r u / 534 / 178

GOVERNO / g o v e r n u / 451.2500 / 25.9792

HAVERÁ / a v e r A / 345 / 115

HORAS / O r a z / 342.6000 / 83.4344

IBGE / i b e j e E / 563 / 12.7279

IMEDIATO / i m e D i a t u / 595 / 198.3333

IMPORTAÇÃO / i n p o r t a s an u / 734 / 244.6667

IMPOSTO / i n p o s t u / 538.5000 / 12.0208

IMPULSIONADO / i n p u s i on n a d u / 969 / 323

IMPULSIONOU / i n p u s i on n o u / 732 / 67.8823

INADEQUADO / i n n a d e k u a d u / 654.2500 / 41.6843

INCOMPLETO / i n k on p l E t u / 798 / 266

INDEXADORES / i n d e k s a d o r i s / 826 / 21.2132

INFORMA / i n f O R m A / 573 / 191

INÍCIO / i n i s u / 337 / 31.1127

INSTALADOS / i n s t a l a d u z / 545 / 181.6667

INSTITUIÇÕES / i n s T i t u i s on i z / 792 / 264

INSUFICIENTE / i n s u f i s i en T i / 950.5000 / 41.2513

INTEGRAÇÃO / i n t e g r a s an u / 752.6667 / 52.5769

INTERCÂMBIO / i n t e R k an b i u / 662 / 40.0214

INTERESSANTE / i n t e r e s an T i / 726 / 242

INTERNO / i n t E r n u / 463 / 154.3333

INVESTINDO / i n v e s t in n / 571 / 25.4558

INVESTIR / i n v e s T i R / 542 / 180.6667

ISTO / i s t u / 330.6667 / 45.0814

JUNHO / j un N u / 364 / 121.3333

JUROS / j u r u z / 330 / 110

-LHES/ L i s / 231 / 77

LOCALIZADO / l o k a l i z a d o / 682 / 227.3333

MAIORIA / m a i o r i A / 514 / 171.3333

MAIS / m a i z / 279.1429 / 41.6990

MAS / m a z / 202 / 67.3333

MEDIDA / m e D i d A / 360 / 120

MELHORES / m e L o r i z / 496 / 165.3333

MERCADO / m e r k a d u / 445 / 27.2489

MÊS / m e s / 364.2000 / 88.6211

MIL / m i u / 202 / 48.3694

MILHÕES / m i L on i s / 464 / 154.6667

MINUTOS / m in n u t u s / 617 / 205.6667

MODIFICAÇÕES / m o d i f i k a s on i z / 758 / 252.6667

MOMENTO / m o m en t u / 537.5000 / 6.3640

MONETÁRIO / m on n e t a r i u / 544.5000 / 6.3640

MUITO / m u in t / 316 / 14.7309

NA / n A / 117 / 48.0833

NACIONAL / n a s on n a u / 500 / 166.6667

NÃO / n an u / 177.5714 / 49.6617

NAQUELE / n a k e l i / 329 / 25.4558

NAS / n A z / 200 / 66.6667

NECESSÁRIO / n e s e s a r i u / 647 / 215.6667

NEM / n e i / 170 / 56.6667

NO / n u / 86.5455 / 15.3842

NOME / n on m i / 263 / 87.6667

NORMALMENTE / n o r m a u m en T / 624 / 208

NOS / n u s / 144 / 48

NOTÍCIA / n o T i s i A / 542 / 180.6667

NOVE / n O v i / 436.5000 / 6.3640

NOVO / n o v u / 254 / 84.6667

NÚMERO / n un m i r u / 360 / 120

O / u / 56.2364 / 17.4164

OITENTA / o i t en t A / 385 / 128.3333

ONTEM / on t e in / 403 / 24.0416

OPERAÇÕES / o p e r a s on i s / 620.3333 / 40.5496

OPINIÃO / o p i n i an u / 477 / 159

OPORTUNIDADE / o p o R t u n i d a D i / 864 / 288

OS / u s / 139.6667 / 29.0287

OU / o / 209 / 69.6667

OUTROS / o t r u s / 386 / 12.7279

PÁGINAS / p a j i n A s / 619 / 206.3333

PAÍS / p a i s / 398 / 132.6667

PARA / p a r A / 217 / 38.1576

PARECE / p a r E s / 435 / 145.6640

PARTIR / p a R T i r / 445.2500 / 26.8499

PASSA / p a s A / 511 / 170.3333

PASSADA / p a s a d A / 562 / 187.3333

PASSADO / p a s a d u / 626 / 208.6667

PASSAGEIROS / p a s a j e r u s / 563 / 187.6667

PASSARÁ / p A s A r / 348 / 116

PAULO / p A u l u / 345 / 115

PELO / p e l u / 217 / 72.3333

PELOS / p e l u z / 331.5000 / 2.1213

PEQUENA / p e k e n a / 368 / 122.6667

PERDA / p e R d A / 467 / 155.6667

PERIGOSA / p e r i g O z A / 732 / 244

PERMITA / p e R m i T / 452 / 150.6667

PERMITE / p e R m i T / 454.1429 / 41.7829

PERMITIDO / p e r m i T i d u / 517 / 172.3333

PESQUISA / p e s k i z A / 612.2000 / 43.3959

PLANO / p l an n u / 312 / 52.8078

POR / p u r / 152 / 11.3137

PORCENTO / p u R s en t u / 601 / 200.3333

PORQUE / p u R k e / 233 / 77.6667

PORTÃO / p o r t an u / 417 / 139

POSSO / p O s / 277 / 92.3333

POUPANÇA / p o u p an s A / 654 / 218

PRECISO / p r e s i z / 556 / 185.3333

PREÇO / p r e s u / 417 / 29.9666

PREÇOS / p r e s u s / 604.6667 / 129.6161

PRESTAÇÃO / p r e s t a s an u / 642 / 214

PREZADO / p r e z a d u / 490 / 163.3333

PROBLEMA / p r o b l en m / 409 / 136.3333

PROJETO / p r o j E t u / 562 / 46.6690

PROVOCANDO / p r o v o k an n / 496 / 165.3333

PRÓXIMO / p r O s i m u / 444 / 148

PÚBLICA / p u b l i k A / 591 / 197

QUADRA / k u a d r A / 457 / 152.3333

QUANDO / k u an d u / 278 / 92.6667

QUARENTA / k u a r en t / 398 / 132.6667

QUATRO / k u a t r u / 485 / 161.6667

QUATROCENTOS / k u a t r u s en t u s / 714 / 238

QUE / k i / 130.4286 / 34.3019

QUEDA / k E d A / 410 / 136.6667

QUINHENTOS / k in en t u s / 723 / 241

QUINZE / k in z i / 336 / 36.0416

RADICAIS / rr a d i k a i z / 485 / 161.6667

REAIS / rr e a i s / 597.3333 / 29.7377

REAL / rr e a u / 365.3333 / 37.2872

RECENTEMENTE / rr e s en T i m en T i / 784 / 97.3858

RECIFE / rr e s i f i / 547 / 182.3333

REGISTRADO / rr e j i s t r a d u / 635 / 211.6667

REGRAS / R E g r A s / 443.7500 / 117.8116

RESOLUÇÃO / rr e z o l u s an u / 621 / 48.0833

REUNIÃO / rr e u n i an u / 479 / 159.6667

RIO / R i u / 299 / 99.6667

SALDO / s a u d u / 421.6667 / 54.1364

SANTA / s an t A / 341 / 113.6667

SÃO / s an u / 230 / 76.6667

SAQUES / s a k i s / 522 / 174

SE / s i / 146.5000 / 7.7782

SEGUINTE / s e g in T i / 458 / 152.6667

SEGUIR / s e g i r / 385.7500 / 38.8705

SEGUNDO / s e g un d u / 410.5000 / 45.9619

SEMANA / s e m an n A / 348 / 116

SEMANAS / s e m an n A s / 640 / 213.3333

SEMPRE / s en p r / 400 / 21.2132

SER / s e r / 287.5000 / 2.1213

SERÁ / s e r a / 287 / 95.6667

SESSENTA / s e s en t A / 492 / 164

SETE / s E T i / 441.2500 / 69.8349

SETENTA / s e t en t / 440 / 146.6667

SEU / s e u / 216 / 86.2670

SEUS / s e u s / 252.5000 / 16.4215

SIM / s in / 365 / 121.6667

SOBRE / s o b r / 323 / 107.6667

SOFRERÁ / s o f r e r A / 467 / 155.6667

SUA / s u A / 222.3333 / 16.0707

SUBINDO / s u b in d u / 429 / 143

SUBSTITUÍDO / s u b i s t i u i d u / 793 / 264.3333

SUFICIENTE / s u f i s i en T i / 852.2500 / 32.6024

SUL / s u u / 320 / 106.6667

TAXAS / t a x A z / 399 / 133

TELEBRÁS / t e l e b r a s / 664.6667 / 83.7874

TELECOMUNICAÇÕES / t e l e k o m u n i k a s on i z / 980 / 326.6667

TELEFÔNICA / t e l e f on n i k / 699.6667 / 71.3045

TELEFÔNICAS / t e l e f on n i k A z / 698 / 25.4558

TELESP / t e l e s p / 536 / 178.6667

TEM / t e in / 175.5000 / 14.8492

TEMPO / t en p u / 296 / 98.6667

TERÁ / t e r a / 268 / 89.3333

TIVEMOS / T i v E m u z / 493 / 164.3333

TODAS / t o d A z / 279 / 93

TODOS / t o d u s / 356.8000 / 6.7602

TRABALHO / t r a b a L u / 507 / 169

TRATA / t r a t A / 447 / 149

TRÊS / t r e s / 289 / 96.3333

TREZENTOS / t r e z en t u z / 685 / 228.3333

TRINTA / t r in t A / 354 / 68.8259

ÚLTIMA / u T i m A / 331.6667 / 21.3854

ÚLTIMAS / u T i m A z / 300 / 100

UM / un / 111.3333 / 47.5894

UMA / u m A / 157 / 19.0263

UNA / u n a / 278 / 92.6667

UNIVERSIDADES / u n i v e r s i d a D i z / 768 / 256

VALOR / v a l o r / 347.5000 / 24.7487

VENCIMENTO / v en s i m en t u / 545 / 48.0833

VIGOR / v i g o r / 380 / 45.2548

VINCULADAS / v in k u l a d A z / 585 / 195

VINTE / v in T i / 258.9000 / 62.9434

VISA / v i s A / 292 / 97.3333

VOCÊ / v o s e / 393 / 131

VÔO / v o / 269 / 89.6667

***fim**

Apêndice C

Algumas frases reconhecidas

Para este anexo foram escolhidas algumas frases reconhecidas pelo sistema híbrido ANN+HMM, ao se utilizar a base de dados segmentada manualmente.

Original: A cotação do dólar aumentou e as bolsas fecharam em baixa

Pré-REMAP: A cotação do dólar aumentou e as bolsas fecharam em baixa

REMAP-1: A cotação do dólar aumentou e as bolsas fecharam em baixa

REMAP-2: A cotação do dólar aumentou e as bolsas fecharam em baixa

REMAP-3: A cotação do dólar aumentou e as bolsas fecharam em baixa

Original: A cotação do dólar aumentou, mas as bolsas fecharam em baixa

Pré-REMAP: A cotação do, dólar aumentou, mais, bolsas fecharam, baixa

REMAP-1: Melhores, bolsas fecharam em baixa

REMAP-2: A cotação do dólar aumentou mas, bolsas fecharam em baixa

REMAP-3: A cotação do dólar aumentou, mas as bolsas fecharam em baixa

Original: A bolsa ficará estável ou sofrerá uma pequena queda

Pré-REMAP: Bolsa ficará, estarão, sofrerá uma pequena queda

REMAP-1: A bolsa ficará estável ou sofrerá uma pequena, queda

REMAP-2: A bolsa ficará estável ou sofrerá, a uma pequena, queda

REMAP-3: A bolsa ficará estável ou sofrerá uma pequena, queda

Original: Não haverá ajustes, nem modificações radicais no plano

Pré-REMAP: Não haverá ajustes nem modificações radicais no plano

REMAP-1: Não haverá ajustes nem modificações radicais no plano

REMAP-2: Não haverá ajustes nem modificações radicais no plano

REMAP-3: Não haverá ajustes nem modificações radicais no plano

Original: Foi detectado um problema em seu cartão, ele deve ser substituído

Pré-REMAP: Foi detectado, problema em seu cartão, ele deve ser substituído
REMAP-1: Foi detectado um problema em seu cartão, ele deve ser substituído
REMAP-2: Foi detectado um problema em seu cartão, ele deve ser, substituído
REMAP-3: Foi detectado um problema em seu cartão, ele deve ser, substituído

Original: É necessário que o convênio permita o intercâmbio
Pré-REMAP: Até, necessário que o convênio permita o intercâmbio
REMAP-1: É necessário que o convênio permita o intercâmbio
REMAP-2: É necessário que o convênio permita o intercâmbio
REMAP-3: É necessário que o convênio permita o intercâmbio

Original: Posso afirmar-lhes que o convênio permite o intercâmbio
Pré-REMAP: Afirmar, convênio permita, intercâmbio
REMAP-1: Posso afirmar-lhes que o convênio permita o intercâmbio
REMAP-2: Posso afirmar-lhes que o convênio permita o intercâmbio
REMAP-3: Posso afirmar-lhes que o convênio permita o intercâmbio

Original: O convênio que foi assinado recentemente permite o intercâmbio
Pré-REMAP: O convênio, foi assinado recentemente, permitido, intercâmbio
REMAP-1: O convênio que foi assinado recentemente permite o intercâmbio
REMAP-2: O convênio que foi assinado recentemente permite o intercâmbio
REMAP-3: O convênio que foi assinado recentemente permite o intercâmbio

Original: O convênio permite o intercâmbio porque visa a integração entre alunos de culturas diferentes
Pré-REMAP: O convênio permita o intercâmbio porque visa a integração entre alunos de culturas diferentes
REMAP-1: O convênio permita o intercâmbio porque visa a integração entre alunos de culturas diferentes
REMAP-2: O convênio permita o intercâmbio porque visa a integração entre alunos de culturas diferentes
REMAP-3: O convênio permita o intercâmbio porque visa a integração entre alunos de culturas diferentes

Original: O convênio permite o intercâmbio quando se trata de universidades vinculadas ao projeto de integração
Pré-REMAP: O convênio permita, intercâmbio quando se trata de universidades vinculadas ao projeto, da, integração
REMAP-1: O convênio permita, intercâmbio quando se trata de universidades vinculadas ao projeto de integração

REMAP-2: O convênio permita, intercâmbio quando se trata de universidades vinculadas ao projeto de integração
REMAP-3: O convênio, permita, intercâmbio quando se, trata de universidades vinculadas ao projeto de integração

Original: O convênio, que foi assinado recentemente, permite o intercâmbio
Pré-REMAP: O convênio que foi assinado recentemente, permitido, intercâmbio
REMAP-1: O convênio que foi assinado recentemente, permita o intercâmbio
REMAP-2: O convênio que foi assinado recentemente, permita o intercâmbio
REMAP-3: O convênio que foi assinado recentemente, permita o intercâmbio

Original: O convênio assinado na última reunião é mais interessante do que o anterior
Pré-REMAP: O convênio assinado na última reunião, mais interessante do que o anterior
REMAP-1: O convênio assinado na última reunião é mais interessante do que o anterior
REMAP-2: O convênio assinado na última reunião é mais interessante do que o anterior
REMAP-3: O convênio assinado na última reunião é mais interessante do que o anterior

Original: A medida que o tempo passa, mais nos convencemos da eficácia do convênio
Pré-REMAP: A medida que o tempo passa, mais nos convencemos da eficácia do, convencemos
REMAP-1: A medida que o tempo passa, mais nos convencemos da eficácia do convênio
REMAP-2: A medida que o tempo, passa, mais nos convencemos da eficácia do convênio
REMAP-3: A medida que o tempo passa, mais nos convencemos da eficácia do convênio

Original: Se o convênio permite o intercâmbio, devemos aproveitar a oportunidade para desenvolver o projeto de integração
Pré-REMAP: Se, convênio permita, intercâmbio devemos aproveitar a oportunidade para desenvolver o projeto, integração
REMAP-1: Se o convênio permita, intercâmbio devemos aproveitar a oportunidade para, dezembro, governo, projeto de integração
REMAP-2: Se o convênio permita o intercâmbio devemos aproveitar a oportunidade para, dezembro, governo, projeto de integração
REMAP-3: Se o convênio permita o intercâmbio devemos aproveitar a oportunidade para desenvolver o projeto de integração

Original: Localizado a uma quadra do centro da cidade, o condomínio permite que você una trabalho e conforto
Pré-REMAP: Localizado, uma quadra do centro da cidade, condomínio permite que você una trabalho e conforto

REMAP-1: Localizado, mas, quadra do centro da cidade, de, o condomínio permite que você una trabalho e conforto
REMAP-2: Localizado, mas, quadra do centro da cidade, o condomínio permite que você una trabalho e conforto
REMAP-3: Localizado, mas, quadra do centro da cidade, o condomínio permite, que você una trabalho e conforto

Original: É suficiente
Pré-REMAP: É suficiente
REMAP-1: É suficiente
REMAP-2: É suficiente
REMAP-3: É suficiente

Original: Isto é suficiente
Pré-REMAP: Que, isto é suficiente
REMAP-1: Isto é insuficiente
REMAP-2: Isto é suficiente
REMAP-3: Isto é suficiente

Original: O saldo é suficiente
Pré-REMAP: O saldo é suficiente
REMAP-1: O saldo é suficiente
REMAP-2: O saldo é suficiente
REMAP-3: O saldo é suficiente

Original: O saldo de sua conta é suficiente
Pré-REMAP: O saldo de sua conta é suficiente
REMAP-1: O saldo de sua conta é suficiente
REMAP-2: O saldo de sua conta é suficiente
REMAP-3: O saldo de sua conta é suficiente

Original: O saldo disponível é insuficiente
Pré-REMAP: O saldo disponível, que, insuficiente
REMAP-1: O saldo disponível é insuficiente
REMAP-2: O saldo disponível é insuficiente
REMAP-3: O saldo disponível é insuficiente

Original: O saldo disponível em sua conta é insuficiente
Pré-REMAP: O saldo disponível em sua conta é insuficiente
REMAP-1: O saldo disponível em sua conta é insuficiente
REMAP-2: O saldo disponível em sua conta é insuficiente
REMAP-3: O saldo disponível em sua conta é insuficiente

Original: Isto parece insuficiente

Pré-REMAP: Isto parece insuficiente

REMAP-1: Isto parece insuficiente

REMAP-2: Isto parece insuficiente

REMAP-3: Isto parece insuficiente

Original: O saldo parece ser insuficiente

Pré-REMAP: Do, saldo parece, ser insuficiente

REMAP-1: O saldo parece, ser insuficiente

REMAP-2: O saldo parece ser insuficiente

REMAP-3: O saldo parece ser insuficiente

Original: O saldo sempre está disponível

Pré-REMAP: O saldo sempre está, a, disponível

REMAP-1: O saldo sempre está disponível

REMAP-2: O saldo sempre está disponível

REMAP-3: O saldo sempre está disponível

Original: O saldo está sempre disponível no início do mês

Pré-REMAP: No, saldo está sempre disponível no início do, milhões

REMAP-1: O saldo está sempre disponível no início do mês

REMAP-2: O saldo está sempre disponível no início do mês

REMAP-3: O saldo está sempre disponível no início do mês

Apêndice D

Ferramenta de avaliação SCLITE

D.1 Introdução

Neste anexo será feita uma descrição do software de avaliação utilizado para extrair as taxas de acerto de palavras presentes no capítulo 5. Maiores detalhes dos procedimentos de avaliação e formatos de arquivos de entrada podem ser encontrados pela internet em [18] ou na documentação que o acompanha.

D.2 Descrição

O SCLITE (do termo em inglês *Score-Lite*), que faz parte da *toolkit* NIST SCTK (*Scoring Toolkit*), é uma ferramenta para contagem de erros e avaliação da saída de sistemas de reconhecimento de fala através do qual são emitidos relatórios (resumidos ou detalhados) de desempenho.

O programa faz uma comparação entre a sentença reconhecida com a sua respectiva sentença de referência em um processo chamado de *alinhamento* e, após esta etapa, são iniciados os levantamentos estatísticos resultantes. Conforme as necessidades do usuário, este poderá escolher dentre os diversos tipos de relatórios (resumidos ou detalhados) e a partir daí verificar o desempenho de seu sistema.

Ao longo deste apêndice serão utilizadas as abreviaturas *REF* e *HIP* quando o texto se referir às sentenças de referência e reconhecidas respectivamente. A seguir é apresentada a sintaxe geral do comando que aciona a avaliação das sentenças de teste:

```
sclite -r arq_ref [ fmt ] -h arq_hip [ fmt [ título ] ] opções
```

onde `-r arq_ref [ fmt ]` é um argumento de entrada obrigatório que indica o arquivo de referência que será comparado com o arquivo de sentenças reconheci-

das. O campo `[ fmt ]` indica o formato do arquivo de referência. Antes deste argumento de entrada poderão vir outras opções [18].

Para indicar o arquivo de sentenças reconhecidas, deve-se indicar no campo `-h arq_hip [ fmt [ título ] ]` que, analogamente ao argumento de entrada do arquivo de referência, indica o nome do arquivo de sentenças reconhecidas, o formato deste arquivo. Por definição o programa gera o(s) arquivo(s) de saída com o mesmo nome do arquivo de hipóteses. Mas se for utilizado o campo `[ título ]`, o arquivo de saída será criado com o nome indicado por este campo.

As demais opções (alinhamento, arquivo de saída e relatórios) são escolhidas pelo usuário para ter a disposição os diversos tipos de relatórios para análise de desempenho do sistema que podem ser verificadas detalhadamente em [18].

Para as análises da saída do sistema de reconhecimento de fala contínua utilizado neste trabalho, foi executada a seguinte linha de comando:

```
sclite -i wsj -r arq_ref -h arq_hyp -O dir_sai -o dtl all
```

Abaixo, segue um exemplo com arquivos e diretórios reais, utilizando a referida sintaxe:

```
sclite -i wsj -r /home/JoseAntonio/frases_corretas.ref -h
/home/JoseAntonio/SMCMD_PA01_NH100_N10.hyp.0 -O /home/JoseAntonio/
-o dtl all
```

Assim, tem-se:

- `/home/JoseAntonio/frases_reconhecidas.ref` é o diretório que contém o arquivo de frases de referência `frases_reconhecidas.ref`
- `/home/JoseAntonio/SMCMD_PA01_NH100_N10.hyp.0` é o diretório que contém o arquivo de frases reconhecidas `SMCMD_PA01_NH100_N10.hyp.0`
- `/home/JoseAntonio/` é o diretório que serão gravados todos os arquivos com os resultados da avaliação

D.3 Alinhamento das sentenças

Quanto ao processo de alinhamento, este é feito em duas etapas. A primeira é chamada de seleção dos textos *REF* e dos textos *HIP*. O SCLITE possui quatro algoritmos que executam esta tarefa e cada um deles é acionado conforme o tipo do formato dos arquivos de entrada. Para a obtenção das taxas de acerto de palavras deste trabalho foi utilizado o formato mais simples que possibilita a

seleção do tipo *identificadores de locução* (*Utterance ID Matching* [18]), sendo cada sentença (*HIP* e *REF*) seguida de um identificador alfa-numérico entre parêntesis pelo qual os pares serão agrupados cada um com o seu respectivo identificador. Assim, define-se o arquivo com as sentenças de referência tendo como extensão `.ref` e para as locuções reconhecidas foi criado um arquivo de extensão `.hip` ao final de cada etapa de reconhecimento.

Na Tabela D.1 temos um exemplo de um arquivo com algumas sentenças de referência (por exemplo, o arquivo `lista_frases_corretas.ref`) e também o resultado da saída do reconhecedor (chamemos de `lista_frases_reconhecidas.hip`):

Tabela D.1: *Exemplo do formato dos arquivos de entrada onde o sclite utiliza o alinhamento do tipo identificador de locução entre sentenças HIP e REF.*

sentenças de referência
A COTAÇÃO DO DÓLAR AUMENTOU , E AS BOLSAS FECHARAM EM BAIXA (1)
A COTAÇÃO DO DÓLAR AUMENTOU , MAS AS BOLSAS FECHARAM EM BAIXA (2)
A BOLSA FICARÁ ESTÁVEL OU SOFRERÁ UMA PEQUENA QUEDA (3)
NÃO HAVERÁ AJUSTES , NEM MODIFICAÇÕES RADICAIS NO PLANO (4)
sentenças reconhecidas
A COTAÇÃO DO DÓLAR AUMENTOU E AS BOLSAS FECHARAM EM BAIXA (1)
A COTAÇÃO DO DÓLAR AUMENTOU MAS , BOLSAS FECHARAM EM BAIXA (2)
, A BOLSA FICARÁ ESTÁVEL OU SOFRERÁ , A UMA PEQUENA QUEDA (3)
NÃO HAVERÁ AJUSTES NEM MODIFICAÇÕES RADICAIS NO PLANO (4)

Após a seleção dos identificadores, é feito o alinhamento das sentenças *REF/ HIP* para que sejam levantadas as estatísticas do desempenho do sistema. A Tabela D.2 mostra o resultado do alinhamento entre sentença de referência e sua respectiva hipótese:

Tabela D.2: *Resultado do alinhamento entre sentenças REF e HIP.*

REF	a	cotação	do	dólar	aumentou	,	mas	AS	bolsas	fecharam	em	baixa
HIP	a	cotação	do	dólar	aumentou	*	mas	,	bolsas	fecharam	em	baixa
Eval						D		S				

Percebe-se que após o alinhamento entre as sentenças o algoritmo detectou um erro de deleção (a vírgula foi suprimida) e um erro de substituição (a palavra *"as"* foi trocada por *","*).

D.4 Contagem dos erros

Após o alinhamento dos pares, é feita a contagem para cada sentença e assim são calculadas as seguintes estatísticas:

$$PC = \frac{N_{pc}}{N_r} \times 100 \quad \text{(D.1)}$$

$$S = \frac{N_S}{N_r} \times 100 \quad \text{(D.2)}$$

$$D = \frac{N_D}{N_r} \times 100 \quad \text{(D.3)}$$

$$I = \frac{N_I}{N_r} \times 100 \quad \text{(D.4)}$$

$$PSE = \frac{N_{rhe}}{N_{rh}} \times 100 \quad \text{(D.5)}$$

onde:

PC: Porcentagem de palavras do conjunto de referência que constam no conjunto de sentenças reconhecidas

S: Taxa de erros de substituição

D: Taxa de erros de deleção

I: Taxa de erros de inserção

PSE: Taxa de sentenças erradas (que contenham pelo menos um dos erros S, D ou I)

N_{pc}: Número de palavras do conjunto de referência que constam no conjunto de sentenças reconhecidas

N_r: Número de palavras do conjunto de referência

N_D: Número de erros de deleção

N_S: Número de erros de substituição

N_I: Número de erros de inserção

N_{rhe}: Número de pares *REF/HIP* errados

N_{rh}: Número de pares *REF/HIP*

Com estas estatísticas, é calculada a taxa de erro de palavras:

$$WER = S + D + I \quad \text{(D.6)}$$

D.5 Interpretação dos resultados

Com isto chega-se à etapa que o SCLITE gera as estatísticas de desempenho do sistema de reconhecimento de fala. Os relatórios fornecem uma análise bastante detalhada das sentenças reconhecidas, frente às sentenças de referência. Um dos relatórios que o programa emite possui extensão `.dtl` (que vem do termo *detailed report*), no qual a primeira seção se refere ao desempenho em termos de reconhecimento de sentenças.

```
SENTENCE RECOGNITION PERFORMANCE

 sentences                                  100
 with errors                    60.0%   (  60)

   with substitions             41.0%   (  41)
   with deletions               23.0%   (  23)
   with insertions              31.0%   (  31)
```

Figura D.1: *Primeira seção do arquivo com relatório detalhado do desempenho do sistema de reconhecimento de fala contínua que mostra o percentual de sentenças com erros.*

Na figura acima, a primeira informação é referente ao número total de sentenças reconhecidas. Na segunda parte são mostradas as estatísticas dos erros para cada sentença testada.

Na próxima seção, mostrada na Figura D.2, é feita uma análise detalhada dos resultados das palavras reconhecidas. A primeira porcentagem é referente à taxa de erro de palavra que corresponde ao somatório dos erros de deleção, substituição ou inserção. Pode-se perceber que não existe uma coerência entre as taxas *Percent Correct* e *Percent Total Error* que possuem valores de 90,7% e 16,1% respectivamente. Isso se deve ao fato de que o parâmetro *Percent Correct* se refere a porcentagem de palavras do conjunto de referência que estão presentes no conjunto de sentenças reconhecidas. A Figura D.2 mostra todos os parâmetros referentes às estatísticas das palavras reconhecidas onde também é mostrado o número de palavras do conjunto de referência, do conjunto de sentenças de reconhecimento e o número de alinhamentos realizados da forma como está exemplificado na Tabela D.2.

```
WORD RECOGNITION PERFORMANCE

Percent Total Error       =   16.1%   ( 161)

Percent Correct           =   90.7%   ( 905)

Percent Substitution      =    6.4%   (  64)
Percent Deletions         =    2.9%   (  29)
Percent Insertions        =    6.8%   (  68)
Percent Word Accuracy     =   83.9%


Ref. words                =           ( 998)
Hyp. words                =           (1037)
Aligned words             =           (1066)
```

Figura D.2: *Seção referente ao desempenho em termos de reconhecimento de palavras.*

A seguir, as ocorrências de cada tipo de erro são enumeradas assim como as palavras que contribuíram.

```
DELETIONS                          Total                (9)
                                   With >=  1 occurances (9)

   1:   11  ->  ,
   2:    6  ->  e
   3:    6  ->  o
   4:    1  ->  a
   5:    1  ->  as
   6:    1  ->  cotação
   7:    1  ->  do
   8:    1  ->  dólar
   9:    1  ->  mas
     -------
        29
```

Figura D.3: *Lista das palavras que contribuíram para a ocorrência do erro de deleção e o número de ocorrência.*

```
INSERTIONS                      Total                 (22)
                                With >=  1 occurances (22)

  1:   35  ->  ,
  2:    5  ->  mas
  3:    5  ->  um
  4:    3  ->  e
  5:    2  ->  na
  6:    2  ->  regras
  7:    1  ->  a
  8:    1  ->  amanhã
  9:    1  ->  aos
 10:    1  ->  da
 11:    1  ->  das
 12:    1  ->  dezembro
 13:    1  ->  dois
 14:    1  ->  dólar
 15:    1  ->  em
 16:    1  ->  estarão
 17:    1  ->  estão
 18:    1  ->  no
 19:    1  ->  prestação
 20:    1  ->  rio
 21:    1  ->  será
 22:    1  ->  visa
    -------
       68
```

Figura D.4: *Lista das palavras que contribuíram para a ocorrência do erro de inserção e o número de ocorrência.*

```
SUBSTITUTIONS                   Total                 (26)
                                With >=  1 occurances (26)

  1:    9  ->  o
  2:    6  ->  as
  3:    6  ->  os
  4:    5  ->  permite
  5:    5  ->  todos
  6:    4  ->  e
  7:    3  ->  está
  8:    3  ->  mês
  9:    2  ->  das
 10:    2  ->  de
 11:    2  ->  horas
 12:    2  ->  no
 13:    2  ->  é
 14:    1  ->  a
 15:    1  ->  amanhã
 16:    1  ->  aumentou
 17:    1  ->  cento
 18:    1  ->  desenvolver
 19:    1  ->  do
 20:    1  ->  em
 21:    1  ->  estação
 22:    1  ->  início
 23:    1  ->  isto
 24:    1  ->  santa
 25:    1  ->  uma
 26:    1  ->  una
    -------
       64
```

Figura D.5: *Lista das palavras que contribuíram para a ocorrência do erro de substituição e o número de ocorrência.*

Outro arquivo que o sclite gera consta de uma lista dos alinhamentos realizados para cada sentença de referência e reconhecida (arquivo de extensão `.pra`), respectivamente. A linha rotulada "`Scores:`" mostra os resultados do reconhecimento com um código de símbolos onde `#C` significa o número de palavras corretas, `#S` é o número de substituições, `#D` é o número de deleções e `#I` é o número de inserções. A linha "`Eval:`" indica quais pares *REF/HIP* cometeram algum tipo de erro.

```
            DUMP OF SYSTEM ALIGNMENT STRUCTURE

System name:
/home/joseamrz/TESE_MESTRADO/FRASES_RECONHECIDAS/COM_GRAMÁTICA/SEGMENTACA
O_MANUAL/N10/SMCMD_PA04_NH100_N10.hyp.3

Speakers:
    0:  1)
    1:  2)
    2:  3)

   •••

Speaker sentences   0:  1)   #utts: 1
id: (1)
Scores: (#C #S #D #I) 11 0 1 0
REF:  a cotação do dólar aumentou , e as bolsas fecharam em baixa
HYP:  a cotação do dólar aumentou * e as bolsas fecharam em baixa
Eval:                            D

Speaker sentences   1:  2)   #utts: 1
id: (2)
Scores: (#C #S #D #I) 5 1 6 0
REF:  A COTAÇÃO DO DÓLAR AUMENTOU , MAS AS bolsas fecharam em baixa
HYP:  * ******* ** ***** AMANHÃ   , *** ** bolsas fecharam em baixa
Eval: D D       D  D     S          D   D

Speaker sentences   2:  3)   #utts: 1
id: (3)
Scores: (#C #S #D #I) 8 1 0 3
REF:  a bolsa ficará estável ou sofrerá * ****** UMA pequena * queda
HYP:  a bolsa ficará estável ou sofrerá , AMANHÃ ,   pequena , queda
Eval:                                  I I      S           I


   •••
```

Figura D.6: *Lista dos alinhamentos executados para cada sentença de reconhecimento.*

Referências Bibliográficas

[1] MARTIN, A., PRZYBOCKI M., The 2001 NIST Evaluation for Recognition of Conversational Speech Over the Telephone, 2001 NIST Large Vocabulary Conversational Speech Recognition Workshop, 2001.

[2] KONIG, Y., BOULARD, H., MORGAN, N., REMAP: Experiments with Speech Recognition, IEEE ICASSP96, Atlanta, pp. 7–10, 1996.

[3] BOURLARD, H., KONIG, Y., MORGAN, N., REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities Application to Transition-Based Connectionist Speech Recognition, PhD Thesis, University of California at Berkeley, 1996.

[4] sctk-1.3 - Speech Recognition Scoring Toolkit SCTK Version 1.3 (Includes the SCLITE Scoring program)
ftp://jaguar.ncsl.nist.gov/pub/sctk-1.3.tgz (11/10/2003).

[5] YNOGUTI, C. A., Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov, Tese de Doutorado, Universidade Estadual de Campinas, 1999.

[6] RABINER, L. R. and JUANG, B. H., Fundamentals of Speech Recognition, Prentice Hall Signal Processing Series, 1993.

[7] MORGAN, N. and BOURLARD, H., Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach, IEEE Signal Processing Magazine, Invited Paper, vol. 12, no. 3, pp. 25–42, May 1995.

[8] MORGAN, N. and BOURLARD, H., Neural Networks for Statistical Recognition of Continuous Speech, Proceedings of the IEEE, Invited Paper, vol. 83, no. 5, pp. 741–770, May 1995.

[9] MORAIS, E. S., Reconhecimento Automático de Fala Contínua Empregando Modelos ANN+HMM, Tese de Mestrado, Universidade Estadual de Campinas, 1997.

[10] TRENTIN, E., GORI, M., A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition, Neurocomputing, 37, pp. 91–126, March 2001.

[11] BOURLARD, H. and WELLEKENS, C. J., Links Between Markov Models and Multilayer Perceptrons, IEEE Transactions on Patter Analysis and Machine Intelligence, vol. 12, no.12, pp. 1167–1178, December 1990.

[12] HAYKIN, S., Neural Networks: A Comprehensive Foundation, Prentice Hall, New Jersey, 1994.

[13] TEBELSKIS, J., Speech Recognition Using Neural Networks, PhD Thesis, Carnegie Mellon University, 1995.

[14] RICHARD, M. D. and LIPPMANN, R. P., Neural Network Classifiers Bayesian a posteriori Probabilities, Neural Computation, vol. 3, no. 4, pp. 461–483, December 1991.

[15] YAH, Y., FANTY, M., COLE, R., Speech Recognition Using Neural Networks With Forward-Backward Probability Generated Targets, Proceedings IEEE ICASSP, 1997.

[16] SILVA, T. C., Fonética e Fonologia do Português – roteiro de estudos e guia de exercícios. Editora Contexto. São Paulo, 2002.

[17] RABINER, L. R., LEVINSON, S. E., A Speaker-Independent, Connected Word Recognition System Based on Hidden Markov Models and Level Building, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 3, no. 33, pp. 561–573, June 1985.

[18] Documentação online sclite
http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm
(10/02/2005).

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo