Dissertação de Mestrado

Inatel
Instituto Nacional de Telecomunicações

# ADAPTAÇÃO AO LOCUTOR USANDO A TÉCNICA MLLR

DANIELA BARUDE FERNANDES

MARÇO/2011

**INSTITUTO NACIONAL DE TELECOMUNICAÇÕES – INATEL**
**MESTRADO EM TELECOMUNICAÇÕES**

# ADAPTAÇÃO AO LOCUTOR USANDO A TÉCNICA MLLR

DANIELA BARUDE FERNANDES

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

ORIENTADOR: Prof. Dr. Carlos Alberto Ynoguti

SANTA RITA DO SAPUCAÍ – MG
**2011**

Dissertação defendida e aprovada em 16/03/2011, pela comissão julgadora:

Prof. Dr. Carlos Alberto Ynoguti – INATEL

Prof. Dr. Fábio Violaro – FEEC/UNICAMP

Prof. Dr. Miguel Arjona Ramírez – USP

**Prof. Dr. Luciano Leonel Mendes**
Coordenador do Curso de Mestrado

À Eduardo Garcia Pina e à minha filha Mariana Barude Pina por estarem ao meu lado nos momentos mais importantes da minha vida.

**AGRADECIMENTOS**

A Deus, por me iluminar em todos os dias da minha vida.

À minha mãe, Maria Terezinha Dias Barude (*in memorian*) que tanto torceu e me incentivou na realização desse trabalho, a meu pai Samuel Jióia Fernandes e às minhas irmãs, Suzana Barude Fernandes e Juliana Barude Fernandes, pelo apoio, pela compreensão e por fazerem meus dias mais felizes.

A meu orientador, professor Doutor Carlos Alberto Ynoguti, pela troca de ideias, pelo incentivo e principalmente pelo empenho na realização deste trabalho.

A todos os meus amigos, pelas palavras de apoio nos momentos difíceis e pelos momentos descontraídos.

Aos professores do Inatel, grandes amigos, por sempre me incentivarem e me ajudarem a transpor obstáculos.

**ÍNDICE**

**LISTA DE FIGURAS**

**LISTA DE TABELAS**

## LISTA DE ABREVIATURAS E SIGLAS

SRF – Sistema de Reconhecimento de Fala

SD – Sistema de Reconhecimento de Fala Dependente de Locutor

SI – Sistema de Reconhecimento de Fala Independente de Locutor

HMM – Modelo Oculto de Markov

MLLR – Regressão Linear de Máxima Verossimilhança

FFT - Transformada Rápida de Fourier

DCT – Transformada Discreta do Cosseno

SA – Sistema Adaptado

**RESUMO**

Neste trabalho realizou-se um estudo da técnica de adaptação ao locutor chamada MLLR, Regressão Linear de Máxima Verossimilhança. Os testes foram realizados utilizando fala contínua e somente as médias das componentes gaussianas dos Modelos Ocultos de Markov (HMMs) foram adaptadas. O ponto fundamental da técnica é a partição dessas médias em classes de regressão para a geração da matriz de transformação. Além disso, a quantidade de material para adaptação de um sistema independente de locutor é muito importante. Sendo assim, diversas alternativas para a formação das classes de regressão foram exploradas. Foram testados métodos baseados em classificação fonética e em medidas de distância, variando-se também o número de classes de regressão. Após a realização dos testes, com um número variado de locuções de adaptação, verificou-se que o melhor resultado foi obtido utilizando-se quatro locuções de adaptação e três classes de regressão, mas pesquisas ainda devem ser feitas na área.

Palavras-chave: Sistemas de Reconhecimento de Fala, Adaptação ao Locutor, Técnica MLLR, Classes de Regressão.

**ABSTRACT**

In this work a study of the technique of speaker adaptation called MLLR, Maximum Likelihood Linear Regression was made. The tests have been done using continuous speech applications and only the means of continuous hidden Markov models (HMM) have been adapted. The basic point of the technique is the partition of these means in regression classes for the generation of the transformation matrix. Moreover, the amount of material for adaptation of a speaker independet system is very important. Being thus, some alternatives for regression classes construction have been explored. Methods based on phonetic classification and based on distance metrics have been tested, varying also the number of regression classes. After tests, with a varied number of adaptation sentences, was verified that the better approach is to use only three regression classes with four adaptation sentences, but research still must be made in the area.

Keywords: Speech Recognition System, Speaker Adaptation, MLLR Technique, Regression Classes.

# CAPÍTULO 1

# INTRODUÇÃO

Sistemas de Reconhecimento de Fala (SRFs) vêm sendo utilizados nas mais diversas aplicações e com diferentes propósitos. O principal objetivo é auxiliar a realização de tarefas, entre as quais podemos citar: acionamento de dispositivos eletrônicos (tetraplégicos controlam o computador usando comandos de voz), conversão de sentenças faladas para sentenças escritas (um escritor dita seu livro inteiro para o computador, que transcreve as palavras com precisão) e realização de tarefas por comando de voz em telefones celulares (em automóveis o motorista fala para seu celular com GPS o endereço de seu destino, ao invés de digitá-lo) [1][2].

A principal função de um SRF é receber um sinal de voz (sinal acústico) em sua entrada e produzir em sua saída uma sequência de fonemas, palavras ou frases correspondentes à entrada. Isso deve ser feito com a mais alta precisão, para qualquer locutor e em qualquer ambiente, possibilitando a comunicação homem-máquina, através da fala, como ocorre entre os seres humanos. Entretanto, a utilização do sinal de fala apresenta algumas desvantagens como por exemplo a interpretação em ambientes muito ruidosos e a variação das características do sinal de voz (diferenças na forma de falar, sotaques, velocidade da fala, condições físicas e emocionais, entre outros).

Existem dois tipos de SRFs: o Sistema de Reconhecimento de Fala Independente de Locutor (SI) e o Sistema de Reconhecimento de Fala Dependente de

Locutor (SD) [3]. Como o próprio nome já diz um SI deve interpretar o sinal de fala independente de quem o está gerando. Já um SD é específico para o sinal de voz de determinada pessoa. De forma intuitiva, notamos que o desempenho de um SD é melhor em relação ao SI, pois para um único locutor teremos menos variabilidade das características do sinal de voz.

Para o treinamento de um SRF é necessária uma quantidade de dados ou de informações, chamada de locuções, relativamente elevada para se ter um sistema robusto, ou seja, que apresenta uma alta taxa de acerto de palavras no reconhecimento. E ainda, dependendo do vocabulário a ser reconhecido, essa quantidade de informação torna-se cada vez mais difícil de ser obtida, e para um único locutor, completamente exaustiva. Dessa forma, várias técnicas vêm sendo utilizadas com a finalidade de melhorar SRFs para que estes possam ser utilizados em locutores que antes não eram bem reconhecidos, sem a necessidade de uma quantidade elevada de material de treinamento ou o treinamento propriamente dito de um SD por completo. Uma dessas técnicas é chamada de adaptação ao locutor.

O processo de adaptação ao locutor nada mais é do que a utilização das informações de um SI combinadas com dados mais específicos de um locutor, obtidos através de algumas locuções desse novo locutor. Esses dados são conhecidos como material de adaptação.

O objetivo desse trabalho é apresentar uma das técnicas de adaptação mais utilizadas, que é a chamada  Regressão Linear de Máxima Verossimilhança, MLLR (do inglês, Maximum Likelihood Linear Regression). Essa técnica utiliza uma matriz de transformação para alterar os parâmetros de um modelo inicial, usado como ponto de partida. Idealmente todos os parâmetros deveriam ser transformados. Entretanto, em muitas aplicações, a quantidade de material disponível para adaptação é limitado, tornando-se inviável esta abordagem. Diante disso, optou-se por adaptar apenas as médias das gaussianas.

O conteúdo da tese se divide da seguinte forma: No capítulo 2 serão

apresentados alguns fundamentos teóricos dos SRFs necessários para o entendimento do processo de adaptação. No capítulo 3 será dada uma visão geral sobre o processo de adaptação e algumas técnicas utilizadas. No capítulo 4, será abordada a técnica MLLR propriamente dita. No capítulo 5, serão apresentados os testes realizados e os resultados obtidos utilizando a técnica MLLR. Por fim, no capítulo 6, teremos a conclusão e algumas sugestões e comentários para pesquisas futuras.

# CAPÍTULO 2

# FUNDAMENTOS TEÓRICOS

## 2.1 - Características do Sinal de Voz

O sinal de voz apresenta diversas variações que podemos classificar em dois tipos: variações intralocutores e variações interlocutores. As variações intralocutores são provenientes de uma mesma pessoa e se caracterizam por mudanças na pronúncia do sinal de voz. Isso ocorre devido ao contexto de uma frase, ao estado emocional do locutor, dicção e grau de clareza ao se pronunciar uma frase e velocidade com que esta é dita. As variações interlocutores são provocadas por diferenças fisiológicas entre os locutores, dentre as quais podemos citar sexo, idade e diferenças culturais (sotaque).

## 2.2 - Sistema de Reconhecimento de Fala

A estrutura de um SRF é representada na figura a seguir [4][5]:

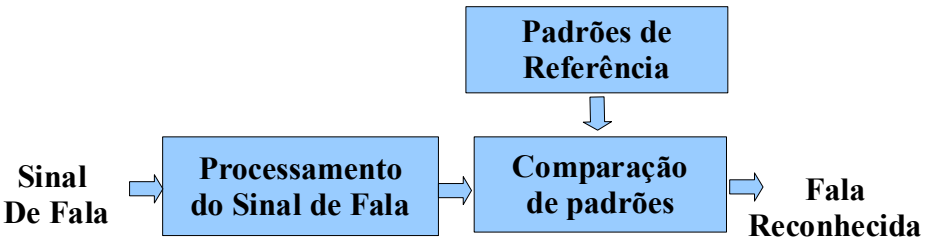


**Figura 1** - *Estrutura Básica de um Sistema de Reconhecimento de Fala*

Inicialmente o sinal de fala é convertido em um conjunto de parâmetros para que seja possível a realização da comparação entre os diversos padrões existentes.

Os padrões de referência são obtidos a partir de amostras do vocabulário a ser reconhecido. Para criá-los o método mais utilizado é constituído pelos Modelos Ocultos de Markov (HMMs).

Na comparação de padrões determina-se a probabilidade de que cada modelo de referência tenha gerado o conjunto de parâmetros de entrada e escolhe-se o mais próximo do sinal de entrada para representar o sinal de fala reconhecido.

### 2.2.1 – Extração dos Parâmetros

Para que seja possível o reconhecimento de um sinal acústico é necessário tratá-lo de forma adequada [4][6][7]. A estrutura básica do processamento de um sinal de fala é representado na figura 2 abaixo:

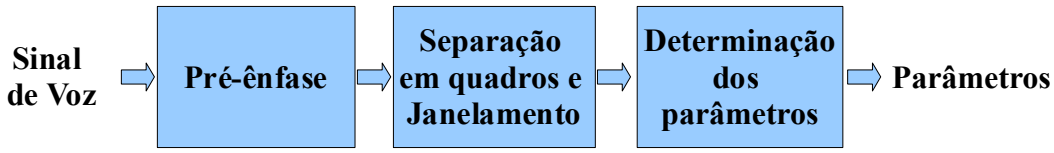


**Figura 2 -** *Processamento do Sinal de Fala*

O processamento do sinal de fala é dividido em três etapas: pré-ênfase, separação em quadros e janelamento e determinação dos parâmetros.

A pré-ênfase, realizada por um filtro passa altas $(1-0{,}95z^{-1})$ , é utilizada para atenuar o efeito combinado do espectro dos pulsos glotais mais a radiação dos lábios depois que o sinal de voz é captado por um microfone [6] [8]. A base de dados utilizada neste trabalho foi feita em ambiente relativamente silencioso, com um microfone direcional de boa qualidade , com uma taxa de amostragem de 11,025

kHz e resolução de 16 bits.

O sinal de voz não é um sinal estacionário. Por isso torna-se necessário dividi-lo em quadros. Essa divisão é feita considerando que o tempo de duração de cada quadro seja especificado de tal forma a garantir que nesse período o sinal de voz seja quase estacionário. Os intervalos de análise, para obtenção dos parâmetros, denominados de janelas, são maiores que os quadros, ou seja, a separação é feita com a superposição das janelas adjacentes. Sendo assim, um intervalo de análise possui amostras do final da janela anterior e amostras do início da janela posterior. Para suavização dos parâmetros (minimizar as descontinuidades do sinal no início e final de cada quadro), os intervalos são multiplicados por um janela, usualmente a janela de Hamming, que dá maior ênfase às amostras localizadas no centro desta. Normalmente em reconhecimento de fala os quadros possuem duração de 10 a 20 ms e as janelas, de 20 a 30 ms. Nos testes foram utilizadas janelas de 20 ms e quadros de 10 ms.

Finalmente, é feita a determinação de parâmetros. Nesta etapa temos a conversão da representação temporal do sinal de voz analisado, em alguma forma de representação espectral. O método de análise espectral por banco de filtros, obtido a partir da Transformada Rápida de Fourier (FFT) foi utilizado na obtenção dos parâmetros Mel Cepstrais [6]. Neste processo, inicialmente é calculado o quadrado do módulo da FFT das amostras pertencentes a cada janela de análise e, em seguida, o sinal é filtrado por um banco de filtros triangulares na escala Mel.

A escala Mel está representada na figura 3 a seguir.

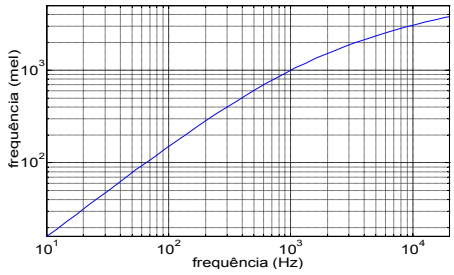


**Figura 3** - *Escala Mel*

A tabela 1, apresenta as frequências centrais do banco de filtros triangulares na escala Mel, com a respectiva largura de faixa utilizada.

**Tabela 1** - *Banco de Filtros na Escala Mel*

| Filtro | Frequência Central (Hz) | BW (Hz) |
|---|---|---|
| 1 | 100 | 100 |
| 2 | 200 | 100 |
| 3 | 300 | 100 |
| 4 | 400 | 100 |
| 5 | 500 | 100 |
| 6 | 600 | 100 |
| 7 | 700 | 100 |
| 8 | 800 | 100 |
| 9 | 900 | 100 |
| 10 | 1000 | 124 |
| 11 | 1149 | 160 |
| 12 | 1320 | 184 |
| 13 | 1516 | 211 |
| 14 | 1741 | 242 |
| 15 | 2000 | 278 |
| 16 | 2297 | 320 |
| 17 | 2639 | 367 |
| 18 | 3031 | 422 |
| 19 | 3482 | 484 |
| 20 | 4000 | 556 |

Após a passagem do sinal pelo banco de filtros, é feito o cálculo do logaritmo da energia na saída de cada filtro e por fim, aplica-se a Transformada Discreta do Cosseno (DCT) sobre estes valores, obtendo-se os 12 coeficientes mel-cepstrais por janela. Os coeficientes mel-cepstrais são obtidos pela seguinte equação:

$$Mel_i = \sum_{j=1}^{F} 10 \log_{10}(X_j) \cos[i(j - \frac{1}{2})\frac{\pi}{F}] \qquad i = 1,2,\ldots,M_c \qquad (2.1)$$

onde:

$X_j$ é a energia na saída do *j*-ésimo filtro.

$M_c$ é o número de coeficientes mel-cepstrais.

Além dos parâmetros mel-cepstrais, também foram utilizados a energia e os parâmetros diferenciais: delta-mel-cepstrais, delta-delta-mel cepstrais e delta-energia, com apenas uma janela de variação de cada lado. Os parâmetros delta foram calculados segundo a seguinte expressão:

$$\Delta_i(n) = \frac{1}{2\mathrm{J}+1} \sum_{j=-J}^{J} j y_{i-j}(n) \qquad (2.2)$$

onde:

$y_i(n)$ é o *n*-ésimo elemento do vetor de parâmetros acústicos do *i*-ésimo quadro da locução.

$\Delta_i(n)$ é o *n*-ésimo elemento do vetor delta correspondente ao vetor de parâmetros acústicos do *i*-ésimo quadro da locução.

$J$ é o número de janelas adjacentes a serem considerados no cálculo dos parâmetros delta do quadro em questão. Neste trabalho, utilizou-se $J = 1$ no cálculo dos parâmetros delta e delta-delta.

Após a obtenção de todos os parâmetros descritos anteriormente, o sinal de fala é representado por uma sequência temporal de vetores, onde cada vetor representa um quadro do sinal de voz, definido anteriormente.

### 2.2.2 – Modelos Ocultos de Markov

Modelo Oculto de Markov, HMM (do inglês, Hidden Markov Model) é um

modelo estatístico que pode ser usado para representar sinais como processos aleatórios [4]. Uma frase, uma palavra ou um fone apresentam variabilidade quando são pronunciados mais de uma vez. Isso significa que podemos considerar uma palavra ou até mesmo um fone como um evento aleatório, que portanto pode ser caracterizado por HMMs.

Um HMM é um conjunto de estados conectados por transições. Cada estado possui uma probabilidade inicial que indica a possibilidade desse estado ser o estado inicial do HMM. Podem ocorrer mudanças entre os estados ou permanência num mesmo estado, de acordo com uma função de probabilidade conhecida como probabilidades de transição. E ainda, associada a cada transição ou a cada estado, existe uma outra função de probabilidade, conhecida como probabilidade de emissão de símbolos de um alfabeto ou conjunto de dados. A cada instante de tempo, ocorre uma mudança de estado ou a permanência no mesmo e a emissão de um símbolo. A sequência de símbolos, conhecida como sequência de observação, é a saída do HMM. A sequência de estados é oculta [9][10].

Os HMMs podem ser discretos ou contínuos. A diferença está no tipo de função que representa a probabilidade de emissão de símbolos. No HMM discreto esta função é representada por uma tabela de probabilidades. No HMM contínuo, é utilizada uma função densidade de probabilidade gaussiana multidimensional ou uma soma de duas ou mais funções densidade de probabilidade gaussianas multidimensionais. Cada função densidade de probabilidade gaussiana é representada por uma média, uma variância e um peso, no caso da utilização de mais de uma dessas funções (esse peso indica o grau de influência de cada função no conjunto de funções existentes). Portanto, os elementos de um HMM são:

- Conjunto de estados $S=\{S_1, S_2, \cdots, S_N\}$ , onde $N$ é o número de estados.
- Função probabilidade do estado inicial $\pi = \{ \pi_i \}$.

$$\pi_i = P[q_1 = S_i] \quad 1 \leqslant i \leqslant N \tag{2.3}$$

onde $q_1$ é o estado inicial (t = 1).

- Função probabilidade de transição A.
- Função de probabilidade de símbolo de saída B.

Podemos definir a matriz A da seguinte forma:

$$A=\{a_{ij}\} \tag{2.4}$$

$$a_{ij}=P[q_t=S_j|q_{t-1}=S_i] \quad 1\leqslant i\,,\,j\leqslant N \tag{2.5}$$

onde $a_{ij}$ é a probabilidade de ocorrer uma transição do estado $S_i$ para o estado $S_j$.

Os coeficientes $a_{ij}$ devem obedecer às seguintes regras:

$$a_{ij}\geqslant 0 \quad 1\geqslant i\,,\,j\geqslant,N \tag{2.6}$$

$$\sum_{j=1}^{N} a_{ij}=1 \quad 1\geqslant i\geqslant N \tag{2.7}$$

A função probabilidade de emissão de símbolos é definida por:

1) Para Modelos Discretos

$$b_i(k)\geqslant 0 \quad \begin{matrix}1\leqslant i\leqslant N\\ 1\leqslant k\leqslant K\end{matrix} \tag{2.8}$$

$$\sum_{k=1}^{K} b_i(k)=1 \quad 1\leqslant i\leqslant N \tag{2.9}$$

onde:

$K$ é o número de símbolos de saída.

$b_i(k)$ é a probabilidade de emitir o símbolo $k$ no estado $S_i$ .

2) Para Modelos Contínuos

$$b_j(O_t)=\sum_{m=1}^{M} c_{jm}G(O_t,\mu_{jm},U_{jm}) \quad 1\leqslant j\leqslant N \tag{2.10}$$

onde:

$O_t$ é o vetor de entrada ou sequência de observação.

$M$ é o número de componentes gaussianas.

$c_{jm}$ é o peso da m-ésima mistura no estado $S_j$ e $\sum_{m=1}^{M} c_{jm}=1$ .

$G$ é uma função densidade de probabilidade Gaussiana multimensional dada por:

$$G(O_t, \mu_{jm}, U_{jm})=\frac{1}{(2\pi)^{\frac{dim}{2}}|U_{jm}|^{\frac{1}{2}}}\exp\{\frac{-(O_t-\mu_{jm})U_{jm}^{-1}(O_t-\mu_{jm})'}{2}\} \quad (2.11)$$

onde:

$dim$ é a dimensão do vetor $O_t$ .

$|U_{jm}|$ é o determinante da matriz covariância $U_{jm}$ .

$U_{jm}^{-1}$ é a matriz covariância inversa.

$\mu_{jm}$ é o vetor média.

Resumindo, um HMM, representado simbolicamente pela letra $\lambda$ , possui três elementos: π, A e B, isto é: $\lambda=(\pi, A, B)$ .

Para que o sinal de voz possa ser modelado por um HMM ainda é preciso mais uma consideração. Sendo o sinal de fala dinâmico e progressivo, as transições entre os estados do HMM irão ocorrer somente num sentido, isto é, não há possibilidade de retorno entre os estados. Nessa consideração, o estado inicial é sempre o estado mais à esquerda do modelo, e a cada transição somente poderá ser ocupado o próximo estado à direita ou permanecer no mesmo estado em que já se encontra. Esse processo é conhecido como esquerda-direita (left-right) [9].

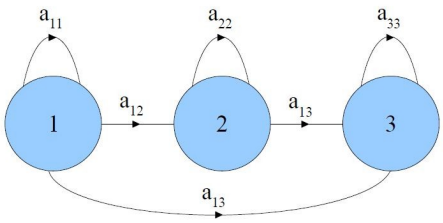


**Figura 4** - *Modelo HMM left-right*

Nesse modelo, temos as seguintes características de acordo com as definições anteriores:

$$a_{ij} = 0 \qquad i > j \tag{2.12}$$

$$\pi = \begin{Bmatrix} 1 & i = 1 \\ 0 & i \neq 1 \end{Bmatrix} \tag{2.13}$$

E para evitar grandes mudanças nos índices dos estados, foi usada a seguinte restrição:

$$a_{ij} = 0, \qquad j > i + \Delta \tag{2.14}$$

onde:

$\Delta$ no exemplo é igual a 2.

Um resumo de como é feito o treinamento de um HMM pode ser encontrado no Anexo A.

# CAPÍTULO 3

# ADAPTAÇÃO AO LOCUTOR

## 3.1 – Introdução

Existem duas formas de melhorar o desempenho de um SI: Adaptação ao locutor e Normalização de locutor. Frequentemente elas são interpretadas como sendo a mesma coisa. Entretanto há uma diferença entre elas.

Na Normalização de locutor [11], todos os locutores usados no re-treinamento do sistema são normalizados, iterativamente, em relação a um locutor médio. O objetivo principal é tentar normalizar as representações paramétricas do sinal de voz, de modo a reduzir os efeitos causados pela variabilidade da fala entre diferentes locutores [12][13]. Entende-se por variabilidade da fala não só as diferenças de comprimento de trato vocal dos locutores, mas também as diferenças linguísticas (sotaque, dialeto) e as condições físicas e emocionais destes locutores. A variabilidade da fala é uma das principais responsáveis pela degradação de desempenho dos sistemas de reconhecimento de fala.

A técnica de distorção (warping) do eixo de frequências é a que proporciona melhores resultados quanto a transformação dos parâmetros acústicos na normalização de locutor. Essa transformação é possível através do escalonamento, feito de forma linear, do sinal de fala no domínio da frequência.

O escalonamento pode ser feito através da compressão ou expansão do sinal de fala, no domínio da frequência, e em seguida, re-amostrando o sinal de fala no

domínio do tempo [12], ou, através da compressão ou expansão do banco de filtros, na escala Mel, na obtenção dos coeficientes Mel Cepstrais, não havendo necessidade de re-amostragem do sinal de fala [13].

Segundo Lee e Rose [13], a segunda maneira apresenta um mecanismo mais simples de implementação e proporciona uma eficiente melhora no desempenho do SRF.

No método proposto em [13], o escalonamento do banco de filtros no eixo das frequências é feito por um fator, chamado fator de distorção. Esse deve ser escolhido de modo que a probabilidade de um conjunto de características acústicas, de um determinado locutor, seja maximizada em relação a um dado modelo acústico tomado como referência.

Desta forma, o fator de distorção ótimo pode ser estimado, para cada locutor, pela máxima probabilidade de se obter um conjunto de características acústicas, dado um modelo λ e as transcrições de todas as locuções desse locutor, ou seja:

$$\hat{\alpha}^i = \underset{\alpha}{argmax}\, P(X_i^{\alpha} / \lambda, U_i) \qquad (3.1)$$

onde:

$X_i^{\alpha} = \{X_{i,1}^{\alpha}, X_{i,2}^{\alpha}, \cdots X_{i,N}^{\alpha}\}$ representa o conjunto de características acústicas de todas as N locuções do locutor i, escalonadas de α.

$U_i = \{U_{i,1}, U_{i,2}, \cdots, U_{i,N}\}$ representa o conjunto de transcrições de todas as N locuções associadas a um dado locutor i.

λ representa o modelo HMM usado como referência.

No trabalho desenvolvido por Dias [11], esse fator de distorção ótimo variou entre os valores de 0,88 e 1,12. Sendo assim, foram obtidos novos bancos de filtros com frequências escalonadas através da expressão:

$$f' = \frac{f}{\alpha} \qquad (3.2)$$

onde:

$f$ representa a frequência original na escala Mel.

$f'$ representa a frequência original na escala Mel.

$\alpha$ representa o fator de distorção.

Portanto, na normalização de locutor, a mudança ocorre na geração dos parâmetros e coeficientes Mel Cepstrais.

A idéia de adaptação ao locutor basicamente consiste em transformar um SI em um SD, usando uma quantidade de dados menor do que a que seria usada para treinar um SD por completo [14]. A maioria das técnicas usam como ponto de partida os HMMs de um SI e com algumas informações do locutor a ser adaptado, transformam um ou mais parâmetros dos HMMs originais, conseguindo melhorar o desempenho do SI para esse locutor adaptado. A figura a seguir ilustra o processo de adaptação ao locutor.

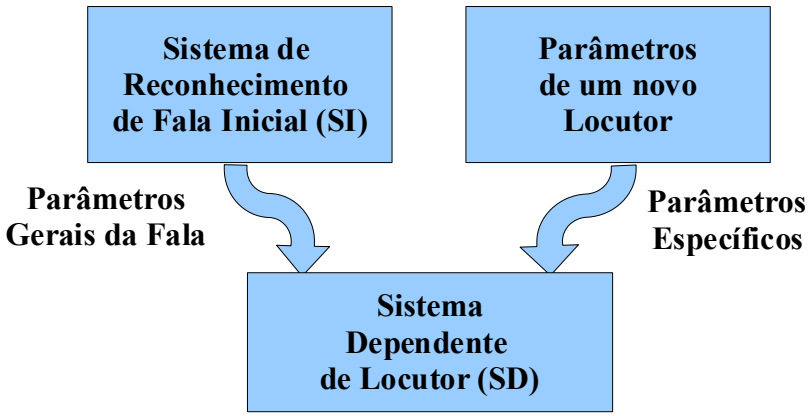


**Figura 5** - *Processo de Adaptação ao Locutor*

Existem várias técnicas de adaptação ao locutor [14]. O princípio básico de adaptação ao locutor, independente da técnica utilizada, é mostrado na figura a seguir:

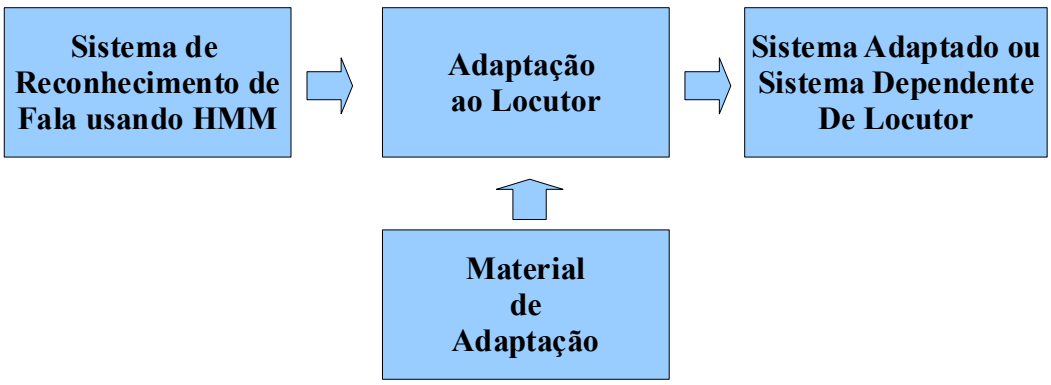


**Figura 6** - *Princípio Básico de Adaptação ao Locutor*

Vários fatores influenciam na escolha da técnica de adaptação [15][16]. Entre eles a quantidade de dados de adaptação e a velocidade de adaptação. Em relação aos dados de adaptação temos a seguinte classificação:

- Adaptação Supervisionada: os dados de adaptação são conhecidos previamente.

- Adaptação Não-Supervisionada: os dados de adaptação são obtidos à medida que o sistema é adaptado.

Em relação à velocidade de adaptação:

- Adaptação Estática: é o modo de adaptação geralmente mais usado, onde uma certa quantidade de dados é fornecida separadamente como material de adaptação e o sistema é adaptado antes de começar a ser usado pelo locutor.

- Adaptação Dinâmica: neste modo de adaptação o sistema de reconhecimento de fala se adapta continuamente ao usuário. Quando uma locução é dita por um usuário o sistema tenta reconhecê-la. Em paralelo a isto, uma adaptação do sistema é realizada para melhorar o desempenho deste, antes da próxima locução ser dita. Este procedimento é repetido para cada locução.

As figuras 7 e 8 ilustram os dois últimos casos:

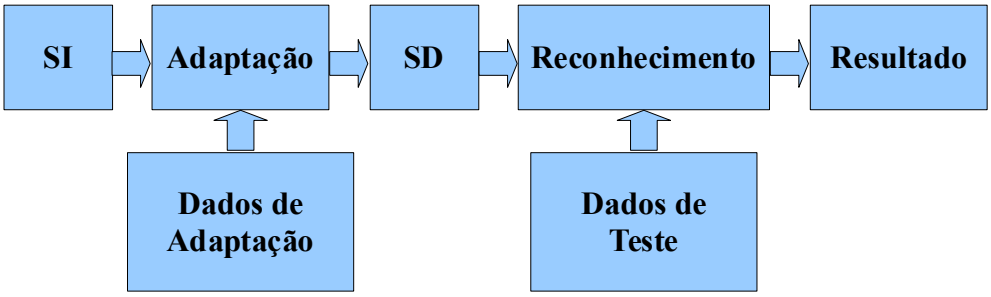


**Figura 7** - *Adaptação Estática*

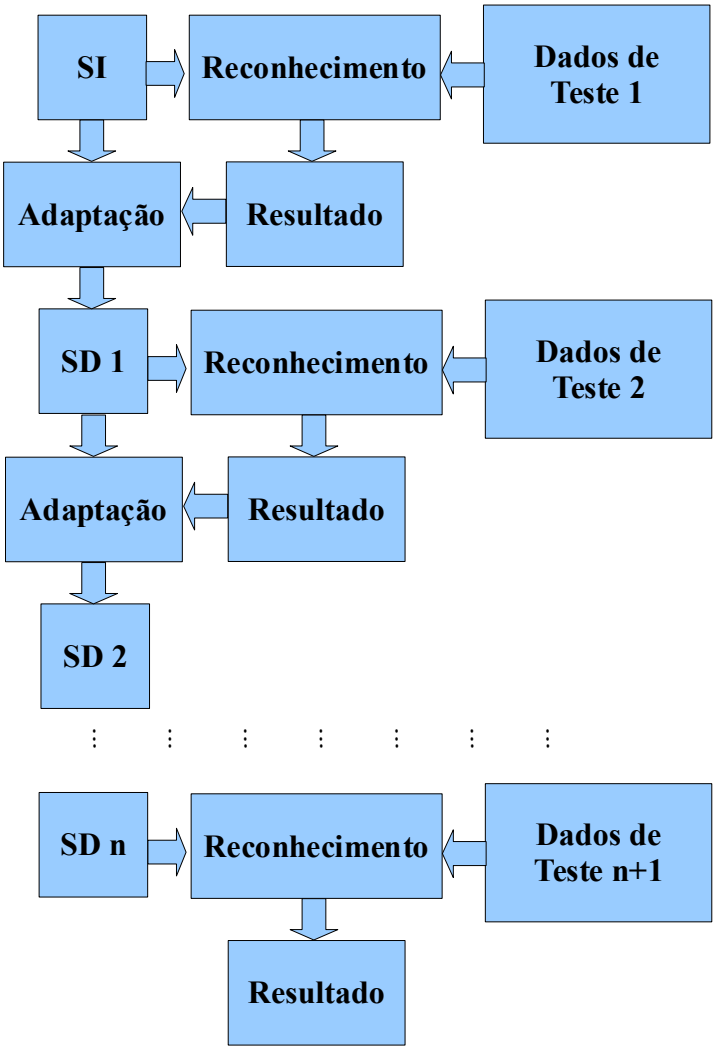


**Figura 8** - *Adaptação Dinâmica*

## 3.2 – Técnicas de Adaptação

As técnicas de adaptação ao locutor aplicadas utilizando-se os HMMs podem ser divididas em três famílias [14]:

- Família MAP (do inglês “Maximum a Posteriori”)
- Família das Transformações Lineares
- Família de Agrupamento de Locutores

Podemos citar como exemplo da família MAP, a técnica que lhe deu o nome, *MAP,* da família de transformações lineares, a *MLLR,* Regressão Linear de Máxima Verossimilhança (do inglês, Maximum Likelihood Linear Regression) e, por fim, da família de agrupamento de locutores, a técnica *Eigenvoices*.

### 3.2.1 – MAP

De acordo com [17][18][19][20] trata-se da técnica mais clássica de adaptação. Tenta maximizar a probabilidade a posteriori, e é baseada na estimação de máxima verossimilhança.

A técnica MAP estima os parâmetros do novo modelo, chamado modelo adaptado, de tal forma que a verossimilhança deste, dado o material de adaptação, seja máxima. Para isso, é feita uma combinação de informações a *priori* provenientes de um modelo independente de locutor, com os dados de adaptação de um novo locutor. Os novos parâmetros (parâmetros adaptados) são estimados e estabelecidos segundo a probabilidade a *posteriori*, daí a origem do nome:

$$P[\lambda|O]=\frac{P[O|\lambda]P_O[\lambda]}{P[O]} \qquad (3.3)$$

onde:

$P_O[\lambda]$ é a probabilidade a priori do modelo independente de locutor usado como

referência, $\lambda$ , antes da observação de qualquer sequência *O*.

Portanto, o vetor média de uma dada gaussiana pode ser reestimado através da expressão:

$$\hat{\mu} = \frac{\tau \mu_o + \sum_{t=1}^{T} \gamma(t) O_t}{\tau + \sum_{t=1}^{T} \gamma(t)} \qquad (3.4)$$

onde:

$\hat{\mu}$ é o vetor média adaptado.

$\mu_o$ é o vetor média a *priori* proveniente do modelo independente de locutor usado como referência.

$\tau$ indica a "inclinação"entre a estimação de máxima verosimilhança da média e a média a *priori*, chamado de meta parâmetro.

$\gamma(t)$ é a probabilidade da gaussiana analisada gerar a sequência de observação $O_t$ .

Essa técnica apresenta como desvantagem uma taxa de convergência lenta e não deve ser utilizada em adaptações onde a quantidade de material de adaptação seja pequena [14].

### 3.2.2 – MLLR

Outra técnica muito utilizada em adaptação ao locutor [21][22]. Nesse caso, transformações lineares de pelo menos algum dos parâmetros dos modelos HMMs dos SIs são realizadas em conjunto com o material de adaptação, gerando um novo modelo adaptado. A grande vantagem é que a mesma transformação pode ser utilizada para transformar, por exemplo, várias gaussianas de um HMM, o que permite uma adaptação mais rápida quando comparada com a técnica MAP [14].

A transformação linear é feita através de uma matriz de transformação obtida de tal forma que a verossimilhança do novo modelo em relação ao material de adaptação seja maximizada. Nas duas técnicas descritas, MAP e MLLR, a convergência cresce à medida que aumenta-se a quantidade do material de adaptação. A técnica MLLR será detalhada no próximo capítulo.

### 3.2.3 – Eigenvoices

A adaptação Eigenvoices é baseada na técnica de processamento de imagem, chamada Eigenface [23], usualmente aplicada como um método de compressão de imagem. A idéia principal é reduzir a dimensão das variáveis de dados, mantendo a maior variabilidade nas variáveis de dados restantes, porque a maior parte dessas têm uma alta correlação entre si [24][25][26]. Como ponto de partida necessita de um conjunto de locutores, chamados de locutores base. Um SD para cada um desses locutores deve ser treinado. Esses modelos irão compor o espaço de locutores de referência.

Para um dado conjunto de diferentes SDs, a análise de componente principal (PCA) [27] irá definir as direções onde se encontra a maior variação dos dados. Tais direções serão chamadas de componentes principais ou autovozes. Essas direções irão representar o novo locutor, realizando a adaptação dos HMMs[28].

# CAPÍTULO 4

# TÉCNICA DE ADAPTAÇÃO MLLR

## 4.1 - Definições

O objetivo da adaptação ao locutor é conseguir melhorar o desempenho de um sistema de reconhecimento de voz, utilizando apenas algumas locuções de um locutor específico. A maioria das técnicas de adaptação usa como ponto de partida os HMMs de um SI e com algumas informações do locutor a ser adaptado, transformam um ou mais parâmetros dos HMMs originais, conseguindo melhorar o desempenho do sistema para esse locutor adaptado.

De acordo com o capítulo 2, um modelo HMM contínuo apresenta uma probabilidade de emissão de símbolos através de uma função densidade de probabilidade gaussiana multidimensional, ou uma soma de duas ou mais funções densidade de probabilidade gaussianas multidimensionais. Por facilidade, a partir de agora, chamaremos de *gaussiana* ao nos referirmos a uma função densidade de probabilidade gaussiana e de *mistura* ao nos referirmos a soma de duas ou mais gaussianas.

A técnica Regressão Linear de Máxima Verossimilhança (MLLR – Maximum Likelihood Linear Regression) faz uma transformação nas médias das componentes gaussianas dos HMMs [22][29][30][31]. Os demais parâmetros do modelo permanecem os mesmos, sem nenhuma alteração.

Idealmente todos os parâmetros deveriam ser transformados. Entretanto, em muitas aplicações, a quantidade de material disponível para a adaptação é limitado, tornando-se inviável a alteração de todos os parâmetros. De acordo com [30], a alteração nas probabilidades de transição de estados e no peso das misturas quase não apresenta melhora no desempenho do sistema. Não é o caso das matrizes de covariância. Porém, é melhor deixá-las inalteradas do que modificá-las com as mesmas transformações utilizadas para as médias. Caso seja possível gerar uma matriz de transformação separada para as matrizes de covariância, isso pode trazer uma melhora no desempenho do sistema.

O vetor média para adaptação das misturas é calculado multiplicando-se o vetor média original (proveniente de um SI, por exemplo) por uma matriz de transformação, da seguinte forma:

$$\hat{\mu}_s = W_s \cdot \xi_s \tag{4.1}$$

onde:

$W_s$ é a matriz de transformação da componente *s* de uma mistura.

$\hat{\mu}_s$ é o vetor média adaptado.

$\xi_s$ é o vetor média estendido para a componente *s* de uma mistura.

Entende-se por vetor média estendido o seguinte vetor:

$$\xi_s = [w, \mu_{s1}, \mu_{s2}, \cdots, \mu_{sn}]' = [w : \mu_s] \tag{4.2}$$

onde:

$\mu_s$ é o vetor média original (de dimensão n)

$w$ é um termo de offset, somente utilizado quando as amostras do material de adaptação forem obtidas em ambientes com características diferentes[29].

4.2 – Classes de Regressão

Uma classe de regressão é um conjunto de médias das componentes

gaussianas a serem adaptadas usando-se a mesma matriz de transformação. É importante ressaltar que uma componente gaussiana de um mesmo estado de um modelo HMM pode ter suas médias divididas em diferentes classes de regressão.

O ponto principal da técnica MLLR é determinar quantas classes de regressão deverão ser utilizadas e quais médias das componentes gaussianas irão pertencer a cada classe.

Teoricamente, o número de classes de regressão pode variar de uma única classe para todas as médias das componentes gaussianas ou uma classe para cada componente gaussiana. O que vai limitar essa quantidade de classes de regressão é o material disponível para adaptação.

Duas diferentes formas de divisão das médias das componentes gaussianas em classes de regressão foram estudadas [30]:

- Divisão por características fonéticas.
- Divisão usando medidas de distâncias.

Nesse trabalho foi feito um estudo considerando as duas formas acima como será mostrado no próximo capítulo.

## 4.3 – Determinação da Matriz de Transformação

A matriz de transformação $\hat{W}_s$ a ser utilizada para adaptar as médias das componentes gaussianas nos HMMs, é obtida seguindo o critério de máxima verossimilhança. Conforme mencionado anteriormente, uma classe de regressão é o conjunto R de componentes gaussianas representado por:

$$s_{1k}, s_{2k}, \cdots, s_{Rk} \qquad 1 \leqslant k \leqslant M \qquad (4.3)$$

onde:

$M$ é o número de componentes gaussianas pertencentes a uma classe de regressão.

Seguindo o princípio de máxima verossimilhança, $\hat{W}_s$ é obtido como sendo a matriz que maximiza a probabilidade de geração das locuções do material de adaptação dado o HMM adaptado, $\hat{\lambda}$ :

$$\hat{W}_s = \max_{W_s} P(O_p|\hat{\lambda}) \tag{4.4}$$

onde:

$O_p$ é o conjunto de $P$ locuções $O^{(1)}, O^{(2)}, \cdots, O^{(P)}$ , cada uma com um número de quadros, $T$. Sendo assim: $O^{(p)} = O_1^{(p)}, \cdots, O_{Tp}^{(p)}$ .

Para resolver a equação (4.4) usamos o mesmo procedimento utilizado na reestimação das equações usadas no treinamento dos HMMs conhecido como algoritmo Baum-Welch. Definimos então, uma função auxiliar:

$$Q(\lambda, \hat{\lambda}) = \sum_{\theta \in \Theta} \sum_{K \in \Omega_b} P(O_P, \theta, K|\lambda) \log(P(O_p, \theta, K|\hat{\lambda})) \tag{4.5}$$

onde:

$K$ representa uma dada sequência de misturas.

$\Omega_b$ é o conjunto de todas as sequências de misturas possíveis.

$\theta$ representa uma sequência de estados.

$\Theta$ conjunto de todas as sequências de estados possíveis.

Definida dessa forma, a função auxiliar tem como propriedade, que se for encontrado um modelo adaptado $\hat{\lambda}$ que a maximize, a seguinte relação deve ser satisfeita [32]:

$$Q(\lambda, \hat{\lambda}) \geqslant Q(\lambda, \lambda) \Rightarrow P(O_p|\hat{\lambda}) \geqslant P(O_p|\lambda) \qquad (4.6)$$

Portanto devemos determinar a matriz de transformação maximizando a função auxiliar em relação a $\hat{W}_s$ . Vamos então expressar $P(O_p, \theta, K|\hat{\lambda})$ como uma função dos parâmetros probabilidade de transição e probabilidade de emissão de símbolos, onde ocorre a matriz $\hat{W}_s$ . O resultado encontrado é o seguinte:

$$\sum_{p=1}^{P} \sum_{t=1}^{T_p} \sum_{r=1}^{R} \gamma_{s_r k}^{(p)}(t) \sum_{s_r k}^{-1} o_t^{(p)} \xi'_{s_r k} = \sum_{p=1}^{P} \sum_{t=1}^{T_p} \sum_{r=1}^{R} \gamma_{s_r k}^{(p)}(t) \sum_{s_r k}^{-1} o_t^{(p)} \hat{W}_s \xi_{s_r k} \xi'_{s_r k} \quad (4.7)$$

onde:

$\sum_{s_r k}$ é a matriz de covariância para a mistura $s_r k$ .

$\gamma_{s_r k}^{(p)}(t)$ é a probabilidade de ocupação da mistura $s_r k$ definida por:

$$\gamma_{s_r k}^{(p)}(t) = \frac{1}{P(O^{(p)}|\lambda)} \sum_{\theta \in \Theta} \sum_{K \in \Omega_b} P(O^{(p)}, \theta_t = s_r, k_t = k|\lambda) \qquad (4.8)$$

O próximo passo é reescrever a equação (4.7) usando duas matrizes, definidas por $V^{(r)}$ e $D^{(r)}$ . São definidas pelas seguintes equações:

$$V^{(r)} = \sum_{p=1}^{P} \sum_{t=1}^{T_p} \gamma_{s_r k}^{(p)}(t) \sum_{s_r k}^{-1} \qquad (4.9)$$

$$D^{(r)} = \xi_{s_r k} \xi'_{s_r k} \qquad (4.10)$$

É muito comum se trabalhar com a matriz de covariância diagonal na realização de testes pois a matriz cheia demanda um esforço computacional elevado e o ganho em desempenho é marginal [4]. Sendo assim, observa-se que quando a utilizamos dessa forma, a matriz $V^{(r)}$ torna-se também diagonal. Além disso a matriz $D^{(r)}$ é uma matriz simétrica.

Portanto, a equação (4.7) fica da seguinte forma:

$$\sum_{p=1}^{P}\sum_{t=1}^{T_p}\sum_{r=1}^{R}\gamma_{s_r k}^{(p)}(t)\sum\nolimits_{s_r k}^{-1} o_t^{(p)}\xi'_{s_r k}=\sum_{r=1}^{R} V^{(r)}\hat{W}_s D^{(r)} \tag{4.11}$$

Definimos por *Y* a matriz que irá representar o lado direito da equação (4.11) . Seus elementos são calculados por:

$$y_{ij}=\sum_{p=1}^{n}\sum_{q=1}^{n+1} w_{pq}\left[\sum_{r=1}^{R} v_{ip}^{(r)} d_{qj}^{(r)}\right] \tag{4.12}$$

Considerando que a matriz $V^{(r)}$ é diagonal e a matriz $D^{(r)}$ é simétrica, podemos simplificar a equação (4.12) para a seguinte forma:

$$y_{ij}=\sum_{q=1}^{n+1} w_{iq}[\sum_{r=1}^{R} v_{ii}^{(r)} d_{jq}^{(r)}] \tag{4.13}$$

A multiplicação das matrizes $V^{(r)}$ e $D^{(r)}$ gera uma nova matriz representada por *G*, conforme abaixo:

$$G^{(i)}=[g_{jq}^{(i)}]=\begin{bmatrix} \sum_{r=1}^{R} v_{ii}^{(r)} d_{11}^{(r)} & \sum_{r=1}^{R} v_{ii}^{(r)} d_{12}^{(r)} & \cdots & \sum_{r=1}^{R} v_{ii}^{(r)} d_{1n+1}^{(r)} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{r=1}^{R} v_{ii}^{(r)} d_{n+11}^{(r)} & \sum_{r=1}^{R} v_{ii}^{(r)} d_{n+12}^{(r)} & \cdots & \sum_{r=1}^{R} v_{ii}^{(r)} d_{n+1n+1}^{(r)} \end{bmatrix} \tag{4.14}$$

Finalmente, usamos a matriz Z para representar os elementos do lado esquerdo da equação (4.11), conforme a seguir:

$$Z=\sum_{p=1}^{P}\sum_{t=1}^{T_p}\sum_{r=1}^{R}\gamma_{s_r k}^{(p)}(t)\sum\nolimits_{s_r k}^{-1} o_t^{(p)}\xi'_{s_r k} \tag{4.15}$$

Simplificando, temos que:

$$z_{ij}=y_{ij}=\sum_{q=1}^{n+1} w_{iq} g_{jq}^{(i)} \tag{4.16}$$

Então, a matriz é determinada linha a linha pelo produto:

$$\hat{w}'_i = G^{(i)-1} z'_i \tag{4.17}$$

onde:

$\hat{w}_i$ e $z_i$ são a i-ésima linha da matriz $\hat{W}_s$ e $Z$ respectivamente.

# CAPÍTULO 5

# TESTES REALIZADOS E RESULTADOS

## 5.1 – Base de Dados

Para os testes realizados neste trabalho foi usada uma base de dados composta por 20 listas de 10 frases foneticamente balanceadas, segundo o português falado no Rio de Janeiro [33]. Estas foram escolhidas segundo o trabalho realizado por Alcaim et.al. [34]. Neste conjunto de frases tem-se um total de 694 palavras distintas.

Para as gravações foram selecionados 40 locutores adultos, sendo 20 homens e 20 mulheres. Cada locutor pronunciou 40 frases distintas e cada frase foi repetida por 8 vezes por diferentes locutores. Portanto, obteve-se um material de treinamento de 1600 locuções. A taxa de amostragem utilizada foi de 11,025 kHz e resolução de 16 bits. Esta base será chamada de *Base de Treinamento*.

Conforme mencionado no Capítulo anterior, a adaptação ao locutor tem como objetivo melhorar o desempenho de um SI para um locutor específico. Isso é feito através de algumas amostras de voz desse locutor. Após a adaptação, o que chamamos de Sistema Adaptado (SA) fica com um desempenho próximo ao de um SD treinado por completo. Portanto, para efeito de comprovação da eficácia da adaptação MLLR foram selecionados mais quatro locutores adultos, sendo três homens e uma mulher. Esses locutores serão chamados daqui pra frente de *locutores de referência*.

Cada locutor de referência fez a gravação das 200 frases 4 vezes, formando uma segunda base de dados, com 800 locuções, que será referida como *Base de Teste.*

A transcrição fonética de ambas as bases de dados foi feita manualmente para cada locução, utilizando-se um programa de visualização gráfica de espectograma e forma de onda do sinal e fones de ouvido para audição da mesma. As subunidades utilizadas (36 ao todo) são mostradas na tabela 2.

**Tabela 2** - *Fones utilizados na Transcrição Fonética das Locuções*

| Fones | Exemplo de palavra em que o fone aparece |
|---|---|
| # | Fone silêncio |
| a | ***a***ç***a***frão |
| an | maç***ã*** |
| e | ***e***l***e***vador |
| E | p***e***le |
| en | s***en***ta |
| i | s***i***no |
| y | fu***i*** |
| in | p***in***to |
| o | b***o***lo |
| O | b***o***la |
| on | s***om***bra |
| u | l***u***a |
| un | ***um*** |
| b | ***b***ela |
| d | ***d***ádiva |
| D | ***d***iferente |
| f | ***f***eira |
| g | ***g***orila |
| j | ***j***iló |

| | |
|---|---|
| k | ***c***achoeira |
| l | ***l***eão |
| L | ***lh***ama |
| m | ***m***ontanha |
| n | ***n***évoa |
| N | i***nh***ame |
| p | ***p***oente |
| r | ce***r***a |
| rr | ce***rr***ado |
| R | ca***r***ta |
| s | ***s***apo |
| t | ***t***empes***t***ade |
| T | ***t***igela |
| v | ***v***erão |
| x | ***ch***ave |
| z | ***z***abumba |

## 5.2 – Sistema de Reconhecimento

Para os testes foi utilizado um software desenvolvido por Ynoguti [33], baseado em modelos ocultos de Markov contínuos. Cada subunidade fonética descrita anteriormente, foi modelada por um HMM de 3 estados. O sistema de reconhecimento utiliza fones independentes do contexto como unidades fundamentais, o One Pass como algoritmo de busca. Foram utilizados os parâmetros mel-cepstrais (12 coeficientes) com suas respectivas derivadas primeira e segunda (parâmetros delta e delta-delta, 12 coeficientes cada um com 1 janela de cada lado). Como modelo de linguagem, foi adotada uma gramática do tipo bigrama.

## 5.3 – Sistemas Independentes e Dependentes de Locutor

Para a realização dos testes foram criados cinco SIs. A diferença entre os cinco sistemas testados foi somente nos parâmetros utilizados e no número de gaussianas em cada estado de cada HMM. A tabela a seguir descreve as características de cada um.

**Tabela 3 -** *Características dos SIs usados na Realização dos Testes*

| Sistema | Parâmetros utilizados | Número de Gaussianas |
|---|---|---|
| 1 | Mel | 1 |
| 2 | Mel | 3 |
| 3 | Mel, Delta Mel, Delta Delta Mel (1 janela) | 3 |
| 4 | Mel | 5 |
| 5 | Mel, Delta Mel, Delta Delta Mel (1 janela) | 5 |

Para comparar a taxa de desempenho também foram utilizados SDs com as mesmas características em relação aos parâmetros e o número de gaussianas de cada estado de cada HMM dos SIs. Ou seja, foram criados cinco SDs para cada um dos quatro locutores de referência, totalizando 20 SDs.

Os cinco SIs foram treinados com as 1600 locuções dos 40 locutores da Base de Treinamento e cada SD de cada locutor de referência, foi treinado com 600 das 800 locuções da Base de Teste. Os resultados foram obtidos através das 200 locuções restantes de cada locutor de referência da Base de teste.

As figuras abaixo ilustram o material de treinamento utilizado nos SIs e nos SDs.

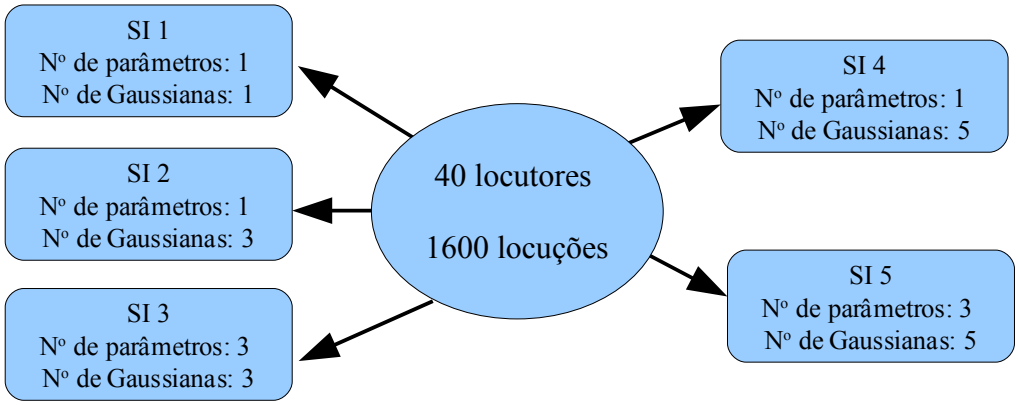


**Figura 9 -** *Material de Treinamento usado nos SIs*

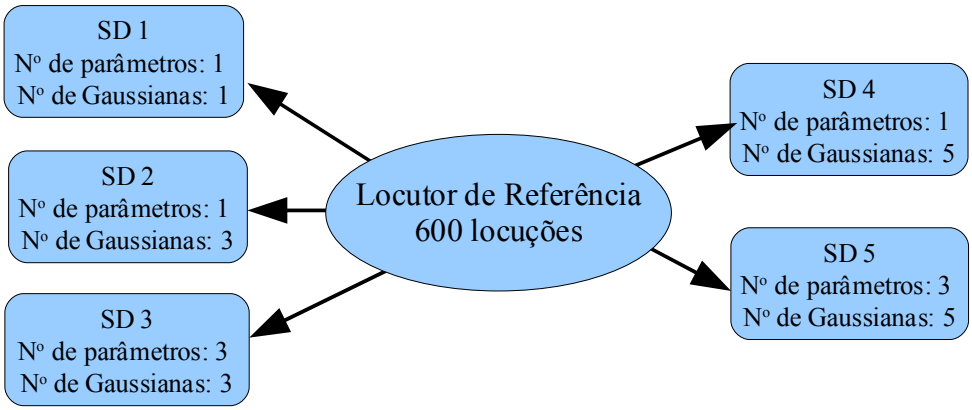


**Figura 10 -** *Material de Treinamento usado em cada SD para cada Locutor de Referência*

Conforme mencionado anteriormente o desempenho de um SI é menor quando comparado ao SD. Utilizando-se os sistemas representados nas figuras 9 e 10, foi possível realizar essa comparação para os quatro locutores de referência. O resultado é apresentado na tabela a seguir:

**Tabela 4 -** *Comparação entre os SIs e os SDs para os quatro Locutores de Referência*

| Sistema | Locutor 1 | | Locutor 2 | | Locutor 3 | | Locutor 4 | |
|---|---|---|---|---|---|---|---|---|
| | SI | SD | SI | SD | SI | SD | SI | SD |
| 1 | 41,30% | 75,40% | 73,40% | 86,70% | 52,60% | 85,90% | 44,20% | 84,70% |
| 2 | 45,70% | 81,00% | 77,40% | 92,00% | 55,00% | 89,40% | 45,60% | 88,50% |
| 3 | 59,90% | 86,20% | 89,50% | 94,60% | 75,90% | 93,30% | 68,90% | 93,40% |
| 4 | 45,10% | 83,30% | 77,30% | 91,70% | 53,70% | 93,10% | 46,50% | 87,40% |
| 5 | 63,30% | 87,50% | 90,10% | 95,10% | 73,30% | 94,80% | 65,20% | 92,90% |

Os valores apresentados como desempenho dos sistemas em todos os testes realizados neste trabalho, foram obtidos através de uma ferramenta que faz parte do pacote SCTK desenvolvido pelo NIST chamado SCLITE [35]. Nesta, faz-se uma comparação entre a locução a ser reconhecida com sua respectiva locução de referência em um processo chamado de alinhamento. A métrica utilizada foi a taxa de erros de palavras (do inglês, Word Error Rate – WER), que calcula e avalia os tipos de erros conforme a equação:

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

onde:

$WER$ é a taxa de erro de palavras.

$N_I$ é o número de erros de inserção.

$N_S$ é o número de erros de substituição.

$N_D$ é o número de erros de deleção.

$N_r$ é o número de palavras presentes no conjunto de referência.

## 5.4 – Sistema desenvolvido para os Testes de Adaptação

Para realizar a adaptação usando a técnica MLLR foram implementados programas na linguagem C++ para a plataforma Windows. O sistema foi desenvolvido utilizando uma interface visual amigável e fácil de ser utilizada.

Inicialmente é criado um arquivo de configuração com as informações necessárias para se realizar a adaptação. Neste arquivo deve ser especificado:

- o número de classes de regressão que será utilizado.
- o nome do arquivo que traz as subunidades fonéticas utilizadas.
- o arquivo com os modelos HMMs do SI utilizado como referência.
- o material de adaptação: quais locuções e a transcrição fonética de cada uma.

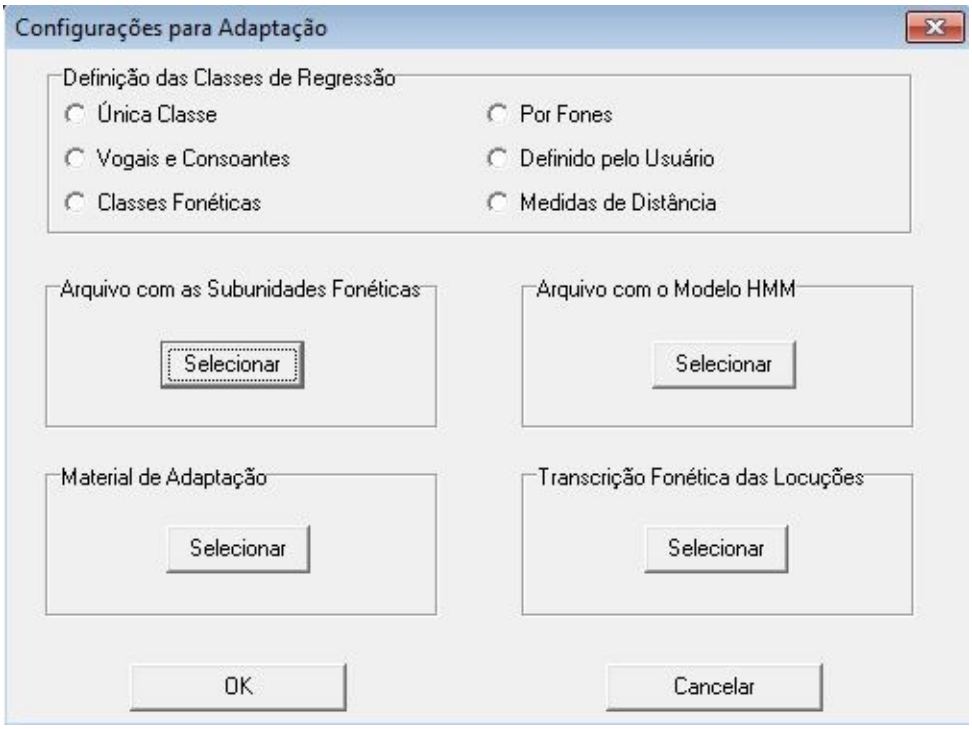


**Figura 11 -** *Interface Gráfica do sistema desenvolvido para especificar os parâmetros de Configuração da Adaptação MLLR*

Um exemplo das informações especificadas em um arquivo de configuração, pode ser visualizado na figura abaixo:

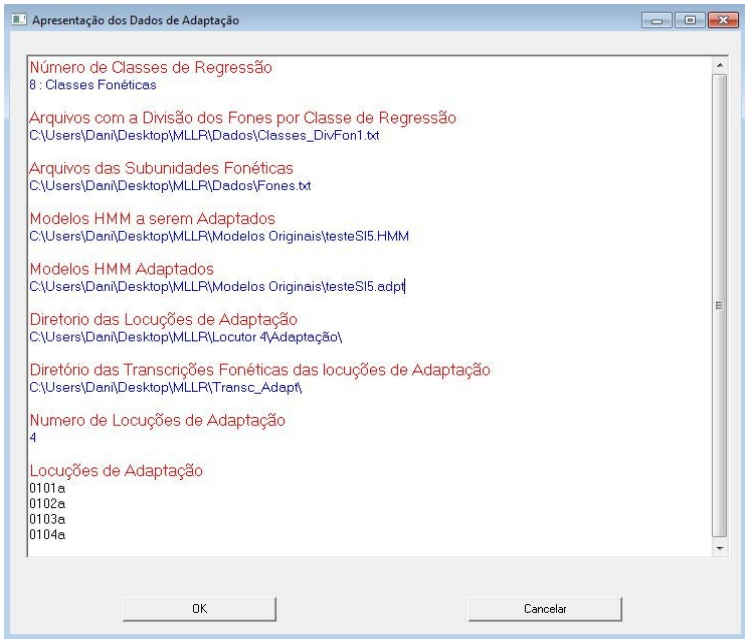


**Figura 12 -** *Exemplo dos parâmetros especificados no arquivo de configuração.*

Após a criação desse arquivo faz-se a adaptação desejada, selecionando no menu da tela principal o item *Adaptar*. Após a adaptação o programa é finalizado.

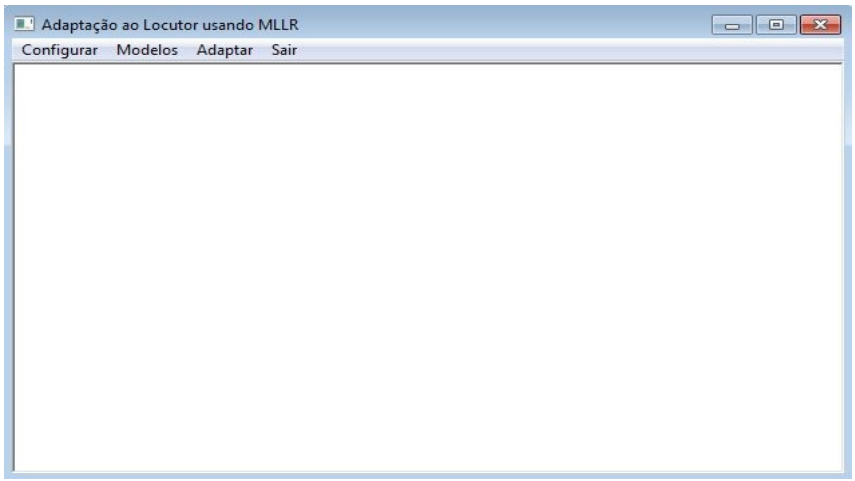


**Figura 13 -** *Tela Principal*

## 5.5 – Experimentos e Resultados

A partir de agora serão feitos testes referentes à adaptação MLLR. Podemos considerar um ganho de desempenho de cada SA se o resultado estiver entre o desempenho de um SI e o desempenho de um SD, de acordo com a tabela 4.

Conforme capítulo anterior, a Técnica de Adaptação MLLR somente transforma as médias das componentes gaussianas dos HMMs. Isso é feito através de uma matriz de transformação determinada utilizando-se material de adaptação e modelos HMMs de um SI. Outro ponto principal desta técnica de adaptação é o número de classes de regressão utilizada e a divisão das médias das componentes gaussianas em cada classe.

A divisão das médias das componentes gaussianas nas classes de regressão a serem utilizadas, foi feita através de duas formas:

- Divisão através das Características Fonéticas.
- Divisão usando Medidas de Distâncias.

Na divisão através das características fonéticas foram criadas quatro possibilidades de classes de regressão. E na divisão utilizando-se medidas de distância, duas possibilidades de classes de regressão, uma através da distância Euclidiana e outra através da distância Bhattacharya[36], conforme será descrito mais adiante.

Na divisão através das características fonéticas as classes de regressão criadas são descritas pela tabela 5.

**Tabela 5** - *Classes de Regressão utilizadas na Divisão por Classes Fonéticas*

| Número de Classes de Regressão | Divisão das Médias das Componentes das Misturas |
|---|---|
| 1 | Todas as médias |
| 3 | Vogais / Consoantes / Fone silêncio |
| 8 | Divisão fonética conforme as tabelas 6 e 7 abaixo |
| 36 | Uma para cada fone (subunidades descritas na tabela 2) |

Ao utilizarmos a divisão em oito classes de regressão foram feitas duas considerações. Na primeira, chamada *Divisão 1*, considerou-se o fone silêncio deixando-o numa classe de regressão. Os demais fones foram divididos nas sete classes de regressão restantes utilizando uma classificação fonética conforme especificado na tabela 6. Na segunda, *Divisão 2*, não se levou em conta o fone silêncio e os demais fones foram divididos nas oito classes, utilizando-se outra classificação fonética, conforme a tabela 7.

**Tabela 6** - *Divisão Fonética levando-se em conta o fone Silêncio*

| Divisão 1 | | |
|---|---|---|
| Classe de Regressão | Classe Fonética | Fones em cada Classe |
| 1 | Silêncio | 01 fone – # |
| 2 | Vogais Orais | 08 fones – a, e, E, i, y, o, O, u. |
| 3 | Vogais Nasais | 05 fones – an, en, in, on, un |
| 4 | Plosivas | 08 fones – p, t, T, k, b, d, D, g |
| 5 | Fricativas | 06 fones – f, s, x, v, z, j |
| 6 | Laterais | 02 fones – l, L |
| 7 | Nasais | 03 fones – n, m, N |
| 8 | Vibrantes | 03 fones – r, rr, R |

**Tabela 7** - *Divisão Fonética sem o fone Silêncio*

| Divisão 2 | | |
|---|---|---|
| Classe de Regressão | Classe Fonética | Fones em cada Classe |
| 1 | Vogais Anteriores | 06 fones – i, y, e, E, en, in |
| 2 | Vogais Médias | 02 fones – a, an |
| 3 | Vogais Posteriores | 05 fones – o, O, u, on, un |
| 4 | Consoantes Labiais | 05 fones – p, b, m, f, v |
| 5 | Consoantes Médias | 07 fones – t, T, d, D, n, s, z |
| 6 | Consoantes Posteriores | 05 fones – k, g, N, x, j |
| 7 | Laterais | 02 fones – l, L |
| 8 | Vibrantes | 03 fones – r, rr, R |

As medidas de distância utilizadas foram:

- Distância Euclidiana

$$d(\mu, \mu_i) = \sqrt{(\mu - \mu_i) \cdot (\mu - \mu_i)^t} \qquad (5.2)$$

onde:

$d(\mu, \mu_i)$ distância euclidiana entre os vetores $\mu, \mu_i$ .
$\mu$ vetor coluna que se deseja quantizar .
$\mu_i$ *i*-ésimo vetor do dicionário.

- Distância Bhattacharya [36]

$$d(\mu_i, \mu_j) = \frac{1}{8}(\mu_i - \mu_j)^t \left\{\frac{(\sum_i + \sum_j)}{2}\right\}^{-1} + \frac{1}{2}\ln\left(\left[\frac{\left|\sum_i + \sum_j\right|}{2}\right]\left[\left|\sum_i\right|^{\frac{1}{2}} \cdot \left|\sum_j\right|^{\frac{-1}{2}}\right]\right) \qquad (5.3)$$

onde:

$d(\mu_i, \mu_j)$ distância bhattacharya entre os vetores $\mu_i, \mu_j$ .
$\mu_i, \mu_j$ vetores utilizados para o cálculo da distância.
$\sum_i, \sum_j$ matrizes de covariância dos vetores *i* e *j*, respectivamente.

Para a divisão das médias das componentes gaussianas usando as medidas de distância primeiro foi estabelecido o número de classes de regressão a ser utilizado.

Para uma comparação com as divisões fonéticas feitas anteriormente, foi escolhido a quantidade de 8 classes de regressão. A partir daí, foi feito um levantamento dos fones mais próximos, utilizando as medidas de distância mencionadas, considerando oito divisões, seguindo o conceito de quantização vetorial [33] . As seguintes distribuições foram feitas:

**Tabela 8** - *Divisão das Médias das Componentes Gaussianas considerando a Distância Euclidiana*

| Distância Euclidiana | |
|---|---|
| Classe de Regressão | Fones em cada Classe |
| 1 | 03 fones – x, T, s |
| 2 | 05 fones – D, f, j, y, t |
| 3 | 04 fones – i, N, in, en |
| 4 | 04 fones – L, v, e, d |
| 5 | 05 fones – E, a, r, n, R |
| 6 | 04 fones – k, rr, p, un |
| 7 | 05 fones – O, o, on, an,u |
| 8 | 05 fones – m, b, g, l, z |

**Tabela 9** - *Divisão das Médias das Componentes Gaussianas considerando a Distância Bhattacharya*

| Distância Bhattacharya | |
|---|---|
| Classe de Regressão | Fones em cada Classe |
| 1 | 04 fones – in, en, i, n |
| 2 | 05 fones – o, O, on, u, un |
| 3 | 04 fones – L, e, R, m |
| 4 | 05 fones – x, s, T, D, j |
| 5 | 03 fones – a, an, rr. |
| 6 | 03 fones – N, E, l |
| 7 | 06 fones – f, p, v, z, b, d |
| 8 | 05 fones – k, t, g, y, r |

Os testes foram feitos com o objetivo de mostrar:

- a influência do tipo de SI utilizado quanto às características relacionadas aos parâmetros envolvidos e o número de gaussianas utilizadas em cada estado de cada HMM;

- a importância da escolha adequada do número de classes de regressão a ser utilizado;

- a influência da forma como é feita a divisão das componentes gaussianas nas classes de regressão;

- a importância da quantidade de material de adaptação utilizado.

A seguir temos a descrição das características de cada teste em relação às locuções de adaptação utilizadas, às classes de regressão e aos sistemas usados como referência. Os resultados obtidos são comparados com o desempenho dos SIs e SDs estabelecidos anteriormente que podem ser verificados na tabela 4.

1º Teste: Influência do SI utilizado como referência

Os cinco SIs foram adaptados utilizando-se as locuções de cada um dos quatro locutores de referência. O propósito deste teste foi verificar qual a influência das características iniciais do SI (parâmetros e número de gaussianas em cada estado do HMM) na adaptação do locutor.

Características:

- Três Classes de Regressão: vogais, consoantes e fone silêncio
- Quatro locuções de adaptação:

  *"A questão foi retomada no Congresso."*

  *"Leila tem um lindo jardim."*

  *"O analfabetismo é a vergonha do país."*

  *"A casa foi vendida sem pressa."*

Resultados:

**Tabela 10** - *Comparação entre os SAs e os SIs*

| Desempenho dos SAs | | | | | | |
|---|---|---|---|---|---|---|
| Sistema | Locutor 1 | | | Locutor 2 | | |
| | SI | SA | SD | SI | SA | SD |
| 1 | 41,30% | 46,20% | 75,40% | 73,40% | 78,10% | 86,70% |
| 2 | 45,70% | 48,80% | 81,00% | 77,40% | 80,50% | 92,00% |
| 3 | 59,90% | 65,40% | 86,20% | 89,50% | 90,90% | 94,60% |
| 4 | 45,10% | 49,50% | 83,30% | 77,30% | 74,90% | 91,70% |
| 5 | 63,30% | 67,20% | 87,50% | 90,10% | 89,10% | 95,10% |
| Sistema | Locutor 3 | | | Locutor 4 | | |
| | SI | SA | SD | SI | SA | SD |
| 1 | 52,60% | 55,20% | 85,90% | 44,20% | 46,20% | 84,70% |
| 2 | 55,00% | 60,30% | 89,40% | 45,60% | 60,50% | 88,50% |
| 3 | 75,90% | 79,60% | 93,30% | 68,90% | 75,50% | 93,40% |
| 4 | 53,70% | 64,20% | 93,10% | 46,50% | 58,80% | 87,40% |
| 5 | 73,30% | 80,10% | 94,80% | 65,20% | 71,60% | 92,90% |

Nota-se que em todos os casos o desempenho dos SAs ficou superior aos SIs e inferior aos SDs. Em todos os casos obteve-se um ganho no desempenho do sistema.

2º Teste: Quantidade de Material de Adaptação

Neste teste foram adaptados dois sistemas: 3 e 5, utilizando três classes de regressão. O objetivo era verificar a influência da quantidade de locuções de adaptação. Portanto foram testadas 1, 3, 4, 5, 6 e 10 frases de adaptação.

Características:

- Três Classes de Regressão: vogais, consoantes e fone silêncio
- Até dez frases de adaptação:

  *"A questão foi retomada no Congresso."*

  *"Leila tem um lindo jardim."*

*"O analfabetismo é a vergonha do país."*

*"A casa foi vendida sem pressa."*

*"Trabalhando com união rende muito mais."*

*"Recebi nosso amigo para almoçar."*

*"A justiça é a única vencedora."*

*"Isso se resolverá de forma tranquila."*

*"Os pesquisadores acreditam nessa teoria."*

*"Sei que atingiremos o objetivo."*

Resultados obtidos com o Sistema 3: Parâmetros Mel, Delta-Mel e Delta-Delta-Mel com 3 gaussianas em cada estado do HMM.

**Tabela 11** - *Comparação entre o SI e o SA variando-se a Quantidade do Material de Adaptação, utilizando-se o Sistema 3*

| Frases | Locutor 1 | Locutor 2 | Locutor 3 | Locutor 4 |
|---|---|---|---|---|
| | SI – 59,90% | SI – 89,50% | SI – 75,90% | SI – 68,90% |
| | SD – 86,20% | SD – 94,60% | SD – 93,30% | SD – 93,40% |
| 1 | 55,80% | 89,40% | 73,60% | 67,00% |
| 3 | 64,20% | 90,70% | 83,80% | 75,90% |
| 4 | 64,60% | 90,60% | 82,70% | 76,10% |
| 5 | 68,70% | 90,10% | 82,60% | 72,50% |
| 6 | 70,10% | 90,50% | 83,20% | 73,10% |
| 10 | 69,10% | 91,40% | 83,20% | 76,10% |

Resultados obtidos com o Sistema 5: Parâmetros Mel, Delta-Mel e Delta-Delta-Mel com 5 gaussianas em cada estado do HMM.

**Tabela 12** - *Comparação entre o SI e o SA variando-se a Quantidade do Material de Adaptação, utilizando-se o Sistema 5*

| Frases | Locutor 1 | Locutor 2 | Locutor 3 | Locutor 4 |
|---|---|---|---|---|
| | SI – 63,30% | SI – 90,10% | SI – 73,30% | SI – 65,20% |

| | SD – 87,50% | SD – 95,10% | SD – 94,80% | SD – 92,90% |
|---|---|---|---|---|
| 1 | 66,00% | 88,70% | 75,80% | 67,90% |
| 3 | 67,10% | 91,80% | 80,70% | 71,60% |
| 4 | 67,20% | 89,10% | 80,10% | 71,60% |
| 5 | 67,80% | 90,00% | 79,50% | 71,90% |
| 6 | 67,70% | 91,20% | 79,30% | 74,20% |
| 10 | 68,80% | 91,90% | 79,80% | 74,60% |

É possível verificar que a quantidade de material disponível para adaptação faz diferença. Utilizando-se somente uma locução e o sistema 3 em nenhum dos locutores houve melhora do desempenho. Já com três frases nos dois sistemas, 3 e 5, em todos eles houve um aumento no desempenho do sistema.

3º Teste: Número de Classes de Regressão

Novamente foram adaptados os sistemas 3 e 5, sendo que 3 frases foram usadas como material de adpatação para o sistema 3 e 4 locuções para o sistema 5. Neste teste verificamos a influência do número de classes de regressão ao se dividir as médias das componentes gaussianas.

Características:

- Até quatro locuções de adaptação:
  *"A questão foi retomada no Congresso."*
  *"Leila tem um lindo jardim."*
  *"O analfabetismo é a vergonha do país."*
  *"A casa foi vendida sem pressa."*

Resultados obtidos com o Sistema 3: Parâmetros Mel, Delta-Mel e Delta-Delta-Mel com 3 gaussianas em cada estado do HMM (3 locuções de adaptação).

**Tabela 13** - *Comparação entre o SI e o SA variando-se o número de classes de regressão, utilizando o Sistema 3*

| Classes de Regressão | Locutor 1 | Locutor 2 | Locutor 3 | Locutor 4 |
|---|---|---|---|---|
| | SI – 59,90% | SI – 89,50% | SI – 75,90% | SI – 68,90% |
| | SD – 86,20% | SD – 94,60% | SD – 93,30% | SD – 93,40% |
| 1 | 65,60% | 90,80% | 82,50% | 72,40% |
| 3 | 64,20% | 90,70% | 83,80% | 75,90% |
| 8 (Div 1) | 61,20% | 84,60% | 81,10% | 66,30% |
| 8 (Div 2) | 41,50% | 56,20% | 80,40% | 62,60% |
| 36 | 43,10% | 72,60% | 52,60% | 62,10% |
| Dist Euclid. | 57,20% | 79,40% | 78,70% | 53,20% |
| Dist Bhattac. | 53,90% | 84,50% | 77,20% | 60,10% |

Resultados obtidos com o Sistema 5: Parâmetros Mel, Delta-Mel e Delta-Delta-Mel com 5 gaussianas em cada estado do HMM (4 locuções de adaptação).

**Tabela 14** - *Comparação entre o SI e o SA variando-se o número de classes de regressão, utilizando o Sistema 5*

| Classes de Regressão | Locutor 1 | Locutor 2 | Locutor 3 | Locutor 4 |
|---|---|---|---|---|
| | SI – 63,30% | SI – 90,10% | SI – 73,30% | SI – 65,20% |
| | SD – 87,50% | SD – 95,10% | SD – 94,80% | SD – 92,90% |
| 1 | 67,80% | 91,10% | 78,80% | 65,20% |
| 3 | 67,20% | 89,90% | 80,10% | 71,60% |
| 8 (Div 1) | 64,20% | 88,50% | 77,00% | 69,10% |
| 8 (Div 2) | 62,60% | 87,10% | 77,80% | 65,30% |
| 36 | 53,00% | 76,10% | 66,60% | 45,70% |
| Dist Euclid. | 62,90% | 85,20% | 78,20% | 62,30% |
| Dist Bhattac. | 57,00% | 87,10% | 78,50% | 65,60% |

Observa-se que à medida que aumentamos o número de classes de regressão

com uma pequena quantidade de material de adaptação, o desempenho do sistema torna-se pior. De fato, ao utilizarmos 36 classes de regressão, tanto no sistema 3 como no sistema 5 houve uma queda no desempenho.

### 4º Teste: Divisão das Médias das Componentes Gaussianas em Oito Classes de Regressão.

Somente sistema 5 foi adaptado dessa vez. Foram utilizadas oito classes de regressão para uma comparação entre as duas divisões fonéticas (divisão 1 e divisão 2) e as duas medidas de distância (Euclidicana e Bhattacharya). O objetivo é mostrar como a mudança de um fone de uma classe de regressão para outra altera o desempenho do sistema.

Características:

- Oito Classes de Regressão:
  Divisão Fonética 1 (Div 1)
  Divisão Fonética 2 (Div 2)
  Distância Euclidiana (Dist E)
  Distância Bhattacharya (Dist B)

- Até dez locuções de adaptação:
  *"A questão foi retomada no Congresso."*
  *"Leila tem um lindo jardim."*
  *"O analfabetismo é a vergonha do país."*
  *"A casa foi vendida sem pressa."*
  *"Trabalhando com união rende muito mais."*
  *"Recebi nosso amigo para almoçar."*
  *"A justiça é a única vencedora."*
  *"Isso se resolverá de forma tranquila."*
  *"Os pesquisadores acreditam nessa teoria."*
  *"Sei que atingiremos o objetivo."*

– Sistema 5: Parâmetros Mel, Delta-Mel e Delta-Delta-Mel com 5 gaussianas em cada estado do HMM

Resultados ao se utilizar 4 frases de adaptação:

**Tabela 15** - *Comparação entre o SI e o SA alterando-se a Divisão das Médias das Componentes Gaussianas nas Oito Classes de Regressão com 4 locuções de Adaptação.*

| Classes | Locutor 1 | Locutor 2 | Locutor 3 | Locutor 4 |
|---|---|---|---|---|
| | SI – 63,30% | SI – 90,10% | SI – 73,30% | SI – 65,20% |
| | SD – 87,50% | SD – 95,10% | SD – 94,80% | SD – 92,90% |
| Div 1 | 64,20% | 88,50% | 77,00% | 69,10% |
| Div 2 | 62,60% | 87,10% | 77,80% | 65,30% |
| Dist E | 62,90% | 85,20% | 78,20% | 62,30% |
| Dist B | 57,00% | 87,10% | 78,50% | 65,60% |

Resultados ao se utilizar 5 frases de adaptação:

**Tabela 16** - *Comparação entre o SI e o SA alterando-se a Divisão das Médias das Componentes Gaussianas nas Oito Classes de Regressão com 5 locuções de Adaptação.*

| Classes | Locutor 1 | Locutor 2 | Locutor 3 | Locutor 4 |
|---|---|---|---|---|
| | SI – 63,30% | SI – 90,10% | SI – 73,30% | SI – 65,20% |
| | SD – 87,50% | SD – 95,10% | SD – 94,80% | SD – 92,90% |
| Div 1 | 64,20% | 88,60% | 78,00% | 69,60% |
| Div 2 | 61,90% | 88,00% | 75,60% | 61,50% |
| Dist E | 65,10% | 87,80% | 79,20% | 69,60% |
| Dist B | 62,40% | 86,80% | 78,50% | 66,10% |

Resultados ao se utilizar 6 frases de adaptação:

**Tabela 17** - *Comparação entre o SI e o SA alterando-se a Divisão das Médias das Componentes Gaussianas nas Oito Classes de Regressão com 6 locuções de Adaptação.*

| Classes | Locutor 1 | Locutor 2 | Locutor 3 | Locutor 4 |
|---|---|---|---|---|

| | SI – 63,30% | SI – 90,10% | SI – 73,30% | SI – 65,20% |
|---|---|---|---|---|
| | SD – 87,50% | SD – 95,10% | SD – 94,80% | SD – 92,90% |
| Div 1 | 65,80% | 89,70% | 78,90% | 70,80% |
| Div 2 | 64,40% | 86,80% | 78,40% | 66,80% |
| Dist E | 65,30% | 88,00% | 78,50% | 71,60% |
| Dist B | 64,90% | 88,00% | 80,10% | 69,80% |

Resultados ao se utilizar 10 frases de adaptação:

**Tabela 18** - *Comparação entre o SI e o SA alterando-se a Divisão das Médias das Componentes Gaussianas nas Oito Classes de Regressão com 10 locuções de Adaptação.*

| Classes | Locutor 1 | Locutor 2 | Locutor 3 | Locutor 4 |
|---|---|---|---|---|
| | SI – 63,30% | SI – 90,10% | SI – 73,30% | SI – 65,20% |
| | SD – 87,50% | SD – 95,10% | SD – 94,80% | SD – 92,90% |
| Div 1 | 65,80% | 90,40% | 81,00% | 70,80% |
| Div 2 | 66,20% | 90,30% | 79,10% | 71,00% |
| Dist E | 66,40% | 90,70% | 78,50% | 71,80% |
| Dist B | 66,20% | 90,50% | 81,80% | 71,40% |

Comparando-se as duas divisões feitas utilizando a classificação fonética, percebe-se que na maior parte dos testes, ao se considerar o fone silêncio como uma classe de regressão, conseguiu-se um desempenho maior. É possível perceber também que a medida de distância Euclidiana trouxe na maioria dos casos um desempenho do sistema superior ao ser comparado à distância Bhattacharya, para uma quantidade de material de adaptação maior. Por fim, nota-se que a taxa de desempenho tende a se estabilizar com o aumento do número de locuções de adaptação. O ganho obtido utilizando-se 10 frases não é muito superior quando comparado com 6 locuções de adaptação.

# CAPÍTULO 6

# CONCLUSÃO

Dos testes realizados usando-se a adaptação MLLR foi possível notar que a quantidade de material de adaptação e o número de classes de regressão estão diretamente relacionados. Para se conseguir um SA com desempenho superior ao SI deve existir um compromisso entre esses dois parâmetros.

Neste trabalho procurou-se utilizar diversas formas de distribuição das médias das componentes gaussianas dos HMMs nas classes de regressão. Para isso foram utilizadas as classificações fonéticas e duas medidas de distância: Euclidiana e Bhattacharya. Entretanto, os testes realizados com essas medidas de distância, foram feitos utilizando-se somente oito classes de regressão. Seria interessante realizarmos testes com outra quantidade de classes de regressão.

Outra conclusão importante é que determinados fones influenciam mais no processo de adaptação que outros. Por exemplo, utilizando a classificação fonética, as subunidades *a* e *an*, formam uma classe de regressão na *divisão 2*, enquanto que na *divisão 1*, elas se encontram associadas a outros fones (verificar as tabelas 6 e 7). De acordo com os resultados, na maioria dos testes comparando as duas divisões, obteve-se um melhor desempenho ao se utilizar a *divisão 1*.

A adaptação de um SI utilizando-se o mesmo número de classes de regressão, apresentou resultados diferentes. Ao se utilizar oito classes de regressão por exemplo,

conforme exposto nas tabelas 15, 16, 17 e 18, com quatro locuções de adaptação o melhor resultado foi obtido utilizando a *divisão 1*. Já com dez locuções de adaptação, a distribuição utilizando-se a distância euclidiana, trouxe um melhor resultado.

De todos os testes realizados foi possível verificar que o melhor desempenho obtido foi com três classes de regressão, utilizando-se pouco material de adaptação, por exemplo, de três a quatro locuções.

Algumas sugestões para trabalhos futuros são:

- Comparação das medidas de distância utilizadas considerando outras quantidades de classes de regressão;

- Um estudo do material de adaptação na escolha dos fones que têm maior influência no processo de adaptação.

- Realização de uma adaptação separando os parâmetros das médias das componentes gaussianas dos HMMs, ou seja, análise individual dos parâmetros mel, delta-mel e delta-delta-mel.

## ANEXO A : TREINAMENTO DE UM HMM

Na fase de treinamento, partimos de um HMM inicial e de sequências de observação $O=\{O_{1,}O_{2,}\cdots,O_T\}$ do evento que se deseja modelar. Quanto maior a quantidade de sequências de observação de um determinado evento, melhor treinado estará o modelo HMM. O treinamento é feito então, de forma iterativa até que se atinja a convergência desejada. Normalmente é utilizado um modelo inicial como ponto de partida e reestimações são feitas a fim de se conseguir um modelo onde a verossimilhança seja maior ou igual a anterior. A cada iteração o modelo gerado servirá como modelo inicial.

Para uma sequência de observação $O=\{O_{1,}O_{2,}\cdots,O_T\}$ de duração *T*, e um modelo λ de *N* estados, devemos calcular a probabilidade de ocorrer a sequência *O* dado o modelo λ, isto é: $P(O|\lambda)$ . Para isso é usado o seguinte procedimento, conhecido como algoritmo "Forward" [4]:

1º Passo – Inicialização:

$$\alpha_1(i)=\pi_i b_i(O_1), \qquad 1\leqslant i\leqslant N \tag{A.1}$$

2º Passo – Indução:

$$\alpha_{t+1}(j)=[\sum_{i=1}^{N}\alpha_t(i)a_{ij}]b_j(O_{t+1}) \qquad \begin{matrix}1\leqslant t\leqslant T-1\\ 1\leqslant j\leqslant N\end{matrix} \tag{A.2}$$

3ºPasso – Término:

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i) \tag{A.3}$$

A variável $\alpha_t(i)$ é chamada variável "forward" e é definida por:

$$\alpha_t(i) = P[O_1 O_2 \cdots O_t, q_t = S_i | \lambda] \qquad \text{(A.4)}$$

Conhecida a sequência de observação determina-se a sequência de estados ótima usando-se o algoritmo de Viterbi [4] descrito abaixo:

1º Passo – Inicialização:

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \leqslant i \leqslant N \qquad \text{(A.5)}$$

$$\psi_1(i) = 0 \qquad \text{(A.6)}$$

2º Passo – Recursão:

$$\delta_t(j) = \max_{1 \leqslant i \leqslant N} [\delta_{t-1}(i) a_{ij}] b_j(O_t) \qquad \begin{matrix} 2 \leqslant t \leqslant T \\ 1 \leqslant j \leqslant N \end{matrix} \qquad \text{(A.7)}$$

$$\psi_t(j) = \underset{1 \leqslant i \leqslant N}{argmax} [\delta_{t-1}(i) a_{ij}] \qquad \begin{matrix} 2 \leqslant t \leqslant T \\ 1 \leqslant j \leqslant N \end{matrix} \qquad \text{(A.8)}$$

3ºPasso – Término:

$$P' = \max_{1 \leqslant i \leqslant N} \delta_T(i) \qquad \text{(A.9)}$$

$$q_t' = arg \max_{1 \leqslant i \leqslant N} \delta_T(i) \qquad \text{(A.10)}$$

4ºPasso – Determinação da Sequência de estados ótima:

$$q_t' = \psi_{t+1}(q'_{t+1}) \qquad t = T-1, T-2, \cdots, 1 \qquad \text{(A.11)}$$

Por fim, para estimarmos os parâmetros do modelo λ que gerou a sequência de observação *O*, usamos o Algoritmo "Forward-Backward", também chamado Algoritmo Baum-Welch [4]. A variável "backward"é definida por:

$$\beta_t(i) = P[O_{t+1} O_{t+2} \cdots O_T | q_t = S_i, \lambda] \qquad \text{(A.12)}$$

E pode ser determinada através do seguinte algoritmo:

1º Passo – Inicialização:

$$\beta_T(i)=1 \qquad 1\leqslant i\leqslant N \tag{A.13}$$

2º Passo – Indução:

$$\beta_t(i)=[\sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\beta_{t+1}(j)] \qquad \begin{matrix} t=T-1,T-2,\cdots,1 \\ 1\leqslant i\leqslant N \end{matrix} \tag{A.14}$$

O Algoritmo Baum-Welch segue os seguintes passos:

1º Passo – Forneça um conjunto inicial de parâmetros para o HMM, {A, B, π}

2º Passo – Calcule $\bar{A}, \bar{B}$ de acordo com as fórmulas de reestimação abaixo:

$$\bar{a}_{ij}=\frac{\sum_{t=1}^{T-1}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{T-1}\alpha_t(i)\beta_t(i)} \tag{A.15}$$

$$\bar{c}_{jm}=\frac{\sum_{t=1}^{T}\gamma_t(j,m)}{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t(j,m)} \tag{A.16}$$

$$\bar{\mu}_{jm}=\frac{\sum_{t=1}^{T}\gamma_t(j,m)O_t}{\sum_{t=1}^{T}\gamma_t(j,m)} \tag{A.17}$$

$$\bar{U}_{jm}=\frac{\sum_{t=1}^{T}\gamma_t(j,m)(O_t-\mu_{jm})(O_t-\mu_{jm})'}{\sum_{t=1}^{T}\gamma_t(j,m)} \tag{A.18}$$

$$\gamma_t(j,m)=\left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{k=1}^{N}\alpha_t(k)\beta_t(k)}\right]\left[\frac{c_{jm}G(O_t,\mu_{jm},U_{jm})}{b_j(O_t)}\right] \tag{A.19}$$

3º Passo – Faça A igual a $\bar{A}$ e B igual a $\bar{B}$ .

4º Passo – Se não ocorrer a convergência desejada, retorne ao passo 2.

As expressões acima são utilizadas para gerar parâmetros quando somente uma sequência de observação é usada. Entretanto, para se ter um SRF robusto, a quantidade de amostras de um fone, uma palavra ou uma frase, o que se queira modelar/parametrizar, deve ser grande. Isso significa que teremos uma sequência de treinamento dos modelos HMMs com múltiplas observações. Quando um HMM é treinado usando sequências com múltiplas observações, os parâmetros do modelo são reestimados após a apresentação de todas as observações. Sendo assim podemos reescrever as expressões anteriores considerando o emprego de sequências com *D* observações, da seguinte maneira:

$$\bar{a}_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d-1} \alpha_t^d(i)\, a_{ij}\, b_j(O_{t+1}^d)\, \beta_{t+1}^d(j)}{\sum_{d=1}^{D}\sum_{t=1}^{T_d-1} \alpha_t^d(i)\, \beta_t^d(i)} \qquad \text{(A.20)}$$

$$\bar{c}_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \alpha_t^d(j)\, \beta_t^d(j)\, N_t^d(j,m)}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \alpha_t^d(j)\, \beta_t^d(j)} \qquad \text{(A.21)}$$

$$\bar{\mu}_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \alpha_t^d(j)\, \beta_t^d(j)\, N_t^d(j,m)\, O_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \alpha_t^d(j)\, \beta_t^d(j)\, N_t^d(j,m)} \qquad \text{(A.22)}$$

$$\bar{U}_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \alpha_t^d(j)\, \beta_t^d(j)\, N_t^d(j,m)\, (O_t^d - \mu_{jm})(O_t^d - \mu_{jm})'}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \alpha_t^d(j)\, \beta_t^d(j)\, N_t^d(j,m)} \qquad \text{(A.23)}$$

onde *D* é o número de observações na sequência de treinamento e a variável $N_t(j,m)$ é dada por:

$$N_t(j,m)=\left[\frac{c_{jm}G(O_t,\mu_{jm},U_{jm})}{\sum_{k=1}^{M}c_{jk}G(O_t,\mu_{jk},U_{jk})}\right] \qquad \text{(A.24)}$$

Existe ainda a necessidade de normalizar os coeficientes $a_{ij}$ e $b_i(k)$ já que estes são menores que 1 e à medida que o instante de tempo *t*, torna-se grande, a variável $\alpha_t(i)$ se aproximará de zero [4]. Para maiores detalhes consultar [5].

## REFERÊNCIAS BIBLIOGRÁFICAS


[1] YARED, G.F,G., *Método para determinação do número de gaussianas em modelos ocultos de Markov para sistemas de fala contínua*. Tese de Doutorado, Campinas, 2006.

[2] Folha.com. Reconhecimento de fala tem inúmeras aplicações. Rafael Capanema. Disponível em: http://www1.folha.uol.com.br/tec/811967-reconhecimento-de-fala-tem-inumeras-aplicacoes.shtml. 10/10/2010. Acesso em 20 de janeiro.

[3] FURUI, S., *Digital Speech Processing, Synthesis and Recognition*. Marcel Dekker, Inc., 1989.

[4] RABINER, L. R. and JUANG, B. H., *Fundamentals of Speech Recognition*, Prentice Hall Press, 1993.

[5] MARTINS, J. A., *Avaliação de Diferentes Técnicas para Reconhecimento de Fala*. PhD thesis, Universidade Estadual de Campinas, Dezembro 1997.

[6] PICONE, J. W., *Signal Modelling Techniques in Speech Recognition.* Proceedings of the IEEE, vol 81, no 9. September 1993.

[7] SELTZER, M., *SPHINX III Signal Processing Front End Specification*. CMU Speech Group, August 1999.

[8] ALENCAR, V. L. S., *Atributos e Domínios de Interpolação Eficientes em Reconhecimento de Voz Distribuído*. Dissertação de Mestrado, Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Março 2005.

[9] DELLER Jr., J. R., PROAKIS, J. G., HANSEN, J.H.L., *Discrete Time Processing of Speech Signals*. MacMillan Publishing Company. New York, 1993.

[10] RABINER, L. R., *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE, 77(2), pages 257-286. February, 1989.

[11] DIAS, R. S. F., *Normalização de Locutor em Sistema de Reconhecimento de Fala.* Dissertação de Mestrado, Universidade Estadual de Campinas, Novembro 2000.

[12] ANDREOU, A., KAMM, T. and COHEN, J., *Experiments in Vocal Tract Normalization*. Proceedings CAIP Workshop: Frontiers in Speech Recognition II, 1994.

[13] LEE, L. and ROSE, R., *A Frequency Warping Approach to Speaker Normalization*. IEEE Transactions on Communications, COM-28(1), pages 49-60, January 1980.

[14] WOODLAND, P., *Speaker Adaptation: Techniques and Challenges.* Proceedings IEEE Automatic Speech Recognition ans Understanding Workshop, pages 85-90, Colorado 2000.

[15] CHRISTENSEN, H., *Speaker Adaptation of hiden Markov Models Using Maximum Likelihood Linear Regression*. Thesis, Aalborg University, Denmark, UK, 1994.

[16] ZAVALIAGKOS, G., SCHWARTZ, R. and MAKHOUL, J., *Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition.* Proceedings ICASSP-95, pages 676-679, Michigan 1995.

[17] GAUVIAN, J. L. and LEE, C.H., *Bayesian Learning for HMM With Gaussian Misture State Observation Densities*. Speech Comm, volume 11, pages 205-213, 1992.

[18] GAUVIAN, J. L. and LEE, C.H., *Speaker Adaptation Based on Map Estimation of HMM Parameters.* IEEE ICASSAP-93, pages 558-561, 1993.

[19] GAUVIAN, J. L. and LEE, C.H., *Maximum a Posteriori Estimation for Multivariate Gassian Mixture Observations of Markov Chains*. IEEE Transactions on Speech and Audio Processing, volume 02, pages 291-298, April 1994.

[20] ZAVALIAGKOS, G., SCHWARTZ, R. and McDONOUGH, J., *Maximum A Posteriori adaptation for large Scale HMM Recognizers.* Int. Conf. Acoustics, Speech, Signal Processing'96, pages 725-728, Atlanta, GA, 1996.

[21] LEGGETTER, C.J. and WOODLAND, P.C., *Speaker Adaptation of Continuous Density HMM's Using Linear Regression.* Int. Conf. Speech Language Processing '94, volume 02, pages 451-454,

Yokohama, Japão 1994.

[22] LEGGETTER, C.J. and WOODLAND, P.C., *Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Markov Models*. Computer Speech and Language, volume 09 (no 02): pages 171-185, April 1995.

[23] TURK, M.; PENTLAND, A., *Eigenfaces for Recognition.* Journal of Cognitive Neiroscience, volume 03 (no 01), pages 71-86, 1991.

[24] KUHN, H, et.al., *Eigenvoices for Speaker Adaptation.* Proc. of ICSLP-98, pages 1771-1774, Sydney, Australia, 1998.

[25] SOUSA, L. C., *Adaptação de Locutor Em Sistemas de Reconhecimento de Fala Contínua Empregando "Eigenvoices"*. Dissertação de Mestrado, Universidade Estadual de Campinas, Setembro 2004.

[26] WESTWOOD, R., *Speaker Adaptation using Eigenvoices.* Master's Thesis. Cambridge University, August, 1999.

[27] RICHARD, J.A.; DEAN, W.W., *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.

[28] NGUYEN, P., *Fast Speaker Adaptation.* Industrial thesis Report, Institut Eurécom, June, 1998.

[29] LEGGETTER, C.J. and WOODLAND, P.C., *Speaker Adaptation of HMM's Using Linear Regression.* Cambridge University, Technical Report, TR.181, June, 1994.

[30] LEGGETTER, C.J., *Improved Acoustic Modelling for HMM's using Linear Transformations*. Ph.D. Thesis, Cambridge University, 1995.

[31] LEGGETTER, C.J. and WOODLAND, P.C., *Flexibe Speaker Adaptation for Large Vocabulary Speech Recognition*. Proceedings EUROSPEECH95, pages1155-1158, 1995.

[32] HUANG, X.D., ARIKI, Y., JACK, M.A., *Hidden Markov Models for Speech Recognition*. Edinburgh University Press, 1990.

[33] YNOGUTI, C. A., *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*. PhD thesis, Universidade Estadual de Campinas, Maio 1999.

[34] ALCAIM, A., SOLEWICZ, J. A., MORAES, J. A., F*reqüência de*

*ocorrência dos fones e lista de frases foneticamente balanceadas no português falado no Rio de Janeiro.* Revista da Sociedade Brasileira de Telecomunicações, 7(1): p 23-41. Dezembro, 1992.

[35] 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 (Fevereiro/2006).

[36] KAILATH, T., *The Divergence and Bhattacharya Distance Measures in Signal Selection,* IEEE Transactions on Communication Technology, 15-1, February 1967, pp. 52-60.