UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE COMUNICAÇÕES

Adaptação de Locutor em Sistema de Reconhecimento de Fala Contínua Empregando "Eigenvoices"

Autor:
Lívio Carvalho Sousa

Orientador:
Prof. Dr. Fábio Violaro

Dissertação submetida à Faculdade de Engenharia Elétrica e de Computação da UNICAMP como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Banca Examinadora:

Prof. Dr. Fábio Violaro (Orientador)	FEEC/UNICAMP
Prof. Dr. Carlos Alberto Ynoguti	Inatel
Prof. Dr. Jaime Portugheis	FEEC/UNICAMP
Prof. Dr. José Antônio Martins	CPqD/Campinas

Campinas, 24 de Setembro de 2004.

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

So85a

Sousa, Lívio Carvalho
Adaptação de locutor em sistema de reconhecimento de fala contínua empregando "Eigenvoices" / Lívio Carvalho Sousa. --Campinas, SP: [s.n.], 2004.

Orientador: Fábio Violaro.
Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Telecomunicações. 2. Processamento de sinais. 3. Sistemas de processamento da fala. 4. Reconhecimento automático da voz. I. Violaro, Fábio. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Resumo

Neste trabalho realizou-se o estudo da técnica via *"eigenvoices"* [13] [16] [17] [18] [30] [31] para adaptação de locutor em um sistema de reconhecimento de fala contínua usando o português do Brasil. Dentre as várias técnicas utilizadas para a adaptação de locutor, incluindo as clássicas *MAP ("Maximum A Posteriori")* e *MLLR ("Maximum Likelihood Linear Regression")*, uma nova técnica, chamada *"eigenvoice technique"*, foi proposta por Kuhn visando tornar mais rápido o processo de adaptação de locutor para aplicação em sistemas operando em tempo real. No início, estudos se concentraram nas aplicações com palavras isoladas, mas várias pesquisas estão sendo realizadas para a análise dessa técnica em fala contínua, como é o caso deste trabalho. A característica principal da técnica de adaptação via "eigenvoices" é a representação do novo locutor como uma combinação linear de parâmetros ("eigenvoices") obtidos a partir de modelos dependente de locutor previamente treinados. Dessa forma, o novo locutor é representado como um ponto dentro do espaço cujos eixos são formados pelos "eigenvoices". O algoritmo de máxima verossimilhança *MLED ("Maximum Likelihood Eigen Decomposition")* foi usado para o cálculo dos coeficientes da combinação linear para a estimação dos parâmetros do novo locutor. Após a realização de testes com número variado de locuções de adaptação e de iterações do algoritmo, foi observado que: para um bom desempenho dos modelos adaptados, 3 a 5 iterações do algoritmo são necessárias; o mais importante não é o número de locuções de adaptação mas sim o seu conteúdo fonético. Em suma, o estudo revelou que a técnica se mostrou eficiente para a aplicação, porém mais pesquisas são necessárias na área.

Abstract

In this work a research was made in order to evaluate the use of the *eigenvoice technique* [13] [16] [17] [18] [30] [31] to speaker adaptation on a continuous speech recognition system. Amongst the several speaker adaptation techniques, like the classical *MAP* and *MLLR*, a new technique, called eigenvoice technique, was proposed by Kuhn for fast speaker adaptation in real time applications. Firstly, researches were made just on isolated words applications, and nowadays they are focused on continuous speech applications, like this work. The main feature of the eigenvoice technique is the representation of the new speaker by a linear combination of parameters (eigenvoices) extracted from speaker dependent models previously trained. The new speaker is represented by a point in a space whose axis are the eigenvoices. The *Maximum Likelihood Eigen Decomposition (MLED)* algorithm was used to calculate the combination coefficients in order to estimate the parameters of the new speaker. After tests varying the number of adaptation sentences and algorithm iterations, it was verified that: for a good adaptation performance, 3 to 5 algorithm iterations are necessary; the number of adaptation sentences is not very important, the more important is the adaptation sentences phonetic content. In conclusion, the eigenvoice technique showed to be efficient for the application on continuous speech, however more studies must be made in the area.

Agradecimentos

À Deus, mesmo eu não possuindo palavras suficientes para expressar um agradecimento verdadeiro, agradeço a Ele pela Sua Providencia em minha vida, pelos caminhos guiados, pelas conquistas e derrotas, pelas pessoas que pôs em minha vida, pelo que fez e faz em mim e pelo que sou hoje.

Aos meus pais, João e Margarida, e à minha irmã Lígia, aqui também eu não podendo expressar o fiel reconhecimento, agradeço pela criação, dedicação, carinho, diálogo, amor, respeito, correções, conselhos, força, incentivo e muitos outros fatores que serviram para mostrar que, sem a companhia deles, eu não teria a capacidade de vencer e conseguir, entre outras coisas, este título de mestre.

Ao prof. Dr. Fábio Violaro que, mais que um orientador, foi um grande amigo e verdadeiro mestre: orientando, esclarecendo, sempre apoiando, sendo paciente, trabalhando, pesquisando em comum união, incentivando nos momentos mais difíceis deste trabalho e fazendo-me aprender um pouco mais sobre a vida além de engenharia.

Ao prof. Dr. Carlos Alberto Ynoguti, agradeço pela sua solicitude, colaboração, paciência, discussões e esclarecimentos fundamentais, auxiliando no manuseio da ferramenta de trabalho utilizada, propondo novas idéias de procedimentos experimentais bem como outras idéias e sugestões.

À todos os meus familiares que apostaram em mim e me deram força e apoio para eu completar mais esta fase em meus estudos.

À minha namorada Cristiane pela companheira, amiga, confidente e conselheira que foi durante todo o período do Mestrado, que, mesmo de longe, sempre me apoiou e incentivou nos diversos momentos de realização deste trabalho, principalmente nos mais difíceis.

Ao meu amigo Watson que, pela sua persistência inspirada e providencial, contribuiu para o meu ingresso

na faculdade e consequentemente é um dos responsáveis por este título de mestre.

Aos meus amigos Helder, Rodrigo e Sérgio pelo companheirismo, amizade, apoio, sugestões, esclarecimentos e convivência agradável durante o período em que compartilhamos a mesma moradia e momentos de vida.

Aos amigos companheiros de laboratório Aline, Ceron, Edmilson, Glauco e Marcos pela saudável convivência, apoio, incentivo, esclarecimentos e discussões, auxiliando na resolução de diversos problemas, propondo variantes de procedimentos e contribuindo para uma maior compreensão do assunto tema de tese.

Aos amigos de departamento Carlos, Flávio, Gabriel, Gilmar, Jaqueline, Leo, Lucas, Luiz Jr., Paulo, Pêpe, Ponchet e Márzio pela amizade, convívio descontraído, solidariedade, dicas e incentivo que possibilitaram um bom relacionamento fraterno e um dia-a-dia saudável durante o período do Mestrado.

Aos amigos prof. Dr. Carlos, Ceron, Ekler, Emerson, prof. Dr. Fábio, Flávio, Gabriel, Gilmar, Glauco, Gleidson, Luiz, Luiz Jr., Marcos, Marzio, Paulo, Rogério e Vicente pela solicitude, consideração, amizade, paciência e disposição nas seções de gravação de voz que foram de fundamental importância neste trabalho.

Aos amigos conterrâneos Cristiano, Gleidson e Paulo Roberto pela amizade, força, consideração e apoio, e que, mesmo distantes, esforçaram-se na configuração e gravação de algumas vozes, auxiliando na elaboração deste trabalho.

Aos amigos de pastoral universitária André, Cristiano, David, Flávia, Germano, Guilherme, Lívia, Lucas, Marnem, Pedro, Rogério, Thaís e Pe. Toninho pela amizade, incentivo, força, partilha, convívio agradável e abertura de conhecimento e fé.

À D. Vera, e à Márcia, Míriam, D. Maria e toda a sua família pela amizade, atenção, apoio, ajuda e principalmente pelo acolhimento, contribuindo para um bom relacionamento de vizinhança e uma convivência harmoniosa.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento deste trabalho através de bolsa de Mestrado (processo 02/05206-1).

À todos aqueles que contribuíram de alguma forma para a conclusão deste trabalho e a aquisição deste título, mesmo eu esquecendo de citar os seus nomes, venho aqui expressar o meu muito obrigado.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos e Acrônimos

Variáveis

T - Número de observações de um evento aleatório, ou número de quadros de uma locução

U - Número de locuções utilizadas no treinamento de HMM's

T_u - Número de quadros da u-ésima locução de treinamento

S - Número de estados de um HMM

M - Número de gaussianas em um mistura

D - Dimensão dos vetores de parâmetros acústicos

L - Número de locutores base que representam o espaço de locutores (número de supervetores)

N - Dimensão dos supervetores

K - Número de "eigenvoices" considerados na representação do espaço de locutores

O - Seqüência de observações de um evento aleatório ou de vetores de parâmetros acústicos

o_t - Observação genérica para um evento aleatório no instante t

$\vec{o}_t$ - Vetor de parâmetros acústicos referente ao quadro t (multidimensional)

q - Seqüência de estados observados de um HMM

q_t - Estado de um HMM observado no instante t

A - Matriz probabilidade de transição de um HMM

a_{ij} - Probabilidade de transição do estado i para o estado j dentro de um HMM

a_{ii} - Probabilidade de transição do estado i para ele mesmo dentro de um HMM

B - Matriz de funções de emissão de estados de um HMM

$b_j(\vec{o}_t)$ - Função de emissão do estado j, para um vetor de parâmetros $\vec{o}_t$, dentro de um HMM

Π - Matriz probabilidade inicial dos estados de um HMM

π_i - Probabilidade inicial do estado i

$c_m^{(s)}$ - Peso da m-ésima gaussiana do s-ésimo estado de um HMM

$\vec{\mu}_m^{(s)}$ - Média gaussiana (multidimensional) da m-ésima gaussiana do s-ésimo estado de um HMM

$\vec{\sigma}_m^{2^{(s)}}$ - Variância gaussiana (multidimensional) da m-ésima gaussiana do s-ésimo estado de um HMM

$\overline{a}_{ij}$ - Probabilidade de transição reestimada do estado i para o estado j de um HMM

$\overline{c}_m^{(s)}$ - Peso reestimado da m-ésima gaussiana do s-ésimo estado de um HMM

$\overline{\vec{\mu}}_m^{(s)}$ - Média gaussiana reestimada da m-ésima gaussiana do s-ésimo estado de um HMM

$\overline{\vec{\sigma}}_m^{2^{(s)}}$ - Variância gaussiana reestimada da m-ésima gaussiana do s-ésimo estado de um HMM

$\hat{\alpha}_t^{(u)}(j)$ - Variável "forward" escalonada, referente ao t-ésimo quadro da u-ésima locução, correspondente ao j-ésimo estado de um HMM

$\hat{\beta}_t^{(u)}(j)$ - Variável "backward" escalonada, referente ao t-ésimo quadro da u-ésima locução, correspondente ao j-ésimo estado de um HMM

$\hat{c}_t^{(u)}$ - Coeficiente de escalonamento das variáveis "forward" e "backward" referente ao t-ésimo quadro da u-ésima locução

$\tilde{\alpha}_t^{(u)}(j)$ - Variável "forward" pré-escalonada, referente ao t-ésimo quadro da u-ésima locução, correspondente ao j-ésimo estado de um HMM

$\tilde{\beta}_t^{(u)}(j)$ - Variável "backward" pré-escalonada, referente ao t-ésimo quadro da u-ésima locução, correspondente ao j-ésimo estado de um HMM

$\alpha_t^{(u)}(j)$ - Variável "forward" original, referente ao t-ésimo quadro da u-ésima locução, correspondente ao j-ésimo estado de um HMM

$\beta_t^{(u)}(j)$ - Variável "backward" original, referente ao t-ésimo quadro da u-ésima locução, correspondente ao j-ésimo estado de um HMM

$H_t^{(u)}(j,m)$ - Probabilidade de ocupação da m-ésima gaussiana do j-ésimo estado de um HMM com respeito ao t-ésimo quadro da u-ésima locução

$G(\vec{o}_t, \vec{\mu}_m^{(j)}, \vec{\sigma}_m^{2(j)})$ - Função densidade de probabilidade da m-ésima gaussiana do j-ésimo estado de um HMM com respeito ao vetor de parâmetros $\vec{o}_t$

Mc - Número de coeficientes Melcepstrais extraídos por quadro da locução

MEL_i - Coeficiente Melcepstral de i-ésima ordem

E_t - Parâmetro Log-energia do t-ésimo quadro da locução

En_t - Parâmetro Log-energia normalizado do t-ésimo quadro da locução

Υ_t - Número de amostras do t-ésimo quadro de uma locução

$\Delta_i(n)$ - n-ésimo elemento do vetor delta de um vetor de parâmetros acústicos do i-ésimo quadro

J - Número de quadros adjacentes a serem considerados no cálculo dos parâmetros delta

τ - Meta-parâmetro (relativo à técnica MAP)

$\mathbf{W}$ - Matriz transformação (relativo à técnica MLLR)

$\vec{\xi}$ - Vetor composto de médias gaussianas (relativo à técnica MLLR)

$\mathbf{V}$ - Matriz de supervetores

$\mathbf{E}$ - Matriz de "eigenvoices"

$\vec{e}(0)$ - "Eigenvoice 0"

$\vec{e}(j)$ - "Eigenvoice" de ordem j

$\vec{e}_m^{(s)}(0)$ - "Sub-eigenvoice 0" referente à m-ésima gaussiana do s-ésimo estado de um HMM

$\vec{e}_m^{(s)}(j)$ - "Sub-eigenvoice" de ordem j referente à m-ésima gaussiana do s-ésimo estado de um HMM

$\mathbf{C}_m^{(s)^{-1}}$ - Matriz covariância inversa da m-ésima gaussiana do s-ésimo estado de um HMM

$\gamma_m^{(s)}(t)$ - Probabilidade da m-ésima gaussiana do s-ésimo estado de um HMM gerar o vetor de parâmetros $\vec{o}_t$

$\vec{\tilde{\mu}}_m^{(s)}$ - Média gaussiana adaptada da m-ésima gaussiana do s-ésimo estado de um HMM

w_i - Coeficiente do i-ésimo "eigenvoice" $e(i)$

$e_{mn}^{(s)}(j)$ - n-ésimo elemento do "eigenvoice" $\vec{e}_m^{(s)}(j)$

o_{tn} - n-ésimo elemento do vetor de parâmetros acústicos $\vec{o}_t$

$\sigma_{mn}^{2^{(s)}}$ - n-ésimo elemento do vetor variância da m-ésima gaussiana do s-ésimo estado de um HMM

$\delta(k, j)$ - Coeficiente do 1º membro do sistema de equações para estimação dos coeficientes w_i's

$\Lambda(j)$ - 2º membro do sistema de equações para estimação dos coeficientes w_i's

Funções

$Q(\lambda_{inicial}, \lambda)$ - Função auxiliar de Baum

Acrônimos

SRF - *Sistema de Reconhecimento de Fala*

HMM - *Modelo Oculto de Markov (do inglês, "Hidden Markov Model")*

PCA - *Análise de Componente Principal (do inglês, "Principal Component Analysis")*

DFT - *Transformada Discreta de Fourier (do inglês, "Discrete Fourier Transform")*

FFT - *Transformada Rápida de Fourier (do inglês, "Fast Fourier Transform")*

DCT - *Transformada Discreta do Cosseno (do inglês, "Discrete Cosine Transform")*

CP - *Critério de Parada*

TEP - *Taxa de Erro de Palavras*

TAP - *Taxa de Acerto de Palavras*

MAP - *"Maximum A Posteriori"*

RMP - *"Regression based Model Prediction"*

SMAP - *"Structural Maximum A Posteriori"*

MLLR - *"Maximum Likelihood Linear Regression"*

CAT - *"Cluster Adaptive Training"*

MLED - *"Maximum Likelihood Eigen Decomposition"*

Capítulo 1

Introdução

O avanço da tecnologia tem possibilitado o desenvolvimento de melhores interfaces entre homem e máquina em diferentes aplicações. As formas de comunicação contidas nas interfaces podem ser das mais diversas, porém as mais observadas são as formas visuais e sonoras. Dentre as formas sonoras, aplicações são desenvolvidas a fim de tentar interagir o sistema com o usuário através de sinais de fala, como por exemplo o reconhecimento de comandos de fala para a realização de determinada tarefa ou a conversão fala-texto para fins de redação. Nesse caso, deve-se fazer uso de sistemas de reconhecimento de fala para compor a camada de interface dessas aplicações.

Sistemas de reconhecimento de fala (SRF´s) vêm sendo utilizados de muitas formas por aplicações destinadas ao uso de muitos usuários. Antes de serem utilizados na decodificação dos sinais de fala, os SRF's precisam ser treinados. Um fato é que um SRF nem sempre está suficientemente treinado para o reconhecimento de fala de qualquer usuário, o que implica em erros de reconhecimento de palavras. Devido a isso, técnicas vêm sendo desenvolvidas com a finalidade de adaptar sistemas de reconhecimento de fala para que possam ser aplicados a usuários que antes não eram bem reconhecidos. Esse processo é chamado de *adaptação de locutor*.

O objetivo deste trabalho de Mestrado é estudar a adaptação de locutor sob uma nova técnica de adaptação chamada *"eigenvoices"*, apresentando alguns resultados de como o desempenho de um SRF se comporta com a aplicação da técnica sobre diversos aspectos.

O conteúdo do trabalho se divide da seguinte forma: No capítulo 2 são abordados os fundamentos dos sistemas de reconhecimento de fala e outros fundamentos teóricos importantes para a compreensão da técnica. No capítulo 3 são comentados alguns experimentos preliminares que tinham o objetivo de definir atributos para a composição da pesquisa. No capítulo 4 são mostradas uma visão geral sobre a adaptação de locutor e algumas técnicas clássicas. No capítulo 5 é explicado a teoria sobre a técnica "eigenvoice", foco deste trabalho de Mestrado. No capítulo 6 é abordado a estrutura do processo de adaptação via "eigenvoices". No capítulo 7 são apresentados os resultados das adaptações realizadas sobre diversas formas. Finalmente no capítulo 8 são apresentados uma conclusão a partir dos resultados e alguns comentários e sugestões para trabalhos futuros.

Capítulo 2

Fundamentos Teóricos

2.1 Modelos Ocultos de Markov

Modelos ocultos de Markov (HMM's, do inglês "Hidden Markov Models") são modelos estatísticos treinados e usados para representar processos aleatórios [3]. Um sinal de fala apresenta variabilidade entre suas repetições, isto é, uma frase, palavra ou até mesmo um fone quando pronunciados duas ou mais vezes apresentam diferenças entre suas pronuncias. Isso leva a considerar uma palavra ou um fone como um evento aleatório por não ser caracterizado deterministicamente sempre que é observado, podendo assim ser modelado por HMM's.

Ao modelar sinais de fala com HMM's, tem-se uma representação do comportamento estatístico do sinal de fala. De posse de modelos específicos para sinais de fala diferentes, palavras ou fones por exemplo, pode-se executar vários procedimentos com o intuito de selecionar o modelo, ou seqüência de modelos, que possui a maior probabilidade de gerar uma certa palavra ou frase, isto é, realizar o *reconhecimento de fala.*

A estrutura de um HMM consiste em uma cadeia de estados. Cada estado do HMM possui uma *função de emissão* que gera parâmetros (ditando o comportamento estatístico daquele estado); uma probabilidade de transição para outro estado do modelo ou de permanência nele próprio; e uma probabilidade inicial que indica quão provável é esse estado ser o estado inicial do HMM. Para entender melhor, considere um HMM específico para a representação de um certo evento aleatório. Suponha que se deseja obter uma seqüência de

observações $\mathbf{O} = \{o_1, o_2, o_3, ..., o_T\}$ desse evento. Uma alternativa seria realizar o evento e colher a seqüência de observações ou utilizar o HMM. Como nem sempre há a possibilidade da realização do evento, o HMM é usado para gerar tal seqüência. A princípio, deve-se escolher um estado pelo qual irá começar a ser gerada a seqüência $\mathbf{O}$, o que é feito segundo a probabilidade inicial de cada estado. No estado inicial, a amostra o_1 é gerada segundo a função de emissão do estado inicial e um *passo* é realizado (transição ou permanência no estado). Se houver uma permanência no estado inicial (segundo a probabilidade de transição desse estado), a segunda amostra o_2 é gerada pelo estado atual e um novo passo é feito. Se houver uma transição, a segunda amostra o_2 é gerada pela função de emissão do novo estado e depois é realizado um novo passo, e assim sucessivamente. À medida que passos são dados, estados do modelos são percorridos gerando uma seqüência de estados visitados $\mathbf{q} = \{q_1, q_2, q_3, ..., q_T\}$, amostras são geradas de acordo com as funções de emissão de cada estado e a seqüência $\mathbf{O}$ é gerada. Dessa forma, um HMM pode "simular" o evento aleatório.

2.1.1 Elementos de um HMM

Um HMM possui alguns elementos que o compõem e que são necessários para representar o seu comportamento estatístico. Antes de abordar os elementos convém falar que os HMM's podem ser *discretos* ou *contínuos*. A diferença básica está na função de emissão dos estados. Enquanto no HMM discreto a função de emissão de um dado estado obedece uma tabela de probabilidade de emissão de símbolos, no HMM contínuo a função de emissão é representada por uma função densidade de probabilidade gaussiana multidimensional ou uma soma de duas ou mais funções densidade de probabilidade gaussianas multidimensionais. Para maior facilidade, o termo função densidade de probabilidade gaussiana será referido como somente *gaussiana* e a soma de duas ou mais gaussianas será referida como *mistura*, no restante desta redação. Cada gaussiana é caracterizada por uma média e uma variância (multidimensionais), e um peso (no caso de misturas). O peso da gaussiana indica o grau de influência daquela gaussiana dentro da mistura, e em uma mistura a soma dos pesos deve ser sempre igual à unidade.

Os elementos de um HMM são:

- Número de estados do modelo - S;

- Matriz probabilidade de transição - $\mathbf{A} = \{a_{ij}\}$, para todo $i \leq S$ e $j \leq S$;
- Função de emissão por estado:
  - $b_j(o_t) = P[o_t|q_t = j]$ para modelos discretos;
  - $b_j(o_t) = \sum_{m=1}^{M} c_m^{(j)} G(o_t, \mu_m^{(j)}, \sigma_{jm}^{2^{(j)}})$ para modelos contínuos.
- Matriz probabilidade inicial dos estados - $\mathbf{\Pi} = \{\pi_i\}$, em que $\pi_i = P[q_1 = i]$.

A matriz probabilidade de transição $\mathbf{A}$ é composta pelos elementos a_{ij}, em que cada elemento é a probabilidade de transição de um estado i para um estado j, ou a probabilidade de permanência nesse estado (a_{ii}). Os termos $c_m^{(j)}$, $\mu_m^{(j)}$ e $\sigma_m^{2^{(j)}}$ são respectivamente o peso, a média e a variância da m-ésima gaussiana do j-ésimo estado do modelo, e M indica o número de gaussianas da mistura. Finalmente a matriz de probabilidade inicial é formada pelos elementos π_i, em que cada elemento é a probabilidade inicial de cada estado do HMM. Como os termos o_t, $\mu_m^{(j)}$ e $\sigma_m^{2^{(j)}}$ podem ser unidimensionais ou multidimensionais, dependendo da aplicação, eles serão representados como vetores no restante desta redação, ou seja, como $\vec{o}_t$, $\vec{\mu}_m^{(j)}$ e $\vec{\sigma}_m^{2^{(j)}}$. No vetor $\vec{\sigma}_m^{2^{(j)}}$ ficam armazenadas somente os elementos da diagonal principal da matriz covariância da m-ésima gaussiana do j-ésimo estado, pois os elementos fora da diagonal principal são considerados nulos, isto é, as componentes das gaussianas multidimensionais são consideradas independentes entre si (ver seção 2.3.1).

Em suma, um HMM, que pode ser simbolicamente representado pela letra λ, é representado por 3 elementos, $\mathbf{\Pi}$, $\mathbf{A}$ e $\mathbf{B}$, no qual este último termo ($\mathbf{B}$) representa o conjunto de todas as funções de emissão de estado do HMM, isto é:

$$\lambda = (\mathbf{\Pi}, \mathbf{A}, \mathbf{B}).$$

Um HMM pode ser implementado sob várias topologias. O sinal de fala, por sua característica dinâmica e progressiva, pode ser modelado por uma topologia cujas transições entre os estados do HMM ocorrem somente em uma sentido, isto é, não há uma caracterização de retrocessos. Esse tipo de topologia é chamado de esquerda-direita (ver Figura 2.1). Nesta topologia, o estado inicial é sempre o estado mais à esquerda do modelo, e cada estado só pode transitar para estados que estão à sua direita ou permanecer nele. Isso caracteriza bem o processo de evolução do sinal de fala.

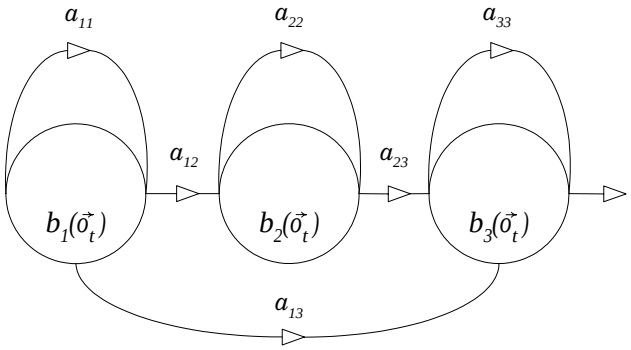

Figura 2.1: HMM esquerda-direita com 3 estados.

2.1.2 Treinamento de HMM's

Para que um HMM possa especificar um evento aleatório, é necessário calcular o trio $(\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$ de tal forma que o modelo possua as características estatísticas do evento. Assim, realiza-se o *treinamento do HMM* a fim de determinar o trio de elementos responsáveis pelo seu comportamento estatístico.

Os métodos usados para o treinamento de HMM's são iterativos e necessitam de um HMM inicial $\lambda_{inicial}$ e de uma ou mais seqüências de observações $\mathbf{O}$ do evento que se quer modelar. Quanto maior o número de seqüências de observações $\mathbf{O}$ usadas para o treinamento, melhor treinado estará o modelo. O HMM inicial é o ponto de partida para que, através do método de treinamento, reestimações sejam feitas a fim de atingir um modelo λ que possua uma verossimilhança tal que: $P[\mathbf{O}|\lambda] \geq P[\mathbf{O}|\lambda_{inicial}]$. A cada iteração, o modelo λ gerado servirá como $\lambda_{inicial}$ para uma nova iteração, e assim sucessivamente até que a convergência seja atingida.

Dois métodos básicos podem ser usados para a geração do $\lambda_{inicial}$:

1. O trio $(\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$ é especificado com valores padrões, por exemplo: variâncias unitárias, pesos igualmente distribuídos entre as misturas, e probabilidades de transição de estados igualmente distribuídas;

2. O trio $(\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$ é calculado através de métodos prévios de estimação.

O primeiro método é usado para fins acadêmicos básicos, isto é, para uma maior compreensão de como se dá o processo de treinamento de HMM's. O segundo método é o padrão utilizado na maioria das aplicações.

Após a estimação do modelo inicial, o método de reestimação propriamente dito é executado a fim de

obter o modelo final λ.

Em reconhecimento de fala, HMM's podem ser treinados para modelar palavras ou fones, dependendo da aplicação. Para aplicação em reconhecimentos de *palavras isoladas*, é treinado um HMM para cada *palavra* do vocabulário do sistema. Para aplicação em *fala contínua*, é treinado um HMM para cada *fone*, seja ele *dependente* ou *independente de contexto*. Neste último caso, os modelos das palavras do vocabulário do sistema são obtidos pela concatenação dos modelos dos fones que a formam. Por exemplo, o HMM da palavra *"bola"* é dado pela concatenação do HMM do *"b"* + HMM do *"o"* + HMM do *"l"* + HMM do *"a"*.

Outra classificação pode ser atribuída aos HMM's quanto aos dados utilizados no treinamento. Se o material de fala é específico de um locutor, o HMM treinado é chamado ser *dependente de locutor*. Se o material de fala oferecido para o treinamento é composto de vários sinais acústicos de locutores diferentes, o HMM treinado é chamado ser *independente de locutor*. As aplicações para HMM's dependentes e independentes de locutor são das mais diversas.

Neste trabalho foram modelados 35 fones mais o "silêncio" (que por facilidade também será referido como um fone). Para cada fone foi treinado um HMM contínuo com 3 estados e topologia esquerda-direita. A função de emissão de cada estado é modelada por uma mistura de 5 gaussianas multidimensionais de dimensão 25 (ver seção 3.1). Todos os fones foram modelados independentemente do contexto. Para a estimação dos HMM's iniciais de cada fone foi usado uma *Segmentação Uniforme* seguida da aplicação do algoritmo *"Segmental K-Means"* [3] e de uma *Segmentação via Viterbi* (as segmentações são abordadas na seção 3.2.1). O método de reestimação usado foi o *Baum-Welch* [3].

Método Baum-Welch

O método proposto por Baum e Welch visa encontrar um HMM com verossimilhança máxima local $P[\mathbf{O}|\lambda]$ para a seqüência $\mathbf{O}$. As equações de reestimação são derivadas da função auxiliar de Baum

$$Q(\lambda_{inicial}, \lambda) = \sum_{\mathbf{q}} P[\mathbf{O}, \mathbf{q}|\lambda_{inicial}] log P[\mathbf{O}, \mathbf{q}|\lambda]. \tag{2.1}$$

Por conveniência, serão apresentados apenas as equações de reestimação para HMM's contínuos com múltiplas seqüências de observação (usadas neste trabalho) encontradas em [29]. São elas:

$$\overline{a}_{ij} = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T_u-1}\hat{\alpha}_t^{(u)}(i)a_{ij}b_j(\vec{o}_{t+1}^{(u)})\hat{\beta}_{t+1}^{(u)}(j)}{\sum_{u=1}^{U}\sum_{t=1}^{T_u-1}\hat{\alpha}_t^{(u)}(i)\hat{\beta}_t^{(u)}(i)/\hat{c}_t^{(u)}}, \tag{2.2}$$

$$\overline{c}_m^{(j)} = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T_u}\hat{\alpha}_t^{(u)}(j)\hat{\beta}_t^{(u)}(j)H_t^{(u)}(j,m)/\hat{c}_t^{(u)}}{\sum_{u=1}^{U}\sum_{t=1}^{T_u}\hat{\alpha}_t^{(u)}(j)\hat{\beta}_t^{(u)}(j)/\hat{c}_t^{(u)}}, \tag{2.3}$$

$$\vec{\overline{\mu}}_m^{(j)} = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T_u}\hat{\alpha}_t^{(u)}(j)\hat{\beta}_t^{(u)}(j)H_t^{(u)}(j,m)\vec{o}_t^{(u)}/\hat{c}_t^{(u)}}{\sum_{u=1}^{U}\sum_{t=1}^{T_u}\hat{\alpha}_t^{(u)}(j)\hat{\beta}_t^{(u)}(j)H_t^{(u)}(j,m)/\hat{c}_t^{(u)}}, \tag{2.4}$$

$$\vec{\overline{\sigma}}_m^{2^{(j)}} = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T_u}\hat{\alpha}_t^{(u)}(j)\hat{\beta}_t^{(u)}(j)H_t^{(u)}(j,m)(\vec{o}_t^{(u)}-\vec{\mu}_m^{(j)})(\vec{o}_t^{(u)}-\vec{\mu}_m^{(j)})^T/\hat{c}_t^{(u)}}{\sum_{u=1}^{U}\sum_{t=1}^{T_u}\hat{\alpha}_t^{(u)}(j)\hat{\beta}_t^{(u)}(j)H_t^{(u)}(j,m)/\hat{c}_t^{(u)}}. \tag{2.5}$$

Nas equações (2.2), (2.3), (2.4) e (2.5), para o treinamento para palavras isoladas, o termo U é o número de pronúncias da mesma palavra que irá ser treinada; para o treinamento para fala contínua, abordado neste trabalho, U é o número de sentenças que contém os fones a serem treinados. O termo T_u é o número de *quadros* da u-ésima palavra ou sentença. Dos quadros são extraídos parâmetros acústicos que constituem as observações $\vec{o}_t$ (ver seção 2.3). Os termos $\hat{\alpha}_t(j)$ e $\hat{\beta}_t(j)$ são chamados de variáveis *"forward"* e *"backward"* escalonadas respectivamente [3]. A variável "forward" original é dada por

$$\alpha_1(j) = \pi_j b_j(\vec{o}_1), \quad 1 \le j \le S, \tag{2.6}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{S}\alpha_t(i)a_{ij}\right]b_j(\vec{o}_{t+1}), \quad \begin{matrix} 1 \le t \le T-1 \\ 1 \le j \le S \end{matrix}, \tag{2.7}$$

e a variável "backward" original é dada por

$$\beta_T(j) = 1, \quad 1 \leq j \leq S, \tag{2.8}$$

$$\beta_t(j) = \sum_{i=1}^{S} a_{ji} b_i(\vec{o}_{t+1}) \beta_{t+1}(i) \qquad \begin{matrix} t = T-1, T-2, \ldots, 1 \\ 1 \leq j \leq S \end{matrix} . \tag{2.9}$$

O escalonamento é feito a fim de reduzir a propagação de erros de arredondamento e evitar a convergência dos resultados das operações para um número pequeno o suficiente que não possa ser representado como um número em ponto flutuante. Esse problema ocorre devido à sucessão de operações aritméticas multiplicativas com termos compreendidos entre 0 e 1 executadas à medida que as iterações em t são incrementadas. Um coeficiente de escalonamento $\hat{c}_t$ é calculado e depois multiplicado às variáveis originais. Da equação (2.6), calcula-se $\hat{c}_1 = 1/\sum_{j=1}^{S} \alpha_1(j)$ e depois calcula-se a variável "forward" escalonada $\hat{\alpha}_1(j) = \hat{c}_1 \alpha_1(j)$. A partir daí, as variáveis escalonadas seguintes são calculadas com base na variável escalonada anterior multiplicada pelos respectivos coeficientes de escalonamento. Assim, a equação (2.7) é reescrita como:

$$\tilde{\alpha}_{t+1}(j) = \left[\sum_{i=1}^{S} \hat{\alpha}_t(i) a_{ij} \right] b_j(\vec{o}_{t+1}), \qquad \begin{matrix} 1 \leq t \leq T-1 \\ 1 \leq j \leq S \end{matrix} , \tag{2.10}$$

o coeficiente de escalonamento é dado por

$$\hat{c}_{t+1} = \frac{1}{\sum_{j=1}^{S} \tilde{\alpha}_{t+1}(j)}, \tag{2.11}$$

resultando

$$\hat{\alpha}_{t+1}(j) = \hat{c}_{t+1} \tilde{\alpha}_{t+1}(j). \tag{2.12}$$

O mesmo acontece para a variável "backward", isto é, as equações (2.8) e (2.9) são reescritas como

$$\hat{\beta}_T(j) = \hat{c}_T \quad 1 \leq j \leq S, \tag{2.13}$$

$$\tilde{\beta}_t(j) = \sum_{i=1}^{S} a_{ji} b_i(\vec{o}_{t+1}) \hat{\beta}_{t+1}(i) \qquad \begin{matrix} t = T-1, T-2, \ldots, 1 \\ 1 \leq j \leq S \end{matrix}, \tag{2.14}$$

em que também

$$\hat{\beta}_t(j) = \hat{c}_t \tilde{\beta}_t(j). \tag{2.15}$$

Finalmente, o termo $H(j,m)$ é dado por:

$$H_t(j,m) = \frac{c_m^{(j)} G(\vec{o}_t, \vec{\mu}_m^{(j)}, \vec{\sigma}_m^{2^{(j)}})}{\sum_{l=1}^{M} c_l^{(j)} G(\vec{o}_t, \vec{\mu}_l^{(j)}, \vec{\sigma}_l^{2^{(j)}})}. \tag{2.16}$$

Cada reestimação pelas equações (2.2), (2.3), (2.4) e (2.5) perfaz uma iteração chamada de *época*. Após cada época, um novo HMM é encontrado e a próxima época utiliza esse novo HMM para a reestimação dos novos valores de probabilidades de transição, médias, variâncias e pesos. Esse processo iterativo continua até que haja a convergência.

2.2 Análise de Componente Principal

A idéia de descrever aqui tal técnica surgiu devido à sua aplicação na redução de dimensão de parâmetros acústicos extraídos do sinal de fala (ver seção 2.3).

A técnica de *Análise de Componente Principal (PCA, do inglês "Principal Components Analysis")* é uma operação matemática que tem como objetivo uma redução de dimensão de variáveis aleatórias n-dimensionais e/ou uma interpretação de resultados [25]. Como já citado, a intenção em utilizar a PCA aqui neste trabalho é, a priori, a primeira citada acima.

De posse de um conjunto de variáveis aleatórias com N componentes, pode-se dizer que a variabilidade total desse conjunto pode ser expressa pelas suas N componentes. Porém existe um número K de componentes ($K < N$) que pode expressar uma variabilidade tão representativa quanto a variabilidade total do conjunto, isto é, as K componentes podem conter uma informação equivalente àquela contida nas N componentes [25]. É através da PCA que se pode dizer quais são essas K componentes.

Realizando a PCA, obtém-se as N componentes principais do conjunto das N variáveis aleatórias, que são uma combinação linear das variáveis aleatórias originais e não correlatas entre si. De um ponto de vista geométrico, as componentes principais formam um novo sistema de coordenadas que é visto como a rotação do sistema de coordenadas formado pelas variáveis aleatórias originais na direção de maior variabilidade [25]. Cada componente principal carrega informação referente à variabilidade do conjunto de variáveis aleatórias. A ordem da componente principal está relacionada à variabilidade que ela representa. Por exemplo, a primeira componente principal é aquela combinação linear das variáveis aleatórias, de um conjunto de variáveis aleatórias, que possui maior variância; a segunda componente principal é a combinação linear das variáveis aleatórias, desse mesmo conjunto, que possui a segunda maior variância e tal que a covariância entre ela e a primeira componente principal seja nula (não correlatas); a i-ésima componente principal é aquela combinação linear que tem a i-ésima maior variância e tal que as covariâncias entre ela e todas as outras $i-1$ componentes principais sejam nulas [25]. Tendo-se calculado todas as N componentes principais, pode-se considerar apenas aquelas K componentes principais que, somando as suas contribuições, representam a maior parte da variabilidade do conjunto de variáveis aleatórias. [25] comenta uma contribuição de K componentes que representem cerca de 80 a 90% da variabilidade total do conjunto. Porém essa faixa nem sempre é adequada.

2.2.1 Cálculo das componentes principais

Tendo posse de um conjunto de N variáveis aleatórias dado pela matriz $\mathbf{X} = (\vec{x}_1\vec{x}_2...\vec{x}_N)^T$, onde cada $\vec{x}_i$ é um vetor de variável aleatória com L observações [L x 1], faz-se necessário, para o cálculo das componentes principais via PCA, a obtenção da matriz de covariância $\mathbf{\Sigma}$ ou matriz correlação $\mathbf{\Gamma}$ de $\mathbf{X}$. Obtida a matriz escolhida ($\mathbf{\Sigma}$ ou $\mathbf{\Gamma}$), calculam-se os pares autovalor-autovetor dessa matriz e, em seguida, realiza-se o produto vetorial de cada autovetor pela matriz $\mathbf{X}$, obtendo-se os conjuntos de componentes principais $\mathbf{Y}$ de $\mathbf{X}$, como se segue:

De posse da matrix $\mathbf{X}$, que pode ser expressa como

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1L} \\ x_{21} & x_{22} & \cdots & x_{2L} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NL} \end{bmatrix}, \tag{2.17}$$

calcula-se o vetor $\vec{\bar{x}}$ cujos elementos $\bar{x}_i$ são as médias de cada vetor $\vec{x}_i$. Disso, calcula-se a matriz $\mathbf{\Sigma}$ ou $\mathbf{\Gamma}$. Escolhendo a matriz $\mathbf{\Sigma}$ por exemplo, tem-se

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_{NN} \end{bmatrix}, \tag{2.18}$$

em que

$$\sigma_{ij} = \frac{1}{L-1} \sum_{k=1}^{L} \left(x_{ik} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_j\right). \tag{2.19}$$

De $\mathbf{\Sigma}$, calcula-se o vetor $\vec{l}$ de autovalores e a matriz $\mathbf{V}$ de autovetores:

$$\vec{l} = \begin{bmatrix} l_1 & l_2 & \cdots & l_N \end{bmatrix} \tag{2.20}$$

e

$$\mathbf{V} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1N} \\ v_{21} & v_{22} & \cdots & v_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ v_{N1} & v_{N2} & \cdots & v_{NN} \end{bmatrix}, \tag{2.21}$$

em que cada linha da matriz $\mathbf{V}$ é um autovetor. Multiplicando-se cada autovetor por $\mathbf{X}$ obtém-se as N

componentes principais:

$$\begin{aligned}
\vec{y}_1^T &= \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1N} \end{bmatrix} * \mathbf{X}, \\
\vec{y}_2^T &= \begin{bmatrix} v_{21} & v_{22} & \cdots & v_{2N} \end{bmatrix} * \mathbf{X}, \\
&\vdots \\
\vec{y}_N^T &= \begin{bmatrix} v_{N1} & v_{N2} & \cdots & v_{NN} \end{bmatrix} * \mathbf{X}.
\end{aligned} \tag{2.22}$$

Ou, de maneira mais compacta,

$$\mathbf{Y} = \begin{bmatrix} \vec{y}_1 & \vec{y}_2 & \cdots & \vec{y}_N \end{bmatrix}^T = \mathbf{V} * \mathbf{X}. \tag{2.23}$$

É importante ressaltar o seguinte: a ordem da componente principal está associada ao autovalor correspondente ao autovetor que gerou tal componente principal, pois a variância da componente principal é igual ao autovalor associado ao autovetor que a gerou [25], isto é, o autovetor correspondente ao maior autovalor gerará a primeira componente principal, o autovetor correspondente ao segundo maior autovalor gerará a segunda componente principal, e assim por diante. Por esse fato, é interessante ordenar o vetor $\vec{l}$ e as linhas de $\mathbf{V}$ de acordo com $\vec{l}$, de tal forma que $\vec{y}_1$ seja a primeira componente principal, $\vec{y}_2$ a segunda componente principal, e assim sucessivamente.

Nota-se que o vetor $\vec{y}_i$ é o transposto do resultado do produto de um vetor de dimensão [1 x N] por uma matriz de dimensão [N x L], resultando em um vetor de dimensão [L x 1], que é a mesma dimensão de cada vetor $\vec{x}_i$ que forma a matriz $\mathbf{X}$.

Como neste trabalho a utilização da PCA será aplicada para parâmetros acústicos multidimensionais extraídos do sinal de fala, o termo N representa a dimensão do vetor de parâmetros (ver seção 2.3) e o termo L representa o número de vetores de parâmetros.

Uma abordagem matemática mais completa sobre PCAs é vista com detalhes em [25]. Na seção 3.1 será abordado a técnica de redução de dimensão PCA aplicada aos vetores de parâmetros.

2.3 Parametrização do sinal acústico

Na seção 2.1 foi abordado a teoria dos HMM's e comentado que, para o treinamento desses modelos, é necessário a obtenção de seqüências de observações **O**. Em reconhecimento de fala, essas observações são representadas por parâmetros extraídos do sinal acústico. Por facilidade, o termo sinal acústico será referido como *locução* no restante desta redação.

Locuções podem ser parametrizadas de diversas maneiras. Para cada parâmetro, o processo de parametrização é peculiar, porém um pré-processamento é realizado antes da extração dos parâmetros a fim de particionar o sinal de fala em *quadros*. A locução inteira é submetida inicialmente a uma pré-ênfase $(1 - 0,95z^{-1})$, e em seguida é "janelada" por janelas de Hamming de 20ms de duração que são deslocadas uma das outras de 10ms. Esse intervalo de 10ms equivale a um *quadro* da locução. A partir dos segmentos de fala janelados é que se obtém os parâmetros acústicos. Cada quadro é representado por um ou mais parâmetros acústicos correspondentes, que foram referidos na equação 2.2 à equação 2.9 pelo termo $\vec{o}_t$. Os parâmetros acústicos abordados aqui são os mais usados para a modelagem de HMM's em reconhecimento de fala: parâmetros Melcepstrais e parâmetros Log-energia normalizados.

2.3.1 Cálculo dos parâmetros acústicos

Parâmetros Melcepstrais

Obtidos os segmentos janelados de uma locução, a Transformada Discreta de Fourier (DFT) é calculada via algoritmo FFT. Os módulos ao quadrado da DFT dos segmentos são passados por um banco de F filtros espaçados na escala Mel [19] e é calculada a energia na saída de cada filtro. Em seguida aplica-se a Transformada Discreta do Cosseno (DCT) sobre o logaritmo das energias obtidas na saída do banco de filtros. Os coeficientes Melcepstrais para cada quadro são dados pela seguinte equação:

$$MEL_i = \sum_{j=1}^{F} 10log_{10}(X_j) \cos\left[i\left(j - \tfrac{1}{2}\right)\tfrac{\pi}{F}\right] \quad i = 1, 2, \ldots, Mc, \tag{2.24}$$

no qual:

- X_j é a energia na saída do j-ésimo filtro;
- Mc é o número de coeficientes Melcepstrais.

Vê-se que para cada quadro da locução é calculado um número Mc de coeficientes Melcepstrais, ou seja, cada quadro da locução é representado por um vetor de coeficientes Melcepstrais de dimensão Mc, sendo esses Mc coeficientes pouco correlacionados entre si, o que leva a considerar a matriz covariância das gaussianas como uma matriz diagonal, e podendo ser representada por um vetor cujos elementos são os elementos da diagonal principal. Neste trabalho, o número de coeficientes usados foi 12 ($Mc = 12$). Por conveniência, o vetor de coeficientes Melcepstrais será chamado de *parâmetro Melcepstral* no restante desta redação.

Parâmetros Log-energia normalizados

Para cada segmento janelado, a energia correspondente é calculada e normalizada pela maior energia dentre todos os segmentos da locução. Os parâmetros Log-energia normalizados são então os logaritmos das energias normalizadas. A energia em dB do t-ésimo quadro é dada por:

$$E_t = 10 \log_{10} \left(\sum_{i=1}^{\Upsilon_t} x^2 (i) \right), \tag{2.25}$$

no qual:

- $x(i)$ é a i-ésima amostra do quadro t;
- Υ_t é o número de amostras do quadro t.

Tomando a máxima energia dentre os quadros de toda a locução

$$E_{\max} = \max_t \{E_t\} \tag{2.26}$$

tem-se o *parâmetro Log-energia normalizado* (unidimensional) para cada quadro dado por:

$$En_t = E_t - E_{\max}. \tag{2.27}$$

Capítulo 3

Experimentos preliminares

3.1 Escolha dos parâmetros acústicos

A seção 2.3 abordou os principais parâmetros acústicos usados em reconhecimento de fala. Neste trabalho, foram adotados os seguintes parâmetros acústicos: parâmetro Log-energia normalizado (unidimensional) e parâmetro Melcepstral (dimensão 12) ; suas derivadas: parâmetro delta Log-energia normalizado e parâmetro delta Melcepstral; e suas segundas derivadas: parâmetro delta delta Log-energia normalizado e parâmetro delta delta Melcepstral.

Os parâmetros delta foram calculados segundo a seguinte expressão [33]:

$$\Delta_i(n) = \frac{1}{2J+1} \sum_{j=-J}^{J} j y_{i-j}(n) \tag{3.1}$$

em que

- $y_i(n)$ 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 quadros adjacentes a serem considerados no cálculo dos parâmetros delta do quadro

em questão. Neste trabalho utilizou-se $J = 1$ tanto para o cálculo dos parâmetros delta como para os delta delta.

Agrupou-se esses 6 parâmetros em um único *parâmetro composto* de dimensão 39 (Log-energia(1) + Melceptral(12) + delta Log-energia(1) + delta Melcepstral(12) + delta delta Log-energia(1) + delta delta Melcepstral(12)) para que se pudesse trabalhar com um único vetor de parâmetros.

A partir dos resultados obtidos em [8] com redução de dimensão de parâmetros compostos (Melcepstral(12) + delta Melcepstral(12) + delta delta Melcepstral(12)) de dimensão 36 com o uso da PCA (seção 2.2), resolveu-se optar por utilizar também essa técnica para a redução de dimensão do parâmetro composto de dimensão 39 e assim reduzir a complexidade computacional do processo de treinamento de HMM's, bem como a complexidade computacional envolvida no processo de adaptação (ver seção 5.2.1).

Após alguns experimentos, obteve-se os resultados apresentados na tabela 3.1 variando-se a dimensão dos parâmetros para o treinamento de HMM's:

Tempo de treinamento e desempenho de modelo independente de locutor com redução de dimensão do vetor de parâmetros acústicos

Dimensão dos parâmetros acústicos	Quantidade de informação associada à dimensão	Tempo de treinamento	Taxa de reconhecimento de palavras
5	89,88%	01h16min	69,89%
10	96,16%	01h36min	83,61%
15	98,47%	01h40min	85,66%
20	99,38%	02h01min	90,03%
25	99,71%	02h17min	90,26%
30	99,89%	02h57min	91,14%
35	99,97%	03h00min	89,16%
39	100,00%	03h25min	91,14%

Tabela 3.1: Tempo de treinamento e desempenho do modelo independente de locutor para diferentes dimensões do parâmetro composto.

Os testes foram realizados em um microcomputador Pentium IV (2,4GHz, 512MB de memória RAM) com 1200 locuções para cada treinamento.

Com base nos resultados mostrados na tabela 3.1, optou-se por adotar como dimensão final dos parâmetros compostos, a ***dimensão 25***. Vale ressaltar que as componentes de cada uma das 25 dimensões são consideradas independentes entre si o que faz com que as matrizes covariância das gaussianas sejam diagonais.

3.2 Treinamento dos HMM's

Antes de abordar o treinamento propriamente dito feito neste trabalho, será feita uma abordagem sobre os passos preliminares na etapa de estimação dos HMM's.

3.2.1 Segmentação da locução

Ao se trabalhar com fala contínua, os dados de treinamento, isto é, os parâmetros acústicos associados às locuções de treinamento precisam passar por um processo que visa separar os parâmetros relativos a cada fone dentro da locução. Esse processo é chamado de *segmentação*, e neste trabalho são utilizados dois tipos: *Uniforme* e *Via algoritmo de Viterbi*.

Imaginando a locução como um agrupamento de fones, pode-se pensar que a locução possui um HMM que é formado pela concatenação dos HMM's dos fones que a compõem. A fim de separar os parâmetros acústicos relativos a cada fone dentre o conjunto total de parâmetros acústicos da locução, pode-se distribuir, de maneira uniforme, os parâmetros acústicos entre todos os estados do suposto HMM da locução. Por exemplo, suponha que se tenha a locução "# *bola* #" (o símbolo "#" representa a unidade silêncio). Suponha também que foram extraídos 900 parâmetros acústicos compostos dessa locução. Pela transcrição da locução e pelo já visto nas seções anteriores, o HMM da locução possui 18 estados, resultado da concatenação dos HMM's dos fones (6 fones x 3 estados por fone). Dividindo 900 parâmetros em 18 estados, têm-se 50 parâmetros por estado. A intensão é armazenar separadamente os parâmetros relativos a um mesmo estado de um mesmo fone, isto é, os parâmetros acústicos relativos ao sétimo estado da locução "# *bola* #" (primeiro estado do fone "*o*") devem ser armazenados juntos com os parâmetros relativos ao décimo terceiro estado da locução "# *pato* #" (também primeiro estado do fone "*o*"). Ao se repetir esse processo com todas as locuções disponíveis para o treinamento estará se separando os parâmetros acústicos relativos a cada estado de cada um dos fones. A esse modo de separação denomina-se *Segmentação Uniforme*, pois os parâmetros são distribuídos igualmente entre os estados do HMM da locução.

Após a segmentação uniforme e de posse do conjunto de parâmetros compostos segmentados para cada estado de cada um dos fones, divide-se cada agrupamento de parâmetros de modo a obter um número de

partições igual ao número de gaussianas em cada estado, que neste trabalho foi feito igual a 5. Para cada partição, estima-se uma média, uma variância e um peso (que é dado pelo número de parâmetros da partição dividido pelo número de parâmetros total do conjunto). Disso obtém-se um HMM inicial para cada um dos fones.

O modelo gerado pela segmentação uniforme possui um agravante, pois a segmentação uniforme divide os parâmetros acústicos da locução igualmente entre os seus fones, o que implica dizer que os fones possuem a mesma duração dentro da locução, o que não é verdade. Por isso, realiza-se uma segunda segmentação tendo como base o alinhamento de Viterbi usando os HMM's advindos da segmentação uniforme. Para cada locução, o HMM dessa locução é montado segundo a sua transcrição e realizado o alinhamento de Viterbi [3], perfazendo assim uma nova separação dos parâmetros acústicos (sempre agrupando separadamente parâmetros acústicos referentes a um mesmo estado de um mesmo fone). Ao se realizar esse procedimento para todas as locuções de treinamento, estará se obtendo novos agrupamentos de parâmetros acústicos para cada estado de cada um dos 36 fones. A esse tipo de separação dá-se o nome de *Segmentação via Algoritmo de Viterbi*.

Fazendo novamente uma divisão em partições para cada agrupamento segmentado e calculando-se novas médias, variâncias e pesos, tem-se novos HMM's, que podem ser considerados mais robustos que aqueles advindos da segmentação uniforme. Esses novos HMM's são usados como modelos iniciais para o treinamento propriamente dito via Método de Baum-Welch (seção 2.1.2).

3.2.2 Software de treinamento dos HMM's

O software utilizado no treinamento de todos os HMM's deste trabalho [7] foi desenvolvido pelo prof. Dr. Carlos Alberto Ynoguti, atualmente professor do Instituto Nacional de Telecomunicações e colaborador do Laboratório Digital de Processamento de Fala - LPDF, onde se desenvolveu este trabalho.

O software possibilita o treinamento de HMM's a partir de três tipos de modelos iniciais:

- Modelos iniciais uniformes;
- Modelos iniciais via "Segmental K-Means";

- Modelos iniciais pré-treinados.

Neste último caso, os HMM's iniciais são HMM's treinados mas que podem servir de modelo inicial para um retreinamento dos mesmos. O software possibilita também o modelamento das funções de emissão de estado com um número fixo de gaussianas por mistura que pode variar entre 1 e 10.

Como já comentado, é treinado um HMM para cada um dos 36 fones a que se refere este trabalho. Os 36 HMM's possuem a mesma configuração, já previamente comentada mas resumida aqui:

- 3 estados;
- 5 gaussianas por estado;
- Cada gaussiana possui dimensão 25 (dimensão escolhida após redução de dimensão dos parâmetros compostos de dimensão 39 via PCA).

O conjunto dos 36 HMM's forma um modelo, que pode ser dependente ou independente de locutor. O software de treinamento tem então, como dados de entrada, os seguintes dados:

- Configuração do modelo;
- Locuções de treinamento (arquivos *.wav*);
- Transcrições das locuções de treinamento (arquivos *.txt*);

Os arquivos com as locuções no formato *.wav* foram gravados em 16 bits, modo mono e a uma freqüência de amostragem de 11025Hz. Os arquivos de transcrição *.txt* das locuções devem possuir os mesmos nomes dos arquivos *.wav* referentes às respectivas locuções.

De posse dos dados para o treinamento, o processo segue os seguintes passos:

1. Extração de parâmetros acústicos;
2. Estimação do modelo pela segmentação uniforme - 1 época;
3. Estimação do modelo pela segmentação via Viterbi - 1 época;
4. Estimação via Método Baum-Welch - várias épocas.

Na etapa de estimação via Método de Baum-Welch, o critério de parada para o término das épocas baseou-se na convergência do modelo. Ao final de cada época, é calculado a verossimilhança média dos HMM's recém treinados sobre 10% das locuções dadas para o treinamento empregando o algoritmo de Viterbi. Essa verossimilhança média é comparada com a verossimilhança média calculada na época anterior, e o processo de treinamento termina até que o seguinte critério seja atingido:

$$CP = \frac{P(O/\lambda_{atual}) - P(O/\lambda_{anterior})}{P(O/\lambda_{anterior})} < 10^{-3}$$

3.2.3 Configuração do treinamento para modelos dependente e independente de locutor

Para o treinamento do modelo independente de locutor, utilizou-se um base de dados formada por 1200 locuções. Um total de 30 locutores (15 homens e 15 mulheres) pronunciou, cada um, 40 frases de uma lista de 200 frases foneticamente balanceadas [11], resultando um total de 1200 locuções, isto é, 6 repetições do conjunto de 200 frases. Essa base de dados foi criada pelo prof. Dr. Carlos Alberto Ynoguti [33]. Para cada frase pronunciada tem-se o arquivo com a locução *.wav* e sua respectiva transcrição fonética *.txt* confeccionada manualmente para cada frase de cada locutor.

Para o treinamento de modelos dependente de locutor, usados no processo de adaptação, cada locutor pronunciou um total de 400 frases, isto é, 2 repetições da lista de frases foneticamente balanceadas [11]. Essa base de dados para o treinamento dos modelos dependente de locutor foi gerada no presente trabalho de tese tomando-se apenas locutores masculinos.

Devido ao tempo estabelecido não permitir e pela complexidade do processo de transcrição fonética manual, não foi possível obter as transcrições fonéticas de todas as 400 frases de cada um dos locutores que se dispuseram a fazer as gravações. Utilizou-se então uma transcrição padrão das 400 frases pronunciadas para todos os locutores. Esse fato tenderá a degradar os resultados, visto que os modelos dependente não foram treinados da maneira ótima.

3.3 Modificações no software de treinamento

No início deste trabalho, algumas novas implementações e modificações tiveram que ser feitas no software de treinamento adotado.

A priori, a ferramenta de treinamento não previa o parâmetro composto de dimensão 39 (Log-energia normalizado + Melcepstral + delta Log-energia normalizado + delta Melcepstral + delta delta Log-energia normalizado + delta delta Melcepstral) (ver seção 3.1), somente o parâmetro composto de dimensão 36. Um primeiro passo do trabalho foi a adaptação do software para o cálculo desse novo parâmetro composto.

Verificou-se que no processo do cálculo das gaussianas, através do algoritmo *LBG*, a divisão das partições era realizada de forma tal que, se o número de partições fosse diferente de uma potência de 2, primeiro se dividia o conjunto total em um número de partições igual à potência de 2 imediatamente superior ao número desejado de partições e depois se realizava um *aglutinamento* a fim de reduzir para o número desejado. Por exemplo, se o número desejado de partições fosse igual a 5 (5 gaussianas por estado), separava-se o conjunto de parâmetros compostos de cada estado para cada fone em 8 partições e, através de um processo de aglutinação, as 3 partições menos densas eram unidas às 3 partições mais próximas resultando em 5 partições finais. Resolveu-se alterar esse método de divisão de tal forma que, agora, é calculado um número de potência de 2 imediatamente inferior ao número desejado de partições, para através de uma *divisão*, atingir tal número desejado. Por exemplo, para o mesmo número de 5 gaussianas do exemplo anterior, divide-se o conjunto total em 4 partições e, a partir da partição com maior número de parâmetros compostos, subdivide-se tal partição em 2 obtendo-se 5 partições finais. Deve ser lembrado que, para cada uma das novas partições, um novo centróide deve ser recalculado. Esta nova maneira de divisão das partições foi motivada para evitar problemas numéricos, causados por partições vazias, quando o material de fala de treinamento é pequeno.

Outras modificações foram realizadas com o intuito de melhorar a interface do programa de treinamento para as implementações realizadas, mas fogem do escopo desta redação.

Capítulo 4

Adaptação de locutor

A idéia da adaptação de parâmetros é de bastante interesse em várias áreas da engenharia de processamento de sinais. Segundo [2], na área de tratamento de fala, a adaptação tem sido aplicada em muitos casos onde ocorre variação de características de canal, mudanças de ambiente, etc. O estudo de adaptação aqui abordado tem o intuito de tentar solucionar o problema da não representação de locutores por parte do sistema de reconhecimento de fala. Esse problema acontece quando nenhum dado de um determinado locutor é apresentado no treinamento de modelos independente de locutor, ou a quantidade de dados desse determinado locutor, quando apresentada no treinamento, não é suficiente para representá-lo.

Atualmente, os sistemas de reconhecimento de fala podem ser usados em muitas aplicações, tais como automação de "call centers", controle de sistemas via fala, discagem por fala, etc. Na maioria das aplicações que utilizam SRF, o número de usuários é relativamente grande. O mais prático para tais aplicações seria a utilização de sistemas de reconhecimento de fala com modelos independente de locutor robustos e que pudessem abranger uma grande variabilidade acústica. Porém, é constatado que modelos independente de locutor têm um desempenho em taxa de erro de palavra (TEP) da ordem de 2 a 3 vezes maior que modelos dependente de locutor [13] [31] [32]. Entretanto é constatado que a quantidade de material de fala necessária torna inviável o treinamento rápido de modelos dependente de locutor para cada novo usuário da aplicação (entre 400 e 600 locuções para aplicações em fala contínua). Por essa razão, surgiu a necessidade de implementar um processo que pudesse gerar um modelo específico para um determinado locutor a partir

de um modelo independente de locutor. Esse processo é chamado de *adaptação de locutor*.

Para a adaptação de locutor é necessário material de fala do locutor a ser adaptado. O material de adaptação pode ser do mais variado, porém a intensão é promover uma boa adaptação de locutor com o menor número de locuções, contribuindo para uma rapidez no processo de adaptação.

Nas aplicações que utilizam modelos independente de locutor, a adaptação de locutor visa a utilização desses sistemas por usuários que não possuem o perfil representado no treinamento do modelo independente de locutor. Em aplicações onde se faz necessário a utilização de um modelo dependente para cada locutor, a adaptação de locutor pode ser realizada dependendo da exigência da aplicação com respeito à TEP. Visto que nem sempre a adaptação de locutor é eficiente, ao ponto de produzir um modelo dependente de locutor que seja superior ao modelo dependente de locutor treinado pelos métodos padrões, a adaptação de locutor pode ser descartável para certas aplicações especiais. Também em sistemas com modelos independente de locutor bastante robustos que conseguem abranger uma grande variabilidade de parâmetros acústicos de locutores, a adaptação de locutor pode não surtir um efeito satisfatório.

Uma vantagem da adaptação de locutor é que nem todos os parâmetros necessitam ser reestimados para o locutor para o qual se deseja encontrar um modelo. Em sua maioria, os processos de adaptação são aplicados apenas para as médias gaussianas dos HMM's, podendo porém ser aplicados às variâncias, pesos das gaussianas, etc [2] [13] [16] [17]. Os parâmetros que não são adaptados (variâncias, pesos e probabilidades de transições entre estados) são herdados de um modelo independente de locutor, de preferência com robustez considerável. O fato de apenas alguns parâmetros serem adaptados faz com que a carga computacional requerida pelo processo de adaptação seja relativamente pequena quando comparada com as técnicas de treinamento de HMM's (ver seção 2.1.2), o que é interessante para sistemas reais que enfrentam restrições com respeito a recursos computacionais e que necessitem de uma rápida adaptação [32].

4.1 Formas de adaptação de locutor

Segundo [2], a adaptação de locutor pode ser realizada por diversos métodos:

- *"Adaptive clustering"*: Um conjunto de modelos independente de locutor é atualizado a partir de novos

dados independentes de locutor;

- *Conversão de locutor*: Um modelo específico para um locutor é convertido em um modelo específico para um novo locutor a partir de pequenas quantidades de dados desse novo locutor;
- *Adaptação de locutor*: Um modelo independente de locutor é adaptado para um modelo dependente de um novo locutor usando dados específicos desse novo locutor;
- *Adaptação sequencial*: Os dados específicos de um locutor são coletados ao longo do tempo e o modelo dependente desse locutor é atualizado à medida que novos dados de adaptação são obtidos.

A forma padrão mais utilizada para adaptação e também utilizada neste trabalho é a terceira: *adaptação de locutor*.

4.2 Modos de adaptação de locutor

4.2.1 Adaptação supervisionada e não-supervisionada

Quanto à disponibilidade da transcrição do material de fala dado para a adaptação, a adaptação de locutor pode ser *supervisionada* ou *não-supervisionada*.

No caso de uma adaptação supervisionada, os dados de adaptação fornecidos como entrada são compostos: pela(s) locução(ões) pronunciada(s) pelo locutor para o qual ser quer gerar um modelo adaptado e pela(s) sua(s) respectiva(s) transcrição(ões) fonética(s). Para adaptações não-supervisionadas, para que a(s) transcrição(ões) seja(m) conhecida(s), realiza-se um reconhecimento de fala da(s) locução(ões) de adaptação e uma posterior transcrição texto-fonética da(s) mesma(s). Neste caso, a confiabilidade do modelo independente de locutor que irá realizar esse reconhecimento deve ser alta, a fim de gerar uma boa estimativa da locução pronunciada.

4.2.2 Adaptação estática e dinâmica

No que diz respeito à maneira com que os dados de adaptação são expostos ao processo, a adaptação de locutor pode ser *estática* ou *dinâmica*.

Na adaptação estática, todo o material de fala para a adaptação é disponível no início do processo para a geração do modelo adaptado. Na adaptação dinâmica, parte do material de adaptação é utilizado para o processo de geração do modelo adaptado e, à medida que novos dados são colhidos, uma nova reestimação é efetuada.

O modo de adaptação de locutor escolhido irá depender da aplicação. Alguns sistemas utilizam um texto padrão, pronunciado pelo locutor que irá fazer uso do sistema. Nesse caso a adaptação é estática e supervisionada, entretanto a transcrição é considerada uma transcrição padrão, sem considerar aspectos peculiares do locutor.

Por conveniência, o locutor candidato à adaptação, visto que ele é considerado um elemento não modelado pelo sistema, será referenciado no restante desta redação como *novo locutor*.

4.3 Técnicas de adaptação de locutor

Existem muitas técnicas para o processo de adaptação de locutor. Segundo [2], essas técnicas podem ser baseadas em:

- Quantização vetorial;
- Características dos parâmetros;
- HMM's discretos;
- HMM's contínuos.

As que se baseiam no último ítem são as mais utilizadas e chamadas de *técnicas baseadas em modelos*.

Em [2] é abordado uma técnica de adaptação de parâmetros para HMM's contínuos, baseada em uma aprendizagem Bayesiana. Segundo [2], a eficiência dos modelos adaptados foi igual ou superior à eficiência dos modelos dependente de locutor, o que leva à conclusão que a adaptação pode tratar dados melhor do que os processos de treinamento padrão de HMM's.

De acordo com [32], as técnicas baseadas em modelos podem ser classificadas em três famílias:

- Família MAP (do inglês, "Maximum a Posteriori");

- Família das Transformações Lineares;
- Famílias dos Agrupamentos ("Clusterizações") de Locutores.

Da família MAP, destaca-se a técnica que lhe deu o nome, *MAP*. Da família das Transformações Lineares, destaca-se a *MLLR (do inglês, "Maximum Likelihood Linear Regression")*. Da família dos Agrupamentos de Locutores, será citado a técnica *CAT (do inglês, "Cluster Adaptive Training")*.

4.3.1 MAP

A técnica MAP é uma das mais clássicas técnicas de adaptação utilizadas [5] [6] [10] [12]. Em suma, a técnica MAP é baseada na estimação de máxima verossimilhança. De posse dos dados de adaptação, a técnica estima os parâmetros do novo modelo tais que a verossimilhança desse modelo, dado os dados de adaptação, seja máxima. Informação *a priori* advinda de modelos independente de locutor são combinadas aos dados de adaptação de um novo locutor para a estimação dos novos parâmetros adaptados que são estabelecidos segundo a probabilidade *a posteriori*, daí a origem do seu nome:

$$P[\lambda|\mathbf{O}] = \frac{P[\mathbf{O}|\lambda]P_0[\lambda]}{P[\mathbf{O}]}, \tag{4.1}$$

em que $P_0[\lambda]$ é a probabilidade a priori do modelo λ antes da observação de qualquer seqüência $\mathbf{O}$.

Para uma dada gaussiana, a sua média pode ser reestimada como

$$\vec{\hat{\mu}} = \frac{\tau\vec{\mu}_0 + \sum_{t=1}^{T}\gamma(t)\vec{o}_t}{\tau + \sum_{t=1}^{T}\gamma(t)}, \tag{4.2}$$

em que τ é o chamado *meta-parâmetro* que indica a "inclinação" entre a estimação de máxima verossimilhança da média e a média a priori μ_0, e $\gamma(t)$ é a probabilidade da gaussiana em questão gerar a observação $\vec{o}_t$.

A vantagem da técnica MAP é que, quanto mais dados de adaptação específicos de um dado locutor são apresentados, mais o modelo adaptado tenderá a se aproximar de um modelo dependente de locutor. Uma desvantagem é que a taxa de convergência da técnica é lenta, além de que a técnica atualiza somente os parâmetros relativos aos dados apresentados. Para sistemas de grande porte, com vários parâmetros para

serem adaptados, mesmo para um conjunto de dados de adaptação razoável, muitos parâmetros não terão representatividade e assim não serão atualizados. Por esse fato, a técnica não é aplicada para adaptações com pequenas quantidades de dados de adaptação.

A técnica MAP possui variantes como a *RMP (do inglês, "Regression Based Model Prediction")* e o *SMAP(do inglês, "Structural MAP")*, que são abordadas em [4] [24].

4.3.2 MLLR

Outra técnica clássica usada em adaptação é a MLLR [20] [22] [23]. Essa técnica utiliza transformações lineares dos parâmetros do modelo, obtidas do modelo independente de locutor, para a adaptação desses parâmetros em conjunto com os dados de adaptação, formando um novo modelo adaptado. De acordo com [32] a vantagem da utilização de transformações lineares é que uma mesma transformação pode ser utilizada por várias gaussianas (ou mesmo todas) de um HMM, e esse compartilhamento permitiria uma adaptação rápida.

Para uma dada gaussiana, a sua média, segundo MLLR, é adaptada da seguinte forma:

$$\hat{\vec{\mu}} = \mathbf{A}\vec{\mu} + \vec{b}, \tag{4.3}$$

em que $\mathbf{A}$ é uma matriz [D x D] (D é a dimensão do vetor observação ou vetor de parâmetros acústicos) e $\vec{b}$ é um vetor D-dimensional. A equação 4.3 pode ser reescrita como

$$\hat{\vec{\mu}} = \mathbf{W}\vec{\xi}, \tag{4.4}$$

em que a matriz $\mathbf{W}$ é de dimensão [D x ($D+1$)] e é chamada de *matriz transformação*, e $\vec{\xi}$ é um vetor composto de médias dado por:

$$\vec{\xi}^T = \begin{bmatrix} 1 & \mu_1 & \dots & \mu_D \end{bmatrix}. \tag{4.5}$$

A matriz $\mathbf{W}$ é estimada de tal forma que a verossimilhança do novo modelo para com os dados de adaptação seja maximizada. No início a matriz de transformação é obtida de um modelo independente de

locutor e em seguida é reestimada a partir dos dados de adaptação. A partir da nova matriz **W** os parâmetros do modelo são reestimados, gerando um novo modelo específico para um dado locutor.

Como muitas gaussianas podem compartilhar uma mesma matriz transformação, muitos parâmetros podem ser adaptados com pouco material, contribuindo para uma adaptação rápida. Uma maneira de classificar as matrizes transformação é a utilização de uma árvore de classes, onde as matrizes são agrupadas de acordo com proximidades acústicas [32]. Esse processo é denominado *"Regression Class Tree"* [21].

Uma vantagem da técnica MLLR com relação a técnica MAP é que ela pode atualizar todos os parâmetros do modelo, representados ou não pelos dados de adaptação [20] [22]. Uma desvantagem da MLLR é a necessidade de uma quantidade de dados razoável para a estimação das matrizes transformação. No caso de um compartilhamento global, isto é, todas as gaussianas compartilharem uma mesma matriz transformação, deve haver dados necessários para a estimação da matriz global; também no caso da regressão por árvore de matrizes ("Regression Class Tree"), a matriz raiz deve ser estimada com uma quantidade de dados razoavelmente grande dependendo do porte do sistema.

Tanto em MAP quanto em MLLR, a convergência cresce à medida que os dados de adaptação também crescem. Alguns estudos foram feitos agrupando as duas técnicas: usando MLLR em uma primeira parte da adaptação e depois aplicando a técnica MAP para um refinamento dos parâmetros cuja representação não foi suficiente [16] [17] [18] [31].

A técnica MLLR também possui algumas variantes como citado nas referências [9] [14].

4.3.3 CAT

A técnica de adaptação CAT [26] [27] faz parte da Família de Agrupamento de Locutores e consiste em representar o locutor, para o qual se quer encontrar o modelo adaptado, como uma combinação linear de modelos relacionados a agrupamentos de locutores. Esse conjunto de modelos forma uma base canônica de representação. De posse da base de modelos e de dados de adaptação provenientes do novo locutor, calculam-se coeficientes baseados em máxima verossimilhança que irão ponderar os modelos da base canônica a fim de encontrar o novo modelo. É assumido que somente os parâmetros médias das gaussianas são adaptados.

Capítulo 5

Técnica de adaptação via "Eigenvoices"

No capítulo anterior foi feita uma abordagem sobre as técnicas mais clássicas para adaptação de locutor (MAP e MLLR). O processo de adaptação via *"eigenvoices"*, base deste trabalho, segue uma linha diferente dessas primeiras pois se baseia em informações sobre peculiaridades de locutores para a estimação do novo locutor, o que não é usado pelas técnicas MAP e MLLR.

5.1 "Eigenfaces"

A idéia da adaptação via "eigenvoices" procede da técnica de reconhecimento de imagens, mais especificamente faces humanas, usando as *"eigenfaces"* [1]. O procedimento usando "eigenfaces" consiste no seguinte: de posse de um conjunto base de imagens de faces humanas, uma nova face pode ser considerada como sendo uma combinação linear de informações advindas de cada uma daquelas que compõem o conjunto base. Cada imagem de face pode ser identificada como um conjunto de pontos próprios, chamado aqui de *pontos dados*. Os pontos dados de todas as imagens do conjunto base formam um espaço dimensional relativamente grande, *espaço de imagens*, dependendo da quantidade de imagens que compõe o conjunto. A intenção é realizar a PCA (ver seção 2.2) sobre todos os pontos dados com o intuito de diminuir a dimensionalidade original do espaço de imagens, como feito em [15]. Obtendo os K autovetores de maior magnitude da matriz covariância ou correlação da matriz formada por todos os pontos dados, estará se obtendo as K direções de

maior variabilidade do espaço de imagens original, isto é, as K direções que contém as maiores informações do espaço. Como os autovetores são ortogonais entre si, esses K autovetores constituem um novo espaço de imagens com dimensão K. A média dos pontos dados que compõem o espaço de imagens é chamada de *"eigenface 0"* e os K autovetores são chamados de *"eigenfaces"*. Em seguida, a idéia é representar uma nova imagem de face humana como a soma da "eigenface 0" mais uma combinação linear das K "eigenfaces".

5.2 "Eigenvoices"

A técnica de adaptação via "eigenvoices" [13] [16] [17] [18] [30] [31] proposta por Kuhn, é baseada na mesma idéia das "eigenfaces" porém voltada para a área de reconhecimento de fala. Ela se enquadra no conjunto de técnicas que se baseiam em algoritmos capazes de, a partir de um conjunto de dados obtidos de locutores de referência, realizar a adaptação de locutor sem a necessidade de um grande montante de material de voz a ser coletado do novo locutor (família das "Clusterização" de Locutores, ver seção 4.3), pois esses algoritmos visam representar esse novo locutor como uma combinação linear de locutores base, semelhante à técnica CAT (ver seção 4.3.3).

Assim como no tratamento de faces, na técnica via "eigenvoices" é necessário um conjunto de locutores, os *locutores base*. Para cada um desses locutores, um modelo dependente de locutor é treinado e esse conjunto de modelos treinados irá compor o espaço de locutores original. Dois pontos devem ser destacados aqui:

- A qualidade dos modelos dependente de locutor;
- A quantidade de modelos dependente de locutor.

O primeiro ponto está relacionado à representação daquele locutor dentro do conjunto. Quanto melhor treinado o modelo dependente de locutor, melhor representado estará o locutor; o segundo ponto está ligado à variabilidade do conjunto. Quanto mais modelos, maior variabilidade de fala estará capturada, facilitando a representação de novos locutores.

5.2.1 Obtenção dos "eigenvoices"

Como já visto, um HMM é representado pelo trio $(\mathbf{\Pi}, \mathbf{A}, \mathbf{B})$ no qual o termo $\mathbf{B}$ representa as funções de emissão dos estados do modelo (ver seção 2.1.1). Para HMM's contínuos, tipo adotado neste trabalho, as funções de emissão são dadas por gaussianas caracterizadas por seus pesos (no caso de misturas), médias e variâncias multidimensionais. As técnicas de adaptação atuais têm foco na reestimação das médias das gaussianas dos HMM's, herdando os demais parâmetros de modelos independente de locutor. Isso se deve ao fato de que as médias são os parâmetros de maior peso na caracterização de um locutor.

O processo de obtenção dos "eigenvoices" não participa do processo de adaptação propriamente dito, pois esses podem, e é bom que sejam, obtidos previamente. Dois fatores contribuem para o cálculo prévio dos "eigenvoices": eles demandam um esforço computacional relativamente grande dependendo da dimensão dos parâmetros acústicos, da configuração do HMM e do número de modelos dependente de locutor; e não há a necessidade de dados de adaptação do novo locutor para o cálculo.

Para cada modelo dependente de locutor, tomam-se todas as médias das gaussianas de todos os HMM's que compõem o modelo e armazenam-se estas médias de maneira sequencial em um único vetor. Assim, cada locutor base é representado por esse único vetor. Esses vetores específicos são chamados de *supervetores*. Neste trabalho, cada modelo é composto de 36 HMM's com 3 estados, 5 gaussianas por estado, em que cada gaussiana possui dimensão 25. Ou seja, cada supervetor é composto por 36 x 3 x 5 x 25 = 13500 elementos. A ordem com que as médias gaussianas são armazenadas nos supervetores não tem importância, o importante é que a ordem seja mantida para todos os supervetores. Por exemplo, o elemento de 5ª ordem da 3ª gaussiana do 2º estado do HMM do fone *"a"* do modelo do locutor 1 dentro do supervetor 1, deve ter a mesma posição do elemento de 5ª ordem da 3ª gaussiana do 2º estado do HMM do fone *"a"* dos outros modelos dependente de locutor dentro dos seus respectivos supervetores. Esse procedimento se aplica aos 13500 elementos.

Depois de montados os L supervetores (L igual ao número de locutores base), é aplicada uma técnica de redução de dimensão, a PCA por exemplo, para encontrar os autovetores e autovalores da matriz covariância ou matriz correlação dos L supervetores. Fazendo uma comparação com o exemplo dado na seção 2.2.1, o termo N é aqui dado pelos 13500 elementos médias dos HMM's que se quer adaptar e o termo L é o

número de supervetores. Neste trabalho, o cálculo tanto da matriz covariância quanto da matriz correlação do conjunto de supervetores exige um esforço computacional muito grande, visto que as matrizes resultantes seriam da ordem de 13500 linhas por 13500 colunas. Devido a esse fato, muitos trabalhos, [28] [31] entre outros, optaram por usar outra técnica para o cálculo dos autovetores e autovalores a partir da matriz $\mathbf{ZZ}^T$, em que $\mathbf{Z}$ é a matriz formada pelos L supervetores, cuja dimensão é [13500 x L].

A técnica *SVD (do inglês, "Singular Value Decomposition")*, baseada no resultado da álgebra linear que diz que qualquer matriz de dados $\mathbf{U}$ de dimensão [M x N] pode ser expressa no seguinte produto de matrizes:

$$\mathbf{U}_{[MxN]} = \mathbf{P}_{[MxN]}.\mathbf{Q}_{[NxN]}.\mathbf{R}^T_{[NxN]}; \tag{5.1}$$

permite calcular as matrizes que compõem esse produto. A matriz $\mathbf{P}$ possui colunas ortogonais o que caracteriza uma base canônica e é formada pelos autovetores da matriz $\mathbf{UU}^T$; a matriz $\mathbf{Q}$ é uma matriz diagonal cujo elementos da diagonal são os autovalores da matriz $\mathbf{UU}^T$. A vantagem da SVD é economia de esforço computacional, visto que não é necessário o cálculo das matrizes quadradas (covariância e correlação) para o cálculo dos autovetores e autovalores. A utilização da matriz $\mathbf{ZZ}^T$ para o cálculo de autovalores e autovetores ao invés da matriz covariância ou correlação de $\mathbf{Z}$ não traz problema algum para a adaptação [28] [31].

Aplicada a técnica SVD sobre a matriz de supervetores $\mathbf{Z}$, de dimensão [13500 x L], tem-se L autovetores de dimensão [13500 x 1] e L autovalores correspondentes. Pela teoria da PCA [25], os autovetores "apontam" para as direções de maior variabilidade do conjunto de dados, cujas variabilidades são representadas pelos autovalores relacionados, ou seja, o autovetor cujo autovalor correspondente possui maior magnitude dentre todos os autovalores "aponta" para a direção de maior variabilidade do conjunto; o autovetor cujo autovalor correspondente possui a segunda maior magnitude dentre todos os autovalores "aponta" para a direção que apresenta a segunda maior variabilidade do conjunto de supervetores, e assim sucessivamente. Com isso, é interessante ordenar os L autovetores em ordem decrescente de acordo com a magnitude dos L autovalores.

Os L autovetores são denominados *"eigenvoices"*, e formam agora um novo espaço, o *espaço de locutores*, onde o novo locutor (locutor para o qual se quer encontrar um modelo adaptado) deverá ser representado.

Um fator importante a se considerar é a média dentre todos os L supervetores chamada de *"eigenvoice 0"*. O propósito da técnica é representar o novo locutor como uma combinação linear dos L "eigenvoices" mais o "eigenvoice 0", ou seja, representar o novo locutor como um "ponto" dentro do espaço de locutores. Esse "ponto" não é nada mais que um vetor de dimensão [13500 x 1] correspondendo ao supervetor do novo locutor onde estão contidas todas as médias gaussianas do modelo adaptado final.

Em suma, na técnica via "eigenvoices", a média adaptada é dada por:

$$\vec{\hat{\mu}} = \vec{e}(0) + \sum_{j=1}^{L} w_j \vec{e}(j), \tag{5.2}$$

em que $\vec{e}(0)$ é o "eigenvoice 0" e $\vec{e}(j)$ são os "eigenvoices" restantes ponderados pelos coeficientes w_j. O objetivo agora é encontrar os coeficientes que ponderam os "eigenvoices".

Para exemplificar, considere o caso hipotético de 100 locutores base, em que cada modelo dependente de locutor seja formado por apenas um HMM que possua apenas um estado e uma gaussiana bidimensional por estado, isto é, cada locutor base, representado pelo seu supervetor, é um par ordenado real (x,y). Montada a matriz $\mathbf{Z}$ [2 x 100] (onde 2 é dimensão do supervetor e 100 é o número de locutores) e aplicada a SVD, tem-se apenas 2 "eigenvoices" (pois de uma matriz $\mathbf{Z}$ [2 x 100], sua matriz $\mathbf{Z}\mathbf{Z}^T$ será de dimensão [2 x 2], que possui somente 2 autovalores e 2 autovetores). O novo locutor é representado pelo ponto $\vec{\hat{\mu}}$ dado pelo resultado da combinação linear dos 2 "eigenvoices", mais o "eigenvoice 0", como visto na figura 5.1.

Observa-se que os "eigenvoices" 1 e 2 possuem magnitudes de ordem diferente em relação à magnitude dos supervetores. Se o novo locutor fosse representado somente como a combinação dos "eigenvoices" ponderados pelos coeficientes w's, esses coeficientes teriam que possui valores de ordem elevada para compensar a baixa magnitude dos "eigenvoices". Uma maneira de realizar o processo é normalizar as médias gaussianas e os parâmetros acústicos subtraindo-os do "eigenvoice 0" correspondente, realizando uma translação da nurvem de dados para valores em torno da origem. Depois de estimar os coeficientes w's e consequentemente as médias adaptadas, basta adicionar o "eigenvoice 0" correspondente a cada nova média estimada, obtendo o valor real. O "eigenvoice 0" é então uma constante que deve ser somada à combinação linear de "eigenvoices", ou pode ser considerada uma translação associada à rotação do sistema de coordenadas. Essa combinação é

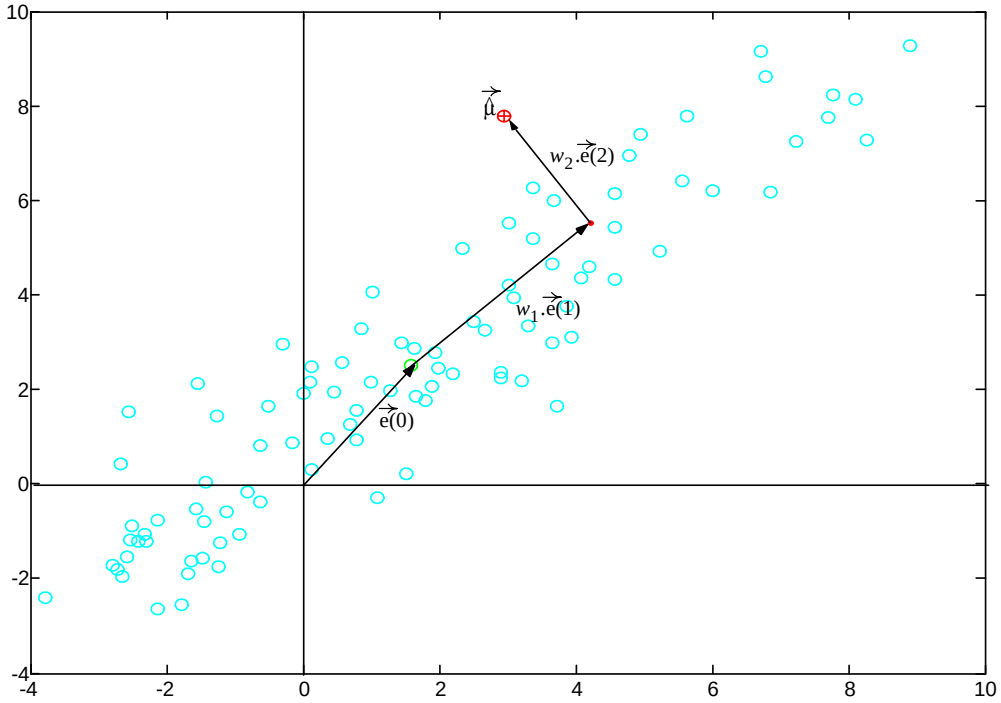

Figura 5.1: Representação do novo locutor no espaço de locutores pelo ponto $\vec{\hat{\mu}}$

responsável pela representação relativa do novo locutor no espaço de locutores, para que a média adaptada, que nada mais é que o supervetor do novo locutor (ponto em destaque), tenha magnitude da ordem dos supervetores. Considerando uma outra analogia, de que um ponto no espaço cartesiano bidimensional (x,y) é identificado pelas suas coordenadas nos eixos x e y, o novo locutor pode ser representado pelas suas coordenadas ao longo dos eixos definidos pelos "eigenvoices".

Assim como na PCA as componentes principais de mais alta ordem podem ser descartadas por não representarem muito da informação do conjunto, "eigenvoices" também podem ser descartados. A razão é que "eigenvoices" são autovetores que indicam as regiões de variabilidade do conjunto. Desse fato, de posse de um conjunto muito grande de "eigenvoices", é comum se descartar os "eigenvoices" que "carregam" menor informação, caracterizando uma redução de dimensão. Como os "eigenvoices" são ordenados de acordo com os autovalores, pode-se considerar apenas os K primeiros "eigenvoices" do total de L. Uma observação que deve ser considerada é que a técnica via "eigenvoices" considera que o novo locutor deve se encontrar dentro do espaço de locutores, o que indica que o número de locutores base, bem como a redução do número de

"eigenvoices", são fatores importantes para a aplicação da técnica.

5.2.2 Estimação dos coeficientes dos "eigenvoices"

Como visto na seção anterior, a técnica de adaptação via "eigenvoices" adapta as médias do modelo para um novo locutor através da equação 5.2.

Para a estimação dos coeficientes dos "eigenvoices" são necessários os seguintes dados:

- $\vec{o}_t$: Vetor de parâmetros acústicos relativo ao t-ésimo quadro extraído da(s) locução(ões) de adaptação;
- $\mathbf{C}_m^{(s)^{-1}}$: Matriz covariância inversa da m-ésima gaussiana do s-ésimo estado de todos os HMM's do modelo independente de locutor;
- $\vec{\mu}_m^{(s)}$: Vetor de médias da m-ésima gaussiana do s-ésimo estado de todos os HMM's do modelo independente de locutor;
- $\gamma_m^{(s)}(t)$: Probabilidade da m-ésima gaussiana do s-ésimo estado do HMM da locução de adaptação gerar o vetor de parâmetros $\vec{o}_t$;
- $\mathbf{E} = \{\vec{e}(0)\vec{e}(1)\vec{e}(2)...\vec{e}(L)\}$: Conjunto de "eigenvoices" previamente calculados.

A função do modelo independente de locutor é servir como condição inicial para o processo de estimação dos coeficientes.

Em trabalhos anteriores aplicados a palavras isoladas [16] [17] [18] [28] [30], a estimação dos coeficientes dos "eigenvoices" foi baseada em palavras. No caso de fala contínua, [31] realiza adaptações com sentenças, mas a publicação não esclarece detalhes de como isso é feito. Neste trabalho, optou-se por realizar a estimação dos coeficientes usando uma locução inteira. O HMM da locução é montado pela concatenação dos HMM's dos fones que a compõem (os HMM's são provenientes do modelo independente de locutor). Para isso faz-se necessário o conhecimento da transcrição fonética da locução a fim de se saber quais fones a compõem, o que caracteriza uma adaptação supervisionada (ver seção 4.2).

A idéia do algoritmo de estimação dos coeficientes dos "eigenvoices" é baseada na estimação de máxima verossimilhança segundo uma "eigen-decomposição" ou *MLED (do inglês "Maximum Likelihood Eigen-Decomposition")*, isto é, tentar maximizar a probabilidade do novo modelo dependente de locutor gerar a

seqüência dos parâmetros acústicos $\vec{o}_t$. Para tal, usa-se a *função auxiliar de Baum*, equação (2.1), [3]. Essa equação pode ser reescrita como [30]:

$$Q(\lambda, \bar{\lambda}) = -\frac{1}{2} P(O/\lambda) . \sum_s \sum_m \sum_t \gamma_m^{(s)}(t) f(\vec{o}_t, s, m), \tag{5.3}$$

em que

$$f(\vec{o}_t, s, m) = [-D \log(2\pi) - \log \left| \mathbf{C}_m^{(s)^{-1}} \right| + h(\vec{o}_t, s, m)] \tag{5.4}$$

e

$$h(\vec{o}_t, s, m) = \left(\vec{\bar{\mu}}_m^{(s)} - \vec{o}_t \right)^T \mathbf{C}_m^{(s)^{-1}} \left(\vec{\bar{\mu}}_m^{(s)} - \vec{o}_t \right), \tag{5.5}$$

e D é a dimensão dos parâmetros acústicos, nesse caso $D = 25$.

As equações de estimação dos coeficientes são baseadas na representação da nova média como uma combinação linear dos "eigenvoices", ou seja,

$$\vec{\bar{\mu}} = \sum_{j=1}^{K} w_j \vec{e}(j), \tag{5.6}$$

em que K ($K \leq L$) é o número de "eigenvoices" escolhidos para compor o espaço de representação do novo locutor. Mais à frente é comentado sobre a compensação para se utilizar a equação (5.2) e não a equação (5.6) no processo de adaptação.

Com relação aos "eigenvoices", sabe-se que cada um é um vetor de dimensão [N x 1], e neste trabalho N = 13500, ou seja, ele possui dimensão igual ao supervetor. Disso, o primeiro elemento de qualquer "eigenvoice" está relacionado com o primeiro elemento do supervetor, que nada mais é que o primeiro elemento, dentre os 25, da primeira média gaussiana armazenada no supervetor, e assim por diante. Com isso, cada "eigenvoice" pode ser considerado como um vetor dado pela equação 5.7, em que M é o número de gaussianas na mistura ($M = 5$), S é o número total de estados do modelo dependente de locutor, isto é, S é o número de estados do HMM multiplicado pelo número de HMM's do modelo (S = 3 x 36 = 108); e cada elemento $\vec{e}_m^{(s)}(j)$ é um vetor de dimensão [25 x 1] relativo a m-ésima gaussiana do s-ésimo estado do modelo dependente de locutor correspondente ao "eigenvoice" de ordem (j). Nesta abordagem, o estado s está diretamente relacionado

à posição que o fone, ao qual o estado está associado, ocupa dentro do supervetor no momento de sua montagem. Para uma melhor compreensão, considere o seguinte exemplo: Seja um modelo composto apenas pelos 5 seguintes fones: "*a*", "*e*", "*i*", "*o*" e "*u*"; e que a montagem dos supervetores obedeça a seguinte ordem: {(médias do primeiro estado do HMM de "*a*") + (médias do segundo estado do HMM de "*a*") + (médias do terceiro estado do HMM de "*a*) + (médias do primeiro estado do HMM de "*e*") + (médias do segundo estado do HMM de "*e*") + ... + (médias do terceiro estado do HMM de "*u*") }. O primeiro estado ($s = 1$) do modelo é o primeiro estado do HMM de "*a*", o segundo estado ($s = 2$) do modelo é o segundo estado do HMM de "*a*", seguindo a ordem de formação do supervetor, o décimo estado ($s = 10$) do modelo é o primeiro estado do HMM de "*o*", e assim sucessivamente até o último estado do modelo ($s = S = 15$) que refere-se ao terceiro estado do HMM de "*u*".

$$\vec{e}(j) = \begin{bmatrix} \vec{e}_1^{(1)}(j) \\ \vec{e}_2^{(1)}(j) \\ \vdots \\ \vec{e}_M^{(1)}(j) \\ \vec{e}_1^{(2)}(j) \\ \vec{e}_2^{(2)}(j) \\ \vdots \\ \vec{e}_M^{(2)}(j) \\ \vdots \\ \vec{e}_m^{(s)}(j) \\ \vdots \\ \vec{e}_1^{(S)}(j) \\ \vec{e}_2^{(S)}(j) \\ \vdots \\ \vec{e}_M^{(S)}(j) \end{bmatrix}. \tag{5.7}$$

Durante o processo de adaptação, a(s) locução(ões) de adaptação não contemplam necessariamente todos

os fones, com isso nem todos os "eigenvoices" serão considerados para a estimação dos coeficientes w's. Isto é, para uma determinada locução, serão considerados somente os "eigenvoices" $\vec{e}_m^{(s)}(j)$'s ($j = 1, 2, ..., K$) correspondentes aos fones que constam na locução. Considerando o exemplo acima, se a locução de adaptação for "*ai*", serão considerados somente os termos $\vec{e}_m^{(s)}(j)$'s para $s = 1$, 2 e 3 (fone "*a*"); e $s = 7$, 8 e 9 (fone "*i*").

Manipulando as equações (5.3), (5.4), (5.5) e (5.6), pode-se chegar a uma equação, resultado da derivada da equação (5.3) com respeito aos coeficientes w's (incorporados pela equação (5.6)) igualada a zero [30], dada por:

$$\begin{aligned} &\sum_s \sum_m \sum_t \left[\gamma_m^{(s)}(t) \left(\vec{e}_m^{(s)}(j) \right)^T \mathbf{C}_m^{(s)^{-1}} \vec{o}_t \right] = \\ &= \sum_s \sum_m \sum_t \left\{ \gamma_m^{(s)}(t) \left[\sum_{k=1}^{K} \left(w(k) \left(\vec{e}_m^{(s)}(k) \right)^T \mathbf{C}_m^{(s)^{-1}} \vec{e}_m^{(s)}(j) \right) \right] \right\}, \end{aligned} \qquad j = 1, 2, ..., K \tag{5.8}$$

em que o termo $\vec{e}_m^{(s)}(j)$ é o "eigenvoice" de ordem j relativo à m-ésima gaussiana do s-ésimo estado do HMM da locução. Vê-se que a equação acima deve ser implementada para $j = 1, 2, ..., K$, obtendo-se assim K equações com K incógnitas, sendo estas incógintas os coeficientes w's.

Fazendo algumas manipulações com a equação (5.8) pode-se chegar a uma forma de escrever os coeficientes w's em função de outras variáveis:

$$\delta(1, j)w_1 + \delta(2, j)w_2 + \cdots + \delta(K, j)w_K = \Lambda(j), \quad j = 1, 2, ..., K \tag{5.9}$$

em que

$$\delta(k, j) = \sum_s \sum_m \left[\sum_{n=1}^{N} \left(\frac{e_{mn}^{(s)}(k).e_{mn}^{(s)}(j)}{\sigma_{mn}^{2^{(s)}}} \right) . \sum_t \gamma_m^{(s)}(t) \right] \tag{5.10}$$

e

$$\Lambda(j) = \sum_s \sum_m \sum_t \left[\sum_{n=1}^{N} \left(\frac{e_{mn}^{(s)}(j).o_{tn}}{\sigma_{mn}^{2^{(s)}}} \right) . \gamma_m^{(s)}(t) \right], \tag{5.11}$$

em que $e_{mn}^{(s)}(j)$ é o n-ésimo elemento do "eigenvoice" $\vec{e}_m^{(s)}(j)$, o_{tn} é o n-ésimo elemento do vetor de parâmetros acústicos $\vec{o}_t$ e $\sigma_{mn}^{2^{(s)}}$ é o n-ésimo elemento da variância da gaussiana m do estado s do HMM da locução de adaptação. Relembra-se aqui que a matriz covariância de cada gaussiana do HMM é diagonal, pois, pela consideração de independência entre os elementos do vetor de parâmetros acústicos (ver seção 2.3.1), as

componentes das gaussianas também são consideradas independentes entre si.

Expandindo-se a equação (5.9) para todo $j = 1, 2, ...K$, tem-se o sistema de equações abaixo:

$$\begin{cases} \delta(1,1)w_1 + \delta(2,1)w_2 \cdots + \delta(K,1)w_K = \Lambda(1) \\ \delta(1,2)w_1 + \delta(2,2)w_2 \cdots + \delta(K,2)w_K = \Lambda(2) \\ \vdots \\ \delta(1,K)w_1 + \delta(2,K)w_2 \cdots + \delta(K,K)w_K = \Lambda(K) \end{cases} \tag{5.12}$$

com K equações e K incógnitas. Aplicando algumas técnicas de resolução de sistemas lineares, como *eliminação gaussiana* aplicada neste trabalho, os K coeficientes são calculados.

Para um melhor entendimento, pode se considerar o seguinte exemplo: Considere a seguinte locução de adaptação: "*b a c o*", constando de 4 fones, e seja 15, por exemplo, o número de vetores de parâmetros acústicos extraídos (relativos a cada quadro da locução). Sabe-se que cada HMM de qualquer fone possui 3 estados e 5 gaussianas multidimensionais (dimensão 25) em cada estado. O HMM da locução de adaptação é então obtido com a concatenação do HMM de "*b*" + HMM de "*a*" + HMM de "*c*" + HMM de "*o*", com 12 estados no total e 5 gaussianas multidimensionais em cada estado. Esses 4 HMM's são todos advindos do modelo independente de locutor. Montado o HMM da locução, obtém-se os termos $\vec{\mu}_m^{(s)}$'s e $\mathbf{C}_m^{(s)^{-1}}$'s, os "eigenvoices" $\vec{e}_m^{(s)}(j)$'s correspondentes (já previamente calculados) e calculam-se os $\gamma_m^{(s)}(t)$ para cada um dos 15 vetores de parâmetros da locução de adaptação. Considere também que, de um total de L supervetores, foram considerados apenas 5 "eigenvoices" para a representação total do espaço de locutores ($K = 5$) mais o "eigenvoice 0" (ver seção 5.2.1). Ao aplicar a equação (5.8) para $j = 1, 2, ..., 5$, considerando o HMM da locução ($s = 1, 2..., 12$; $m = 1, 2..., 5$ e $t = 1, 2..., 15$) e realizando as devidas substituições para cada aplicação, obtém-se um sistema de 5 equações com 5 incógnitas. Resolvendo o sistema, substitui-se os coeficientes solução na equação (5.2) encontrando-se todas as médias gaussianas adaptadas, como pode ser visto na equação (5.13).

$$\vec{\hat{\mu}}_m^{(s)} = \vec{e}_m^{(s)}(0) + \sum_{j=1}^{L} w_j \vec{e}_m^{(s)}(j) \tag{5.13}$$

Querendo-se obter a média de maior peso do primeiro estado do fone "a" do novo locutor, monta-se a equação (5.13) referente a essa média (com os "eigenvoices" correspondentes) aplicando os coeficientes w's encontrados da resolução do sistema de equações (5.12). Os coeficientes w's são comuns na obtenção de todas as médias gaussianas do novo modelo, o que difere são somente os "eigenvoices" junto com o "eigenvoice 0" que são específicos para cada média.

Estimadas todas as médias gaussianas e herdando os parâmetros restantes (pesos e variâncias das gaussianas e probabilidades de transição) do modelo independente de locutor, o novo modelo adaptado é encontrado. A idéia do MLED é ser iterativo, isto é, o novo modelo adaptado encontrado serve como modelo inicial para um outra iteração do algoritmo MLED e uma nova estimação dos coeficientes, continuando até que um critério de parada seja atingido. Aqui ressalta-se que o único parâmetro do sistema que é recalculado a cada iteração do algoritmo é o parâmetro $\gamma_m^{(s)}(t)$. De início esse parâmetro é calculado com base nos pesos, médias e variâncias das gaussianas do modelo independente de locutor, porém da segunda iteração em diante, as novas médias estimadas fazem parte do cálculo das novas probabilidades de ocupação $\gamma_m^{(s)}(t)$'s, mantendo-se constante os pesos e as variâncias.

Quanto às equações (5.2) e (5.6), elas diferem apenas na contribuição do "eigenvoice 0". Como toda a elaboração do algoritmo MLED é baseada na equação (5.6) [30], existem duas maneiras de compensação a fim de se utilizar a equação básica (5.2) na adaptação:

- Normalizar os vetores de parâmetros $\vec{o}_t$'s, subtraindo deles a parcela do "eigenvoice 0" ($\vec{e}_m^{(s)}(0)$) correspondente, antes de aplicá-los na equação (5.11). Isso é equivalente a se retirar a média do conjunto de dados, estimar os coeficientes dos "eigenvoices", achar a posição relativa do novo locutor nesse novo espaço normalizado, e depois somar a média obtendo o valor real da média adaptada.

- Descartar a equação (5.2) considerando somente a equação (5.6) na adaptação das médias, mas considerando que o "eigenvoice" de ordem 1 ($\vec{e}(1)$) é dado pelo "eigenvoice 0" ($\vec{e}(0)$). Essa substituição é geralmente feita [31] para que se possa manter a coerência da ordem de grandeza do valor da nova média adaptada, pois, como já falado anteriormente, os "eigenvoices" (autovetores) possuem módulos pequenos, da ordem da unidade. Portanto, a compensação deve vir de $\vec{e}(0)$, que é uma média.

Neste trabalho utilizou-se o primeiro recurso, porém alguns testes foram feitos mostrando que ambas as maneiras são equivalentes. Desta forma, os parâmetros acústicos são normalizados pelo "eigenvoice 0" nas equações (5.8) e (5.11), isto é, nessas equações, $\vec{o}_t$ e o_{tn} devem ser substituídos por $(\vec{o}_t - \vec{e}_m^{(s)}(0))$ e $(o_{tn} - e_{mn}^{(s)}(0))$ respectvamente.

Finalmente, após algumas iterações, tomam-se os coeficientes finais w's dos "eigenvoices", atualizam-se as médias (13500 em nosso caso) segundo a equação (5.2), obtendo-se assim o modelo adaptado para o novo locutor.

Capítulo 6

Rotina de simulação

A linguagem de programação utilizada para a elaboração da rotina de simulação do algoritmo de adaptação de locutor foi C++. Procurou-se escrever a rotina usando o conceito de linguagem orientada a objetos, a fim de que a mesma pudesse se tornar o mais modular possível.

O primeiro passo a se tomar foi a elaboração de funções que fossem responsáveis pela extração dos parâmetros dos modelos dependente e independente de locutor. Tais modelos são armazenados em arquivos *.hmm*, que são os arquivos de saída do programa de treinamento de HMM's elaborado pelo prof. Dr. Carlos Alberto Ynoguti [7]. Depois de analisar a forma como o arquivo *.hmm* era gerado, elaboraram-se as funções de leitura do arquivo, e funções para a extração das médias das gaussianas, variâncias das gaussianas (extraídas da diagonal da matriz de covariância das gaussianas, pois as componentes dos parâmetros são consideradas independentes entre si), pesos das gaussianas e matrizes de transições entre estados para cada fone. É importante ressaltar a forma como esses parâmetros extraídos são armazenados. Verificou-se que, no cálculo dos coeficientes dos "eigenvoices", para cada estado de cada fone, as gaussianas deviam ser ordenadas de acordo com o seu peso para assegurar uma correspondência entre as gaussianas das misturas de estados equivalentes de modelos de locutores diferentes. Por isso, realizou-se uma modificação na forma com que os arquivos *.hmm* foram gerados (no programa de treinamento de HMM's) a fim de que os arquivos já fossem gravados com as gaussianas ordenadas de acordo com os seus pesos.

O segundo passo foi a criação de funções que pudessem montar os supervetores tendo como entrada as

médias de todas as gaussianas de todos os estados de cada HMM dos modelos dependente de locutor, e calcular o "eigenvoice 0" $\vec{e}(0)$, isto é, o supervetor médio dentre todos os locutores de referência. O terceiro passo foi a criação de uma função que fizessem o cálculo da SVD levando em conta todos os modelos dependente de locutor. Depois de calculados os "eigenvoices", as tarefas "off-line" já estão prontas.

O quarto passo foi obter os parâmetros de fala da locução de adaptação do novo locutor $\vec{o}_t$. Para isso, fez-se uso de funções que constavam no programa de treinamento de HMM's utilizado. Depois de calculados os parâmetros acústicos, verificou-se que para uma maior facilidade no cálculo de $\gamma_m^{(s)}(t)$, era interessante optar por uma segmentação da locução de adaptação via algoritmo de Viterbi. Elaborou-se então uma função de segmentação via Viterbi (baseada na função de segmentação via Viterbi do programa de treinamento de HMM's) que resultava na separação dos vetores de parâmetros acústicos $\vec{o}_t$ em arquivos correspondentes aos estados do HMM da locução de adaptação. Com os vetores de parâmetros acústicos $\vec{o}_t$ separados por estados via algoritmo de Viterbi, o cálculo da probabilidade $\gamma_m^{(s)}(t)$ ficaria mais rápido e fácil. Outra forma do cálculo de $\gamma_m^{(s)}(t)$ é segundo as variáveis "forward" e "backward" [3]. Segundo [30] [31], o cálculo de $\gamma_m^{(s)}(t)$ por ambas as formas não faz diferença.

No quinto passo, uma função para o cálculo da probabilidade $\gamma_m^{(s)}(t)$ foi feita. Para esse cálculo também foram utilizadas certas funções do programa de treinamento de HMM's. Essas funções calculam a probabilidade de uma certa gaussiana multidimensional emitir certa saída. Com os parâmetros acústicos $\vec{o}_t$ separados por estados, devido à segmentação, não há a necessidade de se calcular a probabilidade de uma gaussiana dentro de um certo estado emitir uma observação $\vec{o}_t$ que consta em um outro estado, visto que ela é zero. A função de cálculo de $\gamma_m^{(s)}(t)$ tem como entrada o conjunto de parâmetros acústicos $\vec{o}_t$, já separados por estados do HMM da locução, e os parâmetros das gaussianas dos estados que compõem o HMM da locução (pesos, médias e variâncias).

Depois de elaboradas as funções para calcular os "eigenvoices", obter os supervetores, obter as médias, variâncias das gaussianas e as matrizes de transições de estados do modelo independente de locutor, e calcular os parâmetros acústicos da locução de adaptação, deu-se início à elaboração do algoritmo de adaptação propriamente dito, baseado na equação (5.9). Para a resolução do sistema de equações (5.12), criou-se uma rotina com base no algoritmo de eliminação gaussiana, a mesma técnica de resolução de sistemas de equações

empregada em [30].

Com a primeira estimação dos coeficientes dos "eigenvoices", uma parte da rotina faz as substituições dos coeficientes usando a equação (5.2) (ou se preferir, a equação (5.13)), obtendo assim as primeiras estimações das médias adaptadas.

Após um número de iterações, os últimos parâmetros adaptados são salvos e armazenados em um arquivo *.hmm* através de uma função também elaborada. O novo arquivo *.hmm* criado tem a mesma estrutura do arquivo *.hmm* citado anteriormente. A mesma estrutura do arquivo possibilita a utilização deste último para a verificação do processo de adaptação em uma simulação de reconhecimento de frases, feita pelo programa de reconhecimento de fala também elaborado pelo prof. Dr. Carlos Alberto Ynoguti [7]. Como os únicos parâmetros adaptados são as médias das gaussianas, então as variâncias das gaussianas, os pesos das gaussianas e as matrizes de transições de estado que são armazenados no arquivo *.hmm* do novo locutor são os mesmos do arquivo *.hmm* do modelo independente de locutor.

Para apenas uma locução de adaptação, calcula-se a equação (5.9) referente a essa locução e em seguida estimam-se os coeficientes dos "eigenvoices". Para mais de uma locução de adaptação, antes de se obter a equação (5.9), os segundos membros das equações (5.10) e (5.11) ganham um somatório externo com respeito ao número de locuções de adaptação. Após o processamento de todas as locuções, calcula-se a equação (5.9) a partir das equações (5.10) e (5.11) acumuladas. Há pois dois laços dentro da rotina: um laço de iterações e, dentro desse, um laço de locuções de adaptação. Depois de formulada a equação (5.9) a partir dos membros $\delta(k,j)$ e $\Lambda(j)$ acumulados, os coeficientes dos "eigenvoices" são estimados. A figura 6.1 mostra um esquema do processo de adaptação de locutor.

O microcomputador usado na elaboração de toda a rotina foi um Pentium IV (2,4GHz com 512MB de memória RAM) que foi adquirido com os recursos da reserva técnica da bolsa de Mestrado FAPESP.

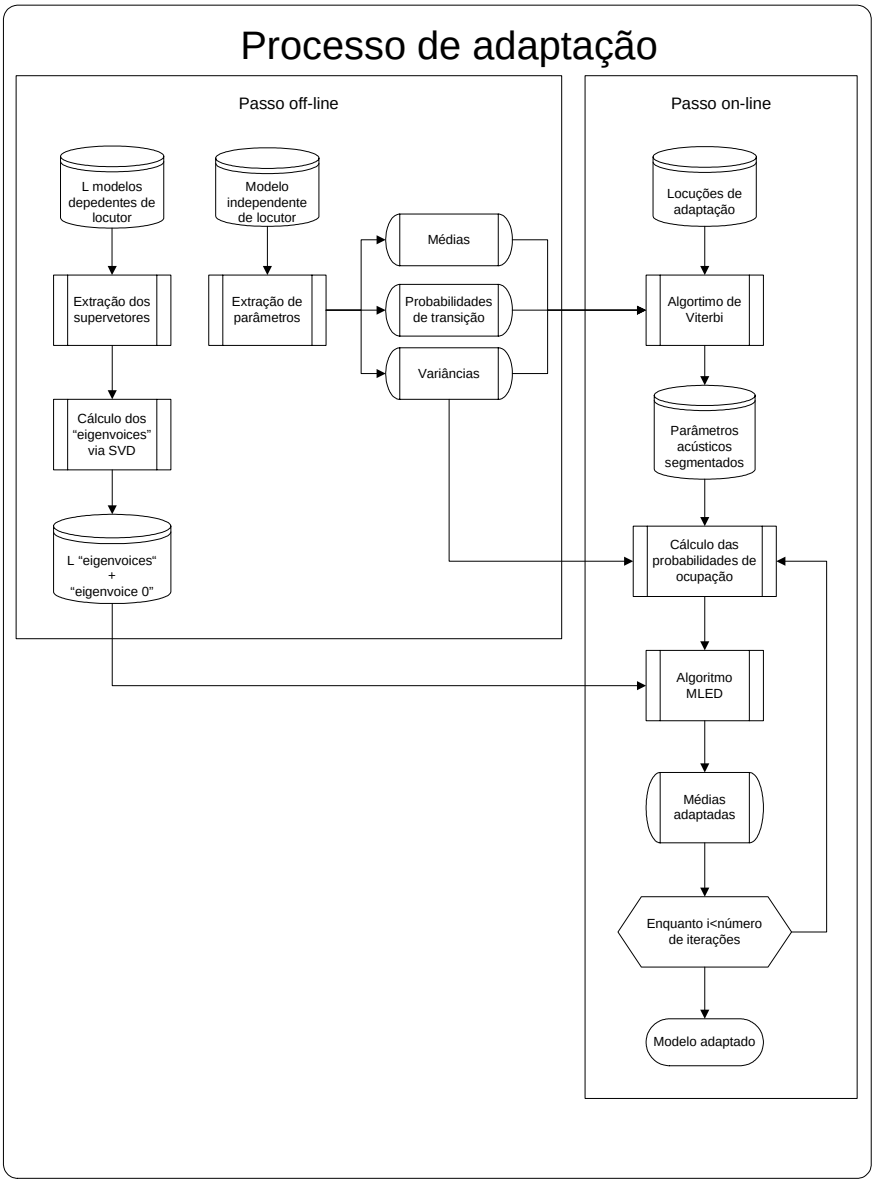

Figura 6.1: Esquema do processo de adaptação de locutor.

Capítulo 7

Resultados e discussões

Foram treinados 18 modelos dependente de locutor (locutores masculinos) para compor a base de locutores. Dos 18 modelos dependente dos locutores base, têm-se 18 "eigenvoices" mais o "eigenvoice 0". Foi decidido não realizar nenhum descarte de "eigenvoices" de alta ordem (autovalores de baixa magnitude) devido ao número reduzido de locutores base, fazendo com que $K = L = 18$.

De uma outra base de dados, elaborada pelo prof. Carlos Alberto Ynoguti (ver seção 3.2.3), foram tomados 10 locutores, onde cada locutor pronunciou 40 locuções. Foi estabelecido um conjunto de adaptação, formado por 10 locuções de cada um dos 10 locutores, e um conjunto de teste, formado pelas restantes 30 locuções de cada locutor. Os conjuntos de locuções são vistos no apêndice B.

Foram realizados dois tipos de testes:

- **Testes com múltiplas locuções de adaptação**: Para cada locutor de teste foram feitos vários testes de adaptação onde, para cada teste, o número de locuções de adaptação foi incrementado de 1 até 10. Isto é, para o primeiro teste usou-se uma locução de adaptação, para o segundo teste usou-se duas locuções de adaptação, e assim por diante até se atingir o número de 10 locuções de adaptação de entrada.

- **Testes com locuções de adaptação separadas**: Para dois locutores de teste foram realizados outros testes de adaptação, onde para cada teste foi usado apenas uma locução de adaptação específica. Isto

é, no primeiro teste usou-se uma locução de adaptação, no segundo teste outra locução de adaptação, e assim por diante até se utilizar todas as 10 locuções de adaptação do conjunto de teste.

Para cada um dos testes acima especificados, foi variado o número de iterações do algoritmo de adaptação de 1 até 5. O intuito dos testes é verificar a influência do número de locuções de adaptação e o número de iterações no processo de adaptação de locutor. O tempo de processamento para cada locução de adaptação foi, em média, *0,9 segundos*.

7.1 Testes com múltiplas locuções de adaptação

Os resultados são mostrados nas figuras 7.1 a 7.10. O eixo vertical indica a taxa de acerto de palavras, o eixo horizontal representa o número de locuções de adaptação de entrada e cada curva indica uma iteração do processo.

Pode ser visto que não há um padrão de comportamento único para todas as curvas. Para alguns locutores de teste, considerando os resultados de todas as iterações a menos da primeira, os valores de taxa de acerto de palavras se concentram dentro de uma certa faixa à medida que se aumenta o número de locuções de adaptação (figuras 7.1, 7.2, 7.4, 7.5, 7.7, 7.8 e 7.9), chegando a diferenças bem pequenas entre as curvas de iteração para alguns desses locutores (figuras 7.7 e 7.8). Para outros locutores, considerando todas os resultados de todas as iterações e aumentando o número de locuções de adaptação, as taxas de acerto de palavras não apresentaram um padrão (figuras 7.3, 7.6 e 7.10), chegando a mostrar um comportamento aleatório (figura 7.6).

Do exposto, a variação do número de iterações não serviu para estimar um número fixo de iterações para o processo de adaptação. De uma maneira geral, três a cinco iterações se mostrou um número razoável. Pode-se observar também que nem sempre o aumento do número de locuções de adaptação implica em um aumento da taxa de acerto de palavras. Nas figuras 7.1 e 7.4, os melhores resultados se deram com apenas 2 locuções de adaptação; nas figuras 7.8, 7.9 e 7.10, e 7.2, 7.6 e 7.7, 4 e 5 locuções de adaptação, respectivamente para cada grupo, são responsáveis pelos maiores desempenhos. Já nas figuras 7.3 e 7.5, os melhores resultados se deram quando são utilizadas 8 e 9 locuções de adaptação respectivamente.

Os testes mostraram que nem todos os locutores conseguiram obter modelos adaptados com desempenho superior ao modelo independente de locutor, usando o mesmo material de teste. Porém, para os locutores cujo o modelo adaptado superou o modelo independente de locutor, bons resultados foram atingidos, chegando a 5,92% de melhoria (figura 7.1).

7.2 Testes com locuções de adaptação separadas

Os resultados são mostrados nas figuras 7.11 e 7.12. O eixo vertical indica a taxa de acerto de palavras e o eixo horizontal representa o número da locução de adaptação usada. Para simplificação, as figuras mostram apenas os resultados para duas e três iterações correspondentes às figuras 7.11 e 7.12 respectivamente.

Pelas figuras 7.11 e 7.12, vê-se que a desempenho do modelo adaptado superou o desempenho do modelo independente de locutor quando se usou uma locução específica de adaptação para cada locutor. Usando a locução de adaptação "07" para o locutor M15, obteve-se uma melhoria de 1% em taxa de acerto de palavras, e usando a locução "02" para o locutor M17 obteve-se uma melhoria de 1,39%. Vale ressaltar que esses locutores não apresentaram nenhuma melhoria na adaptação ao se aplicar até 10 locuções de adaptação nos testes anteriores (figuras 7.8 e 7.9).

Para uma melhor conferência, os valores de taxa de acerto de palavras dos testes de adaptação, correspondentes às figuras 7.1 à 7.12, estão organizados em tabelas no apêndice A. Em cada tabela, os valores de taxa de acerto de palavras (TAP) provindos dos testes de adaptação, que são superiores à TAP do modelo independente de locutor, estão em negrito.

7.3 Testes extras

Em um novo experimento, foram tomadas as duas melhores locuções de adaptação para os locutores M15 e M17, com base nos testes com locuções de adaptação separadas: locuções "04" e "07" para o locutor M15, e locuções "02" e "04" para o locutor M17 (ver figuras 7.11 e 7.12); e foram realizados os seguintes testes:

- Usaram-se as duas locuções de adaptação, respectivas de cada locutor, usando a configuração de teste com múltiplas locuções de adaptação;

- Concatenaram-se as duas locuções de adaptação, respectivas de cada locutor, em uma só locução e depois realizou-se a adaptação com essa locução concatenada.

Para esses novos testes, os resultados obtidos mostraram que a adaptação usando as duas melhores locuções de adaptação dos locutores M15 e M17 separadamente (teste com múltiplas locuções de adaptação) se mostrou melhor que a utilização de uma única locução de adaptação formada pela concatenação de duas. Isso sugere o uso de frases curtas ou mesmo de palavras isoladas no material de adaptação. Também foi observado que, para o locutor M17, a união das duas melhores locuções de adaptação (locuções "02" e "04" obtidas pela análise da figura 7.12) gerou um modelo adaptado cujo desempenho foi superior aos dos modelos gerados por ambas as locuções em separado, e conseqüentemente superior ao modelo independente de locutor, atingido 1,86% de melhoria (em relação ao modelo independente de locutor).

Outro resultado interessante se deu quando, nas fases iniciais de testes, o procedimento de adaptação era realizado da seguinte maneira, ao se utilizar várias locuções de adaptação: realizavam-se as iterações do algoritmo MLED utilizando a primeira locução de adaptação resultando em um modelo adaptado inicial. Esse modelo adaptado servia de modelo inicial para o processo de adaptação da segunda locução que, depois de algumas iterações, resultava em um novo modelo adaptado que servia de modelo inicial para a terceira locução de adaptação, e assim por diante. Esse procedimento depois foi substituído pelo atual, já explicado no capítulo 6. Com esse procedimento antigo, o locutor M03, que pela figura 7.2 não conseguiu bons resultados de adaptação com o procedimento adotado, chegou a atingir uma pequena melhoria de 0,49% com relação ao modelo independente de locutor, quando se usou 5 locuções de adaptação perfazendo 3 iterações.

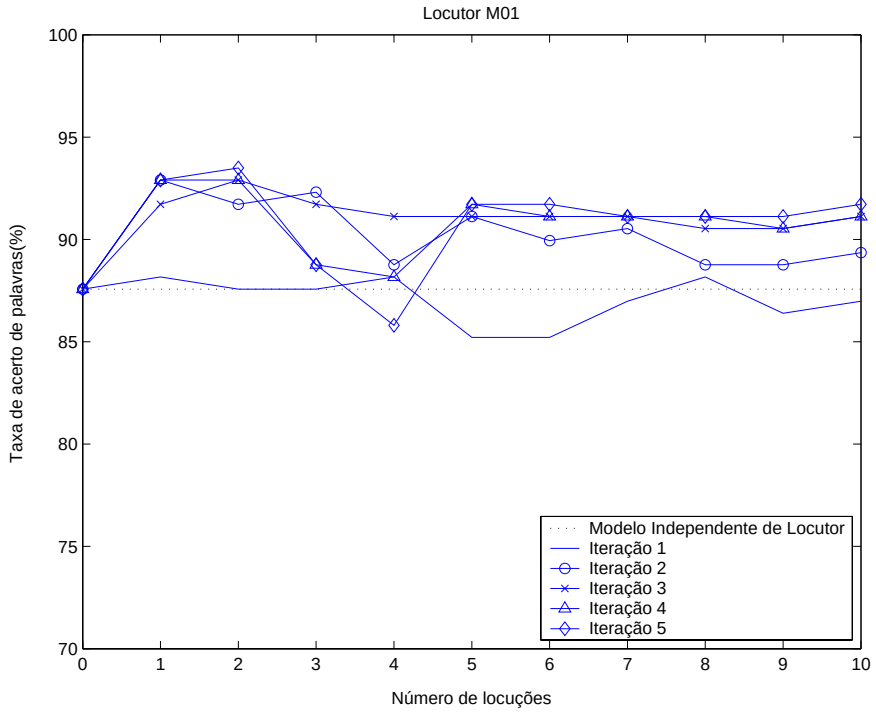

Figura 7.1: Taxa de acerto de palavras para o locutor M01 com número de locuções de adaptação variando entre 1 e 10.

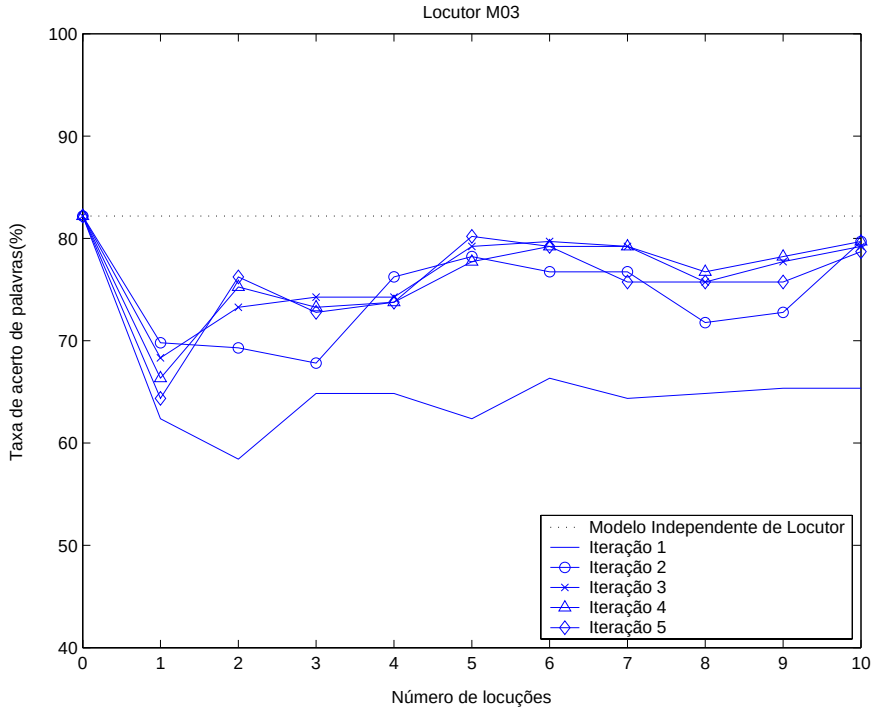

Figura 7.2: Taxa de acerto de palavras para o locutor M03 com número de locuções de adaptação variando entre 1 e 10.

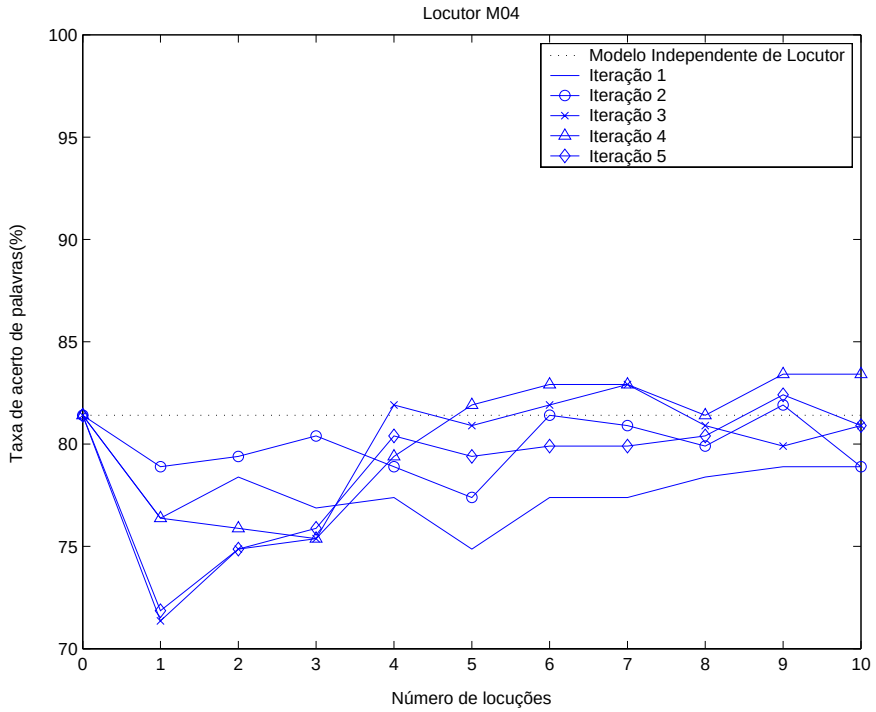

Figura 7.3: Taxa de acerto de palavras para o locutor M04 com número de locuções de adaptação variando entre 1 e 10.

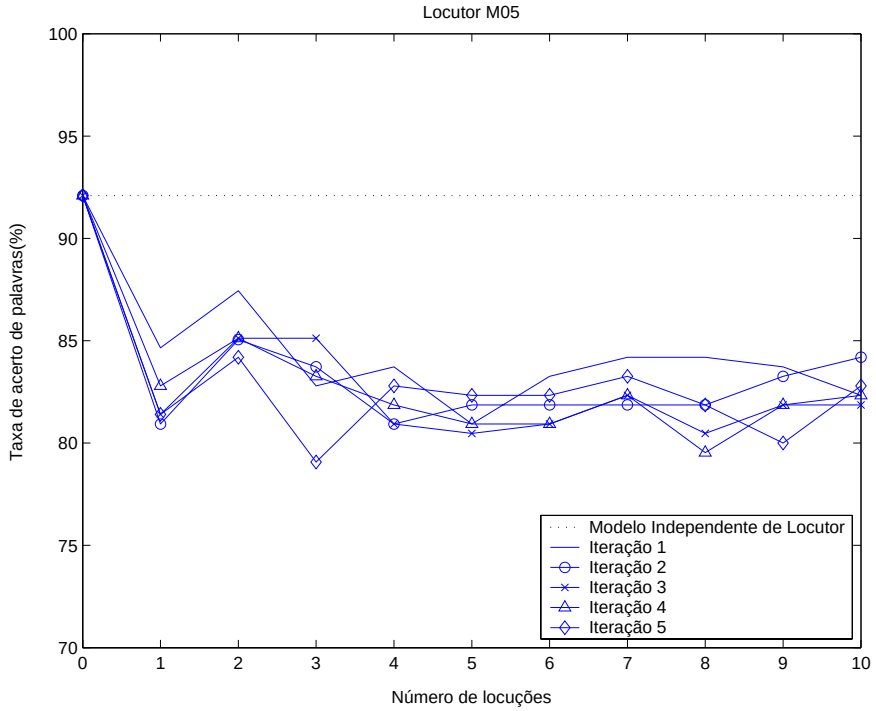

Figura 7.4: Taxa de acerto de palavras para o locutor M05 com número de locuções de adaptação variando entre 1 e 10.

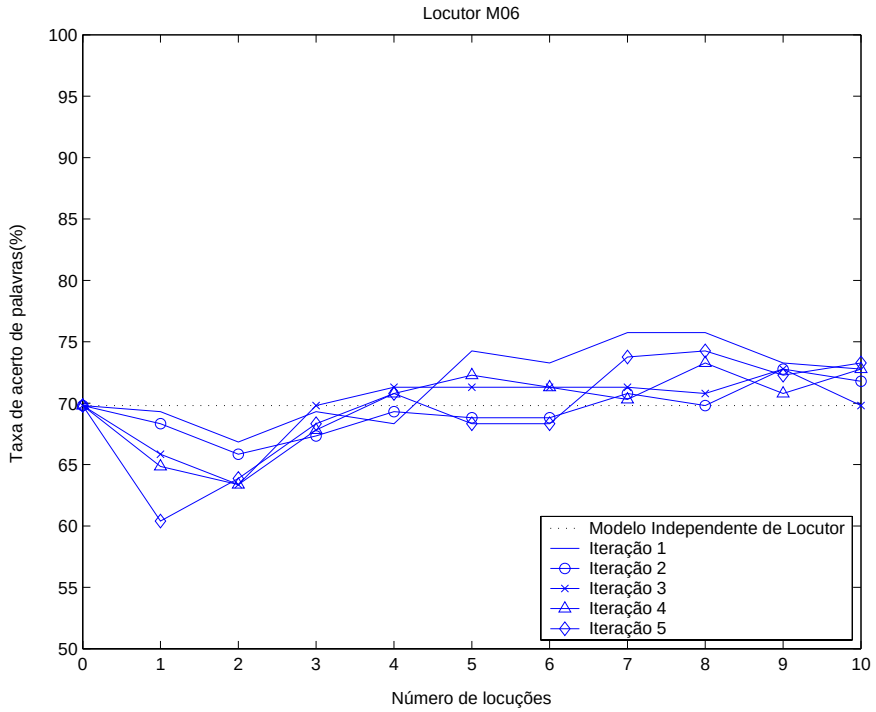

Figura 7.5: Taxa de acerto de palavras para o locutor M06 com número de locuções de adaptação variando entre 1 e 10.

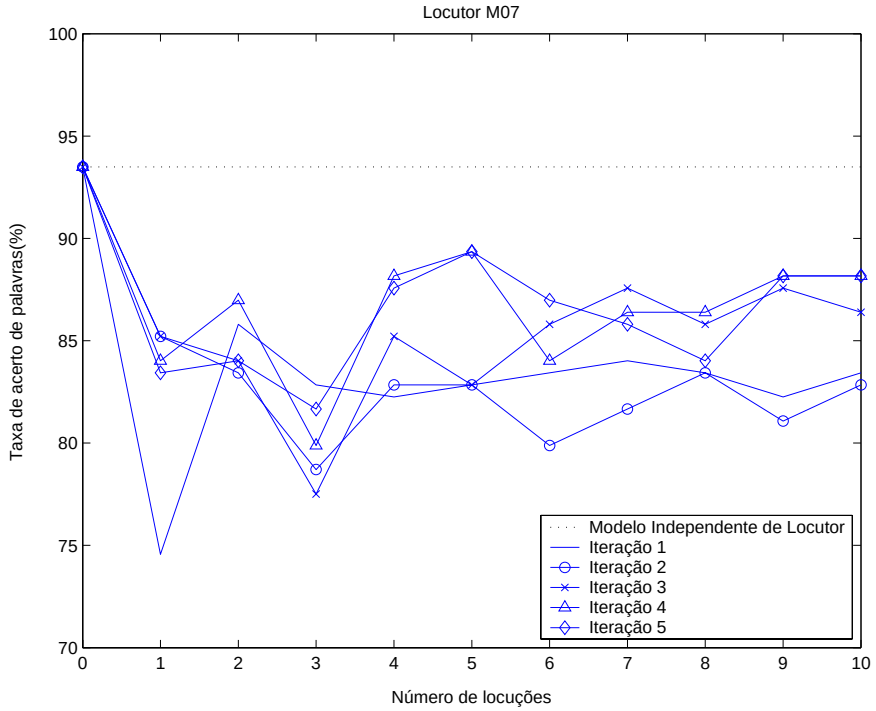

Figura 7.6: Taxa de acerto de palavras para o locutor M07 com número de locuções de adaptação variando entre 1 e 10.

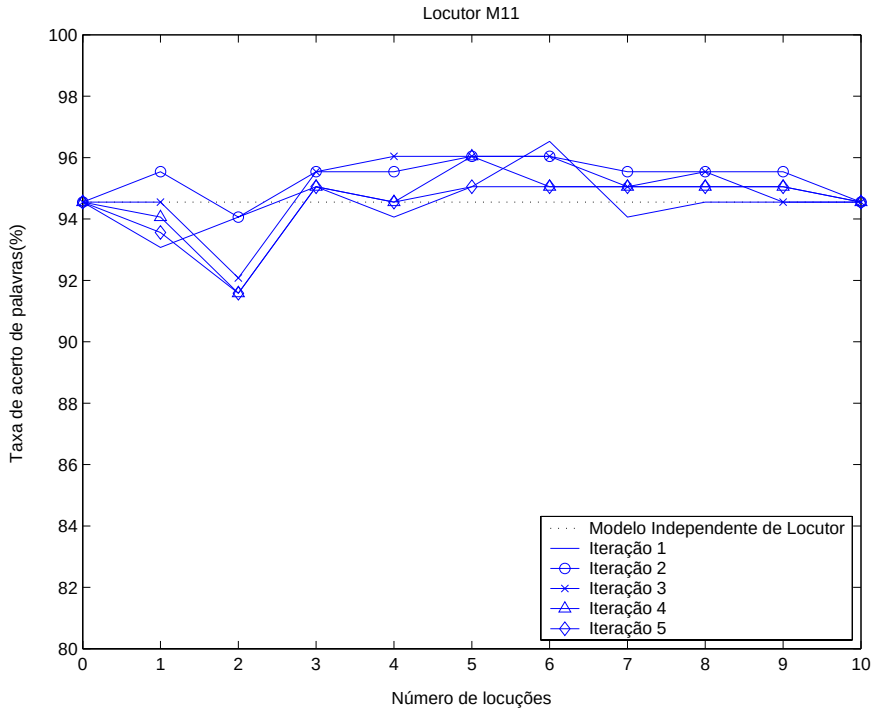

Figura 7.7: Taxa de acerto de palavras para o locutor M11 com número de locuções de adaptação variando entre 1 e 10.

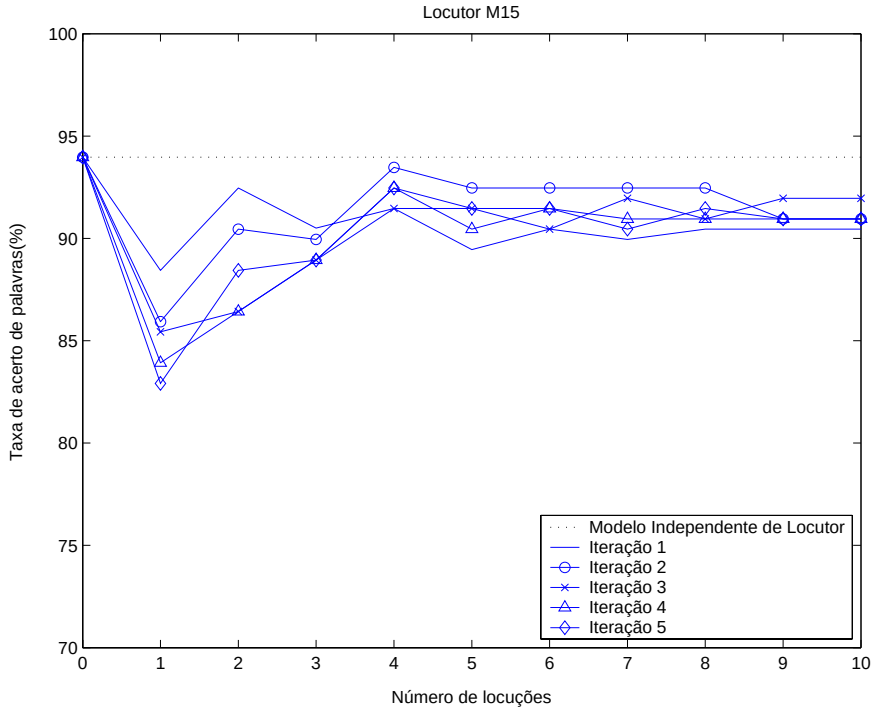

Figura 7.8: Taxa de acerto de palavras para o locutor M15 com número de locuções de adaptação variando entre 1 e 10.

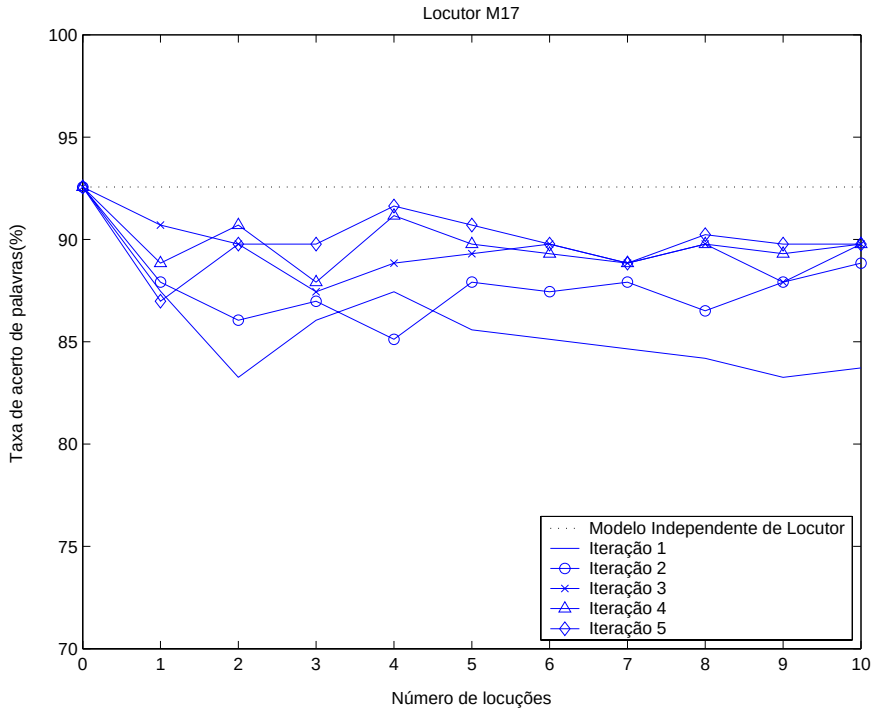

Figura 7.9: Taxa de acerto de palavras para o locutor M17 com número de locuções de adaptação variando entre 1 e 10.

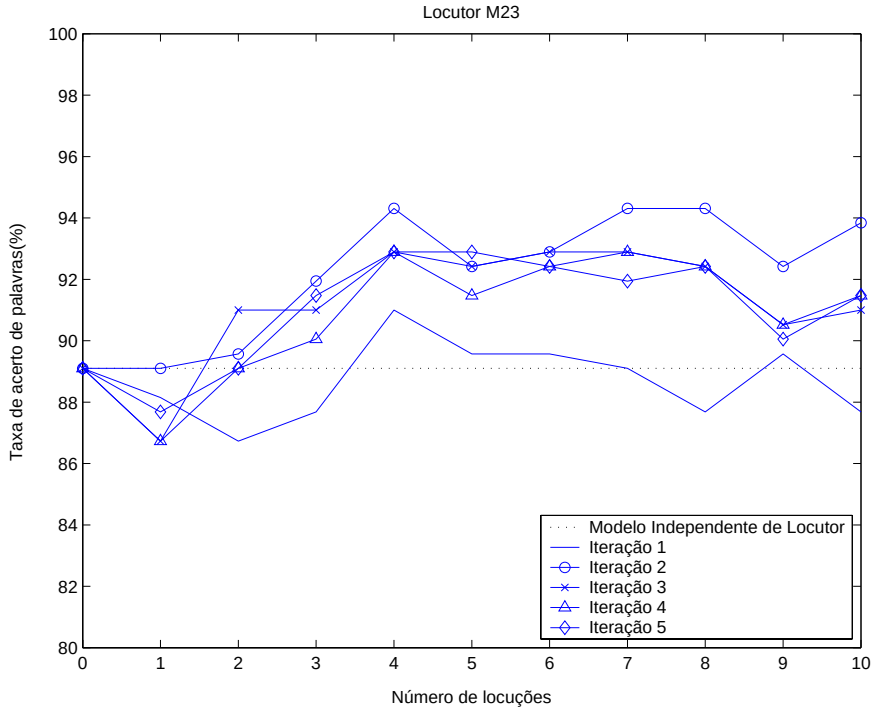

Figura 7.10: Taxa de acerto de palavras para o locutor M23 com número de locuções de adaptação variando entre 1 e 10.

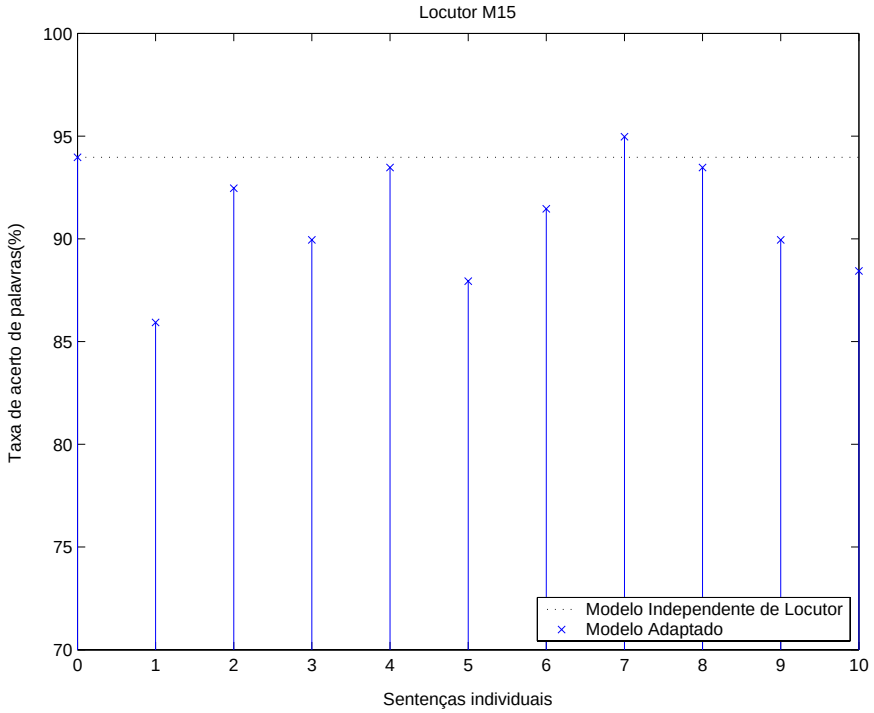

Figura 7.11: Taxa de acerto de palavras para o locutor M15 com 10 locuções de adaptação distintas e isoladas.

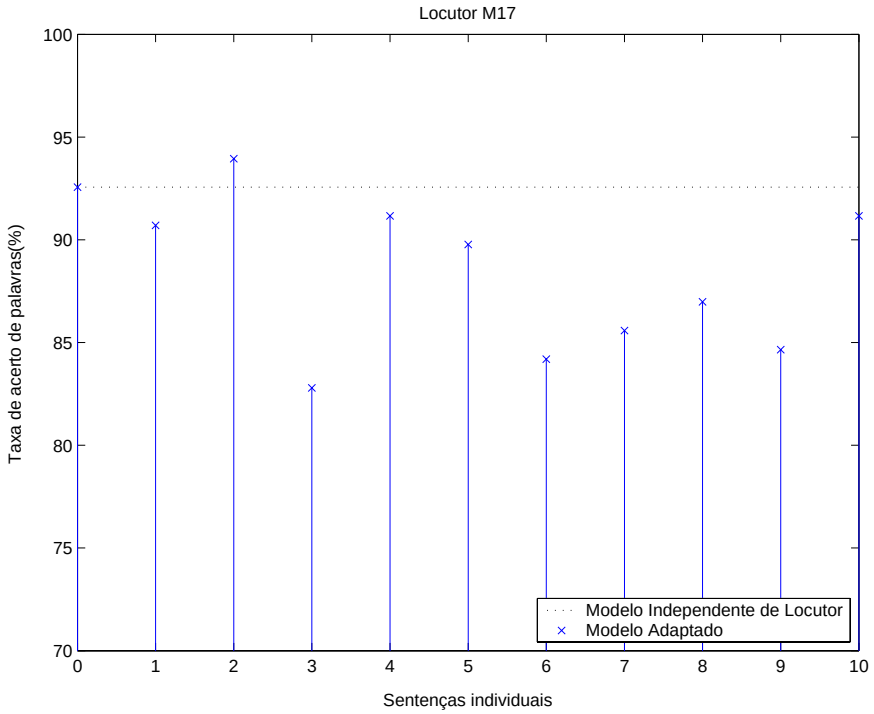

Figura 7.12: Taxa de acerto de palavras para o locutor M17 com 10 locuções de adaptação distintas e isoladas.

Capítulo 8

Conclusão

Do exposto no capítulo 7, é visto que para uma específica metodologia de adaptação (seção 7.1), apenas metade dos locutores de teste tiveram seus respectivos modelos adaptados com desempenho melhor que utilizando modelos independente de locutor. Porém, com a utilização de apenas uma locução de adaptação separadamente (seção 7.2), locutores que antes tiveram modelos adaptados com desempenhos inferiores aos desempenhos do modelo independente correspondente com o mesmo material de teste, agora possuem desempenho superior.

Em outras técnicas como MAP e MLLR, há uma proporção direta entre desempenho e quantidade de material de adaptação, porém, com o uso dos "eigenvoices", essa proporção não é válida, visto que uma limitação do procedimento se refere à dimensão do espaço de locutores, ou seja, quanto maior o número de locutores para compor o espaço de locutores, melhor representado estará o novo locutor. Conclui-se que, para a utilização da técnica via "eigenvoices", é interessante possuir uma certa quantidade de modelos dependente de locutor para a formação de um espaço de locutores robusto, implicando em uma melhor representação do novo locutor; e que a quantidade de material de fala utilizada para a adaptação não é importante, mas sim a qualidade, ou seja, não importa o quanto mas quais as locuções de adaptação que melhor adaptam os modelos, o que faz com que essa técnica (via "eigenvoices") seja aplicada em adaptações rápidas de locutor (pouco material de adaptação).

Outra conclusão é que ainda não há um padrão para a adaptação de locutor via "eigenvoices" utilizando os

procedimentos adotados neste trabalho de Mestrado, o que indica que o estudo não se deu por terminado, mas que serve como ponto de partida para um maior aprofundamento na área. Em suma, mesmo com algumas carências de material de treinamento, o trabalho mostrou que a técnica apresenta resultados satisfatórios e gerou discussões para a realização de mais trabalhos na área.

Algumas sugestões para novos trabalhos são:

- O aumento do espaço de locutores com a geração de mais modelos dependente de locutor;
- A adaptação aplicando coeficientes de "eigenvoices" específicos para grupos de fones e não para todos os fones;
- Um estudo do material de adaptação a fim de estabelecer um padrão para a utilização em adaptação de locutor.

Aqui é importante ressaltar o incentivo à geração de uma base de dados de fala para pesquisa em reconhecimento de fala contínua no Brasil. Neste trabalho, durante a fase de segmentação das locuções de adaptação para o cálculo dos termos $\gamma_m^{(s)}(t)$, foi utilizado o algoritmo de Viterbi (capítulo 6). O algoritmo de Viterbi não é uma boa ferramenta para segmentação de longas locuções e, por isso, a disponibilidade de uma base de dados segmentada contribuiria tanto para uma melhoria dos resultados quanto para o treinamento de mais modelos dependente de locutor.

Apêndice A

Resultados dos testes de adaptação

A.1 Testes com múltiplas locuções de adaptação

Locutor M01					
TAP do locutor para o modelo independente de locutor: 87,57%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	**88,17%**	**92,90%**	**91,72%**	**92,90%**	**92,90%**
02	**87,57%**	**91,72%**	**92,90%**	**92,90%**	**93,49%**
03	**87,57%**	**92,31%**	**91,72%**	**88,76%**	**88,76%**
04	**88,17%**	**88,76%**	**91,12%**	**88,17%**	85,80%
05	85,21%	**91,12%**	**91,12%**	**91,72%**	**91,72%**
06	85,21%	**89,94%**	**91,12%**	**91,12%**	**91,72%**
07	86,98%	**90,53%**	**91,12%**	**91,12%**	**91,12%**
08	**88,17%**	**88,76%**	**90,53%**	**91,12%**	**91,12%**
09	86,39%	**88,76%**	**90,53%**	**90,53%**	**91,12%**
10	86,98%	**89,35%**	**91,12%**	**91,12%**	**91,72%**

Tabela A.1: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M01.

Locutor M03					
TAP do locutor para o modelo independente de locutor: 82,18%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	62,38%	69,80%	68,32%	66,34%	64,36%
02	58,42%	69,31%	73,27%	75,25%	76,24%
03	64,85%	67,82%	74,26%	73,27%	72,77%
04	64,85%	76,24%	74,26%	73,76%	73,76%
05	62,38%	78,22%	79,21%	77,72%	80,20%
06	66,34%	76,73%	79,70%	79,21%	79,21%
07	64,36%	76,73%	79,21%	79,21%	75,74%
08	64,85%	71,78%	75,74%	76,73%	75,74%
09	65,35%	72,77%	77,72%	78,22%	75,74%
10	65,35%	79,70%	79,21%	79,70%	78,71%

Tabela A.2: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M03.

Locutor M04					
TAP do locutor para o modelo independente de locutor: 81,41%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	76,38%	78,89%	71,36%	76,38%	71,86%
02	78,39%	79,40%	74,87%	75,88%	74,87%
03	76,88%	80,40%	75,38%	75,38%	75,88%
04	77,39%	78,89%	**81,91%**	79,40%	80,40%
05	74,87%	77,39%	80,90%	**81,91%**	79,40%
06	77,39%	**81,41%**	**81,91%**	**82,91%**	79,90%
07	77,39%	80,90%	**82,91%**	**82,91%**	79,90%
08	78,39%	79,90%	80,90%	**81,41%**	80,40%
09	78,89%	**81,91%**	79,90%	**83,42%**	**82,41%**
10	78,89%	78,89%	80,90%	**83,42%**	80,90%

Tabela A.3: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M04.

Locutor M05					
TAP do locutor para o modelo independente de locutor: 92,09%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	84,65%	80,93%	81,40%	82,79%	81,40%
02	87,44%	85,05%	85,12%	85,12%	84,19%
03	82,79%	83,72%	85,12%	83,26%	79,07%
04	83,72%	80,93%	80,93%	81,86%	82,79%
05	80,93%	81,86%	80,47%	80,93%	82,33%
06	83,26%	81,86%	80,93%	80,93%	82,33%
07	84,19%	81,86%	82,33%	82,33%	83,26%
08	84,19%	81,86%	80,47%	79,53%	81,86%
09	83,72%	83,26%	81,86%	81,86%	80,00%
10	82,33%	84,19%	81,86%	82,33%	82,79%

Tabela A.4: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M05.

Locutor M06					
TAP do locutor para o modelo independente de locutor: 69,80%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	69,31%	68,32%	65,84%	64,85%	60,40%
02	66,83%	65,84%	63,37%	63,37%	63,86%
03	69,31%	67,33%	**69,80%**	67,82%	68,32%
04	68,32%	69,31%	**71,29%**	**70,79%**	**70,79%**
05	**74,26%**	68,81%	**71,29%**	**72,28%**	68,32%
06	**73,27%**	68,81%	**71,29%**	**71,29%**	68,32%
07	**75,74%**	**70,79%**	**71,29%**	**70,30%**	**73,76%**
08	**75,74%**	**69,80%**	**70,79%**	**73,27%**	**74,26%**
09	**73,27%**	**72,77%**	**72,77%**	**70,79%**	**72,28%**
10	**72,77%**	**71,78%**	**69,80%**	**72,77%**	**73,27%**

Tabela A.5: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M06.

Locutor M07					
TAP do locutor para o modelo independente de locutor: 93,49%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	74,56%	85,21%	85,21%	84,02%	83,43%
02	85,80%	83,43%	84,02%	86,98%	84,02%
03	82,84%	78,70%	77,51%	79,88%	81,66%
04	82,25%	82,84%	85,21%	88,17%	87,57%
05	82,84%	82,84%	82,84%	89,35%	89,35%
06	83,43%	79,88%	85,80%	84,02%	86,98%
07	84,02%	81,66%	87,57%	86,39%	85,80%
08	83,43%	83,43%	85,80%	86,39%	84,02%
09	82,25%	81,07%	87,57%	88,17%	88,17%
10	83,43%	82,84%	86,39%	88,17%	88,17%

Tabela A.6: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M07.

Locutor M11					
TAP do locutor para o modelo independente de locutor: 94,55%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	93,07%	**95,54%**	**94,55%**	94,06%	93,56%
02	94,06%	94,06%	92,08%	91,58%	91,58%
03	**95,05%**	**95,54%**	**95,54%**	**95,05%**	**95,05%**
04	94,06%	**95,54%**	**96,04%**	**94,55%**	**94,55%**
05	**95,05%**	**96,04%**	**96,04%**	**96,04%**	**95,05%**
06	**96,53%**	**96,04%**	**96,04%**	**95,05%**	**95,05%**
07	94,06%	**95,54%**	**95,05%**	**95,05%**	**95,05%**
08	**94,55%**	**95,54%**	**95,54%**	**95,05%**	**95,05%**
09	**94,55%**	**95,54%**	**94,55%**	**95,05%**	**95,05%**
10	**94,55%**	**94,55%**	**94,55%**	**94,55%**	**94,55%**

Tabela A.7: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M11.

Locutor M15					
TAP do locutor para o modelo independente de locutor: 93,97%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	88,44%	85,93%	85,43%	83,92%	82,91%
02	92,46%	90,45%	86,43%	86,43%	88,44%
03	90,50%	89,95%	88,94%	88,94%	88,94%
04	91,46%	93,47%	91,46%	92,46%	92,46%
05	89,45%	92,46%	91,46%	90,45%	91,46%
06	90,45%	92,46%	90,45%	91,46%	91,46%
07	89,95%	92,46%	91,96%	90,95%	90,45%
08	90,45%	92,46%	90,95%	90,95%	91,46%
09	90,45%	90,95%	91,96%	90,95%	90,95%
10	90,45%	90,95%	91,96%	90,95%	90,95%

Tabela A.8: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M15.

Locutor M17					
TAP do locutor para o modelo independente de locutor: 92,56%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	87,47%	87,91%	90,70%	88,84%	86,98%
02	83,26%	86,05%	89,77%	90,70%	89,77%
03	86,05%	86,98%	87,44%	87,91%	89,77%
04	87,44%	85,12%	88,84%	91,16%	91,63%
05	85,58%	87,91%	89,30%	89,77%	90,70%
06	85,12%	87,44%	89,77%	89,30%	89,77%
07	84,65%	87,91%	88,84%	88,84%	88,84%
08	84,19%	86,51%	89,77%	89,77%	90,23%
09	83,26%	87,91%	87,91%	89,30%	89,77%
10	83,72%	88,84%	89,77%	89,77%	89,77%

Tabela A.9: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M17.

Locutor M23					
TAP do locutor para o modelo independente de locutor: 89,10%					
Locução	Iteração 01	Iteração 02	Iteração 03	Iteração 04	Iteração 05
01	88,15%	**89,10%**	86,73%	86,73%	87,68%
02	86,73%	**89,57%**	**91,00%**	**89,10%**	**89,10%**
03	87,68%	**91,94%**	**91,00%**	**90,05%**	**91,47%**
04	**91,00%**	**94,31%**	**92,89%**	**92,89%**	**92,89%**
05	**89,57%**	**92,42%**	**92,42%**	**91,47%**	**92,89%**
06	**89,57%**	**92,89%**	**92,89%**	**92,42%**	**92,42%**
07	**89,10%**	**94,31%**	**92,89%**	**92,89%**	**91,94%**
08	87,68%	**94,31%**	**92,42%**	**92,42%**	**92,42%**
09	**89,57%**	**92,42%**	**90,52%**	**90,52%**	**90,05%**
10	87,68%	**93,84%**	**91,00%**	**91,47%**	**91,47%**

Tabela A.10: Taxas de acerto de palavras dos testes de adaptação com múltiplas locuções de adaptação para o locutor M23.

A.2 Testes com locuções de adaptação separadas

Locutor M15		**Locutor M17**	
TAP para o modelo independente de locutor: 93,97%		TAP para o modelo independente de locutor: 92,56%	
Locução	Iteração 02	Locução	Iteração 03
01	85,93%	01	90,70%
02	92,46%	02	**93,95%**
03	89,95%	03	82,79%
04	93,47%	04	91,16%
05	87,94%	05	89,77%
06	91,46%	06	84,19%
07	**94,97%**	07	85,58%
08	93,47%	08	86,98%
09	89,95%	09	84,65%
10	88,44%	10	91,16%

Tabela A.11: Taxas de acerto de palavras dos testes de adaptação com locuções de adaptação separadas para os locutor M15 e M17.

Apêndice B

Tabelas de locuções

Relação de locuções		
Locutor	Locuções de adaptação	Locuções de teste
M01	A questão foi retomada no congresso	Nosso telefone quebrou
M07	Leila tem um lindo jardim	Desculpe se magoei o velho
	O analfabetismo é a vergonha do pais	Queremos discutir o orçamento
	A casa foi vendida sem pressa	Ela tem muita fome
	Trabalhando com união rende muito mais	Uma índia andava na mata
	Recebi nosso amigo para almoçar	Zé, vá mais rápido
	A justiça é a única vencedora	Hoje dormirei bem
	Isso se resolverá de forma tranquila	João deu pouco dinheiro
	Os pesquisadores acreditam nessa teoria	Ainda são seis horas
	Sei que atingiremos o objetivo	Ela saía discretamente
		Eu vi logo a Ioiô e o Léo
		Um homem não caminha sem um fim
		Vi zé fazer essas viagens seis vezes
		O atabaque do Tito é coberto com pele de gato
		Ele lê no leito de palha
		Paira um ar de arara rara no rio real
		Foi muito difícil entender a canção
		Depois do almoço te encontro
		Esses são nossos times
		Procurei Maria na copa
		A pesca é proibida nesse lago
		Espero te achar bem quando voltar
		Temos muito orgulho da nossa gente
		O inspetor fez a vistoria completa
		Ainda não se sabe o dia da maratona
		Será muito difícil conseguir que eu venha
		A paixão dele é a natureza
		Você quer me dizer a data
		Desculpe, mas me atrasei no casamento
		Faz um desvio em direção ao mar

Tabela B.1: Relações das locuções de adaptação e de teste dos locutores M01 e M07.

Relação de locuções		
Locutor	Relação de locuções de adaptação	Relação de locuções de teste
M03	O grêmio ganhou a quadra de esportes	O tele-jornal termina às sete da noite
M06	Hoje irei à vila sem meu filho	A cabine telefônica fica na próxima rua
M11	Essa magia não acontece todo dia	Defender a ecologia é manter a vida
	Será bom que você estude esse assunto	Nesse verão o calor está insuportável
	O menu incluía pratos bem saborosos	Um jardim exige muito trabalho
	Podia dizer as horas, por favor	O mamão que eu comprei estava ótimo
	A casa é ornamentada com flores do campo	Meu primo falará com a gerência amanhã
	A terra é farta, mas não infinita	De dia apague a luz sempre
	O sinal emitido é captado por receptores	A sociedade uruguaia tem que se mobilizar
	A mensalidade aumentou mais que a inflação	Suas atitudes são bem calmas
		Dezenas de cabos eleitorais buscavam apoio
		A vitória foi paga com muito sangue
		Nossa filha tem amor por animais
		Esse peixe é mais fatal que certas cobras
		O time continua lutando pelo sucesso
		Essa medida foi devidamente alterada
		O estilete é uma arma perigosa
		Aguarde, quinta eu venho jantar em casa
		A mudança é lenta, porém duradoura
		O clima não é mais seco no interior
		A sensibilidade indicará a escolha
		A Amazônia é a reserva ecológica do globo
		O ministério mudou demais com a eleição
		Novos rumos se abrem para a informática
		O capital de uma empresa depende da produção
		Se não fosse ela, tudo seria contido
		A principal personagem no filme é uma gueixa
		Receba seu jornal em sua casa
		A juventude tinha que revolucionar a escola
		A atriz terá quatro meses para ensaiar seu canto

Tabela B.2: Relações das locuções de adaptação e de teste dos locutores M03, M06 e M11.

Relação de locuções		
Locutor	Relação de locuções de adaptação	Relação de locuções de teste
M04	Muito prazer em conhecê-lo	Receba meus parabéns pela apresentação
M15	Eles estavam sem um bom equipamento	Eu planejo uma viagem no feriado
	O sol ilumina a fachada de tarde	No lado de cá do rio há uma boa sombra
	A correção do exame está coerente	A maioria dos visitantes gosta deste monumento
	As portas são antigas	Minha filha é especialista em música sacra
	Sobrevoamos Natal acima das nuvens	A casa só tem um quarto
	Trabalhei mais do que podia	A duração do simpósio é de cinco dias
	Hoje eu acordei muito calmo	Ao contrário de nossa expectativa, correu tranquilo
	Esse canal é pouco informativo	A intenção é obter apoio do governante
	Parece que nascemos ontem	a fila aumentou ao longo do dia
		À noite a temperatura deve ir a zero
		A proposta foi inspecionada pela gerência
		O quadro mostra uma face do cotidiano
		Já era bem tarde quando ele me abordou
		O canário canta ao amanhecer
		A lojinha fica bem na esquina de casa
		Meu time se consagrou como o melhor
		Um instituto deve servir a sua meta
		Ele entende quando se fala pausadamente
		Seu saldo bancário está baixo
		O termômetro marcava um grau
		O discurso de abertura é bem longo
		Eu precisei de microfone na conferência
		Joyce esticou sua temporada até quinta
		Nada como um almoço ao ar livre
		Nossa filha é a primeira aluna da classe
		Gostaria de deitar um pouco
		Não fizemos uma viagem muito cansativa
		Ainda tenho cinco telefonemas para dar
		Os hotéis do sudoeste são fantásticos

Tabela B.3: Relações das locuções de adaptação e de teste dos locutores M04 e M15.

Relação de locuções		
Locutor	Relação de locuções de adaptação	Relação de locuções de teste
M05	Os maiores picos da Terra ficam debaixo d'água	É fundamental chegar a uma solução comum
M17	A inauguração da vila é quarta feira	Há previsão de muito nevoeiro no rio
	Só vota quem tiver o título de eleitor	Muitos móveis virão às cinco da tarde
	É fundamental buscar a razão da existência	A casa pode desabar em algumas horas
	A temperatura só é boa mais cedo	O candidato falou como se estivesse eleito
	Em muitas regiões a população está diminuindo	A idéia é falha, mas interessa
	Nunca se pode ficar em cima do muro	O dia está bom para passear no quintal
	Pra quem vê de fora o panorama é desolador	Minhas correspondências não estão em casa
	É bom te ver colhendo flores	A saída para a crise dele é o diálogo
	Eu me banho no lago ao amanhecer	Finalmente o mau tempo deixou o continente
		Um casal de gatos come no telhado
		A cantora foi apresentar seu último sucesso
		Lá é um lugar ótimo para tomar uns chopinhos
		O musical consumiu sete meses de ensaio
		Nosso baile inicia após as nove
		Apesar desses resultados, tomarei uma decisão
		A verdade não poupa nem as celebridades
		As queimadas devem diminuir este ano
		O vão entre o trem e a plataforma é muito grande
		Infelizmente não compareci ao encontro
		As crianças conheceram o filhote de ema
		A bolsa de valores ficou em baixa
		O congresso volta atrás em sua palavra
		A médica receitou que eles mudassem de clima
		Não é permitido fumar no interior do ônibus
		A apresentação foi cancelada por causa do som
		Uma garota foi presa ontem à noite
		O prato do dia é couve com atum
		Eu viajarei ao Canadá amanhã
		A balsa é o meio de transporte daqui

Tabela B.4: Relações das locuções de adaptação e de teste dos locutores M05 e M17.

Relação de locuções		
Locutor	Relação de locuções de adaptação	Relação de locuções de teste
M23	A velha leoa ainda aceita combater	O jogo será transmitido bem tarde
	É hora do homem se humanizar mais	É possível que ele já esteja fora de perigo
	Ela ficou na fazenda por uma hora	A explicação pode ser encontrada na tese
	Seu crime foi totalmente encoberto	Meu vôo tinha sido marcado para as cinco
	A escuridão da garagem assustou a criança	Daqui a pouco a gente irá pousar
	Ontem não pude fazer minha ginástica	Estou certo que mereço a atenção dela
	Comer quindim é sempre uma boa pedida	Era um belo enfeite todo de palha
	Hoje eu irei precisar de você	O comércio daqui tem funcionado bem
	Sem ele o tempo flui num ritmo suave	É a minha chance de esclarecer a notícia
	A sujeira lançada no rio contamina os peixes	A visita transformou se numa reunião íntima
		O cenário da história é um subúrbio do rio
		Eu tenho ótima razão para festejar
		A pequena nave medirá o campo magnético
		O prêmio será entregue sem sessão solene
		A ação se passa numa cidade calma
		Ela e o namorado vão a Portugal de navio
		O adiamento surpreendeu a mim e a todos
		A gente sempre colhe o que plantou
		Aqui é onde existem as flores mais interessantes
		A corrida de inverno aconteceu com vibração
		Esse empreendimento será de enorme sucesso
		As feiras livres não funcionam amanhã
		Fumar é muito prejudicial à saúde
		Entre com seu código e o número da conta
		Reflita antes e discuta depois
		As aulas dele são bastante agradáveis
		Usar aditivos pode ser desastroso
		O clima não é mau em Calcutá
		A locomotiva vem sem muita carga
		Ainda é uma boa temporada para o cinema

Tabela B.5: Relações das locuções de adaptação e de teste do locutor M23.

Apêndice C

Trabalho Publicado

Este trabalho gerou uma publicação no *International Workshop on Telecommunications 2004*, realizado no Instituto Nacional de Telecomunicações (INATEL) localizado em Santa Rita do Sapucaí - MG, entre os dias 23 e 27 de agosto de 2004. O título da publicação é *"Speaker Adaptation Using Eigenvoices in a Continuous Speech Recognition System"*, e o artigo completo é dado a seguir.

Speaker Adaptation Using Eigenvoices in a Continuous Speech Recognition System

Lívio Carvalho Sousa
State University of Campinas - UNICAMP
P.O. Box 6101 - 13083-852
Campinas - SP - Brazil
livio@decom.fee.unicamp.br

Fábio Violaro
State University of Campinas - UNICAMP
P.O. Box 6101 - 13083-852
Campinas - SP - Brazil
fabio@decom.fee.unicamp.br

***Abstract*— The intention of this work is to present the result of some experiments carried on speaker adaptation using eigenvoices [1, 2, 3] and applied over a continuous speech recognition system (SRS) under development at the Digital Speech Processing Laboratory of the School of Electrical and Computer Engineering (FEEC), University of Campinas (UNICAMP), Brazil. Although not yet fully conclusive, the results of some simulations show the potential of the eigenvoice technique in improving the word recognition rate with only a few sentences of adaptation. Improvements of up to 5% in word recognition rate were obtained. Sometimes just one adaptation sentence showed to be more effective than 10 adaptation sentences. The question remaining is what set of sentences to choose in order to get a better adaptation.**

***Index Terms*—eigenvoices, speaker adaptation, continuous speech recognition system.**

I. INTRODUCTION

Nowadays, speech recognition systems (SRS) are being used in many applications like user identification systems, call centers, embedded devices, voice control systems and other applications. In all these applications the system must present an user independent minimum acceptable performance. Sometimes, to achieve this minimum acceptable performance, a real time fast adaptation is required for each new user.

In SRS the speech subunits are stochastically modeled by Hidden Markov Models (HMM). In continuous SRS these subunits are normally context independent or context dependent phones. The models can be either speaker dependent (SD) or speaker independent (SI). The SD models are trained with the speech data from a single speaker, the one that is supposed to be the SRS user. The SI models are trained with the speech data from many speakers and can be used by any speaker, but at the expense of a higher error rate (typically 2 or 3 times higher) [2,4].

The first step to train the models is to get a large speech database. As more speech material is available, better-trained models will be obtained with a consequent lower word error rate. In many applications where the users change frequently, it is not feasible to retrain the system for each new user with a long lasting training session. Therefore, many studies are been made in order to adapt models, that is, transform a SI model into a SD model, using much less speech data than would be necessary to directly train a SD model. One of the most promising techniques is the one based on eigenvoices.

II. SPEAKER ADAPTATION

The main objective of speaker adaptation is to get a specific model for a new speaker when this speaker is not well represented by the SI model. The generation of this new model demands a lot of training time and a large SD database, what is not feasible in many applications where the user just wants to access a service for a few minutes. Because of that there is a lot of effort in developing adaptation algorithms requiring only a few sentences of adaptation and demanding a low adaptation time, what is a main requirement for SRS accessed by many different people for short periods of use. Normally not all the CDHMMs (Continuous Density Hidden Markov Models) parameters must be adapted, just the gaussian means. The remaining parameters (variances, weighting coefficients and transition probabilities) are inherited from the previous SI system [9,21].

At the present time many speaker adaptation techniques are available for speaker adaptation in SRS [9]:

- MAP (Maximum a Posteriory) [10,11,12,13,14, 15];
- MLLR (Maximum Likelihood Linear Regression) [16,17,18,19];
- CAT (Cluster Adaptive Training) [20].

A disadvantage of the MAP technique is that it has a slow adaptation. Furthermore speech data related to all speech sub-units must be present in the adaptation database. MLLR on the other hand can re-estimate parameters whose data were not observed in the adaptation database, but also requires a large adaptation

database. In the CAT technique, different models are generated and associated to clusters of speakers and a new SD model is represented as a linear combination of cluster models.

III. EIGENVOICE TECHNIQUE

The eigenvoice technique [1,2,3,4,5] represents a new speaker as a linear combination of previously trained SD models, resembling in some aspects the ideas of the CAT technique.

Consider that a SI model is available. Consider also that L speaker databases are available and that a SD model is created for each one of these L speakers. As we are considering in this paper only the adaptation of the gaussian means, from each SD model all the gaussian means are stored into one vector called supervector. In the simulations reported in this paper, a continuous SRS is considered using 36 context independent phones, with each phone represented by a 3 state HMM. The emission density of each state is modeled as a mixture of 5 gaussians of dimension 25 (the dimension of a grouped set of acoustic parameters after dimension reduction by using principal component analysis - PCA). The dimension of each supervector is then 36x3x5x25=13500 and L supervectors with dimension 13500 are then created.

By using dimension reduction techniques, the eigenvalues and eigenvectors of the covariance or correlation matrix associated to the L supervectors are then calculated. As the covariance or correlation matrix has dimension [13500 x 13500], the computational load to calculate the eigenvalues and eigenvectors would be very expensive. So, as an alternative, the SVD (Singular Value Decomposition) was chosen for calculating the eigenvalues and eigenvectors from XX^T, where X is the supervector matrix [13500 x L]. The L eigenvectors [13500 x 1] obtained using SVD are then ordered in a decreasing order of their corresponding eigenvalues. The higher order eigenvectors (lower eigenvalues) can then be discarded, resulting K eigenvectors ($K<L$). These K eigenvectors ($e(j)$, j=1,...,K) are called eigenvoices and the mean value of the supervectors is called eigenvoice $e(0)$. The eigenvoices constitute an orthogonal base of the eigenspace where each new speaker will be represented.

The eigenvoice technique estimates the supervector of the new speaker as $e(0)$ plus a linear combination of the K eigenvectors that point to the directions of maximum variability of the speaker space:

$$\hat{\mu} = e(0) + \sum_{j=1}^{K} w_j e(j) \qquad (1)$$

To get the adapted model for the new speaker it is just necessary to calculate the eigenvoice coefficients w_j. Remember that, as we are adapting only the gaussian means, the variances, weighting coefficients and transition probabilities are kept the same as in the SI model.

The Maximum Likelihood Eigen Decomposition Algorithm (MLED) was proposed by Kuhn [1,3,5] as the iterative algorithm to estimate the w_j coefficients based on a maximum likelihood criterion.

Each eigenvoice can be considered as the association of 36x3x5=540 concatenated sub-eigenvoices, with each sub-eigenvoice having dimension 25 and corresponding to the mean of each gaussian of the 5 gaussians per state, 3 state, 36 CDHMMs:

$$e(j) = \begin{bmatrix} e_1^{(1)}(j) \\ e_2^{(1)}(j) \\ \vdots \\ e_m^{(s)}(j) \\ \vdots \\ e_M^{(S)}(j) \end{bmatrix}, \qquad (2)$$

where $e_m^{(s)}(j)$ is the j^{th} sub-eigenvoice corresponding to gaussian m (m=1,...,M, M=5) in state s (s=1,...,S, S=36x3=108).

The re-estimation equations of the MLED algorithm are obtained from the Baum equation [1,3,22]

$$Q(\lambda,\lambda') = -\tfrac{1}{2} P(O/\lambda) \sum_s \sum_m \sum_t \gamma_m^{(s)}(t) f(o_t, s, m), \quad (3)$$

where

$$f(o_t, s, m) = [-D.\log(2\pi) - \log\left|C_m^s\right| + h(o_t, s, m)], \quad (4)$$

$$h(o_t, s, m) = \left(\mu_m^{(s)} - o_t\right)^T C_m^{(s)^{-1}} \left(\mu_m^{(s)} - o_t\right) \quad (5)$$

and

- $P(O/\lambda)$ is the probability of the observation sequence O, given the model λ;
- D is the dimension of the acoustic parameters o_t;
- $C_m^{(s)^{-1}}$ is the inverse covariance matrix of gaussian m in state s;
- $\mu_m^{(s)}$ is the mean of gaussian m in state s.
- $\gamma_m^{(s)}(t)$ is the emission probability of parameter o_t produced by gaussian m in state s.

Consider the representation of the adapted gaussian means by equation

$$\hat{\mu} = \sum_{j=1}^{K} w_j e(j), \qquad (6)$$

where K is the chosen number of eigenvoices. Substituting equation (6) in (5), deriving equation (3) with respect to w_j and making the derivative equal to zero, we obtain the re-estimation equation

$$\sum_s \sum_m \sum_t [\gamma_m^{(s)}(t)\left(e_m^{(s)}(j)\right)^T C_m^{(s)^{-1}} o_t] = \\ = \sum_s \sum_m \sum_t \left\{ \gamma_m^{(s)}(t) \left[\sum_{k=1}^{K} \left(w_k \left(e_m^{(s)}(k) \right)^T C_m^{(s)^{-1}} e_m^{(s)}(j) \right) \right] \right\} \quad (7)$$

Notice that equation (7) is based on equation (6) and not on equation (1). To compensate this difference, a normalization is provided in the acoustic parameters by subtracting the corresponding sub-eigenvoice $e_m^{(s)}(0)$ [3].

Equation (7) can then be rewritten as

$$d(1,j)w_1 + \cdots + d(K,j)w_K = D(j), \quad (8)$$

where

$$d(k,j) = \sum_s \sum_m \left[\sum_{n=1}^{N} \left(\frac{e_{mn}^{(s)}(k) e_{mn}^{(s)}(j)}{\sigma_{mn}^{2\ (s)}} \right) \sum_t \gamma_m^{(s)}(t) \right] \quad (9)$$

and

$$D(j) = \sum_s \sum_m \sum_t \left[\sum_{n=1}^{N} \left(\frac{e_{mn}^{(s)}(j) o_{tn}}{\sigma_{mn}^{2\ (s)}} \right) \gamma_m^{(s)}(t) \right]. \quad (10)$$

The term $e_{mn}^{(s)}(j)$ is the n^{th} component of sub-eigenvoice $e_m^{(s)}(j)$, o_{tn} is the n^{th} component of the parameters vector o_t, and $\sigma^2_{mn}{}^{(s)}$ is the n^{th} component of the variance of gaussian m in state s.

Making J=1,2,...,K in equation (8), we get a K equations system with K unknown variables that can be solved by using the gaussian elimination method. After calculating the eigenvoice coefficients, new means are re-estimated using equation (5) and a new model is obtained. From this new model, new values of $\gamma_m^{(s)}(t)$ are calculated and a new iteration is done, resulting a new set of coefficients w's. The procedure continues until some stopping criterion is reached.

The reported works about isolated word applications indicate 2 iterations for MLED algorithm [3,5], while [2] applied 3 iterations for his continuous speech application. An objective of this work is also to verify the influence of the number of iterations in the performance of the adapted system.

IV. Acoustic Parameters and Speech Databases

The acoustic parameters used in the CDHMMs are extracted at each 10 ms from a pre-emphasized speech signal (1-0.95z^{-1}) using a 20 ms Hamming window. The employed parameters are: 12 Mel-cepstral coefficients; 12 associated delta-Mel-cepstral coefficients; 12 associated delta-delta-Mel-cepstral coefficients; 1 normalized log-energy-coefficient; 1 associated delta-log-energy coefficient; 1 associated delta-delta-log-energy coefficient [6,7]. In calculating the delta coefficients, only one frame was considered to the left and right of the considered frame. All these 6 parameters were joined into a single vector with dimension 39 and, after a PCA analysis [8], this dimension was reduced to 25.

The speech database for the SI system was created using a sampling frequency of 11.025 kHz, a set of 200 phonetically balanced sentences and 40 speakers, 20 male and 20 female [22]. The training set consisted of 1200 phrases spoken by 30 speakers and the testing set consisted of 400 phrases spoken by other 10 speakers. The phonetic transcription of the speech files was made manually, speaker dependent, file by file, considering 36 context-independent phones for the Brazilian Portuguese language.

For the eigenvoices based adaptation algorithm, 18 SD databases were generated, over the same set of 200 phonetically balanced sentences, with 400 sentences per speaker. Just male speakers were considered. In the SD databases, to reduce the time spent with the preparation of the simulations, a standard phonetic transcription was employed (not speaker dependent).

The whole vocabulary consisted of 700 words. In the recognition section a word-pair grammar was employed. This grammar was created over the set of 200 sentences used in the database.

V. Adaptation Process

From the SD models, 18 supervectors were created and the SVD was applied on this set of supervectors resulting 18 eigenvalues and 18 eigenvectors, the eigenvoices. Because the number of base speakers used in our simulations was very low, the number of eigenvoices was made equal to the number of base speakers (K=L).

The adaptation data consisted of a set of up to 10 sentences with their associated phonetic transcriptions.

The Viterbi algorithm was used to calculate $\gamma_m^{(s)}(t)$, although, as reported in [2], the Forward-Backward algorithm could also be used.

When using just one adaptation sentence, the parameters o_t are extracted and the parameters $\gamma_m^{(s)}(t)$ are calculated using the means of the SI model. The eigenvoices coefficients are estimated by K equations (8) and new means are re-estimated. After that new parameters $\gamma_m^{(s)}(t)$ are calculated using the means obtained in the previous iteration. The iterative procedure continues until a certain number of iterations or a stopping criterion is reached.

For more than one adaptation sentence, equations (9) and (10) require one more summation in the number of sentences. After all sentences are processed, the eigenvoices coefficients are estimated and the new means are re-estimated.

The adaptation was applied for 10 speakers belonging to the SI speech database. The number of adaptation sentences was varied from 1 to 10 for each speaker and the number of iterations was made to vary from 1 to 5.

VI. Results and Comments

At first the speaker adaptation was tested over 10 speakers with a variable number of iterations and adaptation sentences. One set of 10 adaptation sentences was applied for speakers (M01, M07); other different sets were applied for speakers (M03, M06, M11), speakers (M04, M15) and speakers (M05, M17). The results are summarized in Tables I to X. On the top of the tables is presented the average word recognition rate of the considered speaker when using the SI system. The remaining values refer to the adapted system for the specific speaker under different number of adaptation sentences and algorithm iterations. The resulting improved average word recognition rates are presented in bold style.

TABLE I

Speaker M01					SI Model: 87,57%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	**88,17%**	**92,90%**	**91,72%**	**92,90%**	**92,90%**
2	**87,57%**	**91,72%**	**92,90%**	**92,90%**	**93,49%**
3	**87,57%**	**92,31%**	**91,72%**	**88,76%**	**88,76%**
4	**88,17%**	**88,76%**	**91,12%**	**88,17%**	85,80%
5	85,21%	**91,12%**	**91,12%**	**91,72%**	**91,72%**
6	85,21%	**89,94%**	**91,12%**	**91,12%**	**91,72%**
7	86,98%	**90,53%**	**91,12%**	**91,12%**	**91,12%**
8	**88,17%**	**88,76%**	**90,53%**	**91,12%**	**91,12%**
9	86,39%	**88,76%**	**90,53%**	**90,53%**	**91,12%**
10	86,98%	**89,35%**	**91,12%**	**91,12%**	**91,72%**

TABLE II

Speaker M03					SI Model: 82,18%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	62,38%	69,80%	68,32%	66,34%	64,36%
2	58,42%	69,31%	73,27%	75,25%	76,24%
3	64,85%	67,82%	74,26%	73,27%	72,77%
4	64,85%	76,24%	74,26%	73,76%	73,76%
5	62,38%	78,22%	79,21%	77,72%	80,20%
6	66,34%	76,73%	79,70%	79,21%	79,21%
7	64,36%	76,73%	79,21%	79,21%	75,74%
8	64,85%	71,78%	75,74%	76,73%	75,74%
9	65,35%	72,77%	77,72%	78,22%	75,74%
10	65,35%	79,70%	79,21%	79,70%	78,71%

TABLE III

Speaker M04					SI Model: 81,41%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	76,38%	78,89%	71,36%	76,38%	71,86%
2	78,39%	79,40%	74,87%	75,88%	74,87%
3	76,88%	80,40%	75,38%	75,38%	75,88%
4	77,39%	78,89%	**81,91%**	79,40%	80,40%
5	74,87%	77,39%	80,90%	**81,91%**	79,40%
6	77,39%	**81,41%**	**81,91%**	**82,91%**	79,90%
7	77,39%	80,90%	**82,91%**	**82,91%**	79,90%
8	78,39%	79,90%	80,90%	**81,41%**	80,40%
9	78,89%	**81,91%**	79,90%	**83,42%**	**82,41%**
10	78,89%	78,89%	80,90%	**83,42%**	80,90%

TABLE IV

Speaker M05					SI Model: 92,09%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	84,65%	80,93%	81,40%	82,79%	81,40%
2	87,44%	85,05%	85,12%	85,12%	84,19%
3	82,79%	83,72%	85,12%	83,26%	79,07%
4	83,72%	80,93%	80,93%	81,86%	82,79%
5	80,93%	81,86%	80,47%	80,93%	82,33%
6	83,26%	81,86%	80,93%	80,93%	82,33%
7	84,19%	81,86%	82,33%	82,33%	83,26%
8	84,19%	81,86%	80,47%	79,53%	81,86%
9	83,72%	83,26%	81,86%	81,86%	80,00%
10	82,33%	84,19%	81,86%	82,33%	82,79%

TABLE V

Speaker M06					SI Model: 69,80%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	69,31%	68,32%	65,84%	64,85%	60,40%
2	66,83%	65,84%	63,37%	63,37%	63,86%
3	69,31%	67,33%	**69,80%**	67,82%	68,32%
4	68,32%	69,31%	**71,29%**	**70,79%**	**70,79%**
5	**74,26%**	68,81%	**71,29%**	**72,28%**	68,32%
6	**73,27%**	68,81%	**71,29%**	**71,29%**	68,32%
7	**75,74%**	**70,79%**	**71,29%**	**70,30%**	**73,76%**
8	**75,74%**	**69,80%**	**70,79%**	**73,27%**	**74,26%**
9	**73,27%**	**72,77%**	**72,77%**	**70,79%**	**72,28%**
10	**72,77%**	**71,78%**	**69,80%**	**72,77%**	**73,27%**

TABLE VI

Speaker M07					SI Model: 93,49%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	74,56%	85,21%	85,21%	84,02%	83,43%
2	85,80%	83,43%	84,02%	86,98%	84,02%
3	82,84%	78,70%	77,51%	79,88%	81,66%
4	82,25%	82,84%	85,21%	88,17%	87,57%
5	82,84%	82,84%	82,84%	89,35%	89,35%
6	83,43%	79,88%	85,80%	84,02%	86,98%
7	84,02%	81,66%	87,57%	86,39%	85,80%
8	83,43%	83,43%	85,80%	86,39%	84,02%
9	82,25%	81,07%	87,57%	88,17%	88,17%
10	83,43%	82,84%	86,39%	88,17%	88,17%

TABLEI VII

Speaker M11					SI Model: 94,55%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	93,07%	**95,54%**	**94,55%**	94,06%	93,56%
2	94,06%	94,06%	92,08%	91,58%	91,58%
3	**95,05%**	**95,54%**	**95,54%**	**95,05%**	**95,05%**
4	94,06%	**95,54%**	**96,04%**	**94,55%**	**94,55%**
5	**95,05%**	**96,04%**	**96,04%**	**96,04%**	**95,05%**
6	**96,53%**	**96,04%**	**96,04%**	**95,05%**	**95,05%**
7	94,06%	**95,54%**	**95,05%**	**95,05%**	**95,05%**
8	**94,55%**	**95,54%**	**95,54%**	**95,05%**	**95,05%**
9	**94,55%**	**95,54%**	**94,55%**	**95,05%**	**95,05%**
10	**94,55%**	**94,55%**	**94,55%**	**94,55%**	**94,55%**

TABLE VIII

Speaker M15					SI Model: 93,97%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	88,44%	85,93%	85,43%	83,92%	82,91%
2	92,46%	90,45%	86,43%	86,43%	88,44%
3	90,50%	89,95%	88,94%	88,94%	88,94%
4	91,46%	93,47%	91,46%	92,46%	92,46%
5	89,45%	92,46%	91,46%	90,45%	91,46%
6	90,45%	92,46%	90,45%	91,46%	91,46%
7	89,95%	92,46%	91,96%	90,95%	90,45%
8	90,45%	92,46%	90,95%	90,95%	91,46%
9	90,45%	90,95%	91,96%	90,95%	90,95%
10	90,45%	90,95%	91,96%	90,95%	90,95%

TABLE IX

Speaker M17					SI Model: 92,56%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	87,47%	87,91%	90,70%	88,84%	86,98%
2	83,26%	86,05%	89,77%	90,70%	89,77%
3	86,05%	86,98%	87,44%	87,91%	89,77%
4	87,44%	85,12%	88,84%	91,16%	91,63%
5	85,58%	87,91%	89,30%	89,77%	90,70%
6	85,12%	87,44%	89,77%	89,30%	89,77%
7	84,65%	87,91%	88,84%	88,84%	88,84%
8	84,19%	86,51%	89,77%	89,77%	90,23%
9	83,26%	87,91%	87,91%	89,30%	89,77%
10	83,72%	88,84%	89,77%	89,77%	89,77%

TABLE X

Speaker M23					SI Model: 89,10%
Nºof sentences	Nº of Iterations				
	1	2	3	4	5
1	88,15%	**89,10%**	86,73%	86,73%	87,68%
2	86,73%	**89,57%**	**91,00%**	**89,10%**	**89,10%**
3	87,68%	**91,94%**	**91,00%**	**90,05%**	**91,47%**
4	**91,00%**	**94,31%**	**92,89%**	**92,89%**	**92,89%**
5	**89,57%**	**92,42%**	**92,42%**	**91,47%**	**92,89%**
6	**89,57%**	**92,89%**	**92,89%**	**92,42%**	**92,42%**
7	**89,10%**	**94,31%**	**92,89%**	**92,89%**	**91,94%**
8	87,68%	**94,31%**	**92,42%**	**92,42%**	**92,42%**
9	**89,57%**	**92,42%**	**90,52%**	**90,52%**	**90,05%**
10	87,68%	**93,84%**	**91,00%**	**91,47%**	**91,47%**

The general conclusions that can be inferred from the tables above are:

- For some speakers (M01, M04, M06, M11, M23) the adaptation resulted an increase in the word recognition rate (results in bold style), while for others (M03, M05, M07, M15, M17) it resulted a decrease in the recognition rate;
- The maximum improvement was 5.92%, obtained for speaker M01 with 2 adaptation sentences and 5 iterations;
- For some speakers some degradation resulted after using the adaptation algorithm. With speaker M07 the reduction in the average recognition word rate exceeded 18% when using 1 adaptation sentence and 1 iteration;
- Not always increasing the number of iterations is worthwhile. A good compromise value for the number of iterations can be set equal to 3;

- Not always increasing the number of adaptation sentences results an increase in the average word recognition rate;
- Examining the 10 tables, the results may appear disappointing. The improvement was observed just over 5 of the 10 test speakers.

For two speakers to whom the adaptation procedures were not worthwhile, a new test was made, using a single adaptation sentence and trying each one of the 10 adaptation sentences available. In this case it was observed that just one specific adaptation sentence could result a gain over the SI result. These results are summarized in Figs. 1 and 2. In Fig. 1, related to speaker M15, only 2 iterations were employed and with the 7th adaptation sentence the performance exceeded the SI performance by 1%. In Fig. 2, related to speaker M17, 3 iterations were employed and with the 2nd adaptation sentence the performance exceeded the SI performance by 1.5%. It should be reminded that in the previous simulations, with the adaptation being made with up to 10 sentences, speaker M15 was in the best case 0.51% below the SI result and speaker M17 was 0.92% below the SI result. It must be also emphasized that the good adaptation sentences in these two situations are different from the first sentences used in the adaptations shown in Tables VIII and IX. The test with one sentence and 2 iterations in Table VIII (85.93%) corresponds to individual sentence 1 in Fig. 1. The test with one sentence and 3 iterations in Table IX (90.7%) corresponds to individual sentence 1 in Fig. 2.

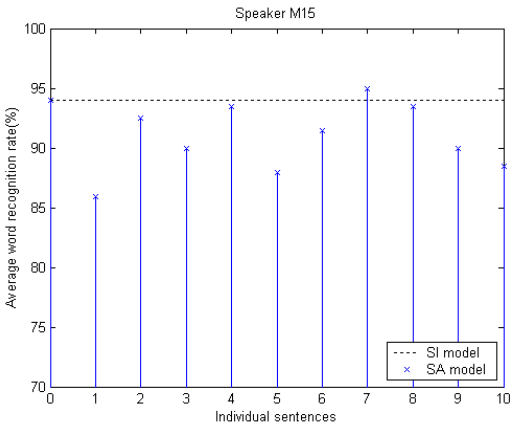

Fig. 1. Adaptation of speaker M15 using a single sentence

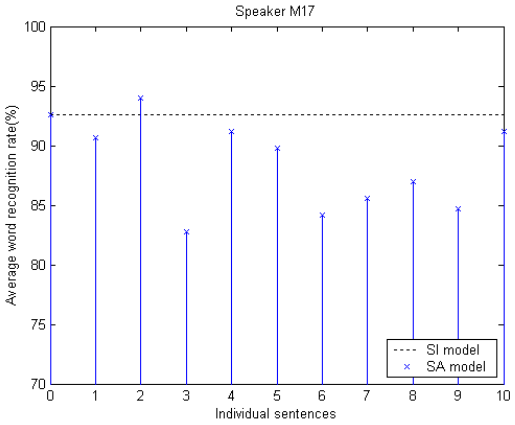

Fig. 2. Adaptation of speaker M17 using a single sentence

In others simulations the number of eigenvoices was reduced ($K<L$) but the performance decreased. This suggests that the number of SD base models used in our simulations was not enough to take into account the acoustic variability of the new speaker. In order to fulfill this deficiency in our simulations, additional SD databases are being recorded in order to result a higher value of base vectors (L around 50) and allow a dimension reduction of the speaker space ($K<L$).

The majority of tests on speaker adaptation using the eigenvoice technique and reported in the literature were made on isolated words SRS. Additionally, not all details are well explained. So, more simulations are necessary in order to understand how to conduct the adaptation so that a gain is always produced in the recognition rate. One point that has to be fully investigated is the high sensitivity of the adaptation process with the number of adaptation sentences and also with the specific sentences used, as shown in Figs. 1 and 2.

Although the results of our simulations are not yet fully conclusive, we hope to have given a little contribution to the research activities in this area.

ACKNOWLEDGEMENTS

The authors would like to express their acknowledgements to FAPESP Agency (Fundação de Amparo à Pesquisa do Estado de São Paulo) for supporting this research under grant 02/05206-1.

REFERENCES

[1] Nguyen, P. “Fast Speaker Adaptation”. *Technical report*, Institut Eurécom, July 1998;

[2] Westwood R.J. “Speaker Adaptation Using Eigenvoices”. Mphil Thesys, University of Cambridge, August 1999;

[3] Kuhn, R.; Junqua, J.-C.; Nguyen, P. and Niedzielski, N. “Rapid Speaker Adaptation in Eigenvoice Space”. *Speech and Audio Processing, IEEE Transactions on*, vol. 8, Edição 6, pp. 695-707, November 2000;

[4] Kuhn, R.; Nguyen, P.; Junqua, J.C. and Goldwasser, L. "Eigenfaces and Eigenvoices: Dimensionality Reduction for Specialized Pattern Recognition", in *Proc. 2nd IEEE Workshop Multimedia Signal Processing*, Redondo Beach, CA, pp. 71-76, December 1998;

[5] Kuhn, R.; Nguyen, P.; Junqua, J.C.; Goldwasser, L.; Niedzielski, N.; Fincke, S.; Field, K. and Contolini, M. "Eigenvoices for Speaker Adaptation", in *Int. Conf. Speech Language Processing `98*, Sydney Austrália, vol. 5, pp. 1771-1774, November 30 – December 04 1998;

[6] Ynoguti, C.A. and Violaro, F. "Desenvolvimento de um Conjunto de Ferramentas para Pesquisas em Reconhecimento de Fala", *Telecomunicações, Revista do Instituto Nacional de Telecomunicações - INATEL*, ISSN 1516-2338, vol. 04, nº 02, pp. 36-43, December 2001;

[7] Davis S. B. and Mermelstein P. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acustic, Speech, and Signal Processing,* ASSP-28, nº 04, August 1980;

[8] Richard J.A. and Dean W.W. "*Applied Multivariate Statistical Analysis*", Prentice Hall 1992;

[9] Woodland P.C. "Speaker Adaptation: Techniques and Challenges", in *Proc IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 85-90, 2000;

[10] Gauvain J.-L. and Lee C.-H. "Bayesian Learning for HMM with Gaussian Mixture State Observation Densities", *Speech Comm.* vol 11, pp. 205-213 1992;

[11] Gauvain J.-L. and Lee C.-H. "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Transactions on Speech and Audio Processing*, vol 02, pp. 291-298, april 1994;

[12] Zavagliakos G., Schwartz R. and McDonough J. "Maximum *a Posteriori* Adaptation for Large Scale HMM Recognizers", in *Int. Conf. Acoustics, Speech, Signal Processing '96*, Atlanta GA, pp. 725-728, 1996.

[13] Lee C.-H. and Gauvain J.-H. "Speaker Adaptation Based on MAP Estimation of HMM Parameters", *IEEE ICASSP-93*, pp. 558-561, 1993;

[14] Shinoda K. and Lee C.-H. "Structural MAP Speaker Adaptation Using Hierarchical Priors", *Proc. IEEE Workshop on Automatic Speech Recognition and Undestanding*, pp. 381-387, 1997;

[15] Ahadi S.M. and Woodland P.C. "Combined Bayesian and Predictive Techniques for Rapid Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language*, vol 11, pp.187-206, 1997;

[16] Leggetter C.J. and Woodland P.C. "Speaker Adaptation of Continuous Density HMM`s Using Linear Regression", in *Int. Conf. Speech Language Processing '94*, Yokohama Japan, vol 02, pp. 451-454, 1994;

[17] Leggetter C.J. and Woodland P.C. "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Markov Models", *Computer Speech and Language*, vol 09, nº 02, pp. 171-185, april 1995;

[18] Gales M. and Woodland P.C. "Mean and Variance Adaptation within the MLLR Framework", *Computer Speech and Language*, vol 10, pp. 250-264, october 1996;

[19] Digilakis V., Ritchev D. and Neumeyer L. "Fast Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures", *IEEE Transactions on Speech and Audio Processing*, vol 03, nº 05, pp. 357-366, september 1995;

[20] Gales M. J. F. "Cluster Adaptive Training for Speech Recognition", in Proc. International Conf. on Spoken Language Processing '98, Sydney Australia, pp. 1783-1786, 1998;

[21] Lee C.-H., Lin C.-H. and Juang B.H. "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models", *IEEE Transactions on Signal Processing*, vol 39, nº 04, pp. 806-814, 1991;

[22] Ynoguti C.A. "*Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*", PhD thesys, Universidade Estadual de Campinas, maio 1999.

Referências Bibliográficas

[1] M. Turk e A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, volume 03(nº 01):71 – 86, 1991.

[2] Chin-Hui Lee; Chih-Heng Lin; e Biing Hwang Juang. A study on speaker adaptation of the parameters of continuous density hidden markov models. *IEEE Transactions on Signal Processing*, volume 39(nº 04):806 – 814, abril 1991.

[3] Lawrence Rabiner e Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[4] K. Shinoda e C.H. Lee. Structural map speaker adaptation using hierarchical priors. In *Proc. IEEE Workshop on Automatic Speech Recognition and Undestanding*, pages 381 – 387, 1997.

[5] Jean-Luc Gauvain e Chin-Hui Lee. Bayesian learning for hmm with gaussian mixture state observation densities. *Speech Comm.*, volume 11:205 – 213, 1992.

[6] Jean Luc Gauvain e Chin-Hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, volume 02:291 – 298, abril 1994.

[7] Carlos Alberto Ynoguti e Fábio Violaro. Desenvolvimento de um conjunto de ferramentas para pesquisa em reconhecimento de fala. *Telecomunicações, Revista do Instituto Nacional de Telecomunicações - INATEL*, volume 04(nº 02):36 – 43, dezembro 2001.

[8] Carlos Alberto Ynoguti e Fábio Violaro. Utilização da análise de componente principal na redução dos vetores de parâmetros em sistemas de reconhecimento de fala contínua. *19º Simpósio Brasileiro de Telecomunicações*, 03 a 06 de setembro 2001.

[9] T. Anastasakos; J. McDonough; R. Schwartz; e J. Makhoul. A compact model for speaker adaptive training. In *Proc. International Conference on Spoken Language Processing '96*, pages 1137 – 1140, Philadelphia, 1996.

[10] G. Zavagliakos; R. Schwartz; e J. McDonough. Maximum *a Posteriori* adaptation for large scale hmm recognizers. In *Int. Conf. Acoustics, Speech, Signal Processing '96*, pages 725 – 728, Atlanta, GA, 1996.

[11] A. Alcaim; J.A. Solewicz; e J.A. Moraes. Frequência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no rio de janeiro. *Revista da Sociedade Brasileira de Telecomunicações*, volume 07(nº 01):23 – 41, dezembro 1992.

[12] Chin-Hui Lee e Jean-Luc Gauvain. Speaker adaptation based on map estimation of hmm parameters. In *IEEE ICASSP-93*, pages 558 – 561, 1993.

[13] R. Kuhn; P. Nguyen; J.-C. Junqua; e L. Goldwasser. Eigenfaces and eigenvoices: Dimensionality reduction for specialized pattern recognition. In *Proc. 2nd IEEE Workshop Multimedia Signal Processing*, pages 71 – 76, Redondo Beach, CA, 7 a 9 de dezembro 1998.

[14] V. Digilakis; D. Ritchev; e L. Neumeyer. Fast speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, volume 03(nº 05):357 – 366, setembro 1995.

[15] M. Kirby e L. Sirovich. Application of the karhunen-loève procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 12(nº 01):103 – 108, janeiro 1990.

[16] R. Kuhn; P. Nguyen; J.-C. Junqua; L. Goldwasser; N. Niedzielski; S. Fincke; K. Field; e M. Contolini. Eigenvoices for speaker adaptation. In *Int. Conf. Speech Language Processing '98*, volume 05, pages 1771 – 1774, Sydney, Austrália, 30 de novembro a 4 de dezembro 1998.

[17] R. Kuhn; P. Nguyen; J.-C. Junqua; R. Boman; N. Niedzielski; S. Fincke; K. Field; e M. Contolini. Fast speaker adaptation using a priori knowledge. In *Int. Conf. Acoustics, Speech, Signal Processing '99*, volume 02, pages 749 – 752, Phoenix, AZ, março 1999.

[18] R. Kuhn; J.-C. Junqua; P. Nguyen; e N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, volume 08(nº 06):695 – 707, novembro 2000.

[19] Steven B. Davis e Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acustic, Speech, and Signal Processing*, ASSP-28(nº 04), agosto 1980.

[20] C.J. Leggetter e P.C. Woodland. Speaker adaptation of continuous density hmm's using linear regression. In *Int. Conf. Speech Language Processing '94*, volume 02, pages 451 – 454, Yokohama, Japão, 1994.

[21] C.J. Leggetter e P.C. Woodland. Flexible speaker adaptation using maximum likelihood linear regression. In *Proc. ARPA Spoken Language Technology Workshop*, pages 104 – 109, 1995.

[22] C.J. Leggetter e P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density markov models. *Computer Speech and Language*, volume 09(nº 02):171 – 185, abril 1995.

[23] M. Gales e P.C. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech and Language*, volume 10:250 – 264, outubro 1996.

[24] S.M. Ahadi e P.C. Woodland. Combined bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, volume 11:187 – 206, 1997.

[25] J.A.Richard e W.W.Dean. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.

[26] S. Furui. Unsupervised speaker adaptation method based on hierarchical spectral clustering. In *Proc. Int. Conf. Acoustics, Speech, Signal Processing '89*, volume 01, pages 286 – 289, Glasgow, maio 1989.

[27] M. J. F. Gales. Cluster adaptive training for speech recognition. In *Proc. International Conf. on Spoken Language Processing '98*, pages 1783 – 1786, Sydney, Australia, 1998.

[28] Liselene. Sistemas de adaptação de locutor utilizando autovozes. Master's thesis, Escola Politécnica da Universidade de São Paulo, 2001.

[29] José Antônio Martins. *Avaliação de Diferentes Técnicas para Reconhecimento de Fala.* PhD thesis, Universidade Estadual de Campinas, dezembro 1997.

[30] P. Nguyen. Fast speaker adaptation. Technical report, Institut Eurécom, julho 1998.

[31] Robert Westwood. Speaker adaptation using eigenvoices. Master's thesis, Cambridge University, agosto 1999.

[32] P.C. Woodland. Speaker adaptation: Techniques and challenges. In *Proc IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 85 – 90, 2000.

[33] Carlos Alberto Ynoguti. *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov.* PhD thesis, Universidade Estadual de Campinas, maio 1999.