Dissertação de Mestrado

Inatel
Instituto Nacional de Telecomunicações

Análise comparativa entre os métodos HMM e GMM-UBM na busca pelo α-ótimo dos locutores crianças para utilização da técnica VTLN

RAMON MAYOR MARTINS

OUTUBRO/2014

Análise comparativa entre os métodos HMM e GMM-UBM na busca pelo α-ótimo dos locutores crianças utilizando a técnica VTLN.

RAMON MAYOR MARTINS

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

ORIENTADOR: Prof. Dr. Carlos Alberto Ynoguti

Martins, Ramon Mayor
M386a
Análise comparativa entre os métodos HMM e GMM-UBM na busca pelo α-ótimo dos locutores crianças para utilização da técnica VTLN. / Ramon Mayor Martins. – Santa Rita do Sapucaí, 2014.
77p.

Orientador: Prof. Dr. Carlos Alberto Ynoguti.
Dissertação de Mestrado – Engenharia de Telecomunicações – Instituto Nacional de Telecomunicações – INATEL.
Inclui bibliografia e anexo.

1. Normalização de locução 2. Reconhecimento de fala 3. Markov 4. Modelos de Mistura Gaussiana 5. VTLN 6. Engenharia de Telecomunicações. I. Ynoguti, Carlos Alberto. II. Instituto Nacional de Telecomunicações – INATEL. III. Título.

CDU 621.39

Santa Rita do Sapucaí
2014

FOLHA DE APROVAÇÃO

Dissertação defendida e aprovada em _____ /_____ /_____ ,
pela comissão julgadora:

_____

Prof. Dr. Carlos Alberto Ynoguti

INATEL

_____

Prof. Dr. Dayan Adionel Guimarães

INATEL

_____

Prof. Dr. Mário Minami

UFABC

_____

Coordenador do Curso de Mestrado
Prof. Dr. José Marcos Câmara Brito

“Deixe o futuro dizer a verdade, e avaliar cada um de acordo com seus trabalhos e suas conquistas”
– Nikola Tesla

À minha família...

AGRADECIMENTOS

Primeiramente, agradeço a Deus por me dar forças para concluir esse trabalho.

Aos meus pais que tanto amo, Antônio Mauro Martins e Maria Aparecida Mayor Martins, pelo incentivo, motivação, apoio e incontáveis outros motivos em todos os momentos desse meu sonho e da minha vida.

A minha avó Fabiana Mayor, minha irmã Manoela e minha sobrinha Lamís pelo carinho de sempre.

A minha namorada Iasmin, pelo companheirismo e por ter me auxiliado em vários momentos nesse trabalho.

Ao Prof. Dr. Carlos Alberto Ynoguti pela enorme paciência, orientação criteriosa, dedicação e excelente profissional que sempre demonstrou ser.

À toda Comunidade INATEL, professores, coordenadores e funcionários pelos conhecimentos adquiridos e atenção.

À Gisele pela atenção que sempre concede aos alunos do mestrado.

Aos eternos amigos Rodrigo Cogliatti, Isackson, Helvécio, Ricardo, Guilherme, Pedro Ivo, Bandiri, Geordan, Simon e tantos outros que fiz no mestrado.

À CAPES e à FAPEMIG pelo auxílio financeiro.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Network*, Rede Neural Artificial
DCT	*Discrete Cosine Transform*, Transformada Discreta do Cosseno
PDF	*Probability Density Function*, Função Densidade Probabilidade
FFT	*Fast Fourier Transform*, Transformada Rápida de Fourier
GMM	*Gaussian Mixture Model*, Modelo de Mistura Gaussiana
HMM	*Hidden Markov Model*, Modelo Oculto de Markov
LFCC	*Linear Frequency Cepstral Coefficients*, Coeficiente Cepstral de Frequência Linear
LPC	*Linear Predictive Coding*, Codificação Preditiva Linear
MFCC	*Mel Frequency Cepstral Coefficient*, Coeficiente de Frequência *Mel* Cepstral
UBM	*Universal Background Model*, Modelo de Base Universal
VT	*Vocal Tract*, Trato Vocal
VTL	*Vocal Tract Length*, Comprimento do Trato Vocal
VTLN	*Vocal Tract Length Normalization*, Normalização de Comprimento do Trato Vocal
WER	*Word Error Rate*, Taxa de Erro de Palavra

RESUMO

Nesta dissertação são abordadas formas de minimizar a alta taxa de erros em sistemas de reconhecimento de fala treinados com locutores adultos e testado com locutores crianças. Propõe-se a utilização do método GMM-UBM como alternativa ao método HMM na busca pelo fator ótimo de escalonamento (α-ótimo) para locutores crianças quando utilizada a técnica de normalização de locutor. A técnica de normalização adotada é a VTLN, que normaliza o trato vocal dos diferentes locutores crianças através do escalonamento de frequências do banco de filtros *mel*. Na avaliação desta técnica, procurou-se também a quantidade de misturas ótimas que melhoram o desempenho do sistema. Desse modo, reduziu-se a taxa de erro no sistema treinado com adultos e testado com crianças de 4,95% para 1,88% quando utilizado a VTLN com os α-ótimos encontrados pelo HMM e 1,92 % quando utilizado a VTLN com os α-ótimos encontrados pelo GMM-UBM. Observou-se que a aplicação da técnica VTLN utilizando os α-ótimos pelo método GMM-UBM obteve desempenho similar ao HMM nos experimentos. Nos experimentos realizados concluiu-se que a escolha do método GMM-UBM se torna mais adequada em virtude da simplicidade de implementação e necessidade de menor custo computacional, sendo assim uma alternativa ao HMM para realizar VTLN em sistemas de reconhecimento de fala para usuários crianças.

Palavras-chave: Normalização de locutor; sistema de reconhecimento de fala; Modelos Ocultos de Markov; Modelos de Mistura Gaussiana; VTLN;

ABSTRACT

The aim of this work is to find means to minimize the high error rate found in speech recognition systems which are trained on adult speakers and tested on children speakers. In this regard, we propose the use of the GMM-UBM method as an alternative to the HMM method to find the optimal warping factor (α-optimal) for children speakers when the speaker normalization technique is used. The adopted normalization technique was VTLN, which normalizes the vocal tract of different children speakers through the use of *mel* filterbank frequency warping. The assessment of this technique also aimed to find the optimal mixture quantity that improves the system performance. Thus, the error rate in the system trained with adults and tested on children was reduced from 4,95% to 1,88% when VTLN was used with α-optimals found by HMM and to 1,92% when VTLN was used with α-optimals found by GMM. It was noticed that the application of VTLN technique using α-optimals found by GMM-UBM method achieved a similar performance to HMM in the experiments. From the experiments it was observed that choosing GMM-UBM method turns to be more suitable due to its implementation simplicity and to the need of lower computational cost, being thus an alternative to HMM in the use of VTLN in Speech Recognition Systems for children speakers.

Keywords: Speaker normalization; speech recognition systems; Hidden Markov Models; Gaussian Mixture Models; VTLN;

Capítulo 1

Introdução

Nos últimos anos, um progresso significativo foi alcançado em sistemas de reconhecimento de fala, os quais vêm sendo aplicados em diversas finalidades e áreas, tais como: na área militar, educacional, entretenimento, médico, pessoal, etc. Hoje é possível redigir um documento simplesmente ditando para a máquina, especificar para um GPS dentro de um automóvel qual o destino que se deseja fazer, realizar uma ligação pelo telefone com um comando de voz, permitir que um piloto e a aeronave conversem entre si, permitir interação entre crianças e máquinas em jogos e inúmeras outras possibilidades. Entretanto, a maior parte dos esforços de pesquisas em reconhecimento de fala foi devotada ao desenvolvimento dos sistemas voltados para locutores adultos. Seguindo os primeiros estudos que despertaram o interesse com relação ao fraco desempenho dos sistemas para usuários crianças, uma crescente atenção foi dada à área das tecnologias de reconhecimento de fala para a fala infantil em uma variedade de cenários de aplicação, tais como: tutores leitores, aprendizado de língua estrangeira, jogos, *sites* educacionais, aplicativos para celulares, entre outros sistemas interativos [1]. Desta forma, um dos desafios dos sistemas de reconhecimento de fala é elevar o desempenho dos sistemas compatíveis de reconhecimento de fala infantil aos mesmos níveis atingidos pelos sistemas "estado da arte" mais modernos de reconhecimento de fala adulta [2].

Algumas dificuldades dos sistemas de reconhecimento de fala estão atreladas às variabilidades na fala entre locutores crianças e adultos. Uma dessas variabilidades é o comprimento do trato vocal (*vocal tract length*, VTL). A compensação das variações acústicas induzidas pela diferença no comprimento do trato vocal entre a fala infantil e a adulta pode ser realizada através da utilização da técnica de normalização de comprimento do trato vocal (*vocal tract length normalization*, VTLN). A VTLN busca reduzir esta variabilidade acústica entre locutores por meio do escalonamento do eixo de frequência do espectro de fala de cada locutor. Algumas investigações utilizando esta técnica mostram que quando um sistema de reconhecimento de fala treinado com locutores adultos é utilizado para reconhecer a fala infantil, a VTLN é capaz de melhorar significativamente o desempenho de reconhecimento [1]. A utilização da técnica VTLN é realizada estimando um fator de escalonamento ótimo (α-ótimo). Este fator de escalonamento é encontrado através de um processo de busca em uma faixa de fatores utilizando métodos estatísticos. Este trabalho explorou o processo de busca a partir de dois métodos: os modelos ocultos de Markov (HMM) e os modelos de mistura Gaussiana (GMM). Objetivou-se realizar uma comparação entre eles e, com isso, avaliar o desempenho do sistema de reconhecimento de fala treinado com locutores adultos e testado com locutores crianças normalizados.

O conteúdo deste trabalho se divide da seguinte forma: no Capítulo 2 são apresentados alguns fundamentos teóricos acerca das características da fala. No Capítulo 3 são abordados os conceitos sobre reconhecimento de fala. No Capítulo 4 é elucidado o processo de normalização de locutor e abordada a técnica de normalização de comprimento do trato vocal (VTLN). No Capítulo 5 são apresentadas os conceitos de caráter introdutório dos métodos estatísticos HMM e GMM. No Capítulo 6 apresentam-se os processos de busca do fator de escalonamento ótimo utilizando os métodos HMM e GMM. No Capítulo 7 é explicado o aparato experimental utilizado. No Capítulo 8 são apresentados os testes realizados, fornecido e discutido os resultados obtidos utilizando o sistema normalizado pela técnica VTLN e o sistema sem normalização. Por fim, no Capítulo 9 tem-se a conclusão, um sumário das principais contribuições e as sugestões para possíveis trabalhos futuros.

Capítulo **2**

Características da Fala

2.1 Produção do sinal de voz

A fala, embora tão natural e sem esforço para os humanos, é uma atividade notavelmente complexa, pois seus sons constituintes revelam grande variabilidade acústica. Esta pode ser de duas naturezas: variações inter-locutores (ou entre locutores) e intra-locutores (ou do próprio locutor) [3].

A variabilidade inter-locutores é o efeito de dissimilaridades nas análises acústicas de locuções por parte de diferentes locutores, ou seja, a pronúncia de uma dada palavra ou frase varia de locutor para locutor. A variabilidade no sinal de fala entre locutores pode ser atribuída em parte às diferenças orgânicas na estrutura do mecanismo vocal e em parte às diferenças aprendidas no uso do mecanismo vocal durante a produção de fala. Diferenças orgânicas podem ser determinadas por hereditariedade, sexo e idade, enquanto diferenças aprendidas podem ser relacionadas a fatores geográficos, sociais e culturais [3]. A variabilidade intra-locutor refere-se ao fato de que o mesmo locutor raramente profere uma dada palavra duas vezes exatamente da mesma forma, ainda que as locuções sejam produzidas sucessivamente. Isso ocorre, por exemplo, em virtude do estado emocional, velocidade da pronúncia, entonação, saúde do locutor, entre outras razões. Neste trabalho há o interesse pelas diferenças orgânicas entre locutores, principalmente por uma fonte física de variabilidade que é a diferença do comprimento do trato vocal (VTL) entre locutores adultos e locutores crianças.

O trato vocal (*vocal tract*, VT) é a fonte principal da geração de fala humana e está representado na região hachurada da Figura 2.1. Ele começa na abertura das cordas vocais, ou glote, e termina nos lábios. De acordo com a dinâmica dos articuladores, diversos sons são produzidos. Os diferentes modos de modificação do fluxo de ar permitem a produção dos sons vocálicos e consonantais.

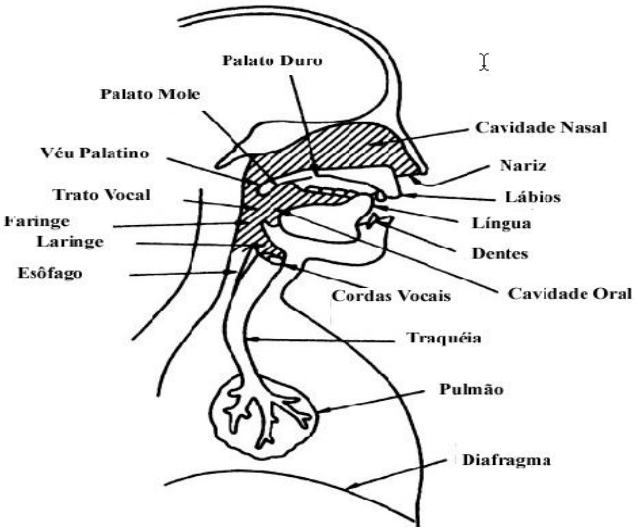

Figura 2. 1: *Trato vocal* [4].

O comprimento do trato vocal é um dos principais responsáveis pela variabilidade acústica entre locutores. A diferença física é mais perceptível entre locutores adultos e crianças. Locutores adultos têm o trato vocal mais longo, enquanto locutores crianças o tem mais curto. Deste modo, tratos vocais maiores tendem a produzir vozes mais graves, enquanto que os menores tendem a produzir vozes mais agudas.

2.2 Características da fala infantil

Além das diferenças anatômicas do trato vocal e o controle menos preciso dos articuladores em locutores crianças, observam-se diferenças importantes nas características espectrais e temporais das vozes das crianças quando comparadas às dos adultos [1]: nas características espectrais, incluem-se frequências formantes e fundamentais mais altas, bem como maior variabilidade espectral. As diferenças temporais incluem durações maiores de segmento de fala, ou seja, a fala das crianças

é mais lenta que a dos adultos [5]. Locutores infantis também demonstram maior variabilidade no sinal da fala e maior esforço para produção de palavras [1]. Além disso, as características acústicas, orgânicas e linguísticas da fala infantil mudam rapidamente com a idade, se tornando um desafio para os sistemas de reconhecimento de fala.

Capítulo **3**

Reconhecimento de fala

3.1 Sistemas de reconhecimento de fala

A função de um sistema de reconhecimento de fala é receber um sinal de voz em sua entrada e produzir em sua saída uma sequência de palavras ou frases que correspondam ao sinal de voz aplicado na entrada.

Os primeiros trabalhos nesta área remontam ao início da década de 50, quando, nos laboratórios Bell, Davis, Biddulph e Balashek construíram um sistema de reconhecimento de dígitos isolados para um único locutor [6]. Nas décadas seguintes muitos outros trabalhos foram realizados para o desenvolvimento desses sistemas, como, por exemplo, os sistemas desenvolvidos na década de 60 por diversos laboratórios japoneses (*Radio Research Lab*. de Tóquio e os laboratórios NEC) e na década de 70 e 80 pelos laboratórios da IBM, AT&T e Laboratórios Bell [6]. Atualmente os sistemas de reconhecimento de fala possuem muitas aplicações, como por exemplo: comandos e controles por voz, ditados, transcrição de discursos e diálogos interativos [7].

Os sistemas de reconhecimento de fala podem ser classificados segundo diversos requisitos. Dentre eles podem ser destacados a habilidade em lidar com locutores específicos e não específicos (dependência de locutor e independência de locutor), com a aceitação de apenas locuções isoladas ou fala fluente (palavra isolada ou palavras conectadas), com o tamanho do vocabulário, com a perplexidade, com o nível de ruído e com a qualidade do transdutor [8].

Para avaliação do desempenho em sistemas de reconhecimento de fala, a métrica mais utilizada é a taxa de erro de palavra [4][5][7]. Essa métrica é importante, pois permite comparar diferentes sistemas, bem como avaliar as melhorias realizadas dentro de um sistema.

Atualmente a tecnologia "estado da arte" em reconhecimento de fala baseada em HMM pode facilmente alcançar uma precisão em reconhecimento de dígitos isolados independente de locutor, atingindo até 3% de taxa de erro de palavra [9]. Além disso, sistemas avançados em ambientes de fala contínua independente de locutor com um vocabulário extenso de palavras e certas restrições gramaticas foram capazes de atingir até 80% de precisão no reconhecimento das palavras [10]. Estes resultados afirmam a utilidade potencial de um sistema de reconhecimento de fala em determinadas aplicações.

3.2 Funcionamento de um sistema de reconhecimento de fala

O processo de reconhecimento de fala consiste em mapear um sinal acústico, capturado por um transdutor (usualmente um microfone), em um conjunto de palavras. As etapas para converter um sinal acústico, em seu correspondente escrito são resumidamente mostradas no diagrama da Figura 3.1 e descritas em seguida.

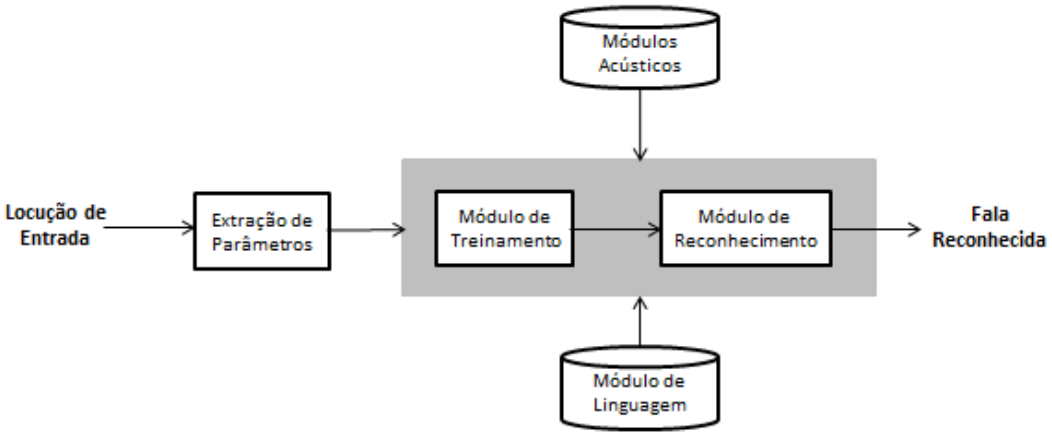

Figura 3. 1: *Diagrama em blocos de um sistema de reconhecimento de fala.*

3.2.1 Extração de parâmetros acústicos

A extração de parâmetros de um sinal de voz é a primeira etapa de um processo de reconhecimento de fala [11]. Uma das finalidades é encontrar um conjunto de propriedades de uma locução que possua uma correlação acústica, ou seja, parâmetros que possam de certa forma ser calculados ou estimados pelo processamento do sinal de fala. Tais parâmetros são denominados características acústicas. Outra finalidade é a redução da quantidade de redundância na informação a ser processada: pretende-se com isso produzir uma representação compacta e eficiente do sinal de fala com o intuito de facilitar o processo de reconhecimento [12].

As observações de saída são definidas por parâmetros representativos do sinal de fala, tais como os coeficientes mel cepstrais (*mel frequency cepstral coefficients*, MFCC), os coeficientes cepstrais de frequência linear (*linear frequency cepstral coefficients*, LFCC), os coeficientes de codificação preditiva linear (*linear predictive coding coefficients*, LPC), entre outros [13][14][15][16]. Neste trabalho serão utilizados os parâmetros *mel-cepstrais* os quais usam conceitos psico-acústicos, que tentam representar a forma como as pessoas percebem os sons.

A estrutura do processamento de sinal de fala, incluindo a etapa para a extração de parâmetros *mel-cepstrais*, é representada na Figura 3.2:

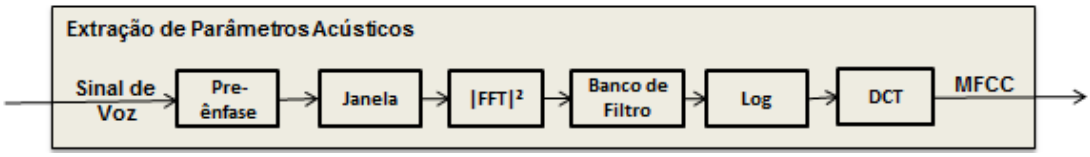

Figura 3. 2: *Processamento do sinal de fala.*

Os passos para a extração de parâmetros consistem na realização da modelagem espectral e, posteriormente, na análise espectral [11]. Na modelagem espectral, o sinal de voz digitalizado passa por um filtro passa-altas denominado pré-ênfase, cuja função de transferência é dada por:

$$H_{pre}(z) = 1 - a_{pre} \, . z^{-1}, \tag{3.1}$$

onde o coeficiente de pré-ênfase a_{pre} varia de 0,9 a 1,0 [4]. O objetivo da pré-ênfase é atenuar as componentes de baixa frequência do sinal de voz, e também, minimizar o efeito dos lábios e da glote [17].

Na etapa seguinte é realizada uma divisão do sinal de voz em intervalos curtos de tempo, denominados janelas, conforme mostrado na Figura 3.3. Isto é feito, pois embora o sinal de voz seja um processo estocástico não estacionário, o fato de o trato vocal mudar muito lentamente durante a fala faz com que seja razoável considerá-lo estacionário em intervalos de curta duração [4].

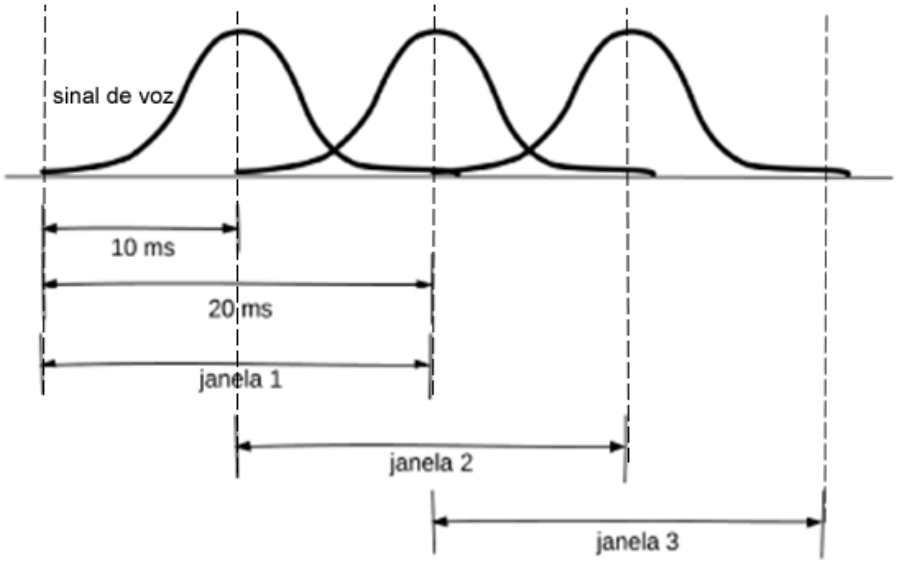

Figura 3. 3 : *Divisão do sinal da voz em janelas.*

O janelamento do sinal tem como função produzir suavização da amplitude do sinal amostrado nos extremos do segmento sob análise, dando maior ênfase às amostras localizadas no centro da janela [4]. Realiza-se este processo com uma superposição parcial entre janelas com o intuito de aumentar a correlação entre as janelas próximas, evitando variações bruscas entre os parâmetros extraídos em janelas adjacentes. Em reconhecimento de fala, um janelamento tipicamente utilizado é o de *Hamming* com janelas de 25 ms, atualizadas a cada 10 ms [4].

Após este processo é realizada a análise espectral. Esta etapa consiste na conversão da representação temporal do sinal de voz em uma representação

espectral. Isto é feito devido ao fato de que a representação espectral é mais correlacionada ao processo de audição e percepção humana do que a representação temporal [17]. Para essa conversão é calculado o quadrado do módulo da transformada rápida de Fourier (*fast Fourier transform*, FFT) das amostras pertencentes a cada janela de análise. Dois métodos de análise espectral são os mais difundidos em aplicações de reconhecimento de fala: o método de banco de filtros e o método de codificação preditiva linear (LPC, *linear predictive coding*) [11]. Este trabalho utilizará o método de banco de filtros. A motivação para a utilização deste método está no fato de que o banco de filtros modela a seletividade em frequência ao longo da membrana basilar na cóclea [18]. Experimentos com a percepção humana mostraram que o nosso sistema auditivo processa os sons em sub-bandas de frequência, chamadas de sub-bandas críticas [11]. A escala *mel*, proposta por *Davis e Melmerstein* [19], implementa esta idéia a partir da expressão:

$$f_{\mathrm{m}} = 2595.\log_{10}.(1+\frac{f}{700}) \tag{3.2}$$

Desta forma, a escala *mel* tenta mapear a frequência percebida de um som, que é logarítmica, em uma escala linear [11]. As frequências centrais para este banco de filtro são dadas na Tabela 3.1. Neste banco, as frequências centrais são atribuídas linearmente de 100 Hz a 1000 Hz. Acima de 1000 Hz, crescem exponencialmente com um fator de $2^{1/5}$ [14].

Tabela 3. 1: *Banco de Filtros baseado na escala mel.*

Índice dos Filtros	Frequência central (Hz)	Índice dos Filtros	Frequência central (Hz)
1	100	13	1516
2	200	14	1741
3	300	15	2000
4	400	16	2297
5	500	17	2639
6	600	18	3031
7	700	19	3482
8	800	20	4000
9	900	21	4595
10	1000	22	5278
11	1149	23	6063
12	1320	24	6964

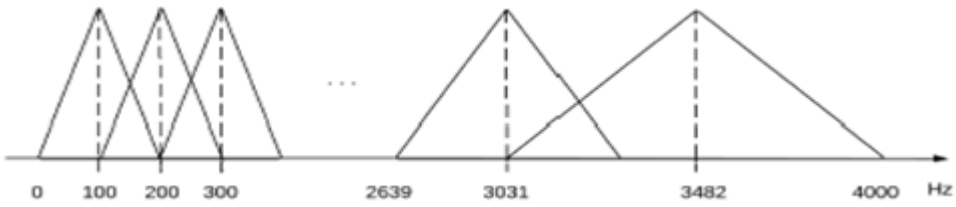

Figura 3. 4: *Representação do banco de filtros na escala mel.*

Após a passagem do sinal pelo banco de filtros é aplicado o logaritmo da energia de saída de cada filtro. Posteriormente aplica-se a transformada discreta do cosseno (*discrete cosine transform*, DCT) sobre os valores, obtendo os parâmetros desejados. Geralmente, no reconhecimento de fala, são utilizados 12 coeficientes *mel-cepstrais* (*mel frequency cepstral coefficients*, MFCC) e 1 coeficiente de energia por janela. Tais coeficientes são obtidos por meio da Equação 3.3:

$$\mathrm{C}_i = \sum_{j=1}^{F} 10.\log_{10}(X_j).\cos\left[i\left(j-\frac{1}{2}\right).\frac{\pi}{F}\right], \tag{3.3}$$

onde X_j é a energia na saída do j-ésimo filtro e $i = 1,2,\ldots,12$, indexa os coeficientes *mel-cepstrais.*

A energia é calculada através da equação:

$$E = \sum_{i=0}^{N-1} s^2(i)\,, \tag{3.4}$$

onde N representa o número de amostras da janela de análise, e $s(i)$ representa o sinal de fala janelado.

Também foi utilizada neste trabalho a energia normalizada, dada por:

$$\mathrm{E}_N = \log(E) - \log(E_{máx})\,, \tag{3.5}$$

onde $E_{máx}$ corresponde à janela de máxima energia para uma determinada frase de treinamento [14].

Também foram utilizados os parâmetros diferenciais (*delta-energia*, *delta-mel-cepstrais* e *delta-delta-mel-cepstrais*). Estes parâmetros são utilizados para caracterizar melhor as variações temporais [14]. Os parâmetros diferenciais foram calculados através da seguinte equação:

$$\Delta_i(n) = \frac{1}{2K+1} \sum_{j=-K}^{K} j y_{i-j}(n)\,, \tag{3.6}$$

onde i é o índice da janela, $y_i(n)$ é o *n*-ésimo elemento do vetor de parâmetros acústicos da *i*-ésima janela da locução, $\Delta_i(n)$ é o *n*-ésimo vetor *delta* correspondente ao vetor de parâmetros $y_i(n)$ calculado no *i*-ésimo quadro e K representa o número de quadros adjacentes empregado no cálculo dos parâmetros diferenciais.

3.2.2 Módulos de Treinamento e Reconhecimento

A maioria dos sistemas de reconhecimento de fala atuais utiliza os Modelos Ocultos de Markov (HMM) ou redes neurais artificiais (ANN). Neste trabalho é utilizado o HMM. Um HMM é um processo estocástico duplo, com um processo estocástico subjacente que não é observável (está oculto) e pode ser visto apenas por meio de outro conjunto de processos estocásticos que produzem uma sequência de símbolos observáveis [20]. Quando aplicado ao reconhecimento de fala, o processo oculto modela a variabilidade temporal, enquanto o processo que produz a sequência de observáveis modela a variação espectral [21].

Um grupo de modelos probabilísticos elementares de subunidades fonéticas de palavras (por exemplo: fonemas, difones ou trifones) é usado para fazer representações das palavras. A sequência de parâmetros acústicos extraída de uma locução é vista como a realização de uma concatenação de processos elementares descritos pelos HMMs [22].

O módulo de treinamento tem por função treinar os modelos HMM das subunidades acústicas a partir de locuções de treinamento e das respectivas transcrições fonéticas [23]. O algoritmo de treinamento comumente utilizado é o *Baum-Welch* [20][24][25].

O módulo de reconhecimento compara o sinal de voz parametrizado com os modelos acústicos previamente treinados e decide qual foi a palavra pronunciada. Pode também fazer uso do modelo de linguagem (se houver) para melhorar a taxa de

acertos. Nesta etapa são fornecidos a locução a ser reconhecida, os modelos HMM das subunidades acústicas e o vocabulário com o universo das palavras que podem ser reconhecidas. Os modelos HMM das subunidades são gerados pelo módulo de treinamento a partir das locuções de treinamento [23]. Utiliza-se também o algoritmo de *Viterbi* [21] para determinar qual o conjunto de HMMs que tem maior probabilidade de ter gerado os dados acústicos oriundos do processo de extração de características.

Capítulo **4**

Normalização de Locutor

4.1 Introdução

Conforme visto no Capítulo 2, uma fonte importante de variabilidade entre os locutores em reconhecimento de fala é a variação do comprimento do trato vocal. O trato vocal possui diferentes formas e comprimentos para cada pessoa, resultando em locuções com diferentes características acústicas. As posições dos picos espectrais para formantes de um determinado som são inversamente proporcionais ao comprimento do trato vocal [26].

As diferenças nos comprimentos de trato vocal se manifestam como escalonamento do espectro para o mesmo som [26]. Desse modo, esta variabilidade é um dos principais responsáveis pela degradação de desempenho em sistemas de reconhecimento de fala. Para minimizar esta variabilidade, podem ser aplicadas técnicas de normalização de locutor. A normalização de locutor não deve ser confundida com a adaptação de locutor. Na primeira, todos os locutores analisados são normalizados em relação a um locutor médio, enquanto na segunda, o treinamento é realizado para um locutor em particular, aquele para o qual o sistema será adaptado [14]. Neste capítulo são discutidos os principais aspectos da técnica de normalização de comprimento do trato vocal (VTLN).

4.2 Normalização de comprimento do trato vocal (VTLN)

A maioria dos sistemas “estado da arte” atualmente incorpora a técnica VTLN para reduzir a variabilidade entre locutores e consequentemente melhorar o desempenho do reconhecimento de fala [27]. A VTLN realiza uma normalização com o intuito de reduzir as variabilidades nos espectros dos sinais de fala que surgem devido a diferenças no comprimento do trato vocal (VTL) dos locutores [27]. O processo de normalização tenta fazer com que todos os locutores tenham tratos vocais de mesmo comprimento, normalizando-os para um comprimento médio. Essas funções são obtidas através da compressão ou ampliação do espectro da fala e são normalmente referidas como escalonamento [27]. A minimização das diferenças do trato vocal é realizada estimando um fator de escalonamento ótimo que será responsável pelo escalonamento do banco de filtros.

4.3 Escalonamento do banco de filtros

A normalização do comprimento do trato vocal é realizada a partir do escalonamento das frequências centrais do banco de filtros. O escalonamento destas frequências é realizado linearmente por um fator de escalonamento α. Este fator é fisicamente representado pela razão entre o comprimento do trato vocal do locutor sendo analisado e o comprimento do trato vocal utilizado como referência [26]. Em locutores masculinos o trato vocal mede cerca de 17 cm e pode chegar a 19 cm ou mais, enquanto em locutores femininos é de cerca de 14,5 cm e em crianças mede cerca de 12 cm [2][27][26].

A faixa de valores para a busca do fator de escalonamento segue a razão explicada anteriormente. Para locutores masculinos, pela proporção de comprimento de trato vocal, a faixa dos fatores varia de 0,85 (resultado da razão entre 14,5/17, sendo 17 cm o comprimento do trato vocal mínimo observado em locutores masculinos e 14,5 cm encontrado em locutores femininos) a cerca de 1,11 (resultado da razão 19/17, sendo 19 cm o comprimento do trato vocal máximo observado em locutores masculinos) [27][26]. Entretanto, se utilizarmos locutores crianças para teste, sabendo que o comprimento do trato vocal em média é 12 cm, a faixa dos

fatores de escalonamento terá que ser diferente. Neste caso, o valor mínimo da faixa dos fatores de escalonamento será 0,70 (da razão 12/17) [2][27]. Este fator é utilizado neste trabalho como mínimo em todos os conjuntos de treinamento, a fim de aumentar a faixa de exploração na busca do fator de escalonamento ótimo.

Determinada a faixa dos fatores de escalonamento, o novo banco de filtros com frequências escalonadas é obtido através de:

$$f' = \beta . f, \tag{4.1}$$

onde f representa a frequência original na escala *mel*, α representa o fator de escalonamento, β representa o fator de escalonamento em frequência $\frac{1}{\alpha}$ para α (variando entre 0,70 e 1,12) e *f'* representa a frequência escalonada.

Dependendo do fator de escalonamento (α) utilizado, as frequências no banco de filtro serão expandidas se ($\alpha < 1$) ou serão comprimidas se ($\alpha > 1$).

Os valores de *f'* são encontrados variando os fatores de escalonamento α em passos de 0,02 na equação (4.1) [26].

A Tabela 4.1 apresenta os valores de *f'* para α = 0,70, α = 1,00 e α = 1,12. Estes fatores representam os extremos de escalonamento utilizados neste trabalho.

Tabela 4. 1: *Valores da frequência central do banco de filtros, em Hz, para: $\alpha = 0,70$, $\alpha = 1,00$ e $\alpha = 1,12$.*

Índice dos Filtros	α = 0,7	α = 1,0	α = 1,12
1	142,86	100	89,29
2	285,71	200	178,57
3	428,57	300	267,86
4	571,43	400	357,14
5	714,29	500	446,43
6	857,14	600	535,71
7	1000	700	625
8	1142,86	800	714,29
9	1285,71	900	803,57
10	1428,57	1000	892,86
11	1641,43	1149	1025,89
12	1885,71	1320	1178,57
13	2165,71	1516	1353,57
14	2487,14	1741	1554,46
15	2857,14	2000	1785,71
16	3281,43	2297	2050,89
17	3770	2639	2356,25
18	4330	3031	2706,25
19	4974,29	3482	3108,93
20	5714,29	4000	3571,43
21	6564,29	4595	4102,68
22	7540	5278	4712,50
23	8661,43	6063	5413,39
24	9948,57	6964	6217,86

Como pode ser observado na Figura 4.1, o processo de escalonamento é realizado diretamente no banco de filtros, é possível gerar diferentes conjuntos de coeficientes *mel-cepstrais* (MFCC) usando apenas uma FFT para cada segmento de fala [26].

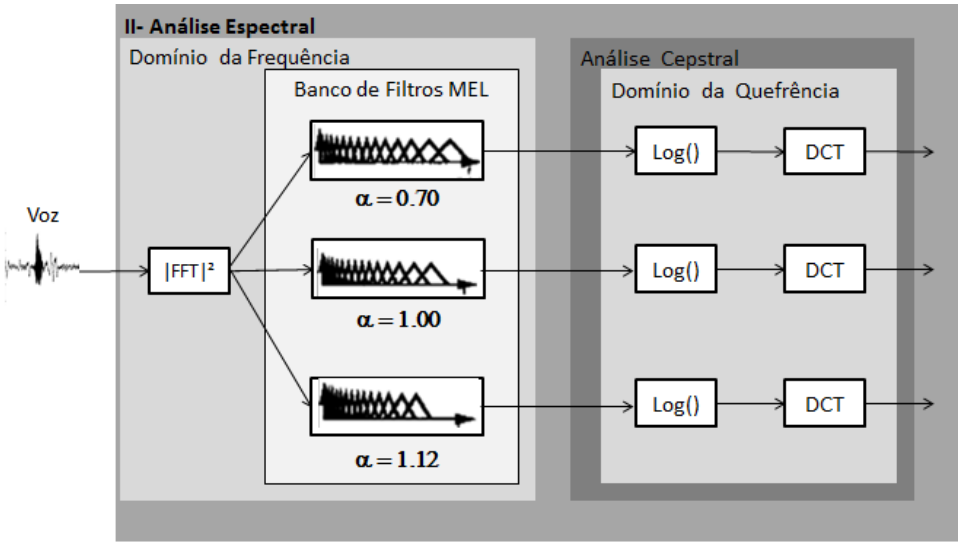

Figura 4. 1: *Análise dos bancos de filtros mel com escalonamento.*

O fator de escalonamento que leva ao melhor desempenho do sistema é chamado de fator de escalonamento ótimo ou α–ótimo, e o processo de escolha deste é detalhado a seguir.

4.4 Fator de escalonamento ótimo (α–ótimo)

O fator de escalonamento ótimo, $\hat{\alpha}_i$, para cada locutor *i*, é obtido pela máxima probabilidade de se obter um conjunto de características acústicas, dado um modelo λ, isto é:

$$\hat{\alpha}_i = \arg\max_{\alpha} \mathrm{P}\left(\mathrm{O}_{\mathrm{i,j}}^{\alpha} \mid \lambda\right), \tag{4.2}$$

onde $\mathrm{O}_{\mathrm{i}}^{\alpha} = \left\{\mathrm{O}_{\mathrm{i,1}}^{\alpha}, \mathrm{O}_{\mathrm{i,2}}^{\alpha}, \ldots, \mathrm{O}_{\mathrm{i,N}}^{\alpha}\right\}$ representa o conjunto de características acústicas de todas as *N* locuções do locutor *i*, escalonada de α. λ representa o modelo estatístico treinado por uma grande população de locutores.

A estimação direta de $\hat{\alpha}$ é difícil pelo fato de corresponder a uma transformação não linear dos parâmetros característicos da fala. Por isso o fator α–ótimo é obtido pela busca em uma faixa de valores.

Para a obtenção da métrica em (4.2), são em geral utilizados modelos ocultos de Markov (*hidden Markov models*, HMMs). Como este trabalho propõe a substituição destes por modelos de misturas Gaussianas (*Gaussian mixture models*, GMMs), ambos são descritos nos capítulos a seguir.

Capítulo **5**

Métodos Estatísticos: HMM e GMM

5.1 Modelos ocultos de Markov (HMM)

Uma série de teorias acerca dos HMM foi fundamentada no final da década de 60 por Leonard E. Baum. No começo dos anos 70, F. Jelinek da IBM e J.K. Baker da CMU foram os pioneiros na utilização do HMM para reconhecimento de fala. Ainda hoje, o HMM é uma das principais ferramentas utilizadas em estudos de sistemas de reconhecimento de fala [27][26][5].

O HMM pode ser entendido como uma máquina de estados finitos conectados. Cada estado está conectado ao outro através de uma probabilidade de transição e cada estado possui também uma probabilidade de permanência. Atrelada a cada estado existe também uma probabilidade de emissão de símbolos observáveis. A cada transição a_{ij} ocorre a emissão de um símbolo, com uma probabilidade $b_j(O_t)$ formando uma sequência de símbolos observáveis. O que é oculto nos HMM, portanto, é a sequência de estados.

Em reconhecimento de fala pode-se relacionar os estados internos de um HMM com a evolução temporal do sinal acústico de fala, e os símbolos externos com o conjunto de possíveis observações (vetores acústicos).

Os HMM podem ser classificados na forma como modelam as probabilidades de emissão de símbolos $b_j(O_t)$,como discretos ou contínuos. Este trabalho trata de modelos contínuos, que possuem uma função densidade probabilidade (fdp)

Gaussiana multidimensional ou uma mistura delas. Cada fdp Gaussiana é representada por um vetor média μ e matriz covariância Σ. Nesta dissertação considera-se que a matriz de covariância é diagonal, o que significa que as variáveis aleatórias de cada dimensão são consideradas independentes entre si, pois o custo computacional é menor, uma vez que terá menos parâmetros a modelar [23].

Os HMMs podem ser classificados de acordo com sua topologia. Estas podem ser lineares, ergódicas, *Bakis* e *left-right* [28]. Aqui utiliza-se a topologia *left-right* representada na Figura 5.1. Nesta, à medida que o tempo aumenta, os índices dos estados aumentam ou permanecem os mesmos. Em reconhecimento de fala esta topologia é adequada, pois uma vez que o sinal de fala é um processo dinâmico e progressivo, as transições entre os estados do HMM irão ocorrer somente em um sentido [29].

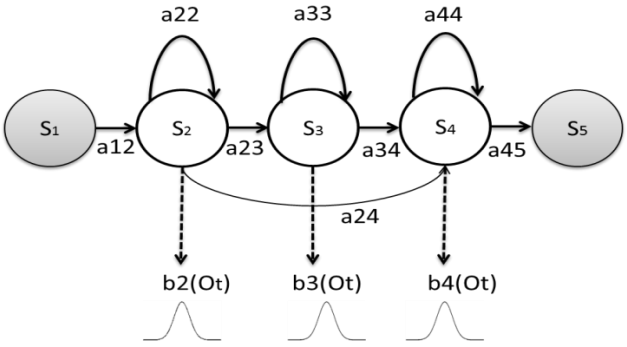

Figura 5. 1: *Estrutura de um HMM left-right de 5 estados.*

Um HMM é caracterizado pelos seguintes elementos:

- Um conjunto $S = \{S_1, S_2, S_3, \dots, S_N\}$ finito de estados, onde N é o número de estados.
- A distribuição da probabilidade de transição de estado $A = a_{ij}$, onde,

$$a_{ij} = P\left[q_{t+1} = S_j \,\middle|\, q_t = S_j\right] \qquad ,1 \le i, j \le N, \tag{5.1}$$

obedecendo às condições:

$$a_{ij} \geq 0 \tag{5.2a}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{5.2b}$$

- A distribuição da probabilidade do símbolo de observação no estado j, $B = \{b_j(O_t)\}$, onde,

$$b_j = P[O_t | \, q_t = S_j] \qquad ,1 \leq j \leq N \, , 1 \leq O_t \leq M \, , \tag{5.3}$$

e onde, O_t representa o símbolo de observação no tempo t.

- A distribuição do estado inicial $\pi = \{\pi_i\}$, onde,

$$\pi_i = P[q_1 = S_i] \qquad , 1 \leq i \leq N \tag{5.4}$$

Por conveniência utiliza-se uma notação compacta para indicar o conjunto de parâmetros completo do modelo HMM, denotado por λ, na expressão (5.5) [21]:

$$\lambda = (A, B, \pi) \tag{5.5}$$

5.2 Modelos de mistura Gaussiana (GMM)

Os vetores acústicos extraídos de um sinal de voz de um locutor podem ser modelados por modelos de mistura Gaussiana (GMM). O GMM consegue modelar qualquer tipo de distribuição de dados, alterando seus parâmetros de mistura [30]. Diferente do HMM, que pode modelar tanto as variabilidades espectrais como as temporais em reconhecimento de fala, o GMM modela somente as variabilidades espectrais. Desta forma, não é um modelo adequado para o problema de reconhecimento de fala, mas é adequado para o problema de reconhecimento de locutor.

Um GMM é um modelo que pode representar distribuições multimodais através da soma ponderada de *M* componentes de densidades Gaussianas [24].

$$P(O|\lambda) = \sum_{i=1}^{M} p_i b_i \, (\vec{O}) \tag{5.6}$$

onde, $\vec{O}$ é um vetor aleatório *D*-dimensional, $b_i(\vec{O})$, $i = 1,2,\dots,M$, são as densidades componentes, p_i são os pesos das misturas, e λ é o modelo correspondente as características acústicas do locutor.

Cada densidade componente é uma função Gaussiana *D*-dimensional com a forma:

$$b_i(\vec{O}) = \frac{1}{2\pi^{\frac{D}{2}}\,|\Sigma_i|^{\frac{1}{2}}} \exp\{-\frac{1}{2}\,(\vec{O}-\vec{\mu}_i)'\sum_{i}^{-1}(\vec{O}-\vec{\mu}_i)\}\,, \tag{5.7}$$

onde, $\vec{\mu}_i$ é um vetor média e Σ_i é a matriz covariância.

Aqui é utilizada a matriz covariância diagonal, pelas mesmas justificativas abordadas no HMM. Os pesos da mistura satisfazem a restrição $\sum_{i=0}^{M} p_i = 1$.

Uma densidade de mistura Gaussiana completa é parametrizada pelo vetor média, pela matriz de covariância e pelos pesos das misturas de todas as densidades componentes. Estes parâmetros são coletivamente representados pela notação λ, representando o modelo GMM.

$$\lambda = \{, \vec{\mu}_i, \Sigma_i, p_i\} \qquad i = 1, \dots, M \tag{5.8}$$

Os parâmetros p, $\vec{\mu}$ e Σ das distribuições de um GMM devem ser ajustados a um conjunto de vetores acústicos de modo a obter a máxima probabilidade de observação dado um modelo GMM, λ. Um método utilizado para esse ajuste é o algoritmo *Expectation-Maximization* (EM) [24]. A idéia básica do algoritmo EM é começar com um modelo inicial λ para estimar um novo modelo $\bar{\lambda}$ de modo que $P(O|\bar{\lambda}) \geq P(O|\lambda)$. O novo modelo então torna-se o modelo inicial para a próxima iteração e o processo é repetido até que algum limite de convergência seja atingido. Esta é a mesma técnica básica utilizada para estimar os parâmetros HMM, via algoritmo de reestimação *Baum-Welch* [24][25].

Em cada iteração EM, as fórmulas de reestimação a seguir são utilizadas, as quais garantem um aumento monotônico no valor da máxima probabilidade de observação dado o modelo.

Pesos da mistura:

$$\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{O}_t, \lambda) \tag{5.9}$$

Média:

$$\overrightarrow{\bar{\mu}_i} = \frac{\sum_{t=1}^{T} p(i|\vec{O}_t, \lambda)\vec{O}_t}{\sum_{t=1}^{T} p(i|\vec{O}_t, \lambda)} \tag{5.10}$$

Variância:

$$\overrightarrow{\bar{\sigma}_i} = \frac{\sum_{t=1}^{T} p(i|\vec{O}_t, \lambda)\overrightarrow{O^2}_t}{\sum_{t=1}^{T} p(i|\vec{O}_t, \lambda)} - \mu_i^2 \tag{5.11}$$

Desta forma, dois fatores críticos ao treinar um GMM são a seleção da quantidade de misturas e a inicialização dos parâmetros do modelo no início do algoritmo EM. A inicialização dos parâmetros é feita utilizando o algoritmo *Segmental K-Means* [31].

Neste trabalho procurou-se encontrar a quantidade ótima de Gaussianas nas misturas para representar o sinal de voz dos locutores, pois isto influencia significativamente na qualidade do modelo. Um cuidado a ser levado em conta é não aumentar muito o número de Gaussianas, pois nesse caso os seus parâmetros passam a ser mal estimados e o desempenho do sistema diminui.

Capítulo **6**

Processos de busca do fator de escalonamento ótimo (α-ótimo) utilizando HMM e GMM

Conforme descrito no Capítulo 4, o método VTLN requer o fator de escalonamento ótimo (α-ótimo) para realizar o escalonamento do banco de filtros. Neste capítulo é descrito o processo de busca que visa encontrar o melhor fator α para cada locutor utilizando os métodos estatísticos HMM e GMM descritos no Capítulo 5.

6.1 Processo de busca do α-ótimo utilizando o método HMM

O primeiro passo deste processo de busca é a obtenção de um HMM treinado a partir de vários locutores, e que pode ser considerado um modelo independente de locutor. Em seguida faz-se a busca pelo fator ótimo de escalonamento (α-ótimo) por varredura para cada locutor. Aqui normaliza-se o comprimento do trato vocal de crianças em relação a um locutor adulto. Para isso, foram elencados 50 locutores crianças (25 meninos e 25 meninas) do conjunto de testes da base de dados TIDIGITS [33].

A ferramenta utilizada para a busca do α-ótimo utilizando o método HMM foi o "*Hidden Markov Model Toolkit"* (HTK) [32].

Seguindo o estudo de [14] acerca da quantidade de locuções necessárias para serem utilizadas na busca do α-ótimo, foram utilizadas 4 locuções de cada locutor

para a busca. Para o método HMM são necessárias as transcrições fonéticas das locuções utilizadas na busca pelo valor ótimo de α. As locuções proferidas pelo locutor foram escolhidas nesse trabalho de modo que contivessem todos os dígitos disponíveis na base de dados TIDIGITS, e para isso procurou-se utilizar locuções de no máximo 3 dígitos conectados para todos os locutores.

Portanto, conforme ilustrado na Figura 6.1, na escolha do α-ótimo a partir do método HMM são levados em consideração o modelo HMM pré-treinado utilizado como referência, as locuções *X* e suas respectivas transcrições fonéticas *W*, proferidas pelos locutores *i* elencados para análise.

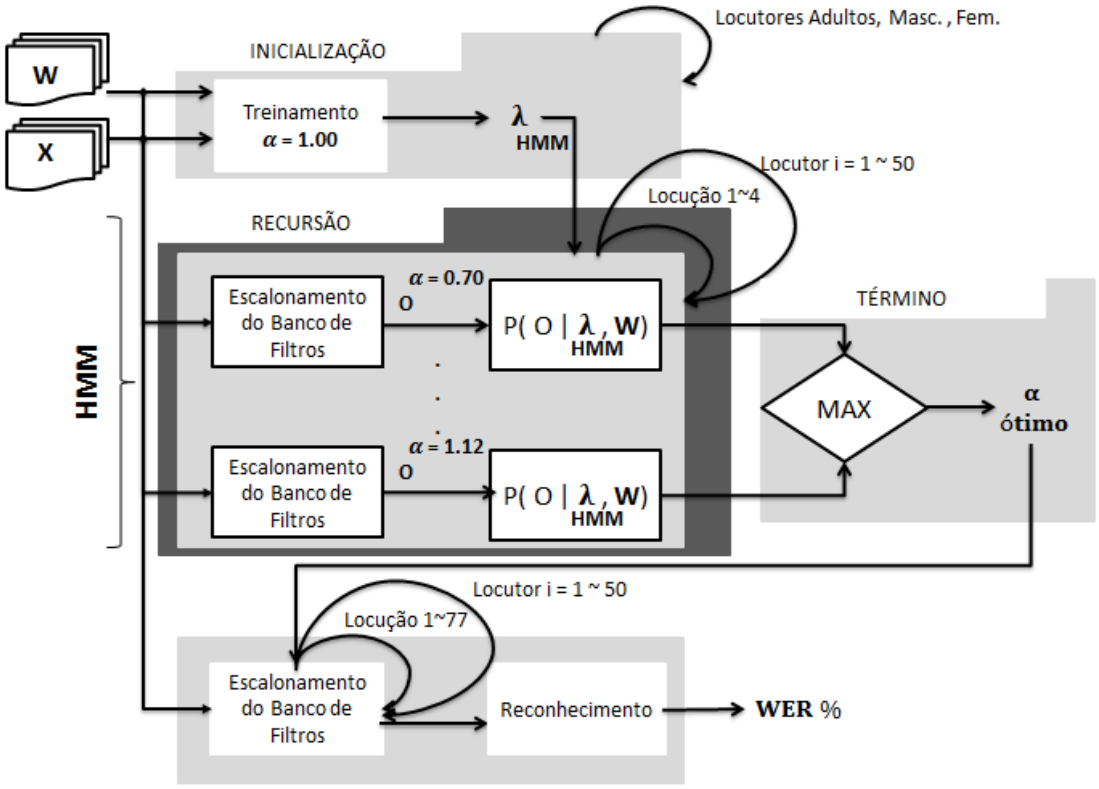

Figura 6. 1: *Escolha do α-ótimo a partir do método HMM.*

O Algoritmo a seguir, foi utilizado na busca do α-ótimo pelo método HMM:

Inicialização:

Inicialmente é criado um modelo sem a utilização da normalização. Este será chamado de λ_{HMM}.

Recursão:

Para essa etapa são definidas as seguintes variáveis: α é o fator de escalonamento (0,70 ≤ α ≤ 1,12), $\{W_i\}$ o conjunto de transcrições fonéticas referentes ao locutor *i*, $\{O_i^\alpha\}$ o conjunto de características espectrais observadas após o escalonamento do filtro *mel* por um fator α, para o locutor *i*, e max $[P(O_i^\alpha|\lambda_{HMM}, W_i)]$ a máxima probabilidade de se obter um conjunto de observação O_i, escalonado de α, dado um modelo λ_{HMM} e um conjunto de transcrições W_i.

• Para cada locutor *i* criança elencado, (*i* = 1 até 50), é realizada uma busca na faixa dos fatores de escalonamento (α), aqui utilizado entre 0,70 e 1,12 (a busca é realizada com espaçamento de 0,02 conforme [26], resultando em 22 fatores α).

• Cálculo de max $[P(O_i^\alpha|\lambda_{HMM}, W_i)]$, máxima probabilidade de observação entre os 22 fatores α. Utiliza-se para este cálculo o Algoritmo de *Viterbi*.

• Para cada α testado é observada sua máxima probabilidade de observação para as 4 locuções.

Término:

• É calculada a mediana da max $[P(O_i^\alpha|\lambda_{HMM}, W_i)]$ para as 4 locuções dos locutores *i*, realizando o mesmo processo para os 22 fatores α.

• O α-ótimo para um determinado locutor será aquele que obtiver a maior probabilidade de observação, max $[P(O_i^\alpha|\lambda_{HMM}, W_i)]$, dentre os 22 fatores α explorados na busca.

• É guardado o fator α que corresponda ao α-ótimo de cada locutor *i*.

Por fim, uma vez encontrados os α-ótimos de todos os locutores em análise, é realizada a extração de característica de todas as locuções proferidas pelos locutores. Desse modo, o filtro na escala *mel* será escalonado para cada locutor de acordo com o seu fator de escalonamento ótimo.

6.2 Processo de busca do α-ótimo utilizando o método GMM-UBM

O segundo método utilizado para a busca do α-ótimo utilizado nesse trabalho é o baseado em modelo de misturas Gaussianas (GMM). Assim como foi feito inicialmente para o método HMM, tem-se a necessidade também nesse processo de um modelo pré-treinado λ para se aplicar o critério de máxima probabilidade de observação dado o modelo, $P(O|\lambda)$. Para isso é utilizado o modelo de base universal (*Universal Background Model*, UBM). Um UBM é um GMM treinado por uma grande quantidade de locutores, pretendendo ser uma representação de toda a variabilidade fonética encontrada para todos os locutores possíveis. As ferramentas utilizadas para o treinamento do UBM e para a verificação utilizando o GMM são os softwares desenvolvidos em Linguagem C pelo professor Carlos Alberto Ynoguti. A aplicação deste método foi dividida em duas etapas: o treinamento do UBM e a verificação do $P(O|\lambda)$ utilizando o GMM.

Treinamento do UBM:

O treinamento do UBM é realizado para todos os locutores adultos de treino. É inicialmente definida a dimensão dos vetores características, neste trabalho foram utilizados 39 parâmetros conforme foi explicado no Capítulo 3.

O próximo passo da configuração é estipular o número de Gaussianas utilizadas no modelo para caracterizar o espectro do sinal de voz dos locutores de treinamento. Foram realizados 8 treinamentos distintos com: 1 Gaussiana (unimodal) e 2, 4, 8, 16, 32, 64 e 128 Gaussianas na mistura. A justificativa para isso é encontrar o número ótimo de Gaussianas para este cenário e posteriormente comparar com as mesmas misturas utilizadas no método HMM. Cada Gaussiana inicial foi configurada com valor de variância σ = 10, média μ = 1 e o peso da mistura p = 1.

A inicialização do treinamento é realizada através do Algoritmo *Segmental K-Means*. Maiores detalhes podem ser encontrados em [23] [31].

O treinamento do modelo UBM é realizado através do cálculo de $P(O|\lambda)$ para os conjuntos de locutores. São realizados re-treinamentos até que se obtenha uma distorção estipulada em 0.0001 (valor de parada definido previamente) calculada através da seguinte equação:

$$d = \frac{-(P(O|\lambda)_{novo} - P(O|\lambda)_{anterior})}{P(O|\lambda)_{novo}} \quad (6.1)$$

O processo continua até que não haja variação substancial de uma iteração para outra.

Após isso, são gerados os dados de treinamento do UBM distintos para cada Gaussiana na mistura. Estes dados serão os modelos UBM pré-treinados, utilizado na etapa a seguir.

Verificação de $P(O|\lambda)$ utilizando o GMM:

A próxima etapa é a aplicação do método proposto por [26], do mesmo modo que o algoritmo descrito na seção 6.2 realizado para o HMM, porém com algumas modificações para utilização do modelo GMM.

A inicialização terá modelo UBM pré-treinado (λ_{UBM}) ao invés do modelo HMM pré-treinado (λ_{HMM}).

A recursão é realizada através da max $[P(O_i^{\alpha}|\lambda_{UBM})]$, isto é, da máxima probabilidade de se obter um conjunto de observação O_i, escalonado de α, dado um modelo λ. Nesse caso a diferença crucial entre os métodos HMM e GMM-UBM é o fato do GMM não necessitar das transcrições fonéticas, o que o torna mais simples. Para o cálculo da max $[P(O_i^{\alpha}|\lambda_{UBM})]$, é utilizado o software "Verificação GMM" fornecido pelo professor Carlos Alberto Ynoguti, configurado da seguinte forma: são carregadas as locuções parametrizadas dos locutores *i*=1 até 50. A parametrização foi realizada do mesmo modo que discutido anteriormente para o treinamento do UBM e HMM.

O fator α-ótimo para um determinado locutor será aquele que obtiver a maior $P(O_i^{\alpha}|\lambda_{UBM})$, dentre os fatores de escalonamento explorados.

O diagrama esquemático do processo de busca, utilizando o método GMM-UBM e as locuções *X* de cada locutor *i* elencado para análise é apresentado na Figura 6.2.

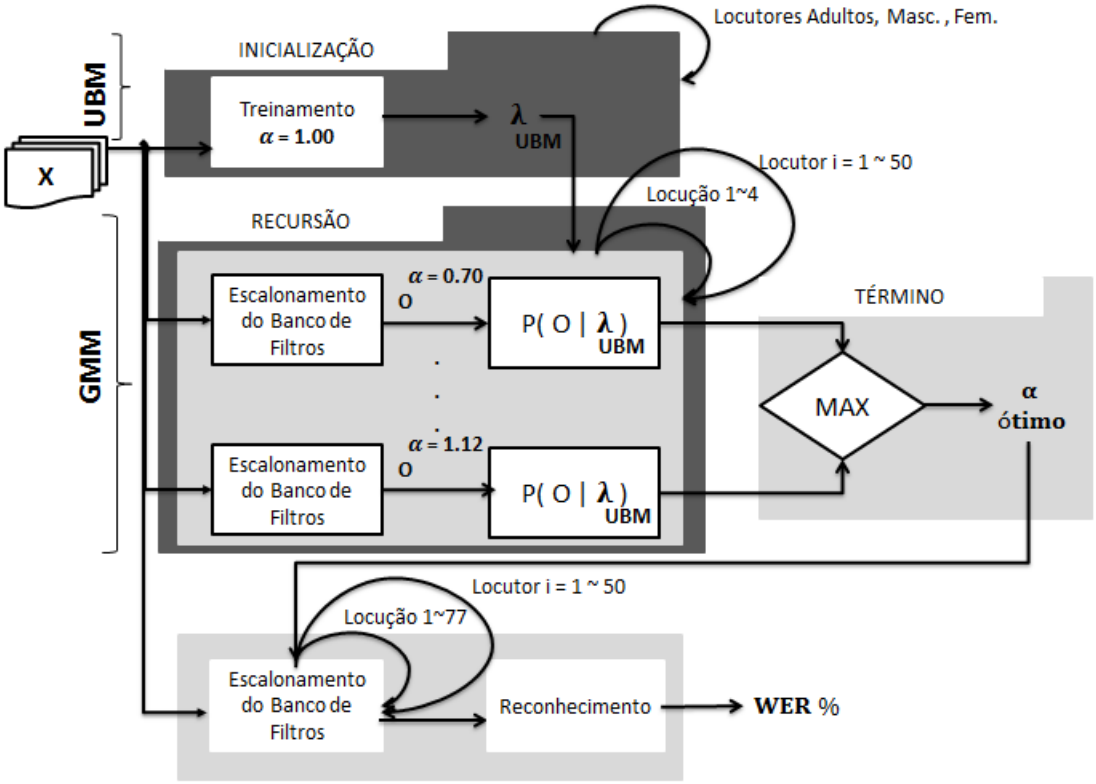

Figura 6. 2: *Escolha do α-ótimo a partir do método GMM-UBM.*

Por fim, uma vez encontrado os α-ótimos de todos os locutores envolvidos na análise, é realizada a extração de característica das locuções para cada locutor de acordo com o α-ótimo encontrado para ele, do mesmo modo como realizado para o HMM.

Capítulo 7

Aparato Experimental

Neste capítulo serão apresentadas a base de dados, as configurações, mecanismos do sistema e os procedimentos para obtenção da taxa de acerto do reconhecimento de fala.

7.1 Base de Dados utilizada

A base de dados TIDIGITS [33] é utilizada nos experimentos realizados. A TIDIGITS tem o propósito de avaliar algoritmos para o reconhecimento de sequência de dígitos conectados independente de locutor [34]. Este *corpus* contém locuções de dígitos no idioma inglês-americano [5]. As locuções são lidas por 325 locutores, sendo eles: 111 homens, 114 mulheres, 50 meninos e 50 meninas. Os locutores são divididos em conjuntos de testes e treino.

A base de dados utilizando locutores adultos possui para treinamento um subconjunto de 112 locutores (55 homens e 57 mulheres) e para teste um subconjunto de 113 locutores (56 homens e 57 mulheres). A base de dados utilizando locutores crianças possui um subconjunto de 50 locutores (25 meninos e 25 meninas) e para teste um subconjunto de 50 locutores (25 meninos e 25 meninas).

Cada locutor na base de dados profere 77 locuções divididas da seguinte forma:

- 11 dígitos isolados (*zero, oh, one, two, three, four, five, six, seven, eight* e *nine*), cada dígito repetido duas vezes;
- 11 sequências de 2, 3, 4, 5 e 7 dígitos.

O *corpus* foi coletado com uma taxa de amostragem de 20 kHz, 16 bits de resolução, em um ambiente silencioso [33][35]. Para os experimentos todas as locuções foram ajustadas para uma taxa de amostragem de 8 kHz, valor utilizado na maioria dos trabalhos na área para fins de comparação de resultados, por exemplo [26][36][37].

7.2 Extrator dos parâmetros acústicos

O sinal de fala foi parametrizado por meio de 39 coeficientes *mel-cepstrais* (MFCC), sendo estes: 12 parâmetros *mel-cepstrais*, 12 derivadas primeira (*delta-mel-cepstrais*) e 12 derivadas segunda (*delta-delta-mel-cepstrais*) dos parâmetros *mel-cepstrais*, 1 parâmetro de energia, 1 derivada primeira (*delta-energia*) e 1 derivada segunda (*delta-delta-energia*) do parâmetro de energia [38][39].

Para o cálculo dos coeficientes foram utilizadas janelas com uma duração de 25 ms, atualizadas a cada 10 ms. Um filtro de pré-ênfase realizado por um filtro passa-altas de primeira ordem com função de transferência:

$$H_{pre}(z) = 1 - 0{,}97 \, . \, z^{-1} \tag{7.1}$$

Foi utilizada uma janela de *Hamming*, aplicadas ao sinal antes do cálculo dos parâmetros. Por fim, foi utilizado um banco de filtros passa-faixa triangular na escala *mel*, com 24 filtros.

7.3 Mecanismo do reconhecimento de fala

Foi utilizada a ferramenta HTK [32], desenvolvida pela Universidade de Cambridge como mecanismo de reconhecimento de fala.

Cada um dos dígitos das locuções foi modelado como uma concatenação de subunidades fonética, retiradas do dicionário fonético ARPAbet [40][41]. O sistema utiliza modelos dependentes do contexto, os trifones, que contêm informação acerca dos fones vizinhos. Esta informação contextual leva em geral a melhores resultados no reconhecimento [42]. As subunidades fonéticas são modeladas através de um HMM de 5 estados com arquitetura "*left-right*". Este modelo utiliza 5 estados no total, sendo o estado de entrada (S_1) e o de saída (S_5) não emissores, ou seja sem função de observação. Os 3 estados restantes (S_2, S_3, S_4) são emissores, considerados estados ativos [43]. Essa forma é utilizada para facilitar a construção de modelos compostos e a união entre modelos [32]. A utilização dessa topologia para o modelo HMM é compatível com os exemplos do HTKbook [32] e com trabalhos da área, por exemplo, [27][5][44][45]. A representação da topologia utilizada pode ser vista na Figura 5.1.

7.4 Processo de treinamento dos HMMs

Após especificadas as configurações é iniciado o treinamento, que consiste basicamente na atualização dos valores dos parâmetros do modelo [46]. Este processo consta de duas fases: i) Inicialização do modelo e ii) Aplicação do algoritmo de *Baum-Welch* [20]. Através da ferramenta HTK, para a inicialização do modelo, é definida a média e a variância de cada componente Gaussiana para ser igual à média e variância global dos dados de treinamento. Após isso, são realizados os ajustes do modelo de silêncio, introdução do modelo para pausas curtas e realinhamento dos dados. Em cada uma destas tarefas é realizado o refinamento do modelo. Isto é feito através da reestimação dos parâmetros dos mesmos através do algoritmo de *Baum-Welch*. É necessário proceder a reestimação até que sejam suficientemente robustos. Neste trabalho foram realizadas 3 reestimações. Depois de realizadas as iterações de reestimação para uma Gaussiana, são inseridas mais Gaussianas nas misturas, desta forma foram realizados vários testes variando o número de Gaussianas. Para o teste, utiliza-se o algoritmo de *Viterbi* [21] para determinar qual o conjunto de HMMs, que tem maior probabilidade de ter gerado os dados acústicos em análise. Para avaliação do modelo, é realizada uma comparação das transcrições obtidas no processo de reconhecimento, com as transcrições

fonéticas inseridas manualmente no sistema, obtendo informações acerca da porcentagem de palavras corretamente reconhecidas, da taxa de erro, precisão, deleções, substituições e inserções. A Figura 7.1 apresenta o processo completo de treinamento de um HMM e teste utilizando a ferramenta HTK.

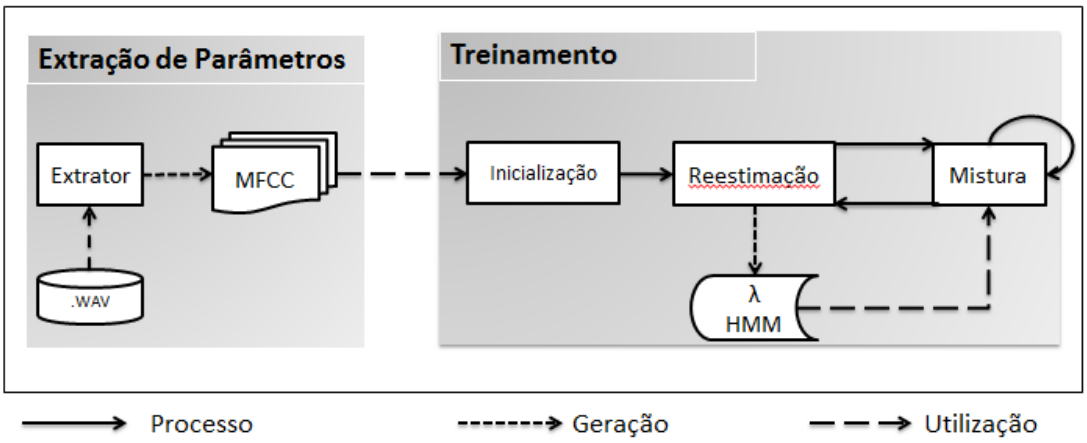

Figura 7. 1: *Processo de treinamento HMM utilizando a ferramenta HTK.*

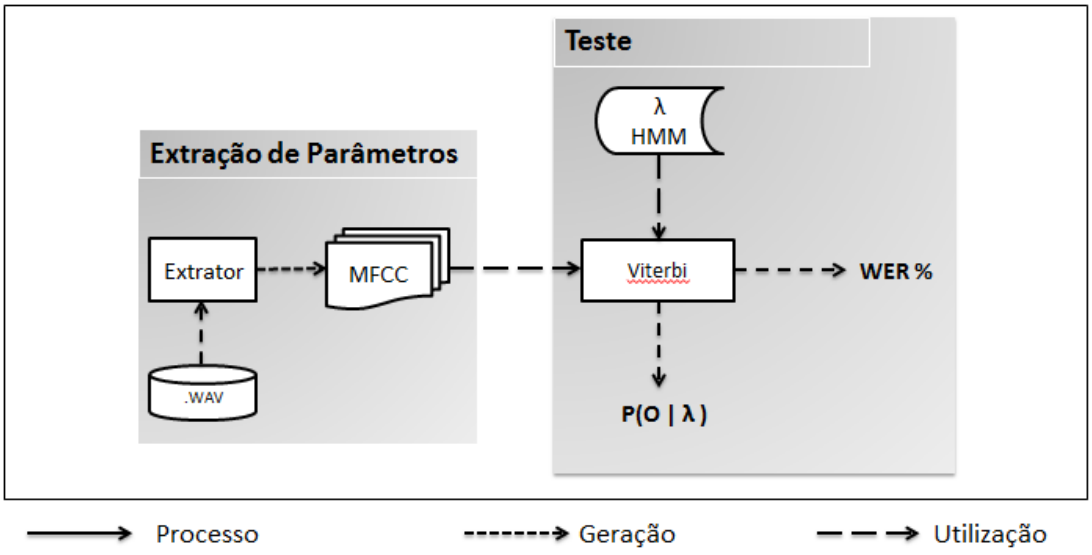

Figura 7. 2: *Processo de teste HMM utilizando a ferramenta HTK.*

Com as configurações e processos descritos nesse capítulo para o reconhecimento de fala, são obtidos os resultados dos testes realizados descritos a seguir.

Capítulo **8**

Resultados Experimentais

Neste capítulo serão apresentados os testes realizados, seus respectivos resultados e a análise dos mesmos. O cenário deste trabalho é o de um reconhecimento de fala treinado por locutores adultos e utilizado com locutores crianças, com o objetivo de medir a melhoria que a técnica de normalização de locutor pode fornecer a tal sistema. Tendo em vista que as aplicações envolvendo locutores crianças podem ir desde sistemas de aprendizado rodando em computadores pessoais (com grandes recursos disponíveis) a brinquedos movidos à pilha (com uma quantidade bastante limitada de recursos), foram realizados vários testes variando-se o número de Gaussianas nas misturas do HMM, para obter uma visão mais ampla do efeito da normalização. Os experimentos foram realizados para três conjuntos: Treinamento com locutores adultos e teste com locutores crianças, treinamento com locutores masculinos e teste com crianças, treinamento com locutores femininos e teste com crianças. Estes conjuntos foram testados com crianças com e sem normalização. Os testes com locutores crianças sem normalização resultou na *baseline* que servirá de comparação com o sistema utilizando crianças normalizado. O cálculo da taxa de erro de palavra para medir o desempenho do sistema de reconhecimento de fala é obtido a partir da ferramenta HTK, através da equação (8.1), levando em consideração as N palavras de referência e as S substituições, D deleções e I inserções na saída do reconhecimento de fala [47].

$$WER = 100 \,.\frac{(S + D + I)}{N}\ \% \tag{8.1}$$

Além da taxa de erros no reconhecimento do sistema, neste capítulo foi realizada uma análise da distribuição dos α-ótimos encontrados para cada locutor utilizando os métodos GMM e HMM. Também foi analisada a curva de máxima $P(O|\lambda)$ por fator de escalonamento. Por fim são apresentados os resultados finais de desempenho e as discussões e comparações entre os métodos HMM e GMM-UBM na busca do α-ótimo para cada locutor.

8.1 Baseline

Foram realizados um total de 8 experimentos para cada conjunto de treinamento e teste na *baseline*, cada qual adicionando mais Gaussianas na mistura (em múltiplos de 2) para verificar e discorrer sobre o comportamento do desempenho dos sistemas e possibilitar a detecção da quantidade ótima de Gaussianas na mistura.

O principal aspecto a ser considerado na *baseline* é o valor do fator de escalonamento (α), fixado em α=1,00, valor padrão do HTK. Este valor representa um banco de filtros na escala *mel* sem escalonamento. O desempenho alcançado na *Baseline* com este valor permitirá uma comparação com os outros sistemas experimentados nesse trabalho, cujos valores de α serão os considerados ótimos decorrentes do processo de busca utilizando os métodos HMM e GMM-UBM para cada locutor que será normalizado. Os resultados encontrados para todas as Gaussianas nas misturas são apresentados nos gráficos a seguir.

O gráfico da Figura 8.1 representa a taxa de erro por Gaussiana na mistura quando realizado treinamento com adultos e testado com crianças sem normalização. A partir da curva traçada, é possível detectar o número ótimo de Gaussiana na mistura. Notou-se que até 64 Gaussianas, a taxa de erro do sistema diminui 2,74 %. Esse comportamento é justificável pelo fato da variabilidade entre locutores adultos ser maior, pois contém locuções de homens e de mulheres. Essas variabilidades representam diferenças significativas na parte orgânica dos locutores, podendo citar: comprimento do trato vocal, frequências formantes, entre outras. Por ser maior essa

variabilidade, mais Gaussianas na mistura no sistema podem ser úteis para representar o sinal de voz dos locutores adultos, no entanto quando utilizada 128 Gaussianas na mistura, o desempenho piorou. Uma possível razão é que com essa quantidade de Gaussianas na mistura os parâmetros passaram a ser mal estimados.

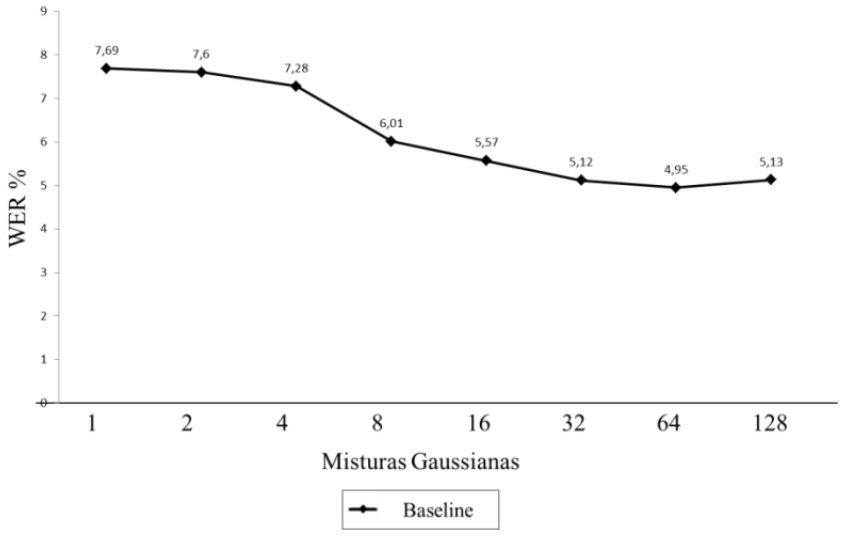

Figura 8. 1: *Curva de desempenho do sistema treinado com locutores adultos e testado com crianças sem normalização.*

A Figura 8.2 representa a taxa de erro por Gaussiana na mistura quando realizado o treinamento com locutores masculinos e testado com crianças sem normalização. Nesse experimento a taxa de erro é maior em relação aos outros experimentos, entre 35,62% utilizando Gaussiana unimodal e 38,46% com 64 e 128 Gaussianas na mistura. O motivo da alta taxa de erro se deve à grande diferença nas características orgânicas entre locutores masculinos e locutores crianças, como: o comprimento do trato vocal e as frequências formantes. Analisando o comportamento da curva é possível observar que a mistura ótima foi alcançada quando utilizadas 4 Gaussianas na mistura, com uma taxa de erro de 35,22%. A razão de se ter encontrado o ponto de mínimo com poucas Gaussianas na mistura pode ser devido ao fato de que a variabilidade acústica é menor entre os locutores homens, portanto para representar o sinal de voz dos locutores adultos, poucas Gaussianas são suficientes.

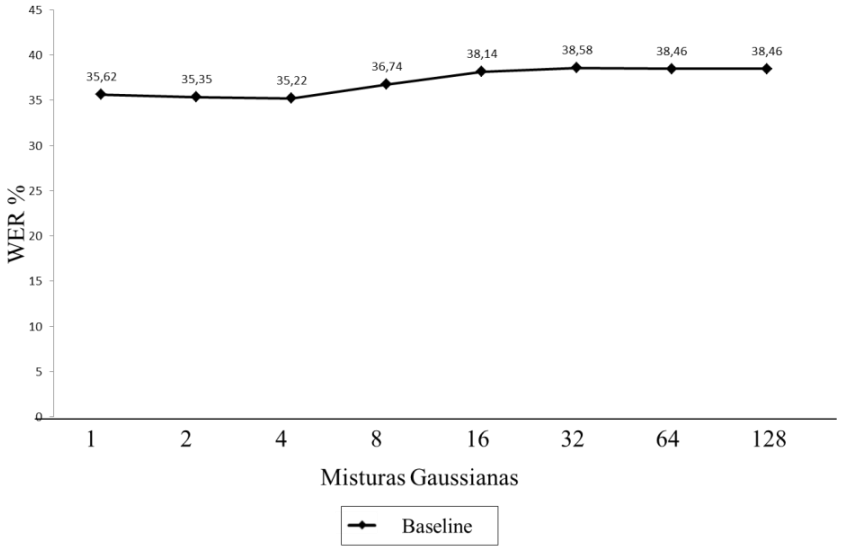

Figura 8. 2: *Curva de desempenho do sistema treinado com locutores masculinos e testado com crianças sem normalização.*

A Figura 8.3 representa o comportamento da taxa de erro por Gaussiana na mistura quando realizado treinamento com locutores femininos e testado com crianças sem normalização. Os valores das taxas de erro são relativamente baixos se comparados com as taxas de erros para sistemas treinados com locutores masculinos. A razão para isso está no fato de que as frequências formantes dos locutores femininos são muito próximas das frequências formantes dos locutores crianças. As taxas de erro ficaram entre 7,31 % utilizando Gaussiana unimodal e 3,87% no ponto de mínimo quando 32 Gaussianas foram adicionadas na mistura. A necessidade de utilizar muitas Gaussianas na mistura para atingir o ponto de mínimo, pode ter acontecido pelo fato de a variabilidade acústica ser maior entre os locutores femininos, por exemplo, alguns locutores femininos podem ter voz mais grave e outros mais agudos.

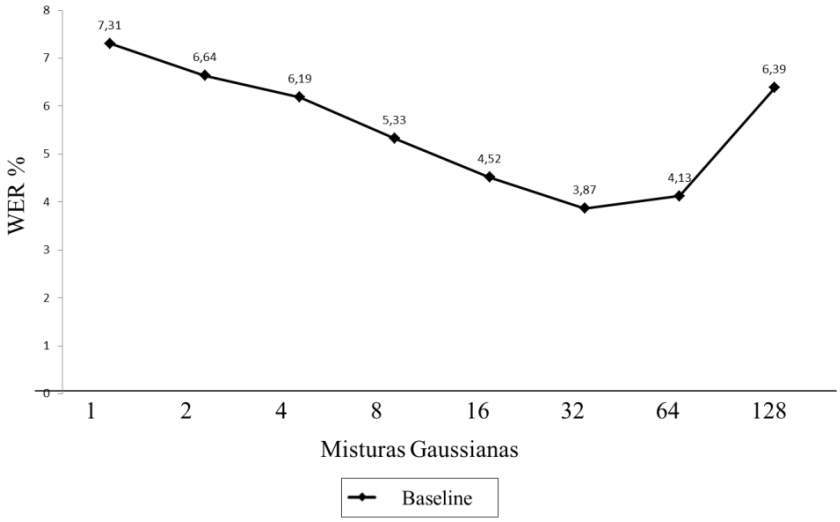

Figura 8. 3: *Curva de desempenho do sistema treinado com locutores femininos e testado com crianças sem normalização.*

A referência para validar o sistema utilizado e a *baseline* alcançada foi o trabalho de Mats Blomberg (KTH Estocolmo) publicado na FONETIK 2011 [5]. Nessa referência, foram utilizadas 32 Gaussianas na mistura por estado e uma configuração semelhante nos ajustes da ferramenta HTK. No gráfico da Figura 8.4 pode ser observada a comparação de desempenho do sistema entre a *baseline* publicada na FONETIK 2011 [5] com a *baseline* utilizada nesse trabalho.

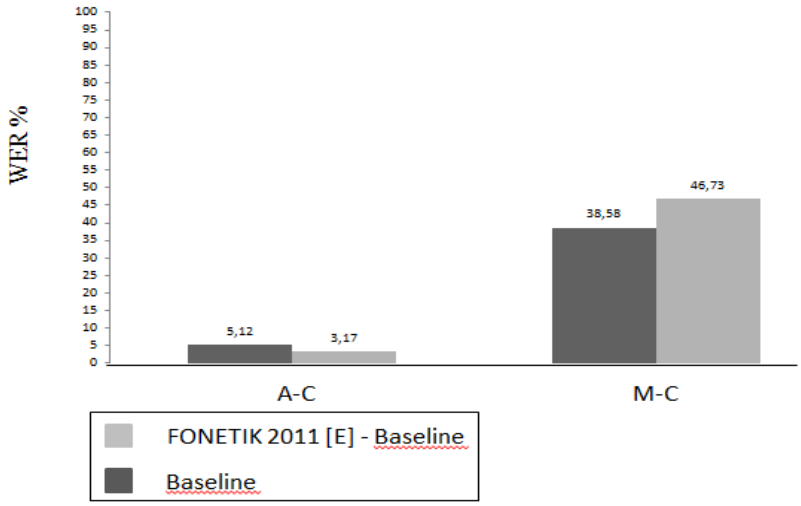

Figura 8. 4: *Gráfico comparativo da WER Baseline vs.WER [5].*

Utilizando um sistema treinado com adultos e testado com crianças, a taxa de erro encontrado por [5] foi 3,17 % ante 5,12 % encontrada na *baseline* deste trabalho. Para o sistema treinado com locutores masculinos e testado com crianças, a taxa de erro encontrado na *baseline* deste trabalho foi de 38,58%, uma taxa menor do que a *baseline* de [5], cuja taxa é 46,73%. Essas taxas de erro possibilitam o sistema de [5] validar o sistema utilizado nesse trabalho, pois não há uma discrepância significativa entre esses valores o que permite concluir que os experimentos ulteriores terão na *baseline* deste trabalho uma referência confiável.

8.2 Distribuição dos α-ótimos encontrados para cada locutor utilizando os métodos GMM-UBM e HMM.

Após a aplicação dos algoritmos que envolvem os processos de busca pelo fator de escalonamento ótimo (α-ótimo) utilizando os métodos HMM e GMM-UBM, os seguintes fatores α da tabela 8.1 e 8.2 foram os que apresentaram a maior probabilidade de observação dado os modelos λ_{HMM} e λ_{UBM} para cada locutor. Na tabela 8.1 são apresentados os α-ótimos para os locutores meninos e na tabela 8.2 os α-ótimos para os locutores meninas. Nas tabelas, também são apresentadas as alterações que ocorreram nos α-ótimos ao se utilizar um método e o outro. Essas alterações são pequenas porem importantes responsáveis pela diferença no desempenho entre os métodos.

Tabela 8. 1: *α-ótimo encontrado para os locutores meninos através dos métodos HMM e GMM-UBM.*

Alphas Otimos em Locutores Meninos			
Locutor	HMM	GMM	alteração
bg	0.94	0.94	
bk	0.94	0.94	
bt	0.96	0.96	
dd	0.94	0.94	
dh	0.94	0.94	
dt	1.00	0.98	X
fb	0.96	0.96	
fw	0.96	0.96	
gf	0.94	0.94	
hk	0.98	0.98	
ic	0.96	0.94	X
jg	0.96	0.96	
lc	0.94	0.94	
lf	0.96	0.96	
ln	0.98	0.96	X
me	0.98	0.98	
mn	0.96	0.98	X
nm	0.94	0.94	
rh	0.96	0.96	
ri	0.92	0.94	X
rj	0.96	0.96	
sb	0.98	0.98	
se	0.94	0.94	
sh	0.94	0.94	
sk	0.96	0.98	X

Tabela 8. 2: α*-ótimo encontrado para os locutores meninas através dos métodos HMM e GMM-UBM.*

Alphas Otimos em Locutores Meninas			
Locutor	HMM	GMM	alteração
ad	0.94	0.94	
af	0.96	0.98	X
db	0.94	0.96	X
di	0.94	0.96	X
fe	0.94	0.94	
hw	0.94	0.94	
ij	0.96	0.96	
ir	0.94	0.94	
js	0.98	0.96	X
ki	0.94	0.94	
lk	0.94	0.96	X
lm	0.94	0.96	X
ma	0.94	0.94	
mf	0.96	0.96	
mg	0.96	0.96	
mt	0.94	0.96	X
nb	0.98	0.96	X
ns	0.94	0.94	
pa	0.94	0.94	
ps	0.96	0.98	X
rb	0.94	0.96	X
rt	0.94	0.96	X
rw	0.94	0.96	X
sc	0.94	0.96	X
ta	0.94	0.94	

A seguir podem ser observados os histogramas da Figura 8.5, Figura 8.6 e Figura 8.7, que mostram graficamente os valores de α-ótimos mais encontrados para meninos, meninas e crianças em ambos os métodos:

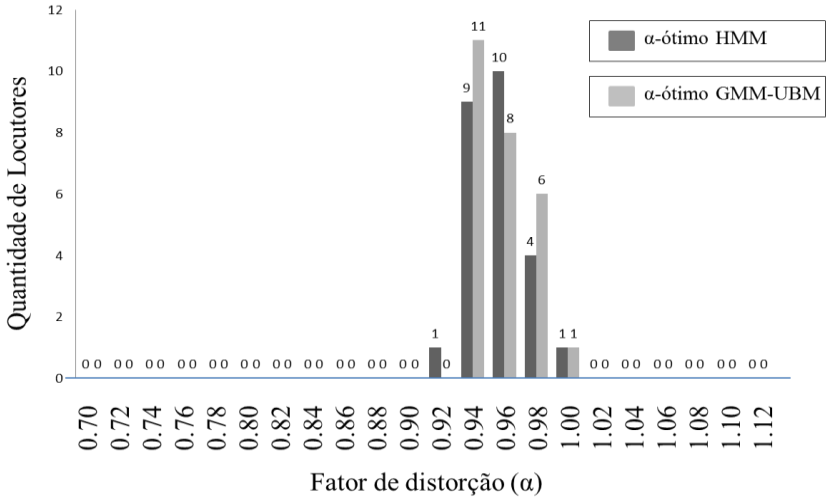

Figura 8. 5: *Comparativo entre os métodos HMM e GMM-UBM da quantidade de Locutores meninos por* α*-ótimo*

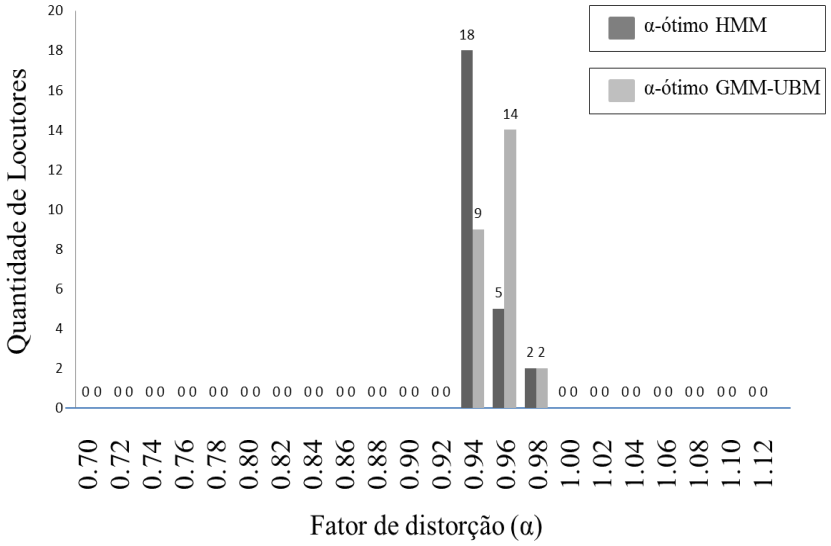

Figura 8. 6: *Comparativo entre os métodos HMM e GMM-UBM da quantidade de Locutores meninas por* α*-ótimo encontrado.*

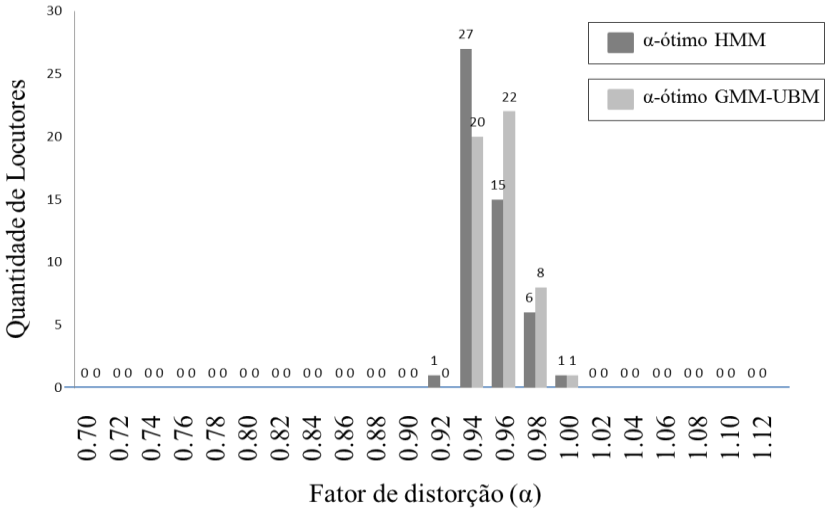

Figura 8. 7: *Comparativo entre os métodos HMM e GMM-UBM da quantidade de Locutores crianças por* α*-ótimo encontrado.*

Com a normalização, o fator de escalonamento (α) que caracterizou a maioria dos locutores crianças, quando utilizado o método HMM de busca foi α = 0,94, enquanto utilizando o método GMM-UBM foi α = 0,96. Em uma análise entre métodos, pode-se dizer que um método ratifica o outro. Os fatores foram praticamente os mesmos para locutores meninos utilizando os dois métodos. Para os locutores meninas, os fatores sofreram alteração em mais da metade do conjunto. No entanto, as alterações dos fatores que ocorreram foram de apenas um espaçamento de 0.02, sendo assim fatores vizinhos, com as frequências escalonadas muito próximas. A razão disso se da por uma questão de implementação do algoritmo de busca, mas que, posteriormente, como será abordado na seção 8.5, acarretará em diferenças com relação ao desempenho dos sistemas.

Em uma análise entre gêneros dos locutores, os fatores α foram satisfatórios, pois possuem uma relação com o fator de escalonamento 0,94 encontrado para locutores adultos femininos em testes realizados por [26], utilizando HMM, uma vez que as frequências formantes das crianças são muito próximas das mulheres, por isso há esse fator comum.

Não foram observadas diferenças significativas entre gêneros de crianças (meninos e meninas), isto poderia ser explicado pelo fato de que pela idade dos locutores, tanto meninos como meninas, ainda são muito parecidos organicamente.

Com relação aos modelos de referência dados na busca a partir da máxima $P(O|\lambda)$, os fatores que representam as crianças, estão distribuídos na faixa de 0,92 e 1,00. A região dos fatores encontrados, comparados com os valores teóricos obtidos da relação entre os comprimentos dos tratos vocais, era esperado que estivessem próximos a α = 0,70 (da razão 12/17 conforme abordado no capítulo 4), uma das razões para os fatores encontrados não estarem nessa região pode ser pelo fato de que esta relação não é linear.

8.3 Análise da curva da máxima probabilidade de observação por fator de escalonamento.

Para cada locutor criança foi realizada uma análise da máxima probabilidade de observação, P($O|\lambda$), por fator de escalonamento com o intuito de avaliar o comportamento da curva na região de busca dos fatores a partir dos métodos HMM e GMM-UBM.

A curva apresentada na Figura 8.8 explora o comportamento partir do método HMM, para o locutor criança "*bg*". É possível observar que a máxima probabilidade de observação aumenta a medida que o fator de escalonamento se aproxima de α = 0,94 atingindo seu pico e diminuindo em sequência até o fator α = 1,12.

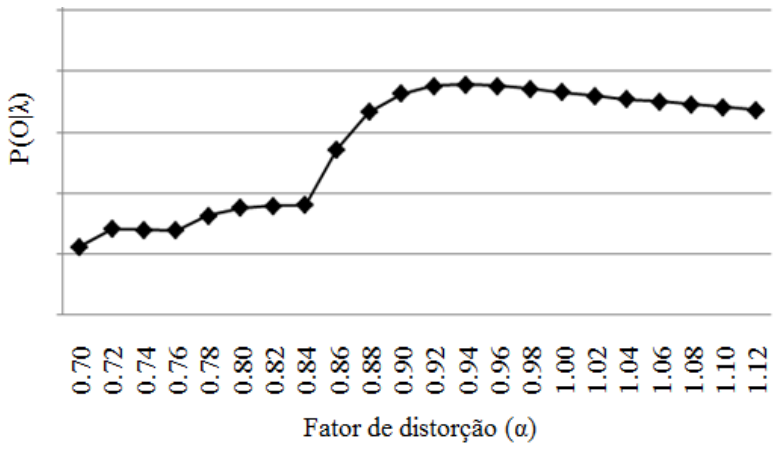

Figura 8. 8: *Curva da máxima P(O|λ) utilizando o método de busca HMM para o locutor "bg".*

A curva apresentada na Figura 8.9 explora o comportamento da máxima P($O|\lambda$) encontrado em cada fator de escalonamento utilizando o método GMM-

UBM, para o locutor criança "*bg*". Nesse gráfico observa-se que a máxima probabilidade de observação também alcança seu pico com o fator α = 0,94 e diminui até o fator α = 1,12.

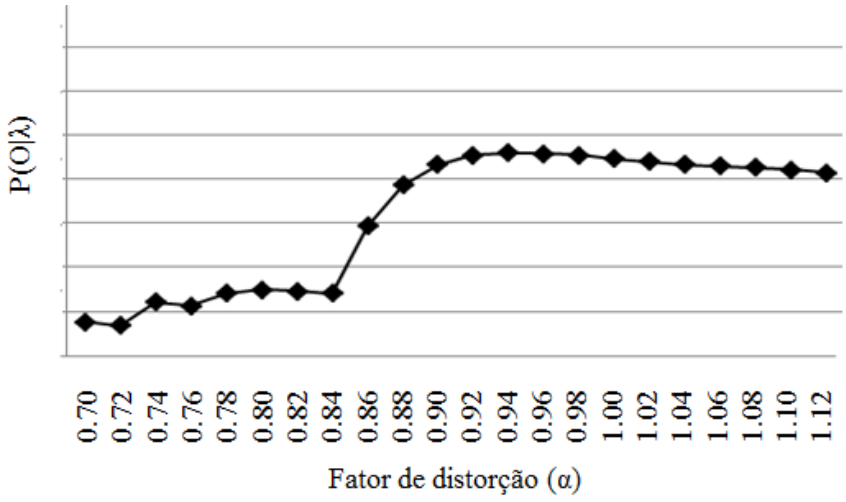

Figura 8. 9: *Curva da máxima P(O|λ) utilizando o método de busca GMM-UBM para o locutor "bg".*

Desses gráficos é possível concluir que a curva da máxima P(*O*|*λ*) utilizando os métodos HMM e GMM-UBM possui comportamento semelhante, com tendência de aumento na região de α= 0,86, atingindo o pico na região entre α = 0,94 e α = 0,96 e com tendência de queda a partir destes fatores.

Em média, os demais locutores analisados apresentaram comportamento similar para ambos os métodos.

8.4 Resultados Finais

Uma vez aplicados os métodos HMM e GMM-UBM utilizados na escolha do α-ótimo para normalização do sistema de reconhecimento de fala e realizados os passos para os testes dos experimentos apresentados no capítulo 7, nesta seção serão apresentados os resultados finais deste processo.

Foram realizados três conjuntos de experimentos: treinamento de locutores adultos, treinamento de locutores masculinos e treinamento de locutores femininos.

Em sequência, foram realizados testes aplicando a técnica de normalização do comprimento do trato vocal (VTLN) em locutores crianças.

Os resultados dessa seção apresentarão uma comparação entre a *baseline* (apresentada na seção 8.2) com o sistema normalizado com a técnica VTLN testado com os α-ótimos encontrados para cada locutor criança de teste através dos métodos HMM e GMM-UBM descritos nesse trabalho.

Os gráficos a seguir contêm os resultados de desempenho do sistema utilizando a métrica (WER %), taxa de erro de palavra, dos experimentos realizados para cada Gaussiana na mistura aplicada.

Nestes pode ser observado o comportamento da curva da taxa de erro (WER) para os conjuntos de treinamento e teste. São representadas três curvas: uma representando a *baseline* e as outras duas representando as curvas utilizando a técnica VTLN em locutores de teste crianças com os α-ótimos encontrados pelos métodos HMM e GMM-UBM. Desse modo, é possível realizar uma análise comparativa do desempenho entre os métodos.

Na Figura 8.10, o gráfico apresenta a curva da taxa de erro para o sistema treinado com locutores adultos e testado com locutores crianças. Com a aplicação da técnica VTLN em locutores crianças, utilizando o método HMM e GMM-UBM de busca do α-ótimo, pode-se observar que o comportamento da curva quando aplicada as Gaussianas na mistura, seguiu o mesmo padrão de comportamento encontrado na curva de taxa de erro da *baseline*. Portanto, a mistura ótima foi de 64 Gaussianas na mistura, pelas mesmas razões que o encontrado na *baseline*, abordadas na seção 8.2.

Utilizando o método HMM, no ponto de mínimo, alcançou-se 1,88% de taxa de erro, uma redução de 3,07% em relação à *baseline*. Utilizando o método GMM-UBM, no ponto de mínimo foi alcançada uma taxa de erro de 1,92%, uma redução de 3,03%. Os dois métodos obtiveram um desempenho semelhante. No ponto de mínimo, foi encontrada uma diferença de 0,04%. A maior diferença encontrada nas taxas de erros encontrados entre os métodos foi 0,34% quando usadas 4 Gaussianas na mistura.

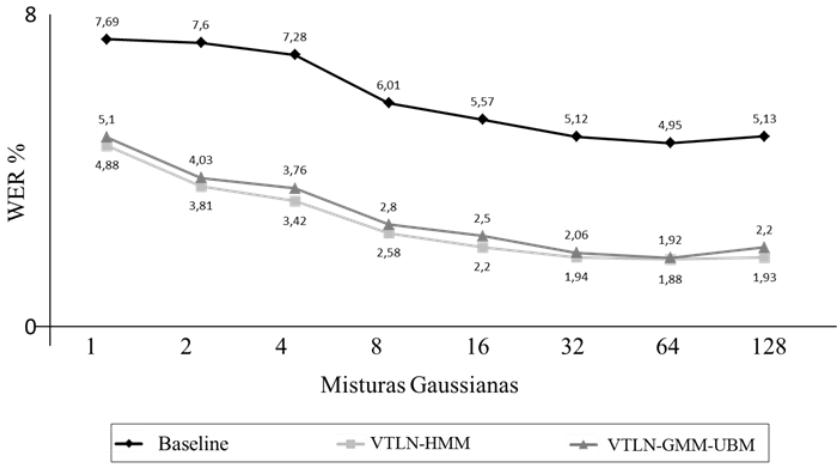

Figura 8. 10: *Curva WER% para o sistema treinado com Adultos e testado com Crianças*.

Para o sistema treinado com locutores masculinos e testado com locutores crianças, é apresentada na Figura 8.11 a curva de taxa de erro por número de Gaussianas na mistura. Nesse sistema, aplicando a VTLN com os α-ótimos encontrados pelo método HMM no ponto de mínimo (utilizando 4 Gaussianas na mistura), alcançou-se 28,39% de taxa de erro, uma redução de 6,83 % em relação à *baseline*.

Quando utilizado o método GMM-UBM, no ponto de mínimo, foi alcançada uma taxa de erro de 29,75%, ou seja, uma redução de 5,47% comparada à *baseline* com a mesma quantidade de Gaussianas na mistura. Os métodos também obtiveram desempenho semelhante no ponto de mínimo. A maior diferença de desempenho entre os dois métodos foi observado fora do ponto mínimo, quando aplicadas 8 Gaussianas na mistura. Nesse ponto a diferença entre os métodos foi de 1,73%.

Quando aplicada a técnica VTLN com os dois métodos, a taxa de erro ainda continuou alta. Uma razão para isso pode estar no fato de haver outras variabilidades entre locutores além do comprimento vocal. No entanto, houve significativa melhora do desempenho, aplicando VTLN.

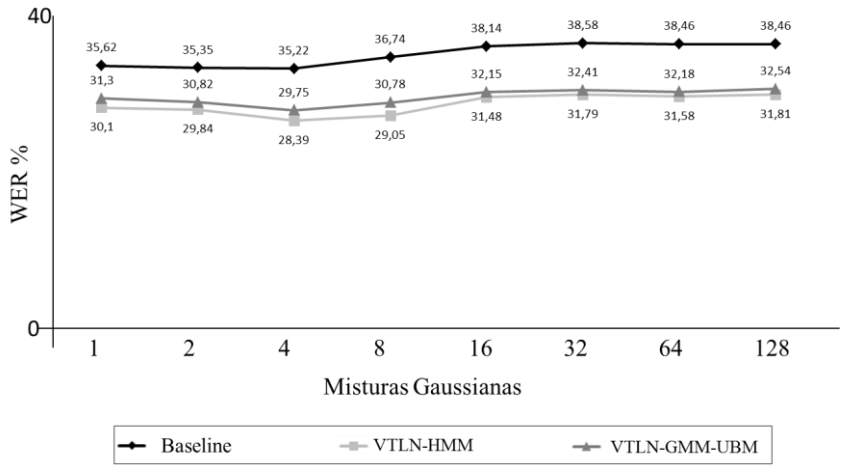

Figura 8. 11: *Curva WER% para o sistema treinado com locutores masculinos e testado com Crianças.*

Na Figura 8.12, é apresentado o comportamento das curvas do sistema treinado com locutores femininos e testado com locutores crianças. A curva que representa o sistema normalizado com a técnica VTLN utilizando os α-ótimos encontrados pelo método HMM, no ponto de mínimo (utilizando 32 Gaussianas na mistura) resultou em uma taxa de erro de 1,47%, uma redução de 2,4 % em relação à *baseline*. Quando utilizados os α-ótimos encontrados pelo método GMM-UBM no ponto de mínimo, foi alcançada uma taxa de erro de 1,58%, uma redução de 2,29% comparada a *baseline* com a mesma quantidade de Gaussianas na mistura. Neste ponto, os métodos HMM e GMM-UBM tiveram uma diferença de 0,11%, comprovando a similaridade entre ambos. A maior diferença de desempenho entre os dois métodos foi observada quando aplicada 2 Gaussianas na mistura. Nesse ponto a diferença entre os métodos foi de 0,38%.

No sistema normalizado com a técnica VTLN, o comportamento das curvas dos dois métodos fugiu a tendência da *baseline* quando aplicada 128 Gaussianas na mistura. Nesse ponto, a maior diferença foi entre HMM e a *baseline*, uma diferença de 4,84%.

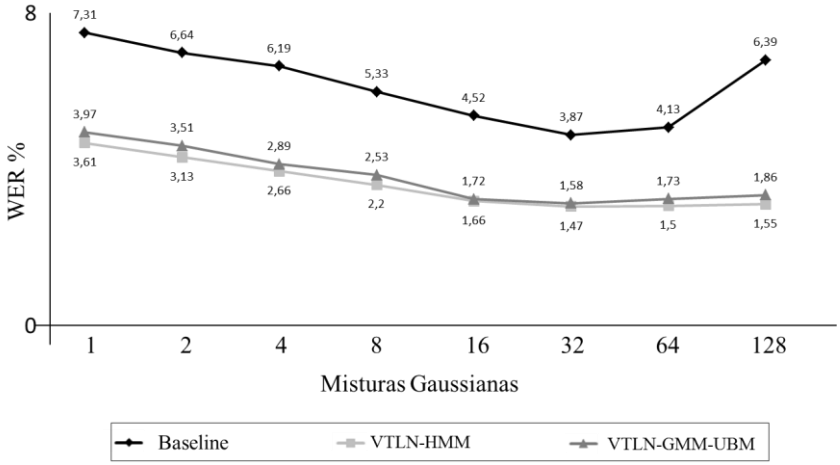

Figura 8. 12: *Curva WER% para o sistema treinado com locutores femininos e testado com Crianças.*

Com estes resultados é possível traçar algumas comparações acerca dos métodos HMM e GMM-UBM utilizados para encontrar o α-ótimo para cada locutor.

8.5 Discussão e Comparação entre os métodos HMM e GMM-UBM na busca do α-ótimo para cada locutor.

Conforme descrito no capítulo 4 e de posse dos resultados finais, foi possível observar que os sistemas treinados com adultos e testado com crianças normalizadas obtiveram uma melhora significativa com relação ao desempenho do sistema de reconhecimento de fala.

Da comparação entre os métodos, comprova-se a partir dos resultados que os locutores crianças normalizados a partir dos α-ótimos encontrados pelo método HMM e pelo método GMM-UBM obtiveram resultados muito próximos.

Outras duas questões foram levadas em consideração: Complexidade de Implementação e Custo Computacional.

Complexidade de Implementação:

Para a busca do α-ótimo, o método HMM demonstra uma complexidade de implementação muito maior do que o GMM-UBM. O HMM requer a construção de uma topologia baseada em estados. Esses estados, por sua vez, possuem uma

probabilidade de transição e permanência, onde esses valores são atualizados durante o processo de treinamento. Desse modo, quanto maior for a locução mais complexo será o modelamento do sistema. Outra complexidade de implementação que permeia o HMM é a necessidade da transcrição fonética. O GMM-UBM utiliza somente um estado, que corresponde a palavra inteira e não utiliza transcrições fonéticas. Com base nesses aspectos, pode-se concluir que o GMM-UBM possui uma implementação mais simples do que o HMM.

Custo Computacional:

O custo computacional observado nos experimentos desse trabalho envolveu duas características: Tempo de processamento e utilização de recursos de memória.

O tempo de processamento refere-se ao tempo de todo processo: Modelo pré-treinado e busca pelos α-ótimos.

A busca pelo α-ótimo utilizando o HMM levou, em média, cerca de 2 minutos por locutor, um total de 1 hora e 40 minutos. A busca utilizando o GMM-UBM levou, em média, cerca de 1 minuto e meio, um total de 1 hora e 15 minutos. A busca pelo α-ótimo é um processo que demandou alto tempo de processamento para ambos os métodos, pois utilizam basicamente o mesmo algoritmo de busca. A diferença fica por conta da questão do software utilizado, HTK para o HMM e o software fornecido pelo professor Carlos Alberto Ynoguti para o GMM-UBM. Com base nas aferições, o método GMM-UBM é mais rápido na busca pelo α-ótimo. Com relação ao modelo pré-treinado, o modelo HMM pré-treinado necessita de reestimações mais demorados para tornar o modelo mais refinado e robusto. Os parâmetros de cada estado do HMM são atualizados a cada reestimação e isso resulta em um tempo de processamento alto. O modelo UBM pré-treinado requer uma convergência para um valor especificado, no entanto, conforme pode ser observado na tabela 8.3, o tempo de processamento é menor do que utilizando o método HMM.

Tabela 8. 3: *Tempo de processamento para os modelos pré-treinados utilizando o método HMM e UBM.*

HMM

mix	Tempo de Processamento (hrs:min) Modelo pré-treinado
1	00:40
2	01:15
4	01:58
8	02:40
16	04:12
32	05:57
64	08:43
128	11:34

UBM

mix	Tempo de Processamento (hrs:min) Modelo pré-treinado
1	00:28
2	00:50
4	01:12
8	01:36
16	03:07
32	04:25
64	05:24
128	08:33

O uso de memória refere-se à quantidade de memória exigida pelo processo ao utilizar os métodos. Para isso procurou-se manter o Sistema Operacional com o menor número de aplicações em execução, para que não influenciasse no desempenho do que se pretende aferir. A ferramenta para medir o custo de memória foi o indicador de memória "*SensorsScreenlet v0.1*" da plataforma Linux Ubuntu 10.04 LTS. A máquina utilizada para a realização dos experimentos foi um “Intel Core 2 Quad CPU 2.83Ghz” com 3Gb de memória RAM. Com base na tabela 8.4 é possível notar que o método GMM-UBM para busca do α-ótimo exige menor quantidade de memória do que o HMM.

Tabela 8. 4: *Comparação do uso de memória no processo de busca do α-ótimo utilizando os métodos HMM e GMM-UBM.*

HMM		
	% de uso da memória RAM	
mix	Modelo pré-treinado	Busca pelo α-ótimo
1	23%	35%
2	25%	34%
4	26%	37%
8	26%	36%
16	27%	37%
32	25%	34%
64	25%	32%
128	24%	33%

GMM-UBM		
	% de uso da memória RAM	
mix	Modelo pré-treinado	Busca pelo α-ótimo
1	13%	23%
2	15%	27%
4	16%	27%
8	17%	26%
16	15%	25%
32	18%	28%
64	17%	25%
128	17%	27%

Desse modo, é possível concluir que utilizando o método GMM na busca do α-ótimo obtiveram-se resultados similares aos encontrados pelo método HMM nos resultados de desempenho do sistema de reconhecimento de fala. O método GMM também possui implementação mais simples e envolve no processo menor custo computacional.

Capítulo **9**

Conclusões e Oportunidades para Pesquisas Futuras

O principal foco desse trabalho foi comparar e avaliar os métodos estatísticos, modelos ocultos de Markov (HMM) e modelos de misturas Gaussianas (GMM) na busca do fator α-ótimo para normalização do comprimento do trato vocal de locutores criança. Os experimentos foram realizados para três conjuntos: Treinamento com locutores adultos e teste com locutores crianças, treinamento com locutores masculinos e teste com crianças, e treinamento com locutores femininos e teste com crianças.

Na comparação entre os métodos de busca do α-ótimo, foi possível observar que os resultados obtidos com o método GMM foram similares aos resultados encontrados pelo método HMM.

Foi realizada uma avaliação acerca do custo computacional. Nos experimentos realizados, quando utilizado o HMM, o tempo de processamento e a utilização de memória do computador para o treinamento dos modelos pré-treinados e o processo de busca do α-ótimo foram maiores em comparação com o GMM-UBM.

Na questão de implementação, o método HMM requer uma topologia de estados que envolvem uma probabilidade de transição e permanência entre eles, bem

como a transcrição fonética. Já o método GMM-UBM é configurado de forma mais simples, por não necessitar de transcrição fonética e possuir somente 1 estado. Desse modo, a busca pelo α-ótimo para cada locutor criança é mais viável se realizada através do método GMM-UBM, pois se condiciona a um desempenho praticamente igual ao HMM e ainda resulta em uma implementação mais simples e custo computacional baixo.

Uma análise foi realizada com relação à curva de máxima $P(O|\lambda)$ encontrada para cada fator de escalonamento, sendo possível observar o comportamento da curva para os métodos HMM e GMM-UBM.

Foi realizada uma investigação do comportamento da curva de taxa de erro da *baseline* e dos experimentos utilizando VTLN com os α-ótimos encontrados com HMM e GMM-UBM, quando aplicadas Gaussianas na mistura. A partir disso, é detectada a mistura ótima para cada experimento. O sistema treinado com adultos e testado com crianças necessita de mais Gaussianas na mistura para alcançar o ponto de mínimo (64 Gaussianas), pois possui maior variabilidade entre os locutores de treinamento. O sistema treinado com locutores masculinos e testado com crianças necessita de poucas Gaussianas na mistura (4 Gaussianas) pois há menor variabilidade entre locutores masculinos. Já no sistema treinado com locutores femininos e testado com crianças, há uma alta variabilidade entre mulheres, por razões de frequências formantes entre elas e, com isso, requerem-se muitas misturas (32 misturas) para atingir o ponto de mínimo do sistema.

Ambos os métodos, com crianças normalizadas reduziram substancialmente a taxa de erro encontrada na *baseline*. Isso endossa o uso da normalização de comprimento do trato vocal realizado pelo escalonamento do banco de filtros e encoraja a utilização desta técnica em sistemas de reconhecimento de fala cujo alvo são crianças.

Uma contribuição deste trabalho, em relação a [26] e [14], foi a avaliação dos métodos estatísticos mais utilizados em reconhecimento de fala, como HMM e GMM-UBM, para a obtenção do α-ótimo de cada locutor criança na aplicação da técnica de normalização do comprimento de trato vocal (VTLN).

Com os métodos analisados visou-se melhorar o desempenho de sistemas de reconhecimento de fala em sistemas específicos para usuários crianças, como jogos, sistemas educacionais, aplicativos para celulares, embarcação em brinquedos etc.

Em suma, as principais contribuições desse trabalho foram:

- Foi apresentada uma forma alternativa, mais simples, de menor custo computacional e de memória, com desempenho similar, para o cálculo dos α-ótimo, para a utilização da técnica VTLN.

- Reduziu-se a taxa de erro em sistemas treinados com adultos e testado com crianças de 4,95% para 1,88% quando utilizado a VTLN com os α-ótimos encontrados pelo HMM e 1,92 % quando utilizado a VTLN com os α-ótimos encontrados pelo GMM-UBM.

Como sugestão de futuras investigações sugere-se:

- No processo de extração de características acústicas, utilizar janelas mais curtas, pois, segundo [48], deve-se proporcionar uma melhor adaptação à fala de crianças que têm tom mais alto, (por exemplo 15 ms).

- Realizar uma análise comparativa dos métodos HMM e GMM-UBM para modelos acústicos dependentes de idades, pois, segundo [1], há uma forte relação entre o fator ótimo de escalonamento e a idade dos locutores crianças.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] GEROSA, M., GIULIANI, D., NARAYANAN, S. and POTAMIANOS, A., "*A Review of ASR Technologies for Children's*", WOCCI'09 - Proceedings of the 2nd Workshop on Child, Computer and Interaction, pp. 3-6, Nov. 2009.

[2] MARTIN, R. and D'ARCY, S., "*Challenges for computer recognition of children's speech*", SLATE - Speech and Language Technology in Education, pp. 108-111, Oct. 2007.

[3] HECKER, M. H., "*Speaker Recognition: An Interpretive Survey of the Literature*", ASHA - American Speech and Hearing, no. 16, pp. 4-5, Jan. 1971.

[4] RABINER, L. R. and JUANG, B. H., Fundamentals of Speech Recognition, Prentice Hall, pp. 16,17,507, Apr. 1993.

[5] BLOMBERG, M., "*Model Space Size Scaling for Speaker Adaptation*", FONETIK'11, vol. 51, pp. 77-80, 2011.

[6] RABINER, L. R. and JUANG, B. H., Automatic Speech Recognition - A Brief History of the Technology, Elsevier Encyclopedia of Language and Linguistics, pp. 1-24, Aug. 2004.

[7] GALES, M. e YOUNG S., "*The Application of Hidden Markov Models in Speech Recognition"*, Fundations and Trends in Signal Processing, vol. 1, pp. 195-304, 2007.

[8] RABINER, L. R. and JUANG, B. H., "*Hidden Markov Models for Speech Recognition*", Technometrics, vol. 33, 3, pp. 251-272, Aug. 1991.

[9] BHATTACHARJEE, U., "*Environment and Sensor Robustness in Automatic Speech Recognition*", International Journal of Innovative Science and Modern Engineering (IJISME), vol. 1,pp. 1-7, Jan. 2013.

[10] ZHAO, Y. and SUN, X., "*Integrated exemplar-based template matching and statistical modeling for continuous speech recognition*", EURASIP Journal on Audio, Speech, and Music Processing, pp. 1-16, Apr. 2014.

[11] PICONE, J. W., "*Signal Modelling Techniques in Speech Recognition*", Proc. IEEE, vol. 81, 9, pp. 1215-1247, Sep. 1993.

[12] STOLCKE, A., SHRIBERG, E., FERRER, L., KAJAREKAR, S., SONMEZ, K., TUR, G., "*Speech Recognition as Feature Extraction for Speaker Recognition*", SAFE, IEEE Workshop, pp. 1-5, Apr. 2007.

[13] KESARKAR, M., "*Feature Extraction for Speech Recognition*", M.Tech. Credit Seminar Report, pp. 1-13, Nov. 2003.

[14] DIAS, R. S. F., YNOGUTI, C. A., VIOLARO, F. *"Normalização de Locutor em Sistema de Reconhecimento de Fala"*, XIX Simpósio Brasileiro de Telecomunicações, pp. 1-6, 2001.

[15] ALENCAR, S. F. V. Atributos e Domínios de Interpolação Eficientes em Reconhecimento de Voz Distribuído. Dissertação de Mestrado Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro. Brasil, 2005.

[16] HECKMANN, M., "*Supervised vs. Unsupervised Learning of Spectro Temporal Speech Features*", ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, pp. 1-5, Sep. 2010.

[17] DELLER, J. R., HANSEN, J. H. and PROAKIS, J. G., "*Discrete-Time Processing of Speech Signals*", IEEE Press, pp. 936, 2000.

[18] ALCAIM A., CUADROS C. D. R. e DA SILVA D. G., *"Reconhecimento Robusto de Locutor Baseado nos Atributos ZCPAC"*, XXV Simpósio Brasileiro de Telecomunicações, pp. 1-5, Sep. 1997.

[19] DAVIS, S. B. and MERMELSTEIN, P., "*Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*", IEEE Trans. Acoust., Speech, Signal Processing, vol. 28, 4, pp. 357-368, Aug. 1980.

[20] RABINER, L. and JUANG, B., "*An Introduction to Hidden Markov Models*". IEEE ASSP Magazine, pp. 4-16, Jan. 1986.

[21] RABINER, L., "*A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*", Proc. IEEE, vol. 77, 2, pp. 257-286, Feb. 1989.

[22] BRUGNARA, F., De MORI, R., "*Survey of the State of the Art in Human Language Technology: HMM Methods in Speech Recognition*", Cambridge University Press, pp. 20-57, 1997.

[23] YNOGUTI, C. A. e VIOLARO, F., "*Desenvolvimento de um conjunto de Ferramentas para Pesquisas em Reconhecimento de Fala*". Revista Cientifica Periodica - Telecomunicações INATEL, vol. 4, 2, Dez. 2001.

[24] REYNOLDS, D. A. and ROSE, R. C., "*Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*", IEEE Trans. Speech Audio Processing, vol. 3, 1, pp. 72-83, Jan. 1995.

[25] Gaussian Mixture Models, Douglas Reynolds. Available at: http://llwebprod2.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/0802_Reynolds_Biometrics-GMM.pdf (last access: on Mar 2014).

[26] LEE, L. and ROSE, R., "*A Frequency Warping Approach to Speaker Normalization*", IEEE Trans. Speech Audio Processing, vol. 6, 1, pp. 49-60, Jan. 2012.

[27] UMESH, S. and SANAND, D. R., "*VTLN Using Analytically Determined Linear-Transformation on Conventional MFCC*", IEEE Trans. Audio Speech and Language Processing, vol. 20, 5, Jul. 2012.

[28] FINK, G. A., Markov Models for Pattern Recognition From Theory to Applications, Springer, pp. 127-128, 2008.

[29] FERNANDES, D. B. Adaptação ao Locutor Usando a Técnica MLLR. Dissertação de Mestrado Instituto Nacional de Telecomunicações. Santa Rita do Sapucaí. Minas Gerais. Brasil, 2011.

[30] REYNOLDS, D. A., "*Speaker Identification and Verification using Gaussian Mixture Speaker Models*", Speech Communication, vol. 17, 2, pp. 91–108, Aug. 1995.

[31] RABINER L. R e JUANG B. H., "*The Segmental K-Means Algorithm for Estimating Parameters of Hidden Markov Models*", IEEE Trans. Acoust., Speech, Signal Processing, vol. 38, 9, pp. 1639-1641, Sep. 1990.

[32] YOUNG, S., EVERMANN, G., GALES, M., HAIN, T., KERSHAW, D., LIU, X. A., MOORE, G., ODELL, J., OLLASON, D., POVEY, D., VALTCHEV, V. and WOODLAND, P., "The HTK Book (for HTK Version 3.4)", Univ. Eng. Dept., Dec. 2006.

[33] TIDIGITS, Joseph Picone. Available at: http://www.isip.piconepress.com/projects/speech/software/tutorials/production/fundamentals/v1.0/section_02/s02_04_p01.html (last access: on Fev 2014).

[34] LEONARD, R. G., "*A Database for Speaker-Independent Digit Recognition*", Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP, Vol. 9, p. 328-331, 1984.

[35] A Speaker-Independent Connected-Digit Database, R. Gary Leonard and George R. Doddington. Available at: http://catalog.ldc.upenn.edu/docs/LDC93S10/tidigits.readme.html (last access: on Dec 2013)

[36] MERTINS, A. and RADEMACHER, J., "*Vocal tract lenght invariant features for automatic speech recognition*", IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 308-312, Jan. 2005.

[37] BURNETT, D. C. and FANTY, M., "*Rapid Unsupervised Adaptation to Children's Speech on a Connected-Digit Task*", Fourth International Conference on Spoken Language - ICSLP Proc., vol. 2, pp. 1145-1148, Oct. 1996.

[38] SILVA, P., NETO, N. e KLAUTAU, A., "*Novos Recursos e Utilização de Locutor no Desenvolvimento de um Sistema de Reconhecimento de Voz para o Português Brasileiro*", XXVII SBRT, pp. 1-6, Oct. 2009.

[39] MARTINS, J. A. Avaliação de Diferentes Técnicas para Reconhecimento de Fala. Tese de Doutorado Universidade de Campinas. Campinas. São Paulo, 1997.

[40] PICONE, J., "*Fundamentals of Speech Recognition: A Short Course*", ISIP, May. 1996.

[41] ARPAbet: Phoneme Set, Carnegie Mellon University. Available at: http://www.speech.cs.cmu.edu/cgi-bin/cmudict#phones (last access: on Jan 2014).

[42] BAHL, L.R., et al., "*Acoustic Markov models used in the Tangora speech recognition system*", ICASSP, vol. 1,pp. 497-500, Apr. 1988.

[43] HTK (v.3.1): Basic Tutorial, Nicolas Moreau. Available at: http://my.fit.edu/~vkepuska/HTK/HTK_basic_tutorial.pdf (last access: on Jan 2014).

[44] WIGGERS, I. P. and ROTHKRANTZ, L. J. M., "*Automatic Speech Recognition using Hidden Markov Models*", Real-time AI & Automatische Spraakherkenning - Delft University of Technology, pp55, Sep. 2003.

[45] MUKUNDAN, S. K., "*'Shreshta Bhasha' Malayalam Speech Recognition using HTK*", IJACCS Journal, vol. 1, pp. 1-5, Mar. 2014.

[46] AGUILLAR, R. C., "*Diseño y Manipulación de Modelos Ocultos de Markov Utilizando Herramientas HTK, Una Tutoría*", Revista Chilena de Ingeniería, vol. 15, 1, pp. 18-26, Apr. 2007.

[47] HARDCASTLE, W. J., LAVER, J. and GIBBON, F. E., The Handbook of Phonetic Sciences, Wiley editor, pp. 831, 2010.

[48] TEIXEIRA, A. D. C. Deteção e Correção de Disfluências em Crianças. Dissertação de Mestrado Faculdade de Ciência e Tecnologia. Coimbra. Portugal, 2012.

APÊNDICE

Locuções utilizadas no processo de busca do α-ótimo para cada locutor criança menino (a) e para cada locutor criança menina (b).

1 bg	**6 dt**	**11 ic**	**16 me**	**21 rj**
791	47z	291	8z3	9o4
869	357	317	316	311
o68	782	928	418	458
z65	762	z54	544	897
2 bk	**7 fb**	**12 jg**	**17 mn**	**22 sb**
5z1	6o8	4z2	9o1	62o
91z	228	810	497	373
573	275	167	o23	856
o36	591	234	z81	o87
3 bt	**8 fw**	**13 lc**	**18 nm**	**23 se**
9o8	79o	5o4	3z7	6z7
262	154	5z7	279	81o
454	386	354	832	541
z55	o73	839	561	z18
4 dd	**9 gf**	**14 lf**	**19 rh**	**24 sh**
3o1	5z3	161	7o2	52z
24z	471	273	173	313
598	782	864	493	654
732	892	o54	625	932
5 dh	**10 hk**	**15 ln**	**20 ri**	**25 sk**
3o5	5z8	3o2	347	271
3z1	510	347	524	312
469	254	641	698	853
725	428	z28	737	941

(a) Locutores criança meninos

1 ad	6hw	11 lk	16 mt	21 rb
1z6	7z9	23o	5z1	9z6
9z9	78o	61z	13o	397
493	164	442	26z	563
z71	841	728	528	782
2af	7ij	12 lm	17 nb	22rt
1o5	2o4	173	186	3o1
362	77z	287	212	448
548	197	388	441	531
617	416	415	851	219
3db	8ir	13 ma	18 ns	23 rw
189	16o	17z	263	49o
582	17z	648	345	158
621	183	959	735	579
793	271	zz6	z64	748
4di	9js	14 mf	19 pa	24 sc
11z	5o4	6o5	191	81z
479	619	9z2	716	84o
z56	935	152	z34	235
z74	o15	489	z89	o7o
5je	10 ki	15 mg	20 os	25 ta
496	7o4	338	98o	339
584	49o	478	196	413
759	875	o12	464	974
o71	zz6	z63	z27	z85

(b) Locutores criança meninas