Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Método para a Determinação do Número de Gaussianas em Modelos Ocultos de Markov para Sistemas de Reconhecimento de Fala Contínua

Autor: Glauco Ferreira Gazel Yared
Orientador: Prof. Dr. Fábio Violaro

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica. Área de concentração: **Engenharia de Telecomunicações**.

Banca Examinadora

Fábio Violaro, Dr. ..... DECOM/FEEC/UNICAMP
Fernando Gil Vianna Resende Junior, PhD. ..... COPPE/POLI/UFRJ
Carlos Alberto Ynoguti, Dr. ..... INATEL
João Bosco Ribeiro do Val, PhD. ..... DT/FEEC/UNICAMP
Amauri Lopes, Dr. ..... DECOM/FEEC/UNICAMP
Jaime Portugheis, PhD. ..... DECOM/FEEC/UNICAMP

Campinas, SP
Abril/2006

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

Y21m

Yared, Glauco Ferreira Gazel
Método para a determinação do número de gaussianas em modelos ocultos de Markov para sistemas de reconhecimento de fala contínua / Glauco Ferreira Gazel Yared. –Campinas, SP: [s.n.], 2006.

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

1. Algoritmos. 2. Markov, Processos de. 3. Reconhecimento automático da voz. 4. Modelos matemáticos. I. Violaro, Fábio. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: A new method for determining the number of gaussians in hidden Markov models for continuous speech recognition systems
Palavras-chave em Inglês: Gaussian elimination algorithm, HMM, Robustness, Model complexity, Hidden Markov models
Área de concentração: Telecomunicações e Telemática.
Titulação: Doutor em Engenharia Elétrica
Banca Examinadora: Fernando Gil Vianna Resende Junior, Carlos Alberto Ynoguti, João Bosco Ribeiro do Val, Amauri Lopes e Jaime Portugheis
Data da defesa: 20/04/2006

Resumo

Atualmente os sistemas de reconhecimento de fala baseados em HMMs são utilizados em diversas aplicações em tempo real, desde telefones celulares até automóveis. Nesse contexto, um aspecto importante que deve ser considerado é a complexidade dos HMMs, a qual está diretamente relacionada com o custo computacional. Assim, no intuito de permitir a aplicação prática do sistema, é interessante otimizar a complexidade dos HMMs, impondo-se restrições em relação ao desempenho no reconhecimento. Além disso, a otimização da topologia é importante para uma estimação confiável dos parâmetros dos HMMs. Os trabalhos anteriores nesta área utilizam medidas de verossimilhança para a obtenção de sistemas que apresentem um melhor compromisso entre resolução acústica e robustez. Este trabalho apresenta o novo Algoritmo para Eliminação de Gaussianas (GEA), o qual é baseado em uma análise discriminativa e em uma análise interna, para a determinação da complexidade mais apropriada para os HMMs. O novo método é comparado com o Critério de Informação Bayesiano (BIC), com um método baseado em medidas de entropia, com um método discriminativo para o aumento da resolução acústica dos modelos e com os sistemas contendo um número fixo de Gaussianas por estado.

Palavras-chave: Algoritmo para Eliminação de Gaussianas, HMM, robustez, complexidade dos modelos.

Abstract

Nowadays, HMM-based speech recognition systems are used in many real time processing applications, from cell phones to automobile automation. In this context, one important aspect to be considered is the HMM complexity, which directly determines the system computational load. So, in order to make the system feasible for practical purposes, it is interesting to optimize the HMM size constrained to a minimum acceptable recognition performance. Furthermore, topology optimization is also important for reliable parameter estimation. Previous works in this area have used likelihood measures in order to obtain models with a better compromise between acoustic resolution and robustness. This work presents the new Gaussian Elimination Algorithm (GEA), which is based on a discriminative analysis and on an internal analysis, for determining the more suitable HMM complexity. The new approach is compared to the classical Bayesian Information Criterion (BIC), to an entropy based method, to a discriminative-based method for increasing the acoustic resolution of the HMMs and also to systems containing a fixed number of Gaussians per state.

Keywords: Gaussian Elimination Algorithm, HMM, robustness, model complexity.

Dedico este trabalho ao bom Deus, ao meu pai Jorge, à minha mãe Fátima, à minha irmã Karen e ao amor da minha vida Luciana. Dedico também ao meu avô Carlos (in memorian), à minha avó Luíza, ao meu avô Jan e à minha avó Mary (in memorian)

Agradecimentos

Agradeço primeiramente à Deus, por ter me dado forças para concluir este trabalho. Agradeço aos meus pais, a minha irmã, a minha esposa, pelo incentivo, apoio e paciência nos momentos mais difíceis.

Agradeço também ao Prof. Dr. Fábio Violaro, pelas orientações e ensinamentos sobre a área de reconhecimento de fala, que foram fundamentais para o progresso do trabalho que desenvolvi no LPDF/DECOM/FEEC/UNICAMP. Agradeço também por sua amizade e seriedade em todos os momentos no desenvolvimento do trabalho.

Agradeço ao Prof. Dr. Carlos Alberto Ynoguti por ter disponibilizado os códigos do sistema de decodificação de fala.

Agradeço à Valter, Fátima e Alexandre Fachini pelo incentivo e amizade ao longo dos últimos onze anos.

Agradeço aos meus amigos de caminhada Reinaldo Alves Araújo, Carlos Henrique R. Barbosa, Bernardo Soares Torres e Marcos Aldred Ramacciotti pela forte amizade.

Agradeço ao CNPq pelo suporte financeiro, o qual viabilizou o trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Glossário

AIC - Akaike Information Criterion

API - Application Program Interface

BIC - Bayesian Information Criterion

BW - Bandwidth

CDF - Cumulative Distribution Function

DC - Discriminative Constant

GEA - Gaussian Elimination Algorithm

GIM - Gaussian Importance Measure

HMM - Hidden Markov Model

LM - Likelihood Maximization

LVCSR - Large Vocabulary Continuous Speech Recognition

MCE - Minimum Classification Error

MDL - Minimum Description Length

MFCC - Mel Frequency Cepstral Coefficient

MLE - Maximum Likelihood Estimation

MMIE - Maximum Mutual Information Estimation

PC - Personal Computer

PDF - Probability Density Function

Perc. Corr. - Percentagem correta ou taxa de reconhecimento de palavras

Reco. Accur. - Recognition Accuracy

SVCSR - Small Vocabulary Continuous Speech Recognition

WGP - Winner Gaussian Probability

Lista de Símbolos

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	-	Função densidade de probabilidade normal, com vetor média μ e matriz de covariância $\boldsymbol{\Sigma}$
$P(\mathbf{O}\|M)$	-	Verossimilhança
$\|\boldsymbol{\Sigma}\|$	-	Determinante da matriz de covariância
dim	-	Dimensão do espaço de características acústicas
F_d	-	Fator de desempenho
$P_{wg}^{(i;j;s)}$	-	Probabilidade da Gaussiana vencedora
$P_{GIM}^{(i;j;s)}$	-	Valor médio da medida de importância da Gaussiana em relação ao estado "s"
$\mathbf{P_f}$	-	Vetor de parâmetros acústicos no instante "f"
H_s	-	Entropia do estado "s"
$\boldsymbol{\mu}_{jm}$	-	Média da Gaussiana "m", que pertence ao estado "j"
$\mathbf{U}_{jm}$	-	Matriz de covariância
$\mathbf{U}_{jm}^{-1}$	-	Matriz de covariância inversa
c_{jm}	-	Peso da Gaussiana
a_{ij}	-	Probabilidade de transição de estados
N_g	-	Número de Gaussianas do estado
$\mathbf{o}_t$	-	Vetor de características no instante "t"

Trabalhos Publicados pelo Autor Durante o Doutorado

1. R. N. Rodrigues, G. F. G. Yared, C. R. N. Costa, J. B. T. Yabu-Uti, F. Violaro, L. L. Ling. "Biometric Access Control Through Numerical Keyboards based on Keystroke Dynamics". *Lecture Notes in Computer Science*, Vol. 3832, pg. 640-646, Janeiro 2006.

2. R. N. Rodrigues, G. F. G. Yared, C. R. N. Costa, J. B. T. Yabu-Uti, F. Violaro, L. L. Ling. "Biometric Access Control Through Numerical Keyboards based on Keystroke Dynamics". *International Conference on Biometrics* (ICB'2006), Hong Kong, China, pg. 640-646, Janeiro 2006.

3. G. F. G. Yared, F. Violaro, L. C. Sousa. "Gaussian Elimination Algorithm for HMM Complexity Reduction in Continuous Speech Recognition Systems". *9th European Conference on Speech Communication and Technology* (Interspeech - Eurospeech 2005), Lisboa, Portugal, pg. 377-380, Setembro 2005.

4. G. F. G. Yared, F. Violaro. "Algoritmo para Redução do Número de Parâmetros de HMMs Utilizados em Sistemas de Reconhecimento de Fala Contínua". *XXII Simpósio Brasileiro de Telecomunicações* (SBrT'2005) , Campinas, São Paulo, Brasil, pg. 423-428, Setembro 2005.

5. C. R. N. Costa, G. F. G. Yared, R. N. Rodrigues, J. B. T. Yabu-Uti, F. Violaro, L. L. Ling. "Autenticação Biométrica via Dinâmica da Digitação em Teclados Numéricos". *XXII Simpósio Brasileiro de Telecomunicações* (SBrT'2005) , Campinas, São Paulo, Brasil, pg. 423-428, Setembro 2005.

6. G. F. G. Yared, F. Violaro. "Finding the More Suitable HMM Size in Continuous Speech Recognition Systems". *3rd International Information and Telecommunications Technology Symposium* (I2TS'2004) , São Carlos, São Paulo, Brasil, pg. 141-146, Dezembro 2004.

7. G. F. G. Yared, F. Violaro. "Determining the Number of Gaussians per State in HMM-based Speech Recognition Systems ". *International Workshop on Telecommunications* (IWT'2004) , Santa Rita do Sapucaí, Minas Gerais, Brasil, pg. 194-201, Agosto 2004.

8. G. F. G. Yared, J. V. Gonçalves, P. Barbosa, L. G. P. Meloni. "Primeiros Experimentos com Dados Articulatórios e sua Relação com a Segmentação Acústica ". *Revista de Estudos da Linguagem*, Vol. 12, No. 1, pg. 39-52, 2004.

9. G. F. G. Yared, J. V. Gonçalves, P. Barbosa, L. G. P. Meloni. "Primeiros Experimentos com Dados Articulatórios e sua Relação com a Segmentação Acústica ". *VII Congresso Nacional de Fonética e Fonologia*, Belo Horizonte, Minas Gerais, Brasil, pg. 194-201, Outubro 2002.

10. P. Feijão, J. Bandeira, G. F. G. Yared, L. G. P. Meloni. “Equalizador Digital de Áudio Empregando a MLT”. *Seminário de Engenharia de Áudio 2002* (SEMEA’2002), Belo Horizonte, Minas Gerais, Brasil, Junho 2002.

Capítulo 1

Introdução

1.1 Importância da Área de Reconhecimento de Fala

Os sistemas de reconhecimento de fala têm sido utilizados em diversas aplicações e com diferentes propósitos. Tais aplicações têm como objetivo auxiliar a realização de várias tarefas, como por exemplo a elaboração de textos a partir da conversão de sentenças faladas para sentenças escritas, o acionamento de dispositivos eletrônicos embarcados em automóveis, a discagem por comando de voz em telefones celulares, etc. Além disso, podem permitir também que tais tarefas possam ser realizadas facilmente por portadores de deficiência física, contribuindo dessa forma para uma melhor qualidade de vida.

É factível que nos próximos anos ou décadas, os sistemas de reconhecimento de fala multi-modais (que utilizam sinais de fala e imagens) juntamente com os sistemas de síntese de fala, tornar-se-ão cada vez mais integrados para a composição de um sistema de informação unificado cujas entradas e saídas serão áudio e vídeo.

Assim, alguns aspectos fundamentais devem ser considerados para que os sistemas de reconhecimento se tornem cada vez mais presentes e integrados, em produtos e serviços freqüentemente utilizados no cotidiano, como uma interface simples e amigável entre a aplicação e o usuário.

1.2 O Contexto

As aplicações de reconhecimento de fala se encontram, em alguns casos, embarcadas em sistemas cujas limitações são dadas pela arquitetura e capacidade de processamento dos dispositivos eletrônicos. Dessa forma, devem existir restrições em relação à complexidade dos sistemas de reconhecimento, e ao mesmo tempo imposições sobre o desempenho do mesmo.

Os sistemas de reconhecimento de fala podem apresentar diversos graus de complexidade, em

função do objetivo para o qual são desenvolvidos. Em geral, para as tarefas de reconhecimento de comandos de voz, é possível se obter sistemas mais simples do que os projetados para o reconhecimento de fala contínua, por exemplo. Porém, em ambas, o reconhecedor precisa operar em tempo real e atender às limitações físicas de memória e de processamento do sistema no qual se encontra implementado (PCs, DSPs, etc.). Dessa forma, independentemente da finalidade, os sistemas de reconhecimento de fala devem possuir a menor complexidade possível, no intuito de atender às limitações práticas de implementação.

A complexidade do sistema deve ser determinada de acordo com o desempenho observado durante o reconhecimento de fala. Em geral, o aumento da complexidade do sistema é acompanhado pelo aumento de desempenho até um certo ponto, além do qual pode-se observar uma diminuição no desempenho. Por outro lado, à medida que se diminui arbitrariamente a complexidade do sistema, o processo de treinamento torna-se mais difícil. Portanto, a determinação da complexidade está diretamente relacionada com o desempenho do sistema.

1.3 A Motivação

O custo computacional do processo de reconhecimento está diretamente relacionado à complexidade dos HMMs que constituem o sistema, e a necessidade de se obter sistemas de reconhecimento de fala compactos e com o maior desempenho possível, motivaram o estudo das técnicas para a determinação da topologia dos HMMs, no intuito de se determinar sistemas de reconhecimento que apresentem um melhor compromisso entre tamanho e desempenho.

Neste contexto serão implementados três métodos presentes na literatura para a determinação da topologia de HMMs, e também será introduzido um novo método para esta finalidade, que incorpora conceitos relacionados com a eliminação do excesso de parâmetros presentes em alguns modelos e ao mesmo tempo com a discriminabilidade dos mesmos.

A idéia básica que sustenta o novo método se aproxima do princípio da parcimônia (Bie03), no sentido que busca a realização do treinamento do sistema sob condição do menor grau de liberdade possível, porém não utiliza o ajuste dos modelos aos dados de treinamento como medida para escolha da topologia. Assim, diferentemente de tal princípio, o novo método busca a determinação da menor complexidade que maximize a discriminabilidade dos modelos e que minimize a existência de parâmetros excedentes nos modelos. Dessa forma, pode-se obter um melhor aproveitamento do algoritmo de treinamento baseado em MLE, e ao mesmo tempo usufruir da simplicidade e velocidade do mesmo.

O novo método será apresentado para o problema específico de reconhecimento de fala contínua e utilizando HMMs, mas os conceitos que serão introduzidos podem ser estendidos, em primeira

instância, para qualquer modelo estatístico baseado em mistura de Gaussianas

1.4 O Embasamento na Literatura

O problema de modelagem estatística possui alguns aspectos que independem da tarefa específica para a qual os modelos são obtidos. Um problema clássico que precisa ser contornado é o da sobre-parametrização (Agu00), que pode ocorrer em modelos com grande grau de liberdade, ou seja, modelos com um número excessivo de parâmetros. Em geral, tais modelos apresentam baixa taxa de erro de treinamento, devido à alta flexibilidade, mas o desempenho do sistema utilizando dados de teste é quase sempre insatisfatório. Por outro lado, modelos com um número de parâmetros insuficiente não podem nem ao menos ser treinados. Neste ponto, observa-se que deve ser atingido um equilíbrio entre a treinabilidade e robustez do modelo, a fim de se obter um sistema com um desempenho elevado. No contexto de reconhecimento de fala, utiliza-se a taxa de reconhecimento e *accuracy* como medidas de desempenho, e no contexto do trabalho, utiliza-se o número total de componentes Gaussianas como medida do tamanho do sistema.

Outro aspecto importante que deve ser considerado é que a estimação confiável de parâmetros é realizada somente quando existem dados suficientes disponíveis para tal tarefa (RJ93). Sabendo-se que a base de dados de treinamento normalmente apresenta um número diferente de amostras de cada fone, é razoável se esperar que o número de amostras disponíveis seja também um fator limitante para o aumento do número de *clusters* (equivalente ao número de componentes Gaussianas) no modelo de uma determinada unidade acústica. Dessa forma, dependendo do número de amostras disponíveis, deve-se aumentar ou diminuir a resolução acústica dos modelos HMM no intuito de se realizar uma estimação de parâmetros confiável. Além disso, a complexidade das fronteiras das distribuições dos parâmetros acústicos também determina o número de componentes necessário para modelar corretamente as diferentes classes.

Existem também argumentos de ordem prática (LLNB04) que sustentam a idéia de se determinarem modelos HMM com um número variado de Gaussianas por estado. O custo computacional está diretamente relacionado com o número de componentes Gaussianas presentes no sistema. Como conseqüência imediata, o número de operações e a memória necessária para a realização das mesmas aumenta com o número de componentes. Portanto, as razões de natureza teórica e prática apresentadas acima servem como base de sustentação para a idéia de se obterem modelos acústicos com um número variado de componentes Gaussianas por estado.

Os métodos mais comuns presentes na literatura para a determinação da topologia de modelos estatísticos se encontram baseados na teoria da informação e codificação, na teoria Bayesiana e na verossimilhança completa (FJ02). O *Minimum Description Length* (MDL) (FJ02; Ris89; KTSS98;

TKS99) e o *Akaike Information Criterion* (AIC) (Aka74; MA94; WL94) são exemplos do primeiro caso. O *Bayesian Information Criterion* (BIC) (Sch78; CS00; CG98; BHS02), que é formalmente equivalente ao MDL, é um exemplo do segundo caso, enquanto os métodos baseados em medidas de entropia (CU93; CS96) são exemplos do terceiro caso.

No problema específico de reconhecimento de fala, os métodos acima ou variações dos mesmos podem ser empregados para a determinação da topologia dos modelos acústicos de fones independentes ou dependentes de contexto, palavras, etc. Além disso, é bastante comum no caso dos fones dependentes de contexto em LVCSR, a utilização de *phonetic decision tree-based state tying*, para fusão de estados de modelos distintos, de acordo com critérios de similaridade fonética (OWW94; YOW94; WOY94; RC00).

1.5 Objetivos

O principal objetivo deste trabalho é a determinação da topologia de HMMs para a obtenção de sistemas de reconhecimento de fala contínua que apresentem um melhor compromisso entre complexidade e desempenho. Além disso, o enfoque do processo para a determinação da topologia pode ser dado na economia de parâmetros, no desempenho durante o reconhecimento, ou em um caso intermediário, na busca pela maior economia desde que atendido um limiar mínimo de desempenho para o sistema.

Neste sentido, um novo Algoritmo de Eliminação de Gaussianas (GEA) será apresentado e os resultados serão comparados com os obtidos através de outros métodos presentes na literatura, e também com a estratégia de se utilizar um número fixo de componentes Gaussianas por estado.

1.6 Descrição do Trabalho

O trabalho se encontra dividido em 7 capítulos. O Capítulo 2 apresentará uma revisão da teoria de modelos de Markov e Modelos Ocultos de Markov, mostrando as principais diferenças entre ambos e justificando a aplicação dos HMMs para o problema de reconhecimento de fala.

O Capítulo 3 apresentará as bases de dados utilizadas nos experimentos e especificará os detalhes da implementação do sistema de treinamento de HMMs contínuos baseado em MLE. Além disso, também serão discutidos alguns resultados observados a partir de dados gerados artificialmente, relacionados com o ajuste dos parâmetros do modelo durante o processo de treinamento.

Na seqüência, o Capítulo 4 apresentará os resultados obtidos através de alguns métodos presentes na literatura, inclusive os fornecidos pelos sistemas que contêm um número fixo de Gaussianas por estado.

A versão inicial do novo método GEA será apresentado no Capítulo 5, e algumas modificações na proposta original serão apresentadas e justificadas no Capítulo 6.

Por fim, o Capítulo 7 apresentará as principais conclusões do trabalho e os pontos que ainda requerem investigações, como motivação para trabalhos futuros.

Capítulo 2

Revisão Teórica

2.1 Introdução

As abordagens iniciais dos sistemas de reconhecimento de fala eram baseadas em técnicas utilizadas em diversos problemas de reconhecimento de padrões, como por exemplo o método de *templates*. A idéia básica era estabelecer um padrão de referência, através da extração de médias, e para cada padrão novo desconhecido, medidas de distância espectral eram utilizadas como forma de comparação com os padrões conhecidos. Além disso, técnicas de programação dinâmica (*Dynamic Time Warping*) eram utilizadas para realizar alinhamentos temporais, no intuito de compensar o efeito de diferentes taxas de locução obtidas por diferentes locutores e para diferentes repetições de um mesmo padrão acústico (fones, palavras, etc.). É importante destacar que a utilização do método de *templates* é bastante eficaz em diversas aplicações dentro da área de reconhecimento de padrões. Entretanto, tal abordagem pode ser classificada como não-paramétrica e a caracterização dos sinais estocásticos de fala inerente às representações dos *templates* é freqüentemente inadequada.

Os sistemas de reconhecimento de fala, atualmente, são baseados em Modelos Ocultos de Markov (HMMs), que foram propostos inicialmente no final da década de 1960 e começo da década de 1970 (BP66; Bak75). Desde então, outras ferramentas matemáticas têm sido utilizadas neste sentido, como por exemplo Redes Neurais Artificiais, porém os sistemas de reconhecimento de fala considerados como referências na literatura (Cam02; LH89b; CEMEG$^+$99) empregam HMMs na modelagem acústica.

A suposição inicial do processo de modelagem estatística é que o sinal de fala pode ser caracterizado como um processo aleatório paramétrico e que os parâmetros podem ser estimados através de técnicas, cujos objetivos são definidos a priori, como por exemplo o método de maximização de verossimilhança (MLE) e o método discriminativo (MMIE, MCE, etc.) (BBdSM86; LER90). Assim, os parâmetros dos HMMs podem ser estimados de forma que, ao término do treinamento, o sistema

seja capaz de realizar o reconhecimento de fala, através de técnicas utilizadas para a decodificação dos sinais acústicos, e apresentar um desempenho satisfatório na tarefa para a qual foi desenvolvido.

Os HMMs modelam de forma eficiente a variabilidade temporal e espacial dos padrões acústicos, ou seja, as diferentes realizações da mesma unidade acústica (fones, palavras, etc.) e as diferentes unidades acústicas parametrizadas e representadas em um espaço de características. Assim, alguns aspectos conceituais dos HMMs devem ser apresentados no sentido de embasar a utilização de tal ferramenta para o problema de reconhecimento de fala. Além disso, deve-se também apresentar uma breve revisão das Cadeias de Markov no intuito de observar as principais diferenças entre tais modelos e os HMMs.

2.2 Cadeias de Markov

Um processo aleatório $X\left(t\right)$ é definido como Markoviano de primeira ordem se o futuro do processo, para uma dada condição do presente, independe do passado, ou seja, para instantes de tempo arbitrários $t_1 < t_2 < \cdots < t_k < t_{k+1}$, o processo de Markov é descrito por

$$P\left[X\left(t_{k+1}\right) = x_{k+1} | X\left(t_k\right) = x_k, \ldots, X\left(t_1\right) = x_1\right] = P\left[X\left(t_{k+1}\right) = x_{k+1} | X\left(t_k\right) = x_k\right], \quad (2.1)$$

se $X\left(t\right)$ for discreto, e é descrito por

$$P\left[a < X\left(t_{k+1}\right) \leq b | X\left(t_k\right) = x_k, \ldots, X\left(t_1\right) = x_1\right] = P\left[a < X\left(t_{k+1}\right) \leq b | X\left(t_k\right) = x_k\right], \quad (2.2)$$

se $X\left(t\right)$ for contínuo.

Além disso, se o processo aleatório de Markov assumir um número finito de estados, é então definido como Cadeia de Markov (Pap84; LG94). Assim, se $X\left(t\right)$ for uma Cadeia de Markov de primeira ordem, então a função distribuição cumulativa (CDF) conjunta, para um caso particular considerando apenas 3 instantes de tempo, pode ser definida como

$$\begin{aligned} P\left[X\left(t_3\right) = x_3, X\left(t_2\right) = x_2, X\left(t_1\right) = x_1\right] &= \\ P\left[X\left(t_3\right) = x_3 | X\left(t_2\right) = x_2, X\left(t_1\right) = x_1\right] \times P\left[X\left(t_2\right) = x_2, X\left(t_1\right) = x_1\right] &= \\ P\left[X\left(t_3\right) = x_3 | X\left(t_2\right) = x_2\right] \times P\left[X\left(t_2\right) = x_2, X\left(t_1\right) = x_1\right], & \end{aligned} \quad (2.3)$$

onde uma simplificação pode ser obtida pela definição de probabilidade condicional (Teorema de Bayes)

$$P\left[A,B\right]=P\left[B,A\right]=P\left[A|B\right]\times P\left[B\right]=P\left[B|A\right]\times P\left[A\right], \tag{2.4}$$

e pelo fato de se assumir que o processo é Markoviano de primeira ordem (o estado atual do processo depende apenas do estado no instante anterior).

Deve-se notar na Equação (2.3) que

$$P\left[X\left(t_2\right)=x_2,X\left(t_1\right)=x_1\right]=P\left[X\left(t_2\right)=x_2|X\left(t_1\right)=x_1\right]\times P\left[X\left(t_1\right)=x_1\right], \tag{2.5}$$

e substituindo (2.5) em (2.3), obtém-se então

$$\begin{aligned}P\left[X\left(t_3\right)=x_3,X\left(t_2\right)=x_2,X\left(t_1\right)=x_1\right]=\\ P\left[X\left(t_3\right)=x_3|X\left(t_2\right)=x_2\right]\times P\left[X\left(t_2\right)=x_2|X\left(t_1\right)=x_1\right]\times P\left[X\left(t_1\right)=x_1\right].\end{aligned} \tag{2.6}$$

Pode-se generalizar a Equação (2.6), por indução, para um caso com $K+1$ instantes de tempo, e dessa forma se obtém

$$\begin{aligned}P\left[X\left(t_{k+1}\right)=x_{k+1},X\left(t_k\right)=x_k,\ldots,X\left(t_1\right)=x_1\right]=\\ P\left[X\left(t_{k+1}\right)=x_{k+1}|X\left(t_k\right)=x_k\right]\ldots P\left[X\left(t_2\right)=x_2|X\left(t_1\right)=x_1\right]\times P\left[X\left(t_1\right)=x_1\right].\end{aligned} \tag{2.7}$$

Assim, é possível calcular a probabilidade da ocorrência de uma seqüência de eventos de uma Cadeia de Markov, na qual cada estado do processo está associado a um evento, e ambos são conhecidos.

2.3 Modelos Ocultos de Markov (HMM)

Os modelos de Markov possuem algumas restrições conceituais, como por exemplo o fato de cada estado estar associado a um evento conhecido de forma determinística, que limitam a utilização de tais modelos em diversos problemas de interesse. Assim, podem-se estender tais conceitos no intuito de incluir o caso onde cada evento observado é uma função probabilística do estado do processo, ou seja,

a seqüência de eventos pode ser observada, mas os estados do processo que geraram tal seqüência são ocultos, o que caracteriza em primeira instância os Modelos Ocultos de Markov (HMMs).

Neste cenário, torna-se então necessário saber como determinar um HMM capaz de explicar a seqüência de eventos observados, como estabelecer a correspondência entre os estados do modelo e os eventos observados e quantos estados devem existir no modelo, dentre outras questões.

Um dos parâmetros que caracterizam os HMMs é o número de estados do modelo, que freqüentemente possui um significado físico. Um exemplo ilustrativo pode ser encontrado na modelagem dos fonemas de uma determinada língua para a finalidade de reconhecimento de fala. Em geral, esta modelagem é realizada por HMMs contendo 3 estados, onde o primeiro e o terceiro estado tem como função modelar as transições entre os fonemas e os efeitos de coarticulação da fala, e o estado intermediário tem a função de modelar a parte mais estável da produção acústica do fonema.

Além disso, as probabilidades de transição de estados A_{ij}, as verossimilhanças $B_j\left(o_t\right)$ obtidas para cada evento observado o_t e as probabilidades iniciais π_i dos estados, para $1 \leq i, j \leq N_s$, em que N_s é o número total de estados nos modelos, também caracterizam os HMMs.

Neste ponto, uma vez estabelecidos os parâmetros que definem os HMMs, surgem três problemas que precisam ser resolvidos a fim de que os modelos possam ser utilizados em aplicações práticas.

- Problema 1
  - Dada uma seqüência de eventos observados e dado um modelo, a primeira questão está ligada ao cálculo da probabilidade da seqüência observada ter sido gerada pelo modelo, ou seja, determinar uma medida de quão próxima a seqüência de observações se encontra do modelo. A solução para este problema pode ser obtida pelo algoritmo *forward* (RJ93).
- Problema 2
  - A segunda questão está relacionada com a determinação da seqüência de estados ocultos, associada aos eventos observados. Na realidade podem existir diversas seqüências de estados possíveis, porém deve-se estabelecer um critério para a escolha da seqüência mais provável, como por exemplo aquela que fornecer a maior verossimilhança $P\left(\mathbf{O}\left|\lambda\right.\right)$, em que $\mathbf{O}$ é a seqüência de observações e λ é o conjunto de parâmetros que define o modelo. A solução para este problema pode ser obtida pelo algoritmo de Viterbi (RJ93).
- Problema 3
  - A terceira questão está relacionada com a estimação dos parâmetros do modelo, os quais devem ser ajustados de acordo com um método de treinamento (MLE, MMIE, MCE, dentre outros), de tal forma que o sistema apresente um desempenho satisfatório para a

aplicação prática. A solução para este problema, via MLE, pode ser obtida pelo algoritmo Baum-Welch (RJ93; BJM83; BBdSM86; PB02).

Os algoritmos citados como soluções para os três problemas dos Modelos Ocultos de Markov serão apresentados no Capítulo 3.

2.4 Metodologia

O primeiro passo para realização dos experimentos no intuito de atingir o principal objetivo deste trabalho, que consiste na determinação de sistemas de reconhecimento de fala que apresentem um melhor compromisso entre tamanho e desempenho, é a escolha da base de dados para o treinamento e para os testes. Assim, utilizaram-se duas bases de dados disponíveis no Laboratório de Processamento Digital de Fala do DECOM/FEEC/Unicamp. As bases possuem características diferentes, desde a língua de origem até quantidade de dados disponíveis. A partir de tais bases será possível avaliar a eficiência do novo método que será apresentado, de acordo com os diferentes aspectos de cada base.

Um conjunto de dados gerados artificialmente também será utilizado no intuito de permitir a visualização do comportamento do sistema de treinamento de HMMs contínuos, do ponto de vista do ajuste dos parâmetros dos modelos, assim como do ponto de vista da inicialização dos parâmetros.

Além disso, um sistema de treinamento de HMMs contínuos será desenvolvido no intuito de permitir uma maior assimilação dos problemas práticos de implementação e também para facilitar os testes posteriores com as técnicas de seleção de topologias dos modelos, que constituem o principal foco deste trabalho.

Três métodos presentes na literatura serão implementados para a determinação da topologia dos modelos HMMs. Esses métodos, juntamente com os sistemas contendo um número fixo de Gaussianas por estado (referência), serão utilizados para comparações com o novo método proposto neste trabalho. Serão avaliados o desempenho e o tamanho dos sistemas obtidos através de cada método.

O processo de decodificação será realizado a partir da adaptação de um sistema desenvolvido em um trabalho anterior (Yno99), e também através das ferramentas de um sistema de reconhecimento de fala que é utilizado como referência na literatura (Cam02). Também serão empregados diferentes modelos de linguagem no processo de decodificação, impondo-se dessa forma diferentes restrições ao processo de reconhecimento, a fim de se avaliar a robustez dos HMMs e separar melhor o efeito do modelo de linguagem do efeito da seleção de topologia.

O novo método será apresentado, desde a proposta inicial até a última versão, no intuito de mostrar a evolução da pesquisa no sentido de se determinar medidas mais apropriadas para o processo de seleção de estrutura, e também estimular a continuidade do trabalho em diversos aspectos.

No final do trabalho, serão avaliados alguns resultados importantes do ponto de vista fonético, no que diz respeito à segmentação automática via alinhamento forçado de Viterbi, e a relação dos mesmos com o reconhecimento de fala.

Os detalhes da metodologia e dos experimentos, assim como a análise dos resultados, serão apresentados em cada capítulo, à medida que os novos conceitos forem introduzidos.

Capítulo 3

Base de Dados e o Sistema de Treinamento de HMMs Contínuos Desenvolvido

3.1 Introdução

O ponto de partida para qualquer processo de modelagem estatística consiste na aquisição de dados do sistema para o qual se deseja obter um modelo e no processamento inicial dos mesmos, de tal forma que as características mais relevantes sejam extraídas. Neste trabalho foram utilizadas duas bases de dados: uma base de fala contínua em português do Brasil (Yno99), e outra base de fala contínua do inglês dos Estados Unidos (TIMIT) (Zha93; KVY93; ZWZ91; Nil94; LH89b; PB02). Tais bases possuem um conjunto de frases foneticamente balanceadas de forma que os fonemas da língua de origem apareçam com aproximadamente a mesma freqüência. Além disso, a quantidade de frases disponíveis para o treinamento é um fator importante que deve ser considerado durante o processo de treinamento, pois a limitação na quantidade de informação disponível acerca do sistema é um fator restritivo para alguns pontos do processo de modelagem, tais como a determinação da complexidade dos modelos e a escolha das unidades acústicas (fones independentes ou dependentes de contexto). Sob este aspecto, a base de dados em Português pode ser considerada pequena, enquanto a TIMIT pode ser considerada uma base grande (aproximadamente 3 vezes maior do que a base em Português). As principais características de cada base serão apresentados neste Capítulo, assim como as implicações de tais características para o treinamento e reconhecimento de fala contínua.

Os principais objetivos deste trabalho estão relacionados com o processo de treinamento de HMMs contínuos para sistema de reconhecimento de fala contínua e, dessa forma, desenvolveram-se todas as etapas necessárias para a obtenção dos modelos acústicos, desde a extração de parâmetros até o algoritmo de Baum-Welch, que realiza um treinamento baseado em MLE. Os algoritmos de decodificação foram adaptados de um trabalho anterior (Yno99) e das ferramentas do HTK (Cam02), a fim de

se ter uma referência para comparação com outros trabalhos presentes na literatura. Neste Capítulo também será descrito o algoritmo de Viterbi, além dos módulos de treinamento de HMMs contínuos.

3.2 Base de Dados

Os experimentos realizados neste trabalho utilizaram duas bases de dados, no intuito de se avaliar o processo de treinamento de acordo com a quantidade de dados disponíveis e de se ter segmentações de referência, permitindo a avaliação da qualidade das segmentações geradas automaticamente e que serão utilizadas no processo de treinamento. Sabe-se que a insuficiência de dados se torna um fator crítico para a modelagem à medida que se aumenta a resolução acústica dos modelos, e neste ponto os métodos para a determinação de topologia passam a desempenhar uma função de grande importância no processo de treinamento: ajustar o tamanho dos modelos à quantidade de informação disponível para o treinamento e à complexidade das distribuições no espaço de características acústicas.

3.2.1 Base de Dados Pequena em Português do Brasil

A base de dados pequena de fala contínua é constituída de 1600 sentenças foneticamente balanceadas, das quais 1200 são utilizadas para o treinamento e 400 para testes. É importante notar que no conjunto de 1600 sentenças, existem apenas 200 frases distintas (ASM92), elaboradas a partir de um vocabulário contendo 694 palavras.

As sentenças foram produzidas por 40 locutores, sendo 20 do sexo masculino e 20 do sexo feminino. Além disso, as sentenças de treinamento e as de teste foram produzidas por diferentes locutores. Entretanto, deve-se destacar que todas as frases utilizadas nas sentenças de teste foram utilizadas também nas sentenças de treinamento. Os sinais acústicos foram amostrados a 11025Hz, com 16 bits por amostra. Outro ponto importante diz respeito à origem dos locutores presentes na base que, neste caso, são predominantemente do estado de São Paulo e, portanto, a base não cobre aspectos fonéticos e lingüísticos de todas a regiões do Brasil.

As 1600 sentenças possuem as respectivas transcrições fonéticas geradas manualmente, utilizando 35 unidades fonéticas distintas da língua portuguesa. Assim, no processo de modelagem, serão determinados 36 HMMs para representar os 35 fones independentes de contexto e mais o silêncio. A Tabela 3.1 mostra os fonemas da língua portuguesa que serão modelados, de acordo com a classificação dos mesmos. Além disso, cada fone será modelado por um HMM com três estados, os quais também se encontram especificados.

Tal base foi utilizada nos experimentos para a obtenção de sistemas de reconhecimento de fala contínua independentes de locutor, considerando apenas fones independentes de contexto durante a

Tab. 3.1: Símbolos fonéticos, símbolos utilizados nas transcrições fonéticas e exemplos.

Classe	Símbolo Fonético	Símbolo Utilizado nas Transcrições	Exemplo	Estados do HMM
Vogal	a	a	ação = **a** s an un	4-6
Vogal	e	e	eleito = **e** l **e** y t u	10-12
Vogal	ɛ	E	pele = p **E** l y	13-15
Vogal	i	i	sido = s **i** d u	19-21
Semi-Vogal	j	y	flui = f l u **y**	22-24
Vogal	o	o	boa = b **o** a	28-30
Vogal	ɔ	O	copa = k **O** p a	31-33
Vogal	u	u	luz = l **u** s	37-39
Vogal	ɜ̃ː	an	amanhã = a m **an** N **an**	7-9
Vogal	ẽː	en	lenta = l **en** t a	16-18
Vogal	ĩː	in	informática = **in** f o R m a T i k a	25-27
Vogal	õː	on	sombra = s **on** b r a	34-36
Vogal	ũː	un	um = **un**	40-42
Laterais	l	l	lado = **l** a d u	64-66
Laterais	ʎ	L	falha = f a **L** a	67-69
Não-Laterais	ɾ	r	irá = i **r** a	82-84
Não-Laterais	ɣ	rr	rua = **rr** u a	85-87
Não-Laterais	ɹ	R	inverno = in v E **R** n u	88-90
Nasais	m	m	maratona = **m** a r a t o n a	70-72
Nasais	n	n	nove = **n** O v y	73-75
Nasais	ɲ	N	conheceram = k on **N** e s e r an un	76-78
Oclusivas	b	b	belo = **b** E l u	43-45
Oclusivas	d	d	deve = **d** E v y	46-48
Oclusivas	g	g	garota = **g** a r o t a	55-57
Oclusivas	k	k	calmo = **k** a u m u	61-63
Oclusivas	p	p	palha = **p** a L a	79-81
Oclusivas	t	t	tempo = **t** en p u	94-96
Fricativas	f	f	feiras = **f** e y r a z	52-54
Fricativas	ʒ	j	já = **j** a	58-60
Fricativas	s	s	seco = **s** e k u	91-93
Fricativas	v	v	vila = **v** i l a	100-102
Fricativas	ʃ	x	chegar = **x** e g a R	103-105
Fricativas	z	z	zé = **z** E	106-108
Africadas	ʤ	D	diálogo = **D** i a l o g u	49-51
Africadas	ʧ	T	título = **T** i t u l u	97-99
Silêncio		#	bom dia = # b on D i a #	1-3

modelagem.

Por fim, deve-se notar que a base pequena não possui segmentações acústicas de referência e, portanto, as segmentações geradas automaticamente a partir de alinhamentos forçados de Viterbi serão utilizadas sempre que necessárias.

3.2.2 Base de Dados em Inglês dos Estados Unidos (TIMIT)

A base de dados TIMIT é formada por 6300 sentenças foneticamente balanceadas, das quais 5040 são utilizadas para o propósito de reconhecimento de fones contínuos (LH89a; LG93; ZWZ91), sendo que as locuções foram produzidas por 630 locutores. O conjunto de treinamento é composto de 3696 sentenças, e o conjunto de teste pode ser completo, contendo 1344 sentenças (*complete test*), ou básico, contendo 192 sentenças (*core test*). Nos experimentos realizados com a TIMIT foi escolhido o conjunto completo de sentenças de teste. Deve-se observar que, dentre as 1344 sentenças de teste, existem 624 frases distintas. Os sinais acústicos referentes às 6300 sentenças da base encontram-se amostrados a 16KHz e com uma resolução de 16 bits por amostra.

A TIMIT cobre os 8 maiores dialetos do inglês americano. Além disso, existem três tipos de sentenças:

- de dialeto (*" 1260 SA sentences"*): geradas para expor as variações de dialetos dos locutores.
- foneticamente compactas (*"3150 SX sentences"*): geradas para fornecer boa cobertura de pares de fones, com ocorrências extras de contextos fonéticos considerados difíceis ou de interesse particular.
- foneticamente diversificado (*"1890 SI sentences"*): selecionadas de fontes de texto existentes para acrescentar diversidade em tipos de sentenças e em contextos fonéticos.

As sentenças 1260 SA não são utilizadas pelo fato de serem produzidas por todos os 630 locutores da base, o que deve ser evitado no problema de reconhecimento de fala, onde os locutores da base de teste não podem se encontrar ao mesmo tempo na base de treinamento. Além disso, não existem frases utilizadas simultaneamente nas sentenças de treinamento e de teste. Deve-se destacar também que a TIMIT é constituída de 2342 frases distintas (2 SA + 450 SX + 1890 SI), formadas por um vocabulário de 6240 palavras.

As transcrições fonéticas da TIMIT possuem um conjunto de 64 fonemas distintos, o que por convenção é simplificado para 48 fonemas que são utilizados durante o treinamento dos modelos acústicos e, posteriormente, reduzidos novamente para 39 fonemas, que são utilizados no reconhecimento (LH89a). Os fones independentes de contexto utilizados durante o treinamento encontram-se indicados nas Tabelas 3.2 e 3.3.

Tab. 3.2: Símbolos utilizados nas transcrições fonéticas de oclusivas, fricativas, nasais, africadas e silêncios.

Classe	Símbolo	Exemplo
Stops	b	bee = **vcl b** iy
Stops	d	day = **vcl d** ey
Stops	g	gay = **vcl g** ey
Stops	p	pea = **cl p** iy
Stops	t	tea = **cl t** iy
Stops	k	key = **cl k** iy
Stops	dx	dirty = vcl d er **dx** iy
Fricativas	s	sea = **s** iy
Fricativas	sh	she = **sh** iy
Fricativas	z	zone = **z** ow n
Fricativas	zh	azure = ae **zh** er
Fricativas	f	fin = **F** ih n
Fricativas	th	thin = **th** ih n
Fricativas	v	van = **v** ae n
Fricativas	dh	then = **th** e n
Nasais	m	mom = **m** aa **m**
Nasais	n	noon = **n** uw **n**
Nasais	ng	sing = s ih **ng**
Nasais	en	button = vcl b ah cl t **en**
Africadas	jh	joke = **vcl jh** ow cl k
Africadas	ch	choke = **cl ch** ow cl k
unvoiced closure	cl	
voiced closure	vcl	
Silêncio epentético	epi	
Silêncio	sil	

Deve-se notar que as transcrições utilizam duas representações para o silêncio:“epi” (silêncio epentético) e “sil” (silêncio prolongado, geralmente no começo, término das sentenças e pausas durante a produção acústica, associadas às pontuações da sentença). Além disso, as transcrições originais estabelecem diferentes símbolos para os *unvoiced closure* (“pcl”, “tcl” e “kcl”), que precedem os fonemas “p”, “t” e “k”. Para efeito de simplificação, todos os *unvoiced closure* passam a ser representados pelo símbolo “cl”. De forma semelhante, também estabelecem diferentes símbolos para os *voiced closure* (“bcl”, “dcl” e “gcl”), que precedem os fonemas “b”, “d” e “g”. Os *voiced closure* passam a ser representados pelo símbolo “vcl”.

Outras substituições também são realizadas nas transcrições originais no intuito diminuir a quantidade de fones para o processo de modelagem. Os fones “ax-h”, “ux”, “axr”, “em”, “nx”, “eng”, “j”, “hv” e “pau” são substituídos por “ax”, “uw”, “er”, “m”, “n”, “ng”, “jh”, “hh” e “sil”, respec-

Tab. 3.3: Símbolos utilizados nas transcrições fonéticas, de vogais, semi-vogais e *glides*.

Classe	Símbolo	Exemplo
Vogais	iy	beet = vcl b **iy** cl t
Vogais	ih	bit = vcl b **ih** cl t
Vogais	eh	bet = vcl b **eh** cl t
Vogais	ey	bait = vcl b **ey** cl t
Vogais	ae	bat = vcl b **ae** cl t
Vogais	aa	bott = vcl b **aa** cl t
Vogais	aw	bout = vcl b **aw** cl t
Vogais	ay	bite = vcl b **iy** cl t
Vogais	ah	but = vcl b **ah** cl t
Vogais	ao	bought = vcl b **ao** cl t
Vogais	oy	boy = vcl b **oy**
Vogais	ow	boat = vcl b **ow** cl t
Vogais	uh	book = vcl b **uh** cl k
Vogais	uw	boot = vcl b **uw** cl t
Vogais	er	bird = vcl b **er** vcl d
Vogais	ax	about = **ax** vcl b aw cl t
Vogais	ix	debit = vcl d eh vcl b **ix** cl t
Semi-Vogais/Glides	l	lay = **l** ey
Semi-Vogais/Glides	r	ray = **r** ey
Semi-Vogais/Glides	w	way = **w** ey
Semi-Vogais/Glides	y	yacht = **y** aa cl t
Semi-Vogais/Glides	hh	hay = **hh** ey
Semi-Vogais/Glides	el	bottle = vcl b aa cl t **el**

tivamente. Além disso, a *glottal stop* "q" é removida das transcrições originais. Todas as alterações são realizadas em conformidade com os padrões estabelecidos pela Carnegie Mellon University/Massachusetts Institute of Technology (CMU/MIT).

Por fim, após o processo de reconhecimento, novos agrupamentos são realizados de tal forma que fones semelhantes sejam associados ao mesmo símbolo nas transcrições obtidas com o reconhecimento, obtendo-se 39 fones distintos. A Tabela 3.4 apresenta tais agrupamentos.

A TIMIT possui a segmentação acústica manual, em termos de fones e de palavras, de todas as 6300 sentenças, que pode ser utilizada durante o processo de treinamento dos modelos, além de permitir também a análise das segmentações geradas pelos sistemas de reconhecimento, a partir de alinhamentos forçados de Viterbi.

Em última análise, o treinamento de sistema de reconhecimento a partir da TIMIT será realizado com uma quantidade de informações consideravelmente maior do que as disponíveis na base pequena em Português, o que pode permitir a obtenção de modelos acústicos mais complexos, quando neces-

Tab. 3.4: Agrupamentos realizados após o reconhecimento.

Fones utilizados no treinamento	Nova representação após o reconhecimento
“sil”, “cl”, “vcl” e “epi”	“sil”
“el” e “l”	“l”
“en” e “n”	“n”
“sh” e “zh”	“sh”
“ao” e “aa”	“aa”
“ih” e “ix”	“ih”
“ah” e “ax”	“ah”

sário, e mais robustos. Em contrapartida, utilizaram-se 48 fones no processo de modelagem acústica com a TIMIT, enquanto eram utilizados apenas 36 com a base pequena em Português. A TIMIT cobre todos os dialetos do Inglês americano, enquanto a base pequena cobre basicamente os regionalismos do estado de São Paulo. Dessa forma, os experimentos explorarão bases com quantidade e variabilidade de informação diferentes, sendo tais características de fundamental importância para a eficácia do processo de treinamento visando a obtenção de modelos robustos, que apresentem um elevado desempenho no reconhecimento.

3.3 O Sistema Desenvolvido para o Treinamento de HMMs Contínuos

O sistema desenvolvido para o treinamento de HMMs contínuos para o reconhecimento de fala, consiste basicamente de 5 módulos interligados: extração de parâmetros, inicialização, algoritmo de Viterbi, Baum-Welch e seleção de topologia. Neste Capítulo, serão abordados os cinco módulos citados, com embasamento na teoria presente na literatura, enquanto o novo método de seleção de topologia será discutido nos Capítulos seguintes. Além disso, alguns aspectos relacionados com a decodificação também serão apresentados, de forma a esclarecer as estratégias adotadas de acordo com a base de dados utilizada: base pequena em Português (reconhecimento de fala contínua) e TIMIT (reconhecimento de fones contínuos).

3.3.1 Módulo de Extração de Parâmetros

A extração de características dos sistemas físicos é o ponto de partida para o processo de modelagem e, portanto, a escolha apropriada de tais características é de fundamental importância para a obtenção de modelos matemáticos representativos dos fenômenos físicos observados. Nos experi-

mentos, seguiram-se as convenções adotadas na área de reconhecimento de fala, de forma a tornar o sistema mais próximo possível do estado da arte.

Front-end

Os sinais acústicos receberam o seguinte tratamento inicial:

- Subtração do nível DC de cada sentença.
- Pré-ênfase através do filtro passa-altas $\frac{Y[z]}{X[z]} = (1 - c_f z^{-1})$, onde $c_f = 0,95$ nos experimentos com a base pequena e $c_f = 0,97$ nos experimentos com a TIMIT (valores escolhidos de acordo com a literatura).
- Janelamento de Hamming $v[n] = y[n] \cdot j[n]$, utilizando uma janela de 20ms, com deslocamentos de 10ms, de acordo com a Equação (3.1). Para o sinal amostrado a 11025Hz (base em Português) utilizou-se N=221, e para o sinal amostrado a 16KHz (TIMIT), utilizou-se N=320.

$$j[n] = \begin{cases} 0,54 - 0,46\cos\left(2\pi n/(N-1)\right), & 0 \leq n \leq N-1. \\ 0, \quad \text{para } \forall n \text{ fora do intervalo.} \end{cases} \tag{3.1}$$

Neste ponto, deve-se observar que a última janela sobre o sinal acústico dificilmente possuirá o número exato de amostras que devem estar contidas na janela e, dessa forma, acrescenta-se zeros no intuito de se ter o número correto de amostras.

O sinal janelado $v[n]$ é submetido a um banco de filtros cujas saídas serão utilizadas para os cálculos dos parâmetros espectrais. Neste sentido, calcula-se a FFT $V(k)$ do sinal janelado $v[n]$ (512 pontos para o sinal amostrado a 11025Hz e 1024 pontos para o sinal amostrado a 16KHz), para então se obter $|V(k)|^2$.

Na seqüência, realiza-se o produto espectral $F(k) = |V(k)|^2 \cdot H(k)$, em que $H(k)$ é a amplitude do banco de filtros triangulares nas freqüências discretas de interesse, que se encontra representada na Figura 3.1.

A freqüência central e a largura de banda (BW) de cada filtro do banco (Pic93) estão indicadas na Tabela 3.5. Deve-se destacar que, para uma freqüência de amostragem do sinal acústico de 11025Hz, apenas os 21 primeiros filtros são utilizados pelo banco, enquanto que, para a freqüência de amostragem de 16KHz, os 23 primeiros filtros são empregados no banco. Nos experimentos realizados com a TIMIT, cujos sinais acústicos se encontram amostrados a 16KHz, utiliza-se um banco de filtros ligeiramente diferente do apresentado na Tabela 3.5, que foi implementado através das ferramentas do HTK (26 filtros são utilizados, ao invés de 23).

O logaritmo da energia na saída de cada filtro do banco é então calculado pela Equação (3.2).

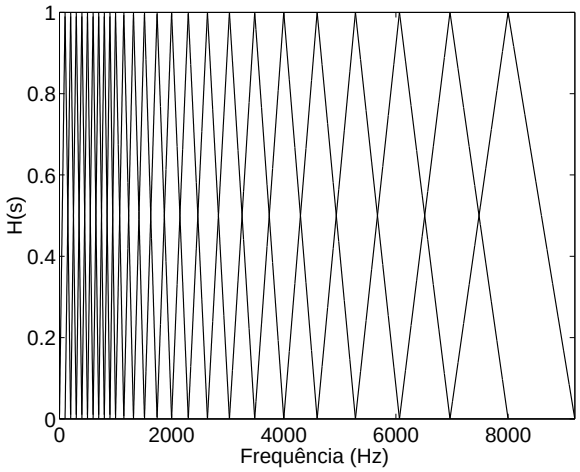


Fig. 3.1: Função de Transferência do Banco de Filtros.

Tab. 3.5: Especificações do Banco de filtros triangulares.

Freqüência Central (Hz)	BW (Hz)	Freqüência Central (Hz)	BW (Hz)
100	200	1741	484
200	200	2000	556
300	200	2297	640
400	200	2639	734
500	200	3031	844
600	200	3482	968
700	200	4000	1112
800	200	4595	1278
900	200	5278	1468
1000	248	6063	1686
1148	320	6964	1938
1320	368	8000	2226
1516	422	9190	2558

$$E_m = \log_{10}\left(\sum F\left(k\right)\right). \tag{3.2}$$

A partir da energia calculada na saída de cada filtro, obtêm-se então os valores dos coeficientes mel-cepstrais, através da Equação (3.3 (DM80)).

$$MFCC_i = \sum_{m=1}^{M}\left(E_m.\cos\left[i.\left(m-\frac{1}{2}\right).\frac{\pi}{M}\right]\right), \quad i = 1,2,3,\ldots,12, \tag{3.3}$$

em que “M” é o número total de filtros do banco, “i” é o número do coeficiente mel-cepstral (nos experimentos, utilizaram-se 12 coeficientes mel-cepstrais), e E_m é a energia calculada na saída do *m*-

ésimo filtro. Uma vez calculados os coeficientes MFCC, realiza-se então a subtração do vetor média de MFCCs, no intuito de diminuir o efeito do ruído ambiente presentes nos dados (remoção da média espectral). Deve-se notar que no caso de se processar cada quadro antes do término da locução, o que ocorre em aplicações de tempo real, não é possível se efetuar a remoção da média espectral.

Outro parâmetro calculado para a composição do vetor de características acústicas é o log-energia, que é definido pela Equação (3.4).

$$\text{LogEnergia} = 10\log_{10}\left(\sum_{n=0}^{N-1}\left(v[n]\right)^2\right) \tag{3.4}$$

De forma semelhante aos coeficientes MFCC, determina-se o quadro com maior log-energia para uma dada locução e subtrai-se tal valor dos log-energia calculados para os demais quadros (normalização da energia).

Por fim, calculam-se os parâmetros Δ e $\Delta\Delta$, que são aproximações das derivadas de primeira e segunda ordem dos coeficientes MFCC e log-energia, de acordo com a Equação (3.5).

$$\Delta_f\left(i\right) = \frac{1}{2K+1}\sum_{k=-K}^{K} k.p_{f+k}\left(i\right), \tag{3.5}$$

em que k é o número de quadros adjacentes utilizados no cálculo dos parâmetros Δ e "i" é o *i-ésimo* coeficiente MFCC ou o parâmetro log-energia. Nos experimentos realizados com a base pequena utilizou-se $K = 1$, enquanto nos experimentos com a TIMIT, utilizou-se $K = 2$, de acordo com trabalhos anteriores encontrados na literatura (Yno99; Val95).

O vetor de parâmetros acústicos pode então ser composto pelos 39 parâmetros calculados, conforme indicado na Equação (3.6).

$$\mathbf{P_f} = [1\ \text{LogEnergia},\ 12\ \text{MFCC},\ 1\ \Delta_{\text{LogEnergia}},\ 12\ \Delta_{\text{MFCC}},\ 1\ \Delta\Delta_{\text{LogEnergia}},\ 12\ \Delta\Delta_{\text{MFCC}}] \tag{3.6}$$

Portanto o espaço de características acústicas tem dimensão 39 e, para efeito de simplificação na modelagem, assume-se que os parâmetros são estatisticamente independentes entre si.

3.3.2 Módulo de Inicialização dos Parâmetros do Modelo

A estratégia de inicialização dos parâmetros do modelo é escolhida de acordo com a disponibilidade da segmentação acústica das sentenças de treinamento (segmentações de referência). A primeira estratégia é adotada no caso de se ter a segmentação acústica manual das sentenças de treinamento ou no caso de se ter uma segmentação acústica gerada automaticamente por um sistema de reconhe-

cimento de fala. A segunda estratégia é adotada no caso em que a base de dados de treinamento não é segmentada. Na seqüência são apresentados os principais aspectos de cada estratégia.

Inicialização dos Parâmetros dos Modelos para uma Base de Dados Segmentada

A disponibilidade de uma segmentação acústica das sentenças de treinamento, permite que os parâmetros extraídos dos quadros correspondentes ao intervalo de duração de um dado fone sejam separados em três partes, cada uma associada a um dos três estados do HMM, e utilizados no algoritmo *Segmental K-means* para a determinação inicial dos parâmetros de cada estado do modelo.

O algoritmo *Segmental K-means* consiste basicamente na determinação iterativa de clusters. De forma resumida, tal algoritmo pode ser descrito nos seguintes passos:

- 1 - Estabelecer o número de clusters desejados (número de Gaussianas associado a cada estado) e o número de divisões n_d a cada iteração.
- 2 - A partir do conjunto de amostras, determinar o valor médio inicial "μ".
- 3 - Somar uma perturbação "$\pm\epsilon$" (tipicamente $0,01 \leq \epsilon \leq 0,05$) às médias dos n_d maiores *clusters* , obtendo-se então novos centros $\mu \pm \epsilon$.
- 4 - Calcular a distância Euclidiana de cada amostra para cada um dos centros existentes.
- 5 - Associar cada amostra ao centro mais próximo, de acordo com a medida de distância Euclidiana.
- 6 - Após o término das associações, re-calcular o valor médio de cada um dos clusters obtidos no item anterior.
- 7 - Se o tamanho desejado para o sistema não foi atingido, identificar os n_d maiores *clusters*, os quais serão divididos na seqüência.
- 8 - Repetir os itens de 3 a 8 se o número de Gaussianas desejado para o sistema não foi atingido, caso contrário finalizar o algoritmo.

Deve-se notar que algumas variações em relação ao número de divisões (*splitting*) realizadas em cada iteração do algoritmo podem ser adotadas. Além disso, trabalhos presentes na literatura recomendam que, para o problema de reconhecimento de fala, deve-se adotar $n_d \leq 2$ e realizar algumas épocas de treinamento via MLE após as divisões em cada iteração (Cam02).

Em geral, os modelos associados a cada unidade acústica consistem de três estados e, portanto, o conjunto de amostras de um determinado fone extraídas da base segmentada é divido por três, de

forma que o algoritmo *Segmental K-means* seja aplicado para a inicialização dos parâmetros de cada estado. Uma vez atingido o número de Gaussianas (*clusters*) desejado para cada estado, as médias, variâncias e pesos de cada Gaussiana (*cluster*) que pertence a um determinado estado do modelo de um fone são calculados pelas Equações (3.7)-(3.9)

$$\boldsymbol{\mu}_{\,gaussiana} = \frac{1}{N_{cluster}} \sum_{n=1}^{N_{cluster}} \mathbf{x}_{\,n}, \tag{3.7}$$

$$\boldsymbol{\sigma}^2_{gaussiana} = \frac{1}{N_{cluster}-1} \sum_{n=1}^{N_{cluster}} (\mathbf{x}_{\,n} - \boldsymbol{\mu}_{\,})^2, \tag{3.8}$$

$$c_{gaussiana} = \frac{N_{cluster}}{N_{estado}}, \tag{3.9}$$

em que $N_{cluster}$ é o número de amostras associadas ao cluster, N_{estado} é o número de amostras associadas a um determinado estado do modelo, e $\mathbf{x_n}$ é o *n-ésimo* vetor amostra. Assim, cada estado possui N_{estado} amostras, as quais são agrupadas em clusters, cada qual com um próprio número $N_{cluster}$ de amostras.

As probabilidades iniciais de transição de estados para um HMM do tipo *left-to-right*, com três estados, estão representadas na Figura 3.2.

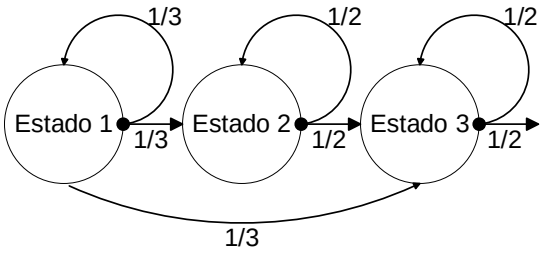


Fig. 3.2: HMM do tipo *left-to-right* com três estados.

A TIMIT possui uma segmentação de referência e, portanto, será utilizada para a inicialização dos HMMs.

Inicialização dos Parâmetros dos Modelos para uma Base de Dados Não-Segmentada

A inicialização dos HMMs no caso em que não se dispõe de sentenças de treinamento segmentadas, assume que todos os fones da sentença têm a mesma duração e, portanto, realiza um segmentação uniforme de tal forma que todos os fones possuam aproximadamente o mesmo número de quadros. Tal suposição é certamente incorreta, porém, na ausência de outra alternativa, é utilizada como ponto de partida para a obtenção dos parâmetros iniciais dos modelos.

Assim, uma vez realizada a segmentação uniforme das sentenças de treinamento, aplica-se o algoritmo *Segmental K-means*, conforme apresentado anteriormente, para então se obter os parâmetros iniciais dos modelos. Na seqüência, realiza-se um alinhamento forçado de Viterbi utilizando-se os HMMs iniciais, no intuito de se obter uma nova segmentação para as sentenças. A partir desta nova segmentação, aplica-se o algoritmo *Segmental K-means*, pelo menos mais uma vez, para então se obter os modelos iniciais que serão utilizados na primeira época de treinamento do algoritmo Baum-Welch.

É importante notar que, neste caso, o erro inicial cometido faz com que algumas Gaussianas pertencentes ao modelo de um determinado fone se encontrem sobre a distribuição real de outro fone. Assim, se houver excesso de Gaussianas em tais modelos, o processo de treinamento baseado em MLE pode levar à obtenção de sistemas pouco robustos e com baixo desempenho no reconhecimento.

O algoritmo de Baum-Welch tem se mostrado bastante eficiente para o treinamento baseado em MLE, mesmo quando se parte da segmentação uniforme das sentenças de treinamento. Uma estratégia de inicialização presente na literatura é a de gerar modelos com apenas uma Gaussiana por estado e, aplicar algumas épocas (em geral 2 ou 3) do algoritmo de Baum-Welch, para então se realizar o *splitting* das Gaussianas dos modelos, que é seguido novamente de algumas épocas de treinamento. O processo é repetido iterativamente até que o número de Gaussianas dos modelos seja atingido.

A base de dados pequena não possui uma segmentação de referência para as sentenças de treinamento e, dessa forma, os experimentos realizados com tal base utilizaram a segmentação uniforme como ponto de partida, seguido dos passos apresentados anteriormente, os quais se encontram resumidos abaixo.

- 1 - Segmentação uniforme das sentenças de treinamento.
- 2 - Aplicação do algoritmo *Segmental K-means*.
- 3 - Alinhamento forçado de Viterbi a partir dos modelos gerados no passo anterior.
- 4 - Aplicação novamente do algoritmo *Segmental K-means* para a obtenção dos modelos iniciais que serão utilizados pelo treinamento via MLE.

3.3.3 Módulo do Algoritmo de Viterbi

O algoritmo de Viterbi tem função fundamental no processo de decodificação realizado por Modelos Ocultos de Markov, pois somente a partir da decodificação é possível descobrir a seqüência de estados mais provável, para um dado conjunto de observações $\mathbf{O} = [\mathbf{o}_1,\ \mathbf{o}_2, \ldots, \mathbf{o}_t]$. No caso de reconhecimento de fala, em que se modelam as unidades acústicas através de uma seqüência de três estados de um HMM, os fones independentes ou dependentes de contexto e, conseqüentemente, as

palavras produzidas durante a locução, são descobertas a partir da seqüência de estados com maior verossimilhança obtida através de métodos de decodificação baseados no algoritmo de Viterbi. Além disso, mesmo quando se conhece a transcrição fonética da sentença produzida, é possível se utilizar o algoritmo de Viterbi no intuito de se descobrir as fronteiras entre as unidades acústicas, ou seja, realizar o alinhamento forçado de Viterbi (segmentação da sentença).

Assim, em um problema genérico, o algoritmo de Viterbi determina a seqüência mais provável de estados $\mathbf{Q} = [q_1,\ q_2, \ldots, q_t]$, para um dado conjunto de observações **O** até o instante "t", a partir do valor de maior probabilidade definido pela Equação (3.10)

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P\left[q_1, q_2, \ldots, q_{t-1}, q_t = i, \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t | M\right], \tag{3.10}$$

em que M é o conjunto de parâmetros que define o HMM utilizado pelo algoritmo de Viterbi.

Por indução, o melhor caminho para o instante $t+1$ pode ser obtido através da Equação (3.11)

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i)\, a_{ij}\right] .b_j(\mathbf{o}_{t+1}), \tag{3.11}$$

em que a_{ij} é a probabilidade de transição do estado "i" para o estado "j", e $b_j(\mathbf{o}_{t+1})$ é a verossimilhança calculada pela mistura de N_g Gaussianas do modelo do estado "j", para o vetor de parâmetros acústicos $\mathbf{o}_{t+1}$. O valor de verossimilhança obtido a partir da mistura de Gaussianas multidimensionais que formam o modelo do estado "j" é calculada pela Equação (3.12)

$$b_j(\mathbf{o}_t) = \sum_{m=1}^{N_g} c_{jm}.G\left(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm}\right), \tag{3.12}$$

sendo $G\left(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm}\right)$ a PDF Gaussiana multidimensional definida pela Equação (3.13)

$$G\left(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm}\right) = \frac{1}{(2\pi)^{\dim/2}\left|\mathbf{U}_{jm}\right|^{1/2}} \exp\left[-\left(\mathbf{o}_t - \boldsymbol{\mu}_{jm}\right) \cdot \mathbf{U}_{jm}^{-1} \cdot \left(\mathbf{o}_t - \boldsymbol{\mu}_{jm}\right)' \Big/ 2\right] \tag{3.13}$$

em que $\mathbf{U}_{jm}$ e $\boldsymbol{\mu}_{jm}$ são, respectivamente, a matriz de covariância e o vetor média, da componente Gaussiana "m", do modelo do estado "j". Além disso, $|\mathbf{U}_{jm}|$ é o determinante da matriz de covariância, $\mathbf{U}_{jm}^{-1}$ é o inverso da matriz de covariância e c_{jm} é o peso da componente "m".

Deve-se observar que a suposição de independência estatística entre as características acústicas, resulta em uma matriz de covariância diagonal, e o valor de verossimilhança calculado para cada componente é resultado do produto de "dim" Gaussianas unidimensionais (neste trabalho, dim=39).

Assim, assumindo que o vetor de observações **O** possua T elementos, e que existam N_s estados distintos no modelo, o algoritmo de Viterbi pode ser resumido em 4 passos:

- 1 - Inicialização

$$\delta_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N_s \tag{3.14}$$

$$\psi_1(i) = 0 \tag{3.15}$$

- 2 - Parte Recursiva

$$\delta_t(j) = \max_{1 \leq i \leq N_s} [\delta_{t-1}(i) a_{ij}] . b_j(o_t), \quad \begin{cases} 2 \leq t \leq T \\ 1 \leq j \leq N_s \end{cases} \tag{3.16}$$

$$\psi_t(j) = \arg \max_{1 \leq i \leq N_s} [\delta_{t-1}(i) a_{ij}], \quad \begin{cases} 2 \leq t \leq T \\ 1 \leq j \leq N_s \end{cases} \tag{3.17}$$

- Finalização

$$P^* = \max_{1 \leq i \leq N_s} [\delta_T(i)] \tag{3.18}$$

$$q_T^* = \arg \max_{1 \leq i \leq N_s} [\delta_T(i)] \tag{3.19}$$

- Seqüência de Estados mais Provável

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right), \quad t = T-1,\ T-2, \ldots, 1 \tag{3.20}$$

Finalmente, a seqüência de estados mais provável é fornecida pela variável q_t^*. Um exemplo ilustrativo encontra-se na Figura 3.3, que mostra para a sentença “Olá !”, cuja transcrição é “# o l a #”, os estados dos modelos de cada fone, as observações que correspondem aos vetores de parâmetros acústicos, e os possíveis caminhos (sequência de estados) do alinhamento forçado de Viterbi.

É importante notar que, no caso de reconhecimento de fala, em geral o começo e o término de todas as sentenças são marcados por um intervalo de silêncio no sinal acústico e, portanto, $\pi_i = 1$ para o primeiro estado do modelo do silêncio e $\pi_i = 0$ para os demais estados. Assim, no processo de decodificação, o primeiro estado da sequência (primeiro estado do modelo do silêncio) e o último estado da sequência (último estado do modelo do silêncio) são conhecidos.

Um aspecto prático que deve ser considerado está relacionado com os valores assumidos pela variável δ, que podem ser significativamente menores que 1, podendo exceder até mesmo a precisão

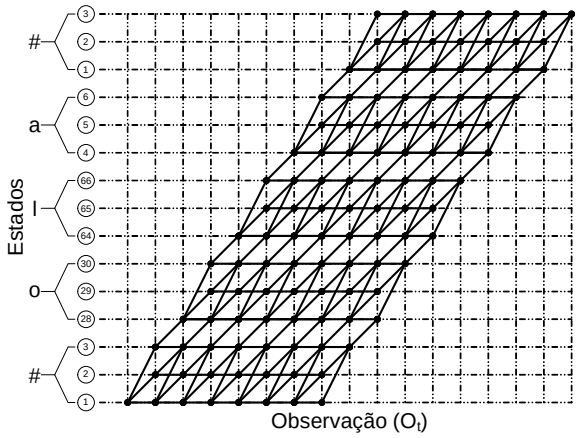


Fig. 3.3: Exemplo de aplicação do algoritmo de Viterbi.

numérica *double* da máquina, resultando em *underflow*. No intuito de contornar tal problema, utiliza-se o logaritmo da expressão (3.16), ou seja,

$$\delta_t(j) = \log_{10}\left\{\left\{\max_{1\leq i\leq N_s}\left[\delta_{t-1}(i)\, a_{ij}\right].b_j(o_t)\right\}\right\}.$$

3.3.4 Módulo do Algoritmo de Baum-Welch

Os experimentos utilizaram um algoritmo de treinamento baseado na maximização da verossimilhança (MLE) (BJM83) das sentenças de treinamento. Os parâmetros iniciais dos modelos, obtidos através de uma das estratégias apresentadas anteriormente, são submetidos aos ajustes determinados pelo algoritmo de treinamento e, para cada época, verifica-se o valor da verossimilhança média de 10% das sentenças de treinamento. O processo é repetido até que a diferença percentual entre a verossimilhança média obtida na iteração atual e a obtida na iteração anterior seja menor que 10^{-3}.

É importante observar que a estimação dos parâmetros é realizada de forma a maximizar a verossimilhança das sentenças de treinamento, sem preocupação com a capacidade de discriminação dos modelos. Assim, em primeira análise, um peso baixo de uma determinada Gaussiana indica, por exemplo, que tal componente contribui pouco para a verossimilhança total do modelo, ou seja, para a cobertura dos dados de treinamento. Entretanto, tal valor não tem relação com a contribuição da Gaussiana para a capacidade de classificação do modelo.

No intuito de se definir as equações de re-estimação dos parâmetros do modelo através do algoritmo Baum-Welch, antes é necessário definir os algoritmos *Forward* e *Backward*.

Algoritmo Forward

Inicialmente, define-se a variável *forward* $\alpha_t(i)$ como

$$\alpha_t(i) = P(\mathbf{o}_1,\ \mathbf{o}_2, \ldots,\ \mathbf{o}_t,\ q_t = i|M), \tag{3.21}$$

que corresponde à probabilidade da seqüência de observações parciais $\mathbf{o}_1,\ \mathbf{o}_2, \ldots,\ \mathbf{o}_t$, passando pelo estado "i" no instante "t", para um dado conjunto de modelos definidos por $M = (\boldsymbol{\mu},\ \mathbf{U},\ C,\ \mathbf{A})$, sendo $\boldsymbol{\mu}$ as médias, $\mathbf{U}$ as matrizes de covariância, C os coeficientes de ponderação ou pesos das Gaussianas e $\mathbf{A}$ a matriz de transição de estados.

Na sequência calcula-se as variáveis *forward* para todos os estados, e para todo o conjunto de observações de tamanho T. Assim, o algoritmo pode ser resumido por:

- 1 - Inicialização

$$\alpha_1(i) = \pi_i b_i(\mathbf{o}_1), \quad 1 \leq i \leq N_s \tag{3.22}$$

- 2 - Parte Recursiva

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N_s} \alpha_t(i).a_{ij}\right].b_j(\mathbf{o}_{t+1}), \quad \begin{cases} 1 \leq t \leq T-1 \\ 1 \leq j \leq N_s \end{cases} \tag{3.23}$$

Por fim, a verossimilhança $P(\mathbf{O}|M)$ pode ser calculada pela Equação (3.24).

$$P(\mathbf{O}|M) = \sum_{i=1}^{N_s} \alpha_T(i) \tag{3.24}$$

Neste ponto, existem duas observações importantes para o problema de reconhecimento de fala: $\pi_i = 1$ para o primeiro estado do silêncio e $\pi_i = 0$ para os demais estados do modelo, e $\alpha_T(i) \neq 0$ para o terceiro estado do silêncio enquanto $\alpha_T(i) \longrightarrow 0$ para os demais estados. Portanto, o valor $P(\mathbf{O}|M)$ é uma medida da probabilidade da sentença formada pelas observações $\mathbf{O}$ ter sido produzida pela sequência de estados $\mathbf{Q} = [q_1,\ q_2,\ \ldots,\ q_t, \ldots,\ q_T]$.

Algoritmo Backward

De forma semelhante, define-se a variável *backward* $\beta_t(i)$ como

$$\beta_t(i) = P(\mathbf{o}_{t+1},\ \mathbf{o}_{t+2}, \ldots,\ \mathbf{o}_T|q_t = i,\ M), \tag{3.25}$$

que corresponde à probabilidade da sequência de observações parciais do instante t+1 até a última observação no instante T, dado que o caminho passa pelo estado "i" no instante t, e para um dado conjunto de modelos definidos por $M = (\boldsymbol{\mu},\ \mathbf{U},\ C,\ \mathbf{A})$.

Assim, o algoritmo pode ser resumido por:

- 1 - Inicialização

$$\beta_T(i) = 1, \quad 1 \leq i \leq N_s \tag{3.26}$$

- Parte Recursiva

$$\beta_t(i) = \sum_{j=1}^{N_s} a_{ij}.b_j(\mathbf{o}_{t+1}).\beta_{t+1}(j), \quad \begin{cases} t = T-1,\ T-2, \ldots, 1 \\ 1 \leq i \leq N_s \end{cases} \tag{3.27}$$

Escalonamento das Variáveis Forward e Backward

Em geral, os valores fornecidos pelas variáveis *forward* e *backward*, que são medidas de probabilidades calculadas de forma recursiva a partir de valores de verossimilhança e de probabilidades de transição de estados, tendem a se tornar significativamente menores que 1 à medida que a sequência de observações é processada, excedendo dessa forma a precisão *double* das variáveis, e resultando freqüentemente em *underflow*. Portanto, é importante realizar um escalonamento de tais variáveis, para cada instante de tempo, a fim de se evitar o problema numérico de *underflow*. As alterações nos algoritmos associadas ao escalonamento das variáveis *forward* e *backward* encontram-se resumidas a seguir:

- 1 - Definir os valores iniciais da nova variável $\tilde{\alpha}$.

$$\tilde{\alpha}_1(i) = \alpha_1(i) \tag{3.28}$$

- 2 - Definição da constante de escalonamento $\hat{c}_1$.

$$\hat{c}_1 = \frac{1}{\sum_{i=1}^{N_s} \tilde{\alpha}_1(i)} \tag{3.29}$$

- 3 - Definição da variável $\hat{\alpha}$.

$$\hat{\alpha}_1(i) = \tilde{\alpha}_1(i).\hat{c}_1 \tag{3.30}$$

- 4 - Parte Recursiva.

$$\tilde{\alpha}_{t+1}(j) = \left[\sum_{i=1}^{N_S} \hat{\alpha}_t(i).a_{ij}\right].b_j(\mathbf{o}_{t+1}), \quad \begin{cases} 1 \leq t \leq T-1 \\ 1 \leq j \leq N_s \end{cases} \tag{3.31}$$

$$\hat{c}_t = \frac{1}{\sum_{i=1}^{N_S} \tilde{\alpha}_t(i)} \tag{3.32}$$

$$\hat{\alpha}_t(i) = \tilde{\alpha}_t(i) .\hat{c}_t \tag{3.33}$$

$$\hat{\beta}_t(i) = \tilde{\beta}_t(i) .\hat{c}_t \tag{3.34}$$

Novamente, deve-se notar que no problema específico de reconhecimento de fala, $\tilde{\alpha}_1(i) \neq 0$ para o primeiro estado do silêncio e $\tilde{\alpha}_1(i) = 0$ para os demais estados.

As variáveis normalizadas $\hat{\alpha}$ e $\hat{\beta}$ podem então ser utilizadas nas equações de re-estimação dos parâmetros dos HMMs.

Estimação dos Parâmetros dos Modelos (Algoritmo Baum-Welch)

Os parâmetros dos HMMs contínuos (Mar97), para uma seqüência de observações múltiplas **O**, são calculados a partir das variáveis *forward* e *backward* escalonadas, das constantes de escalonamento $\hat{c}$, dos vetores de parâmetros acústicos $\mathbf{o}_t$, dos próprios parâmetros atuais $(\boldsymbol{\mu},\ \mathbf{U},\ C,\ \mathbf{A})$ do modelo, e dos valores de verossimilhança normalizada definida pela Equação (3.35)

$$N_t(j,m) = \frac{c_{jm}.G\left(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm}\right)}{\sum_{\kappa=1}^{N_g} c_{j\kappa}.G\left(\mathbf{o}_t, \boldsymbol{\mu}_{j\kappa}, \mathbf{U}_{j\kappa}\right)}, \tag{3.35}$$

em que N_g é o número de Gaussianas presentes no estado "j" do modelo. Assim, a média, matriz de covariância, peso e matriz de transição de estados são estimados pelas Equações (3.36)-(3.39).

- Média

$$\boldsymbol{\mu}_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j) .\hat{\beta}_t^d(j) .N_t^d(j,m) .\mathbf{o}_t^d \Big/ \hat{c}_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j) .\hat{\beta}_t^d(j) .N_t^d(j,m) \Big/ \hat{c}_t^d} \tag{3.36}$$

- Matriz de Covariância

$$\mathbf{U}_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j).\hat{\beta}_t^d(j).N_t^d(j,m).\left(\mathbf{o}_t^d - \boldsymbol{\mu}_{jm}\right)\cdot\left(\mathbf{o}_t^d - \boldsymbol{\mu}_{jm}\right)' \Big/ \hat{c}_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j).\hat{\beta}_t^d(j).N_t^d(j,m) \Big/ \hat{c}_t^d} \tag{3.37}$$

- Peso ou Coeficiente de Ponderação

$$c_{jm} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j).\hat{\beta}_t^d(j).N_t^d(j,m) \Big/ \hat{c}_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T_d} \hat{\alpha}_t^d(j).\hat{\beta}_t^d(j) \Big/ \hat{c}_t^d} \tag{3.38}$$

- Probabilidade de Transição de Estados

$$a_{ij} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T_d-1} \hat{\alpha}_t^d(i).a_{ij}.b_j\left(\mathbf{o}_{t+1}^d\right).\hat{\beta}_{t+1}^d(j)}{\sum_{d=1}^{D}\sum_{t=1}^{T_d-1} \hat{\alpha}_t^d(i).\hat{\beta}_t^d(i) \Big/ \hat{c}_t^d}, \tag{3.39}$$

em que D é o número de sentenças de treinamento e T_d é o número de quadros extraídos da sentença "d". Deve-se notar que no problema de reconhecimento de fala, para um determinada sentença de treinamento, as variáveis $\hat{\alpha}$ e $\hat{\beta}$ são calculadas apenas para os estados dos modelos acústicos correspondentes aos fones presentes na sentença. Além disso, duas restrições devem ser seguidas durante o treinamento, sendo que a primeira,

$$\sum_{m=1}^{N_g} c_{jm} = 1, \tag{3.40}$$

deve ser satisfeita para que a mistura de Gaussianas seja uma PDF, e a segunda,

$$\sum_{j=1}^{N_s} a_{ij} = 1, \tag{3.41}$$

estabelece que a probabilidade de transição de um estado para os demais ou para a permanência no próprio estado é de 100%.

3.3.5 Módulo de Seleção de Topologia

O algoritmo de treinamento baseado em MLE, apresentado na seção anterior, assume inicialmente uma topologia que pode ter sido definida de forma arbitrária, como por exemplo uma topologia com um número fixo de Gaussianas por estado, ou uma topologia com um número variado de Gaussianas

por estado, que pode ter sido determinada através de uma dentre as diversas técnicas presentes na literatura (BIC, AIC, etc.), e também assume um conjunto de parâmetros obtidos através de um processo de inicialização. Dessa forma, tal algoritmo atua apenas na re-estimação dos parâmetros do sistema, desconsiderando se a quantidade de parâmetros disponíveis para os ajustes é suficiente, insuficiente, ou excessiva. No caso de se ter um número insuficiente de parâmetros nos modelos, o processo de ajuste será impossibilitado de determinar um sistema capaz de modelar até mesmo os dados da base de treinamento, o que, do ponto de vista de reconhecimento de fala, corresponde a modelos com baixa resolução acústica ou com poucas Gaussianas por estado. No caso de se ter um sistema com um número excessivo de parâmetros, o processo de ajuste pode levar à ocorrência de sobre-parametrização, ou seja, o sistema se torna super ajustado aos dados de treinamento, além de propiciar a interferência entre modelos concorrentes, conduzindo a erros de classificação. A inicialização dos parâmetros do modelo através de uma segmentação uniforme, por exemplo, pode contribuir para que algumas Gaussianas, de modelos contendo um número excessivo de componentes, convirjam para as distribuições erradas no espaço acústico durante o treinamento.

A determinação da topologia dos modelos é, portanto, parte do processo de treinamento e exerce uma função de grande importância para a obtenção de sistemas robustos. Assim, vários métodos foram propostos no sentido de possibilitar a determinação da topologia mais apropriada para um determinado modelo estatístico que, no caso do HMM, corresponde à determinação do número de Gaussianas por estado. Porém, é importante destacar que os métodos para determinação da topologia possuem formulações gerais, que permitem a aplicação dos mesmos para quaisquer modelos estatísticos, como por exemplo Redes Neurais Artificiais, modelos Polinomiais, HMMs, etc.

O treinamento discriminativo é menos sensível à existência de parâmetros em excesso no modelo, pois além de realizar a maximização da verossimilhança para os padrões corretos, também minimiza a interferência entre os modelos de diferentes unidades acústicas. A maior parte dos algoritmos de treinamento discriminativo atua na etapa de estimação de parâmetros, enquanto existem poucos trabalhos na literatura que abordam aspectos discriminativos durante a tarefa de seleção de topologia.

Na sequência, serão apresentados alguns dentre os métodos presentes na literatura para a determinação da topologia de HMMs.

Critério de Informação Bayesiano (BIC)

O critério BIC tem sido amplamente utilizado para a seleção de estruturas no processo de modelagem estatística em diversas áreas. O conceito fundamental que sustenta o critério BIC é o Princípio da Parcimônia, o qual determina que o modelo selecionado deve ser aquele que apresentar a menor complexidade e ao mesmo tempo tenha uma elevada capacidade para modelar os dados de treinamento. Tal princípio pode ser observado claramente na equação (3.42)

$$BIC\left(M_l^j\right) = \sum_{t=1}^{N_j} \log P\left(\mathbf{o}_t^j | M_l^j\right) - \lambda \frac{\nu_l^j}{2} \log N_j, \tag{3.42}$$

onde M_l^j é o modelo candidato "*l*" do estado "*j*", N_j é o número de amostras do estado "*j*", $\mathbf{o}_t^j$ é a *t-ésima* amostra do estado "*j*", ν_l^j é o número de parâmetros livres presentes em M_l^j e o parâmetro λ controla o termo de penalização.

De acordo com tal critério, o modelo selecionado deve ser aquele que apresentar o maior valor de BIC dentre todos os modelos candidatos. Pode-se notar então que a topologia do modelo de cada estado é obtida sem levar em consideração os modelos dos demais estados existentes. Entretanto, algumas modificações no critério BIC (Bie03) já foram propostas no intuito de levar em consideração todos os estados existentes durante a seleção de topologia.

Algoritmo Discriminativo para o Aumento da Resolução Acústica

O critério discriminativo proposto em (PB00) tem como princípio a determinação de quais estados dos modelos possuem baixa resolução acústica (número insuficiente de Gaussianas no modelo), de acordo com um limiar previamente estabelecido. A idéia consiste basicamente em decodificar os dados de treinamento, fazer um alinhamento de Viterbi usando as transcrições das sentenças decodificadas e comparar tal alinhamento com o alinhamento correto (BP98; GJPP99). Assim, é possível verificar quais estados estão sendo confundidos com outros e, desta forma, elaborar uma lista de confusão para cada estado, a qual é definida por $F(\mathbf{x}_t)$.

Neste sentido, a Equação (3.43) mede o comportamento do modelo de um determinado estado em relação às amostras associadas ao próprio estado em questão

$$P_c^l = \frac{1}{N_{fr}^l} \sum_{t \ni C(\mathbf{x}_t)=l} \frac{P(\mathbf{x}_t|M_l)}{P(\mathbf{x}_t|M_l) + \sum\limits_{j \in F(\mathbf{x}_t)} P(\mathbf{x}_t|M_j)}, \tag{3.43}$$

onde $\mathbf{x}_t$ são os quadros associados à classe $C(\mathbf{x}_t)$ que corresponde ao estado "l", N_{fr}^l é o número de quadros associados ao estado "l", $P(\mathbf{x}_t|M_l)$ é o logaritmo da verossimilhança dada pelo modelo "M_l" do estado "l" e "j" são os estados da lista de confusão $F(\mathbf{x}_t)$.

Uma vez calculado o P_c^l para cada estado, deve-se encontrar todos os estados que apresentem P_c^l inferior a um limiar pré-definido e substituir tais modelos por correspondentes que foram treinados em um sistema que utiliza uma maior resolução acústica (maior número de Gaussianas por estado).

É importante notar que, neste método, parte-se de um sistema menor que possui modelos inicialmente com "X" Gaussianas por estado e que, após a análise, terá modelos com "X" ou "Y" Gaussianas por estado, onde "Y" é o número de componentes presentes em cada mistura do sistema de maior complexidade. O aumento do número de Gaussianas tem como objetivo aumentar a resolução acús-

tica dos HMMs onde for necessário, de acordo com o critério discriminativo, e conseqüentemente a discriminabilidade do mesmos.

Trabalhos anteriores mostram que este critério pode fornecer resultados que superam os obtidos pelo clássico BIC, dependendo dos limiares escolhidos (PB00).

Método Baseado na Medida de Entropia dos Estados

O método baseado em medidas de entropia tem como estratégia o aumento gradativo do tamanho do sistema (CU93), através de divisões (*splitting*) sucessivas das Gaussianas dos modelos. Basicamente, todos os modelos acústicos são inicializados com uma Gaussiana por estado, e em seguida a entropia de cada Gaussiana é calculada pela Equação (3.44)

$$H_{si} = \sum_{d=1}^{dim} \frac{1}{2} \log_{10} \left(2\pi e \sigma_{sd}^2\right), \tag{3.44}$$

em que "dim" é a dimensão do espaço de características acústicas (neste trabalho é 39) e "σ_{sd}^2" é a variância da *i-ésima* componente da mistura do estado "*s*", ao longo da dimensão "*d*".

A entropia de cada estado (H_s) pode então ser calculada de acordo com a Equação (3.45)

$$H_s = \sum_{i=1}^{N_g} N_{si} H_{si}, \tag{3.45}$$

em que N_g é o número de componentes Gaussianas no estado "*s*" e N_{si} é o número de amostras associadas à *i-ésima* componente da mistura do estado "*s*".

Na seqüência o algoritmo *Segmental k-means* é utilizado no intuito de realizar a divisão de todas as Gaussianas dos modelos de forma que, após as divisões, cada estado possuirá uma Gaussiana a mais na mistura. Calcula-se então a entropia de cada estado para esta nova configuração $\hat{H}_s$, e também a variação de entropia de acordo com a Equação (3.46)

$$d_s = H_s - \hat{H}_s. \tag{3.46}$$

A idéia consiste em arranjar os estados em ordem decrescente de variação de entropia, e manter os N primeiros estados com a nova configuração (com a Gaussiana acrescentada após as divisões), e retornar os demais estados para a configuração anterior (retirando a Gaussiana acrescentada após as divisões). O processo é repetido até que o tamanho desejado para o sistema seja atingido.

Critério de Informação de Akaike (AIC)

O Critério de Informação de Akaike (Aka74; SIK86) assume que os dados foram gerados como realizações de uma variável aleatória, cuja função densidade de probabilidade real $g\left(x|\theta^*\right)$ é desconhecida, e seleciona a mistura de Gaussianas $g\left(x|\hat{\theta}\right)$ que mais se aproxima da função real para descrever os dados, em que θ^* e $\hat{\theta}$ são o conjunto de parâmetros que definem a PDF real e a estimada, respectivamente. Além disso, sabe-se que $g\left(x|\hat{\theta}\right)$ aumenta juntamente com o tamanho do sistema, e portanto o AIC deve avaliar se o aumento do número de parâmetros livres do sistema e da verossimilhança para os dados de treinamento representam um ganho considerável no ajuste dos modelos aos dados.

A topologia selecionada deve ser aquela que minimizar a Equação (3.47), que corresponde ao AIC,

$$AIC\left(M_l^j\right) = -\sum_{t=1}^{N_j} \log P\left(\mathbf{o}_t^j | M_l^j\right) + \nu_l^j. \tag{3.47}$$

onde M_l^j é o modelo candidato "*l*" do estado "*j*", N_j é o número de amostras do estado "*j*", $\mathbf{o}_t^j$ é a *t-ésima* amostra do estado "*j*", ν_l^j é o número de parâmetros livres presentes em M_l^j.

Critério Minimum Description Length (MDL)

O *Minimum Description Length* (SiI02) pertence a um grupo de técnicas para a determinação da topologia de modelos baseada na teoria da informação e de codificação, enquanto o BIC pertence a um grupo de técnicas baseadas na teoria bayesiana (LGW03). Assim, o MDL considera que a melhor topologia é aquela que permitir a extração da maior quantidade de informação dos dados, o que pode ser medido pela capacidade de compressão dos dados, que é determinada pela topologia do modelo. Considerando-se a teoria de codificação e os códigos de Huffman, é possível demonstrar que, utilizando misturas Gaussianas, os dados podem ser comprimidos para um comprimento de código de $\sum_{j=1}^{N_j} \log \frac{1}{g\left(x_j|\hat{\theta}\right)}$. Além disso, é importante notar que o método de compressão também precisa ser transmitido juntamente com os dados comprimidos. Dessa forma, os parâmetros $\hat{\theta}$ da mistura de Gaussianas $g\left(x|\hat{\theta}\right)$ precisam ser codificados e, utilizando uma precisão p, o comprimento obtido é de $p \times \nu$ (MA94), em que ν é o número de parâmetros livres do modelo. (Ris89) demonstrou que para minimizar o comprimento total do código, é necessário que $p = \frac{1}{2}\log N_j$. Portanto, o critério MDL que deve ser minimizado é dado por

$$MDL\left(M_l^j\right) = -\sum_{t=1}^{N_j} \log P\left(\mathbf{o}_t^j | M_l^j\right) + \frac{\nu_l^j}{2} \log N_j, \tag{3.48}$$

sendo ν o número de parâmetros livres das N_g Gaussianas do modelo, representadas em um espaço acústico de dimensão "dim", que é calculado por

$$\nu = \left\{ (N_g - 1) + N_g . \left[dim + \frac{1}{2} . dim . (dim + 1) \right] \right\}. \tag{3.49}$$

3.4 O Decodificador

Os algoritmos de decodificação são baseados no algoritmo de Viterbi, podendo utilizar também modelos de linguagem ou gramáticas, que consistem basicamente de probabilidades de transições entre *labels* (fones ou palavras) para uma dada língua. Para a geração de um modelo de linguagem, é necessário saber quantos *labels* distintos existem nas sentenças de treinamento e, no caso dos modelos do tipo *bigram*, deve-se contar o número de ocorrências de cada par adjacente de *labels* "i" e "j", definido por $N(i,j)$. Assim, o número total de ocorrências de um *label* "i" é dado por $N(i) = \sum_{j=1}^{L} N(i,j)$.

As probabilidades $p(i,j)$ do modelo de linguagem *bigram* são definidas por

$$p(i,j) = \begin{cases} \alpha . N(i,j) / N(i), & \text{se } N(i) > 0 \\ 1/L, & \text{se } N(i) = 0 \end{cases} \tag{3.50}$$

em que α é calculado de tal forma que $\sum_{j=1}^{L} p(i,j) = 1$.

As probabilidades *unigram* $p(i)$, calculadas para o modelo de linguagem *Back-off bigram*, são dadas por

$$p(i) = \begin{cases} N(i)/N, & \text{se } N(i) > u \\ u/N, & \text{se } N(i) \leq u \end{cases} \tag{3.51}$$

em que u é um contador de truncamento da *unigram* e $N = \sum_{i=1}^{L} \max[N(i), u]$.

As probabilidades do modelo de linguagem *Back-off bigram* são calculadas por

$$p(i,j) = \begin{cases} [N(i,j) - D]/N(i), & \text{se } N(i,j) > t \\ b(i) . p(j), & \text{se } N(i,j) \leq t \end{cases} \tag{3.52}$$

sendo D um desconto (em geral 0,5) e t é um limiar de contagem do modelo de linguagem *bigram*. O peso $b(i)$ calculado de tal forma a garantir que $\sum_{j=1}^{L} p(i,j) = 1$, é dado pela Equação (3.53)

$$b(i) = \frac{1 - \sum_{j \in B} p(i,j)}{1 - \sum_{j \in B} p(j)}, \tag{3.53}$$

em que B é o conjunto de todas as palavras para as quais $p(i,j)$ possui um *bigram*.

Os experimentos realizados com a base de dados pequena utilizaram modelos de linguagem do

tipo *Back-off bigram* e *Word-pairs*, sendo que o último é um caso determinístico do modelo *bigram* apresentado na Equação (3.50), e os *labels* adotados foram palavras. O modelo de linguagem *Word-pairs* é descrito por

$$p(i,j) = \begin{cases} 1, & se\ N(i,j) \neq 0 \\ 0, & se\ N(i,j) = 0 \end{cases} \tag{3.54}$$

O decodificador utilizado nos experimentos com a base de dados pequena foi adaptado do sistema desenvolvido por (Yno99).

Por outro lado, os experimentos realizados com a TIMIT utilizaram o modelo de linguagem *bigram*, onde os *labels* adotados foram fones independentes de contexto, e o decodificador foi implementado através das ferramentas do HTK (Cam02). É importante destacar que, de acordo com trabalhos anteriores, a obtenção de ganhos em termos de *accuracy* de fones implica diretamente em ganhos em termos de *accuracy* de palavras, além de permitir uma melhor identificação das fontes de erros (LG93). Dessa forma, o reconhecimento de fones contínuos pode ser empregado para uma melhor avaliação do desempenho de sistemas de reconhecimento de fala.

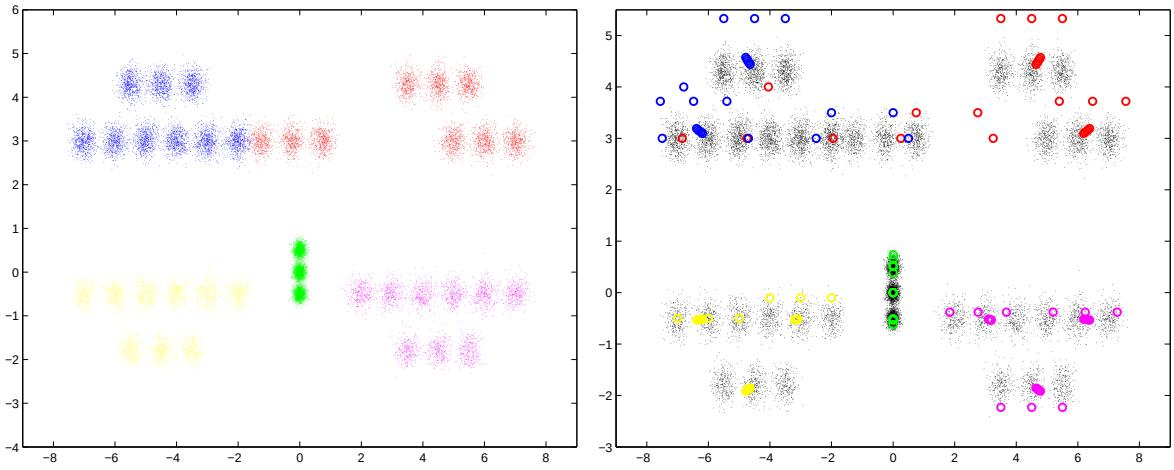
Em última análise, utilizou-se um fator multiplicativo que pondera a importância do modelo de linguagem durante a decodificação, o qual assumiu os valores 1 e 2 (Val95), nos experimentos realizados com a base de dados em Português e a base em Inglês, respectivamente.

3.5 Dados Artificiais

Os parâmetros extraídos dos sinais acústicos se encontram representados em um espaço de características de dimensão 39. Assim, não é possível visualizar alguns pontos relacionados ao processo de treinamento, tais como localização inicial das Gaussianas em relação a distribuições dos diferentes padrões, movimentação das Gaussianas no espaço com o decorrer do treinamento, atuação dos algoritmos para determinação do número de componentes por estado, e etc. Uma estratégia que permite a visualização de tais aspectos, consiste na geração de um conjunto de dados artificiais, onde as distribuições dos padrões são todas conhecidas. A utilização de dados artificiais para a verificação de aspectos relacionados ao comportamento e desempenho de novos métodos aplicados à modelagem de sistemas físicos é bastante comum, desde problemas relacionados com predições de séries temporais até problemas de reconhecimento de padrões. Neste sentido, gerou-se artificialmente 5 distribuições de dados bidimensionais, para simular 5 unidades acústicas, as quais se encontram indicadas na Figura 3.4(a) pelas cores azul, vermelho, amarelo, roxo e verde.

Deve-se notar que existe sobreposição das distribuições azul e vermelha, a qual foi estabelecida propositalmente, pois, no caso real dos parâmetros acústicos, freqüentemente ocorre sobreposição

entre as distribuições obtidas para diferentes unidades acústicas.



(a) Distribuições artificiais. (b) Inicialização das médias das Gaussianas

Fig. 3.4: Treinamento via MLE realizado com dados artificiais.

Os valores iniciais das média das Gaussianas utilizadas na modelagem de cada distribuição foram determinados arbitrariamente, e se encontram representadas pelos símbolos “o” da Figura 3.4(b), cada qual com cor associada à distribuição que deve ser modelada pela Gaussiana. Deve-se observar que algumas Gaussianas vermelhas foram inicializadas sobre o espaço da distribuição azul e, da mesma forma, algumas Gaussianas azuis foram inicializadas sobre o espaço da distribuição vermelha. Tal inicialização foi realizada no intuito de simular o efeito de uma inicialização uniforme. Na modelagem foram empregadas 8 Gaussianas por estado, o que resultou em um sistema contendo 120 Gaussianas (5 distribuições, cada qual modelada por um HMM com 3 estados do tipo left-to-right).

Uma vez geradas as 5 distribuições artificiais e definidos os valores iniciais dos parâmetros das Gaussianas, pode-se então iniciar o processo de treinamento pelo algoritmo Baum-Welch e observar a movimentação das Gaussianas no espaço das características artificiais. A Figura 3.5, indica as localizações das Gaussianas após 20 épocas de treinamento.

Pode-se observar que 3 Gaussianas vermelhas convergiram para o espaço da distribuição azul, encontrando-se distantes das fronteiras entre as distribuições. Além disso, uma Gaussiana azul convergiu para o espaço da distribuição vermelha, porém, neste caso, em uma região de fronteira. Por outro lado, as demais Gaussianas convergiram para as respectivas distribuições, o que era desejado. Assim, é possível verificar que a inicialização dos parâmetros dos modelos pode permitir que as Gaussianas convirjam para as distribuições indevidas durante o treinamento, resultando em uma maior interferência entre os modelos e portanto em mais erros de classificação durante a decodificação.

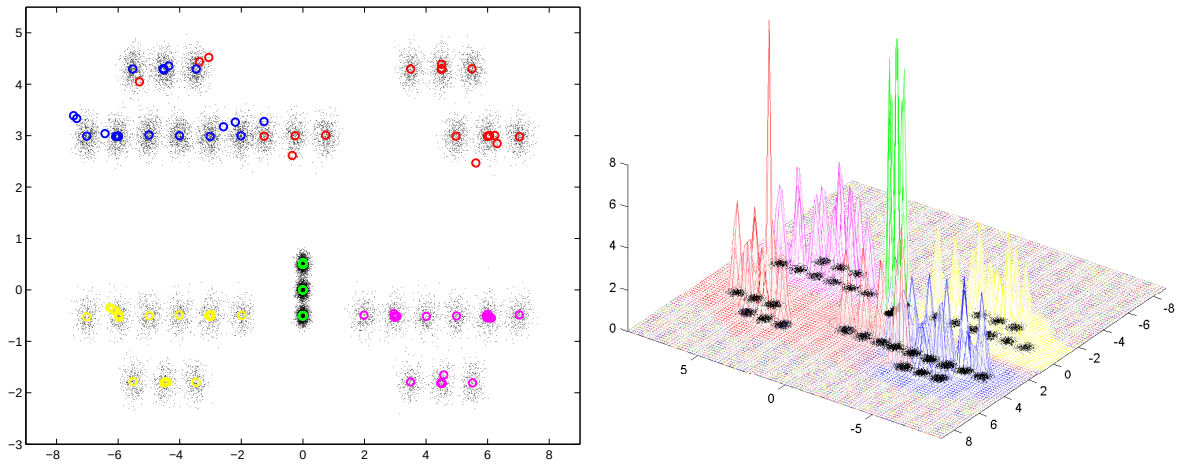


Fig. 3.5: Localização das Gaussianas após o treinamento. Visualização 3D das Gaussianas após o treinamento.

Deve-se notar também que existe um número exagerado de Gaussianas no sistema, e neste caso diversas Gaussianas se encontram localizadas praticamente sobre o mesmo ponto no espaço das características artificiais, ou seja, Gaussianas com médias muito próximas, e tal redundância terá provavelmente pouco efeito sobre o desempenho do sistema do ponto de vista do reconhecimento, mas terá um grande efeito sobre o custo computacional do sistema.

Em última análise, foi possível verificar a eficiência do algoritmo MLE para o ajuste das médias das Gaussianas, porém dependendo da inicialização dos parâmetros e da existência de sobreposição entre distribuições diferentes, algumas Gaussianas podem convergir para posições incorretas no espaço de características. O próximo passo consiste em avaliar a utilização de técnicas de determinação da topologia dos modelos, capazes de encontrar sistemas que apresentem um melhor compromisso entre complexidade e desempenho. Tal experimento será considerado novamente no Capítulo 6, com a aplicação do novo método proposto neste trabalho.

3.6 Discussão

A utilização de duas bases de dados permitirá a avaliação do novo método proposto neste trabalho em diferentes aspectos, tais como a quantidade de dados disponíveis para o treinamento e a inicialização dos parâmetros dos modelos a partir de uma segmentação uniforme e a partir de uma segmentação manual, dentre outros.

O sistema de treinamento de HMMs contínuos foi implementado em parte através de *scripts* do Matlab, e em parte através de códigos compilados. A estratégia de se utilizar o Matlab como plataforma para a implementação do sistema se deve ao fato da existência de diversas ferramentas

relacionadas com processamento de sinais, estatística e modelagem de sistemas, os quais podem facilitar e agilizar os testes de novas idéias. Entretanto, a linguagem de *scripts* do Matlab, em geral, resulta em um aumento considerável no tempo de processamento de trechos de programa com custo computacional elevado, como por exemplo os *loops*. A solução adotada para contornar tal problema foi a utilização dos *mex files*, que são arquivos escritos em uma linguagem de programação, e posteriormente compilados, de tal forma que o tempo de execução dos trechos de programa com elevado custo computacional é dado basicamente pelo tempo de execução da função compilada.

As funções compiladas foram implementadas na linguagem C e correspondem aos trechos do programa com alto custo computacional. Tais funções podem ser acessadas pelo Matlab através das APIs, que são um conjunto de bibliotecas em C que permitem a interface entre o código compilado e as linhas de comando dos *scripts* do Matlab. O tempo de processamento de um *script* do Matlab que utiliza *mex files* para a implementação das funções com elevado custo computacional é comparável ao tempo de processamento do mesmo código implementado em C e compilado.

O decodificador baseado nas ferramentas do HTK também utilizou o Matlab como linguagem de *scripts* para a execução das linhas de comando referentes às funções compiladas fornecidas pelo HTK. Assim, a elaboração dos arquivos de configuração, o processamento de texto sobre as sentenças de treinamento e reconhecimento, o processamento dos arquivos contendo as sentenças parametrizadas e os arquivos contendo os parâmetros dos modelos, podem ser facilmente manipulados através de ferramentas existentes no Matlab para o processamento de texto.

Em última análise, foi possível conciliar as facilidades proporcionadas pelas ferramentas disponíveis no Matlab para o teste de novas idéias com a velocidade de processamento das funções compiladas. Além disso, obteve-se um sistema de treinamento e reconhecimento integrado sobre a mesma plataforma, parte desenvolvida durante este trabalho (treinamento) e parte adaptada de um sistema de reconhecimento utilizado como referência na literatura (HTK).

3.7 Conclusões

Os testes iniciais com os dados gerados artificialmente mostraram a eficiência do treinamento via MLE e ao mesmo tempo indicaram que alguns problemas podem ocorrer, no que diz respeito à convergência de Gaussianas para as distribuições incorretas no espaço de características, dependendo da inicialização dos parâmetros dos modelos. Além disso, o experimento com dados artificiais será abordado novamente no Capítulo 6, porém com técnicas para a determinação da topologia dos modelos, visto que diversas Gaussianas convergiram para praticamente o mesmo ponto no espaço de características, além de algumas componentes terem convergido para as distribuições incorretas.

A utilização dos *mex files* viabilizou a implementação do sistema de treinamento de HMMs con-

tínuos para o reconhecimento de fala através do Matlab, o que facilitará o teste de novas idéias no decorrer do trabalho.

Capítulo 4

Determinação do Número de Componentes em Modelos com Misturas de Gaussianas

4.1 Introdução

O processo para a obtenção de sistemas de reconhecimento de fala contínua, a partir de bases de dados pequenas, possui alguns problemas inerentes à quantidade de dados disponíveis. A insuficiência de informação devido ao tamanho da base de treinamento limita a precisão da modelagem e pode diminuir a confiabilidade dos modelos resultantes. Tal fato pode ser observado em qualquer problema de modelagem estatística através da qual se pretende extrair informações relevantes para a geração de modelos representativos do sistema físico em questão. Dessa forma, várias técnicas têm sido propostas no intuito de ajustar o processo de modelagem à quantidade de dados disponíveis, e ao mesmo tempo diminuir a influência de termos espúrios contidos nos dados, na obtenção dos parâmetros dos modelos. Tal ajuste é realizado através da escolha apropriada da complexidade dos modelos, de acordo com um critério específico. Têm-se como exemplos de tais técnicas o critério BIC (*Bayesian Information Criterion*), o critério AIC (*Akaike Information Criterion*), métodos baseados em medidas de entropia, além de outros. Recentemente, dentro do contexto de reconhecimento de padrões e de reconhecimento de fala, alguns métodos discriminativos têm sido propostos neste sentido, utilizando informações relacionadas à capacidade de classificação dos modelos (PB00; DKW00; Bie03), dentre os quais se pode citar o novo GEA (*Gaussian Elimination Algorithm*), contribuição do presente trabalho.

Não obstante, os sistemas obtidos a partir de bases pequenas podem ser gerados com menor custo e de forma mais rápida, devido ao tamanho reduzido da base de dados e ao esforço computacional necessário para processá-la. O vocabulário reduzido, juntamente com uma gramática restritiva, simplificam a busca durante a decodificação e ao mesmo tempo permitem que um sistema de reconhe-

cimento gerado a partir de uma base de dados pequena possa atender a algumas aplicações práticas. Neste caso, o ajuste de complexidade do sistema também se torna interessante no sentido de diminuir o custo computacional global do sistema e viabilizar sua implementação em dispositivos com memória e capacidade de processamento limitados.

Neste Capítulo serão apresentados os resultados obtidos através de três métodos utilizados para a determinação da complexidade dos HMMs que correspondem aos modelos acústicos das unidades fonéticas da língua portuguesa adotadas neste trabalho, empregando apenas a base em Português.

4.2 Determinação da Complexidade de HMMs

No processo de modelagem estatística, algumas etapas se sucedem até a obtenção dos modelos finais. De forma resumida, o primeiro passo consiste em extrair as informações relevantes sobre o sistema físico em questão, o que no contexto de reconhecimento de fala é realizado através da parametrização do sinal acústico. Na seqüência, deve-se escolher a ferramenta matemática através da qual os modelos serão obtidos, que neste caso corresponde aos Modelos Ocultos de Markov (HMMs). No passo seguinte, realiza-se a seleção da topologia mais apropriada para os modelos, seguido da estimação dos parâmetros do modelo. Por fim, realizam-se testes visando avaliar a robustez e precisão dos modelos encontrados.

Durante a fase de seleção da topologia, a abordagem mais simples, do ponto de vista da modelagem estatística, consiste em utilizar um número fixo de Gaussianas por estado para os HMMs. Entretanto, alguns problemas de ordem prática e de estimação de parâmetros podem surgir em decorrência da escolha de tal abordagem, como por exemplo o aumento demasiado do tamanho do sistema e a sobre-parametrização respectivamente.

A determinação da complexidade mais apropriada de cada HMM é realizada independentemente da gramática ou algoritmo utilizado na decodificação, pois tal processo ainda pertence à fase de treinamento dos modelos. Por este motivo, não se deve utilizar informações extraídas da base de testes nesta etapa da modelagem. Uma vez que a base de dados utilizada nos experimentos se encontra dividida em base de treinamento e base de teste, todos os métodos discutidos utilizam apenas informações da base de treinamento no processo de determinação da complexidade dos modelos.

O desempenho dos sistemas de reconhecimento é medido em termos de taxa de reconhecimento e *accuracy*, de acordo com as Equações 4.1 e 4.2 respectivamente,

$$T_{Reconhecimento} = \frac{N_{corretos}}{N_{labels}} \times 100, \tag{4.1}$$

$$Accuracy = \frac{N_{corretos} - N_{ins}}{N_{labels}} \times 100, \tag{4.2}$$

em que $N_{corretos}$ é o número de *labels* corretos, N_{labels} é o número total de *labels* da transcrição correta, N_{ins} é o número de inserções e os *labels* podem ser fones ou palavras.

Os resultados foram obtidos utilizando-se duas gramáticas distintas, a *Word-pairs* e a *Back-off bigram*, a fim de se avaliar os modelos em uma condição bastante restritiva e em outra bastante flexível na decodificação, respectivamente. Os sistemas que utilizam um número fixo de Gaussianas por estado são utilizados como referência para comparação de desempenho entre os métodos. Assim, obtiveram-se sistemas de referência contendo de 5 até 15 Gaussianas por estado.

Tab. 4.1: Sistemas de referência com um número fixo de Gaussianas por estado (de 5 até 15 Gaussianas por estado).

#	#	*Word-pairs*		*Back-off bigram*	
Número de Gaussianas por estado	Número total de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	Percentagem Correta (%)	Reco. Accur. (%)
5	540	91,75	88,63	77,34	52,76
6	648	92,66	89,88	78,37	54,16
7	756	92,47	89,39	78,74	56,01
8	864	91,75	88,28	78,52	54,95
9	972	93,08	90,41	80,33	57,98
10	1080	93,69	90,64	79,65	57,45
11	1188	**94,07**	**91,25**	80,14	56,73
12	1296	94,03	91,1	**81,01**	**58,25**
13	1404	93,53	90,34	79,58	57,72
14	1512	93,5	90,34	79,01	56,28
15	1620	92,93	89,96	80,18	57,90

Conforme pode ser observado na Tabela 4.1, o desempenho dos sistemas de referência tende a cair quando se utiliza mais de 12 Gaussianas por estado, o que pode ser um indicativo da ocorrência de sobre-parametrização. Por esse motivo, a análise foi realizada até o sistema contendo 15 Gaussianas por estado. Os sistemas de referência que apresentaram os melhores desempenhos, para as gramáticas *Word-pairs* e *Back-off bigram*, possuem 11 e 12 Gaussianas por estado respectivamente.

Deve-se notar que o comportamento dos sistemas de referência é aproximadamente o mesmo, independentemente da gramática utilizada, a menos de um fator de escala, como pode ser observado na Figura 4.1. Tal fato sugere, em uma análise inicial, que os resultados das alterações nos modelos acústicos podem ser observados de forma independente da gramática utilizada na decodificação.

Na seqüência, serão apresentados os resultados obtidos através de três métodos para seleção da

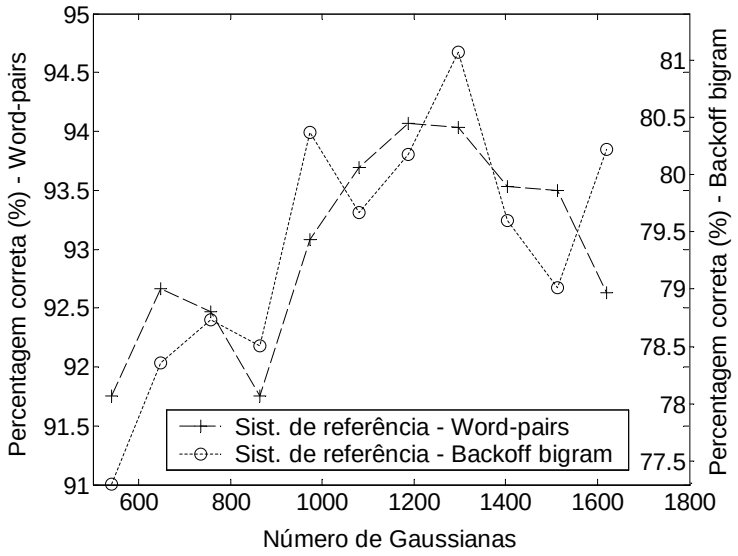


Fig. 4.1: Taxa de reconhecimento de palavras dos sistemas de referência (de 5 até 15 Gaussianas por estado).

complexidade dos modelos.

4.2.1 Critério de Informação Bayesiano (BIC)

O Critério de Informação Bayesiano (BIC) é utilizado freqüentemente na seleção de topologia de modelos estatísticos. Tal método consiste basicamente na escolha do modelo que melhor se ajuste aos dados de treinamento e que ao mesmo tempo possua o menor número de parâmetros possível, seguindo dessa forma o princípio da parcimônia. Neste sentido, varia-se o parâmetro λ, que pondera o termo de penalização do tamanho do modelo, até se determinar o modelo cuja topologia maximize o BIC.

O BIC pode ser considerado como uma aproximação do critério MDL (FJ02; CS00), e trabalhos anteriores mostram que o MDL é mais eficiente na determinação do número de componentes em misturas Gaussianas do que o critério AIC (MA94). Por este motivo, os experimentos foram realizados apenas com o BIC.

É importante notar que a quantidade de topologias candidatas aumenta consideravelmente à medida que se incrementa o tamanho do modelo. No problema em questão, por exemplo, tem-se 108 estados e permite-se que cada estado tenha até 11 Gaussianas. Assim, o número de total de possíveis topologias candidatas é de 11^{108}. Entretanto, não é necessário e nem viável testar todas as possibilidades, pois os estados são analisados individualmente. Assim, iniciando a análise pelo maior modelo, deve existir um tamanho abaixo do qual não é possível aumentar o valor do BIC e, neste ponto, a

análise deve ser interrompida.

Os tamanhos máximos de 1188 e 1296 Gaussianas foram escolhidos para as análises utilizando as gramáticas *Word-pairs* e *Back-off bigram* respectivamente, pois correspondem aos sistemas com número fixo de Gaussianas por estado com os melhores desempenhos.

Word-pairs

Os resultados obtidos através do BIC, utilizando a gramática *Word-pairs*, assim como os valores do parâmetro λ, encontram-se na Tabela 4.2.

Tab. 4.2: Desempenho dos modelos obtidos através do BIC. A comparação foi realizada com o sistema de referência contendo 1188 Gaussianas (11 por estado). Os valores entre parênteses indicam a diferença entre os sistemas obtidos pelo BIC e o de referência.

λ	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	Economia de Parâmetros (%)
0,03	1034	93,95 (-0,12)	90,91 (-0,34)	13
0,05	1016	93,99 (-0,08)	90,79 (-0,46)	14,5
0,07	1006	94,18 (+0,11)	91,25 (0)	15,3
0,1	989	93,8 (-0,27)	90,79 (-0,46)	16,8
0,2	946	93,42 (-0,65)	90,41 (-0,84)	20,4
0,3	910	93,57 (-0,5)	90,34 (-0,91)	23,4
0,4	884	93,88 (-0,19)	90,72 (-0,53)	25,6

O sistema contendo 1006 Gaussianas, obtido através do BIC para $\lambda = 0,07$, apresentou o melhor resultado do ponto de vista da taxa de reconhecimento de palavras. Tal sistema tem 15,3% menos parâmetros do que o de referência (1188 Gaussianas), e no entanto possui praticamente o mesmo desempenho. É possível obter sistemas menores, apresentado uma maior economia de parâmetros, mas neste caso com uma degradação no desempenho.

A questão inicial na busca da topologia mais apropriada é definir uma prioridade, ou seja, privilegiar o desempenho em detrimento da economia ou privilegiar a economia em detrimento do desempenho, desde que estabelecido um limiar mínimo aceitável para a taxa de reconhecimento. Deve-se ressaltar que, no primeiro caso, é possível se obter um sistema cujo desempenho pode superar consideravelmente aqueles obtidos pelos sistemas de referência com aproximadamente o mesmo tamanho. Por exemplo, o sistema obtido pelo BIC para $\lambda = 0,4$, contendo 884 Gaussianas, tem aproximadamente o mesmo tamanho do sistema de referência com 864 Gaussianas, porém apresenta um desempenho superior, em termos de taxa de reconhecimento de palavras e *accuracy*, de 2,13% e 2,44% respectivamente.

Back-off bigram

Os resultados obtidos com a gramática *Back-off bigram* mostram a mesma tendência daqueles obtidos com a *Word-pairs*, a menos de um fator de escala, como era esperado. À medida que se aumenta gradativamente o peso do termo de penalização λ, obtêm-se sistemas menores que o de referência e cujo desempenho aumenta até um determinado valor. Além deste ponto, nota-se uma diminuição no desempenho dos sistemas obtidos, e portanto a análise é interrompida.

A Tabela 4.3 apresenta os resultados obtidos e também as comparações relativas ao sistema de referência com 1296 Gaussianas (12 por estado).

Tab. 4.3: Desempenho dos modelos obtidos através do BIC. A comparação foi realizada com o sistema de referência contendo 1296 Gaussianas (12 por estado). Os valores entre parênteses indicam a diferença entre os sistemas obtidos pelo BIC e o de referência.

λ	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	Economia de Parâmetros (%)
0,03	1123	80,52 (-0,49)	56,92 (-1,33)	13,3
0,05	1110	80,75 (-0,26)	57,60 (-0,65)	14,4
0,07	1092	80,71 (-0,3)	57,30 (-0,95)	15,7
0,1	1072	80,26 (-0,75)	55,90 (-2,35)	17,3
0,15	1042	79,46 (-1,55)	54,77 (-3,48)	19,6
0,2	1010	79,61 (-1,4)	55,26 (-2,99)	22,1
0,3	957	80,03 (-0,98)	55,79 (-2,46)	26,2

É importante notar que os sistemas obtidos pelo BIC, utilizando-se a gramática *Back-off bigram* na decodificação, apresentam uma degradação em relação ao melhor sistema de referência maior do que aquela observada anteriormente empregando a gramática *Word-pairs*.

O sistema obtido pelo BIC, cujo desempenho mais se aproxima daquele de referência, contém 1110 Gaussianas, possuindo portanto 14,4% menos parâmetros. No entanto, a taxa de reconhecimento de palavras e o *accuracy* obtidos foram 80,75% e 57,60% respectivamente. Nesta condição, observou-se então uma queda no desempenho (-0,26% e -0,65%) em relação ao sistema de referência, enquanto no caso anterior, utilizando-se a gramática *Word-pairs*, foi possível se encontrar um sistema com 15,3% menos parâmetros, apresentando pelo menos o mesmo desempenho (+0,11% e 0%) do sistema de referência. Tal fato sugere, em uma primeira análise, que a obtenção de sistemas menores que os de referência e que possuam pelo menos o mesmo desempenho destes seja mais difícil quando realizada em uma condição mais flexível durante a decodificação.

Sabe-se que a utilização do BIC visa obter um sistema cuja topologia seja mais apropriada do que aquela contendo um número fixo de Gaussianas por estado. Se a comparação for realizada entre

os sistemas obtidos pelo BIC e os de referência com aproximadamente o mesmo número de Gaussianas, ao invés de se comparar com sistemas maiores, torna-se mais clara a vantagem da utilização do método de determinação da complexidade do modelo, independentemente da gramática utilizada. Por exemplo, o sistema contendo 1092 Gaussianas, indicado na Tabela 4.3, tem aproximadamente o mesmo tamanho do sistema de referência contendo 1080 Gaussianas (10 por estado), indicado na Tabela 4.1, e possui um desempenho superior em termos de taxa de reconhecimento de +1,06%. Portanto, é possível se obter um sistema contendo um número variado de Gaussianas por estado, através de um dos métodos de determinação da complexidade de modelos HMM, que supere o sistemas de referência com aproximadamente o mesmo tamanho, conforme esperado.

Uma vez realizada a observação em relação à eficácia do método para determinação da complexidade dos modelos, e a menos que seja relevante realizá-la novamente, os resultados serão apresentados incluindo apenas comparações em relação ao melhor sistema com número fixo de Gaussianas por estado, pois o principal objetivo deste trabalho é a obtenção de sistemas menores do que o melhor de referência e que apresentem pelo menos o mesmo desempenho, conforme mencionado anteriormente.

4.2.2 Medida de Entropia

O método para determinação da complexidade dos HMMs baseado na entropia, utiliza medidas sobre a quantidade de dados associados a cada *cluster*, sobre a variância de cada Gaussiana e, finalmente, avalia a variação da entropia do estado após a introdução de uma nova componente no modelo, ou seja, verifica a quantidade de informação contida no aumento do tamanho do modelo.

Resumidamente, as variações de entropia de todos os estados são ordenadas em ordem decrescente e os N primeiros estados, de acordo com tal ordenação, são incrementados, enquanto os demais permanecem com o mesmo número de parâmetros. O processo se repete até que o tamanho desejado para o sistema seja atingido.

Tal método necessita de mais condições iniciais do que aquele apresentado anteriormente pelo BIC, o que confere uma maior flexibilidade ao algoritmo, porém aumenta consideravelmente a busca pela topologia mais apropriada. A definição arbitrária a priori do tamanho do sistema pode ser interessante no caso de um sistema embarcado, no qual o limite físico do dispositivo é conhecido e bastante reduzido. Porém, ajustar o tamanho do sistema à capacidade do dispositivo, não garante que esta seja a melhor estratégia do ponto de vista do compromisso entre complexidade e desempenho, pois é possível que sistemas menores possuam ainda um desempenho superior. Portanto, a questão inicial é como definir o tamanho do sistema para que se possa executar o método baseado na entropia dos estados. Optou-se arbitrariamente por utilizar os tamanhos dos melhores sistemas obtidos pelo novo método GEA, que será introduzido no próximo capítulo, pois dessa forma será possível fazer uma comparação mais direta entre tais métodos.

Outro aspecto importante reside no fato de que o método da entropia implementa um incremento gradativo no tamanho do sistema, juntamente com um treinamento baseado no algoritmo *Segmental K-means*. Tal algoritmo é mais simples que o de Baum-Welch, e a cada iteração utiliza o próprio modelo para realizar um alinhamento de Viterbi. Dessa forma, as iterações iniciais são realizadas utilizando uma segmentação pouco confiável, devido ao estágio inicial precário do modelo. À medida que o sistema é incrementado e treinado, as segmentações passam a ser mais confiáveis. Neste ponto, surge idéia de se utilizar uma segmentação mais elaborada na primeira iteração do algoritmo, obtida através de uma segmentação manual ou a partir de um sistema previamente treinado, ou até mesmo, em um caso extremo, utilizar tal segmentação durante todo o processo de incremento do tamanho do sistema. Neste último caso, o algoritmo se torna mais rápido devido ao fato de não se realizar mais o alinhamento de Viterbi durante o crescimento dos HMMs. Assim, alguns testes foram realizados com a alteração no algoritmo original, utilizando uma segmentação fixa durante todas as iterações do método da entropia.

Os resultados foram obtidos para as duas gramáticas utilizadas na decodificação, pelas mesmas razões apresentadas anteriormente, e as comparações foram realizadas com os sistemas contendo 1188 Gaussianas e 1296 Gaussianas, que correspondem aos sistemas de referência (número fixo de Gaussianas por estado) com os melhores desempenhos, quando utilizadas as gramáticas *Word-pairs* e *Back-off bigram* respectivamente.

Word-pairs

Os experimentos iniciais com o método da entropia foram realizados utilizando a gramática *Word-pairs*, e se encontram presentes na Tabela 4.4. Os sistemas com número variado de Gaussianas por estado possuem um total de 996 Gaussianas e os incrementos de N estados a cada iteração variaram de 40 até 100.

O melhor resultado foi obtido para o incremento de 100 estados a cada iteração do algoritmo. Tal sistema possui um desempenho que supera o de referência em 0,57% e 1,1%, em termos de taxa de reconhecimento de palavras e *accuracy* respectivamente, e ao mesmo tempo possui 16,2% menos parâmetros.

Neste método, todos os sistemas obtidos possuem o mesmo tamanho, pois tal variável foi definida no início da análise. Portanto, a escolha do melhor modelo ocorre baseada exclusivamente nos desempenhos. É importante ressaltar novamente que a escolha do tamanho final do sistema segue um critério arbitrário, de acordo com o algoritmo. Porém, neste trabalho, a escolha foi baseada em resultados obtidos através de outro método, que será introduzido posteriormente. Portanto, na ausência de outro método capaz de fornecer um valor apropriado para o tamanho do sistema, a busca para a obtenção da topologia capaz de fornecer tais resultados certamente seria mais custosa.

Tab. 4.4: Desempenho dos modelos obtidos através do método da entropia. A comparação foi realizada com o sistema de referência contendo 1188 Gaussianas (11 por estado). Os valores entre parênteses indicam a diferença entre os sistemas obtidos por tal método e o de referência. O alinhamento de Viterbi é realizado a cada iteração do algoritmo.

Incremento de N estados	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	Economia de Parâmetros (%)
40	996	93,8 (-0,27)	91,86 (+0,61)	16,2
50	996	93,91 (-0,16)	91,48 (+0,23)	16,2
60	996	94,6 (+0,53)	91,78 (+0,53)	16,2
70	996	94,14 (+0,07)	91,94 (+0,69)	16,2
80	996	94,03 (-0,04)	92,09 (+0,84)	16,2
90	996	94,56 (+0,49)	92,24 (+0,99)	16,2
100	996	94,64 (+0,57)	92,35 (+1,1)	16,2

Na seqüência, obtiveram-se os resultados através do método da entropia utilizando uma segmentação fixa, gerada pelo melhor sistema de reconhecimento de fala disponível (obtido pelo novo método GEA que será introduzido no próximo Capítulo), ao invés de realizar o alinhamento de Viterbi a cada iteração do algoritmo. A Tabela 4.5 mostra os resultados em tal condição.

Tab. 4.5: Desempenho dos modelos obtidos através do método da entropia, utilizando uma segmentação fixa, ao invés do alinhamento de Viterbi em cada iteração. A comparação foi realizada com o sistema de referência contendo 1188 Gaussianas (11 por estado). Os valores entre parênteses indicam a diferença entre os sistemas obtidos por tal método e o de referência.

Incremento de N estados	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	Economia de Parâmetros (%)
40	996	93,53 (-0,54)	90,79 (-0,46)	16,2
50	996	93,5 (-0,57)	90,95 (-0,3)	16,2
60	996	93,84 (-0,23)	91,59 (+0,34)	16,2
70	996	93,72 (-0,35)	91,21 (-0,04)	16,2
80	996	94,64 (+0,57)	91,82 (+0,57)	16,2
90	996	94,29 (+0,22)	91,63 (+0,38)	16,2
100	996	94,79 (+0,72)	91,71 (+0,46)	16,2

A alteração no método da entropia também permitiu a obtenção de sistemas com 16,2% menos parâmetros que o de referência, com desempenho superior e fornecendo o maior valor da taxa de reconhecimento de palavras, que foi de 94,79%. Entretanto, tal sistema obtido para $\mathrm{N} = 100$, forneceu um valor de *accuracy* de 91,71%, apresentando portanto um desempenho inferior àquele obtido anteriormente para $\mathrm{N} = 100$ (vide Tabela 4.4), seguindo o método da entropia original.

É importante notar que, em ambos os casos, os melhores resultados foram obtidos com os maiores valores de incremento N testados, próximos ao valor máximo possível ($\mathrm{N} = 108$). Tal fato traz indícios de que um incremento mais detalhado, onde o extremo seria de um estado a cada iteração ($\mathrm{N} = 1$), talvez seja menos eficiente, tornando o processo muito oneroso.

Os valores médios de taxa de reconhecimento e *accuracy*, obtidos pela proposta original do método baseado na entropia, são de (94,24 $\pm$ 0,33)% e (91,96 $\pm$ 0,29)%, respectivamente, com nível de confiança de 95%. Os resultados obtidos com a utilização da segmentação fixa no método da entropia fornecem (94,04 $\pm$ 0,49)% e (91,39 $\pm$ 0,37)% como valores médios de taxa de reconhecimento e *accuracy*, respectivamente. Pode-se observar então através dos valores médios de *accuracy*, que é mais interessante utilizar o método da entropia seguindo a proposta original, ao invés de utilizar uma segmentação fixa. No entanto, os valores médios de taxa de reconhecimento de palavras foram bem próximos entre si, mostrando que a idéia de se utilizar uma segmentação fixa para diminuir o tempo de treinamento também é válida.

Back-off bigram

Os últimos resultados obtidos através do método da entropia, utilizando a gramática *Back-off bigram*, estão indicados nas Tabelas 4.6 e 4.7. Novamente os tamanhos dos sistemas foram definidos de acordo com os tamanhos daqueles que apresentaram os melhores desempenhos obtidos pelo novo GEA, no intuito de se realizar comparações posteriores entre os métodos. Dessa forma, os sistemas foram determinados com 1040 e 1055 Gaussianas. Nestes experimentos o alinhamento de Viterbi é realizado a cada iteração do algoritmo.

Tab. 4.6: Desempenho dos modelos obtidos através do método da entropia. A comparação foi realizada com o sistema de referência contendo 1296 Gaussianas (12 por estado). Os valores entre parênteses indicam a diferença entre os sistemas obtidos por tal método e o de referência. O alinhamento de Viterbi é realizado a cada iteração do algoritmo.

Incremento de N estados	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	Economia de Parâmetros (%)
40	1040	80,3 (-0,71)	59,8 (+1,55)	19,8
50	1040	81,2 (+0,19)	60,67 (+2,42)	19,8
60	1040	80,56 (-0,45)	58,7 (+0,45)	19,8
70	1040	81,62 (+0,61)	60,51 (+2,26)	19,8
80	1040	81,47 (+0,46)	60,44 (+2,19)	19,8
90	1040	81,05 (+0,04)	61,01 (+2,76)	19,8
100	1040	81,24 (+0,23)	61,57 (+3,32)	19,8

Tab. 4.7: Desempenho dos modelos obtidos através do método da entropia. A comparação foi realizada com o sistema de referência contendo 1296 Gaussianas (12 por estado). Os valores entre parênteses indicam a diferença entre os sistemas obtidos por tal método e o de referência. O alinhamento de Viterbi é realizado a cada iteração do algoritmo.

Incremento de N estados	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	Economia de Parâmetros (%)
40	1055	80,3 (-0,71)	58,85 (+0,6)	18,6
50	1055	80,56 (-0,45)	59,38 (+1,13)	18,6
60	1055	81,16 (+0,15)	59,83 (+1,58)	18,6
70	1055	81,58 (+0,57)	60,67 (+2,42)	18,6
80	1055	81,09 (+0,08)	59,98 (+1,73)	18,6
90	1055	80,98 (-0,03)	61,31 (+3,06)	18,6
100	1055	81,39 (+0,38)	61,50 (+3,25)	18,6

Deve-se notar que os sistemas obtidos neste ponto, apresentam sensível ganho de desempenho, quando comparados ao sistema de referência e também quando comparados àqueles obtidos pelo BIC (vide Tabela 4.3).

4.2.3 Método Discriminativo

A utilização de medidas discriminativas para a seleção de topologias tem sido pouco explorada na literatura, apesar da existência de diversos métodos para a determinação da complexidade de modelos estatísticos.

O método em questão utiliza informações extraídas da segmentação da base de dados de treinamento, realizada através do alinhamento forçado de Viterbi. Além disso, identifica dois tipos de modelos, de acordo com a complexidade: "não agressivos", que correspondem aos modelos com baixa resolução acústica e que necessitam de mais Gaussianas, e os "invasivos", correspondendo aos modelos com alta resolução acústica e que interferem em outros modelos.

A idéia básica consiste em aumentar a complexidade dos HMMs de um determinado sistema, de acordo com uma medida discriminativa que indica quais estados do HMM necessitam de uma maior resolução acústica. Dessa forma, os modelos dos estados identificados como "não agressivos" são substituídos por modelos correspondentes extraídos de um sistema maior. Portanto, tal método requer inicialmente pelo menos dois sistemas contendo um número fixo de componentes por estado, um com baixa resolução e outro com alta resolução acústica. Além disso, outra condição inicial para o algoritmo é o limiar abaixo do qual os estados são considerados "não agressivos".

A notação do trabalho original (PB00) $M_X \times M_Y$-lim, foi utilizada para representar os sistemas obtidos através deste método, em que M_X é o sistema menor contendo "x" Gaussianas por estado,

M_y é o sistema maior contendo "y" Gaussianas por estado, e "lim" é o limiar adotado.

Neste ponto, um aspecto importante que deve ser discutido quando se deseja encontrar a topologia que apresente o melhor compromisso entre complexidade e desempenho, é a escolha dos sistemas M_x e M_y. Admitindo a hipótese de se realizar a escolha mais adequada de tais sistemas, é razoável esperar que a topologia mista contendo apenas dois números distintos de componentes por estado para o modelo final não seja a mais plausível do ponto de vista do compromisso desejado neste trabalho, pois a complexidade de cada estado pode variar continuamente de um até um determinado número inteiro de Gaussianas.

A estratégia proposta no trabalho original afirma, em geral, ser suficiente apenas o tratamento dos modelos dos estados considerados "não agressivos", e portanto nenhuma atenção é dedicada aos modelos "invasivos". A determinação dos modelos "invasivos" é utilizada para reforçar a idéia de se aumentar a resolução acústica dos modelos "não agressivos". Porém, deve-se notar que, apesar da complementaridade do problema, a existência de modelos "invasivos" pode ser indício da ocorrência de sobre-parametrização e, neste caso, uma abordagem envolvendo a diminuição do número de parâmetros do modelo é mais apropriada do que o enriquecimento dos modelos considerados com baixa complexidade, principalmente do ponto de vista da robustez do sistema. Resumidamente, o método sugere que o sistema final contendo dois números distintos de Gaussianas por estado tenha um tamanho menor do que o do sistema M_y e desempenho maior do que o do sistema M_x.

Os sistemas de referência (número fixo de Gaussianas por estado) contendo 1188 e 1296 Gaussianas, que apresentaram os melhores desempenhos no reconhecimento, foram utilizados como os sistemas M_y, nos testes com as gramáticas *Word-pairs* e *Back-off bigram*, respectivamente.

Word-pairs

Os resultados obtidos através do método discriminativo, utilizando-se a gramática *Word-pairs*, encontram-se presentes na Tabela 4.8.

Os resultados mostram que os sistemas mistos obtidos com menos parâmetros que o melhor de referência, apresentam pelo menos uma pequena perda no desempenho, não tendo ocorrido portanto aumento no desempenho em caso algum. Entretanto, se a comparação for realizada entre o sistema de referência e o sistema misto com aproximadamente o mesmo tamanho, pode-se observar ganho de desempenho. Por exemplo, comparando-se o sistema $M_5 \times M_{11}$-$0,60$ que contém 756 Gaussianas com o de referência de mesmo tamanho, nota-se um ganho de desempenho do sistema misto em relação ao de referência de 0,84% e 1,29% em termos de taxa de reconhecimento e *accuracy*, respectivamente. É importante ressaltar que o método discriminativo apresenta vantagens quando a comparação é feita entre o sistema com número variado de componentes e o de referência com aproximadamente o mesmo tamanho, mas não mostrou vantagens quanto ao principal objetivo deste trabalho, ou seja,

Tab. 4.8: Desempenho dos modelos obtidos através do método discriminativo. A comparação foi realizada com o sistema de referência contendo 1188 Gaussianas (11 por estado). Os valores entre parênteses indicam a diferença entre os sistemas obtidos por tal método e o de referência.

Sistema misto $M_X \times M_Y$-lim	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	Economia de Parâmetros (%)
$M_5 \times M_{11}$-$0,45$	714	92,89 (-1,18)	90,03 (-1,22)	39,9
$M_5 \times M_{11}$-$0,60$	756	93,31 (-0,76)	90,68 (-0,57)	36,4
$M_5 \times M_{11}$-$0,85$	792	93,69 (-0,38)	90,99 (-0,26)	33,3
$M_7 \times M_{11}$-$0,45$	872	93,08 (-0,99)	90,22 (-1,03)	26,6
$M_7 \times M_{11}$-$0,60$	888	92,81 (-1,26)	90,11 (-1,14)	25,3
$M_7 \times M_{11}$-$0,85$	908	93,08 (-0,99)	90,22 (-1,03)	23,6
$M_9 \times M_{11}$-$0,45$	1014	93,72 (-0,35)	90,99 (-0,26)	14,6
$M_9 \times M_{11}$-$0,60$	1024	93,69 (-0,38)	90,91 (-0,34)	13,8
$M_9 \times M_{11}$-$0,85$	1036	93,38 (-0,69)	90,45 (-0,8)	12,8
$M_{10} \times M_{11}$-$0,45$	1094	93,95 (-0,12)	90,95 (-0,3)	7,9
$M_{10} \times M_{11}$-$0,60$	1097	93,91 (-0,16)	90,64 (-0,61)	7,7
$M_{10} \times M_{11}$-$0,85$	1101	93,88 (-0,19)	90,68 (-0,57)	7,3

determinar sistemas menores que o melhor de referência, mas que apresentem pelo menos o mesmo desempenho. Tal fato pode ser explicado provavelmente pela própria restrição do método, que permite ao sistema final ter apenas dois números possíveis de Gaussianas por estado.

O sistema misto que mais se aproximou do objetivo desejado foi o $M_{10} \times M_{11}$-$0,45$, que possui 7,9% menos parâmetros que o de referência e apresenta a menor perda de desempenho, de 0,12% e 0,3% em termos de taxa de reconhecimento e *accuracy*, respectivamente.

Back-off bigram

Os resultados finais obtidos com a utilização da gramática *Back-off bigram* estão indicados na Tabela 4.9. Neste caso, entretanto, tentou-se avaliar uma faixa maior de limiares em relação àqueles escolhidos anteriormente.

O sistema que mais se aproximou do objetivo desejado foi o $M_7 \times M_{12}$-$0,8$, que possui 27,8% menos parâmetros do que o melhor de referência, apresentando uma perda de 0,71% na taxa de reconhecimento e um ganho no *accuracy* de 1,24%. Tais experimentos também mostram uma tendência de aumento na taxa de reconhecimento à medida que aumenta o valor do limiar, o que não foi possível de ser observado nos experimentos com a gramática *Word-pairs*. É importante mencionar novamente que, apesar de não ser o foco deste trabalho, a eficiência do método pode ser notada de forma mais clara quando os sistemas mistos são comparados com os de referência contendo aproximadamente o mesmo tamanho.

Tab. 4.9: Desempenho dos modelos obtidos através do método discriminativo. A comparação foi realizada com o sistema de referência contendo 1296 Gaussianas (12 por estado). Os valores entre parênteses indicam a diferença entre os sistemas obtidos por tal método e o de referência.

Sistema misto $M_X \times M_Y$-lim	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	Economia de Parâmetros (%)
$M_5 \times M_{12}$-0,2	617	76,66 (-4,35)	52,08 (-6,17)	52,4
$M_5 \times M_{12}$-0,4	736	77,38 (-3,63)	52,76 (-5,49)	43,2
$M_5 \times M_{12}$-0,6	792	78,71 (-2,3)	54,69 (-3,56)	38,9
$M_5 \times M_{12}$-0,8	820	78,63 (-2,38)	55,64 (-2,61)	36,7
$M_7 \times M_{12}$-0,2	806	78,18 (-2,83)	55,41 (-2,84)	37,8
$M_7 \times M_{12}$-0,4	876	78,48 (-2,53)	55,79 (-2,46)	32,4
$M_7 \times M_{12}$-0,6	921	79,39 (-1,62)	57,68 (-0,57)	28,9
$M_7 \times M_{12}$-0,8	936	80,3 (-0,71)	59,49 (+1,24)	27,8
$M_9 \times M_{12}$-0,2	1002	79,46 (-1,55)	57,26 (-0,99)	22,7
$M_9 \times M_{12}$-0,4	1038	79,46 (-1,55)	57,11 (-1,14)	19,9
$M_9 \times M_{12}$-0,6	1071	80,18 (-0,83)	59,27 (+1,02)	17,4
$M_9 \times M_{12}$-0,8	1086	80,33 (-0,68)	58,66 (+0,41)	16,2
$M_{11} \times M_{12}$-0,2	1197	79,99 (-1,02)	57,68 (-0,57)	7,6
$M_{11} \times M_{12}$-0,4	1209	80,41 (-0,6)	58,55 (+0,3)	6,7
$M_{11} \times M_{12}$-0,6	1213	79,8 (-1,21)	58,36 (+0,11)	6,4
$M_{11} \times M_{12}$-0,8	1218	79,99 (-1,02)	57,79 (-0,46)	6,0

4.3 Discussão

No intuito de atingir o principal objetivo deste trabalho, os três métodos apresentados enfrentaram uma maior dificuldade na determinação da topologia dos HMMs durante a utilização da gramática *Back-off bigram*, o que pode ser explicado possivelmente pela maior flexibilidade conferida ao processo de decodificação neste caso.

Portanto, a comparação entre os métodos pode ser feita de forma mais rigorosa através dos resultados obtidos com a gramática *Back-off bigram*. Neste sentido, definiu-se um fator de desempenho (F_d) em função dos ganhos obtidos em termos da taxa de reconhecimento de palavras (G_{TR}) e do *accuracy* (G_{acc}), conforme a Equação 4.3.

$$F_d = G_{TR} + G_{acc}. \tag{4.3}$$

Dessa forma, a Tabela 4.10 mostra resumidamente os melhores sistemas obtidos através de cada método, de acordo com o fator de desempenho.

O melhor desempenho, independentemente da economia, foi obtido então pelo método da entropia. Entretanto, se a busca for realizada no intuito de se escolher o modelo que apresente a maior

Tab. 4.10: Melhores sistemas obtidos através do BIC, método da entropia e método discriminativo, de acordo com os valores de F_d. A comparação foi realizada com o sistema de referência contendo 1296 Gaussianas (12 por estado).

Método	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	F_d (%)	Economia de Parâmetros (%)
BIC	1110	80,75 (-0,26)	57,60 (-0,65)	-0,91	14,4
Entropia	1055	81,39 (+0,38)	61,50 (+3,25)	+3,63	18,6
Discriminativo	936	80,3 (-0,71)	59,49 (+1,24)	+0,53	27,8

economia e pelo menos o mesmo desempenho do sistema de referência, ou seja, para $F_d \geq 0$, obtém-se então a Tabela 4.11.

Tab. 4.11: Sistemas mais econômicos obtidos através do BIC, método da entropia e método discriminativo, para a condição $F_d \geq 0$. A comparação foi realizada com o sistema de referência contendo 1296 Gaussianas (12 por estado).

Método	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)	F_d (%)	Economia de Parâmetros (%)
BIC	-	-	-	-	-
Entropia	1040	81,24 (+0,23)	61,57 (+3,32)	+3,55	19,8
Discriminativo	936	80,3 (-0,71)	59,49 (+1,24)	+0,53	27,8

Deve-se notar que apesar do método discriminativo ter fornecido o sistema mais econômico e ao mesmo tempo ter atendido à condição $F_d \geq 0$, não se pode obter conclusões acerca de qual o método mais eficiente para esta finalidade, visto que o menor sistema gerado pelo método da entropia possui 1040 Gaussianas, e o BIC não forneceu modelos que satisfizessem a condição desejada. Porém, verificou-se que o desempenho do BIC foi inferior ao dos demais métodos apresentados.

Os resultados mostram também que é possível se obter um ganho considerável, em detrimento da economia de parâmetros, ou uma economia considerável, em detrimento do ganho de desempenho.

Portanto, é importante se definir inicialmente as prioridades para o processo de determinação da topologia dos HMMs. Neste sentido, a fim de simplificar as referências posteriores, definem-se três objetivos que podem ser priorizados durante o processo de determinação da complexidade das misturas Gaussianas:

- Objetivo I - Determinação de um sistema que apresente um desempenho significativamente superior, quando comparado ao sistema de referência com aproximadamente o mesmo tamanho.
- Objetivo II - Determinação de um sistema menor que o melhor sistema de referência e que apresente o maior desempenho possível.

- Objetivo III - Determinação de um sistema que apresente pelo menos o mesmo desempenho do melhor sistema de referência ($F_d \geq 0$), e que possua a menor complexidade possível.

Neste trabalho o enfoque principal é atingir um melhor compromisso entre complexidade e desempenho do que o melhor sistema de referência, o que corresponde aos objetivos II e III.

Em última análise, os três métodos apresentados são comprovadamente eficazes na determinação de sistemas contendo um número variado de Gaussianas por estado, quando os desempenhos de tais sistemas são comparados com os de referência que possuem aproximadamente o mesmo tamanho (Objetivo I), o que pode ser observado na Tabela 4.12. Entretanto, deve-se destacar também que essa tarefa é mais simples do que a da determinação de sistemas com número variado de componentes que superem o desempenho do melhor sistema de referência (Objetivo II), ou que apresentem pelo menos o mesmo desempenho do melhor sistema de referência mas que possuam consideravelmente menos parâmetros (Objetivo III).

Tab. 4.12: Comparação entre sistemas com número variado de componentes por estado, obtidos através do BIC, método da entropia, método discriminativo, e o sistema de referência com aproximadamente o mesmo tamanho (1080 Gaussianas).

Método	Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)
BIC	1092	80,71 (+1,06)	57,3 (-0,15)
Entropia	1055	81,39 (+1,74)	61,50 (+4,05)
Discriminativo	1086	80,33 (+0,68)	58,66 (+1,21)
Sistema de Referência	1080	79,65	57,45

4.4 Conclusões

Os três métodos se mostraram eficazes na determinação de sistemas que apresentam um melhor compromisso entre desempenho e complexidade dos modelos. Tal fato pode ser facilmente observado pelas comparações entre os sistemas com número variado de componentes por estado e os sistemas de referência com aproximadamente o mesmo tamanho. Entretanto, nem todos os métodos tiveram desempenho satisfatório na determinação de sistemas de acordo com os principal objetivo do trabalho, que consiste na busca de sistemas que possuam pelo menos o mesmo desempenho do melhor de referência e que possuam consideravelmente menos parâmetros.

O Critério de Informação Bayesiano (BIC), dentre os métodos analisados, foi o que apresentou as maiores dificuldades, principalmente quando priorizado o Objetivo III. Tal método é o mais simples do ponto de vista de algoritmo e das condições iniciais.

Em contrapartida, o método da entropia apresentou resultados satisfatórios, independentemente do objetivo definido. Tal método possui o algoritmo mais custoso, e a ausência da informação inicial sobre o tamanho do sistema, que no caso foi extraída através da utilização de outro método, poderia ter dificultado consideravelmente a busca pela topologia mais apropriada. A utilização de uma segmentação fixa pode ser utilizada no intuito de diminuir o custo computacional de tal algoritmo, porém os sistemas obtidos desta forma podem apresentar desempenho menor do que aqueles obtidos com a proposta original do algoritmo.

Por último, o método discriminativo também forneceu resultados satisfatórios para os três objetivos definidos, mesmo na ausência de uma segmentação manual da base de dados. Utilizou-se então uma segmentação obtida pelo alinhamento forçado de Viterbi, realizado a partir do sistema de reconhecimento disponível com o melhor desempenho.

As gramáticas utilizadas forneceram resultados semelhantes, a menos de um fator de escala. Contudo, verificou-se que a *Back-off bigram* requer um sistema mais robusto para que o desempenho durante a decodificação seja aceitável e, neste sentido, pode mostrar mais claramente a eficácia dos métodos analisados.

Capítulo 5

O Algoritmo de Eliminação de Gaussianas (GEA)

5.1 Introdução

A análise discriminativa tem sido bastante explorada no contexto de modelagem estatística em problemas de classificação. Neste sentido, existem diversos algoritmos de treinamento capazes de realizar a tarefa de estimação de parâmetros de forma a maximizar a discriminabilidade dos modelos.

Em linhas gerais, o processo de modelagem estatística consiste de 4 etapas interligadas: análise dos dados (coleta e tratamento inicial dos dados, extração de variáveis características do sistema), detecção de estrutura (determinação da topologia dos modelos), estimação de parâmetros e validação. Neste trabalho, uma análise discriminativa é utilizada na fase de detecção de estrutura no intuito de auxiliar a determinação da topologia mais apropriada do ponto de vista da relação entre complexidade e desempenho do sistema.

Conforme mencionado anteriormente, tal análise tem sido pouco explorada nesta etapa do processo de modelagem e, diferentemente de outras abordagens existentes, este trabalho utiliza medidas discriminativas para a redução da complexidade de HMMs no intuito de evitar o problema de sobreparametrização. Assim, sistemas contendo um número fixo de componentes por estado (referência) podem ser utilizados como ponto de partida para a obtenção de sistemas menores contendo um número variado de Gaussianas e com desempenho superior, sendo que o enfoque pode ser dado na economia de parâmetros ou no ganho de desempenho, de acordo com a prioridade estabelecida. É importante destacar que, apesar dos experimentos terem sido realizados apenas com Modelos Ocultos de Markov, a metodologia proposta pode ser estendida diretamente para qualquer modelo com mistura de Gaussianas.

Outro aspecto que deve ser notado a fim de se generalizar os conceitos do método proposto, en-

volve a questão da segmentação dos dados de fala. Para que o método possa ser aplicado, é necessário se ter uma base de treinamento segmentada, podendo-se dessa forma diferenciar os padrões existentes e, no caso de sinais de fala, a realização de tal tarefa é bastante onerosa. Em outros problemas de reconhecimento de padrões, é possível se obter tal segmentação de forma mais simples e precisa.

A base de dados pequena em Português do Brasil (não segmentada) foi utilizada nos experimentos durante a elaboração da versão inicial do novo método proposto e, posteriormente, os experimentos com a versão final foram realizados com a base de dados TIMIT (segmentada).

5.2 Proposta Inicial de uma Medida Discriminativa

A primeira questão que deve ser analisada no intuito de se obter informações sobre a discriminabilidade entre modelos é a necessidade de se ter uma base de dados segmentada, de tal forma que seja possível medir quantitativamente a capacidade de classificação de cada modelo, através de taxas de acerto e/ou *accuracy*. No caso específico de sinais de fala, existem algumas padronizações lingüísticas que permitem estabelecer fronteiras entre as menores unidades acústicas (fonemas) e, mesmo neste caso, algumas variações podem também ser aceitas. Tais fronteiras, em geral, são determinadas manualmente, como um refinamento de uma segmentação automática e, portanto, são encontradas com muito custo.

Na ausência de tal segmentação, uma alternativa é utilizar a segmentação automática obtida pelo alinhamento de Viterbi realizado contra a transcrição correta de cada sentença, e gerada a partir de um sistema de reconhecimento que possua um desempenho elevado. Uma vez que a base de dados pequena em Português não é segmentada, adotou-se tal estratégia a fim de se obter uma segmentação para essa base. Em contrapartida, a base de dados TIMIT possui uma segmentação manual precisa e, neste caso, permitirá avaliar a diferença existente entre a segmentação manual e a segmentação obtida simplesmente pelo alinhamento de Viterbi.

É mais complicado tratar os padrões observados no sinal de fala do que em outros sinais, no que tange à determinação da segmentação de referência, e portanto pode ser um teste bastante rigoroso para avaliar a eficácia do novo método.

Outra questão fundamental para a determinação da medida discriminativa, assumindo-se a disponibilidade de uma base de treinamento segmentada, é definir como calcular tal medida. Neste sentido, a verossimilhança pode ser utilizada para quantificar o quão ajustadas as variáveis do modelo se encontram em relação aos dados de treinamento. No caso de HMMs contínuos, que utilizam PDFs Gaussianas, isto corresponde à verificação da cobertura de cada Gaussiana sobre o espaço de parâmetros. Após concluído o processo de treinamento, a cobertura de um modelo sobre o espaço de características de outro modelo pode implicar na ocorrência de erros de reconhecimento durante a

decodificação. Tal fato se deve, em geral, à existência de parâmetros em excesso, o que caracteriza a sobre-parametrização.

A idéia inicial consiste basicamente em se calcular a cobertura de uma determinada Gaussiana sobre os dados associados ao estado cujo modelo contém tal Gaussiana, e na seqüência se calcular a cobertura da mesma em relação aos dados dos demais estados. Assim, quanto maior a cobertura sobre os dados associados ao próprio estado e menor a cobertura sobre os dados dos demais estados, maior a importância da Gaussiana para a discriminabilidade do modelo. O processo é então repetido para cada Gaussiana de cada estado.

Em um problema genérico, em que se deseja classificar dois padrões,"A" e "B", através de dois modelos com mistura de Gaussianas, "M_1" e "M_2" respectivamente, é desejável que as Gaussianas de M_1 apresentem valores de verossimilhança mais elevados do que os valores obtidos com M_2, quando avaliados com os dados rotulados como A e, da mesma forma, é desejável que as Gaussianas de M_2 apresentem valores de verossimilhança mais elevados do que os valores obtidos com M_1, quando avaliados com os dados rotulados como B. No caso prático, devido à existência de sobreposição das distribuições dos padrões, é possível que Gaussianas de M_1 apresentem valores de verossimilhança menores do que os valores obtidos com M_2, quando avaliados com os dados rotulados como A, e também que Gaussianas de M_2 apresentem valores de verossimilhança menores do que os valores obtidos com M_1, quando avaliados com os dados rotulados como B.

Dessa forma, utilizou-se como medida de importância de uma Gaussiana em relação a cada estado, a Probabilidade da Gaussiana Vencedora (WGP) (YV04a; YV04b), definida pela Equação (5.1).

$$P_{wg}^{(i;j;s)} = \frac{N_{wg}^{(i;j;s)}}{N_{frames}^{(s)}}, \tag{5.1}$$

em que $N_{wg}^{(i;j;s)}$ corresponde ao número de vezes em que a Gaussiana "i", que pertence ao estado "j", apresenta a maior verossimilhança utilizando dados do estado "s", e $N_{frames}^{(s)}$ é o número de quadros associados ao estado "s". Deve-se notar que, dessa forma, os valores dos coeficientes de ponderação das Gaussianas são desconsiderados.

Assim, analisando ainda o caso genérico para efeito de ilustração, idealmente as Gaussianas do modelo M_1 devem fornecer valores de P_{wg} próximos de 1, quando avaliadas com dados rotulados como A, e valores próximos de 0 quando avaliadas com dados rotulados como B. Pode-se repetir a análise facilmente e encontrar conclusões análogas para o modelo M_2.

Uma vez quantificada a importância de cada Gaussiana, definiu-se então uma medida da discriminabilidade de cada modelo, de tal forma que quanto maior o valor obtido através desta medida, maior a capacidade de classificação do modelo, e quanto menor o valor obtido, menor a capacidade de classificação do modelo. Dessa forma é possível avaliar o comportamento de cada componente dentro do

modelo e então decidir sobre a permanência ou exclusão da mesma. A constante discriminativa DC pode ser calculada pela Equação (5.2)

$$DC^{(j)} = \frac{\left[\sum_{i=1}^{N_j} P_{wg}^{(i;j;j)}\right]^K}{\sum_{s\neq j}^{N_s}\sum_{i=1}^{N_j} P_{wg}^{(i;j;s)}}, \tag{5.2}$$

em que "N_j" é o número de Gaussianas do estado "j", "N_s" é o número total de estados e "K" é o expoente de rigor.

O algoritmo proposto é baseado em um processo de eliminação de Gaussianas de um determinado modelo, uma de cada vez, e na observação do valor da constante discriminativa obtida após cada eliminação. Se o valor da constante discriminativa diminuir após a eliminação, então a contribuição da Gaussiana eliminada para o próprio modelo é maior do que a interferência da mesma em outros modelos, e portanto deve ser re-introduzida no sistema. Se o valor da constante discriminativa aumentar após a eliminação, então a contribuição da Gaussiana eliminada para o próprio modelo é menor do que a interferência da mesma em outros modelos, e portanto tal componente deve ser mantida fora do sistema.

Dessa forma, deve-se buscar as eliminações que maximizem a constante discriminativa, e repetir o processo iterativamente até que não haja novo aumento no valor de DC após a eliminação de qualquer Gaussiana restante no modelo. Neste ponto, é importante destacar que o expoente de rigor torna o critério de eliminação mais rigoroso, ou seja, à medida que se aumenta o valor de "K", torna-se mais difícil eliminar Gaussianas de acordo com tal critério.

O modelo obtido após o processo de eliminação de Gaussianas é treinado finalmente com o algoritmo baseado em MLE e, uma vez atingido o critério de convergência, espera-se que o sistema reduzido tenha um desempenho consideravelmente superior ao original ou pelo menos que apresente o mesmo desempenho. A explicação reside no fato de que é mais difícil a ocorrência de sobre-parametrização durante o treinamento realizado com um sistema com menos graus de liberdade, e portanto é mais difícil que Gaussianas modelando uma determinada distribuição convirjam para posições relativas a outras distribuições no espaço acústico. Em resumo e de forma bastante simplificada, o algoritmo de treinamento baseado em MLE tem como finalidade maximizar a cobertura de uma determinada distribuição, pelas Gaussianas que a modelam, mesmo que para isso haja o aumento na interferência entre modelos diferentes e conseqüentemente o aumento de erros de classificação. Em contrapartida, os algoritmos de treinamento discriminativo priorizam a discriminabilidade dos modelos, mas o custo computacional para este fim é maior do que o necessário para o treinamento baseado em MLE. Portanto, a seleção discriminativa da topologia dos modelos pode atenuar o ponto fraco do treinamento baseado em MLE, que não leva em conta a capacidade de discriminação dos

modelos, e ao mesmo tempo usufruir da principal vantagem que é simplicidade e velocidade de tal algoritmo, comparado ao treinamento discriminativo. Uma outra possibilidade é utilizar o modelo simplificado obtido com o processo discriminativo de redução de Gaussianas em um treinamento discriminativo, que neste caso pode ser realizado em um tempo inferior ao necessário para treinar o sistema de referência original, devido à existência de menos parâmetros nos modelos.

A Figura 5.1 ilustra de forma esquemática o novo algoritmo discriminativo proposto. Deve-se notar que, neste algoritmo, o valor do expoente de rigor é a única condição que precisa ser definida inicialmente.

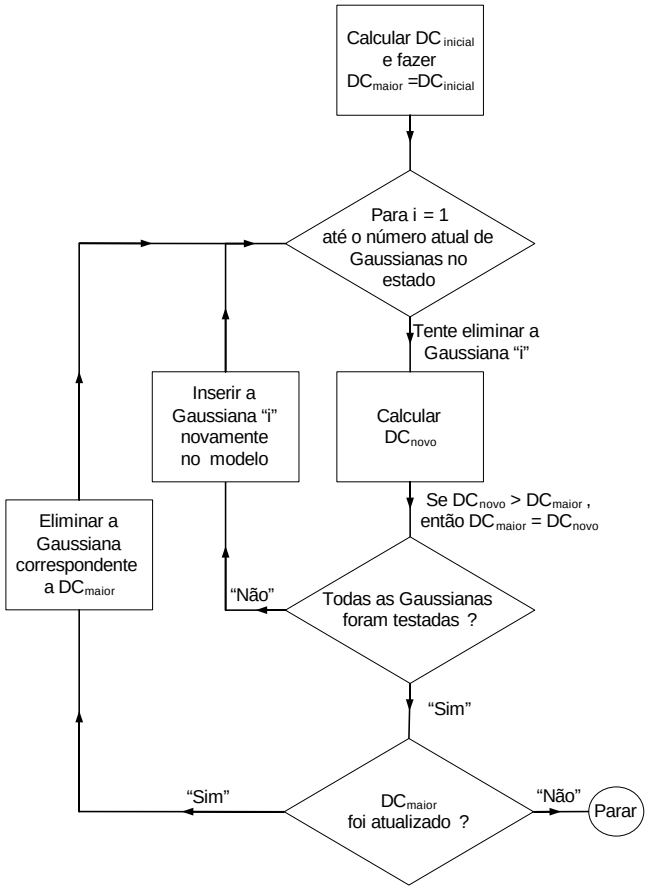


Fig. 5.1: Algoritmo discriminativo para eliminação de Gaussianas.

A questão agora é encontrar o valor de "K" que resulte no melhor sistema reduzido. A princípio, não se sabe qual o valor mais apropriado e, da mesma forma como nos demais métodos existentes na literatura, é necessário testar uma faixa de valores e observar a relação entre o tamanho e o desempenho dos sistemas obtidos. Dessa forma, escolhe-se então o sistema reduzido que se encontre mais próximo do objetivo desejado.

É intuitivo esperar que o valor do expoente aumente até um determinado valor máximo, além do

qual o rigor do critério se torna muito elevado, não permitindo a retirada de Gaussianas do modelo. Por outro lado, deve existir um valor mínimo de "K", abaixo do qual todos os sistemas obtidos apresentam uma elevada degradação no desempenho. Assim, é interessante determinar os valores extremos do expoente de rigor e dessa forma limitar a busca pelos sistemas reduzidos.

O método discriminativo proposto não impõe restrições em relação ao sistema cuja complexidade se deseja reduzir e, portanto, os experimentos iniciais foram realizados com diversos sistemas de referência. Entretanto, uma estratégia que pode ser adotada para efeito de simplificação do problema, por exemplo, é otimizar apenas o sistema de referência que apresente o melhor desempenho.

5.3 Redução da Complexidade de Sistemas com Número Fixo de Gaussianas por Estado

Os primeiros experimentos com o método discriminativo baseado na medida de WGP foram realizados com a base de dados pequena em Português, empregando a gramática *Word-pairs* no processo reconhecimento. Neste sentido, utilizaram-se diversos sistemas de referência como ponto de partida para o algoritmo, a fim de avaliar a eficácia do mesmo em diferentes condições iniciais.

As Figuras 5.2 a 5.7 mostram os desempenhos dos sistemas obtidos a partir do novo método discriminativo proposto. Os valores do expoente de rigor utilizados nos experimentos iniciais estão compreendidos no intervalo $2 \leq \mathrm{K} \leq 14$. Os sistemas obtidos para $\mathrm{K} = 2$ apresentam um desempenho consideravelmente inferior ao do sistema de referência original, enquanto o tamanho do sistema reduzido tende a saturar para valores de K próximos de 14. No último caso, pode-se interpretar que, a existência de Gaussianas eliminadas mesmo com o aumento elevado do rigor do critério, implica que tais componentes Gaussianas de fato não contribuem para a discriminabilidade dos modelos e portanto devem ser eliminadas. A Figura 5.8 mostra o tamanho dos sistemas reduzidos em função do valor do expoente de rigor empregado e, conforme pode ser observado, o tamanho dos sistemas obtidos pelo algoritmo discriminativo proposto segue um padrão exponencial.

Os resultados mostram que é possível obter sistemas reduzidos que apresentem pelo menos o mesmo desempenho do sistema de referência original ou que apresentem um ganho considerável de desempenho.

É importante notar que, na abordagem adotada anteriormente, o melhor sistema contendo um número fixo de Gaussianas por estado era utilizado como referência para efeito de comparação, enquanto nos experimentos iniciais com o novo algoritmo discriminativo, o sistema contendo um número fixo de Gaussianas por estado a partir do qual são obtidos os sistemas reduzidos é utilizado como referência. Portanto, a referência varia em cada um dos gráficos apresentados (Figuras 5.2 a 5.7).

A mudança de referência tem como única finalidade mostrar que os sistemas de referência contendo um número fixo de Gaussianas por estado possuem, em geral, componentes em excesso e que se encontram modelando os padrões incorretos. Assim, o algoritmo discriminativo é capaz de detectar e eliminar tais componentes, permitindo a diminuição da complexidade do sistema original e também o aumento no desempenho.

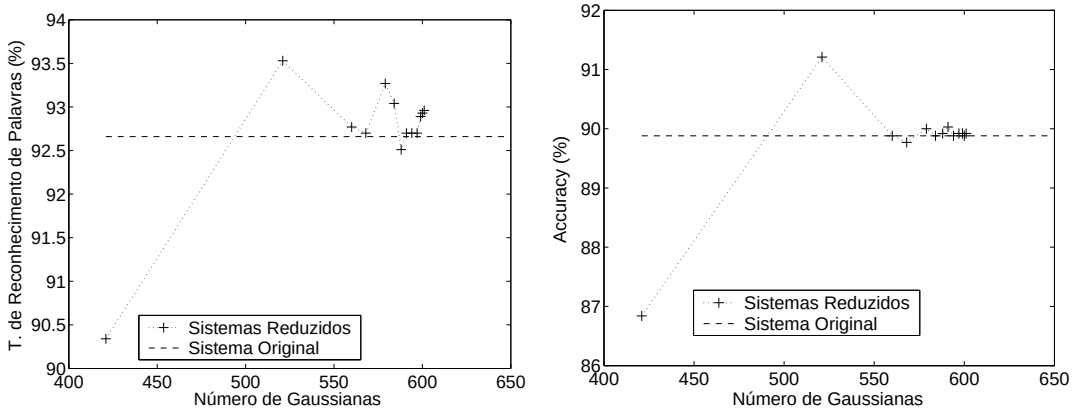


Fig. 5.2: Sistemas obtidos a partir da redução da complexidade do sistema de referência contendo 6 componentes por estado.

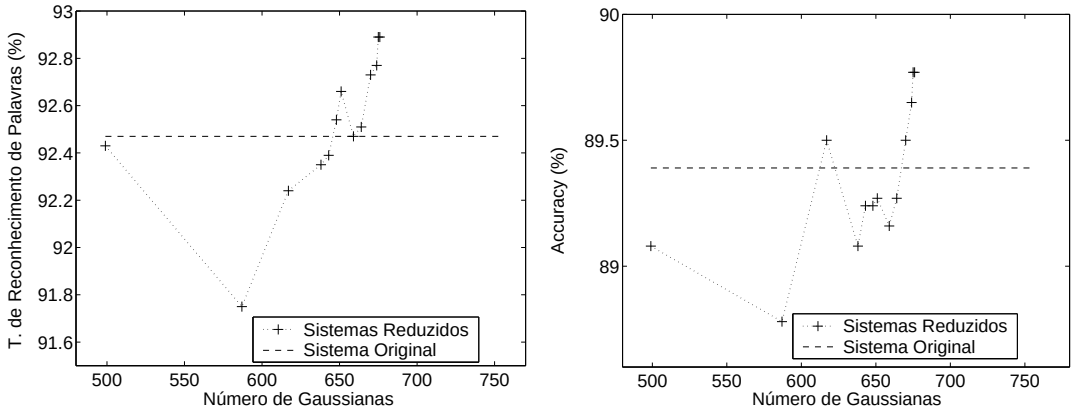


Fig. 5.3: Sistemas obtidos a partir da redução da complexidade do sistema de referência contendo 7 componentes por estado.

Além disso, mesmo no caso de se utilizar o melhor sistema contendo um número fixo de componentes (1188 Gaussianas) como referência para comparação, que corresponde ao principal objetivo deste trabalho e à estratégia adotada no Capítulo 4, também é possível se obter sistemas reduzidos e que tenham aproximadamente o mesmo desempenho ($F_d \geq 0$), como pode ser observado na Tabela 5.1.

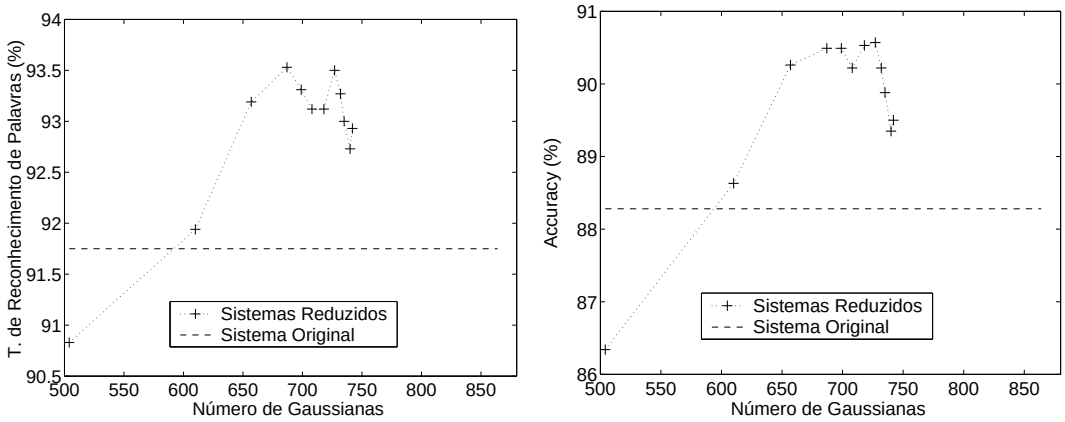


Fig. 5.4: Sistemas obtidos a partir da redução da complexidade do sistema de referência contendo 8 componentes por estado.

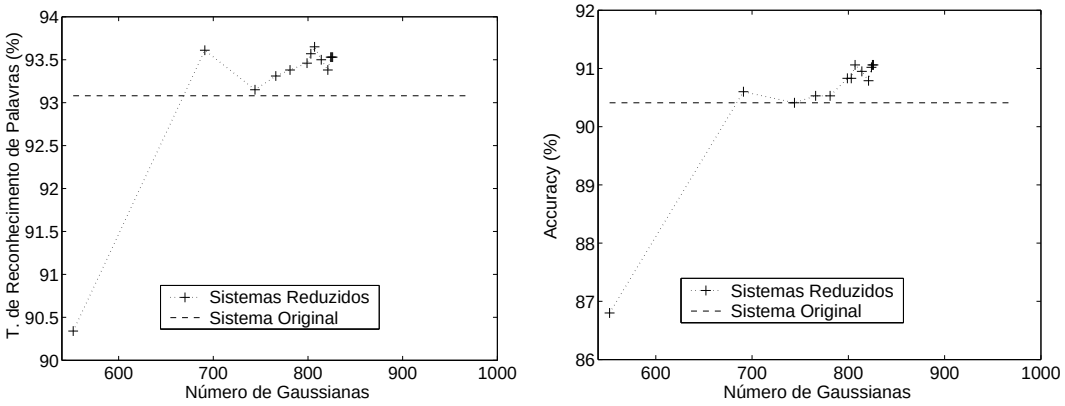


Fig. 5.5: Sistemas obtidos a partir da redução da complexidade do sistema de referência contendo 9 componentes por estado.

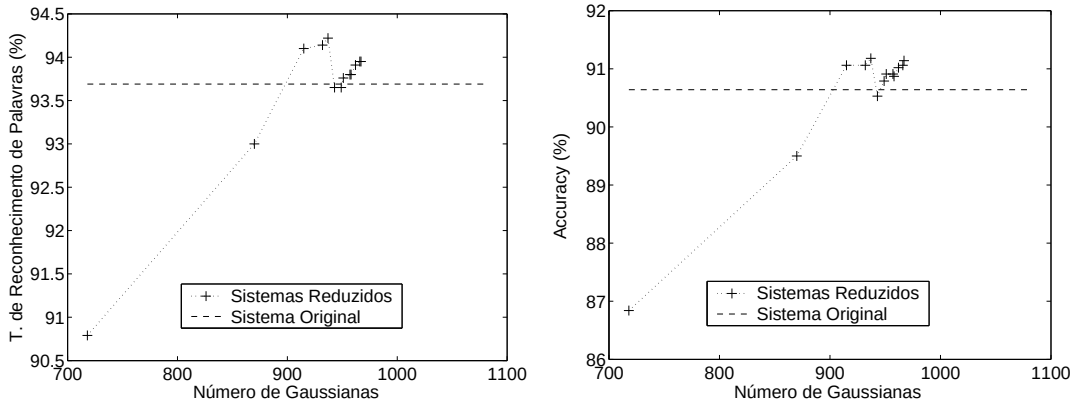


Fig. 5.6: Sistemas obtidos a partir da redução da complexidade do sistema de referência contendo 10 componentes por estado.

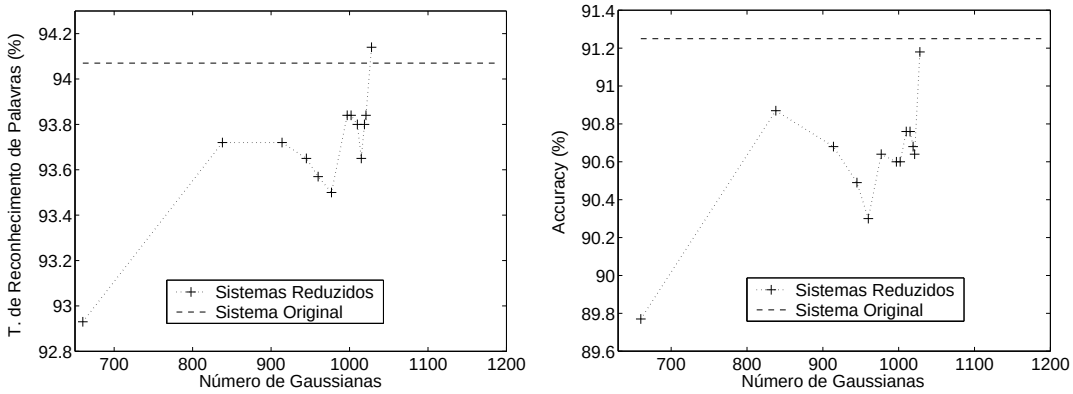


Fig. 5.7: Sistemas obtidos a partir da redução da complexidade do sistema de referência contendo 11 componentes por estado.

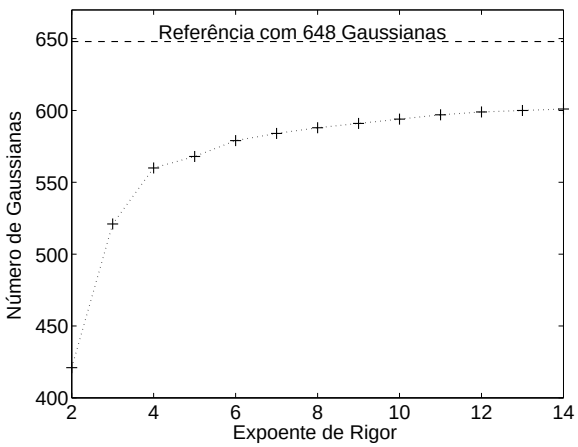
Tab. 5.1: Desempenho dos modelos obtidos através do método discriminativo que utiliza a medida de WGP. A comparação foi realizada com o sistema de referência contendo 1188 Gaussianas (11 Gaussianas por estado). Os valores entre parênteses indicam a diferença entre os sistemas obtidos por tal método e o de referência.

Sistema Original	Expoente de Rigor	Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d (%)	Economia de Parâmetros (%)
1080	6	937	94,22 (+0,15)	91,18 (-0,07)	+0,08	21,1
1188	14	1028	94,14 (+0,07)	91,18 (-0,07)	0	13,5

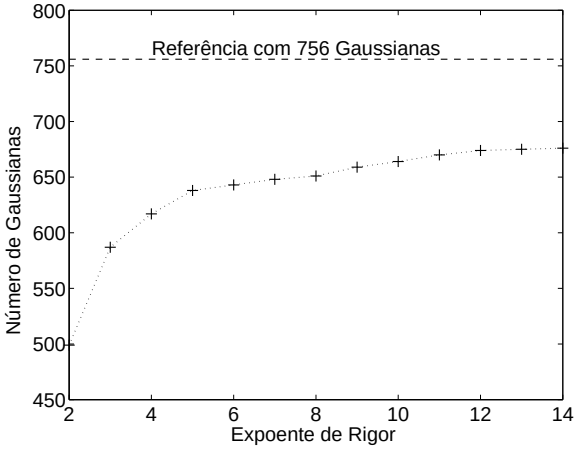
Entretanto, não foi possível se determinar um sistema reduzido e com um ganho de desempenho considerável, tendo como referência o melhor sistema com número fixo de componentes por estado. Uma possível explicação para este fato é a permanência de Gaussianas em excesso nos modelos mesmo após a aplicação do novo algoritmo discriminativo que, neste caso, podem ter convergido para as distribuições incorretas no espaço acústico durante o re-treinamento do sistema.

Neste ponto surge então uma dúvida com relação à eficácia do processo de eliminação de Gaussianas. Apesar do algoritmo discriminativo eliminar componentes responsáveis por erros de modelagem, tornando dessa forma o sistema menos flexível e mais robusto, ainda devem existir componentes que precisam ser eliminadas e que não são detectadas pelo novo critério discriminativo. Uma hipótese intuitiva é que algumas componentes se encontrem bastante próximas entre si no espaço acústico, sendo responsáveis por uma modelagem redundante e não tendo função primordial para a capacidade de classificação do modelo. Tal fato pode ocorrer, por exemplo, quando as componentes redundantes se encontram na parte central da distribuição que está sendo modelada.

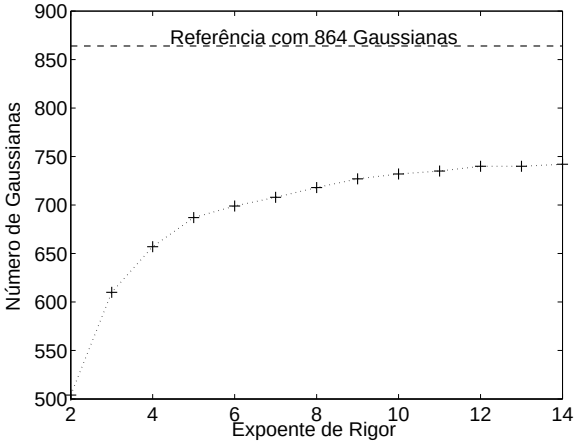
A solução mais simples que pode ser utilizada para evitar uma possível redundância na modelagem é se determinar a distância entre as Gaussianas de um modelo no espaço de características, e



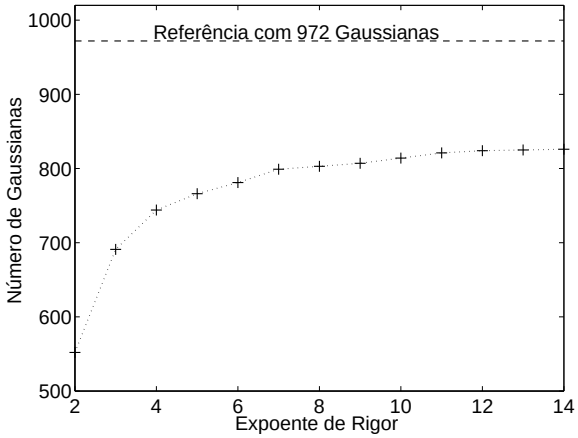
(a) Sistema original com 648 Gaussianas



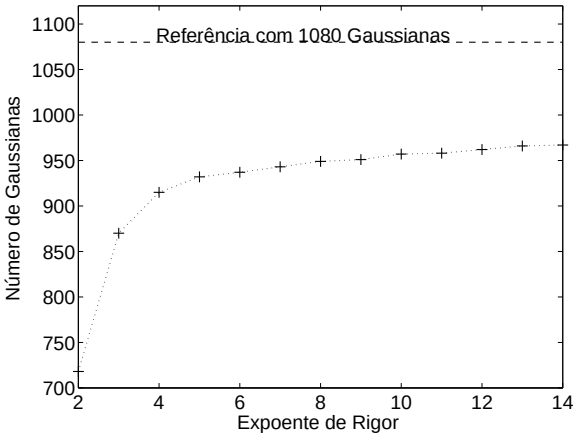
(b) Sistema original com 756 Gaussianas



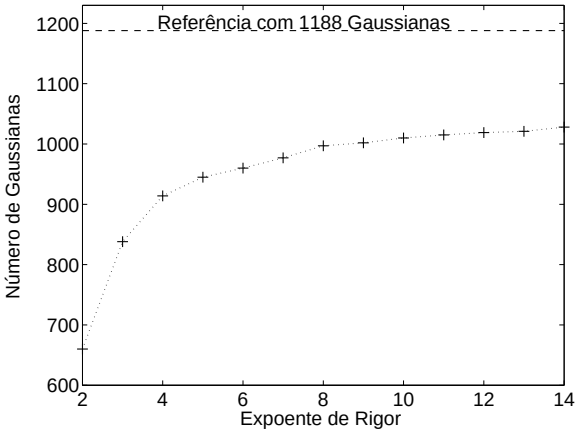
(c) Sistema original com 864 Gaussianas



(d) Sistema original com 972 Gaussianas



(e) Sistema original com 1080 Gaussianas



(f) Sistema original com 1188 Gaussianas

Fig. 5.8: Número de Gaussianas dos sistemas reduzidos. Os sistemas de referência contêm de 6 até 11 componentes por estado.

substituir as componentes que se encontrem próximas entre si, a menos de um limiar de distância pré-estabelecido, por uma única componente.

Deve-se notar também que uma modelagem mais detalhada, onde as Gaussianas podem se encontrar bastante próximas entre si, é interessante e desejável nas fronteiras das distribuições, pois dessa forma a elevada resolução acústica do modelo pode contribuir para uma melhor discriminabilidade do mesmo. Em contrapartida, o mesmo nível de resolução é desnecessário na parte central da distribuição, onde devem ocorrer menos erros de classificação. A questão é definir uma medida capaz de permitir diferentes níveis de resolução acústica para o modelo em função da localização das componentes no espaço de características.

5.4 Eliminação de Gaussianas Baseada na Análise Discriminativa e na Análise Interna

A medida mais simples para a determinação de grupos de componentes redundantes é a da distância Euclidiana. Neste sentido, as distâncias entre as componentes de um modelo são calculadas e aquelas que se encontrarem a menos de um limiar L_d pré-definido são substituídas pela Gaussiana que apresentar o maior determinante da matriz de covariância. A idéia é que quanto maiores as variâncias ao longo das dimensões acústica, maior o valor do determinante da matriz de covariância, e portanto maior é a cobertura da Gaussiana multidimensional escolhida, sobre o conjunto de dados que antes era coberto pelo grupo de Gaussianas.

Então, definiu-se como análise interna do modelo o processo de eliminação de Gaussianas baseado em medidas de distância Euclidiana, onde cada modelo é analisado separadamente dos demais, diferentemente da análise discriminativa que se baseia em medidas extraídas de todos os modelos simultaneamente. Na realidade, a análise interna e a análise discriminativa são duas faces do mesmo problema, pois são utilizadas para resolver dois aspectos diferentes do problema de sobreparametrização.

Portanto, a análise interna tem como foco a economia de parâmetros, visando apenas eliminar o excesso de Gaussianas localizadas na parte central da distribuição e evitando também que tais componentes convirjam para distribuições erradas durante o re-treinamento do sistema. É importante destacar que a análise interna não tem por finalidade a eliminação de Gaussianas responsáveis por erros classificação e, dessa forma, deve ser utilizada juntamente com a análise discriminativa.

Assim, partindo-se do melhor sistema reduzido através do novo método discriminativo, é possível se determinar o valor do limiar de distância para o qual é possível se obter a maior economia de parâmetros sem ocorrer degradação no desempenho em relação ao sistema utilizado como ponto de partida.

A Tabela 5.2 mostra os sistemas obtidos a partir da utilização conjunta do novo critério discriminativo baseado na WGP e da análise interna baseada em medidas de distância Euclidiana, tendo como ponto de partida para o processo de eliminação de Gaussianas o sistema de referência contendo 11 componentes por estado. Pode-se observar então que o melhor sistema reduzido possui 26,9% menos Gaussianas que o sistema de referência (1188 Gaussianas) e um fator de desempenho de +0,98%, o que corresponde a uma maior economia de parâmetros e um maior ganho de desempenho em relação aos sistemas obtidos anteriormente utilizando-se apenas o método discriminativo.

Tab. 5.2: Desempenho dos modelos obtidos através da análise discriminativa e interna dos modelos. As comparações foram realizadas com o melhor sistema original (11 Gaussianas por estado). Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência contendo 1188 Gaussianas.

K	L_d	Número de Gaussianas no Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
2	3,4	592	93,04 (-1,03)	90 (-1,25)	-2,28	50,2
3	3,4	754	93,95 (-0,12)	91,25 (0)	-0,12	36,5
4	3,4	825	94,26 (+0,19)	90,99 (-0,26)	-0,07	30,6
5	3,4	856	94,03 (-0,04)	90,91 (-0,34)	-0,38	27,9
6	3,4	869	94,48 (+0,41)	91,82 (+0,57)	+0,98	26,9
7	3,4	882	93,38 (-0,69)	90,41 (-0,84)	-1,53	25,8
8	3,4	902	93,69 (-0,38)	90,79 (-0,46)	-0,84	24,1
9	3,4	907	93,91 (-0,16)	91,21 (-0,04)	-0,2	23,7
10	3,4	914	93,65 (-0,42)	90,79 (-0,46)	-0,88	23,1
11	3,4	919	93,65 (-0,42)	90,76 (-0,49)	-0,91	22,6
12	3,4	921	93,61 (-0,46)	90,6 (-0,65)	-1,11	22,5
13	3,4	923	93,76 (-0,31)	90,64 (-0,61)	-0,92	22,3
14	3,4	930	94,03 (-0,04)	91,02 (-0,23)	-0,27	21,7

O único resultado obtido para a condição $F_d \geq 0$ forneceu um sistema contendo 26,9% menos parâmetros do que o de referência (1188 Gaussianas).

A análise interna utiliza a medida de distância Euclidiana para a eliminação das Gaussianas em excesso presentes nos modelos a partir de um limiar estabelecido. Uma alternativa para permitir que as Gaussianas localizadas nas fronteiras da distribuição sejam analisadas por limiares diferentes daqueles empregados para as Gaussianas localizadas na parte central da distribuição, é utilizar uma nova medida de distância Euclidiana modificada, definida pela Equação (5.3)

$$M_{d_{xy}} = \frac{d_{xy}}{\frac{\sum\limits_{i=x,y} P_{wg}^{(i;j;j)}}{\sum\limits_{i=x,y} P_{wg}^{(i;j;j)} + \sum\limits_{s \neq j} \sum\limits_{i=x,y} P_{wg}^{(i;j;s)}}}, \tag{5.3}$$

onde d_{xy} é dado por

$$d_{xy} = \sqrt{\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right) \cdot \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right)^T},$$

onde $\boldsymbol{\mu}_x$ e $\boldsymbol{\mu}_y$ são os vetores de média das Gaussianas *x* e *y* respectivamente.

A distância modificada utiliza os valores de P_{wg} como medida indireta da localização das Gaussianas no espaço acústico. À medida que a localização de uma determinada Gaussiana se aproxima das fronteiras da distribuição, os valores de P_{wg} calculados em relação aos demais estados aumentam, de tal forma que o denominador da Equação (5.3) diminui e portanto o valor de $M_{d_{xy}}$ é calculado com um aumento aparente na distância Euclidiana convencional d_{xy}. Por outro lado, à medida que a localização de uma determinada Gaussiana se aproxima da parte central da distribuição, os valores de P_{wg} calculados em relação aos demais estados diminuem, de tal forma que o denominador da Equação (5.3) se aproxima do valor 1 e portanto $M_{d_{xy}}$ se aproxima da distância Euclidiana convencional. Na prática, as distâncias calculadas para as Gaussianas da fronteira recebem uma correção de tal forma que tais Gaussianas "percebam" um limiar de distância inferior ao "percebido" pelas Gaussianas localizadas na parte central da distribuição. Com isso, pode-se valorizar diferentes resoluções acústicas em diferentes partes das distribuições no espaço de características.

A Tabela 5.3 mostra os resultados obtidos através da utilização conjunta do novo critério discriminativo e da análise interna com a medida de distância Euclidiana modificada.

O único resultado obtido para a condição $F_d \geq 0$ forneceu um sistema contendo 21% menos parâmetros do que o de referência (1188 Gaussianas). Apesar de tal resultado implicar em uma menor economia de parâmetros em relação ao resultado obtido com a utilização da distância Euclidiana convencional, pode-se observar que os sistemas obtidos através da associação da análise discriminativa com a análise interna utilizando a medida Euclidiana modificada apresentaram um maior desempenho para os maiores valores do expoente de rigor. Em contrapartida, os sistemas obtidos através da associação da análise discriminativa com a análise interna utilizando a medida Euclidiana convencional, apresentaram um maior desempenho para os menores valores do expoente de rigor. A Figura 5.9 ilustra os fatores de desempenho dos sistemas gerados pela análise conjunta, quando utilizadas a medida Euclidiana convencional e a modificada na etapa de análise interna dos modelos.

Os resultados não mostraram de forma clara qual a medida de distância mais vantajosa para a análise interna. Entretanto, é fundamental destacar que tais resultados mostram a evidente importância da realização da análise interna, após a análise discriminativa, para a obtenção de sistemas com um melhor compromisso entre tamanho e desempenho, independentemente da medida de distância adotada.

A análise interna também pode ser utilizada, em princípio, separadamente da análise discriminativa. Assim, adotando os mesmos limiares 3,4 e 6 para a medida de distância Euclidiana convencional

Tab. 5.3: Desempenhos obtidos através da análises discriminativa e interna dos modelos. As comparações foram realizadas com o melhor sistema de referência (11 Gaussianas por estado). Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência contendo 1188 Gaussianas.

K	L_d	Número de Gaussianas no Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
2	6	593	92,01 (-2,06)	88,93 (-2,32)	-4,38	50,1
3	6	756	93,38 (-0,69)	90,49 (-0,76)	-1,45	36,4
4	6	830	93,57 (-0,5)	90,68 (-0,57)	-1,07	30,1
5	6	861	93,61 (-0,46)	90,41 (-0,84)	-1,3	27,5
6	6	874	93,8 (-0,27)	90,87 (-0,38)	-0,65	26,4
7	6	888	93,65 (-0,42)	90,72 (-0,53)	-0,95	25,3
8	6	908	93,88 (-0,19)	90,99 (-0,26)	-0,45	23,6
9	6	913	93,95 (-0,12)	90,91 (-0,34)	-0,46	23,1
10	6	920	93,84 (-0,23)	90,83 (-0,42)	-0,65	22,6
11	6	925	93,76 (-0,31)	90,87 (-0,38)	-0,69	22,1
12	6	929	93,8 (-0,27)	90,91 (-0,34)	-0,61	21,8
13	6	931	93,88 (-0,19)	91,18 (-0,07)	-0,26	21,6
14	6	938	94,52 (+0,45)	91,78 (+0,53)	+0,98	21

e modificada, respectivamente, durante a realização da análise interna apenas, obtiveram-se os resultados apresentados na Tabela 5.4 para a otimização da complexidade do sistema de referência contendo 1188 Gaussianas. Comparando-se então os sistemas reduzidos com 951 e 966 Gaussianas, é possível verificar que a diferença entre o fator de desempenho do sistema obtido através da utilização da medida Euclidiana modificada e do sistema obtido através da utilização da medida Euclidiana convencional, é de 0,45. Deve-se notar que mesmo neste caso, é difícil obter uma conclusão a respeito de qual a medida mais apropriada para a análise interna.

A análise interna, quando realizada separadamente da análise discriminativa, fornece sistemas que apresentam relações entre desempenho e economia inferiores às obtidas pela análise conjunta. Dessa forma, pode-se mostrar a importância da complementaridade das análises durante o processo de eliminação de Gaussianas para a obtenção de sistemas reduzidos que apresentem um melhor compromisso entre tamanho e desempenho, estabelecendo como referência de comparação o sistema contendo um número fixo de componentes por estado que apresente o melhor desempenho durante o reconhecimento.

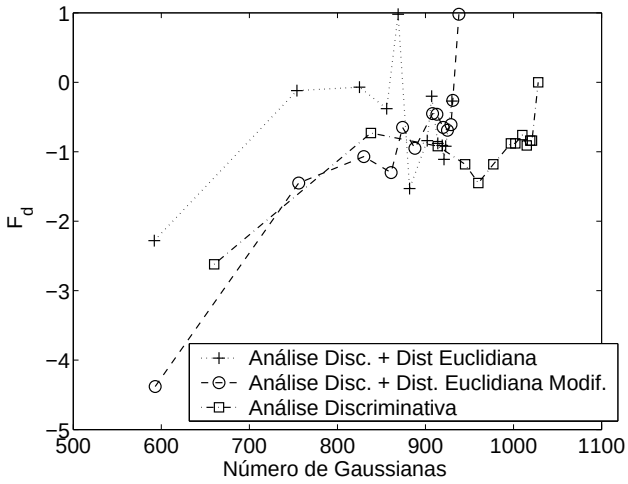


Fig. 5.9: Sistemas obtidos a partir da análise conjunta, utilizando-se a medida Euclidiana convencional e a modificada (na etapa de análise interna), e também através de apenas a análise discriminativa.

Tab. 5.4: Desempenho dos modelos obtidos através da análise interna dos modelos. As comparações foram realizadas com o sistema de referência contendo 11 Gaussianas por estado. Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o original.

Distância Euclidiana Convencional						
L_d	Número de Gauss. no Sist. Original	Número de Gauss. no Sist. Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
3,4	1188	951	94,22 (+0,15)	91,25 (0)	+0,15	19,9
Distância Euclidiana Modificada						
L_d	Número de Gauss. no Sist. Original	Número de Gauss. no Sist. Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
6	1188	966	94,33 (+0,26)	91,59 (+0,34)	+0,6	18,7

5.5 Discussão

Em uma análise inicial, os sistemas obtidos através do algoritmo discriminativo mostram que é possível se obter uma redução no número de Gaussianas presentes no modelo e ao mesmo tempo satisfazer pelo menos a condição $F_d \geq 0$, quando tais sistemas são comparados ao de referência utilizado como ponto de partida para o processo de otimização. Além disso, impondo a comparação entre os sistemas reduzidos e o melhor sistema de referência (1188 Gaussianas), também é possível obter uma economia de parâmetros e ao mesmo tempo satisfazer a mesma condição para o fator de desempenho.

A análise interna também se mostrou eficiente no processo de eliminação de Gaussianas, mesmo

quando realizada separadamente da análise discriminativa. Porém, não se determinou de forma clara qual a medida de distância mais apropriada para a análise interna: distância Euclidiana convencional ou modificada.

Por fim, os melhores resultados foram obtidos para a associação da análise discriminativa com a análise interna, independentemente da medida de distância empregada na análise interna, conforme pode ser observado através das Tabelas 5.4 e 5.5. Tal fato evidencia a complementaridade das análises e ilustra a importância da utilização conjunta de tais algoritmos para a obtenção de sistemas que apresentem um melhor compromisso entre complexidade e desempenho.

Tab. 5.5: Desempenho dos modelos obtidos através das análises discriminativa, interna e conjunta. As comparações foram realizadas com o sistema de referência contendo 1188 Gaussianas. Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o original.

Análise Discriminativa							
K		Número de Gauss. no Sist. Original	Número de Gauss. no Sist. Reduzido	Percent. Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâm. (%)
14		1188	1028	94,14 (+0,07)	91,18 (-0,07)	0	13,5
Análise Interna (Análise Discriminativa + Interna (Distância Euclidiana Convencional)							
K	L_d	Número de Gauss. no Sist. Original	Número de Gauss. no Sist. Reduzido	Percent. Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâm. (%)
6	3,4	1188	869	94,48 (+0,41)	91,82 (+0,57)	+0,98	26,9
Análise Interna (Análise Discriminativa + Interna (Distância Euclidiana Modificada)							
K	L_d	Número de Gauss. no Sist. Original	Número de Gauss. no Sist. Reduzido	Percent. Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâm. (%)
14	6	1188	938	94,52 (+0,45)	91,78 (+0,53)	+0,98	21

Uma estratégia que pode ser adotada na busca da melhor topologia, é partir do melhor sistema de referência e aplicar a análise conjunta para se obter o sistema reduzido na condição menos flexível possível, e para a qual seja possível se observar pelo menos o mesmo desempenho do sistema original ou, em alguns casos, ganho de desempenho. Deve-se notar que, mesmo desconhecendo o melhor sistema de referência, é possível realizar a análise conjunta para a otimização de qualquer sistema contendo um número fixo de Gaussianas por estado, a fim de se obter sistemas reduzidos que apresentem um melhor compromisso entre complexidade e desempenho.

É importante notar que neste ponto os melhores resultados gerados através da análise conjunta são comparáveis, em termos de desempenho, àqueles gerados pelo BIC, onde se obteve por exemplo o sistema contendo 1006 Gaussianas, que apresentou 94,18% e 91,25% de taxa de reconhecimento e *accuracy*, respectivamente, e pelo método baseado na medida de entropia dos estados, onde se

obteve o sistema contendo 996 Gaussianas, que apresentou taxas de 94,64% e 92,35%. Além disso, o novo método discriminativo, juntamente com a análise interna dos estados, superou claramente, em termos de desempenho, os sistemas obtidos pelo método discriminativo para o aumento da resolução acústica. Deve-se destacar que, neste ponto, as comparações entre os métodos são realizadas a partir dos resultados obtidos com a gramática *Word-pairs*.

A utilização da gramática *Word-pairs* durante os experimentos realizados com a proposta inicial do método de eliminação de Gaussianas proposto, baseado na nova análise discriminativa e na análise interna dos estados, pode ter dificultado a avaliação da medida de distância mais apropriada para a análise interna, pelo fato de ser bastante restritiva. Entretanto, justifica-se o emprego de tal gramática na fase inicial dos experimentos, pelo efeito de simplificação do processo de decodificação.

No intuito de se avaliar os modelos acústicos em uma condição mais flexível, e portanto que exija uma maior robustez dos sistemas obtidos a partir do novo método, pode-se então empregar a gramática *Back-off bigram* nos experimentos seguintes.

Em última análise, a medida de WGP utilizada para o cálculo da constante discriminativa do modelo, assim como outros métodos existentes na literatura, é baseada em valores de verossimilhança, que podem estar compreendidos entre 0 e $+\infty$. Além disso, tal medida calcula valores discretos de probabilidade para definir a importância de cada Gaussianas em relação aos modelos. Dessa forma, é possível questionar se uma medida mais suave e contínua para o cálculo da importância de cada Gaussiana poderia ser mais eficaz para a determinação da constante discriminativa de cada modelo e resultar em eliminações mais apropriadas. Neste sentido, uma proposta alternativa para o cálculo da importância de cada Gaussiana será apresentada no próximo capítulo.

5.6 Conclusões

Os resultados iniciais com o novo algoritmo discriminativo baseado na WGP mostram que é possível se encontrar sistemas reduzidos, a partir dos sistemas de referência, que apresentem pelo menos o mesmo desempenho. Entretanto, tal método não é eficiente na eliminação de componentes redundantes presentes nos modelos, as quais não são detectadas pela constante discriminativa.

Assim, a análise interna dos modelos foi proposta visando permitir a eliminação das componentes redundantes, e é baseada em medidas de distância. A distância Euclidiana foi utilizada em uma primeira análise e, na seqüência, foi proposta uma modificação em tal medida no intuito de atribuir indiretamente pesos diferentes para as distâncias calculadas entre Gaussianas localizadas nas fronteiras da distribuição de dados e para as distâncias calculadas entre Gaussianas localizadas na parte central da distribuição. Os resultados não mostraram de forma clara qual a medida de distância mais vantajosa para a análise interna. Entretanto, evidenciaram que a utilização conjunta da análise dis-

criminativa e da análise interna fornece sistemas reduzidos que apresentam um melhor compromisso entre complexidade e desempenho. Portanto, deve-se utilizar a análise conjunta, ao invés da utilização separada da análise discriminativa ou da análise interna.

Além disso, a análise conjunta permitiu a obtenção de sistemas com desempenhos comparáveis aos sistemas gerados pelo BIC e pelo método baseado na entropia dos estados, mas superiores aqueles gerados pelo método discriminativo para o aumento da resolução acústica. Entretanto, algumas questões com relação à WGP empregada no cálculo da constante discriminativa devem ser investigadas, no intuito de se determinar uma medida contínua para o cálculo da importância de cada Gaussiana.

Capítulo 6

O Novo GEA Utilizando uma Nova GIM

6.1 Introdução

Os algoritmos de treinamento discriminativo presentes na literatura, que atuam na fase de estimação de parâmetros do processo de modelagem estatística, utilizam medidas de verossimilhança para o cálculo da capacidade de classificação dos modelos. De forma semelhante, os algoritmos existentes na literatura que atuam na etapa de detecção da topologia dos modelos (Bie03; PB00; GJPP99) também se baseiam em medidas de verossimilhança. A nova WGP, apresentada no Capítulo 5, utiliza informações de verossimilhança para o cálculo da importância de cada Gaussiana em relação aos modelos existentes. Entretanto, pode-se questionar a utilização de uma função discreta de probabilidade como medida de importância, visando aumentar a eficiência do algoritmo de eliminação de Gaussianas (GEA).

Assim, uma nova proposta para a medida de importância das Gaussianas (GIM) será introduzida, utilizando uma função contínua para os cálculos. Basicamente, a idéia é levar em consideração a contribuição de todas as amostras para o cálculo da GIM, ao invés de apenas aquelas amostras para as quais a Gaussiana fornece o maior valor de verossimilhança, que corresponde à medida adotada para o cálculo da WGP.

Outro ponto que será abordado está relacionado com a segmentação acústica utilizada no GEA, pois a eficiência do algoritmo discriminativo está diretamente relacionada com a qualidade da segmentação utilizada. Nos experimentos realizados com a base de dados pequena em Português, devido à ausência de uma segmentação de referência, utilizou-se o sistema de reconhecimento com maior desempenho para a obtenção do alinhamento forçado de Viterbi contra as transcrições corretas, no intuito de se associar cada amostra com um estado dos modelos. Os últimos experimentos realizados com a base de dados TIMIT, que possui uma segmentação de referência, mostraram indícios de que a correlação entre a qualidade da segmentação acústica e o desempenho durante o reconhecimento,

do sistema que a gerou, não é alta. Portanto, do ponto de vista do reconhecimento de fala, a melhor segmentação acústica talvez não seja a mais apropriada para o GEA, pelo contrário, aquela obtida através do alinhamento forçado de Viterbi realizado por um sistema de reconhecimento que apresente um desempenho elevado pode contribuir para uma maior eficácia do método. Neste sentido, a relação entre o desempenho no reconhecimento e na segmentação será investigada.

Por fim, as topologias obtidas para os sistemas com número variado de Gaussianas por estado, a partir do novo GEA, serão analisadas no intuito de se avaliar a complexidade da modelagem de cada classe fonética utilizada nos experimentos.

6.2 Probabilidades Hipervolumétricas para o Cálculo da GIM

No problema de reconhecimento de fala, em que PDFs Gaussianas multidimensionais são utilizadas para modelar diferentes padrões no espaço de características acústicas, os valores de verossimilhança são utilizados durante o processo de decodificação e reconhecimento. A verossimilhança pode assumir valores entre 0 e $+\infty$, de acordo com a proximidade de cada amostra em relação à média e de acordo com a variância da PDF. Assim, para uma determinada PDF, quanto menor a variância e quanto mais próxima da média se encontrar uma determinada amostra, maior será o valor de verossimilhança calculado. A utilização de medidas de verossimilhança para o cálculo da importância das Gaussianas pode permitir a atribuição de um peso maior para Gaussianas com variância pequena, responsável pela cobertura de poucos dados, em detrimento de Gaussianas com variâncias maiores, responsáveis pela cobertura de um número maior de dados. Supondo a existência duas PDFs normais $N\left(5;1\right)$ e $N\left(3;0,3\right)$, como exemplo ilustrativo de cálculos de importância para Gaussianas com variâncias diferentes, utilizadas para modelar 30 amostras, conforme indicadas na Figura 6.1, pode-se verificar que 99% da PDF $N\left(5;1\right)$ contém as 30 amostras, enquanto 99% da PDF $N\left(3;0,3\right)$ contém apenas 13 amostras. No caso de se utilizar a soma das verossimilhanças obtidas a partir de tais amostras como medida de importância das Gaussianas, obtêm-se 6,08 e 10,9 respectivamente. Apesar da Gaussiana $N\left(3;0,3\right)$ ser responsável pela cobertura de uma menor quantidade de dados do que a Gaussiana $N\left(5;1\right)$, apresentará uma importância maior se a soma das verossimilhanças for utilizada como GIM.

Uma alternativa que leva em consideração a proximidade da amostra em relação à média e à cobertura da PDF ao mesmo tempo, é a medida de probabilidade. Assim, se a medida baseada no cálculo de probabilidades hipervolumétricas, que será apresentada na próxima seção, for adotada para este exemplo ao invés da soma das verossimilhanças, a importância de cada Gaussianas $N\left(5;1\right)$ e $N\left(3;0,3\right)$ será de 10,16 e 5,98, respectivamente, o que parece mais razoável.

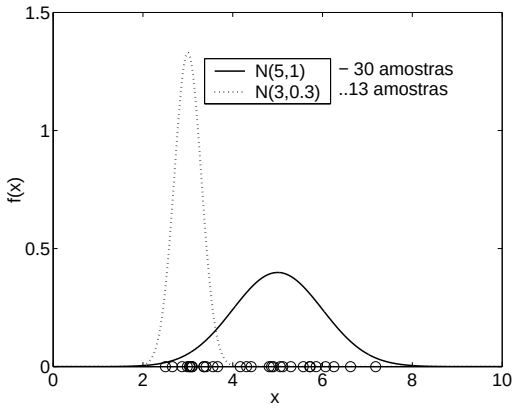


Fig. 6.1: Exemplo de cálculos de importância para Gaussianas com variâncias diferentes.

6.2.1 Cálculo do Hipervolume

Neste trabalho, todas as Gaussianas multidimensionais $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ estão representadas num espaço acústico de dimensão 39, e a função densidade de probabilidade (PDF) é dada pela Equação (6.1)

$$f(\mathbf{O}_t) = \frac{1}{(2\pi)^{\dim/2}\,|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)/2}, \tag{6.1}$$

em que $|\boldsymbol{\Sigma}|$ é o determinante da matriz de covariância. Se os parâmetros forem estatisticamente independentes (matriz de covariância diagonal), então a PDF pode ser escrita na forma da Equação (6.2)

$$f(\mathbf{O}_t) = \prod_{d=1}^{\dim} \frac{1}{\sqrt{2\pi\sigma_d^2}} e^{-\left[(x_d-\mu_d)^2/2\sigma_d^2\right]}. \tag{6.2}$$

Além disso, pode-se definir a contribuição de cada amostra para a GIM, ao longo de cada dimensão acústica, pelas áreas indicadas nas Figuras 6.2(a) e 6.2(b).

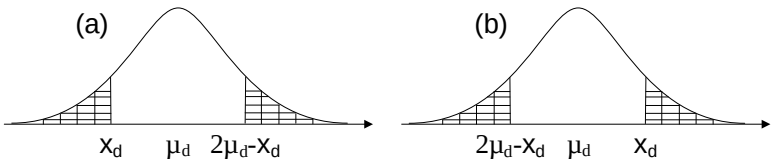


Fig. 6.2: *Contribuição de cada amostra para a GIM. (a) para $x_d \leq \mu_d$. (b) para $x_d > \mu_d$.*

Assim, para cada Gaussiana, a contribuição ao longo de todas as dimensões para o GIM é calculada pela Equação (6.3)

$$GIM\left(\mathbf{O}_t\right)^{(i;j;s)} = \prod_{d=1}^{\dim}\left(1 - \left[\frac{2}{\sqrt{\pi}}\int_0^{\|Z_d\|} e^{-(z_d)^2} dz_d\right]\right), \tag{6.3}$$

em que "dim" é a dimensão do vetor de características, $z_d = \frac{x_d - \mu_{dij}}{\sqrt{2}\sigma_{dij}}$, e $\mathbf{O}_t = (x_1, x_2, \ldots, x_{\dim})$ é o vetor de características. Os valores μ_{dij} e σ_{dij} correspondem à média e ao desvio padrão respectivamente, ao longo da dimensão "*d*", da Gaussiana "*i*" que pertence ao estado "*j*".

O GIM da Gaussiana "*i*", que pertence ao estado "*j*", é calculado a partir de cada amostra do estado "*s*", de tal forma que o valor médio do GIM pode ser obtido em relação a cada estado, conforme definido na Equação (6.4) abaixo

$$P_{GIM}^{(i;j;s)} = \frac{\sum_{t=1}^{N_s} GIM\left(\mathbf{O}_t\right)^{(i;j;s)}}{N_s}, \tag{6.4}$$

onde N_s é o número de amostras do estado "*s*".

No intuito de se calcular o valor GIM, é necessário que se tenha uma base de dados segmentada, uma vez que a Equação (6.3) requer que as amostras tenham sido previamente rotuladas. Conforme mencionado anteriormente, a segmentação pode ser obtida, por exemplo, pelo alinhamento de Viterbi realizado a partir das transcrições corretas de cada sentença, utilizando-se o melhor sistema HMM disponível.

A nova medida proposta (GIM) se baseia portanto na probabilidade das amostras se encontrarem fora do intervalo

$$\mu_d - \|x_d - \mu_d\| < x < \mu_d + \|x_d - \mu_d\|.$$

Pode-se notar que, quanto mais próxima a amostra se encontra da média da Gaussiana, maior é a contribuição para o GIM ao longo da dimensão analisada.

O $P_{GIM}^{(i;j;s)}$ pode ser então utilizado como medida de importância de cada Gaussiana em relação a cada estado. Assim, é possível implementar um método discriminativo de seleção de Gaussianas baseado em tal medida, em que o principal objetivo é maximizar a relação discriminativa, de tal forma que cada modelo obtido após a análise apresente o máximo $P_{GIM}^{(i;j;s)}$ para os padrões correspondentes ao estado "j" (s = j) e o mínimo $P_{GIM}^{(i;j;s)}$ para os demais padrões (s $\neq$ j) ao mesmo tempo. A relação discriminativa que deve ser maximizada é dada pela Equação (6.5)

$$DC^{(j)} = \frac{\left[\sum_{i=1}^{M_j} P_{GIM}^{(i;j;j)}\right]^K}{\left[\sum_{s\neq j}^{N}\sum_{i=1}^{M_j} P_{GIM}^{(i;j;s)}\right] \Big/ N-1} \tag{6.5}$$

onde K é o expoente de rigor, M_j é o número de Gaussianas do estado "*j*" e N é o número total de estados. Se o logaritmo da Constante Discriminativa (DC) for calculado, a expressão resultante é dada pela Equação (6.6), que é similar à Equação (3.42), no sentido que a primeira parcela mede a capacidade de modelar os padrões corretos e a segunda parcela é um termo de penalização.

$$\log DC^{(j)} = K \log \sum_{i=1}^{M_j} P_{GIM}^{(i;j;j)} - \log \frac{\sum_{s \neq j}^{N} \sum_{i=1}^{M_j} P_{GIM}^{(i;j;s)}}{N-1}. \tag{6.6}$$

No entanto, as expressões diferem no sentido que o termo de penalização da Equação (3.42) somente considera aspectos inerentes ao modelo analisado, enquanto o termo de penalização na Equação (6.6) leva em consideração aspectos dos modelos de todos os estados presentes no sistema.

A principal idéia do método é a de eliminar Gaussianas de um sistema previamente treinado com um número fixo de componentes por estado e observar então o novo valor DC obtido. O valor da Constante Discriminativa (DC) pode aumentar ou diminuir dependendo da relevância da Gaussiana eliminada. Dessa forma, o expoente de rigor tem uma função importante na seleção de Gaussianas, uma vez que torna o critério discriminativo mais restritivo: quanto maior o valor do expoente de rigor, mais rigoroso se torna o critério e portanto menos Gaussianas são eliminadas.

O procedimento descrito acima é aplicado para cada estado dos modelos HMM. Uma vez concluído o processo de eliminação discriminativa, os modelos resultantes são treinados novamente pelo algoritmo de Baum-Welch, porém agora em uma condição bem menos flexível (menos parâmetros nos modelos).

É importante destacar novamente que o algoritmo discriminativo detecta apenas Gaussianas que pertencem a um dado estado e no entanto interferem na modelagem de dados associados a outros estados. Portanto, ainda podem existir Gaussianas excedentes no modelo após a aplicação do algoritmo discriminativo. Apesar de não serem detectadas pelo critério discriminativo, tais componentes precisam ser descartadas, uma vez que este procedimento pode ser realizado sem degradação da capacidade de classificação do sistema. Neste sentido, deve-se utilizar na seqüência a análise interna, baseada na Equação (5.3), a fim de se eliminar o excesso de Gaussianas que ainda pode existir nos modelos. A Figura 6.3 ilustra o algoritmo proposto.

Assim, os primeiros resultados obtidos através do GEA com a utilização da nova GIM (YVS05), para a gramática *Word-pairs* e com a base em Português, encontram-se indicados na Figura 6.4. Então, conforme pode ser observado, obtiveram-se sistemas reduzidos que superaram o desempenho do melhor sistema de referência contendo 1188 Gaussianas. Entretanto, tais resultados ainda não mostram de forma clara se a nova medida para quantificar a GIM é mais apropriada do que a WGP apresentada anteriormente.

A utilização da gramática *Word-pairs* pode ser, possivelmente, a responsável pela dificuldade

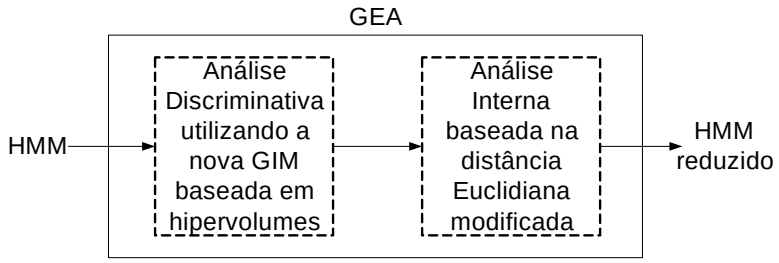


Fig. 6.3: Algoritmo de Eliminação de Gaussianas (GEA)

para a determinação de forma clara da medida mais apropriada para a GIM, visto que o processo de decodificação se torna bastante restritivo neste caso e, portanto, a exigência de uma maior robustez dos modelos acústicos se torna atenuada. Resumidamente, para se obter um mesmo desempenho, utilizando-se gramáticas diferentes, é necessário que os modelos acústicos empregados juntamente com a gramática menos restritiva possuam uma maior robustez do que os modelos acústicos empregados com a gramática mais restritiva, pois as limitações impostas pela gramática mais restritiva podem compensar a utilização de modelos menos robustos.

Outro fator que contribui diretamente para a eficácia do GEA e que também deve ser analisado é a segmentação utilizada na fase de análise discriminativa dos modelos. Devido à ausência de uma segmentação manual da base de dados pequena em Português, os dados foram rotulados automaticamente através do alinhamento forçado de Viterbi, realizado por um sistema de reconhecimento de fala. Assim, na seqüência será avaliado o efeito de diferentes segmentações, obtidas a partir de sistemas de reconhecimento de fala com desempenhos diferentes, sobre o GEA.

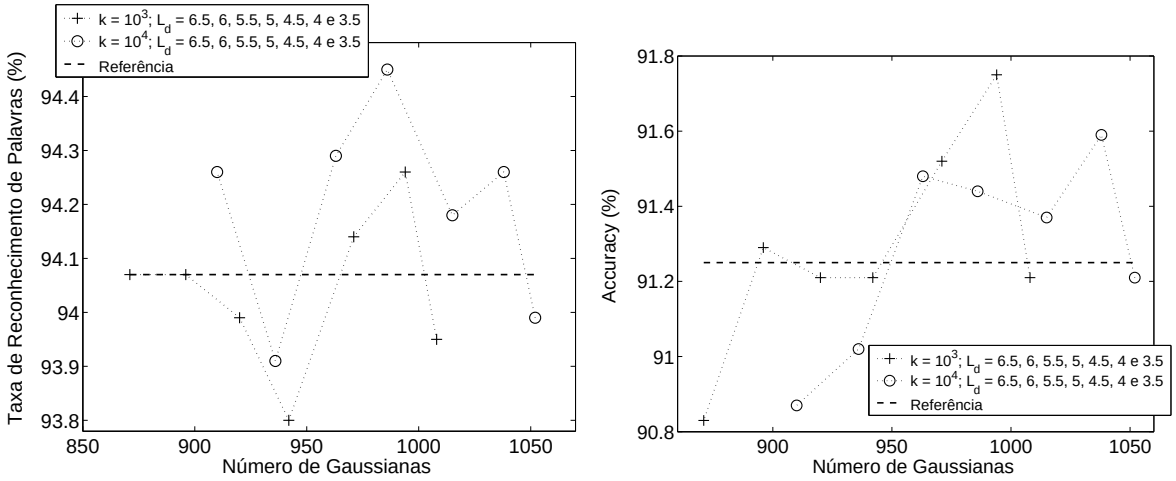


Fig. 6.4: Sistemas obtidos através do GEA utilizando a nova GIM.

6.3 Avaliação do GEA para Diferentes Segmentações Acústicas

A utilização de diferentes segmentações acústicas tem por finalidade avaliar a estratégia da escolha da segmentação mais apropriada para o GEA, baseada exclusivamente no desempenho dos sistemas durante o reconhecimento. Assim, utilizaram-se três sistemas de reconhecimento de fala (Sistema I, Sistema II e Sistema III), com número variado de Gaussianas por estado, que possuem as seguintes taxas (reconhecimento de palavras; accuracy): $(94,64\%;92,35\%)$, $(94,71\%;92,2\%)$ e $(94,79\%;91,71\%)$, para a obtenção de três novas segmentações (Segment I, Segment II e Segment III).

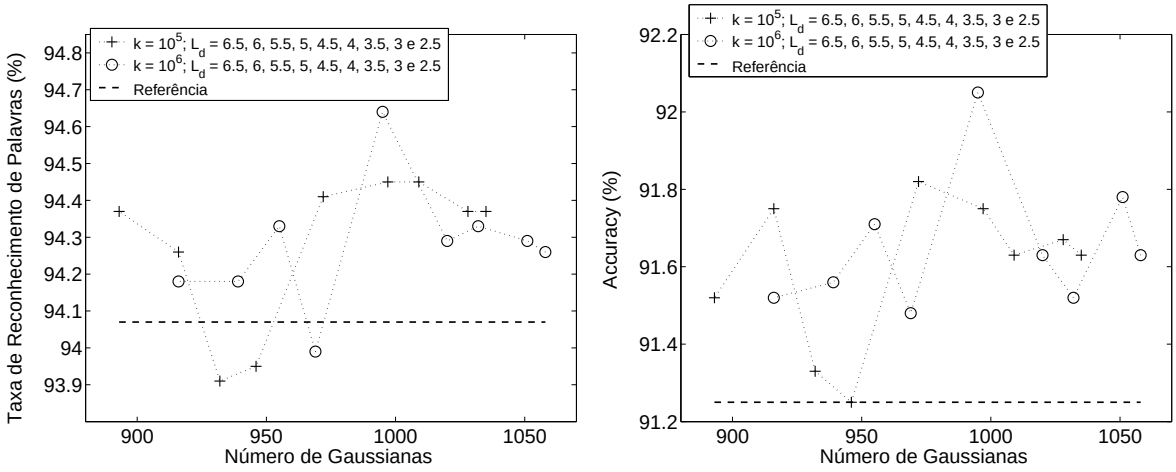


Fig. 6.5: Resultados obtidos através do GEA, a partir da segmentação do Sistema I.

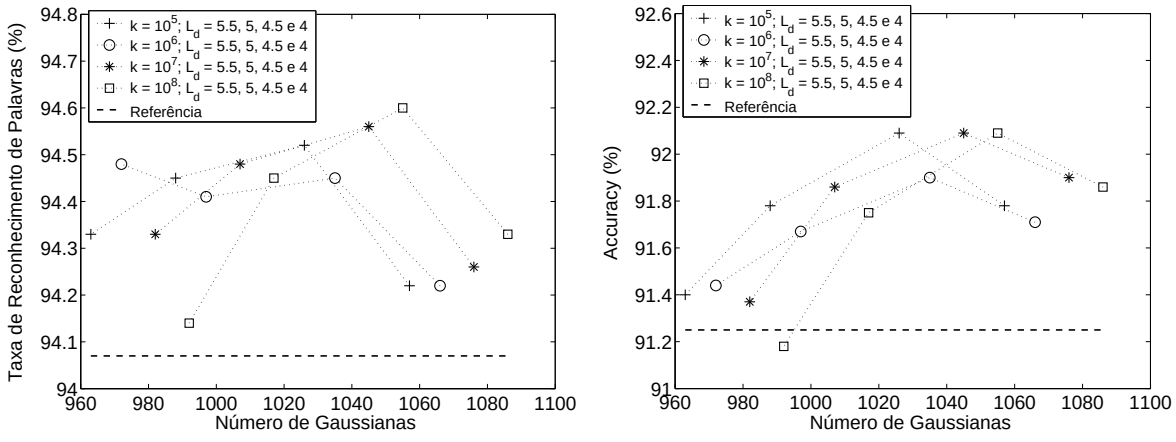


Fig. 6.6: Resultados obtidos através do GEA, a partir da segmentação do Sistema II.

Deve-se destacar que todos os experimentos anteriores utilizam uma segmentação (Segment 0) obtida da mesma forma, por um sistema de reconhecimento que fornece taxas de $(94,52\%;91,78\%)$.

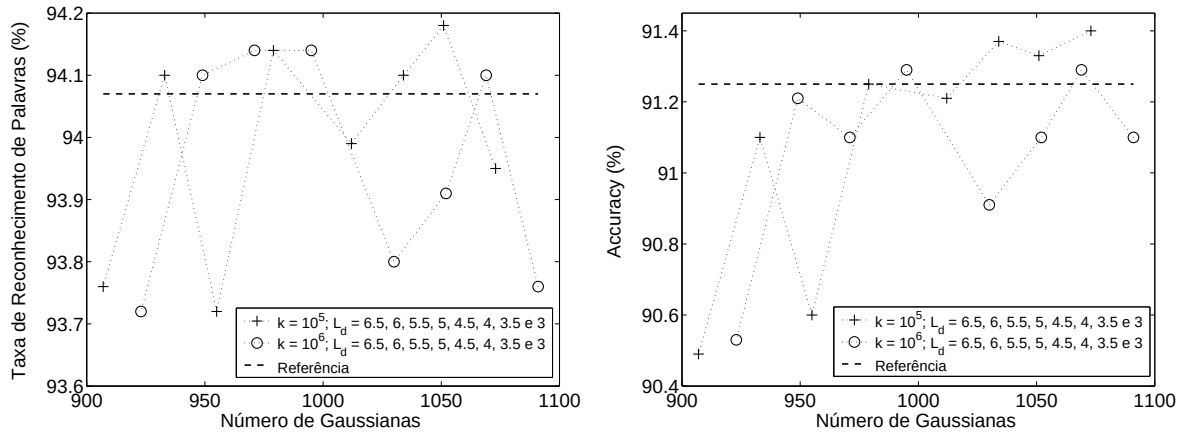


Fig. 6.7: Resultados obtidos através do GEA, a partir da segmentação do Sistema III.

As Figuras 6.5, 6.6 e 6.7 mostram os resultados obtidos em cada caso.

Então, é importante avaliar qual é a segmentação que permite a obtenção das melhores topologias, de acordo com o principal objetivo desejado. Neste sentido, a Tabela 6.1 apresenta as topologias mais apropriadas quando o enfoque da busca se concentra na economia, ou seja, na determinação de topologias que apresentem a maior economia possível, desde que a condição $F_d \geq 0$ seja satisfeita.

Tab. 6.1: Desempenho dos modelos mais econômicos (para $F_d \geq 0$) obtidos através do GEA. As comparações foram realizadas com o melhor sistema de referência contendo 1188 Gaussianas (11 por estado). Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência.

Segmentação	K	L_d	Número de Gauss. no Sist. Reduzido	Percent. Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâm. (%)
Segment 0	10^3	6	896	94,07 (0)	91,29 (+0,04)	+0,04	24,6
Segment I	10^5	6,5	893	94,37 (+0,3)	91,52 (+0,27)	+0,57	24,8
Segment II	10^5	5,5	963	94,33 (+0,26)	91,4 (+0,15)	+0,41	18,9
Segment III	10^5	5	979	94,14 (+0,07)	91,25 (0)	+0,07	17,6

Os sistemas reduzidos podem ser selecionados também visando a obtenção do maior desempenho possível no reconhecimento. Neste sentido, a Tabela 6.2 apresenta os sistemas que fornecem os maiores fatores de desempenho.

O sistema que apresentou a maior economia (24,8%) foi obtido pelo GEA a partir da segmentação Segment I. De forma semelhante, o maior fator desempenho (1,37) foi obtido a partir das segmentações Segment I e Segment II. Portanto, em uma análise geral, a segmentação Segment I foi a que forneceu os sistemas com o melhor compromisso entre desempenho e complexidade. Deve-se notar que a melhor segmentação para o GEA, neste caso, foi gerada pelo sistema de reconhecimento que

Tab. 6.2: Desempenho dos modelos com os melhores desempenhos no reconhecimento, obtidos através do GEA. As comparações foram realizadas com o melhor sistema de referência contendo 1188 Gaussianas (11 por estado). Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência.

Segmentação	K	L_d	Número de Gauss. no Sist. Reduzido	Percent. Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâm. (%)
Segment 0	10^3	4	994	94,26 (+0,19)	91,75 (+0,5)	+0,69	16,3
Segment I	10^6	4,5	995	94,64 (+0,57)	92,05 (+0,8)	+1,37	16,2
Segment II	10^8	4,5	1055	94,6 (+0,53)	92,09 (+0,84)	+1,37	11,2
Segment III	10^5	3,5	1051	94,18 (+0,11)	91,33 (0,08)	+0,19	11,5

apresentou o maior *accuracy*, mas não apresentou a maior taxa de reconhecimento de palavras. Além disso, também é claro que a segmentação Segment III forneceu resultados inferiores aos obtidos com as outras segmentações testadas, e foi gerada a partir do sistema de reconhecimento de fala que forneceu o menor *accuracy* e a maior taxa de reconhecimento. Dessa forma, há uma indicação de que a segmentação mais apropriada para o GEA pode ser gerada a partir do sistema de reconhecimento de fala que apresentar o maior valor de *accuracy*. Entretanto, ainda são necessários mais testes a fim de se obter uma conclusão mais precisa sobre a relação entre a segmentação mais apropriada para o GEA e as medidas de desempenho dos sistemas de reconhecimento utilizados para a geração de tais segmentações.

Os experimentos realizados com o GEA, a partir deste ponto, passaram a utilizar a segmentação Segment I na fase de análise discriminativa, ao invés de utilizar a segmentação Segment 0, que foi empregada até o momento.

No intuito de se relacionar a qualidade das segmentações geradas pelos sistemas de reconhecimento de fala, do ponto de vista acústico, com as medidas de desempenho de tais sistemas, alguns testes serão realizados com a base de dados TIMIT, que possui uma segmentação de referência (manual).

É importante notar que, neste ponto, os melhores resultados obtidos com a segmentação Segment I são comparáveis aos melhores resultados obtidos através do método de seleção de topologia baseado na entropia dos estados. Além disso, mesmo analisando apenas os resultados das decodificações realizadas com o emprego da gramática *Word-pairs*, os sistemas obtidos pelo GEA apresentam claramente um melhor compromisso entre complexidade e desempenho, em relação aos resultados obtidos através do BIC e do método discriminativo para o aumento da resolução acústica.

Por fim, utilizando também a base em Português, empregou-se a gramática *Back-off bigram* na decodificação a fim de se avaliar os sistemas reduzidos em uma condição mais flexível e, portanto, que exige uma maior robustez dos modelos acústicos. Assim, a Tabela 6.3 mostra os resultados ob-

tidos pelo GEA com a utilização da gramática *Back-off bigram*. Deve-se lembrar que a estratégia adotada no intuito de diminuir a busca pelas topologias mais apropriadas é a de otimizar a complexidade do melhor sistema contendo um número fixo de componentes por estado e, dessa forma, o ponto de partida para os testes realizados com tal gramática é o sistema de referência contendo 1296 Gaussianas.

Tab. 6.3: Desempenho dos modelos obtidos através do GEA, utilizando a gramática *Back-off bigram* na decodificação. Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência contendo 1296 Gaussianas (12 por estado).

L_d	K	Número de Gaussianas no Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
4	10^5	1068	81,32 (+0,31)	61,31 (+3,06)	+3,37	17,6
4,5	10^5	1038	81,85 (+0,84)	61,65 (+3,4)	+4,24	19,9
5	10^5	1009	81,62 (+0,61)	61,31 (+3,06)	+3,67	22,2
5,5	10^5	985	80,82 (-0,19)	58,59 (+0,34)	+0,15	24
4	10^6	1085	81,77 (+0,76)	62,03 (+3,78)	+4,54	16,3
4,5	10^6	1055	82,34 (+1,33)	62,37 (+4,12)	+5,45	18,6
5	10^6	1026	81,69 (+0,68)	60,93 (+2,68)	+3,36	20,8
5,5	10^6	1002	80,9 (-0,11)	59,49 (+1,24)	+1,13	22,7
4	10^7	1118	81,35 (+0,34)	61,12 (+2,87)	+3,21	13,7
4,5	10^7	1088	81,69 (+0,68)	60,85 (+2,6)	+3,28	16,1
5	10^7	1059	81,05 (+0,04)	59,34 (+1,09)	+1,13	18,3
5,5	10^7	1035	80,37 (-0,64)	57,68 (-0,57)	-1,21	20,1

Os experimentos com a gramática *Back-off bigram* mostraram que os sistemas obtidos através do GEA apresentaram os maiores valores de ganho de desempenho e economia de parâmetros, em relação aos obtidos anteriormente pelo BIC, método baseado na entropia e análise discriminativa para o aumento da resolução acústica dos modelos.

A utilização da segmentação Segment I certamente contribuiu para uma maior eficácia da análise discriminativa e no cálculo das distâncias Euclidianas modificadas. Entretanto, deve-se notar que ainda são necessários novos experimentos a fim de se avaliar qual a medida de importância mais apropriada para o algoritmo discriminativo (WGP ou GIM baseada em medidas de hipervolume), assim como para a determinação da medida de distância mais apropriada (distância Euclidiana convencional ou modificada) para a análise interna. Por outro lado, é evidente a eficácia da estratégia de eliminação de Gaussianas visando a obtenção de sistemas que apresentem um melhor compromisso entre complexidade e desempenho.

6.4 Experimentos Realizados com a Base de Dados TIMIT

A base de dados TIMIT tem sido bastante explorada na literatura para a avaliação de sistemas de reconhecimento de fones contínuos (LH89a; LG93). Além disso, todas as sentenças de treinamento e de teste possuem uma segmentação de referência, o que pode permitir a utilização de tais informações durante o processo de obtenção dos modelos acústicos e também a avaliação dos alinhamentos forçados de Viterbi, gerados por sistemas de reconhecimento.

O passo inicial que precede os experimentos com o novo GEA, utilizando-se a TIMIT como base de dados, consiste na obtenção de sistemas de referência, os quais podem ser utilizados como ponto de partida para o novo método proposto. Neste sentido, foram gerados três sistemas de referência, contendo 1152, 2304 e 4608 Gaussianas, correspondendo a 8, 16 e 32 componentes por estado, respectivamente. A Tabela 6.4 apresenta os desempenhos de tais sistemas de referência, durante o reconhecimento de fones contínuos.

Tab. 6.4: Sistemas de referência com um número fixo de componentes por estado (8, 16 e 32 Gaussianas por estado).

Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)
1152	70,94	63,7
2304	73,03	65,87
4608	73,11	66,14

Na seqüência pode-se aplicar o GEA visando a obtenção de sistemas que apresentem um melhor compromisso entre tamanho e desempenho. Assim , as Tabelas 6.5, 6.6 e 6.7 apresentam os resultados gerados pelo GEA, utilizando como pontos de partida os sistemas contendo 1152, 2304 e 4608 Gaussianas.

É possível observar que o GEA não forneceu sistemas com economia de parâmetros ou ganho de desempenho considerável, partindo-se do sistema de referência contendo 1152 Gaussianas. Tal fato pode ser explicado pela quantidade de unidades acústicas utilizadas nos experimentos e pela complexidade intrínseca do problema que, neste caso, possui conjuntos diferentes de sentenças de treinamento e de teste. Além disso, a TIMIT possui um conjunto de dados 3 vezes maior do que a quantidade contida na base de dados pequena em Português e, portanto, permite que a resolução acústica dos modelos não seja limitada pela quantidade de informação disponível, mas principalmente pela complexidade das distribuições dos parâmetros. Dessa forma, um sistema contendo 8 Gaussianas por estado no problema de reconhecimento de fala com a TIMIT é comparativamente menor do que um sistema contendo 8 Gaussianas por estado no problema de reconhecimento de fala com a base de dados pequena em Português.

Os resultados fornecidos pelo GEA, partindo-se do sistema de referência com 2304 Gaussianas, apesar de não terem apresentado sistemas reduzidos com fator de desempenho $F_d \geq 0$, forneceram uma economia de parâmetros maior do que a observada nos experimentos com o sistema de referência contendo 1152 Gaussianas. Outro ponto que deve ser notado está no fato de que os resultados obtidos para um limiar de distância $L_d \neq 0$ apresentaram uma grande perda de desempenho e, portanto, foram desconsiderados. Dessa forma, há uma indicação de que o sistema de referência com 2304 Gaussianas ainda deve ser menor do que o melhor sistema de referência, pois a análise interna não detectou excesso de Gaussianas nos modelos.

Tab. 6.5: Desempenho dos modelos obtidos através do GEA, utilizando a base de dados TIMIT. Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência contendo 1152 Gaussianas (8 por estado).

L_d	K	Número de Gaussianas no Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
0	10	1080	70,56 (-0,38)	63,23 (-0,47)	-0,85	6,25
5	10	1078	71,15 (+0,21)	63,58 (-0,12)	+0,09	6,42
0	20	1095	71,21 (+0,27)	63,66 (-0,04)	+0,23	4,95
5	20	1093	71,14 (+0,2)	63,57 (-0,13)	+0,07	5,12
0	50	1116	71,29 (+0,35)	63,78 (+0,08)	+0,43	3,13
5	50	1114	71,34 (+0,4)	63,72 (+0,02)	+0,42	3,3

Tab. 6.6: Desempenho dos modelos obtidos através do GEA, utilizando a base de dados TIMIT. Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência contendo 2304 Gaussianas (16 por estado).

L_d	K	Número de Gaussianas no Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
0	5	2056	72,68 (-0,35)	65,91 (+0,04)	-0,31	10,8
0	6	2068	72,68 (-0,35)	65,98 (+0,11)	-0,24	10,2
0	7	2083	72,77 (-0,26)	66,04 (+0,17)	-0,09	9,59
0	8	2090	72,75 (-0,28)	66,08 (+0,21)	-0,07	9,29
0	9	2097	72,69 (-0,34)	65,99 (+0,12)	-0,22	8,98
0	10	2106	72,61 (-0,42)	65,93 (+0,06)	-0,36	8,59

Por fim, o GEA forneceu as maiores economias e ganhos de desempenho para os sistemas obtidos a partir do sistema de referência com 4608 Gaussianas. Porém, é intuitivo esperar que o número de componentes excedentes ou que sejam responsáveis por erros de classificação aumente, à medida que o tamanho do sistema cresce e, dessa forma, os resultados gerados através de qualquer método para a determinação da topologia dos modelos tendem a ser mais expressivos.

Tab. 6.7: Desempenho dos modelos obtidos através do GEA, utilizando a base de dados TIMIT. Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência contendo 4608 Gaussianas (32 por estado).

L_d	K	Número de Gaussianas no Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
0	10	4037	73,49 (+0,38)	66,49 (+0,35)	+0,73	12,4
14	10	3825	73,73 (+0,62)	66,51 (+0,37)	+0,99	17
0	100	4384	73,55 (+0,44)	66,51 (+0,37)	+0,81	4,86
10	100	4357	73,62 (+0,51)	66,77 (+0,63)	+1,14	5,45
12	100	4281	73,87 (+0,76)	66,71 (+0,57)	+1,33	7,1
17	100	3924	73,35 (+0,24)	66,36 (+0,22)	+0,46	14,8

Deve-se destacar que nos experimentos realizados com a TIMIT, diferentemente dos realizados com a base pequena em Português, não se determinou o melhor sistema contendo um número fixo de componentes por estado. A estratégia adotada foi a de otimizar três sistemas de referência, sendo que o maior sistema possui quatro vezes mais parâmetros do menor sistema, a fim de se avaliar a eficiência do novo método.

Os resultados presentes na literatura para sistemas com um número fixo de Gaussianas por estado, que empregam fones independentes de contexto para o problema de reconhecimento de fones contínuos usando a TIMIT, diferem dos obtidos neste trabalho, devido à utilização de outras configurações experimentais, como por exemplo os resultados indicados na Tabela 6.8, onde o autor (Val95) não utilizou saltos entre os estados do HMM durante a modelagem das unidades acústicas, dentre outros pontos.

Tab. 6.8: Sistemas de referência com um número fixo de componentes por estado (8 e 16 Gaussianas por estado), extraídos de (Val95).

Número de Gaussianas	Percentagem Correta (%)	Reco. Accur. (%)
1152	70,55	64,38
2304	72,52	67,22

6.4.1 Uma Medida Simplificada para a GIM

A nova medida para a GIM baseada no cálculo de hipervolumes não mostrou claramente se é mais apropriada do que a medida de WGP, proposta inicialmente. Porém, deve-se destacar que as comparações foram realizadas a partir dos resultados obtidos com a gramática *Word-pairs*, que simplifica

consideravelmente o processo de decodificação. Além disso, ambas foram propostas no intuito de se evitar algumas problemas intrínsecos da medida de verossimilhança, do ponto de vista da atribuição de importância para as Gaussianas dos modelos, como por exemplo a faixa de valores de verossimilhança calculados para PDFs com diferentes variâncias.

No intuito de se obter uma medida mais próxima da verossimilhança e ao mesmo tempo que seja mais simples do que a GIM baseada em cálculos de hipervolumes e a WGP, definiu-se a contribuição de cada amostra para a medida de importância da Gaussiana como a distância ponderada das amostras em relação a média da Gaussiana, de acordo com a Equação (6.7)

$$\text{Contribuição} = \sum_{i=1}^{dim} \frac{1}{|D_i|/\sigma_i}, \tag{6.7}$$

em que "$|D_i|$" é o módulo da diferença entre o valor da amostra e o valor da média da Gaussiana ao longo da dimensão "i" do espaço de parâmetros acústicos, e "dim" é o tamanho do vetor de parâmetros (no caso é 39).

Assim, utilizando-se a Equação (6.7) ao invés da Equação (6.3), e aplicando-se o GEA para otimizar os sistemas de referência contendo 1152, 2304 e 4608 Gaussianas, obtiveram-se os resultados indicados nas Tabelas 6.9, 6.10 e 6.11, para a TIMIT.

Tab. 6.9: Desempenho dos modelos obtidos através do GEA, utilizando a GIM baseada em medidas de distância ponderada. Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência contendo 1152 Gaussianas (8 por estado).

L_d	K	Número de Gaussianas no Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
0	10^5	1105	70,64 (-0,3)	63,32 (-0,38)	-0,68	4,08
0	2.10^5	1117	71,32 (+0,38)	63,83 (+0,13)	+0,51	3,04
5	2.10^5	1115	71,38 (+0,44)	63,78 (+0,08)	+0,52	3,21
10	2.10^5	1101	71,34 (+0,4)	63,8 (+0,1)	+0,5	4,43
12	2.10^5	1086	71,26 (+0,32)	63,54 (-0,16)	+0,16	5,73
0	3.10^5	1121	71,33 (+0,39)	63,84 (+0,14)	+0,53	2,69
0	5.10^5	1128	71,31 (+0,37)	63,8 (+0,1)	+0,47	2,08
0	8.10^5	1131	71,33 (+0,39)	63,81 (+0,11)	+0,5	1,82

Pode-se notar que, em geral, os resultados obtidos anteriormente através dos cálculos de hipervolume como medida da contribuição de cada amostra para a GIM forneceram uma maior economia de parâmetros do que a observada nos resultados obtidos através dos cálculos de distância ponderada como medida da contribuição de cada amostra para a GIM. Do ponto de vista do cálculo da contribuição de cada amostra para a GIM, a utilização da distância ponderada também apresenta o problema

Tab. 6.10: Desempenho dos modelos obtidos através do GEA, utilizando a GIM baseada em medidas de distância ponderada. Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência contendo 2304 Gaussianas (16 por estado).

L_d	K	Número de Gaussianas no Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
0	10^3	1908	71,38 (-1,65)	64,13 (-1,74)	-3,39	17,2
0	10^4	2101	70,98 (-2,05)	63,66 (-2,21)	-4,26	8,81
0	10^5	2197	71,29 (-1,74)	63,73 (-2,14)	-3,88	4,64
0	10^6	2257	72,67 (-0,36)	65,14 (-0,73)	-1,09	2,04
0	10^7	2281	73,02 (-0,01)	65,91 (+0,04)	+0,03	1
5	10^7	2279	72,65 (-0,38)	65,9 (+0,03)	-0,35	1,09
10	10^7	2265	72,06 (-0,97)	65,11 (-0,76)	-1,73	1,69
0	10^8	2292	71,82 (-1,21)	64,57 (-1,3)	-2,51	0,52

Tab. 6.11: Desempenho dos modelos obtidos através do GEA, utilizando a GIM baseada em medidas de distância ponderada. Os valores entre parênteses indicam a diferença de desempenho entre os sistemas reduzidos e o de referência contendo 4608 Gaussianas (32 por estado).

L_d	K	Número de Gaussianas no Sistema Reduzido	Percentagem Correta (%)	Reco. Accur. (%)	F_d	Economia de Parâmetros (%)
0	10^4	4212	73,02 (-0,09)	65,95 (-0,19)	-0,28	8,59
0	10^5	4470	73,67 (+0,56)	66,67 (+0,53)	+1,09	2,99
16	10^5	4159	73,49 (+0,38)	66,43 (+0,29)	+0,67	9,74
17	10^5	4064	73,53 (+0,42)	65,80 (-0,34)	+0,08	11,8
18	6.10^5	4069	73,58 (+0,47)	66,51 (+0,37)	+0,84	11,7
19	6.10^5	3978	73,64 (+0,53)	66,11 (-0,03)	+0,5	13,7
0	10^6	4578	73,75 (+0,64)	66,88 (+0,74)	+1,38	0,65
0	10^7	4600	73,71 (+0,6)	66,85 (+0,71)	+1,31	0,17

da larga faixa de valores que podem ser assumidos por tal medida, de acordo com o desvio padrão utilizado, o que também ocorre com a medida de verossimilhança.

Portanto, apesar da utilização da distância ponderada para o cálculo da contribuição de cada amostra para a GIM também permitir a obtenção de sistemas reduzidos e com ganho de desempenho, a eficácia do GEA é maior com a utilização da medida de hipervolume para o cálculo da contribuição de cada amostra para a GIM.

6.5 Análise da Complexidade dos HMMs para cada Classe de Fonemas

A utilização de métodos para a determinação da complexidade dos HMMs tem como principal finalidade evitar a ocorrência de sobre-parametrização durante o processo de treinamento, aumentando dessa forma a robustez do sistema, e adequando os tamanhos dos modelos à quantidade de dados disponíveis para o processo de treinamento. Tem-se também como consequência a obtenção de sistemas menores do que os de referência, contendo um número fixo de componentes por estado, e que apresentem pelo menos o mesmo desempenho. Outro ponto que pode ser explorado através dos resultados obtidos pelos métodos de determinação de topologias, é a análise da complexidade dos modelos utilizados para representar cada unidade acústica ou conjuntos de fonemas agrupados segundo as características acústicas. Neste sentido, para cada um dos métodos apresentados neste trabalho (GEA, método baseado na entropia dos estados, BIC e método discriminativo para o aumento da resolução acústica) determinou-se o sistema com o maior desempenho, analisando-se os resultados obtidos com a utilização da gramática *Back-off bigram*. Assim, o número de Gaussianas por estado para cada grupo de fonemas encontram-se indicados nas Tabelas 6.12-6.15.

Tab. 6.12: Número de componentes por estado para o sistema obtido através do GEA, que apresenta o desempenho de 82,34% e 62,37% em termos de taxa de reconhecimento de palavras e *accuracy*, respectivamente. Os modelos acústicos correspondem aos fones adotados na base em Português

Fonemas	Estado 1	Estado 2	Estado 3	Média
Vogais/Semi-Vogais	$10,69 \pm 0,05$	$8,9 \pm 0,1$	$10,2 \pm 0,1$	$10,0 \pm 0,4$
Laterais	9 ± 3	$9,0 \pm 0,7$	$10,5 \pm 0,3$	$9,3 \pm 0,5$
Não-laterais	$11,0 \pm 0,3$	$9,7 \pm 0,1$	$10,67 \pm 0,09$	$10,4 \pm 0,2$
Nasais	$8,7 \pm 0,4$	$8,3 \pm 0,3$	$10,0 \pm 0,3$	$9,0 \pm 0,3$
Oclusivas	$11,0 \pm 0,1$	$7,3 \pm 0,3$	$9,5 \pm 0,2$	$9,3 \pm 0,5$
Fricativas	$10,2 \pm 0,1$	$7,7 \pm 0,4$	$11,2 \pm 0,1$	$9,7 \pm 0,4$
Africadas	$9,5 \pm 0,3$	$10,0 \pm 0,6$	$11,5 \pm 0,3$	$10,3 \pm 0,2$
Silêncio	10	11	12	11
Média	$10,31 \pm 0,05$	$8,58 \pm 0,08$	$10,42 \pm 0,05$	

Assim, é possível notar que:

- Nas vogais, a ordem decrescente dos estados em termos do número de componentes utilizadas na modelagem, para qualquer um dos quatro métodos apresentados, foi Estado 1 > Estado 3 > Estado 2.
- Nas laterais, o Estado 3 apresentou um maior número de componentes, em relação aos Estados 1 e 2, para qualquer um dos quatro métodos apresentados.

Tab. 6.13: Número de componentes por estado para o sistema obtido através do BIC, que apresenta o desempenho de 80,75% e 57,6% em termos de taxa de reconhecimento de palavras e *accuracy*, respectivamente. Os modelos acústicos correspondem aos fones adotados na base em Português

Fonemas	Estado 1	Estado 2	Estado 3	Média
Vogais/Semi-Vogais	$11,23 \pm 0,05$	$9,5 \pm 0,1$	$11,08 \pm 0,07$	$10,6 \pm 0,3$
Laterais	9 ± 1	10 ± 1	$10,5 \pm 0,9$	$9,8 \pm 0,4$
Não-laterais	$11,0 \pm 0,3$	$10,3 \pm 0,3$	$11,67 \pm 0,08$	$11,0 \pm 0,2$
Nasais	$11,67 \pm 0,08$	$9,0 \pm 0,6$	8 ± 1	$9,8 \pm 0,6$
Oclusivas	$11,0 \pm 0,1$	$10,0 \pm 0,1$	$9,8 \pm 0,3$	$10,3 \pm 0,4$
Fricativas	$11,33 \pm 0,06$	$8,0 \pm 0,4$	$10,1 \pm 0,3$	$9,8 \pm 0,6$
Africadas	$11,5 \pm 0,3$	$9,5 \pm 0,3$	$10,0 \pm 0,6$	$10,3 \pm 0,2$
Silêncio	12	3	12	9
Média	$11,14 \pm 0,03$	$9,19 \pm 0,08$	$10,50 \pm 0,07$	

Tab. 6.14: Número de componentes por estado para o sistema obtido através do método baseado na entropia dos estados, que apresenta o desempenho de 81,39% e 61,5% em termos de taxa de reconhecimento de palavras e *accuracy*, respectivamente. Os modelos acústicos correspondem aos fones adotados na base em Português

Fonemas	Estado 1	Estado 2	Estado 3	Média
Vogais/Semi-Vogais	$10,08 \pm 0,05$	$9,62 \pm 0,08$	$9,92 \pm 0,07$	$9,9 \pm 0,2$
Laterais	8 ± 2	8 ± 2	8 ± 2	$8,0 \pm 0,5$
Não-laterais	$9,7 \pm 0,3$	$9,0 \pm 0,3$	10 ± 0	$9,6 \pm 0,2$
Nasais	$9,0 \pm 0,5$	$9,0 \pm 0,5$	$10,33 \pm 0,09$	$9,4 \pm 0,4$
Oclusivas	$10,50 \pm 0,04$	$10,2 \pm 0,1$	$10,67 \pm 0,04$	$10,4 \pm 0,1$
Fricativas	$9,5 \pm 0,1$	$10,00 \pm 0,05$	$9,5 \pm 0,2$	$9,7 \pm 0,2$
Africadas	10 ± 0	10 ± 0	10 ± 0	10 ± 0
Silêncio	11	11	7	9,7
Média	$9,83 \pm 0,04$	$9,64 \pm 0,04$	$9,83 \pm 0,04$	

- A ordem decrescente das classes em termos do número de componentes utilizadas na modelagem, para cada método foi:
  - GEA: Não-laterais > Africadas > Vogais/Semi-Vogais > Fricativas > Oclusivas $\simeq$ Laterais > Nasais.
  - BIC: Não-laterais > Vogais/Semi-Vogais > Oclusivas $\simeq$ Africadas > Fricativas $\simeq$ Laterais $\simeq$ Nasais.
  - Entropia: Oclusivas > Africadas > Vogais/Semi-Vogais > Fricativas > Não-laterais > Nasais > Laterais.
  - Discriminativo: Vogais/Semi-Vogais > Oclusivas $\simeq$ Fricativas > Nasais > Laterais $\simeq$ Afri-

Tab. 6.15: Número de componentes por estado para o sistema obtido através do método discriminativo para o aumento da resolução acústica, que apresenta o desempenho de 80,3% e 59,49% em termos de taxa de reconhecimento de palavras e *accuracy*, respectivamente. Os modelos acústicos correspondem aos fones adotados na base em Português

Fonemas	Estado 1	Estado 2	Estado 3	Média
Vogais/Semi-Vogais	$10,1 \pm 0,1$	$7,8 \pm 0,1$	$9,7 \pm 0,1$	$9,2 \pm 0,5$
Laterais	7 ± 0	7 ± 0	9 ± 2	$7,8 \pm 0,5$
Não-laterais	$8,7 \pm 0,6$	7 ± 0	7 ± 0	$7,6 \pm 0,4$
Nasais	$8,7 \pm 0,6$	7 ± 0	$8,7 \pm 0,6$	$8,1 \pm 0,5$
Oclusivas	$11,2 \pm 0,2$	$7,8 \pm 0,2$	$7,8 \pm 0,2$	$8,9 \pm 0,5$
Fricativas	$9,5 \pm 0,2$	$7,8 \pm 0,2$	$9,5 \pm 0,2$	$8,9 \pm 0,5$
Africadas	9 ± 2	7 ± 0	7 ± 0	$7,8 \pm 0,4$
Silêncio	7	7	7	7
Média	$9,64 \pm 0,07$	$7,56 \pm 0,06$	$8,81 \pm 0,08$	

cadas > Não-laterais.

- Em geral, a ordem decrescente dos estados em termos do número de componentes utilizadas na modelagem foi Estado1 > Estado 3 > Estado2.

É possível associar a quantidade de componentes necessárias para a modelagem com a complexidade das distribuições de parâmetros acústicos. Dessa forma, foram alocadas mais componentes para as consoantes do que para as vogais, o que se deve a maior complexidade das distribuições acústicas das consoantes.

É intuitivo esperar que a parte central das vogais, modelada pelo Estado 2, apresente a menor dificuldade para o processo de modelagem desta classe, visto que a parte central está associada à região de maior estabilidade do processo de produção de tais fonemas.

Em última análise, o fato de se utilizar, em geral, um maior número de componentes no Estado 1 e um menor número de componentes no Estado 2, implica que os maiores efeitos de transição ou de coarticulação entre as unidades acústicas devem ocorrer na parte inicial dos fones, e que a parte central corresponde à região de maior estabilidade do processo de produção.

6.6 Análise do Desempenho no Reconhecimento e o Alinhamento Forçado de Viterbi

Neste trabalho, os sistemas utilizados para gerar as segmentações acústicas, empregadas nos cálculos discriminativos do GEA, foram escolhidos de acordo com o fator de desempenho. O sistemas

com maior fator de desempenho foram os responsáveis pela geração das segmentações utilizadas nos experimentos com a base de dados pequena em Português e também com a TIMIT. Assim, é interessante avaliar do ponto de vista acústico se tal critério de escolha está de acordo com a precisão da segmentação acústica. Neste sentido, obtiveram-se 20 sistemas de reconhecimento, alguns contendo um número fixo e outros um número variado de Gaussianas por estado, sendo que os últimos foram gerados através do GEA (pela variação do expoente de rigor K e pelo limiar de distância L_d). Para cada sistema foram calculadas as percentagens de erros de segmentação inferiores a um determinado limiar (TGG03; TG01).

As Figuras 6.8-6.12 mostram as relações obtidas entre os erros de segmentação e o desempenho no reconhecimento dos 20 sistemas gerados.

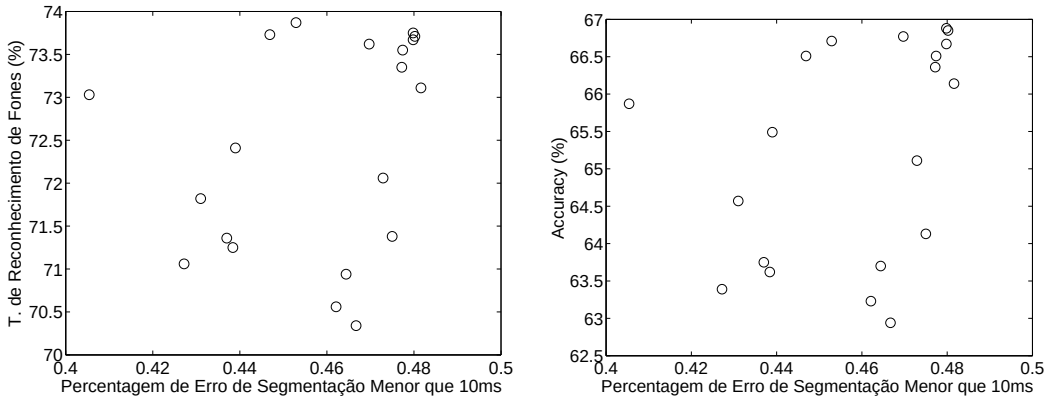


Fig. 6.8: Relação entre percentagem de erros menores que 10ms e o desempenho do sistema correspondente.

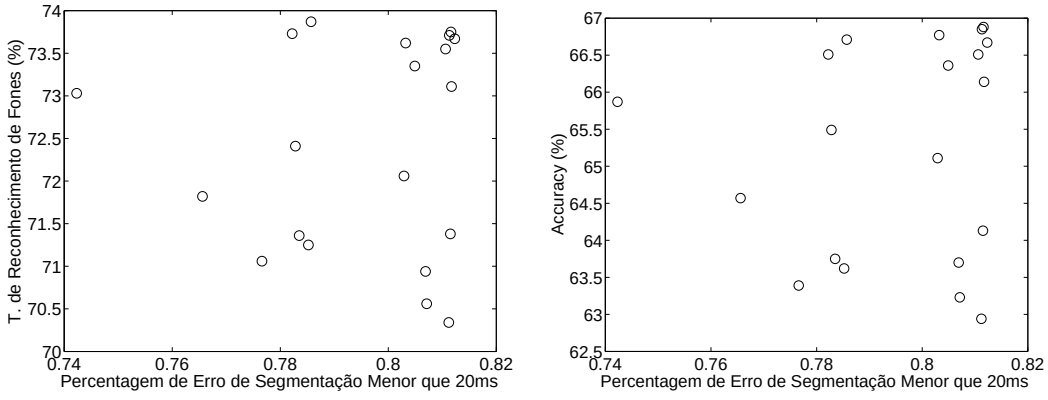


Fig. 6.9: Relação entre percentagem de erros menores que 20ms e o desempenho do sistema correspondente.

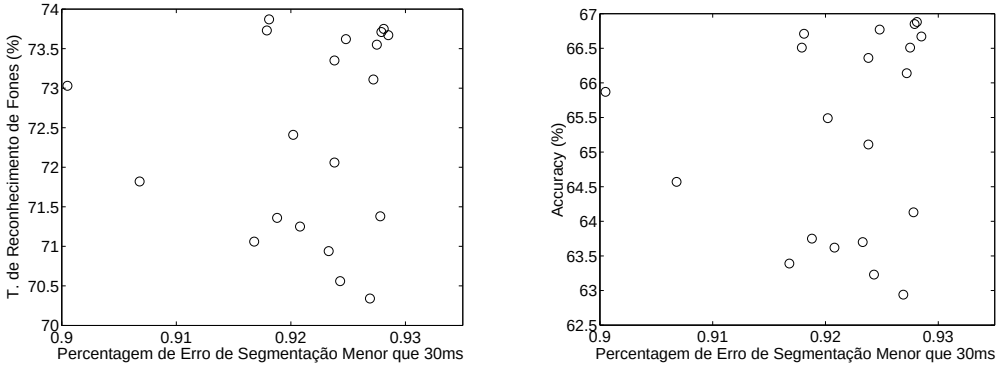


Fig. 6.10: Relação entre percentagem de erros menores que 30ms e o desempenho do sistema correspondente.

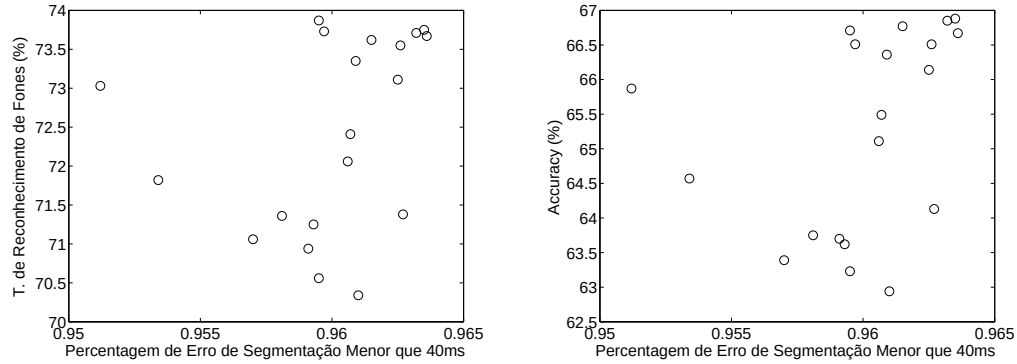


Fig. 6.11: Relação entre percentagem de erros menores que 40ms e o desempenho do sistema correspondente.

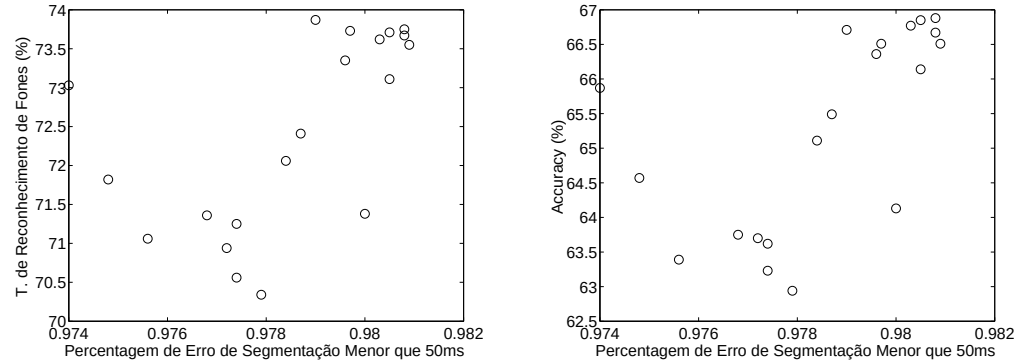


Fig. 6.12: Relação entre percentagem de erros menores que 50ms e o desempenho do sistema correspondente.

Qualitativamente, é possível observar que a correlação aumenta com o valor do limiar de erro de segmentação tolerado. Quantitativamente, a Tabela 6.16 mostra os coeficientes de correlação entre a percentagem de erros abaixo do limiar tolerado e o desempenho dos sistemas utilizados para gerar tais segmentações.

Tab. 6.16: Coeficiente de correlação (c.c.) entre a percentagem de erros abaixo do limiar tolerado e o desempenho do sistema correspondente, em termos da taxa de reconhecimento de fones, do *accuracy* e fator de desempenho F_d.

Erro tolerado	Taxa de Reconhecimento de Fones	*accuracy*	F_d
$\leq$10ms	c.c = 0,28	c.c. = 0,34	c.c. = 0,32
$\leq$20ms	c.c = 0,05	c.c. = 0,11	c.c. = 0,08
$\leq$30ms	c.c = 0,07	c.c. = 0,12	c.c. = 0,10
$\leq$40ms	c.c = 0,29	c.c. = 0,34	c.c. = 0,32
$\leq$50ms	c.c = 0,58	c.c. = 0,61	c.c. = 0,60

Os valores do coeficiente de correlação obtidos foram relativamente baixos (no máximo 61%), o que não implica em uma forte correlação entre o desempenho do sistema durante o reconhecimento de fones e o desempenho do mesmo para a segmentação acústica. Além disso, comparando-se as três medidas de desempenho durante o reconhecimento de fones contínuos, o *accuracy* foi a que apresentou a maior correlação com a precisão da segmentação acústica.

Portanto, um sistema que possui um elevado desempenho durante o reconhecimento de fones contínuos não possui necessariamente uma elevada precisão de segmentação, e o *accuracy* medido no reconhecimento apresenta a maior correlação com a precisão de segmentação, em relação a outras medidas de desempenho de reconhecimento calculadas.

6.7 Experimentos com Dados Artificiais

Na seção 3.5, apresentaram-se os experimentos iniciais realizados com dados gerados artificialmente. Observou-se que após o treinamento via MLE, diversas Gaussianas convergiram praticamente para o mesmo ponto no espaço de características e algumas convergiram para as distribuições incorretas. Assim, aplicou-se o GEA ($L_d = 0,5$ e $k = 1$) a fim de avaliar sua atuação sobre as Gaussianas obtidas após o treinamento.

O sistema gerado após o treinamento possui 120 Gaussianas, enquanto o sistema resultante da aplicação do GEA possui 46 Gaussianas. É possível verificar na Figura 6.13, que as Gaussianas vermelhas que convergiram para a distribuição azul foram eliminadas, assim como o número de Gaussianas do sistema reduziu consideravelmente, conforme esperado, uma vez que diversas Gaussianas se encontravam praticamente no mesmo ponto no espaço de características.

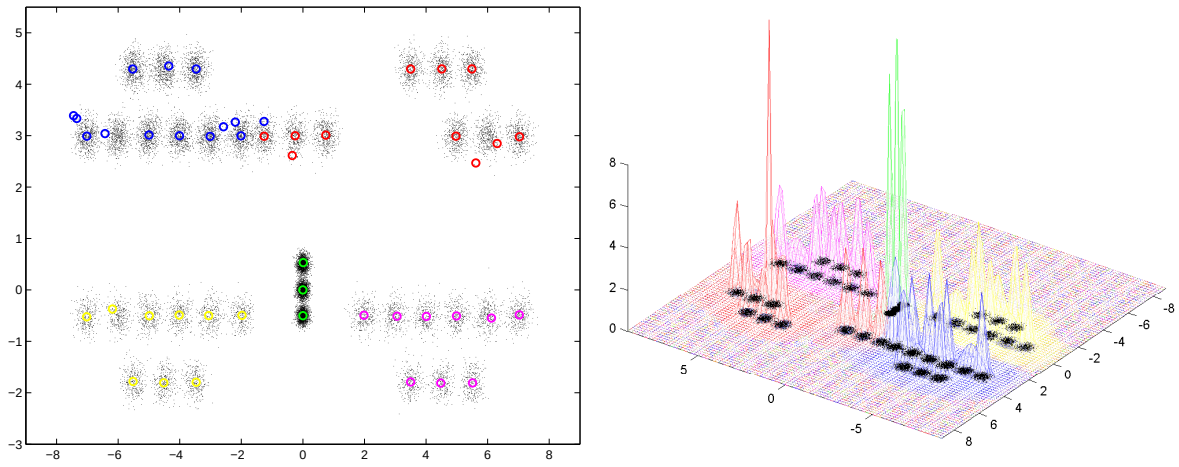


Fig. 6.13: Aplicação do GEA para o sistema treinado via MLE, utilizando dados artificiais.

6.8 Discussão

Os resultados fornecidos pelo novo GEA foram comparáveis aos obtidos pelo método da baseado na entropia dos estados, e apresentou ganhos consideráveis em relação ao BIC e ao método discriminativo para o aumento da resolução acústica, mesmo quando as comparações são realizadas a partir dos resultados de reconhecimento com a gramática *Word-pairs*. Tal gramática dificulta a avaliação da robustez dos modelos, devido às restrições impostas no processo de decodificação. A Tabela 6.17 mostra os sistemas obtidos através dos quatro métodos apresentados, que forneceram os os sistemas com as maiores economias (para $F_d \geq 0$) e maiores desempenhos, nos experimentos com a base em Português.

Se as comparações forem realizadas a partir dos resultados obtidos com a utilização da gramática *Back-off bigram* na decodificação, pode-se observar um desempenho superior do novo GEA em relação aos demais métodos apresentados, o que pode ser observado pela Tabela 6.18.

O novo GEA se mostrou eficiente na determinação de sistemas com um melhor compromisso entre desempenho e complexidade, quando se prioriza o desempenho no reconhecimento. Além disso, os resultados tornaram evidentes a importância de se realizar a análise discriminativa juntamente com a análise interna. Entretanto, ainda são necessários novos testes a fim de se determinar a medida mais apropriada para a contribuição de cada amostra para a GIM (WGP ou cálculo de hipervolumes) e também da medida de distância mais apropriada para a análise interna (distância Euclidiana convencional ou modificada).

Os resultados obtidos através de quatro segmentações distintas (Segment 0, Segment I, Segment II e Segment III), geradas a partir de sistemas de reconhecimento com diferentes desempenhos, mostraram que há uma indicação de que a escolha da segmentação mais apropriada para a análise discri-

Tab. 6.17: Comparação entre os sistemas obtidos pelo GEA (algoritmo discriminativo utilizando medidas de hipervolume + análise interna utilizando a medida de distância Euclidiana modificada), BIC, método baseado na entropia dos estados (com re-alinhamento de Viterbi a cada iteração do algoritmo) e método discriminativo para o aumento da resolução acústica dos modelos, com o emprego da gramática *Word-pairs* na decodificação. Os valores entre parênteses correspondem à diferença entre o desempenho do sistema com número variado de componentes por estado e o de referência (1188 Gaussianas).

Sistemas com as maiores economias, desde que $F_d \geq 0$ (Objetivo III)					
Método	Parâmetros	Número de Gauss. no Sist. Reduzido	Percent. Correta (%)	Reco. Accur. (%)	F_d
GEA	$L_d = 6,5$ K=10^5	893	94,37 (+0,3)	91,52 (+0,27)	+0,57
BIC	$\lambda = 0,07$	1006	94,18 (+0,11)	91,25 (0)	+0,11
Entropia	incremento=100	996	94,64 (+0,57)	92,35 (+1,1)	+1,67
Discriminativo	-	-	-	-	-

Sistemas com os maiores desempenhos (Objetivo II)					
Método	Parâmetros	Número de Gauss. no Sist. Reduzido	Percent. Correta (%)	Reco. Accur. (%)	F_d
GEA	$L_d = 4,5$ K=10^6	995	94,64 (+0,57)	92,05 (+0,8)	+1,37
BIC	$\lambda = 0,07$	1006	94,18 (+0,11)	91,25(0)	+0,11
Entropia	incremento=100	996	94,64 (+0,57)	92,35 (+1,1)	+1,67
Discriminativo	$M_{10} \times M_{11}$-$0,45$	1094	93,95 (-0,12)	90,95 (-0,3)	-0,42

minativa realizada pelo GEA, pode ser baseada nos valores de *accuracy*.

Os experimentos realizados com a TIMIT, no intuito de avaliar a qualidade das segmentações geradas pelos sistemas obtidos pelo GEA, também mostraram a eficiência de tal método na obtenção de sistemas reduzidos que superem os sistemas de referência originais, em termos de desempenho. Com relação às segmentações geradas pelos sistemas reduzidos, não foi observada uma forte correlação entre desempenho no reconhecimento de fones contínuos (coeficiente de correlação máximo de 61%) e a precisão do alinhamento forçado de Viterbi. Além disso, dentre as medidas de desempenho observadas, o *accuracy* é a que apresenta a maior correlação com a precisão da segmentação acústica.

Por fim, analisando as classes de fonemas utilizadas nos experimentos com a base de dados pequena em Português, em geral, os modelos acústicos das consoantes foram os que consumiram um maior número de componentes. Uma possível explicação reside no fato de que as vogais apresentam uma região de estabilidade durante o processo de produção acústico e podem ser caracterizadas principalmente por componentes de baixa freqüência, enquanto as consoantes são caracterizadas predominantemente por componentes de freqüências elevadas. Portanto, as vogais devem possuir características espectrais mais estáveis do que as consoantes, facilitando dessa forma o processo de

Tab. 6.18: Comparação entre os sistemas obtidos pelo GEA (algoritmo discriminativo utilizando medidas de hipervolume + análise interna utilizando a medida de distância Euclidiana modificada), BIC, método baseado na entropia dos estados (com re-alinhamento de Viterbi a cada iteração do algoritmo) e método discriminativo para o aumento da resolução acústica dos modelos, com o emprego da gramática *Back-off bigram* na decodificação. Os valores entre parênteses correspondem a diferença entre o desempenho do sistema com número variado de componentes por estado e o de referência (1296 Gaussianas).

Sistemas com as maiores economias, desde que $F_d \geq 0$ (Objetivo III)					
Método	Parâmetros	Número de Gauss. no Sist. Reduzido	Percent. Correta (%)	Reco. Accur. (%)	F_d
GEA	$L_d = 5,5$ K=10^5	985	80,82 (-0,19)	58,59 (+0,34)	+0,15
BIC	-	-	-	-	-
Entropia	incremento=100	1040	81,24 (+0,23)	61,57 (+3,32)	+3,55
Discriminativo	$M_7 \times M_{12}$-$0,8$	936	80,3 (-0,71)	59,49 (+1,24)	+0,53

Sistemas com os maiores desempenhos (Objetivo II)					
Método	Parâmetros	Número de Gauss. no Sist. Reduzido	Percent. Correta (%)	Reco. Accur. (%)	F_d
GEA	$L_d = 4,5$ K=10^6	1055	82,34 (+1,33)	62,37 (+4,12)	+5,45
BIC	$\lambda = 0,05$	1110	80,75 (-0,26)	57,6(-0,65)	-0,91
Entropia	incremento=100	1055	81,39 (+0,38)	61,5 (+3,25)	+3,63
Discriminativo	$M_7 \times M_{12}$-$0,8$	936	80,3 (-0,71)	59,49 (+1,24)	+0,53

classificação.

6.9 Conclusões

O processo de eliminação de Gaussianas (GEA) proposto neste trabalho, mostrou que o treinamento de sistemas reduzidos contendo um número variado de Gaussianas por estado é mais eficiente do que o treinamento de sistemas obtidos através de outros três métodos presentes na literatura, quando uma gramática menos restritiva é utilizada na decodificação. Dois dentre os métodos apresentados adotam a estratégia de partir de sistemas com baixa complexidade e aumentar o número de Gaussianas do modelo de acordo com o critério proposto pelo método: discriminativo ou baseado na entropia dos estados. O terceiro método apresentado, o BIC, avalia diversas topologias e escolhe aquela que melhor se ajuste aos dados de treinamento e que possua o menor número de parâmetros possível, de acordo com o princípio da parcimônia.

Os resultados mostraram, portanto, que é mais fácil eliminar Gaussianas no intuito de se aumentar a robustez dos modelos, quando se utiliza o treinamento baseado em MLE, do que introduzir

novas componentes no intuito de aumentar a resolução acústica dos modelos. Tal cenário talvez seja invertido quando o processo de estimação de parâmetros do treinamento é discriminativo.

Nos experimentos realizados com a TIMIT, observou-se que a correlação entre o desempenho do sistema no reconhecimento e no alinhamento forçado de Viterbi para a obtenção de segmentações acústicas não é forte (coeficiente de correlação $\leq 61\%$), sendo o que o *accuracy* é a medida de desempenho mais correlacionada com a precisão na segmentação. Além disso, há indícios de que a segmentação mais apropriada para o GEA deve ser gerada pelo sistema de reconhecimento que apresentar o maior *accuracy*.

Em última análise, os resultados obtidos com os dados artificiais tem um caráter qualitativo, porém podem ser considerados representativos de problemas práticos encontrados durante a modelagem de dados reais de fala. Neste sentido, o GEA também se mostrou eficiente para eliminação das Gaussianas que convergiram para as distribuições incorretas durante o treinamento e para a redução do número de Gaussianas excedentes presentes nos modelos.

Capítulo 7

Conclusões

O trabalho permitiu a observação de vários aspectos práticos e teóricos relacionados com o treinamento de HMMs contínuos através de um algoritmo baseado em MLE (Baum-Welch), desde questões relacionadas com a precisão numérica das variáveis, até as questões envolvendo o ajuste de parâmetros dos modelos e a determinação das topologias mais apropriadas para o sistema. Além disso, outros aspectos importantes tanto do ponto de vista da modelagem acústica para o reconhecimento de fala, quanto do ponto de vista lingüístico, foram observados em relação à complexidade dos modelos de acordo com a classe fonética, e em relação à proximidade da segmentação automática realizada pelo alinhamento forçado de Viterbi e a segmentação manual.

Na seqüência serão apresentadas as principais conclusões do trabalho e também os pontos que exigem ainda maiores investigações, no intuito de motivar trabalhos futuros.

7.1 O Treinamento Baseado em MLE com Dados Artificiais

Os testes realizados com os dados artificiais permitiram a observação do comportamento do algoritmo de treinamento baseado em MLE, em relação ao efeito da inicialização dos parâmetros dos modelos, principalmente quando o sistema se encontra super-dimensionado. Foi possível notar que, dependendo da inicialização das médias das Gaussianas de um determinado modelo, algumas componentes podem convergir para posições no espaço de características pertencentes a distribuições associadas a outros modelos. Tal fato contribui para uma menor discriminabilidade dos modelos, o que pode resultar em erros durante a decodificação.

Outro ponto observado está relacionado com a obtenção de Gaussianas próximas entre si após o treinamento e em localizações distantes das fronteiras das distribuições. Tal redundância provavelmente tem pouco efeito sobre a capacidade de discriminação do modelo, mas certamente pode ter um efeito considerável sobre o custo computacional do sistema.

Em última análise, a aplicação do novo GEA permitiu a eliminação das Gaussianas que convergiram para as distribuições incorretas no espaço de características e também permitiu a redução do excesso de Gaussianas presentes nos modelos.

7.2 Os Modelos de Linguagem

As gramáticas *Word-pairs* e *Back-off bigram* foram utilizadas para o reconhecimento de fala contínua com a base de dados em Português, enquanto a *Bigram* foi utilizada para o reconhecimento de fones contínuos com a base TIMIT.

Foi possível observar que a gramática *Back-off bigram* permitiu a realização de testes mais rigorosos em relação à robustez dos sistemas obtidos, quando comparada com a gramática *Word-pairs*, o que se deve ao fato da primeira permitir uma maior flexibilidade ao processo de decodificação, enquanto a segunda é mais restritiva. Dessa forma, é possível separar melhor a contribuição da gramática, da contribuição da seleção de topologia, para o desempenho do sistema durante o reconhecimento.

Por último, os testes realizados com a TIMIT utilizaram a gramática *Bigram* ao nível de fones e, neste caso também foi possível se obter sistemas com um número variado de Gaussianas por estado e que apresentem um melhor compromisso entre tamanho e desempenho, em relação aos sistemas que utilizam um número fixo de Gaussianas por estado. Os experimentos visando o reconhecimento de fones contínuos têm por finalidade avaliar a eficácia do processo de modelagem e, neste caso, o novo GEA também foi validado.

7.3 Métodos para a Determinação da Complexidade dos HMMs

Nos experimentos foram implementados três métodos presentes na literatura para a determinação da complexidade dos HMMs: o BIC, o método baseado na entropia dos estados e o método discriminativo para o aumento da resolução acústica dos modelos. Além disso, também foi proposto um novo método de eliminação de Gaussianas (GEA), baseado em análises entre modelos (discriminativa) e na análise de cada modelo separadamente (análise interna).

As medidas de taxa de reconhecimento e *accuracy* foram utilizadas simultaneamente para avaliar o desempenho dos sistemas, visto que, dependendo da aplicação, pode ser dada uma maior ênfase em uma medida em detrimento da outra. Dessa forma, utilizou-se o fator de desempenho F_d como forma de considerar os ganhos ou perdas observados nas duas medidas ao mesmo tempo. Nos experimentos envolvendo o reconhecimento de fala contínua, a medida de taxa de reconhecimento pode ser mais enfatizada, enquanto nos experimentos envolvendo reconhecimento de fones contínuos, a medida de *accuracy* recebe maior ênfase.

Os resultados obtidos a partir da utilização da gramática mais restritiva durante a decodificação (*Word-pairs*), mostram que o método baseado na medida de entropia e o novo GEA forneceram os sistemas com número variado de componentes por estado que apresentaram os melhores desempenhos, fornecendo a taxa de reconhecimento de palavras e *accuracy* de 94,64% e 92,35% para o primeiro caso, e 94,64% e 92,05% para o segundo caso. Assim, os dois métodos forneceram praticamente os mesmos resultados, superando os resultados obtidos pelo BIC e pelo método discriminativo para o aumento da resolução acústica.

Por outro lado, os resultados obtidos a partir da utilização da gramática mais flexível durante a decodificação (*Back-off bigram*), mostraram que o GEA permitiu a determinação do sistema que apresentou o melhor desempenho, fornecendo a taxa de reconhecimento de palavras e *accuracy* de 82,34% e 62,37%, respectivamente, seguido do método baseado em medidas de entropia, que forneceu 81,39% e 61,5% para as mesmas medidas de desempenho.

Portanto, o melhor sistema obtido pelo GEA de acordo com o Objetivo II (vide seção 4.3), foi equivalente ao melhor sistema obtido através dos demais métodos discutidos, quando se utilizou uma gramática mais restritiva, e foi superior aos demais quando se utilizou uma gramática mais flexível durante a decodificação.

Pode-se constatar também que somente o BIC não possibilitou a determinação de sistemas de acordo com o Objetivo III. Além disso, não se pode afirmar qual o método que fornece a maior economia de parâmetros, visto que apenas três tamanhos diferentes foram testados no método baseado na entropia dos estados. Neste sentido, são necessários mais experimentos a fim de se avaliar o método mais apropriado do ponto de vista do Objetivo III (vide seção 4.3).

Em última análise, todos os métodos se mostraram eficientes com relação ao Objetivo I (vide seção 4.3), onde foram observados ganhos de desempenho em relação aos sistemas com número fixo de componentes por estado de mais de 2%, independentemente da medida de desempenho adotada.

7.4 Importância da Segmentação Acústica para o GEA

As segmentações acústicas das sentenças de treinamento são de fundamental importância para a análise discriminativa realizada pelo GEA. Os resultados mostraram que as segmentações (Segment I e Segment II) geradas pelos sistemas de reconhecimento de fala (Sistema I e Sistema II) que apresentaram os maiores valores de *accuracy*, permitiram a obtenção dos sistemas com número variado de Gaussianas por estado, através do GEA, que apresentaram os maiores desempenhos.

Portanto, há indícios de que a medida de *accuracy* deve ser a mais apropriada para a escolha do sistema de reconhecimento de fala a partir do qual serão geradas as segmentações através do alinhamento forçado de Viterbi. Porém ainda são necessários mais testes a fim de se obter resultados

mais conclusivos.

Neste sentido, foi importante avaliar a relação entre a precisão das segmentações acústicas obtidas pelo alinhamento forçado de Viterbi e o desempenho dos sistemas de reconhecimento utilizados para essa tarefa. Os resultados mostraram que a correlação entre o desempenho no reconhecimento e no alinhamento forçado de Viterbi é maior quando o *accuracy* é utilizado, apesar de mesmo neste caso não ser alta (coeficiente de correlação $\leq 61\%$).

7.5 Complexidade dos Modelos por Classe Fonética

Os modelos das vogais mostraram claramente que o número de Gaussianas utilizadas no primeiro estado é maior do que o número utilizado no terceiro estado, e o segundo estado é o que apresenta o menor número de Gaussianas. A parte central da produção acústica das vogais é bastante estável, a qual é modelada pelo segundo estado do HMM, o que pode justificar o emprego de um menor número de Gaussianas neste caso. O primeiro e o terceiro estado modelam as transições entre os fonemas e, assumindo que o número de Gaussianas pode ser um indicativo da dificuldade para a modelagem da distribuição, a parte inicial das vogais, que é modelada pelo primeiro estado, é a que evidencia os maiores efeitos de coarticulação (transição entre os fones) do processo de produção acústico.

Em última análise, os modelos das consoantes apresentaram um maior número de Gaussianas do que as vogais. Porém, no intuito de se obterem resultados mais conclusivos, ainda são necessárias investigações para se determinar qual a ordem de complexidade dos modelos por classe fonética.

7.6 Trabalhos Futuros

O novo método proposto se mostrou eficiente para a determinação da topologia dos HMMs de acordo com os Objetivos I, II e III, para o problema de reconhecimento de fala contínua. Além disso, a aplicação do GEA sobre dados gerados artificialmente mostrou a atuação do mesmo para a eliminação de Gaussianas que modelam as distribuições incorretas e para a redução do excesso de Gaussianas presentes nos modelos. Assim, é interessante estender os conceitos do GEA para a determinação da complexidade de qualquer modelo que utilize misturas de Gaussianas para problemas relacionados com reconhecimento de padrões.

A modelagem foi realizada para fones independentes de contexto, mas pode ser estendido para o caso dos fones dependentes de contexto. Neste caso, o GEA pode auxiliar a determinação da complexidade dos HMMs após a utilização do *phonetic decision tree-based state tying*, por exemplo.

A utilização do GEA para a implementação de um sistema de reconhecimento de fala embarcado pode permitir que sistemas com número variado de Gaussianas por estado possam atender as limi-

tações físicas dos dispositivos eletrônicos, tais como memória e capacidade de processamento em tempo real, e ao mesmo tempo apresentar um desempenho no mínimo equivalente ao dos sistemas maiores com número fixo de Gaussianas por estado.

As aplicações de reconhecimento de fala têm se tornado relativamente comuns e, neste sentido, os aspectos práticos abordados durante este trabalho podem contribuir para a implementação de sistemas de informação integrados de reconhecimento e síntese áudio-visual de fala, que facilitem cada vez mais as tarefas do cotidiano.

Referências Bibliográficas

[Agu00] Luís Antônio Aguirre. *Introdução à Identificação de Sistemas - Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais*. Editora UFMG, 2000.

[Aka74] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.

[ASM92] A. Alcaim, J. A. Solewicz, and J. A. Moraes. Frequência de ocorrência dos fones e lista de frases foneticamente balanceadas no português falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações.*, 7(1):23–41, 1992.

[Bak75] J. K. Baker. The dragon system-an overview. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(1):24–29, 1975.

[BBdSM86] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. Maximum mutual information estimation. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1986.

[BHS02] Alain Biem, Jin-Young Ha, and Jayashree Subrahmonia. A bayesian model selection criterion for HMM topology optimization. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2002.

[Bie03] Alain Biem. Model selection criterion for classification: Application to HMM topology optimization. In *7-th International Conference on Document Analysis and Recognition (ICDAR'03)*, 2003.

[BJM83] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelyhood approach to continuous speech recognition. *IEEE Transactions on Pattern Anal. Machine Intell.*, 5:179–190, 1983.

[BP66] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Stat.*, 37:1554–1563, 1966.

[BP98] L. R. Bahl and M. Padmanabhan. A discriminant measure for model complexity adaptation. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1998.

[Cam02] Cambridge University Engineering Departament. *The HTK Book*, 2002.

[CEMEG+99] S. S. Chen, M. J. F. Gales E. M. Eide, R. A. Gopinath, D. Kanevsky, and P. Olsen. Recent improvements to IBM's speech recognition systems for automatic transcription of broadcast news. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1999.

[CG98] Scott Shaobing Chen and P. S. Gopalakrishnan. Clustering via the bayesian information criterion with applications in speech recognition. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1998.

[CS96] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Classification Journal*, 13:195–212, 1996.

[CS00] Imre Csiszár and Paul C. Shields. The consistency of the bic markov order estimator. In *IEEE International Symposium on Information Theory*, 2000.

[CU93] Y. J. Chung and C. K. Un. Use of different number of mixtures in continuous density hidden markov models. *Electronics Letters*, 29(9):824–825, 1993.

[DKW00] Jacques Duchateau, KrisDemuynck, and Patrick Wambacq. Discriminative resolution enhancement in acoustic modelling. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2000.

[DM80] Steven B. Davis and Paul Mermelstein. Comparison of parametric representation for monosyllabic word recognition in continuous spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[FJ02] Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 24(3):381–396, 2002.

[GJPP99] Yuqing Gao, Ea-Ee Jan, Mukund Padmanabhan, and Michael Picheny. HMM training based on quality measurement. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1999.

[KTSS98] Mineichi Kudo, Hiroshi Tenmoto, Satoru Sumiyoshi, and Masaru Shimbo. A subclass-based mixture model for pattern recognition. In *International Conference on Pattern Recognition*, 1998.

[KVY93] S. Kapadia, V. Valtchev, and S. J. Young. MMI training for continuous phoneme recognition on the TIMIT database. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1993.

[LER90] A. Ljolje, Y. Ephraim, and L. R. Rabiner. Estimation of hidden markov model parameters by minimizing empirical error rate. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1990.

[LG93] L. F. Lamel and J. L. Gauvain. High performance speaker-independent phone recognition using CDHMM. In *Eurospeech 1993*, 1993.

[LG94] Alberto Leon-Garcia. *Probability and Random Process for Electrical Engineering*. Addison-Wesley, 1994.

[LGW03] X. Liu, M. J. F. Gales, and P. C. Woodland. Automatic complexity control for hlda systems. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2003.

[LH89a] Kai-FU Lee and Hsiao-Wuen Hon. Speaker-independent phone recognition using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1641–1648, 1989.

[LH89b] Kai-Fu Lee and Hsiao-Wuen Hon Mei-Yuh Hwang. Recent progress in the sphinx speech recognition system. In *Workshop on Speech and Natural Language*, 1989.

[LLNB04] Christophe Lévy, Georges Linarès, Pcascal Nocera, and Jean-Grançois Bonastre. Reducing computational and memory cost for cellular phone embedded speech recognition system. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2004.

[MA94] Patricia McKenzie and Michael Alder. Selecting the optimal number of components for a gaussian mixture model. In *IEEE International Symposium on Information Theory*, 1994.

[Mar97] José Antônio Martins. *Avaliação de Diferentes Técnicas para Reconhecimento de Fala*. PhD thesis, Universidade Estadual de Campinas, 1997.

[Nil94] Les T. Niles. Acoustic modeling for speech recognition based on spotting of phonetic units. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1994.

[OWW94] J. J. Odell, P. C. Woodland, and P. C. Woodland. Tree-based state clustering for large vocabulary speech recognition. In *International Symposium on Speech, Image Processing and Neural Networks*, 1994.

[Pap84] Athanasios Papoulis. *Brownian Movement and Markoff Processes*. McGraw-Hill, 1984.

[PB00] M. Padmanabhan and L. R. Bahl. Model complexity adaptation using a discriminant measure. *IEEE Transactions on Speech and Audio Processing*, 8(2):205–208, 2000.

[PB02] Darryl William Purnell and Elizabeth C. Botha. Improved generalization of MCE parameter estimation with application to speech recognition. *IEEE Transactions on Speech and Audio Processing*, 10(4):232–239, 2002.

[Pic93] Joseph W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1223, 1993.

[RC00] Wolfgang Reichl and Wu Chou. Robust decision tree state tying for continuous speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 8(5):555–566, 2000.

[Ris89] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.

[RJ93] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[SiI02] Koichi Shinoda and Ken ichi Iso. Efficient reduction of gaussian components using mdl criterion for HMM-based speech recognition. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2002.

[SIK86] Y. Sakimoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics*. KTK Scientific Publishers, 1986.

[TG01] Doroteo Torre Toledano and Luis A. Hernández Gómez. Local refinement of phonetic boundaries: A general framework and its application using different transition models. In *Eurospeech 2001*, 2001.

[TGG03] D. T. Toledano, L. A. H. Gómez, and L. V. Grande. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6):617–625, 2003.

[TKS99] Hiroshi Tenmoto, Mineichi Kudo, and Masaru Shimbo. Determination of the number of components based on class separability in mixture-based classifiers. In *International Conference on Knowledge-based Intelligent Information Engineering Systems*, 1999.

[Val95] Valtcho Valtchev. *Discriminative Methods in HMM-based Speech Recognition*. PhD thesis, University of Cambridge, 1995.

[WL94] Liang Wang and Gaëtan A. Libert. Combining pattern recognition techniques with akaike's information criteria for identifying ARMA models. *IEEE Transactions on Signal Processing*, 42(6):1388–1396, 1994.

[WOY94] P. C. Woodland, J. J. Odell, and S. J. Young. Large vocabulary continuous speech recognition using HTK. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1994.

[Yno99] Carlos Alberto Ynoguti. *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*. PhD thesis, Universidade Estadual de Campinas, 1999.

[YOW94] S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modeling. In *ARPA Human Language Technology Workshop*, 1994.

[YV04a] Glauco F. G. Yared and Fábio Violaro. Determining the number of gaussians per state in HMM-based speech recognition system. In *International Workshop on Telecommunications*, 2004.

[YV04b] Glauco F. G. Yared and Fábio Violaro. Finding the more suitable HMM size in continuous speech recognition systems. In *International Information and Telecommunications Technologies Symposium*, 2004.

[YVS05] Glauco F. G. Yared, Fábio Violaro, and Lívio C. Sousa. Gaussian elimination algorithm for HMM complexity reduction in continuous speech recognition systems. In *Interspeech - Eurospeech 2005*, 2005.

[Zha93] Yunxin Zhao. A speaker-independent continuous speech recognition system using continuous mixture gaussian density HMM of phoneme-sized units. *IEEE Transactions on Speech and Audio Processing*, 1(3):345–361, 1993.

[ZWZ91] Yunxin Zhao, Hisashi Wakita, and Xinhua Zhuang. An HMM based speaker-independent continuous speech recognition system with experiments on the TIMIT database. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1991.