UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

TÉCNICAS EFICIENTES DE IDENTIFICAÇÃO AUTOMÁTICA DE LOCUTORES

FREDERICO QUADROS D'ALMEIDA

ORIENTADOR: FRANCISCO ASSIS DE OLIVEIRA NASCIMENTO

TESE DE DOUTORADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: PPGENE.TD – 037/09
BRASÍLIA, MARÇO DE 2009

FICHA CATALOGRÁFICA

D'ALMEIDA, FREDERICO QUADROS
Técnicas Eficientes de Identificação Automática de Locutores [Distrito Federal] 2009.
146p., 210 x 297 mm (ENE/FT/UnB, Doutor, Engenharia Elétrica, 2009)
Tese de Doutorado – Universidade de Brasília. Faculdade de Tecnologia.
Departamento de Engenharia Elétrica
1. Reconhecimento Automático de Locutores 2. Modelos de Misturas de Gaussianas
3. Modelos Multicondicionais 4. Eficiência Computacional
5. Processamento Digital de Sinais
I. ENE/FT/UnB II. Título

REFERÊNCIA BIBLIOGRÁFICA

D'ALMEIDA, F. Q. (2009). Técnicas Eficientes de Identificação Automática de Locutores. Tese de Doutorado em Engenharia Elétrica, Publicação PPGENE.TD-037/09, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 146p.

CESSÃO DE DIREITOS

AUTOR: Frederico Quadros D'Almeida

TÍTULO: Filtração em Múltiplas Etapas Aplicada ao Tratamento de Água com Presença de Algas: Avaliação de Variáveis Operacionais.

GRAU: Doutor ANO: 2009

_____

Frederico Quadros D'Almeida
Universidade de Brasília, Faculdade de Tecnologia, Departamento de Engenharia Elétrica
Campus Universitário Darcy Ribeiro, Caixa Postal 4386
70.910-900 - Brasília – DF - Brasil

RESUMO

TÉCNICAS EFICIENTES DE IDENTIFICAÇÃO AUTOMÁTICA DE LOCUTORES

Autor: Frederico Quadros D'Almeida

Orientador: Francisco Assis de Oliveira Nascimento

Programa de Pós-graduação em Engenharia Elétrica

Brasília, março de 2009

Os sistemas de identificação automática de locutor têm despertado crescente interesse científico atualmente. A aplicação de novas formas de modelagem da voz dos locutores tem melhorado de modo significativo a robustez desses sistemas a ruído, tornando sua aplicação prática viável em situações reais nas quais não se dispõe de áudio de boa qualidade. Contudo, essa crescente qualidade na modelagem e a consequente melhora no desempenho dos sistemas de identificação têm promovido, como efeito colateral, o aumento no custo computacional das tarefas de identificação. Em muitas situações, seja pelo grande número de locutores a serem testados, seja pela necessidade de uma resposta rápida do sistema, esse custo elevado torna proibitiva a aplicação efetiva das ferramentas de identificação automática de locutor.

Neste trabalho são propostas, implementadas, avaliadas e validadas novas técnicas que buscam reduzir significativamente o custo computacional associado a tarefas de identificação automática de locutores sem, contudo, afetar o desempenho do sistema no que concerne às taxas de identificações corretas.

Os métodos apresentados exploram características próprias dos modelos multicondicionais de mistura de gaussianas (GMM – *Gaussian Mixture Models*), modelagem comumente aplicada nos sistemas de identificação de locutores robustos a variações na qualidade do áudio questionado. O foco principal das novas técnicas apresentadas é reduzir o número de componentes gaussianas a serem calculadas no processo de identificação, o que possibilita a consequente redução do custo computacional.

Os resultados obtidos com as técnicas introduzidas neste trabalho demonstram que é possível obter reduções superiores a 90% no custo computacional das tarefas de identificação de locutores sem afetar o desempenho do sistema.

ABSTRACT

EFFICIENT AUTOMATIC SPEAKER IDENTIFICATION TECHNIQUES

Author: Frederico Quadros D'Almeida

Supervisor: Francisco Assis de Oliveira Nascimento

Programa de Pós-graduação em Engenharia Elétrica

Brasília, March, 2009

Automatic speaker identification systems are a very attractive research field currently. The application of new voice modeling techniques have significantly increased the noise robustness of the systems, making it possible to develop practical applications suited to real audio conditions, where one cannot guarantee high audio quality. However, these advances in voice modeling and the consequent improvement on the identification have caused, as a side effect, a relevant increas on the computational cost of the task. In many situations, the large number of speakers in the database or the need for a fast identification makes it prohibitive to accept this much elevated cost of the new modeling schemes.

On this work, new techniques to reduce significantly the computational effort associated with automatic speaker identification tasks without affecting the system identification performance are presented, implemented, evaluated and validated.

The presented methods explore some characteristics typical of the multiconditional Gaussian Mixture Models (GMM), a very commonly used modeling technique on noise robust speaker identification systems. The main goal of the new presented techniques is to reduce the number of gaussian components to be calculated during the speaker identification process, so that its computational cost is minimized.

Results show that, by using a combination of the novel techniques, it is possible to surpass a 90% reduction on the effort of a speaker identification task without affecting the system performance.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIAÇÕES

GMM	*Gaussian Mixture Models* (Modelos de Mistura de Gaussianas)
HMM	*Hidden Markov Models* (Modelos Ocultos de Markov)
MCM	Método da Condição Máxima
MSC	Método da Soma das Condições
MCP	Método da Condição Persistente
MCP-LV	Método da Condição Persistente Linearmente Variada
$MCP\text{-}LV_{dB}$	Método da Condição Persistente Linearmente Variada pelo Valor da SNR em Decibéis
MMA	Modelos Multicondicionais Adaptativos
MGD	Método das Gaussianas Dominantes
MR-GMM	Modelos de Mistura de Gaussianas Multirresolução
$MTP(^{+}/_{-})$	Método do Treinamento Progressivo (positivo/negativo)
RAL	Reconhecimento Automático de Locutor
SNR	*Signal-to-Noise Ratio* (Relação Sinal-Ruído)
TO	Treinamento Ótimo
TS	Treinamento Sintonizado

LISTA DE SÍMBOLOS

b_i	Densidade componente gaussiana de um modelo GMM
λ	Modelo GMM de um locutor qualquer
λ_s	Modelo GMM do locutor *s*
M	Número total de componentes gaussianas de um GMM
$\vec{\mu}_i$	Vetor de médias de um GMM
$\vec{\mu}_i^{[n]}$	Vetor de médias de um GMM para o ciclo *n* de treinamento (*n=0* representa os valores de inicialização)
n	Índice de um modelo GMM componentes de um modelo multicondicional
N	Número total de modelos GMM compondo um modelo multicondicional
p_i	Pesos das componentes gaussianas de índice *i* de um GMM
$p_i^{[n]}$	Peso da componente gaussiana de índice *i* de um GMM, para o ciclo *n* de treinamento (*n=0* representa os valores de inicialização)
s	Índice de um locutor em particular
S	Número total de locutores no banco de dados (ou universo)
$\vec{\sigma}_i$	Vetor de variâncias de um GMM (diagonal principal da matriz de covariâncias)
$\vec{\sigma}_i^{[0]}$	Vetor de variâncias de um GMM (diagonal principal da matriz de covariâncias) para o ciclo *n* de treinamento (*n=0* representa os valores de inicialização)
Σ_i	Matriz de covariâncias de um GMM
t	Índice da janela de análise de um trecho de áudio (de teste ou de treino)
T	Número total de janelas de análise em um trecho de áudio (de teste ou treino)
U	Universo de locutores do banco de dados
$\vec{x}_t$	Vetor de parâmetros da voz para a janela de análise *t* do áudio de treino

X_s Coleção de vetores de parâmetros da voz, $\vec{x}_t$, para todas as janelas de análises do áudio de treino do locutor *s*

$y[k]$ Valor da amostra k do sinal *y*

$\vec{y}_t$ Vetor de parâmetros da voz para a janela de análise *t* do áudio de teste

Y_s Coleção de vetores de parâmetros da voz, $\vec{y}_t$, para todas as janelas de análises do áudio de teste do locutor *s*

1 INTRODUÇÃO

É sabido, pela experiência diária, que é possível identificar uma pessoa por sua voz com grande índice de acerto. Por exemplo, quando se fala ao telefone, poucas palavras são necessárias para que se identifique a voz de um interlocutor conhecido, mesmo que ele não se apresente. Essa capacidade de determinar a identidade de uma pessoa pela análise de sua voz é extremamente interessante e possui diversas aplicações como o controle de acesso a ambientes restritos, aplicações de banco por telefone, além de aplicações na área de investigações policiais e em processos judiciais.

A fim de explorar essas possibilidades, desde a década de 1930 (McGehee, 1937), esforços vêm sendo dedicados para o desenvolvimento de técnicas capazes de, a partir da análise de um trecho de voz, reconhecer o locutor que a produziu. Posteriormente, especialmente após o surgimento dos computadores digitais, os primeiros sistemas automatizados de reconhecimento de locutor começaram a surgir (Luck, 1969; Atal, 1972; Lummis, 1973).

Os sistemas de reconhecimento automático de locutor (RAL) encontram-se atualmente bastante desenvolvidos, tendo alcançado um patamar de desempenho para situações de áudio de boa qualidade (baixo ruído, boa codificação) que não permite grandes evoluções adicionais (Campbell, 1997). De fato, hoje, boa parte dos esforços se dirige ao desenvolvimento de sistemas cada vez mais resistentes a variações na qualidade do áudio questionado, especialmente para a construção de sistemas robustos ao ruído (Ming, Stewart, Vaseghi, 2005; Xu *et* al, 2005; Ming *et al*, 2007). As abordagens utilizadas para esse fim são variadas; contudo, um ponto comum a muitos dos trabalhos é a utilização de alguma forma de modelagem multicondicional (Matsui, Kanno e Furui, 1996; Ming, Stewart e Vaseghi, 2005; Yang e Gong, 2006; Ming *et al*, 2007). A modelagem multicondicional (também denominada de modelagem multi-SNR) se baseia na construção de diferentes modelos para um mesmo locutor, sendo que cada modelo é treinado para uma condição específica (de nível de ruído, por exemplo). Dessa forma, havendo, para cada locutor, modelos adaptados a diferentes situações, consegue-se como resultado final um sistema mais abrangente e mais robusto a variações nas condições do áudio questionado.

Embora a modelagem multicondicional seja uma técnica que produz bons resultados no aumento da robustez do sistema de RAL ao ruído e a outras variações nas características do áudio de entrada, ela provoca, como efeito colateral, um significativo incremento no

custo computacional do sistema. Na realidade, como cada locutor passa a ser representado por um número *N* de modelos multicondicionais, tem-se basicamente um aumento de *N* vezes no custo do processo de reconhecimento. Esse aumento, especialmente nas tarefas de identificação de locutor, onde têm que ser examinados todos os modelos da base de dados a fim de se determinar o mais ajustado ao áudio questionado, pode tornar inviável a efetiva implantação desse tipo de sistema. A situação se agrava ainda mais no caso de sistemas com bases de dados com muitos locutores, ou quando há necessidade de respostas rápidas.

Nesse sentido, o presente trabalho apresenta sua contribuição no desenvolvimento de sistemas de RAL que mantêm as características de desempenho e de robustez ao ruído dos sistemas multicondicionais atualmente em uso, mas que exigem esforço computacional significativamente menor, permitindo o desenvolvimento de aplicações de identificação em bancos de dados com grande quantidade de locutores.

1.1 HISTÓRICO

Os estudos sobre o reconhecimento de locutores a partir da sua voz são muito antigos. As primeiras referências de pesquisas bem estruturadas sobre esse tema remontam à década de 1930, com os estudos de Frances McGehee (1937) decorrentes do caso do sequestro e morte do filho do aviador Charles Lindberg em 1932 (Hollien, 2002). Nessa época, contudo, os estudos se baseavam apenas em identificações feitas por ouvintes, sem a utilização de qualquer ferramenta ou técnica específica, e as pesquisas se direcionavam mais para a determinação da confiabilidade dessas identificações que para o aperfeiçoamento dos métodos utilizados.

Somente após o fim da Segunda Guerra Mundial, com a publicação de trabalhos sobre o recém-desenvolvido espectrograma[*] (Gray e Kopp, 1944), o desenvolvimento de técnicas de reconhecimento de locutores ganhou verdadeiro impulso. Nesse momento, imaginava-se que o reconhecimento de pessoas por sua voz, comparando visualmente espectrogramas, seria análogo à identificação pelas impressões digitais, tanto em simplicidade de execução como em confiabilidade de resultados. Por essa razão, cunhou-se o termo *voiceprint* ("impressão" de voz), para designar os espectrogramas, numa alusão ao termo *fingerprint* (impressão digital). Não demorou muito, contudo, para que se percebesse que os métodos es-

[*] De fato, o desenvolvimento do espectrograma deu-se anos antes, mas essa pesquisa era de interesse militar e não foi publicada até o final da guerra (Lindh, 2004; Ericksson, 2005).

pectrográficos eram mais adequados para a observação de similaridades entre as pronúncias de uma mesma palavra por diferentes locutores que para a percepção de características únicas capazes de diferenciar um locutor dos demais. Com isso, o uso do espectrograma foi dominantemente redirecionado para aplicações de ensino de línguas estrangeiras e para treinamento de fala para deficientes auditivos (Kopp e Green, 1946), e o termo *voiceprint* praticamente caiu em desuso.

Uma virada na utilização dos espectrogramas para o reconhecimento de locutores ocorreu em 1962, com a publicação do trabalho "*Voiceprint Identification*", de Lawrence Kersta, na revista *Nature* (Kersta, 1962). Kersta defendia a infalibilidade de seu método[*] e relatava taxas de identificações corretas de 99% ou superiores. Os métodos e os resultados de Kersta, contudo, sempre foram controvertidos, e sua aceitação na comunidade científica em geral foi restrita (Vanderslice e Ladefoged, 1967). As restrições ocorrerem especialmente porque a maioria dos outros estudos realizados em condições semelhantes obtinha resultados muito menos expressivos (Bricker e Pruzansky, 1966; Young e Campbell, 1967; Stevens *et al*, 1968), embora houvesse também estudos que apresentavam resultados corroborando o método de Kersta (Tosi *et al*, 1972).

Independentemente da polêmica em torno das técnicas de reconhecimento de locutor baseados na comparação visual de espectrogramas, é importante destacar que, até esse momento, o processo de reconhecimento era uma tarefa executada "manualmente" por um especialista treinado. Não havia uma técnica capaz de automatizar esse processo, o que se devia, em grande parte, pela falta de ferramentas computacionais.

Entre o final da década de 1960 e o início da década de 1970, os trabalhos na área de reconhecimento automático de locutores (RAL) ganharam nova força com a utilização de computadores digitais (Luck, 1969; Atal, 1972). Se iniciou, nesse período, o processo de automatização das tarefas de reconhecimento de locutor (Atal, 1974; Atal, 1976). A maioria dos trabalhos ainda enfocava o reconhecimento do locutor com base na pronúncia de um texto pré-definido (reconhecimento dependente do texto[†]), analisando a evolução temporal de certos parâmetros da voz, especialmente da frequência fundamental, formantes, intensidade, e coeficientes do preditor linear (Doddington, 1970; Atal, 1972; Lummis, 1973; Atal, 1974; Rosemberg, 1976). Um dos grandes problemas enfrentados nessas técnicas era a

[*] Em 1962, Lawrence Kersta proferiu uma palestra intitulada "*Voiceprint-identification infallibility*" no encontro anual da Sociedade de Acústica da América (*Acoustical Society of America*)
[†] Ver seção 2.1.

necessidade do perfeito alinhamento dos eventos sonoros para que se realizassem as comparações (Doddington, Flanagan e Lummis, 1972), o que não ocorre normalmente mesmo em duas pronuncias consecutivas de um mesmo texto por um mesmo locutor. Outros trabalhos desse período abordavam o reconhecimento automático de locutores pela análise das médias de parâmetros em trechos longos de voz (Furui, Itakura e Saito, 1972; Furui, 1974; Markel, Oshika e Gray, 1977; Markel e Davis, 1979), sendo, dessa maneira, os primeiros métodos independentes do texto pronunciado. Nessa época também se começou a utilizar os coeficientes cepstrais para caracterizar as vozes, inicialmente com Atal (1974), que demonstrou a superioridade da representação cepstral frente a diversos outros tipos de parâmetros, e posteriormente com Furui (1981).

A partir da década de 1980, com o aumento do poder computacional disponível, as técnicas de RAL ficaram progressivamente mais complexas, proporcionando, consequentemente, melhorias de desempenho dos sistemas. Em 1985, Soong *et al* e também Buck, Burton e Shore propuseram sistemas de RAL baseados em técnicas de Quantização Vetorial (*Vector Quantization* – VQ), após verificar o sucesso desse tipo de modelagem para aplicações de reconhecimento de fala (Shore e Burton, 1983). Também surgiram sistemas baseados em modelos ocultos de Markov (*Hidden Markov Models* - HMM), como o de Poritz (1982), embora essa técnica de modelagens e outras dela derivadas somente se popularizassem na década de 1990.

Embora os experimentos com VQ demonstrassem bons resultados, suas aplicações, em geral, eram limitadas a situações de vocabulário restrito, devido a características da própria modelagem. Visando superar essa limitação, o novo passo nos sistemas de RAL, foi a popularização das modelagens probabilísticas, ocorrida a partir da década de 1990, que proporcionou o desenvolvimento de sistemas realmente independentes do texto. Outra vez, a melhora no desempenho do reconhecimento veio acompanhada por um aumento na complexidade dos modelos e, consequentemente, de um significativo aumento no custo computacional associado.

Os modelos baseados em HMM, que demonstravam bons resultados em aplicações de reconhecimento de fala, foram os primeiros modelos probabilísticos largamente utilizados (Rosenberg, Lee e Soong, 1990; Webb e Rissanen, 1993). Os HMM são modelos que incorporam informações sobre a evolução temporal dos parâmetros, o que é muito relevante para reconhecimento de fala ou mesmo para reconhecimento de locutor em sistemas dependentes do texto. Essa informação temporal, contudo, provou não adicionar qualquer

vantagem nos sistemas de RAL independentes do texto (Tisby, 1991; Matsui e Furui, 1992), de forma que, rapidamente, a modelagem dominante para essas situações passou a ser a dos modelos de mistura de gaussianas (*Gaussian Mixture Models* – GMM), introduzidos por Reynolds (1992) e utilizados neste trabalho. Os GMM são basicamente HMM desprovidos da informação temporal de transições entre estados.

Recentemente, grande parte dos esforços nas pesquisas com sistemas de RAL busca aumentar a robustez dos modelos a situações de ruído (Ming, Stewat e Vaseghi, 2005; Ming *et al*, 2007). Nesse sentido, embora haja algumas variações, boa parte dos sistemas utiliza alguma forma de modelagem multicondicional (Matsui, Kanno e Furui, 1996) como forma de dotar o modelo de informações do locutor em diversas condições. Como não poderia deixar de ocorrer, a incorporação de condições de treinamento variadas aos modelos provocou um significativo aumento em sua complexidade, o que gera, inevitavelmente, o crescimento do custo computacional das tarefas de reconhecimento.

É nesse contexto que foi desenvolvido o presente trabalho, cujo foco principal é a redução do custo computacional das tarefas de identificação de locutores em sistemas baseados em GMM multicondicionais.

1.2 JUSTIFICATIVA

A utilização de escutas telefônicas como forma de investigação policial tem crescido de forma significativa recentemente. O alvo específico das investigações varia de país a país, dependendo de suas particularidades. No Brasil, as ações se dirigem, em geral, para organizações criminosas envolvidas com fraudes, corrupção, contrabando e outras modalidades de crimes. Em outros países, como nos Estados Unidos e em muitos países europeus, os esforços concentram-se na identificação de organizações terroristas. Seja qual for o tipo de investigação em andamento, é de essencial importância a identificação dos participantes dos diálogos, para que todos os envolvidos sejam localizados e respondam judicialmente por suas ações; ou mesmo para que sejam imediatamente detidos, caso haja ameaças iminentes.

A primeira abordagem para solucionar essa questão, sem dúvida, é a identificação dos proprietários das linhas telefônicas envolvidas. Esse método, entretanto, não tem se mostrado viável, entre outros fatores porque, ao menos no Brasil, se consideram protegidas por sigilo

também as informações cadastrais das linhas telefônicas (Brasil, 2006), o que dificulta sobremaneira a determinação dos seus proprietários[*]. Adicionalmente, deve-se considerar as imperfeições dos cadastros de titulares, decorrentes da popularização da telefonia e, principalmente, nas novas modalidades de cobrança criadas, especialmente a de telefones pré-pagos. Por fim, mesmo que essas questões fossem solucionadas, ainda haveria a questão da utilização das linhas telefônicas por outras pessoas que não seus efetivos titulares, algo comum, especialmente dentro de organizações criminosas.

Diante desse quadro, a alternativa mais viável é a identificação direta dos envolvidos nos diálogos a partir da análise de suas vozes, visto que hoje a tecnologia para isso está suficientemente madura. Contudo, embora a solução tecnológica exista, seu custo computacional ainda é proibitivo, especialmente porque o volume de identificações a serem realizadas será elevado, haja visto a quantidade de escutas simultaneamente realizadas, e porque a base de dados de locutores tende a crescer muito rapidamente.

A identificação pela análise direta da voz possibilita também outras aplicações policiais, como a identificação de locutores em materiais gravados por meio de escutas ambientais, e mesmo por gravações telefônicas que chegam à polícia sem qualquer informação dos números telefônicos envolvidos. São aplicações secundárias, de fato, mas não deixam de ser relevantes visto que não exigem alterações no sistema desenvolvido.

Há também grande uso para sistemas de RAL em aplicações de controle de acesso por voz, especialmente em aplicações acessadas por telefone, nas quais não há outra característica biométrica disponível (banco por telefone, etc.).

1.3 OBJETIVO E PRINCIPAIS CONTRIBUIÇÕES

A fim de viabilizar a proposta de identificação direta dos locutores envolvidos nos diálogos telefônicos acompanhados no curso de investigações policiais, é necessário reduzir tanto quanto possível o custo computacional desse processo sem, contudo, prejudicar o desempenho do sistema de RAL. É nesse sentido que foi desenvolvido o presente trabalho, onde

[*] Nesse caso, a questão reside na determinação dos proprietários das outras linhas telefônicas que entram em contato com a linha que está sendo monitorada, para a qual houve autorização judicial de quebra de sigilo e cujo proprietário é conhecido. Essa informação, a identidade das pessoas que entram em contato com o suspeito inicial, é de grande valor para a investigação porque pode revelar outros participantes da organização criminosa, inicialmente desconhecidos, permitindo uma abordagem muito mais eficaz no sentido de deter toda a organização e não apenas um de seus integrantes.

são apresentadas diversas técnicas capazes de minorar o esforço computacional associado às identificações automáticas e, ao mesmo tempo, manter, ou mesmo melhorar, o desempenho dos sistemas normalmente empregados.

Deve-se destacar que os sistemas aplicados nesse trabalho visam apenas, como qualquer sistema de identificação automática de locutores[*], apresentar uma lista dos melhores candidatos dado um trecho de voz questionado. Para a confirmação final da identificação, caso seja necessário utilizar essa informação como prova num processo judicial, por exemplo, será necessária a análise da voz em questão por um especialista (perito), que observará, além dos fatores acústicos, outras características de alto nível (vocabulário, correção gramatical das sentenças, vícios de pronuncia etc.) ainda não incluídas nas análises automáticas. São, portanto, nesse aspecto, sistemas semelhantes aos populares AFIS[†] (*Automatic Fingerprint Identification Systems*), porém voltados à identificação pela voz.

O foco deste trabalho, portanto, é o desenvolvimento de sistemas de RAL baseados em modelos de mistura de gaussianas (GMM) multicondicionais com desempenho de identificação semelhante aos dos sistemas atualmente em uso, mas com significativa redução do custo computacional nas tarefas de identificação dos locutores. Nesse sentido, os principais resultados alcançados nesta tese foram quatro métodos propostos, implementados, avaliados e validados, capazes de reduzir significativamente o custo computacional de identificações em sistemas de RAL multicondicional sem afetar o desempenho do sistema; são eles: o Método da Condição Persistente, os Modelos Multicondicionais Adaptativos, o Método das Gaussianas Dominantes e os Modelos de Mistura de Gaussianas Multirresolução.

Alguns desses métodos, o Método da Condição Persistente (MCP) e os Modelos Multicondicionais Adaptativos (MMA), exploram a correlação temporal das características do áudio questionado como forma de prever a condição futura desse áudio e, dessa maneira, reduzir o número efetivo de condições de treinamento que compõem o modelo.

No Método das Gaussianas Dominantes (MGD), por outro lado, foi desenvolvida uma forma alternativa de treinamento dos modelos multicondicionais (Método do Treinamento Progressivo – MTP), que promove uma forte correlação entre as componentes correspon-

[*] Ou mesmo como a maioria dos sistemas de identificação automática em geral.

[†] Nesse ponto se trata da semelhança na questão da forma de entrada de dados no sistema (impressão digital ou fragmento de impressão digital questionada x trecho de voz questionado) e de saída ou resultado do sistema (lista de candidatos mais semelhantes). Não se trata, portanto, de semelhança em termos de método de análise ou mesmo de desempenho de identificação, visto que as impressões digitais têm forma fixa e imutável, enquanto que a voz (fala) é uma manifestação dinâmica, variável e mesmo voluntariamente alterável.

dentes dos diferentes GMM que compõem o modelo multicondicional. Dessa maneira, após serem determinadas, para o GMM de uma única condição, quais as componentes gaussianas que dominam o resultado final da verossimilhança do modelo, descartam-se as demais componentes em todos os outros modelos multicondicionais do mesmo locutor a serem avaliados e, consequentemente, reduz-se a complexidade computacional do processo.

Por fim, os Modelos de Mistura de Gaussianas Multirresolução (MR-GMM) exploram o processo de identificação em etapas sucessivas de descarte de locutores, iniciando com modelos simples e progredindo para modelos mais detalhados, como forma de reduzir o universo de locutores a serem avaliados com os modelos complexos. Com isso, a complexidade média dos modelos utilizados é reduzida e é diminuído o custo computacional da identificação.

Deve-se ainda destacar que a maior parte do processamento realizado para tarefas de identificação ou de verificação de locutores é idêntica, de modo que as novas técnicas propostas e validadas nesta tese no sentido de diminuir o custo computacional de tarefas de identificação podem ser igualmente aplicadas a sistemas de verificação de locutores.

1.4 ORGANIZAÇÃO DO TRABALHO

Para facilitar a leitura e a localização dos tópicos abordados neste trabalho, esta tese foi organizada em capítulos da maneira descrita a seguir:

O capítulo 2 se inicia com uma discussão sobre os conceitos básicos do reconhecimento automático de locutores, diferenciando os problemas de identificação de locutor, foco deste trabalho, dos problemas de verificação de locutor; diferenciando ainda os sistemas dependentes do texto dos independentes do texto. Em seguida, na seção 2.2, é introduzida a modelagem por mistura de gaussianas (Reynolds, 1992), a forma de treinamento desses modelos e de identificação do locutor a quem pertence um determinado trecho de voz questionado. A modelagem com misturas de gaussianas é utilizada em todo este trabalho, por ter se mostrado a melhor solução para sistemas independentes de texto (Tisby, 1991; Matsui e Furui, 1992). Ainda no capítulo 2, na seção 2.3, é tratada a questão da seleção dos parâmetros da voz utilizados para a modelagem dos locutores. Os sistemas de RAL, em geral, utilizam a representação cepstral ou mel-cepstral, pois estudos realizados demonstraram a

superioridade desse tipo de parâmetro (Atal, 1974; Furui, 1981). Contudo, como, num trabalho mais recente (Souza e Souza, 2001), essa conclusão foi contestada, optou-se por reavaliar a questão, tendo sido feitos experimentos com diversos tipos de parâmetros de representação da voz que concluíram pela reafirmação da superioridade da representação cepstral.

O capítulo 3 trata da robustez dos sistemas de RAL a variações na qualidade do áudio questionado, especialmente com relação à variações no nível de ruído e a alterações na taxa de codificação de áudio no formato MP3. Inicialmente, na seção 3.1, se demonstra como o desempenho dos sistemas de identificação se deteriora rapidamente com a alteração da condição do áudio questionado, no caso de sistemas unicondicionais (treinados com áudio de apenas uma situação de ruído ou de codificação). A seção 3.2 demonstra uma forma tradicional (embora muito limitada) de se superar essa degradação, qual seja a utilização de sistemas "sintonizados" para a condição do áudio a ser testado. Na seção 3.3 é explanada a técnica da modelagem multicondicional (Matsui, Kanno e Furui, 1996), que permite a utilização simultânea de várias condições de treinamento num único sistema de RAL, dessa forma conferindo ao sistema grande robustez a variações na qualidade do áudio questionado, mas provocando, colateralmente, um significativo aumento no custo computacional demandado. Nessa mesma seção 3.3, são demonstrados resultados de simulações com diferentes tipos de modelagens multicondicionais que permitem concluir que a utilização da condição de máxima verossimilhança a cada janela de análise (método da condição máxima) promove os melhores resultados entre os tipos de modelagens multicondicionais. Por fim, na seção 3.3.2.1, é apresentada uma solução, desenvolvida durante as pesquisas desta tese, capaz de aumentar significativamente o desempenho de sistemas multicondicionais para as condições de ruído elevado, ao custo de um aumento no custo computacional do sistema[*].

No capítulo 4, são apresentadas as novas técnicas para a redução do custo computacional dos sistemas de RAL baseados em GMM multicondicionais, desenvolvidas durante a elaboração desta tese. Algumas das técnicas introduzidas exploram a coerência temporal da qualidade do áudio questionado como forma de minimizar os cálculos exigidos para a identificação do locutor, como o Método da Condição Persistente (MCP), apresentado na seção 4.2, e os Modelos Multicondicionais Adaptativos (MMA), apresentados na seção

[*] O aumento do custo computacional proposto pode ser eliminado com a utilização da técnica da modelagem multicondicional adaptativa (MMA) introduzida no capítulo 4.

4.3. Na seção 4.4, é exposto o Método das Gaussianas Dominantes (MGD), que explora a coerência entre componentes correspondentes dos modelos de condições distintas dentro de um único modelo multicondicional. Por fim, na seção 4.5, são apresentados os Modelos de Mistura de Gaussianas Multirresolução (MR-GMM), que exploram a possibilidade de identificação em etapas sucessivas de descarte dos locutores menos ajustados ao áudio questionado como forma de minimização do custo computacional total da identificação.

O capítulo 5 contém a conclusão dos trabalhos realizados, além de uma discussão geral dos resultados obtidos e algumas propostas de continuidade para pesquisas.

2 RECONHECIMENTO AUTOMÁTICO DE LOCUTOR

Antes de serem discutidas as técnicas de Reconhecimento Automático de Locutor (RAL), é importante diferenciar os conceitos de identificação, verificação e de reconhecimento de locutor. Usualmente, utiliza-se a expressão "identificação de locutor" para designar a tarefa de, dada uma amostra de voz, determinar quem é o seu autor, dentre um conjunto pré-definido de candidatos. Nesse caso, não há uma indicação prévia da identidade do falante. A expressão "verificação de locutor", por sua vez, é geralmente empregado quando se tem uma amostra de voz e é dada uma suposta identidade do seu autor. A tarefa, nesse caso, é confirmar se a identidade alegada é verdadeira. O "reconhecimento de locutor" é a expressão utilizada para designar genericamente uma tarefa de identificação ou de verificação.

Em princípio, a tarefa de verificação de locutor é computacionalmente mais simples que a de identificação. Como, na verificação, existe um locutor suspeito, caso sejam identificadas características divergentes entre sua voz e a amostra de voz questionada, o resultado da comparação é negativo. Caso todas as características observadas sejam compatíveis, o resultado é positivo. A verificação de locutor é uma tarefa "um para um", ou seja, são comparadas as características da amostra de voz questionada apenas com as características da voz do locutor suspeito[*].

Na identificação, como não existe um locutor em particular indicado, a amostra de voz questionada precisa ser comparada com as vozes de todos os locutores do banco de dados, a fim de determinar aquele cujas características mais se aproximam das da voz questionada. Existe ainda a possibilidade de a amostra de voz questionada não ter sido originada por nenhum dos locutores do banco de dados. Nos casos em que isso é possível, diz-se que a

[*] De fato, a maioria dos sistemas automáticos de verificação de locutor utiliza uma abordagem diferente, que aproxima a tarefa de verificação daquela do reconhecimento. Em geral, para a verificação, compara-se o áudio questionado tanto com o modelo do locutor indicado (suspeito) quanto com modelos de outros locutores da base de dados (uma amostra significativa da população). A verificação, então, é feita em termos da comparação entre o grau de semelhança (verossimilhança) da voz questionada com o modelo do locutor suspeito e com os modelos de outros locutores da população (Reynolds, Quatieri e Dunn, 2000). Esse tipo de abordagem é necessário porque todo o trabalho com o reconhecimento de locutores (automático ou não) deve ser tomado de forma estatística, em virtude das próprias características da voz, especialmente pelo fato de ser um fenômeno dinâmico e variável. Não é possível estabelecer, portanto, pela simples comparação "um para um" da voz, uma identificação fidedigna; ao contrário do que se pode realizar com análises de impressões digitais, por exemplo. Na realidade, mesmo no caso das impressões digitais, a comparação "um para um" é uma simplificação da análise estatística mais completa, viável apenas porque existe um grau de variabilidade tamanho das formas das impressões digitais que, ao se observar um determinado número de coincidências entre uma amostra questionada e um padrão, e não havendo qualquer divergência, a probabilidade de a amostra não pertencer à pessoa indicada é extremamente reduzida, podendo mesmo ser desprezada (Champod, 1995).

identificação é de conjunto aberto (*open-set*). Ao contrário, se a amostra de voz é, obrigatoriamente, de um dos locutores do banco de dados, diz-se que a identificação é em conjunto fechado (*closed-set*). A tarefa de identificação é do tipo "um para *N*", ou seja, as comparações devem ser feitas entre uma amostra de voz questionada e todos os *N* locutores do banco de dados.

Claramente, a tarefa de identificação é computacionalmente mais complexa que a de verificação, pois exige a pesquisa de todos os registros do banco de dados[*]. A tarefa de identificação em conjunto aberto é ainda mais complexa, pois, além de analisar os *N* locutores do banco de dados, o sistema precisa decidir se a voz é de um outro locutor, não cadastrado.

Essa tese enfoca o problema do desenvolvimento de técnicas computacionalmente eficientes de identificação de locutor, pois, em todos os casos, não foi indicado o pretenso autor do trecho de voz questionado. Além disso, este trabalho não lida com a questão de locutores não pertencentes ao banco de dados, sendo as análises realizadas sempre em conjunto fechado.

2.1 SISTEMAS DE RECONHECIMENTO AUTOMÁTICO DE LOCUTOR DEPENDENTES E INDEPENDENTES DO TEXTO

Os sistemas RAL podem ser divididos em duas classes principais, a dos dependentes do texto e a dos independentes do texto. Os sistemas RAL dependentes do texto possuem uma restrição muito forte: eles exigem que a amostra de voz questionada possua o mesmo conteúdo textual das vozes presentes no banco de dados. Essa restrição é necessária porque, nesses sistemas, a comparação é feita entre eventos vocais correspondentes da amostra de voz questionada e das vozes do banco de dados. Nos sistemas independentes do texto, as comparações são realizadas entre modelos que representam uma ampla gama de realizações vocais de determinado indivíduo. Assim, não é necessário que as vozes padrão e questionada possuam o mesmo conteúdo para sua comparação.

A voz, como se sabe, é um fenômeno dinâmico e variável. Para contornar a variabilidade das características acústicas da voz em um sistema de reconhecimento, uma das alternati-

[*] Apesar de as verificações de locutores exigirem, de fato, a análise de outros locutores do banco de dados, como exposto na nota anterior, não é estritamente necessário que todos sejam analisados, sendo suficiente a observação de uma amostra significativa da população. Dessa forma, a tarefa de verificação continua sendo computacionalmente menos exigente que a de identificação.

vas é realizar a comparação entre trechos que possuem exatamente o mesmo conteúdo textual. Esse é o princípio dos sistemas RAL dependentes de texto (Doddington, 1970; Atal, 1974; Park e Hazen, 2002). Nesses sistemas, são comparados diretamente parâmetros extraídos das vozes padrão e questionada em instantes equivalentes da produção de determinada sequência sonora.

Vale destacar que, nesse caso, para realizar as comparações, não é suficiente que as amostras padrão e questionada tenham o mesmo conteúdo textual; isso é apenas uma condição necessária. É preciso, ainda, obter um alinhamento perfeito entre cada realização sonora dos dois trechos de áudio, a fim de comparar blocos com conteúdo acústico equivalente (Doddington, Flanagan, Lummis, 1972; Aronowitz, 2006). A exigência de um alinhamento preciso entre os trechos de áudio a serem comparados provoca uma dificuldade adicional na construção desses sistemas, pois, mesmo quando um mesmo locutor lê um determinado texto predefinido e fixo, a velocidade da fala e os intervalos de pausa em duas execuções distintas são diferentes. Com isso, ocorre um desalinhamento das realizações vocais e é preciso todo um pré-processamento para reajustar os pontos correspondentes.

Os métodos de RAL dependentes do texto são particularmente úteis em aplicações onde o locutor é cooperativo e a amostra de voz questionada pode ser obtida sob demanda. Enquadram-se nessas características, por exemplo, aplicações de controle de acesso e de banco por telefone. Nesses casos, pode-se solicitar ao locutor que se deseja identificar que pronuncie uma determinada frase, e é mesmo possível solicitar que a frase seja repetida, caso sua realização seja muito diferente das amostras padrão armazenadas. Para muitas outras aplicações, entretanto, os sistemas de RAL dependentes do texto são inviáveis. Nesses casos, como, muitas vezes, as amostras de voz são obtidas sem a ciência do locutor, é impossível solicitar que esse pronuncie uma determinada frase específica.

A maior vantagem dos sistemas RAL dependentes do texto é que a duração da amostra de áudio questionada pode ser menor que nos sistemas independentes do texto. Nesses últimos, como é necessário modelar toda a voz do locutor, e não apenas aquelas realizações do texto predefinido, são necessárias amostras mais longas, com mais realizações vocais, a fim de permitir uma modelagem global da voz do locutor.

Os sistemas RAL independentes do texto são capazes de trabalhar com amostras de voz independentemente de seu conteúdo textual. Esses sistemas permitem uma maior flexibilidade e se adéquam propriamente a aplicações forenses e investigativas, em que não se po-

de esperar pela pronuncia de uma frase específica nas amostras questionadas. Por aceitarem qualquer conteúdo textual, esses sistemas exigem que as amostras de voz sejam relativamente longas, aproximadamente 30 segundos (Reynolds e Rose, 1995). Com amostras mais longas, tem-se uma grande quantidade de sons produzidos, permitindo uma melhor caracterização acústica da voz, e uma modelagem mais completa da voz do locutor.

Existem diversas técnicas de RAL independente do texto. Algumas, mais simples e já em desuso, utilizam apenas as médias de parâmetros como frequências formantes ou frequência fundamental (Furui, Itakura e Saito, 1972; Furui, 1974; Markel, Oshika e Gray, 1977; Markel e Davis, 1979). Outras utilizam vetores de parâmetros quantizados (Soong *et al*, 1985; Buck, Burton e Shore, 1985). As técnicas mais recentemente utilizadas e que têm demonstrado os melhores resultados são a dos Modelos de Mistura de Gaussianas (*Gaussian Mixture Models* - GMM), introduzidos por Reynolds (1992) e a dos Modelos Ocultos de Markov (*Hidden Markov Models* - HMM) (Rosenberg, Lee e Soong, 1990).

Em todos os experimentos utilizados nesse trabalho foram adotados os Modelos de Mistura de Gaussianas, pois os Modelos Ocultos de Markov são computacionalmente mais complexos e não têm apresentado melhorias significativas de desempenho em aplicações independentes do texto se comparados a outros tipos de modelagem (Tisby, 1991; Matsui e Furui, 1992).

2.2 MODELOS DE MISTURA DE GAUSSIANAS

A utilização de distribuições gaussianas para a modelagem de eventos é bastante comum nas ciências em geral. Contudo, esse tipo de modelagem é apropriado para sistemas que possuem um comportamento definido com variações em torno de um ponto médio. A voz humana não pode ser bem modelada por uma única gaussiana, pois, para a produção dos diferentes sons, são alteradas as características do trato vocal. Consequentemente, não existe uma distribuição em torno de um ponto. Contudo, cada classe de sons (fonema) da voz humana pode ser modelada por uma função gaussiana, visto que, dentro de cada uma dessas classes, o trato vocal tem uma forma determinada, com pequenas variações. Assim, pode-se utilizar um conjunto de gaussianas, cada uma modelando uma classe de sons, para modelar a voz de um determinado locutor.

A utilização de Modelos de Mistura de Gaussianas (*Gaussian Mixture Models* - GMM) nos sistemas RAL foi introduzida por Reynolds em 1992. A idéia básica dos GMMs é modelar a voz dos locutores não por uma, mas por um conjunto (mistura) de gaussianas. Dessa forma, um GMM representa uma densidade formada pelo somatório ponderado de *M* densidades gaussianas componentes e pode ser representado matematicamente pela equação:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad , \tag{2.1}$$

em que $\vec{x}$ é um vetor de parâmetros (variáveis) *D*-dimensional, $b_i(\vec{x})$, i = 1, ..., *M*, são as *M* densidades gaussianas que compõem o modelo, também chamadas de componentes da mistura, e p_i , i = 1, ..., *M*, são os pesos de cada componente, também denominados de coeficientes de mistura.

Cada densidade componente do GMM, $b_i(\vec{x})$, é uma densidade gaussiana *D*-dimensional de forma

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{ -\frac{(\vec{x} - \vec{\mu}_i)^T \, \boldsymbol{\Sigma}_i^{-1} (\vec{x} - \vec{\mu}_i)}{2} \right\} \quad , \tag{2.2}$$

com vetor de média $\vec{\mu}_i$ e matriz de covariância $\boldsymbol{\Sigma}_i$.

Os pesos das componentes da mistura são propriamente normalizados de forma que

$$\sum_{i=1}^{M} p_i = 1 \tag{2.3}$$

Na equação (2.1), λ representa a descrição completa de um GMM, incluindo suas médias, pesos e matriz de covariância.

$$\lambda = \{ p_i, \vec{\mu}_i, \boldsymbol{\Sigma}_i \}, i = 1, ..., M \tag{2.4}$$

A vantagem da utilização de GMM, e não de simples modelos gaussianos, pode ser verificada nas ilustrações a seguir. Alguns fenômenos, muitos na realidade, são tais que seus resultados têm uma distribuição simples, aproximadamente normal, em torno de um valor médio, como ilustrado no gráfico da Figura 2.1:

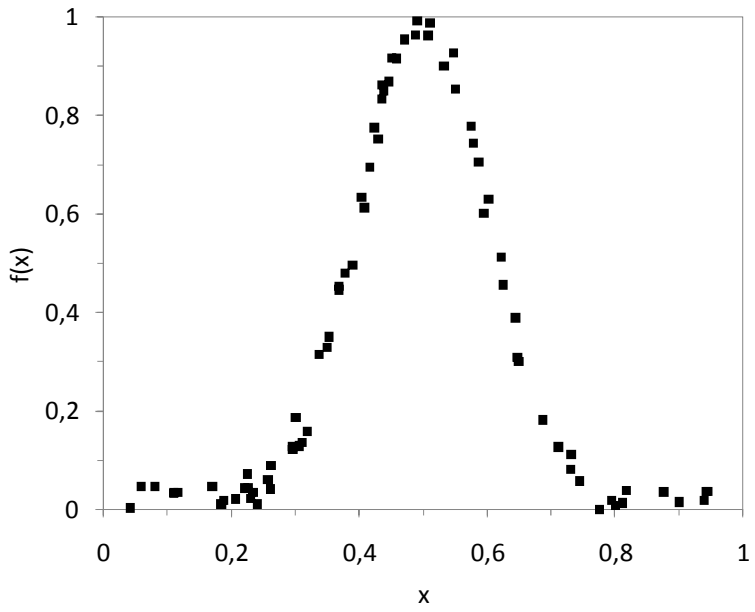

Figura 2.1: Distribuição normal de valores.

Nesses casos, a modelagem do processo por uma única gaussiana é suficiente para aproximar bem os resultados obtidos, como se pode verificar na Figura 2.2.

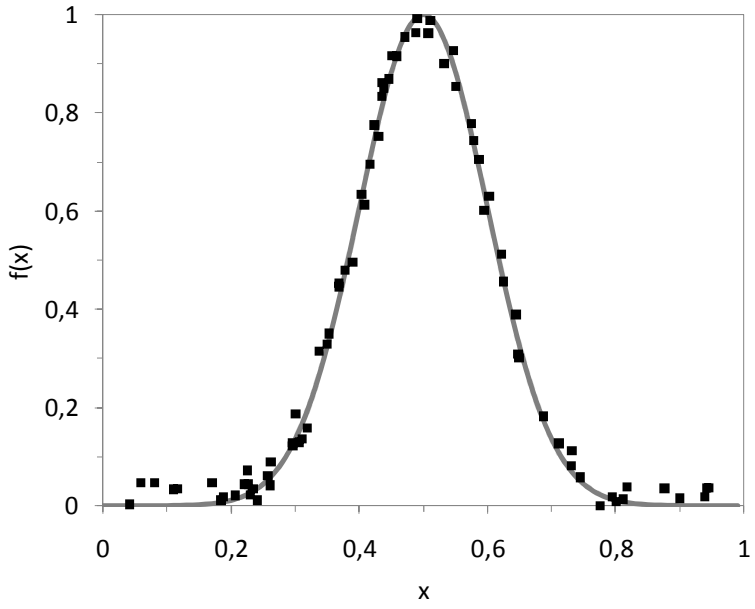

Figura 2.2: Distribuição normal de valores e aproximação por modelo de uma gaussiana.

Em outros casos, contudo, como no caso da voz, o sistema é, na realidade, uma composição de sistemas distintos (uma configuração do trato vocal para cada fonema). Nesses casos, a tentativa de modelar o sistema por uma única gaussiana não produz resultados satisfatórios, como demonstrado na Figura 2.3:

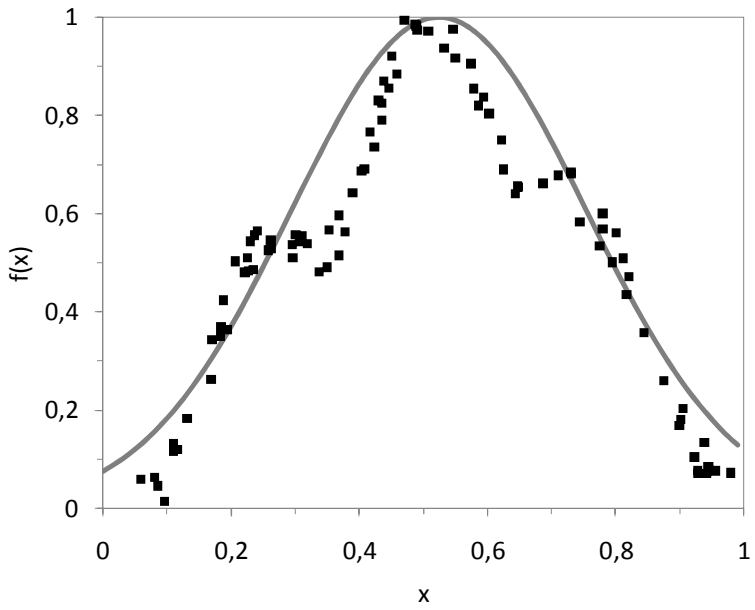

Figura 2.3: Distribuição composta de valores e aproximação por modelo de uma gaussiana.

Para processos desse tipo, a aproximação por uma soma ponderada de distribuições gaussianas promove resultados muito melhores.

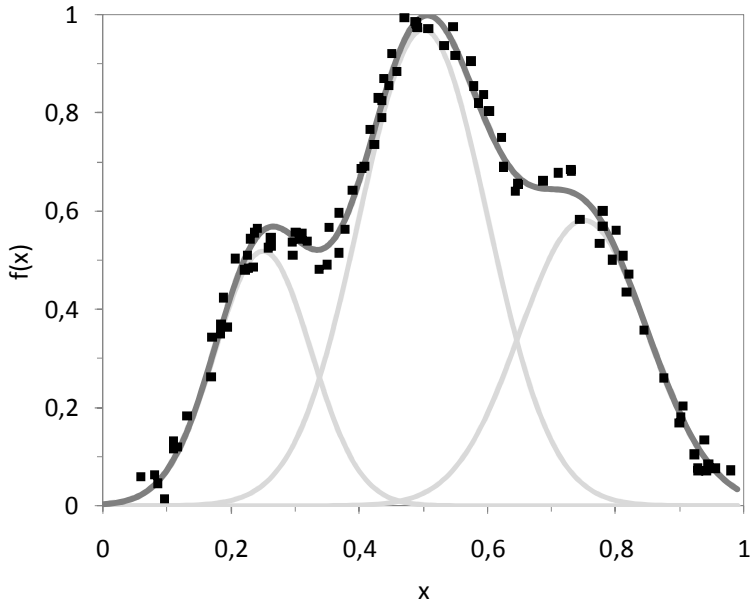

Figura 2.4: Distribuição composta de valores e aproximação por GMM (exibindo também as três componentes gaussianas individualmente).

Na Figura 2.4, foram utilizadas três gaussianas distintas para modelar os dados do processo fictício f(x). Embora o exemplo utilizado nessa ilustração seja extremamente simplificado, tanto por tratar de um processo unidimensional quanto por apresentar apenas três compo-

nentes, ele é suficiente para demonstrar como a modelagem gaussiana torna-se inadequada para processos mais complexos como a voz humana.

Nos sistemas RAL, a voz de cada locutor é modelada por um GMM distinto, dando origem a um modelo λ_s, $s = 1, ..., S$; sendo S o número total de locutores modelados. O universo de locutores modelados é representado por U:

$$U = \{\lambda_s, s = 1, ..., S\} \tag{2.5}$$

Em geral, para simplificação dos cálculos, as matrizes de covariância, Σ_i, são feitas diagonais, transformando-se em simples vetores de variâncias, $\vec{\sigma}_i$. Experimentos realizados por Reynolds (2000) demonstraram que essa simplificação do modelo não causa perda de desempenho na identificação.

Ao se restringir as matrizes de covariâncias a matrizes diagonais, ignora-se o efeito da correlação entre os diferentes parâmetros das gaussianas D-dimensionais. Com isso, as hipersuperfícies de probabilidade constante das gaussianas são, obrigatoriamente, elipsóides com seus eixos principais alinhados aos eixos coordenados do espaço de parâmetros. Os gráficos da Figura 2.5 à Figura 2.7 ilustram, de forma simplificada, para um caso bidimensional, o efeito da restrição de diagonalidade da matriz de covariância do modelo gaussiano multidimensional.

Na Figura 2.5, pode-se observar uma distribuição de dados num espaço bidimensional. Os pontos se distribuem de maneira normal nas duas dimensões, sem correlação. Em situações como essa, a modelagem dos dados por uma gaussiana bidimensional irá resultar num modelo com matriz de covariância diagonal, de forma que os eixos das elipses de probabilidade constante do modelo são alinhados aos eixos coordenados do espaço.

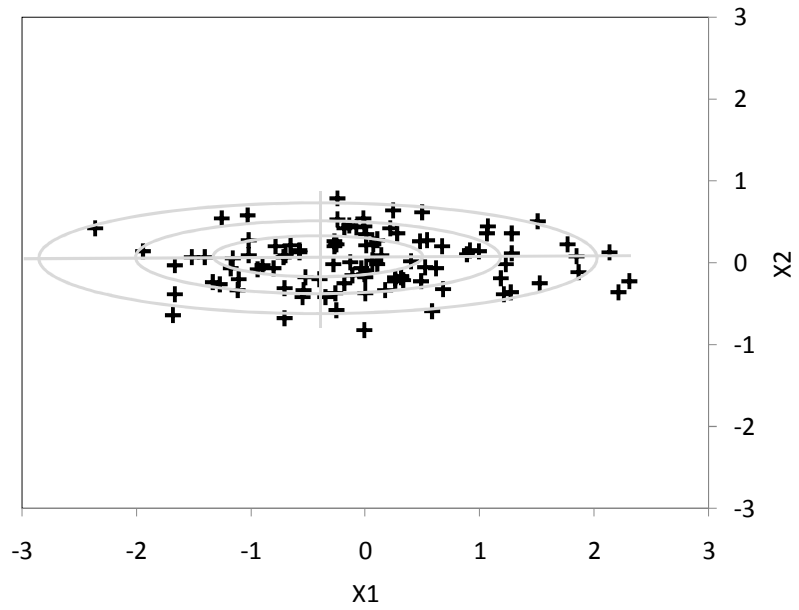

Figura 2.5: Distribuição bidimensional sem correlação e elipses de probabilidade constante de modelo gaussiano bidimensional com matriz de covariância diagonal.

A Figura 2.6 contém outra distribuição de pontos num espaço bidimensional. Dessa vez, os pontos distribuem-se de modo normal, mas há, claramente, uma correlação positiva entre os valores das variáveis *X1* e *X2*. Para representar corretamente esse tipo de distribuição de valores por um modelo gaussiano bidimensional, é necessário utilizar uma matriz de covariância completa, de modo a permitir a rotação dos eixos das elipses de probabilidade constante.

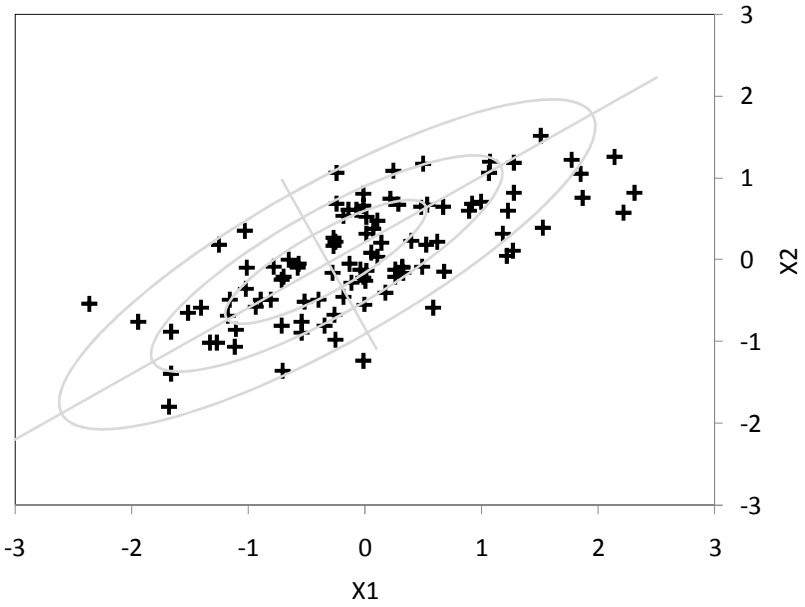

Figura 2.6: Distribuição bidimensional com correlação e elipses de probabilidade constante de modelo gaussiano bidimensional com matriz de covariância completa.

Se, para o mesmo conjunto de dados ilustrado no gráfico da Figura 2.6, for tentada uma modelagem gaussiana bidimensional com matriz de covariância diagonal, se estará impondo ao modelo que mantenha os eixos das elipses de probabilidade constante alinhados com os eixos coordenados, como se pode observar no gráfico da Figura 2.7.

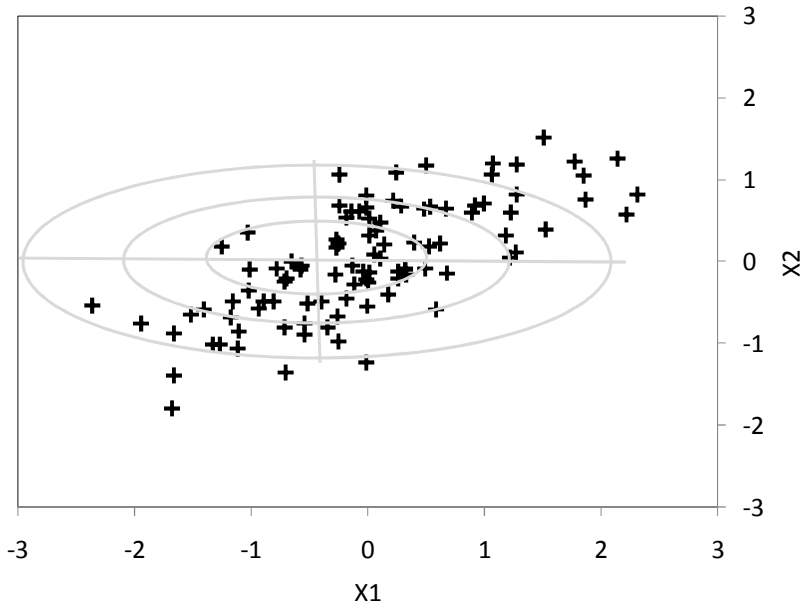

Figura 2.7: Distribuição bidimensional com correlação e elipses de probabilidade constante de modelo gaussiano bidimensional com matriz de covariância diagonal.

Essa restrição, certamente, limita a representação dos dados a um modelo menos fiel. Contudo, como se está trabalhando com GMM, não com gaussianas únicas, é possível utilizar mais de uma gaussiana para representar os dados, contornando a limitação imposta pela restrição da matriz de covariância, como se pode observar no gráfico da Figura 2.8.

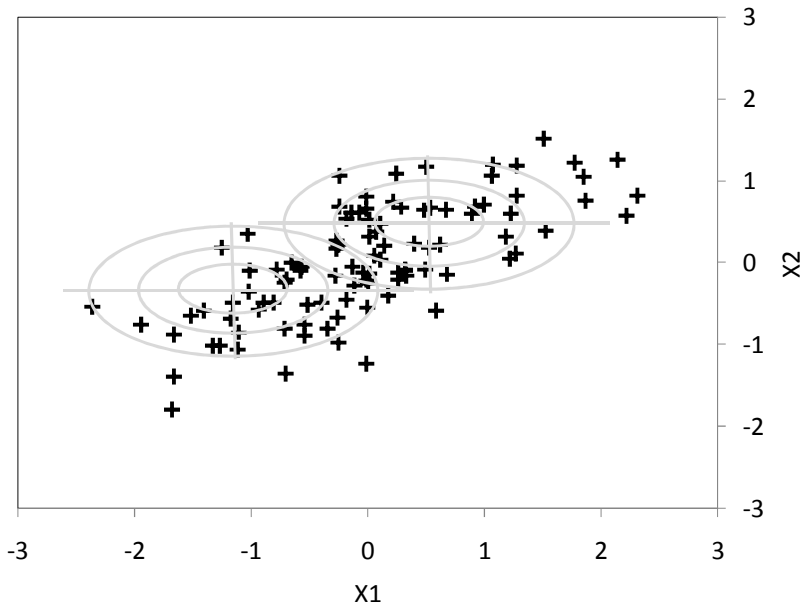

Figura 2.8: Distribuição bidimensional com correlação e elipses de probabilidade constante das componentes de GMM bidimensional com matriz de covariância diagonal.

2.2.1 Treinamento dos Modelos

O primeiro passo para a construção de um RAL baseado em GMM é o treinamento dos modelos. Para esse treinamento é necessário, para cada locutor a ser modelado, um arquivo de áudio contendo gravações de sua voz; esses arquivos são chamados de arquivos de treinamento. Para cada arquivo de treinamento, são calculados diversos vetores de parâmetros, $\vec{x}_t$, para diferentes instantes de tempo[*], *t*. O conjunto desses vetores de parâmetros de treinamento para um determinado locutor *s* é representado por:

$$X_s = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_T\} \tag{2.6}$$

Em (2.6), para simplificar a notação, não foi adicionado, no lado direito, o índice *s* indicativo do locutor.

O objetivo do treinamento do GMM é ajustar os parâmetros do modelo, λ_s, de forma a maximizar a verossimilhança $p(X_s \mid \lambda_s)$. Para simplificar o problema, considera-se que cada vetor de parâmetros, $\vec{x}_t$, é independente dos demais, de forma que é possível escrever:

[*] Neste trabalho, será utilizada a representação *t* para indicar o índice da janela da análise em questão. Esse índice, portanto, não representa a variável tempo contínuo, mas pode ser relacionado com ela, desde que se multiplique o valor de *t* pelo tamanho do passo das janelas de análise. Dessa maneira, a expressão "diferentes instantes de tempo", utilizada no texto, deve ser compreendida como "diferentes intervalos de amostras de áudio, com inícios nos instantes *k.t*, *k = 0, 1, ..., K,* e de comprimento (*t*) igual ao comprimento da janela de análise".

$$p(X_s \mid \lambda_s) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda_s) \tag{2.7}$$

A equação (2.7) é uma função não-linear dos parâmetros do modelo λ_s, o que impede uma maximização direta. Em geral, a maximização de (2.7) é realizada com o algoritmo *expectation-maximization* (EM) descrito por Dempster, Laird e Rubin (1977).

O algoritmo EM funciona em duas etapas distintas. Na primeira etapa, chamada de etapa E (*expectation*), são calculados os valores de $b_i(\vec{x}_t)$ de cada uma das componentes do GMM, para cada um dos vetores de parâmetros; esses valores são comumente denominados de ativações das componentes. Na segunda etapa, chamada de etapa M (*maximization*), o modelo é atualizado da seguinte forma:

$$\begin{aligned}
p_i^{[n+1]} &= \frac{1}{T}\sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda_s^{[n]}\right) \\
\vec{\mu}_i^{[n+1]} &= \frac{\sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda_s^{[n]}\right)\vec{x}_t}{\sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda_s^{[n]}\right)} \\
\vec{\sigma}_i^{2\,[n+1]} &= \frac{\sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda_s^{[n]}\right)\vec{x}_t^{2}}{\sum_{t=1}^{T} p\left(i \mid \vec{x}_t, \lambda_s^{[n]}\right)} - \vec{\mu}_i^{2\,[n]}
\end{aligned} \quad , \tag{2.8}$$

sendo que os índice entre colchetes indicam se se trata dos parâmetros atualizados (ciclo de treinamento *n+1*) ou dos parâmetros antigos (ciclo de treinamento *n*).

Além disso, em (2.8), *I* representa a matriz identidade, $\vec{\sigma}_i$ representa um vetor de variâncias que compõe a diagonal de Σ_i (os demais coeficientes são zero, se for utilizada matriz diagonal) e

$$p\left(i \mid \vec{x}_t, \lambda_s^{[n]}\right) = \frac{p_i^{[n]} b_i^{[n]}(\vec{x}_t)}{\sum_{k=1}^{M} p_k^{[n]} b_k^{[n]}(\vec{x}_t)} \tag{2.9}$$

Com essas atualizações, a cada passo do algoritmo EM, a verossimilhança expressa em (2.7) é maior que no passo anterior.

$$p\left(X_s \mid \lambda_s^{[n+1]}\right) \geq p\left(X_s \mid \lambda_s^{[n]}\right) \tag{2.10}$$

A finalização do treinamento é realizada, em geral, quando se atinge um número máximo de ciclos ou quando o valor de (2.7) se estabiliza (de acordo com algum critério especifica-

do). Na prática, de dez a vinte ciclos de treinamento são suficientes para a estabilização do modelo.

Deve-se destacar que a necessidade da utilização do algoritmo EM, ou de outro semelhante, decorre das características próprias dos GMM. Como se trata de um modelo composto por elementos distintos (componentes gaussianas), antes de se realizar o ajuste dos parâmetros para a maximização de $p(X_s \mid \lambda_s)$, é necessário determinar a que elemento componente do GMM o vetor $\vec{x}_t$ pertence (ou determinar em que grau esse vetor pertence a cada um dos elementos, como ocorre na formulação da equação (2.9))

Outra questão importante com relação ao treinamento é a forma de inicialização dos modelos[*]. Não existe um procedimento ideal para o arbitramento das condições iniciais da modelagem e, como em qualquer otimização não linear, uma escolha inadequada dessas condições pode levar o modelo a ficar "preso" em um máximo local muito distante do máximo global, ocasionando uma representação deficiente. Neste trabalho, todos os modelos foram inicializados utilizando:

$$p_i^{[0]} = \frac{1}{M}, i = 1,...,M \tag{2.11}$$

Os vetores de médias, $\vec{\mu}_i$, foram inicializados com valores aleatórios escolhidos entre os limites[†] dos valores observados no conjunto de dados de treinamento, X_s. Os vetores de variâncias, $\vec{\sigma}_i$, foram inicializados com valores aleatórios escolhidos entre os limites dos valores das distâncias entre os pontos médios iniciais, $\vec{\mu}_i^{[0]}$, e os pontos do conjunto de dados de treinamento, X_s.

2.2.2 Identificação do Locutor

Para identificar o locutor a quem pertence a voz em um arquivo de teste é necessário determinar qual dos modelos, λ_s, do universo, U, apresenta a maior probabilidade *a posteriori* para o conjunto de parâmetros calculado a partir do arquivo questionado dado:

[*] Veja como a inicialização pode afetar o modelo final no caso do Método do Treinamento Progressivo Negativo, seção 4.4.1.

[†] Foram considerados os limites individuais para cada coordenada, tanto no caso da inicialização das médias como no caso da inicialização das variâncias.

$$\tilde{s} = \arg\max_{\lambda_s \in U} p(\lambda_s \mid Y) = \arg\max_{\lambda_s \in U} \frac{p(Y \mid \lambda_s)\, p(\lambda_s)}{p(Y)} \quad , \tag{2.12}$$

sendo Y o conjunto de vetores de parâmetros calculados a partir do arquivo de teste, $Y = \{\vec{y}_1, \vec{y}_2, ..., \vec{y}_T\}$, e utilizando, na última passagem, a regra de Bayes.

Supondo que todos os locutores são igualmente prováveis, e observando que $P(Y)$ depende unicamente do áudio testado (e, portanto, é o mesmo para todos os locutores do universo), conclui-se que a identificação do locutor pode ser realizada simplesmente calculando:

$$\tilde{s} = \arg\max_{\lambda_s \in U} p(Y \mid \lambda_s) \tag{2.13}$$

Admitindo a independência entre os elementos do vetor de parâmetros de teste, tal como formulada em (2.7) para os parâmetros de treinamento, e utilizando o logaritmo, o cálculo de (2.13) é realizado por:

$$\tilde{s} = \arg\max_{\lambda_s \in U} \sum_{t=1}^{T} \log p(\vec{y}_t \mid \lambda_s) \tag{2.14}$$

O uso do logaritmo é um artifício para evitar problemas numéricos, pois os valores envolvidos são muito baixos.

Considera-se que a identificação foi correta quando o locutor que maximiza a expressão (2.14), $\tilde{s}$, é, de fato, o locutor correto $\hat{s}$, ou seja, quando:

$$\tilde{s} = \hat{s} \tag{2.15}$$

2.3 SELEÇÃO DE PARÂMETROS DE MODELAGEM

A utilização dos GMM ou de qualquer outra técnica de modelagem para representar os locutores do banco de dados exige que os modelos sejam treinados com conjuntos de parâmetros extraídos das vozes desses locutores, como visto na seção 2.2.1. Da mesma forma, depois de treinados os modelos, a identificação de uma voz questionada é realizado com base na utilização de parâmetros extraídos dessa voz, como descrito na seção 2.2.2. Observa-se, portanto, que o tipo de parâmetro a ser utilizado no sistema de RAL é fundamental para o desempenho do sistema.

Experimentos com diversos tipos de parâmetros na modelagem dos locutores têm sido realizados desde o início do desenvolvimento das técnicas de RAL por Luck (1969) e Sambur (1972). Comparações visando determinar o melhor tipo de parâmetro para representar as características da voz de um locutor foram feitas desde esse período por Wolf (1972), e Atal (1974) e Sambur (1978), demonstrando a relevância de uma boa escolha desses parâmetros para o desempenho global do sistema de RAL.

Atualmente, um dos tipos de parâmetros mais frequentemente utilizados nos sistemas de RAL são os coeficientes cepstrais ou sua variante os coeficientes mel-cepstrais (*Mel Frequency Cesptral Coefficients* - MFCC), como detalhado em Reynolds, Quatieri e Dunn (2000). Os coeficientes mel-cepstrais, especialmente quando calculados diretamente por bancos de filtros, tem se mostrado como um tipo de parâmetro eficiente para aplicações de RAL; ao contrário do que ocorre se calculados a partir de modelos de predição linear, quando apresentam grande vulnerabilidade ao ruído (Tierney, 1980). Como se sabe, o espectro da voz humana apresenta muitos picos, decorrentes dos pulsos glóticos (para sons vozeados ou sonoros), além de uma curva envoltória suave, resultado da modulação do espectro dos pulsos glóticos pelas ressonâncias e anti-ressonâncias do trato vocal do locutor. Nas aplicações de RAL, o interesse principal da utilização dos coeficientes cepstrais está na sua capacidade de caracterização dessa envoltória (Childers, Skinner e Kemerait, 1977), pois ela está diretamente relacionada com as formas (comprimentos, larguras) dos diversos segmentos do trato vocal dos locutores. Assim, através de informações da envoltória do espectro de voz, é possível caracterizar a forma do trato vocal e, desse modo, individualizar um locutor.

Contudo, recentemente, tentando explorar a relação entre as características biométricas do trato vocal dos locutores e suas vozes, Souza e Souza (2001) realizaram um trabalho demonstrando bons resultados em sistemas de RAL que utilizam parâmetros intimamente relacionados com essa biometria do trato vocal. As simulações realizadas levaram a concluir que parâmetros como as relações entre as áreas dos tubos do trato vocal e os coeficientes de reflexão apresentam desempenho superior aos coeficientes cepstrais comumente utilizados. Os experimentos realizados por Souza e Souza (2001), entretanto, utilizaram um universo com número de locutores reduzido, apenas cinco, as simulações foram realizadas utilizando arquivos de treinamento e de teste de durações curtas (5 segundos) e com os mesmos conteúdos textuais para treino e teste (apesar de serem realizações vocais diferentes, o texto lido era o mesmo). Devido a essas restrições, a fim de confirmar a validade das

conclusões apresentadas e de utilizar o tipo de parâmetro que apresenta melhores resultados, foi realizada uma longa série de avaliações comparativas de desempenho de sistemas de RAL utilizando diversos tipos de parâmetros.

2.3.1 Descrição do Banco de Dados

Para realizar as avaliações a fim de verificar o melhor tipo de parâmetro para aplicações de RAL, foi utilizado um banco de dados de vozes contendo 30 locutores distintos, sendo 15 do sexo masculino e 15 do sexo feminino ($S = 30$). Cada locutor foi gravado lendo um texto pré-definido. A fim de reproduzir as condições do experimento de Souza e Souza (2001) foi utilizado o mesmo texto para todos os locutores[*]. Posteriormente, cada uma dessas gravações foi fracionada em 21 arquivos; os pontos de início e final dos cortes dos arquivos são os mesmo para todos os locutores. Dessa maneira, foram gerados 21 arquivos para cada locutor, que serão indicados por $A_{n,s}$, onde s indexa o locutor e n indexa o trecho do arquivo originalmente gravado. Como os pontos de corte dos arquivos foram os mesmos com relação ao conteúdo textual lido, tem-se que os arquivos $A_{k,s}$, k = cte. e $s = 1, ..., S$ contêm as vozes dos 30 locutores lendo um mesmo trecho do texto.

Essa forma de divisão dos arquivos foi escolhida por duas razões. Primeiramente, para que todos os trechos tivessem uma duração aproximadamente igual (a duração média aproximada dos trechos é de 30 segundos). Além disso, como o objetivo é avaliar o desempenho de diferentes parâmetros na identificação de locutores, todas as simulações foram realizadas utilizando trechos de mesmo conteúdo textual. Dessa forma, tentou-se minimizar a influência que diferentes conteúdos textuais dos arquivos poderiam ter na identificação do locutor[†].

Todas as gravações foram realizadas em ambientes acusticamente preparados, com microfones e placas de captura de áudio profissionais. Os arquivos foram gerados a uma taxa de amostragem de 22 kHz, com quantização de 16 bits por amostra, em modo monaural.

---

[*] Destaque-se que, apesar dessa restrição, o processo continua sendo de reconhecimento independente do texto. Em que pese que todos os locutores tenham sido gravados lendo um mesmo texto, foram utilizados trechos distintos do áudio (com conteúdo textual diferente) para os procedimentos de treinamento dos modelos e de teste. De fato, a necessidade de se utilizar o mesmo texto para todos os locutores somente se justifica para possibilitar a aplicação da condição adicional de sucesso descrita na equação (2.17).

[†] Essa preocupação tem papel especial no caso particular porque os experimentos de Souza e Souza (2001) tratavam a identificação correta de uma forma mais restritiva que a habitual, como detalhado na equação (2.19). Dessa forma, visando reproduzir tanto quanto possível as condições desse trabalho, utilizou-se esse mesmo critério.

Todos os arquivos de áudio $A_{n,s}$ foram pré-processados da seguinte forma. Inicialmente, todos os arquivos foram normalizados de forma que sua amplitude de pico correspondesse a 100% do valor máximo de quantização. Posteriormente, foram excluídos os trechos de silêncio dos arquivos. Essa eliminação foi realizada com o uso de um detector automático de silêncio baseado na medida da energia do sinal em janelas de 20 ms, com sobreposição de 15 ms (avanços de 5 ms) e limiar de silêncio definido manualmente (um único valor para todos os arquivos) de forma a otimizar o processo. Por fim, todos os arquivos foram submetidos a um filtro de pré-ênfase do tipo

$$y'[k] = y[k] - 0{,}95\,y[k-1], \tag{2.16}$$

sendo $y[k]$ o valor da amostra de áudio original de índice *k*, e $y'[k]$ o valor da amostra de áudio filtrada de índice *k*.

2.3.2 Figuras de Mérito

As simulações computacionais para avaliação dos diferentes tipos de parâmetros foram realizados utilizando sistemas de RAL idênticos (GMMs com 16 componentes, $M = 16$, e matrizes de covariância, Σ, diagonal, como discutido na seção 2.2), alterando apenas o tipo de parâmetro utilizado para a modelagem dos locutores. Foram avaliados com os seguintes parâmetros:

a) coeficientes mel-cepstrais;

b) coeficientes cepstrais;

c) coeficientes de reflexão;

d) relação entre áreas;

e) coeficientes de áreas;

f) logaritmo das relações entre áreas;

g) logaritmo dos coeficientes de áreas;

h) coeficientes de autocorrelação;

i) coeficientes de predição linear;

j) coeficientes de autocorrelação do filtro inverso;

k) pares de linhas espectrais (*Line Spectrum Pairs* - LSP)

A escolha dos parâmetros avaliados seguiu basicamente aqueles examinados no trabalho de Souza e Souza (2001).

O treinamento dos modelos, λ_s, foi limitado a 50 ciclos. Em geral, após 10 ciclos de treinamento não há melhoras apreciáveis na modelagem; contudo, como essa etapa exige um tempo de processamento comparativamente pequeno, optou-se por realizar mais ciclos de treinamento a fim de garantir a boa adequação dos modelos aos dados. Para o treinamento dos modelos λ_s foram utilizados os arquivos $A_{21,s}$ do banco de dados. Os demais arquivos $A_{n,s}$, $n = 1, ..., 20$, foram utilizados para as identificações, de modo que, para cada tipo de parâmetro, foram realizados 20 procedimentos de identificação para cada um dos 30 locutores do banco de dados, totalizando 600 procedimentos de identificação por parâmetro avaliado.

O cálculo de todos os parâmetros foi realizado utilizando o pacote de ferramentas Voicebox para Matlab, desenvolvido por Brookes (2001). Os parâmetros mel-cepstrais foram calculados diretamente através de banco de filtros, como descrito em Reynolds e Rose (1995), sendo usados 12 coeficientes. Todos os demais parâmetros foram calculados a partir dos coeficientes de predição linear (LPC), sendo usados doze coeficientes ($D = 12$), computados em janelas de 20 ms de duração sem sobreposição.

Para determinar se a identificação do locutor foi bem sucedida, visando reproduzir os experimentos da melhor maneira possível, adotou-se o critério definido em Souza e Souza (2001). Segundo esse critério, não basta apenas que a voz testada produza a máxima verossimilhança para o locutor correto, como definido na equação (2.13). É preciso ainda que, para vozes de todos os locutores, o modelo apresente o valor máximo de verossimilhança para o locutor correto:

$$\underset{\sim}{s} = \arg \max_{1 \le r \le S} p\left(Y_r \mid \lambda_s\right), \tag{2.17}$$

sendo $Y_r = \{\vec{y}_1, \vec{y}_2, ..., \vec{y}_T\}$ o conjunto de parâmetros extraído do arquivo de teste $A_{n,r}$.

Como, em (2.17), são comparados diferentes arquivos (diferentemente do que ocorre em (2.13)), é necessário normalizar a expressão com relação ao tempo para compensar variações de duração nos arquivos testados:

$$\underset{\sim}{s} = \arg \max_{1 \le r \le S} \frac{\sum_{t=1}^{T} \log p\left(\vec{y}_{r,t} \mid \lambda_s\right)}{T}, \tag{2.18}$$

Com base no critério estabelecido, considera-se uma identificação correta apenas se

$$\tilde{s} = \hat{s} = \underset{\sim}{s} \tag{2.19}$$

Destaque-se que essa segunda condição imposta (igualdade à direita da equação (2.19)) torna a identificação correta mais difícil, porém mais robusta. Na realidade, não se pode considerar a equação (2.17) como propriamente uma condição de identificação. Trata-se, na realidade, de uma condição de robustez do modelo contra falsos locutores, pois avalia o modelo de um locutor, λ_s, contra as vozes de todos os locutores, Y_r, $r = 1,...,S$.

2.3.3 Resultados

Os resultados dos procedimentos de identificação com sistemas baseados nos diversos parâmetros foram organizados na Tabela 2.1. Nessa tabela, a primeira coluna indica o tipo de parâmetro utilizado para a modelagem dos locutores e a segunda coluna contém o percentual de identificações corretas obtidas, segundo o critério definido na equação (2.19)

Tabela 2.1: Desempenho do sistema de RAL para os diferentes tipos de parâmetros da voz utilizados na modelagem dos locutores.

Tipo de Parâmetro da Voz	**Taxa de Identificações Corretas (%)**
Coeficientes Cepstrais	95,8
Coeficientes Mel-Cepstrais	95,5
Pares de Linhas Espectrais	94,8
Coeficientes de Reflexão	89,3
Logaritmo das Relações entre Áreas	87,5
Relação entre Áreas	83,9
Logaritmo dos Coeficientes de Áreas	75,3
Coeficientes de Predição Linear	57,0
Coeficientes de Áreas	31,9
Coeficientes de Autocorrelação do Filtro Inverso	19,0
Coeficientes de Autocorrelação	16,4

Os resultados exibidos na Tabela 2.1 demonstram que a utilização de parâmetros cepstrais, quer sejam os coeficientes cepstrais ou os mel-cepstrais, proporcionam a maior quantidade de identificações corretas, 95,8% e 95,5%, respectivamente. Alguns parâmetros mais intimamente relacionados com a biometria do trato vocal, como os coeficientes de reflexão, a relação entre áreas e o logaritmo da relação entre áreas também apresentaram bom desempenho, identificando corretamente 89,3%, 87,5% e 83,9% dos locutores, respectivamente.

Esse resultado permite concluir que, em situações de maior variabilidade no conteúdo textual[*] e quando existe um maior número de locutores a serem comparados (neste caso, 30 locutores), os parâmetros cepstrais realmente apresentam o melhor desempenho em sistemas de RAL. Por essa razão, e tendo em vista que o uso dos coeficientes mel-cepstrais é o atual padrão nos sistemas de RAL, todos os demais sistemas de identificação deste trabalho utilizarão esses parâmetros para a modelagem dos locutores.

Outra conclusão interessante, não prevista antes dos experimentos, foi o excelente resultado apresentado no sistema utilizando os pares de linhas espectrais. O percentual de identificações corretas nesse caso atingiu 94,8%, apenas 1% inferior ao obtido com os coeficientes cepstrais e 0,7% inferior ao obtido com os coeficientes mel-cepstrais. O bom desempenho do sistema baseado no LSP, apesar de não ter sido previsto anteriormente, não é de todo inesperado. Esses parâmetros possuem propriedades muito interessantes para sistemas de RAL, dentre as quais se destaca sua pequena sensibilidade (pequenas alterações na voz modelada provocam pequenas alterações nos parâmetros), como demonstrado em Bultheel (1984). Essa característica é indispensável em sistemas RAL que utilizam os GMM, pois esses modelos admitem como hipótese implícita que a distribuição dos parâmetros se dá de maneira gaussiana, ou seja, que seus valores se distribuem de forma contínua em torno de um valor médio. Parâmetros como os coeficientes de predição linear e os coeficientes de autocorrelação (normais ou do filtro inverso) que apresentam grande sensibilidade a pequenas variações na composição espectral da voz (Tierney, 1980) apresentaram desempenho insatisfatório, como esperado.

[*] Nesse contexto, a maior variabilidade no conteúdo textual significa que as frases utilizadas pelos locutores nos procedimentos de treinamento dos modelos e de testes eram temporalmente mais longas e com mais realizações vocais que as utilizadas originalmente por Souza e Souza (2001).

3 TÉCNICAS ROBUSTAS DE RECONHECIMENTO AUTOMÁTICO DE LOCUTOR

Os sistemas de reconhecimento automático de locutor (RAL), como visto na seção 2.2.2, baseiam-se na maximização da verossimilhança dos modelos dos locutores com relação aos parâmetros extraídos de um trecho de voz questionado (áudio de teste), de acordo com a equação (2.13). Devido a essa formulação, é de se esperar que, à medida que o áudio questipnado se degrada, modificando suas características com relação ao áudio utilizado para o treinamento dos modelos, haja uma progressiva diminuição do desempenho da identificação.

Dessa maneira, embora os Modelos de Mistura de Gaussianas (GMM) permitam uma representação dos locutores que possibilita o desenvolvimento de sistemas de RAL com alto desempenho de reconhecimento, algumas situações podem comprometer ou mesmo inviabilizar as identificações dos locutores. Dois fatores que ocorrem frequentemente em situações práticas prejudicando os resultados dos sistemas de RAL são a presença de ruído e a codificação do áudio em formato MP3 a baixas taxas.

3.1 SENSIBILIDADE DOS SISTEMAS DE RAL AO RUÍDO E À CODIFICAÇÃO MP3

3.1.1 Sensibilidade ao Ruído

Para demonstrar a sensibilidade do desempenho dos sistemas de RAL baseados em GMM a variações na relação sinal-ruído (SNR) foi realizado o seguinte procedimento. Inicialmente, os modelos dos locutores foram treinados utilizando como dados de treinamento os arquivos $A_{21,s}$ do banco de dados original descrito na seção 2.3.1. Para simplicidade de notação, será utilizada a representação Φ_0 para designar o banco de dados original descrito na seção 2.3.1.

O sistema de RAL foi elaborado com modelos GMM de 16 componentes gaussianas e utilizado para a modelagem da voz dos locutores parâmetros mel-cepstrais[*]. Foi ainda utilizada, tanto na etapa de treinamento quanto na de identificação, a técnica de normalização por remoção do valor médio dos parâmetros descrita em Reynolds e Rose (1995), visto que

[*] Calculados como descrito na seção 2.3.2.

esse procedimento demonstrou proporcionar ganhos significativos no desempenho do sistema.

Os modelos foram inicializados conforme descrito na seção 2.2.1 e foram treinados por 50 ciclos. Após o treinamento dos modelos, passaram a ser realizadas as identificações segundo o método descrito na seção 2.3.2, com a diferença de que foi utilizada como condição de sucesso (identificação correta) apenas a expressa nas equações (2.14) e (2.15), repetidas a seguir nas equações (3.1) e (3.2) por conveniência:

$$\tilde{s} = \arg\max_{\lambda_s \in U} \sum_{t=1}^{T} \log p\left(\vec{y}_t \mid \lambda_s\right) \tag{3.1}$$

Portanto, considerou-se que a identificação foi correta quando o locutor que maximiza a expressão (3.1), $\tilde{s}$, é, de fato, o locutor correto $\hat{s}$, ou seja, quando:

$$\tilde{s} = \hat{s} \tag{3.2}$$

Outra diferença fundamental implementada para a avaliação presente foi a utilização não apenas do banco de dados de vozes original, Φ_0, mas também de versões desse banco de vozes geradas partir do áudio "sem ruído" (áudio originalmente captado, Φ_0) pela adição computacional de ruído branco (média zero e distribuição normal de amplitudes). Apesar de a injeção de ruído aos sinais ter sido realizada de forma simulada, em ambiente computacional, simulações com sistemas de RAL realizados por Ming *et al* (2007) confirmam que esse tipo de processo aproxima-se bastante da adição física (acústica) de ruído ao áudio no momento de sua captação.

Os valores de SNR utilizados nas simulações foram de 60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB e 10 dB[*], de modo que foram geradas as versões do banco de dados Φ_{50dB}, Φ_{40dB}, Φ_{30dB}, Φ_{26dB}, Φ_{16dB} e Φ_{10dB}. O valor de SNR igual a 60 dB corresponde ao valor de SNR intrínseco do sistema de aquisição de áudio, ou seja Φ_0, = Φ_{60dB}. Esse valor foi estimado a partir da energia média dos sinais nos momentos de silêncio (ausência de voz) e de fala. Assim, o áudio de SNR igual a 60 dB não teve inserção adicional de ruído.

Nos demais casos, a adição de ruído foi feita de modo a se manter uma determinada SNR média ao longo de todo o trecho. Para isso, foi calculada a energia média do sinal, E_s, na amostra de áudio:

[*] Ressalta-se que, em todas as simulações, foi utilizada a frequência de amostragem do banco de vozes original, 22 kHz, e também a codificação de 16 bits por amostra.

$$E_s = \sum_k y_s^2[k] \quad , \tag{3.3}$$

sendo y_s o vetor sinal de áudio e k o índice temporal das amostras.

Foi gerado um vetor ruído, y_n, de mesma dimensão do vetor de sinal de áudio, y_s, contendo amostras com amplitudes distribuídas de maneira gaussiana e média zero, e foi calculada a energia desse vetor de ruído:

$$E_n = \sum_k y_n^2[k] \tag{3.4}$$

Posteriormente, as amplitudes do vetor ruído foram ajustadas de modo a se ter a SNR desejada

$$y'_n[k] = y_n[k].\sqrt{10^{\frac{-SNR}{10}}.\frac{E_s}{E_n}} \tag{3.5}$$

Por fim, o vetor ruído com amplitudes ajustadas foi adicionado ao vetor sinal, gerando um vetor de áudio resultante que passou a ser utilizado

$$y[k] = y_s[k] + y'_n[k] \tag{3.6}$$

Vale destacar que o cálculo da energia do áudio sem ruído foi realizado no áudio após a remoção dos trechos de silêncio, pois essa remoção foi parte do pré-processamento dos sinais. Com isso, a potência média (energia total por tempo total) calculada desse sinal é maior do que seria se fosse considerado o sinal integral, com os trechos de silêncio, visto que os trechos de baixa energia (sem fala) foram excluídos do cálculo. Consequentemente, para a obtenção de uma SNR estabelecida, a potência média do sinal de ruído a ser adicionado ao sinal, y'_n, é maior do que seria se considerado o arquivo de áudio integral (com os trechos de silêncio). Em avaliações realizadas, constatou-se que os valores de SNR indicados nesse trabalho equivalem, em média, a valores cerca de 10% superiores ao que seriam obtidos se o processo de adição de ruído fosse realizado sobre os áudios integrais.

Importante mencionar que a metodologia de medida de SNR utilizada nesse trabalho foi escolhida pelos seguintes motivos: 1) como o ritmo de locução e os tempos de pausas entre palavras e entre frases variam para indivíduos distintos, a medição da SNR global considerando esses tempos de silêncio acarreta a influência dessas características nos resultados finais. Desse modo obtêm-se valores de SNR instantâneo distintos mesmo se a potência média do sinal (nos trechos em que há voz) e do ruído são mantidas fixas; 2) a detecção

dos trechos com fala e dos trechos de silêncio é muito mais simples em áudios sem ruído, onde se pode utilizar um detector de energia simples.

Os resultados dos experimentos realizados para demonstrar a sensibilidade do sistema à variação da SNR do áudio questionado são exibidos na Tabela 3.1, para frequência de amostragem de 22 kHz.

Tabela 3.1: Degradação do desempenho de sistema de RAL devido à diminuição da SNR, para frequência de amostragem de 22 kHz, quantização de 16 bits, e modelos GMM de 16 componentes.

SNR do Áudio de Teste (dB)	**Taxa de Identificações Corretas (%)**
60	100,0
50	100,0
40	99,3
30	93,7
26	83,0
20	54,7
16	28,5
10	13,8
8	12,0
5	10,0

Para melhor visualização, os resultados da Tabela 3.1 foram traçados no gráfico da Figura 3.1.

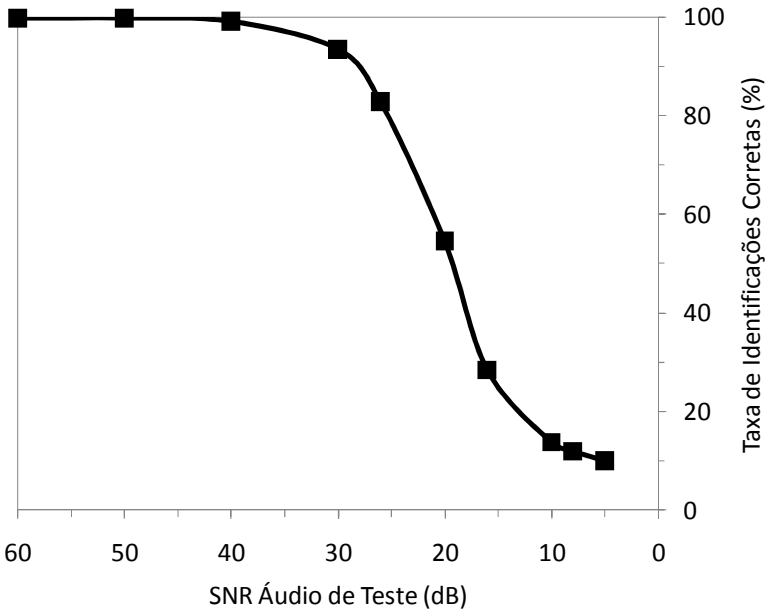

Figura 3.1: Degradação do desempenho de sistema de RAL devido à diminuição da SNR, para frequência de amostragem de 22 kHz, quantização de 16 bits, e modelos GMM de 16 componentes.

Como se observa, o desempenho do sistema se degrada rapidamente com o aumento da SNR do áudio, especialmente após a marca dos 30 dB.

Foram realizados experimentos semelhantes, para determinação da sensibilidade ao ruído, com áudio de outras frequências de amostragem, 16 kHz, 11 kHz e 8 kHz, e também com outras codificações, 8 bits μ-law e 8 bits linear, a fim enriquecer os resultados. A utilização de frequências de amostragem mais baixas leva a uma corrupção maior das frequências ocupadas pela voz, mesmo se mantido o valor da SNR. Isso ocorre porque a energia da voz concentra-se basicamente na faixa até a frequência de 4 kHz e o ruído utilizado tem espectro plano (ruído branco). Dessa maneira, quanto menor a frequência de amostragem, mais concentrada espectralmente estará a energia do ruído e, consequentemente, maior contaminação haverá nas frequências de relevância para a identificação dos locutores. Os experimentos com frequências de amostragem diferentes de 22 kHz foram realizados por meio da reamostragem do áudio original, descrito na seção 2.3.1.

A variação na codificação do áudio, por outro lado, afeta o áudio por alterar os erros de quantização. Dessa forma, a utilização de codificações mais limitadas introduz, no sinal, outra fonte de ruído, além do ruído adicionado diretamente. Deve-se destacar, nesse ponto, que, como forma de evitar o acúmulo de erros de quantização, todo o procedimento de

adição de ruído foi efetuado com aritmética de 32 bits. Somente após a obtenção do áudio ruidoso da equação (3.6), $y[k]$, o áudio foi recodificado para a forma final adequada.

Os resultados dos experimentos não apresentaram grandes variações para qualquer frequência de amostragem ou codificação testada. Em todos os casos, observou-se o padrão semelhante de degradação progressiva de desempenho do sistema. Os resultados para o caso de áudio de frequência de amostragem 8 kHz e quantização de 8 bits linear, caso mais restritivo utilizado, são exibidos na Tabela 3.2.

Tabela 3.2: Degradação do desempenho de sistema de RAL devido à diminuição da SNR, para frequência de amostragem de 8 kHz, quantização de 8 bits linear, e modelos GMM de 16 componentes.

SNR do Áudio de Teste (dB)	**Taxa de Identificações Corretas (%)**
60	97,7
50	97,5
40	97,3
30	92,0
26	81,3
20	54,0
16	32,5
10	14,2
8	12,2
5	9,3

Para melhor visualização, os resultados da Tabela 3.2 foram traçados no gráfico da Figura 3.2.

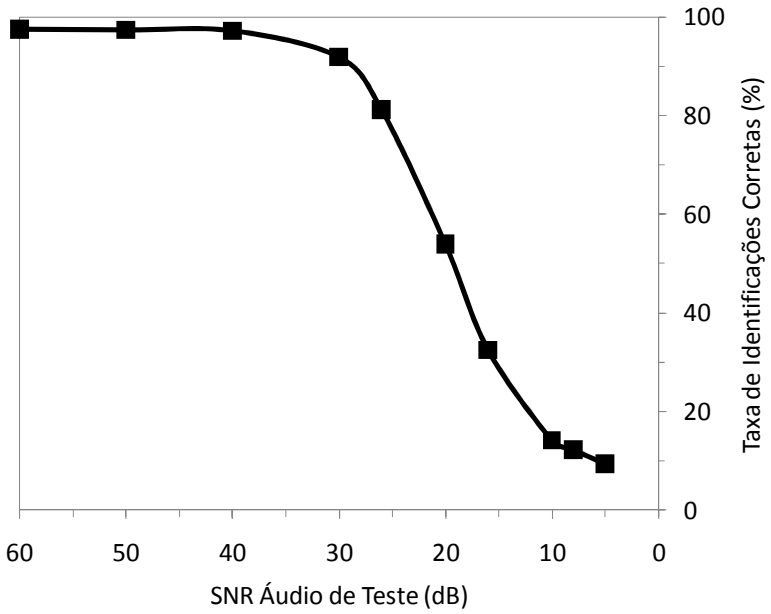

Figura 3.2: Degradação do desempenho de sistema de RAL devido à diminuição da SNR, para frequência de amostragem de 8 kHz, quantização de 8 bits linear, e modelos GMM de 16 componentes.

Como se verifica, não houve alterações significativas com relação aos resultados obtidos para o caso de frequência de amostragem de 22 kHz e quantização de 16 bits.

Embora os resultados exibidos tenham sido obtidos para um sistema baseado em GMM de 16 componentes, experimentos realizados com sistemas baseados em GMM de 2, 4, 8 e 32 componentes demonstraram resultados semelhantes (com relação à degradação pelo ruído). A inclinação do gráfico de degradação de desempenho é aproximadamente constante, havendo, contudo, uma antecipação do ponto de início da queda e também uma redução global de desempenho à medida que o número de componentes do modelo é reduzido.

3.1.2 Sensibilidade à Codificação MP3

Da mesma forma que o sistema de RAL apresenta degradação de desempenho quando o áudio questionado tem SNR reduzida, também ocorre diminuição na taxa de locutores corretamente identificados quando se utiliza áudio codificado em formatos MP3 (com perdas) a baixas taxas de bits, como foi relatado no trabalho realizado durante a elaboração desta tese D'Almeida *et al* (2007).

Para demonstrar a sensibilidade dos sistemas de RAL baseados em GMM à codificação do áudio de teste no formato MP3, foi realizada avaliação computacional análoga à descrita no caso da sensibilidade ao ruído. O sistema foi treinado com o áudio do banco de vozes

original, Φ_0, e avaliado para áudio de versões do banco de vozes que foram submetidos a uma codificação no formato MP3 com variadas taxas de codificação: 56 kbit/s, 40 kbit/s, 32 kbit/s, 24 kbit/s, 20 kbit/s, 16 kbit/s e 8 kbit/s[*], de modo que foram geradas as versões do banco de dados Φ_{56kbps}, Φ_{40kbps}, Φ_{32kbps}, Φ_{24kbps}, Φ_{20kbps}, Φ_{16kbps} e Φ_{8kbps}. Em todos os casos, foi mantida a frequência original de amostragem de 22 kHz.

Os resultados das avaliações realizadas para demonstrar a sensibilidade do sistema à codificação do áudio questionado no formato MP3 com baixas taxas são exibidos na Tabela 3.3.

Tabela 3.3: Degradação do desempenho de sistema de RAL devido à codificação do áudio no formato MP3, para frequência de amostragem de 22 kHz e modelos GMM de 16 componentes.

Taxa de Codificação (kbit/s)	**Taxa de Identificações Corretas (%)**
S/C	100,0
56	100,0
40	98,5
32	89,7
24	73,5
20	75,3
16	63,2
8	10,5

Para melhor visualização, os resultados da Tabela 3.3 foram traçados no gráfico da Figura 3.3, onde, apenas para efeitos de visualização, atribuiu-se o valor de 65 kbit/s ao áudio sem codificação MP3[†].

[*] Ressalta-se que, em todas as simulações, foi utilizada a frequência de amostragem do banco de vozes original, 22 kHz, e também a codificação de 16 bits por amostra.

[†] De fato, o valor correto seria de 352 kbit/s (352.000 bits/s = 22.000 amostras/s x 16 bits/amostra). Entretanto, a utilização desse valor exigiria um redimensionamento do gráfico que dificultaria a visualização da parte que realmente é relevante. Além disso, por se tratar de áudio sem compressão, a comparação de sua taxa de codificação com a taxa do áudio com codificação MP3 não tem qualquer sentido na presente aplicação de reconhecimento de locutores. Essa comparação somente seria razoável no contexto de compressão de dados.

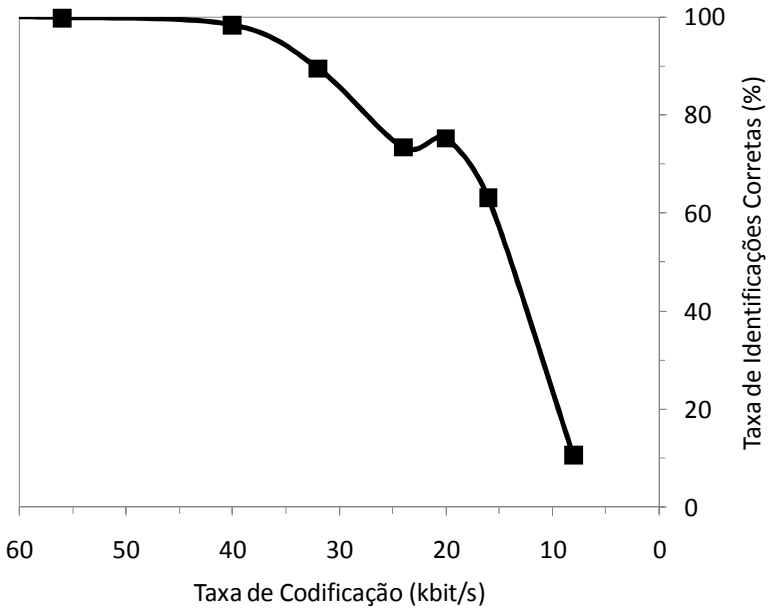

Figura 3.3: Degradação do desempenho de sistema de RAL devido à codificação do áudio no formato MP3, para frequência de amostragem de 22 kHz e modelos GMM de 16 componentes.

Verifica-se que o desempenho do sistema se degrada rapidamente com a redução da taxa de codificação utilizada, especialmente para taxas inferiores a 40 kbit/s.

Embora o resultado exibido tenha sido obtido para um sistema baseado em GMM de 16 componentes, experimentos realizados com sistemas baseados em GMM de 2, 4, 8 e 32 componentes demonstraram resultados semelhantes. Também, foram realizadas avaliações com áudio amostrado com frequências de 16 kHz e 11 kHz. Os resultados obtidos foram semelhantes, havendo, em todos os casos, uma intensa diminuição no desempenho do sistema à medida que se reduz a taxa de codificação do áudio.

3.2 TREINAMENTO SINTONIZADO

Como se verificou na seção 3.1, a degradação do desempenho dos sistemas de RAL para condições de áudio ruidoso ou de codificação MP3 a baixas taxas é intensa. As taxas de identificações corretas caem de 100%, para o áudio sem ruído ou com codificação MP3 a altas taxas; para aproximadamente 10%, no caso áudio com SNR igual a 5 dB ou codificação MP3 abaixo de 10 kbit/s. Uma forma conhecida de minimizar essa degradação e de manter desempenhos elevados em situações de ruído elevado ou de codificação a baixas taxas é treinar os modelos GMM com amostras de áudio de características semelhantes ao

do áudio a ser testado, ou seja, realizar um treinamento sintonizado (TS) (Matsui, Kanno e Furui, 1996).

3.2.1 Treinamento Sintonizado em Áudio Ruidoso

A fim de demonstrar os ganhos de desempenho do sistema de RAL com o treinamento sintonizado (TS) em situação de áudio ruidoso, foram realizados procedimentos de identificação com áudio de todos os valores de SNR utilizados na seção 3.1 (60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 10 dB) utilizando modelos treinados com áudio de mesma SNR. Os resultados desses procedimentos são exibidos na Tabela 3.4, onde também foram repetidos os resultados da Tabela 3.1, referentes ao desempenho de um sistema de RAL com modelos treinados com áudio sem ruído.

Tabela 3.4: Taxas de identificações corretas de sistema de RAL com treinamento sintonizado (TS), em função do nível de ruído, para frequência de amostragem de 22 kHz, quantização de 16 bits, e modelos GMM de 16 componentes.

SNR do Áudio de Teste (dB)	**Taxa de Identificações Corretas (%)**	
	Treinamento Sintonizado (TS)	**Treinamento Sem Ruído**
60	100,0	100,0
50	99,7	100,0
40	98,0	99,3
30	91,7	93,7
26	87,7	83,0
20	79,8	54,7
16	78,5	28,5
10	70,8	13,8
8	67,8	12,0
5	64,0	10,0

A fim de permitir uma melhor visualização dos resultados, os resultados da Tabela 3.4 foram traçados no gráfico da Figura 3.4.

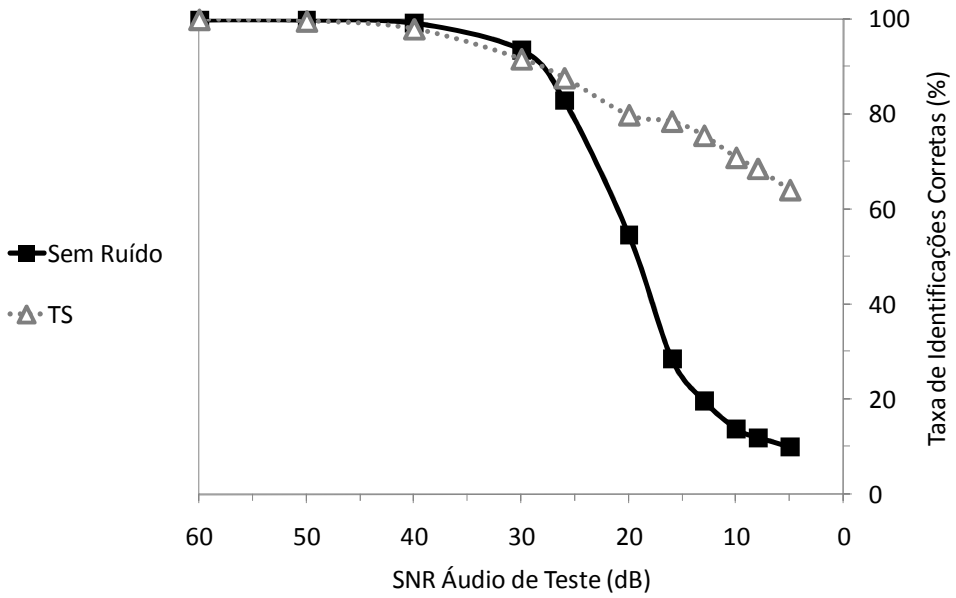

Figura 3.4: Comparação da degradação de desempenho entre sistemas de RAL com treinamento sem ruído e com treinamento sintonizado (TS), para frequência de amostragem de 22 kHz, quantização de 16 bits linear, e modelos GMM de 16 componentes.

Como se pode observar, para valores de SNR inferiores a 30 dB, o TS tem desempenho significativamente superior ao treinamento com áudio sem ruído, sendo que a diferença de desempenho entre os sistemas cresce rapidamente com a diminuição da SNR, superando 50% de diferença de taxa de identificações corretas para valores de SNR iguais ou inferiores a 16 dB. Dessa maneira, verifica-se que o treinamento sintonizado é uma técnica altamente eficaz na melhora do desempenho de sistemas de RAL para condições de áudio ruidoso.

Entretanto, para áudio com valores de SNR entre 50 dB e 30 dB, verifica-se que a utilização do TS provoca uma leve degradação do desempenho do sistema, quando comparado com sistemas treinados com áudio sem ruído. Essa constatação leva a concluir que o treinamento dos modelos com áudio de nível de ruído semelhante ao do áudio de teste não é necessariamente a melhor opção. A fim de explorar de modo mais aprofundado essa questão e de determinar o nível de ruído ótimo para o treinamento dos modelos para cada condição de ruído, foram realizados experimentos com diversos sistemas de RAL, cada um treinado com áudio de determinado valor de SNR (60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB e 5 dB). Para cada condição de treinamento, foram realizadas identificações com áudio de diversos valores de SNR (60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB e 5 dB). Com esse procedimento, foram analisa-

das todas as combinações de SNR de áudio de treinamento e de SNR de áudio questionado. Os resultados dessas análises estão expostos na Tabela 3.5:

Tabela 3.5: Taxas de identificações corretas de sistemas de RAL treinados com diferentes condições de SNR, para frequência de amostragem de 22 kHz, quantização de 16 bits, e modelos GMM de 16 componentes.

Taxas de Identificações Corretas (%)		**SNR Teste (dB)**						
		60	**50**	**40**	**30**	**26**	**20**	**16**
SNR Treinamento (dB)	**60**	100,0	100,0	99,3	93,7	83,0	54,7	28,5
	50	98,7	99,7	99,5	95,8	87,2	58,7	30,0
	40	96,3	96,5	98,0	93,2	87,5	53,7	26,3
	30	73,5	76,7	85,5	91,7	90,8	72,0	37,7
	26	60,8	62,7	69,2	84,2	87,7	81,2	57,3
	20	30,3	31,3	41,2	62,2	71,7	79,8	75,0
	16	17,7	18,2	22,3	40,7	54,0	72,0	78,5
	13	13,8	13,8	16,7	22,7	34,8	58,8	70,0
	10	5,8	6,2	7,7	14,7	19,2	39,0	51,0
	8	6,7	4,2	8,3	6,2	15,5	22,5	37,0
	5	3,8	3,7	6,7	10,2	13,5	14,3	25,0

Continuação da Tabela 3.5

Taxas de Identificações Corretas (%)		**SNR Teste (dB)**			
		13	**10**	**8**	**5**
SNR Treinamento (dB)	**60**	19,7	13,8	12,0	10,0
	50	25,7	15,7	10,0	8,3
	40	19,0	14,2	8,3	5,3
	30	21,7	9,2	8,2	4,5
	26	31,5	14,7	9,2	5,5
	20	56,5	33,5	21,5	8,8
	16	72,3	53,2	37,2	18,5
	13	75,5	67,8	53,2	27,8
	10	67,3	70,8	66,8	47,3
	8	66,8	66,3	68,5	58,3
	5	37,0	46,7	59,8	64,0

Como se verifica na Tabela 3.5, embora o TS sempre produza elevados valores da taxa de identificações corretas, nem sempre essa situação maximiza o desempenho do sistema. Verifica-se, em muitos casos, que o treinamento dos modelos com áudio de valor de SNR maior que o do áudio questionado é que otimiza o desempenho do sistema de RAL. Na Tabela 3.6, foram listadas as taxas de identificações corretas para o treinamento sintonizado (TS) e para o treinamento ótimo (TO).

Tabela 3.6: Taxas de identificações corretas de sistema de RAL com treinamento sintonizado (TS), para frequência de amostragem de 22 kHz, quantização de 16 bits, e modelos GMM de 16 componentes.

SNR do Áudio de Teste (dB)	**Taxa de Identificações Corretas (%)**		
	Treinamento Ótimo (TO)		**Treinamento Sintonizado (TS)**
60	100,0	(60 db)	100,0
50	100,0	(60 db)	99,7
40	99,5	(50 db)	98,0
30	95,8	(50 db)	91,7
26	90,8	(30 db)	87,7
20	81,2	(26 db)	79,8
16	78,5	(16 db)	78,5
13	75,5	(13 db)	75,5
10	70,8	(10 db)	70,8
8	68,5	(8 db)	68,5
5	64,0	(5 db)	64,0

A fim de permitir uma melhor visualização dos resultados, os resultados da Tabela 3.4 foram traçados no gráfico da Figura 3.4, juntamente com os resultados do sistema treinado com áudio sem adição de ruído, para comparações.

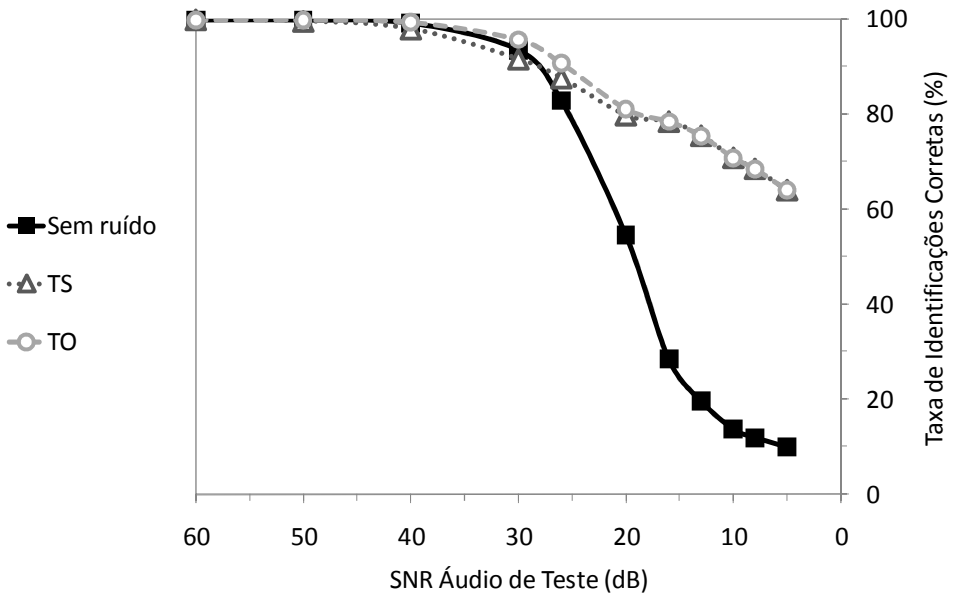

Figura 3.5: Comparação da degradação de desempenho entre sistemas de RAL com treinamento sintonizado (TS) e com treinamento ótimo (TO), para frequência de amostragem de 22 kHz, quantização de 16 bits linear, e modelos GMM de 16 componentes.

Como se pode observar, o treinamento ótimo (TO) permite um aumento significativo no desempenho do sistema de RAL na faixa SNR entre 50 dB e 20 dB; até 4,1%, para áudio de teste com SNR de 30 dB, e média de 2,1%. Para valores de SNR inferiores a 20 dB, o treinamento ótimo teve desempenho igual ao treinamento sintonizado (ou seja, o treinamento ótimo é o treinamento sintonizado).

3.2.2 Treinamento Sintonizado em Áudio com Codificação MP3

Foram também realizados experimentos com o treinamento sintonizado (TS) para o caso de áudio codificado no formato MP3, sendo realizados procedimentos de identificação com áudio de todos os valores de codificação utilizados na seção 3.1.2 (56 kbit/s, 40 kbit/s, 32 kbit/s, 24 kbit/s, 20 kbit/s, 16 kbit/s e 8 kbit/s) utilizando modelos treinados com áudio de mesma codificação. Os resultados dessas avaliações são exibidos na Tabela 3.7, onde também foram repetidos os resultados da Tabela 3.3, referentes ao desempenho de um sistema de RAL com modelos treinados com áudio sem codificação MP3.

Tabela 3.7: Taxas de identificações corretas de sistema de RAL com treinamento sintonizado (TS), em função da taxa de codificação MP3, para frequência de amostragem de 22 kHz, quantização de 16 bits, e modelos GMM de 16 componentes.

Taxa de Codificação (kbit/s)	**Taxa de Identificações Corretas (%)**	
	Treinamento Sintonizado (TS)	**Treinamento Sem Codificação**
S/C*	100,0	100,0
56	100,0	100,0
40	100,0	98,5
32	99,8	89,7
24	99,5	73,5
20	99,5	75,3
16	97,8	63,2
8	91,5	10,5

A fim de permitir uma melhor visualização, os resultados da Tabela 3.7 foram traçados no gráfico da Figura 3.6, onde, apenas para efeitos de visualização, atribuiu-se o valor de 60 kbit/s ao áudio sem codificação MP3.

* Sem codificação MP3.

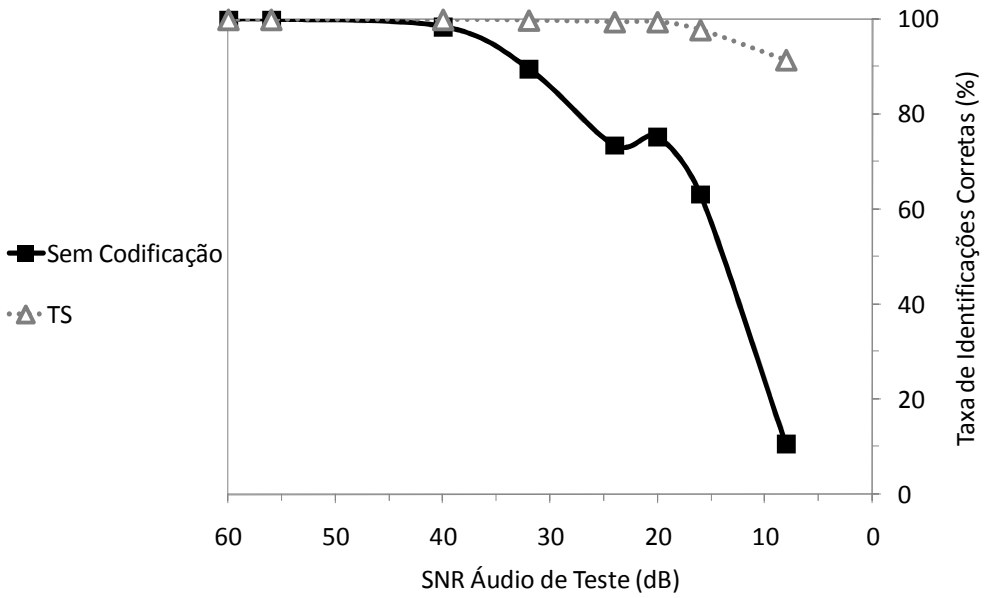

Figura 3.6: Comparação da degradação de desempenho entre sistemas de RAL com treinamento sem codificação e com treinamento sintonizado (TS), para frequência de amostragem de 22 kHz, quantização de 16 bits linear, e modelos GMM de 16 componentes.

Como se pode observar, para valores de taxa de codificação inferiores a 40 kbit/s, o TS tem desempenho significativamente superior ao treinamento com áudio sem codificação, sendo que a diferença de desempenho entre os sistemas cresce rapidamente com a diminuição da taxa de codificação. Dessa maneira, verifica-se que o TS é uma técnica altamente eficaz na melhora do desempenho de sistemas de RAL para condições de áudio codificado no formato MP3.

Apesar de não se ter verificado qualquer degradação de desempenho no TS em comparação com o treinamento sem codificação, tendo em vista o que se constatou no caso de áudio ruidoso, optou-se por realizar experimentos com diversos sistemas de RAL, cada um treinado com áudio de um determinado valor de taxa de codificação (56 kbit/s, 40 kbit/s, 32 kbit/s, 24 kbit/s, 20 kbit/s, 16 kbit/s e 8 kbit/s). Para cada condição de treinamento, foram realizadas identificações com áudio de diversos valores de taxa de codificação (56 kbit/s, 40 kbit/s, 32 kbit/s, 24 kbit/s, 20 kbit/s, 16 kbit/s e 8 kbit/s). Com esse procedimento, foram analisadas todas as combinações de taxa de codificação de áudio de treinamento e de taxa de codificação de áudio questionado (a exemplo do que foi feito com o nível de ruído). Os resultados dessas análises estão expostos na Tabela 3.8

Tabela 3.8: Taxas de identificações corretas de sistemas de RAL treinados com diferentes taxas de codificação MP3, para frequência de amostragem de 22 kHz, quantização de 16 bits, e modelos GMM de 16 componentes.

Taxas de Identificações Corretas (%)		Taxa de Codificação de Teste (kbit/s)							
		S/C*	**56**	**40**	**32**	**24**	**20**	**16**	**8**
Taxa Codificação de Treino (kbit/s)	**S/C**	100,0	100,0	98,5	89,7	73,5	75,3	63,2	10,5
	56	100,0	100,0	98,5	89,7	73,5	75,3	63,2	10,5
	40	98,7	99,8	100,0	98,8	68,2	61,3	66,8	3,8
	32	97,7	98,8	99,8	99,8	90,8	82,8	70,5	18,2
	24	90,2	86,2	88,3	96,2	99,5	96,0	68,5	14,7
	20	80,7	76,7	67,8	72,0	91,3	99,5	90,0	12,0
	16	57,7	71,2	60,5	51,7	72,8	91,2	97,8	13,2
	8	24,0	17,8	20,3	13,0	20,7	31,7	21,0	91,5

3.3 MODELOS MULTICONDICIONAIS

Os sistemas de RAL baseados em Modelos de Mistura de Gaussianas (GMM), como visto na seção 3.1, apresentam bom desempenho na identificação dos locutores desde que o áudio testado não tenha nível de ruído elevado nem tenha sido codificado no formato MP3 com uma taxa de bits por segundo muito baixa. Entretanto, muitas vezes, esse tipo de degradação no áudio a ser testado é inevitável e pode até ocorrer em intensidade imprevisível ou variável ao longo do tempo (particularmente no caso do ruído), de modo que se torna necessário desenvolver um sistema de RAL mais robusto a esses fatores. A fim de diminuir a sensibilidade dos sistemas de RAL a variações na qualidade do áudio de teste, tornando o sistema mais robusto ao ruído ou a codificações com perdas a baixas taxas, é comumente utilizada a técnica de modelagem multicondicional (Matsui, Kanno e Furui, 1996; Yang e Gong, 2006), que tira proveito das propriedades do treinamento sintonizado detalhado na seção 3.2.

* Sem codificação MP3.

A idéia fundamental da modelagem multicondicional é a criação de diversos modelos para um mesmo locutor, $\lambda_{s,n}$, sendo que cada modelo será treinado com áudio de bancos de vozes de diferentes características (nível de ruído, taxa de codificação etc.), Φ_n. Dessa maneira, como há uma variedade de modelos do locutor disponíveis, espera-se que haja um modelo cujo treinamento tenha sido feito com áudio de características não tão distintas das do áudio questionado, de forma que a degradação do desempenho da identificação seja reduzida. O conjunto de modelos para um determinado locutor é então combinado em um único modelo multicondicional, que é representado matematicamente por:

$$p(Y \mid s) = \sum_{n=0}^{N-1} p(Y \mid s, \Phi_n) P(\Phi_n \mid s) \quad , \tag{3.7}$$

onde $Y = \{\vec{y}_1, \vec{y}_2, ..., \vec{y}_T\}$ representa o conjunto de vetores de parâmetros extraídos dos arquivos de voz, como detalhado na equação (2.6); s representa o locutor; $p(Y \mid s, \Phi_n)$ é a verossimilhança do modelo do locutor s treinado para a condição Φ_n com relação ao conjunto de vetores de parâmetros Y; e $P(\Phi_n \mid s)$ é a probabilidade *a priori* de ocorrência de áudio com característica Φ_n para o locutor s.

Deve-se observar que os valores de $P(\Phi_n \mid s)$ são restritos a:

$$\sum_n P(\Phi_n \mid s) = 1 \tag{3.8}$$

Na realidade, a expressão (3.7) expressa a modelagem multicondicional de forma muito restritiva, pois fixa o valor de $P(\Phi_n \mid s)$ para todas as janelas de análise do áudio e mesmo para qualquer áudio questionado, pois representa uma probabilidade *a priori* da ocorrência da condição de ruído Φ_n para o locutor s. Uma maneira mais flexível de expressar os modelos multicondicionais, e que é comumente utilizada, é permitir o ajuste dos valores de $P(\Phi_n \mid s)$ a cada janela de análise, de modo que se passa a ter um modelo matematicamente expresso por:

$$p(\vec{y}_t \mid s) = \sum_{n=0}^{N-1} p(\vec{y}_t \mid s, \Phi_n) P(\Phi_n \mid s, \vec{y}_t) \quad , \tag{3.9}$$

sendo $p(\vec{y}_t \mid s, \Phi_n)$ a verossimilhança do modelo do locutor s treinado para a condição Φ_n com relação ao vetor de parâmetros $\vec{y}_t$. É, portanto, o mesmo que $p(\vec{y}_t \mid \lambda_{s,n})$, a verossi-

milhança do modelo $\lambda_{s,n}$ diante do vetor de parâmetros de $\vec{y}_t$. Assim, trata-se de um modelo GMM normal, tal como descrito nas equações (2.1) a (2.4):

$$p\left(\vec{y}_t \mid \lambda_{s,n}\right) = \sum_{i=1}^{M} p_{n,i} b_{n,i}\left(\vec{y}_t\right) \quad , \tag{3.10}$$

onde, do lado direito, foram omitidos os índices *s* referentes ao locutor para simplificar a notação.

O cálculo de $P\left(\Phi_n \mid s, \vec{y}_t\right)$ pode ser feito de diversas maneiras. Pode-se, por exemplo, realizar uma análise diretamente no áudio (não necessariamente no vetor de parâmetros) para estimar suas características (nível de ruído), como realizado por Xu, Dalsgaard e Lindberg (2005). Essa forma apresenta a vantagem de diminuir o custo computacional do processo, pois, após a determinação das características do áudio, passa-se a utilizar o modelo GMM unicondicional mais adequado, não sendo calculados os demais GMM do modelo multicondicional[*]. Essa opção pode ser representada matematicamente por:

$$\begin{aligned} P\left(\Phi_n \mid s, \vec{y}_t\right) &= 1, \; se \; n = \tilde{n} \\ P\left(\Phi_n \mid s, \vec{y}_t\right) &= 0, \; se \; n \neq \tilde{n} \end{aligned} \quad , \tag{3.11}$$

sendo que, na equação (3.11), a determinação do modelo a ser utilizado, $\tilde{n}$, pode ser feita, por exemplo, por estimativas da SNR do áudio.

No caso de a estimativa das características do áudio ser realizada uma única vez, o modelo retorna à forma de um GMM unicondicional, apenas com a escolha do GMM mais adequado disponível[†]:

$$\begin{aligned} p\left(\vec{y}_t \mid s\right) &= \sum_{n=0}^{N-1} p\left(\vec{y}_t \mid s, \Phi_n\right) P\left(\Phi_n \mid s, \vec{y}_t\right) \\ &= p\left(\vec{y}_t \mid s, \Phi_{\tilde{n}}\right) = p\left(\vec{y}_t \mid \lambda_{s,\tilde{n}}\right) \end{aligned} \tag{3.12}$$

Dessa maneira, essa abordagem dos modelos multicondicionais assemelha-se ao treinamento ótimo detalhado na seção 3.2.

[*] Essa vantagem é relativa, pois depende de como é feita a estimativa das características do áudio e, principalmente, de com que frequência essa estimativa é reavaliada.

[†] Para maximizar o desempenho do sistema, nesse caso, deve-se utilizar o modelo ótimo para o nível de ruído do áudio de teste, que não necessariamente é o modelo treinado com áudio de mesmas características (sintonizado), conforme detalhado na seção 3.2.

Apesar de apresentar vantagem sobre a modelagem unicondicional (uma vez que o treinamento ótimo é superior ao treinamento sem ruído), esse método, se tomado na forma de (3.12), tem a limitação de não acompanhar evoluções temporais nas características do sinal, resultando, portanto, em desempenho inferior ao dos modelos multicondicionais completos. Portanto, essa abordagem somente é possível se as características do áudio a ser identificado são constantes[*].

A forma mais comumente empregada de modelagem multicondicional envolve a cálculo de $p\left(\vec{y}_t \mid \lambda_{s,n}\right)$ para todos os modelos do locutor *s*, $\lambda_{s,n}$, e a seleção daquele que apresenta a maior verossimilhança dado o vetor de parâmetros $\vec{y}_t$ (Matsui, Kanno e Furui, 1996; Yang e Gong, 2006). Essa abordagem será denominada de Método da Condição Máxima (MCM). Matematicamente, nesse caso, tem-se que $P\left(\Phi_n \mid s, \vec{y}_t\right)$ é definido por:

$$\begin{aligned} &P\left(\Phi_n \mid s, \vec{y}_t\right) = 1, se\ n = \tilde{n}[t] = \arg \max_{n=0,\ldots,N-1} p\left(\vec{y}_t \mid \lambda_{s,n}\right) \\ &P\left(\Phi_n \mid s, \vec{y}_t\right) = 0, se\ n \neq \tilde{n}[t] \end{aligned}, \qquad (3.13)$$

onde foi utilizada a notação $\tilde{n}[t]$ para explicitar que o modelo a ser utilizado depende do instante (da janela de análise) *t* considerado.

A principal vantagem desse tipo de abordagem é a utilização do modelo ótimo para cada janela de análise do áudio, possibilitando melhor desempenho do sistema de RAL. Deve-se ainda perceber que, por ser baseado em análises de curto tempo, esse método alcança bons resultados mesmo em situações de variação rápida nas características do áudio, como ocorre, por exemplo, em gravações de áudio ambiental em locais com trânsito de veículos ou utilização de máquinas. Por outro lado, esse tipo de sistema apresenta um aumento no custo computacional proporcional ao número de modelos multicondicionais utilizados, pois, efetivamente, para cada locutor, são calculados *N* funções de verossimilhança. Na realidade, esse acréscimo no custo computacional é a maior dificuldade para a utilização generalizada dos modelos multicondicionais, e a minimização desse custo é o objetivo central desta tese, como se verá no capítulo 4.

[*] No caso de haver variação lenta das características do áudio de teste, é possível segmentar o áudio em trechos de características aproximadamente constantes e realizar o teste com os modelos mais adequados para cada trecho.

Também é possível, ao invés de utilizar apenas o modelo relativo à condição de treinamento de melhor resultado, utilizar a soma dos resultados de todas as condições. Dessa maneira, as probabilidades *a priori* utilizadas são:

$$P\left(\Phi_n \mid s, \vec{y}_t\right) = P\left(\Phi_n \mid s, Y\right) = \frac{1}{N} \quad , \tag{3.14}$$

onde foi inserido o termo $P\left(\Phi_n \mid s, Y\right)$ para explicitar que esse termo será constante no tempo. Essa abordagem será denominada de Método da Soma das Condições (MSC) e sofre do mesmo problema de aumento de custo computacional verificado no MCM. Entretanto, deve-se destacar que o MSC não comporta algumas das otimizações propostas nesta tese, como será visto no capítulo 4.

Recentemente, Ming *et al* (2007) propuseram uma forma alternativa de modelagem multicondicional. Nesse novo método, os diferentes bancos de vozes, Φ_n, são unidos num único grande banco de vozes multicondicional, $\Phi = \Phi_0 + \Phi_1 + ... + \Phi_{N-1}$, e todo o treinamento e testes são realizados considerando o modelo multicondicional com se fosse, na realidade, simplesmente um grande modelo GMM[*]. Essa nova abordagem, embora dificulte a visualização da modelagem multicondicional subjacente, é, na realidade, apenas um novo modo de tratar o treinamento multicondicional. Como se pode verificar pela inserção da equação (3.10) na equação (3.9), o modelo multidimensional pode ser escrito como:

$$p\left(\vec{y}_t \mid s\right) = \sum_{n=0}^{N-1}\left[\sum_{i=1}^{M} p_{n,i} b_{n,i}\left(\vec{y}_t\right)\right] P\left(\Phi_n \mid s, \vec{y}_t\right) = \sum_{n=0}^{N-1}\sum_{i=1}^{M} p'_{n,i} b_{n,i}\left(\vec{y}_t\right) \quad , \tag{3.15}$$

onde

$$p'_{n.i} = p_{n,i} P\left(\Phi_n \mid s, \vec{y}_t\right) \tag{3.16}$$

Alterando a indexação de *n* e *i* para um novo índice *k*:

$$p\left(\vec{y}_t \mid s\right) = \sum_{k=1}^{N \cdot M} p'_k b_k\left(\vec{y}_t\right) \tag{3.17}$$

A equação (3.17) é, claramente, a definição de um modelo GMM com $N \cdot M$ componentes gaussianas.

[*] A proposta de Ming, de fato, além dessa reformulação, utiliza o processamento do áudio em sub-bandas de frequência, o que confere maior robustez do modelo a ruídos não uniformemente espalhados no espectro.

A abordagem proposta por Ming *et al* (2007) tem a vantagem de apresentar apenas uma restrição para os coeficientes do modelo:

$$\sum_{k=1}^{N \cdot M} p_k^{'} = 1 \quad , \tag{3.18}$$

enquanto que a modelagem multicondicional tradicional apresenta $N+1$ restrições, uma indicada na equação (3.8) e outras *N* decorrentes das normalizações

$$\sum_{i=1}^{M} p_{n,i} = 1, n = 0, ..., N-1 \quad , \tag{3.19}$$

para cada um dos *N* modelos GMM da forma de (3.10). Consequentemente, na nova abordagem, o modelo é mais flexível e pode redistribuir as componentes gaussianas de forma a maximizar a qualidade global da modelagem. Em contrapartida, essa abordagem apresenta dois inconvenientes. Primeiro, como todas as componentes gaussianas são treinadas conjuntamente, pode ocorrer de algumas delas (as que representam uma determinada condição do áudio de treino) receberem pesos maiores que outras, fazendo o modelo tender a valorizar mais uma determinada condição (média) de áudio. Essa situação pode ser evitada ou ao menos minimizada pela construção cuidadosa do banco de vozes global de treino[*], Φ. O outro inconveniente é que, na abordagem de Ming *et al* (2007) não é possível aplicar algumas das otimizações propostas nesta tese para reduzir o custo computacional do modelo pela redução do número de gaussianas efetivamente computadas, como será discutido no capítulo 4.

3.3.1 Simulações com Modelos Multicondicionais

A fim de observar o desempenho do sistema de RAL com modelos GMM multicondicionais, foram realizados alguns procedimentos de identificação com essa modelagem. Foram analisados modelos multicondicionais baseados na condição de (3.13), que utiliza apenas o resultado do modelo GMM de melhor desempenho (Método da Condição Máxima - MCM); modelos multicondicionais baseados na soma dos resultados de todos os GMM (Método da Soma das Condições - MSC) e também utilizando a metodologia proposta em

[*] A construção de um banco de vozes realmente balanceado é mais complexa do que aparenta ser a princípio, pois, mesmo que sejam utilizadas quantidades semelhantes de trechos de áudio com as *N* diferentes características, é possível que duas ou mais dessas condições sejam bem representadas por um único conjunto de gaussianas (um GMM), como visto na seção 3.2, o que fará esse conjunto de componentes receber pesos mais elevados que as demais, desbalanceando o modelo.

Ming *et al* (2007), que consiste na utilização de um modelo GMM único, com grande número de componentes, treinado com áudio de diferentes condições. Vale ainda destacar que os resultados da seção 3.2, que tratam do treinamento sintonizado e do treinamento ótimo, também podem ser interpretados como resultados de um sistema GMM multicondicional que utiliza:

$$\begin{aligned} P\left(\Phi_n \mid s, \vec{y}_t\right) &= P\left(\Phi_n \mid s, Y\right) = 1, se\ n = \hat{n} \\ P\left(\Phi_n \mid s, \vec{y}_t\right) &= P\left(\Phi_n \mid s, Y\right) = 0, se\ n \neq \hat{n} \end{aligned} \quad , \tag{3.20}$$

sendo $\hat{n}$ o índice que representa a condição de treinamento do modelo GMM* dentro do modelo multicondicional. De fato, esse é o modelo utilizado em Xu, Dalsgaard e Lindberg (2005), com a diferença de que não foi necessário estimar o nível de ruído do áudio, pois esse era conhecido.

Deve-se destacar que, apesar das semelhanças entre os modelos multicondicionais expressos nas equações (3.13) e (3.20), há diferenças significativas entre essas modelagens. A abordagem da equação (3.13) avalia a condição ótima, $\tilde{n}$, para cada vetor de parâmetros, $\vec{y}_t$, individualmente. Dessa maneira, é possível que a condição ótima estimada varie ao longo do tempo para um mesmo modelo de locutor. A abordagem da equação (3.20), por outro lado, utiliza uma condição ótima fixa, $\hat{n}$, para toda a simulação, não havendo variação com relação ao tempo nem com relação aos diferentes locutores.

A Tabela 3.9 apresenta o resultado dos experimentos realizados usando modelos multicondicionais treinados com condições de SNR igual a 60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB e 5 dB. A linha "Ming" lista os resultados das simulações utilizando o modelo multicondicional proposto em Ming *et al* (2007) e utilizando 16 componentes gaussianas por condição (totalizando 11 condições x 16 componentes/condição = 176 componentes gaussianas); a linha "MCM", relaciona os resultados das simulações usando o Método da Condição Máxima da equação (3.13) e modelos de 16 componentes por condição; a linha "MSC", lista os resultados das identificações utilizando o Método da Soma das Condições, conforme a equação (3.14), e modelos

* Ao se interpretar a técnica do treinamento sintonizado ou do treinamento ótimo como condições para determinação de $P\left(\Phi_n \mid s, Y\right)$, deve-se considerar que, nos casos apresentados na seção 3.2, as características (SNR) do áudio de teste eram precisamente conhecidas e, consequentemente, o índice $\hat{n}$ era corretamente determinado. Em aplicações práticas, o valor da SNR será estimado a partir do áudio disponível, o que pode provocar a escolha de modelos não exatamente sintonizados ou ótimos (erro na estimativa de $\hat{n}$), levando, consequentemente, a uma perda de desempenho. Portanto, os resultados da seção 3.2 devem ser considerados como limites máximos de desempenho.

de 16 componentes por condição; e a linha "TO" contém os resultados do Treinamento Ótimo da Tabela 3.6, que, como disposto na equação (3.20), também pode ser interpretado como uma modelagem multicondicional. Todas as simulações com modelos multicondicionais foram realizadas com áudio de frequência de amostragem 22 kHz e quantização de 16 bits. Todos os modelos foram treinados com áudio de aproximadamente 30 s para cada condição. No caso da modelagem proposta por Ming *et al* (2007), como se trata de um modelo único, o treinamento foi realizado com um único áudio gerado pela concatenação dos áudios de treinamento para cada uma das condições utilizadas.

Tabela 3.9: Taxas de identificações corretas de sistemas de RAL multicondicionais utilizando as abordagens de Ming, MCM, MSC e TO, todas com 11 condições de treinamento.

Taxas de Identificações Corretas (%)		**SNR do Áudio de Teste (dB)**										
		60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
Modelagem Multicondicional	**Ming**	65,2	67,2	73,7	85,2	88,0	91,8	86,3	69,7	42,5	28,7	18,3
	MCM	97,8	98,2	98,2	97,8	97,2	95,0	92,5	89,0	83,0	75,5	60,0
	MSC	97,3	97,5	97,7	96,2	96,2	94,8	91,2	89,3	83,0	76,0	58,2
	TO	100,0	100,0	99,5	95,8	90,8	81,2	78,5	75,5	70,8	68,5	64,0

Para permitir uma melhor visualização, os resultados da Tabela 3.9 foram traçados nos gráficos das Figura 3.7 e da Figura 3.8.

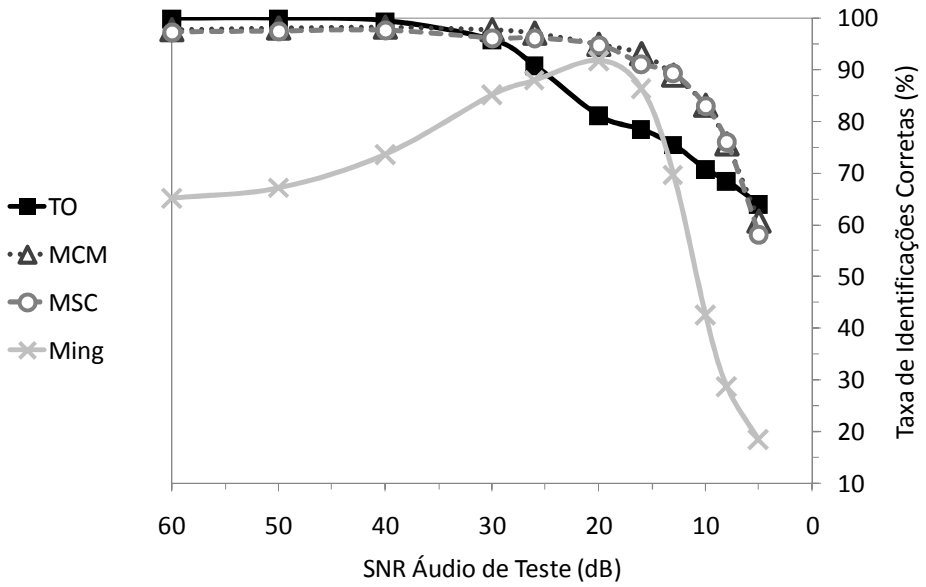

Figura 3.7: Comparação do desempenho de sistemas de RAL multicondicionais utilizando as abordagens de Ming, MCM, MSC e TO, todas com 11 condições de treinamento.

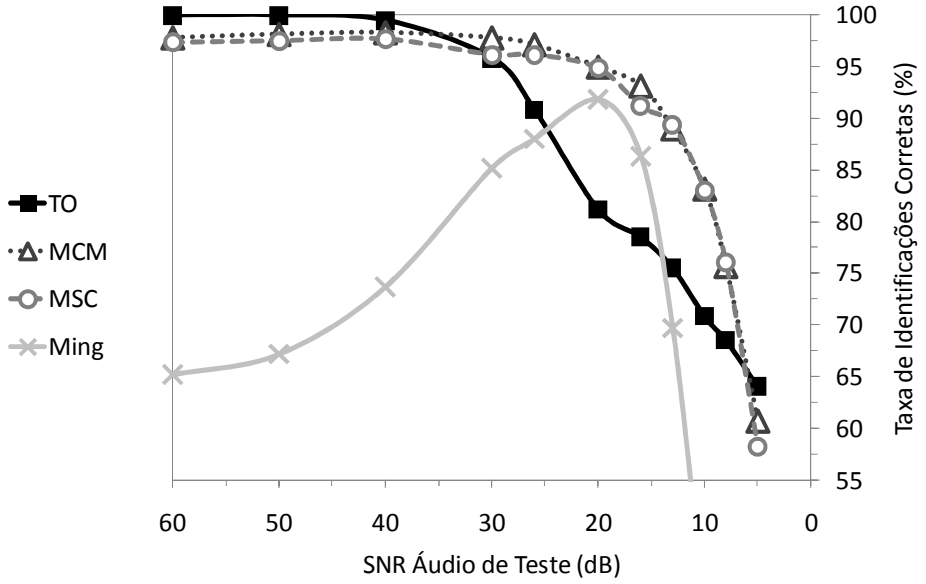

Figura 3.8: Comparação do desempenho de sistemas de RAL multicondicionais utilizando as abordagens de Ming, MCM, MSC (11 condições de treinamento) e treinamento ótimo.

Como se pode observar, o método de Ming teve desempenho sempre inferior aos MSC, equação (3.14), e também ao MCM, equação (3.13). O método de Ming teve resultados superiores ao do treinamento ótimo apenas para valores de SNR entre 20 dB e 16 dB (em média, houve perda de 18,9% em comparação com o treinamento ótimo). O MSC e o MCM obtiveram resultados muito semelhantes em toda a faixa de SNR, embora o MCM

tenha demonstrado leve superioridade[*] (0,9%, na média). Comparados com o treinamento ótimo, o MSC e o MCM tiveram resultados superiores na faixa de 30 dB a 8 dB. Nos valores extremos de SNR, tanto no extremo superior (60 dB a 40 dB) quanto no inferior (5 dB), o treinamento ótimo demonstrou resultados superiores. Na média[†], entretanto, o MSC teve resultados 4,8% superiores ao treinamento ótimo e o MCM, resultados 5,6% superiores ao treinamento ótimo.

Um fato que merece destaque nos resultados apresentados é o baixo desempenho do modelo de Ming. Como destacado no texto compreendido entre as equações (3.15) e (3.19), embora esse modelo seja, de fato, um único GMM com grande número de componentes, ele pode ser interpretado como um modelo multicondicional sem algumas das restrições presentes nos modelos multicondicionais tradicionais. A maior flexibilidade do modelo de Ming, entretanto, não resulta em melhoria de desempenho de identificação; pelo contrário, a diferença média na taxa de identificações corretas entre o modelo de Ming e o MCM é de 24,5%. Para possibilitar uma visão mais detalhada da diferença de desempenho entre o método de Ming e o treinamento ótimo (TO), foi traçado o gráfico da Figura 3.9, que contém as diferenças nas taxas de identificações corretas entre esses dois métodos (treinamento ótimo menos método de Ming) para cada uma das condições de ruído analisadas. Foram ainda adicionados, nesse gráfico, traços com a diferença de desempenho entre o TO e o MCM e entre o TO e o MSC.

[*] Deve-se ainda destacar que o MCM, por utilizar efetivamente apenas um GMM nos cálculos para a determinação do modelo de locutor mais ajustado aos dados, permite a elaboração de técnicas que minimizam o custo computacional do processo, como será descrito no capítulo 4.

[†] Deve-se destacar que, como os valores de SNR escolhidos para a realização das simulações foi arbitrário, o valor da média não contém realmente um significado no sentido absoluto, pois, se forem alterados os valores de SNR utilizados, o valor dessa média certamente será modificado. Dessa maneira, devem também ser sempre observados os resultados para cada valor de SNR.

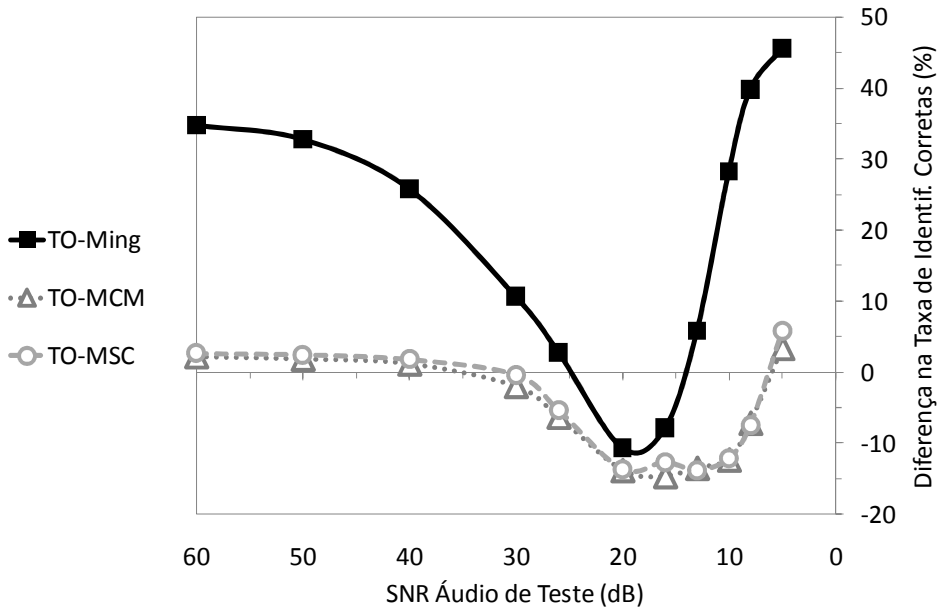

Figura 3.9: Diferenças de desempenho de sistemas de RAL multicondicionais: TO - método de Ming, TO - MCM e TO - MSC.

Como se pode observar no gráfico da Figura 3.9, em todos os casos, o treinamento ótimo teve desempenho superior aos demais modelos multicondicionais nas faixas extremas de ruído muito baixo ou muito alto, o que é ilustrado pelos valores positivos das diferenças traçadas. Na zona de ruído médio, entretanto, o treinamento ótimo apresentou resultados inferiores aos demais métodos.

No caso particular do modelo de Ming, observa-se claramente que somente houve ganhos para as SNR de 20 dB e 16 dB, havendo aumento progressivo da degradação à medida em que se afasta dessa faixa central. Esse comportamento pode ser uma indicação de que, como alertado anteriormente, houve uma supervalorização das componentes gaussianas que modelaram o áudio de ruído médio, em detrimento das componentes que modelaram as condições extremas. Essa distorção ocorre porque a modelagem de todas as condições de ruído é realizada simultaneamente e, mesmo havendo áudio com níveis de ruído alto e baixo, há uma predominância de situações de ruído médio.

Os resultados com o MCM e com o MSC também apresentaram ganhos com relação ao treinamento ótimo apenas para a zona central de ruído. Contudo, nesses casos, a zona de ganho foi mais larga, de 30 dB a 8 dB, e a degradação nas faixas extremas foi muito menor que a observada no modelo de Ming.

3.3.2 Análise da Perda de Desempenho com o Método da Condição Máxima

O ganho de desempenho do sistema baseado no MCM em comparação com o TO, observado na região de SNR intermediário, é o comportamento esperado, pois, ao menos no caso do MCM, é certo que a verossimilhança $p(Y \mid s)$ da equação (3.9) para o MCM será sempre maior que a obtida com o TO. Isso ocorre porque, com o método da condição máxima (MCM), a cada janela de análise, é escolhido o resultado do modelo treinado com a condição que maximiza o resultado de $p(\vec{y}_t \mid \lambda_{s,n})$, de acordo com a equação (3.13). Portanto, a cada vetor de parâmetros (ou janela de análise), tem-se que:

$$p(\vec{y}_t \mid s)_{MCM} \geq p(\vec{y}_t \mid s)_{TO} \tag{3.21}$$

Consequentemente, para o áudio integral, tem-se:

$$p(Y \mid s)_{MCM} \geq p(Y \mid s)_{TO} \tag{3.22}$$

Dessa forma, a perda de desempenho de identificação verificada para o MCM nas faixas extremas de SNR não pode ser atribuída a problemas na modelagem[*]. De fato, a única possível explicação para essa perda de desempenho seria o fato de, com o MCM, o aumento da verossimilhança $p(Y \mid s)$ para o locutor correto, $\tilde{s}$, ser inferior ao que ocorre para os locutores incorretos, s':

$$\begin{aligned} &p(Y \mid \tilde{s})_{MCM} - p(Y \mid \tilde{s})_{TO} \leq p(Y \mid s')_{MCM} - p(Y \mid s')_{TO} \\ &\Delta p(Y \mid \tilde{s}) \leq \Delta p(Y \mid s') \end{aligned} \tag{3.23}$$

Desse modo, embora não se viole a condição estabelecida em (3.22), a identificação do locutor, dada pela maximização expressa em (3.1) e repetida a seguir por conveniência:

$$\tilde{s} = \arg\max_{\lambda_s \in U} \sum_{t=1}^{T} \log p(\vec{y}_t \mid \lambda_s) \quad , \tag{3.24}$$

pode ser alterada de modo a se identificar incorretamente, com o MCM, trecho de áudio que o sistema com o TO identificou com sucesso.

Pode-se teorizar que esse fenômeno ocorre porque, no caso do locutor correto, como o modelo do TO está bem ajustado aos dados, não há grandes possibilidades de ganhos com a

[*] No sentido de os modelos não representarem bem os locutores e, consequentemente, levarem a valores baixos de probabilidades $p(Y \mid s)$.

utilização dos demais modelos, treinados com outras condições de ruído. De fato, existe algum aumento no valor da verossimilhança calculada com o uso do MCM, mesmo para o locutor correto, pois o treinamento ótimo gera um modelo ajustado ao nível de ruído médio do áudio de treinamento, o que não é o modelo mais adequado a cada uma das janelas de análises do áudio questionado consideradas individualmente, visto que as variações normais na intensidade da voz ocorridas durante a fala fazem com que a SNR instantânea varie ao longo do áudio. Contudo, para os locutores incorretos, como o modelo não está bem ajustado aos dados, a utilização de outros $N - 1$ modelos, treinados com outras condições de ruído, promoveria mais chances de que, casualmente, um desses novos modelos esteja mais bem ajustado aos dados que o modelo inicial.

3.3.2.1 Minimizando os Efeitos da Perda de Desempenho

Tendo em vista a perda de desempenho verificada com o MCM nos experimentos com áudio de SNR extremos, idealizou-se, durante este trabalho, uma forma de minimizar essa perda, ao menos para os casos de ruído muito elevado. A solução consiste na utilização, no treinamento do modelo multicondicional, de áudio com SNR menor do que o valor mínimo de SNR do áudio a ser identificado. Espera-se, com essa abordagem, melhorar o desempenho do sistema de identificação baseado em modelos multicondicionais e no MCM por duas razões. Inicialmente, espera-se uma melhora pelo deslocamento do que se considera "valores extremos" de SNR para além dos valores a serem testados, o que deve minimizar a perda de desempenho em função do discutido na seção 3.3.2.

Além disso, deve-se considerar que os valores de SNR calculados, tanto para os áudios utilizados para o treinamento dos modelos multicondicionais como para os áudios a serem identificados, são valores médios ao longo de todo o trecho considerado. Dessa forma, é possível que ocorram, no áudio questionado, momentos em que a intensidade da voz do locutor é menor que a média[*], quando, portanto, o valor instantâneo da SNR estará também abaixo do valor médio. Nesse caso, pode ocorrer desse valor estar além do valor mínimo de SNR utilizado no treinamento do modelo multicondicional, de modo que não haverá modelo especificamente treinado para essa condição. Assim, a inclusão, no modelo multicondicional, de modelos treinados com áudio mais degradado (em média) do que aqueles

[*] O mesmo efeito poderia ocorrer por flutuações na intensidade instantânea do ruído. Contudo, nos experimentos realizados ao longo deste trabalho, a intensidade do ruído foi sempre mantida constante.

que se pretende identificar torna o modelo mais apto a lidar com situações como a descrita, promovendo, em princípio, uma melhora no seu desempenho.

A fim de validar essa proposta de minimizar a degradação observada no desempenho do sistema, foram realizados análises com dois outros modelos multicondicionais, utilizando o MCM, e treinamento com áudio de SNR igual a 60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB, para o primeiro modelo, e SNR igual a 60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB, 3 dB e 2 dB, para o segundo. Destaque-se a inclusão da condição de SNR igual a 3 dB no treinamento do primeiro novo modelo multicondicional e a inclusão das condições de SNR iguais a 3 dB e 2 dB no treinamento do segundo modelo. Os resultados dessas avaliações estão expostos na Tabela 3.10, nas linhas MCM (3 dB) e MCM (2 dB). Nessa tabela, também foram incluídos, para facilitar as comparações, os resultados do sistema anterior, baseado MCM e treinado para valores de SNR até 5 dB, linha MCM (5 dB); e o resultado do treinamento ótimo (TO), ambos constantes da Tabela 3.9.

Tabela 3.10: Taxas de identificações corretas de sistemas de RAL multicondicionais utilizando MCM com treinamento até 5 dB SNR (MCM 5 dB), MCM com treinamento até 3 dB SNR (MCM 3 dB), MCM com treinamento até 2 dB (MCM 2 dB) e treinamento ótimo (TO).

Taxas de Identificações Corretas (%)		**SNR do Áudio de Teste (dB)**										
		60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
Modelagem Multicondicional	**MCM (5 dB)**	97,8	98,2	98,2	97,8	97,2	95,0	92,5	89,0	83,0	75,5	60,0
	MCM (3 dB)	97,8	98,2	98,2	97,8	97,2	95,5	94,2	91,5	86,6	80,8	68,5
	MCM (2 dB)	97,8	98,2	98,3	97,8	97,2	96,0	95,0	92,3	87,0	83,2	72,2
	TO	100,0	100,0	99,5	95,8	90,8	81,2	78,5	75,5	70,8	68,5	64,0

A fim de permitir uma melhor visualização, os resultados da Tabela 3.10 foram traçados no gráfico da Figura 3.10.

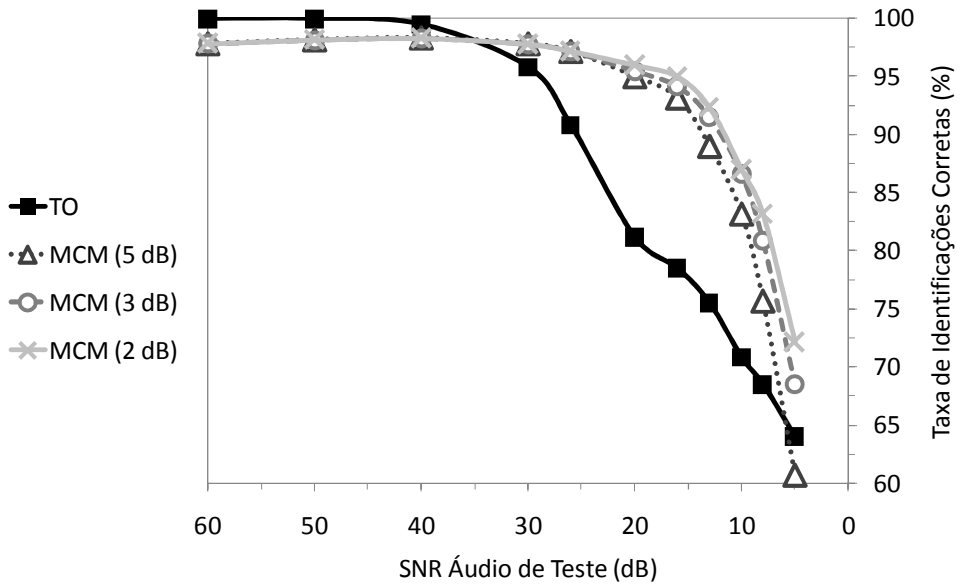

Figura 3.10: Comparação do desempenho de sistemas de RAL com TO, 11 condições de treinamento, e MCM, com condições de treinamento diversas.

Como se pode observar, a inclusão da condição de treinamento de SNR igual a 3 dB promoveu uma significativa melhora no desempenho do sistema MCM, não apenas para valores "extremos" de SNR, mas para todos os valores SNR inferiores a 20 dB; e isso sem causar qualquer prejuízo nas demais situações. Vale também mencionar que a utilização da condição de treinamento de SNR igual a 3 dB fez com que o modelo MCM tivesse desempenho superior ao treinamento ótimo para a condição de teste de SNR igual a 5 dB, o que não ocorria no modelo multicondicional treinado sem essa condição. A inclusão da condição de treinamento de SNR igual a 2 dB promoveu ainda mais alguma melhora em toda a faixa de 20 dB até 5 dB, em comparação com os resultados do modelo multicondicional treinado até 3 dB.

Deve-se destacar que a inclusão de novas condições de treinamento torna os modelos multicondicionais mais complexos computacionalmente, aumentando o esforço necessário para o processo de identificação. Contudo, os métodos desenvolvidos ao longo da elaboração desta tese, particularmente o dos Modelos Multicondicionais Adaptativos (MMA), apresentados na seção 4.3, apresentam uma solução para essa questão.

4 TÉCNICAS EFICIENTES DE RECONHECIMENTO AUTOMÁTICO DE LOCUTOR

A utilização de modelos multicondicionais, como visto na seção 3.3, é um meio eficaz para aumentar o desempenho dos sistemas de RAL com relação a variações na qualidade do áudio questionado (Matsui, Kanno e Furui, 1996; Yang e Gong, 2006),. Embora esse tipo de modelagem proporcione ganhos significativos no desempenho dos sistemas de RAL, ele provoca também um aumento no custo computacional associado às tarefas de identificação, que cresce linearmente com o número de condições de treinamento utilizadas. Dessa maneira, foram feitos, ao longo do desenvolvimento desta tese, estudos buscando métodos de construir sistemas que mantivessem a característica de robustez aos ruídos dos modelos multicondicionais, mas que, ao mesmo tempo, tivessem seu custo computacional reduzido.

O custo computacional associado a uma tarefa de identificação de locutor é função do número de locutores do universo, *S*, pois todos os modelos terão que ser simulados para determinar aquele que maximiza a probabilidade $p(Y \mid s)$ da equação (3.7); da duração do trecho de áudio questionado, uma vez que os vetores de parâmetros da voz, $\vec{y}_t$, são extraídos a intervalos de tempo fixos; da dimensão das gaussianas utilizadas no modelo, *D*; e do número total de componentes gaussianas dos modelos, que, no caso de modelos multicondicionais, é o produto do número de condições de treinamento, *N*, pelo número de componentes em cada modelo unicondicional, *M*.

Com exceção do número de locutores do universo, *S*, que não pode ser alterado para uma dada aplicação, e do número de condições de treinamento, *N*, todos os demais parâmetros foram estudados no sentido de minimizar o custo computacional associado à tarefa de identificação por Reynolds e Rose (1995). Em particular com relação ao número de componentes dos modelos, *M*, Reynolds e Rose (1995) determinaram que são necessárias, no mínimo, de 8 a 16 para o bom desempenho do sistema, no caso de áudio sem ruído, fato confirmado durante estudos preliminares à elaboração deste trabalho. Com relação ao número de condições de treinamento, *N*, observou-se, na seção 3.3.2.1, que o desempenho do sistema cresce com o aumento do número de condições utilizadas, mesmo quando são incluídas condições de ruído além do valor médio do ruído do áudio questionado. Dessa

forma, em princípio, qualquer redução no número de condições de treinamento do modelo ocasiona perda de desempenho[*].

Observa-se, portanto, que, para se obter uma redução do custo computacional do processo de identificação, não se pode simplesmente reduzir a complexidade dos modelos utilizados, sob pena de ser afetado o desempenho do sistema. É possível, contudo, explorar propriedades do áudio ou do próprio modelo GMM multicondicional como forma de eliminar alguns cálculos que, em tese, não comprometem o resultado final da identificação. Nesse sentido, as técnicas eficientes propostas, implementadas, avaliadas e validadas neste capítulo guardam semelhanças em suas abordagens com as técnicas de codificação ou de compressão de dados comumente utilizadas.

São introduzidos, neste capítulo, métodos que exploram a correlação temporal das características do áudio questionado como o Método da Condição Persistente (MCP) e Modelos Multicondicionais Adaptativos (MMA). O Método da Condição Persistente é uma variante do Método da Condição Máxima (MCM) tradicionalmente utilizado que realiza a estimação da condição de treinamento que maximiza a verossimilhança com base em valores passados e futuros dessa condição máxima. Dessa maneira, somente é realmente usado o modelo multicondicional em algumas janelas de análise, nas quais é determinada a condição máxima do mesmo modo que se faz no MCM. Nas demais janelas, a condição máxima é estimada a partir dos valores anteriormente calculados e, dessa forma, nessas janelas, utiliza-se de fato apenas um modelo unicondicional. Os Modelos Multicondicionais Adaptativos (MMA), apesar de também se fundamentarem na correlação temporal da qualidade do áudio, não se propõem a estimar diretamente a condição de treinamento que maximiza a verossimilhança. Esse método apenas admite que, entre janelas de análise adjacentes, a qualidade do áudio varia de forma limitada. Dessa maneira, uma vez conhecida a condição de treino que maximiza a verossimilhança em uma janela, não é necessário calcular todos os modelos multicondicionais para a subsequente. Basta, nesse caso, calcular os modelos vizinhos (mais próximos) ao anterior.

Esses dois métodos admitem como hipótese a existência de correlação temporal das características do áudio questionado. Essa é uma condição normalmente atendida e, por esse

[*] Durante a elaboração desta tese, foram feitos experimentos aumentando o intervalo entre os níveis de ruído das condições de treinamento. Verificou-se que é possível utilizar intervalos maiores que os usualmente empregados, de 2 dB. Essa solução, contudo, é limitada em termos de redução potencial do custo computacional total do processo. Além disso, verificou-se que o intervalo ótimo a ser empregado depende de diversos fatores como: intensidade do ruído, frequência de amostragem e codificação do áudio etc. Dessa maneira, esse tipo de abordagem foi descartada.

motivo, foram propostas essas técnicas para a redução do custo das tarefas de identificação. Observe-se que está se admitindo a simples correlação temporal da qualidade do áudio, ou seja, a correlação temporal da relação sinal-ruído. Como as fontes de ruído normalmente presentes nesse tipo de aplicação têm intensidade relativamente constante, como condicionadores de ar ou ventiladores, ou de variação lenta[*], como motores de automóveis, a condição imposta pelos métodos propostos não deve afetar significativamente o desempenho do sistema.

É também introduzido, neste capítulo, um método que explora a coerência entre os vários GMMs que compõem um modelo multicondicional: o Método das Gaussianas Dominantes (MGD). O Método das Gaussianas Dominantes foi idealizado com o objetivo de explorar uma característica típica dos modelos multicondicionais, que é a existência de vários modelos representando dados relativamente semelhantes. De fato, os modelos não representam a mesma informação, pois são treinados com áudio de qualidades diferentes; entretanto, a variação na qualidade do áudio é gradual, e, dessa forma, modelos adjacentes são treinados com dados razoavelmente semelhantes[†]. Tomando essa semelhança entre os dados utilizados para o treinamento dos modelos, pode-se construir um modelo multicondicional formado por GMM cujas componentes gaussianas correspondentes modelam regiões "próximas" do espaço de parâmetros[‡]. Essa "proximidade" entre as componentes dos diferentes GMM é explorada pelo MGD para, após calcular um GMM completo para uma determinada janela de análise (calculando o valor de todas as componentes gaussianas), determinar quais são as componentes mais significativas na composição do valor da verossimilhança do modelo e, dessa forma, descartar as demais componentes em todos os outros modelos multicondicionais do mesmo locutor.

Por fim, é apresentada uma metodologia que utiliza eliminações progressivas dos locutores no processo de identificação, os Modelos de Mistura de Gaussianas Multirresolução (MR-GMM). Essa técnica, na realidade, pode ser igualmente aplicada a modelos multicondicionais ou unicondicionais, e se fundamenta na possibilidade de utilizar modelos mais simplificados numa primeira etapa com o único objetivo de eliminar os locutores cujas vozes são muito distintas da presente no trecho questionado. Com isso, os modelos mais detalhados e

[*] Comparadas à duração de uma janela de análise, cerca de 20 ms.
[†] Em geral, utilizam-se intervalos de SNR de 2 dB a 5 dB entre treinamentos de um modelo multicondicional.
[‡] Os modelos normalmente utilizados em sistemas de RAL não têm essa propriedade; contudo, com algumas alterações no processo de treinamento, é possível estabelecer essa relação.

mais complexos são utilizados apenas na avaliação de um número restrito de locutores, fazendo com que a complexidade média do sistema seja reduzida.

Os Modelos de Mistura de Gaussianas Multirresolução não dependem da coerência temporal da qualidade do áudio nem também de qualquer coerência entre os modelos multicondicionais. Seu funcionamento se baseia unicamente na representação dos locutores por uma modelagem progressivamente mais detalhada. Com isso, realiza-se um processo de identificação também é progressivo.

4.1 CUSTO COMPUTACIONAL DA IDENTIFICAÇÃO

Para permitir uma melhor análise da redução do custo computacional obtido com cada um dos métodos a serem a seguir detalhados, é interessante, antes, estudar a relação entre o número de componentes gaussianas que compõem um modelo e o custo computacional da identificação.

Pela definição dos GMM, equação (2.1), constata-se que o custo computacional total de uma tarefa de identificação, *W*, é, dentro de certas condições, a seguir detalhadas, proporcional a *M*, o número de gaussianas nos modelos dos locutores, λ.

$$W_{GMM} \propto M \tag{4.1}$$

Para o caso multicondicional, equação (3.7), o número total de componentes gaussianas é $N \cdot M$, o produto do número de gaussianas de cada modelo, *M*, pelo número de condições de treinamento, *N*.

Observe-se que, para cada componente gaussiana existente nos modelos, é necessário o cálculo de uma nova gaussiana, equação (2.2), a multiplicação pelo coeficiente da mistura e a soma desse produto para a totalização da probabilidade do modelo, equação (2.1). Os custos computacionais com a normalização temporal e com a identificação do locutor que maximiza a probabilidade do áudio (função *argmax*), expressos na equação (2.14), são independentes do número de componentes dos modelos. Esses custos são pouco significativos, pois os cálculos são executados apenas uma vez para cada locutor, no caso da normalização temporal, e apenas uma vez para todo o processo, no caso da identificação do máximo, enquanto que os cálculos das expressões (2.1) e (2.2) são executados uma vez para cada vetor de parâmetros, $\vec{y}_t$, e para cada locutor. Consequentemente, mesmo que

sejam utilizados trechos de curta duração, com apenas 1 s de comprimento[*], serão realizados 50 cálculos[†] relativos às expressões (2.1) e (2.2) para cada cálculo relativo à expressão (2.14). O custo com o cálculo do logaritmo também independe do número de componentes e, apesar de ser realizado a cada vetor de parâmetros, é pouco significativo se comparado aos demais, pois é uma única operação escalar. No caso multicondicional, considerando o Método da Condição Máxima (MCM), há, ainda, o cálculo de outra função *argmax* para cada vetor de parâmetros, $\vec{y}_t$, de acordo com a equação (3.13). Entretanto, essa é uma identificação de um máximo entre os N modelos multicondicionais de um único locutor. Como N é relativamente pequeno, da ordem de 10, esse custo também não é muito significativo.

Dessa maneira, dentro de certas condições normalmente atendidas pelos sistemas de RAL, pode-se estabelecer, com relativa precisão, que o custo da identificação num modelo GMM multicondicional tradicional (W_{MCM}) é proporcional à quantidade total de componentes gaussianas dos modelos:

$$W_{MCM} \propto N \cdot M \tag{4.2}$$

Durante o restante deste capítulo, serão apresentados métodos de redução do custo computacional das tarefas de identificação desenvolvidos durante a elaboração desta tese. Em todos os casos, serão utilizadas, como medida de redução da complexidade computacional, comparações entre o número total efetivo de gaussianas a serem computadas, tendo em vista a aproximação da equação (4.2).

Para facilitar a visualização dos métodos propostos, será utilizada a seguinte representação gráfica para indicar os modelos GMM e suas componentes:

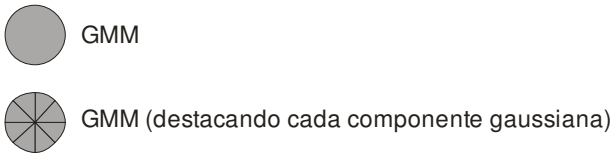

Figura 4.1: Representação gráfica dos GMM.

[*] Normalmente, são utilizados arquivos com mais de 10 s, o que aumenta ainda mais a relação entre o número de cálculos.

[†] Considerando que a extração dos parâmetros foi realizadas em janelas de 20 ms sem sobreposição, condição tipicamente utilizada nesse tipo de aplicação.

Para representar os modelos multicondicionais, será utilizada a seguinte representação gráfica:

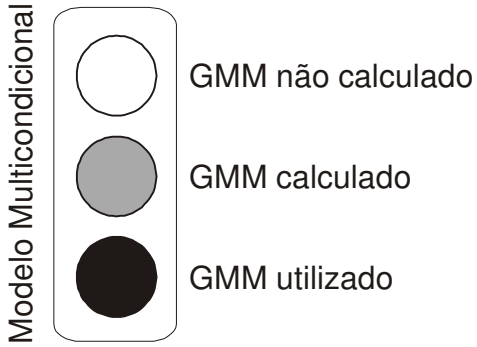

Figura 4.2: Representação gráfica de um GMM multicondicional.

Os círculos (setores circulares) brancos simbolizam GMM (componentes gaussianas) constantes do modelo multicondicional, mas que não foram calculados; círculos (setores circulares) cinza simbolizam GMM (componentes gaussianas) que foram calculados e círculos (setores circulares) pretos simbolizam GMM (componentes gaussianas) que foram calculadas e que foram utilizadas para a determinação da verossimilhança do modelo.

Seguindo essa simbologia, a representação de uma identificação utilizando o MCM seria a exposta na Figura 4.3:

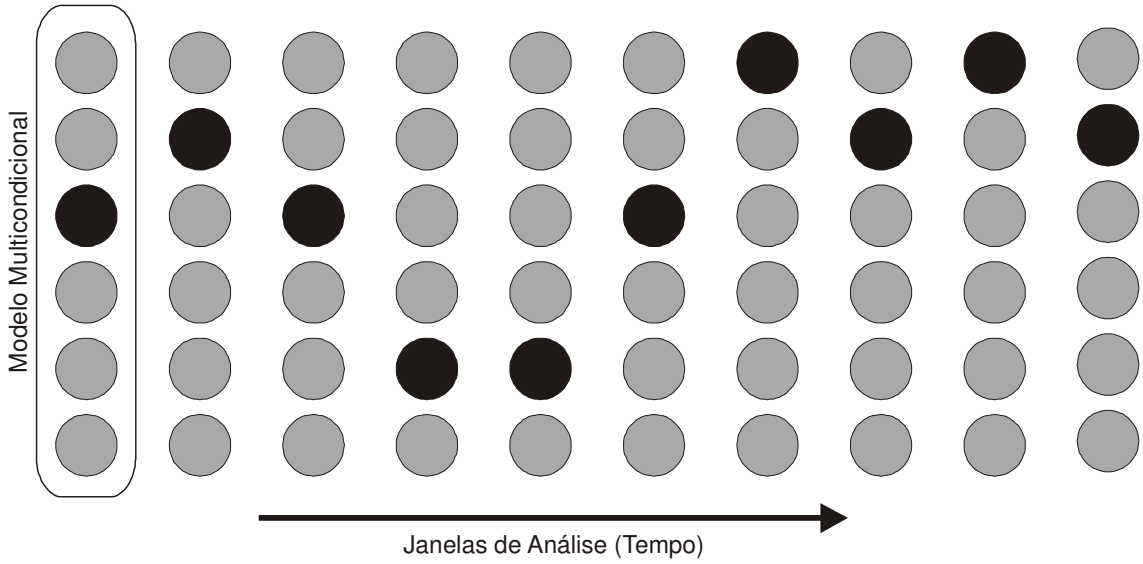

Figura 4.3: Ilustração do Método da Condição Máxima (MCM).

Como se verifica, no MCM, todos os modelos GMM são calculados para todas as janelas de análise, o que gera o grande custo computacional da identificação.

4.2 MÉTODO DA CONDIÇÃO PERSISTENTE

Uma forma de diminuir o custo computacional das tarefas de identificação de locutores em sistemas com modelos multicondicionais é a diminuição do número de condições de treinamento utilizadas. Contudo, para que se mantenha um bom nível de robustez a condições variáveis de ruído, é necessário que se mantenha um número razoavelmente elevado de condições de treinamento no modelo multicondicional, de modo que a simples redução do número de condições de treinamento é uma solução limitada. Além disso, como observado na seção 3.3.2.1, a inclusão de novas condições de treinamento com valores de SNR abaixo daqueles do áudio questionado promove um significativo aumento no desempenho da identificação, de modo que a opção pela simplificação do modelo multicondicional é, de fato, inviável.

Entretanto, mesmo se forem mantidas todas as condições de treinamento, há formas de se reduzir o valor efetivo de N na equação (4.2) se forem exploradas características como a forte correlação temporal que existe no áudio a ser identificado. Pode-se, por exemplo, considerando que as janelas de análise do áudio questionado são relativamente pequenas (usualmente 20 ms) se comparadas ao tempo de variação da condição de ruído do áudio, reutilizar por mais de uma janela a condição de ruído que resultou na máxima verossimilhança. Dessa forma, de fato, apenas em algumas janelas de análise se utilizaria o modelo multicondicional; nas demais, seriam calculados apenas os modelos unicondicionais correspondentes à condição de treinamento máxima previamente determinada. Com isso, na realidade, se está explorando a correlação existente nos sinais de áudio questionados para estimar as condições do áudio em janelas de análise futuras a partir do conhecimento das condições na janela de análise presente.

Nesse sentido, considerando a formulação dos modelos multicondicionais, é proposta a utilização do Método da Condição Persistente (MCP), que consiste na modelagem multicondicional com o MCM, equação (3.13), com a reutilização da condição determinada por mais de uma janela de análise, ou seja, para mais de um vetor de parâmetros. Matematicamente, o MCP pode ser expresso por:

$$
\begin{aligned}
&p\left(\lambda_{s,n} \mid s, \vec{y}_{kp+1}\right) = p\left(\lambda_{s,n} \mid s, \vec{y}_{kp+1+q}\right) = 1, se\ n = \tilde{n}\left[kp+1\right] = \arg\max_{n=0,\ldots,N} p\left(\vec{y}_{kp+1} \mid \lambda_{s,n}\right) \\
&p\left(\lambda_{s,n} \mid s, \vec{y}_{kp+1}\right) = p\left(\lambda_{s,n} \mid s, \vec{y}_{kp+1+q}\right) = 0, se\ n \neq \tilde{n}\left[kp+1\right] \\
&\forall k = 0,1,2,.... \\
&\forall q = 1,..., p-1
\end{aligned}
\quad , \tag{4.3}
$$

sendo p uma número natural constante denominado de persistência.

A vantagem computacional do MCP consiste no fato de que a análise dos diferentes modelos multicondicionais somente ocorre a cada p vetores de parâmetros[*]. Dessa maneira, somente a cada p janelas de análise é feito o cálculo de todas as $N \cdot M$ gaussianas do modelo multicondicional. Nas demais *p-1* janelas, somente são calculadas as M componentes do modelo da condição previamente determinada, como ilustrado graficamente na Figura 4.4.

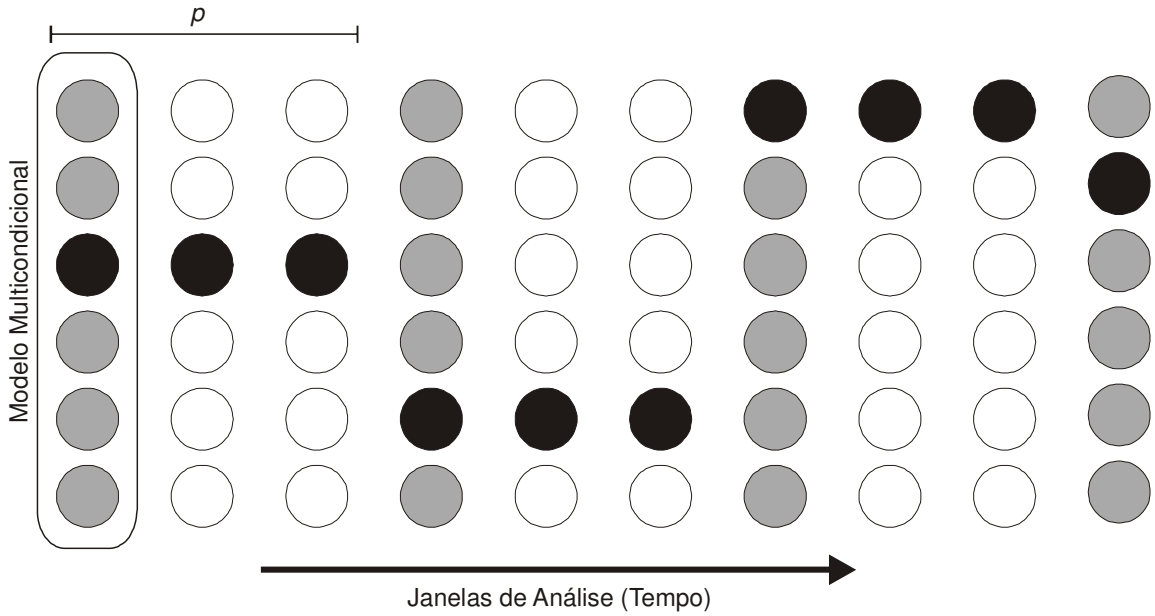

Figura 4.4: Ilustração do Método da Condição Persistente (MCP).

Em adição à ilustração da Figura 4.4, é inserida no quadro da Figura 4.5, a programação básica do MCP.

[*] De fato, só há alteração se $p \geq 2$, pois para $p = 1$ tem-se simplesmente o MCM descrito na seção 3.3.

```
p(Y|λs) = 0 % verossimilhança do modelo multicondicional do
locutor s dado o conjunto de vetores de parâmetros Y
Para t = 1 a T, passo tamanho p % loop de janelas de análise
  Para n de 1 a N % loop de condições de treinamento
    Calcule p(y[t]|λ[s,n]) % verossimilhança do modelo do
    locutor s para a condição n dado o vetor de parâmetros
    y[t]
  Fim para n
  n'[t] = argmax(p(y[t]|λ[s,n]) em relação a n) % determinar
  a condição de treinamento que maximiza a verossimilhança
  p(Y|λs) = p(Y|λs) + p(y[t]|λ[s,n'[t]])
  Para t2 = 1 a (p - 1)
    Calcule p(y[t+t2]|λ[s,n'[t]])
    p(Y|λs) = P(Y|λs) + p(y[t+t2]|λ[s,n'[t]])
  Fim para t2
Fim para t
```

Figura 4.5: Programação básica do Método da Condição Persistente (MCP).

Dessa maneira, o custo computacional (número de gaussianas) médio por janela de análise do MCP passa a ser:

$$W_{MCP}(p) \propto M\left(\frac{N+(p-1)}{p}\right) = M\left(1+\frac{N-1}{p}\right) \tag{4.4}$$

Quanto maior o valor de p, menor o custo computacional, sendo que:

$$\lim_{p\to\infty} W_{MCP}(p) = M \tag{4.5}$$

Ou seja, no limite, o custo computacional é o custo do cálculo do modelo unicondicional. Na realidade, não faz sentido aumentar demasiadamente o valor de p, ou se estaria se utilizando um modelo mais parecido com o treinamento ótimo ou com o modelo de (Xu, Dalsgaard e Lindberg, 2005), expresso na equação (3.11). A definição do valor da persistência, p, a ser utilizada numa situação prática depende fundamentalmente da característica do ruído do áudio questionado. Entretanto, em geral, mesmo para condições de ruído constante, valores de p superiores a 10 não trazem reduções significativas no custo computacional e, considerando que o tempo usual para a janela de análise de áudio é de 20 ms, fixam a condição de ruído por 0,2 s, o que relativamente elevado para algumas aplicações. Dessa maneira, entende-se que, em geral, o valor de p a ser utilizado deve estar entre 2 e 10.

A redução do custo computacional em comparação ao modelo multicondicional tradicional (MCM) é dada por:

$$W_{MCM} - W_{MCP}(p) \propto M \cdot N - M\left(\frac{N+(p-1)}{p}\right) = M(N-1)\left(\frac{p-1}{p}\right), \qquad (4.6)$$

e diferença relativa entre os custos computacionais é calculada por:

$$\frac{\Delta W}{W} = \frac{W_{MCM} - W_{MCP}(p)}{W_{MCM}} \propto \left(\frac{N-1}{N}\right)\left(\frac{p-1}{p}\right) \qquad (4.7)$$

Observa-se, portanto, que a diferença relativa de custo computacional não depende do número de gaussianas utilizadas em cada modelo unicondicional, *M*, e que também não é muito sensível ao parâmetro *N*, salvo para valores de *N* muito pequenos, que não são usuais (normalmente, um modelo multicondicional é composto por cerca de dez ou mais modelos unicondicionais). Por outro lado, a sensibilidade com relação ao parâmetro da persistência da condição, *p*, é muito grande para valores pequenos de *p*. Os gráficos da Figura 4.6 e da Figura 4.7 ilustram a queda do custo computacional absoluto e relativo com o uso da condição persistente para os casos de modelos multicondicionais de 10 e de 20 condições de treinamento (*N*), para *p* = 1, 2, ...,10. No caso do gráfico da Figura 4.6, a escala vertical utilizada é em número de vezes *M*, a quantidade de componentes gaussianas de cada modelo unidimensional. No caso do gráfico da Figura 4.7, a escala vertical utilizada é em relação ao custo computacional original sem utilização da condição persistente.

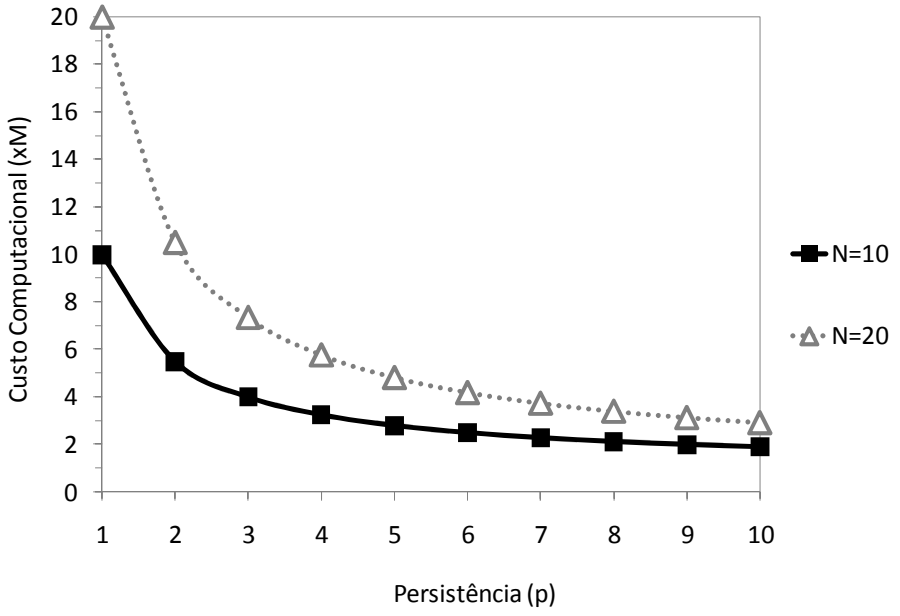

Figura 4.6: Custo computacional absoluto de sistemas com MCP, em função da persistência, *p*, para diferentes números de condições de treinamento, *N*.

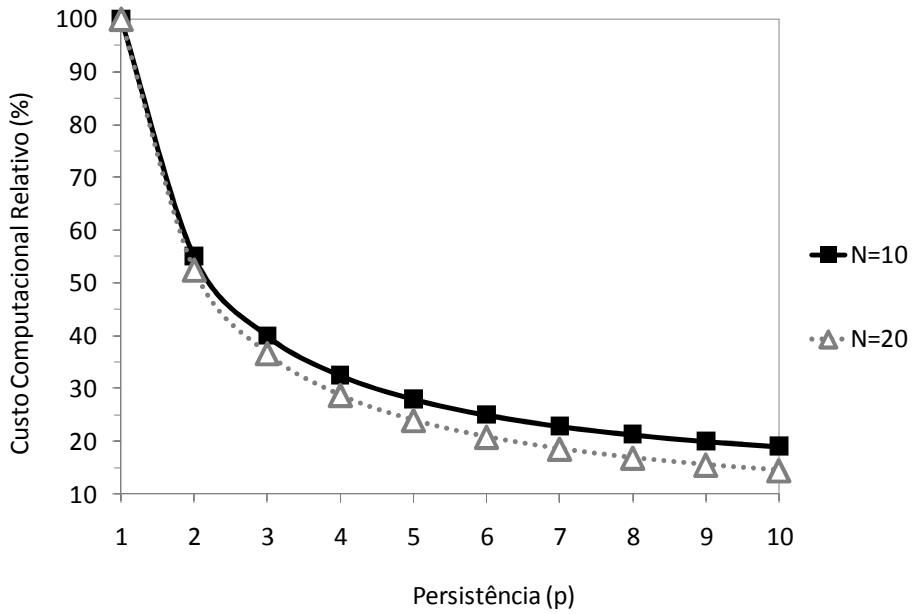

Figura 4.7: Custo computacional relativo de sistemas com MCP em função da persistência, *p*, para diferentes números de condições de treinamento, *N*.

Como se observa no gráfico da Figura 4.7, e como era esperado, pela forma da equação (4.4), a queda no custo computacional é mais acentuada para valores pequenos de *p*, por causa do termo $(p-1)/p$. Também se verifica que não há grandes diferenças para os casos de $N = 10$ e de $N = 20$, pois o termo dependente de *N*, $(N-1)/N$, varia pouco para valores grandes de N ($N > 10$).

A fim de validar o método da condição persistente proposto, foram realizadas simulações com o modelo multicondicional usado na seção 3.3.1 (12 condições de treinamento: 60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB), utilizando valores de *p* iguais a 1, 2, 3, 5 e 10. Os resultados desses procedimentos são apresentados na Tabela 4.1.

Tabela 4.1: Taxas de identificações corretas de sistemas de RAL multicondicional utilizando o MCP, para diferentes valores de persistência, *p*.

Taxas de Identificações Corretas (%)		SNR do Áudio de Teste (dB)										
		60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
Persistência (*p*)	**1***	97,8	98,2	98,2	97,8	97,2	95,5	94,2	91,5	86,6	80,8	68,5
	2	98,2	98,7	98,0	97,0	96,2	94,3	92,8	91,2	88,0	83,2	76,8
	3	99,5	99,5	98,7	96,2	94,3	92,5	90,7	87,0	82,7	79,2	74,8
	5	98,3	98,3	97,7	94,7	93,0	88,7	86,0	78,8	77,8	74,2	71,3
	10	96,3	96,5	95,8	91,8	89,5	83,3	78,3	72,0	71,3	67,7	66,3

Para permitir uma melhor visualização, os resultados da Tabela 4.1 foram traçados no gráfico da Figura 4.8.

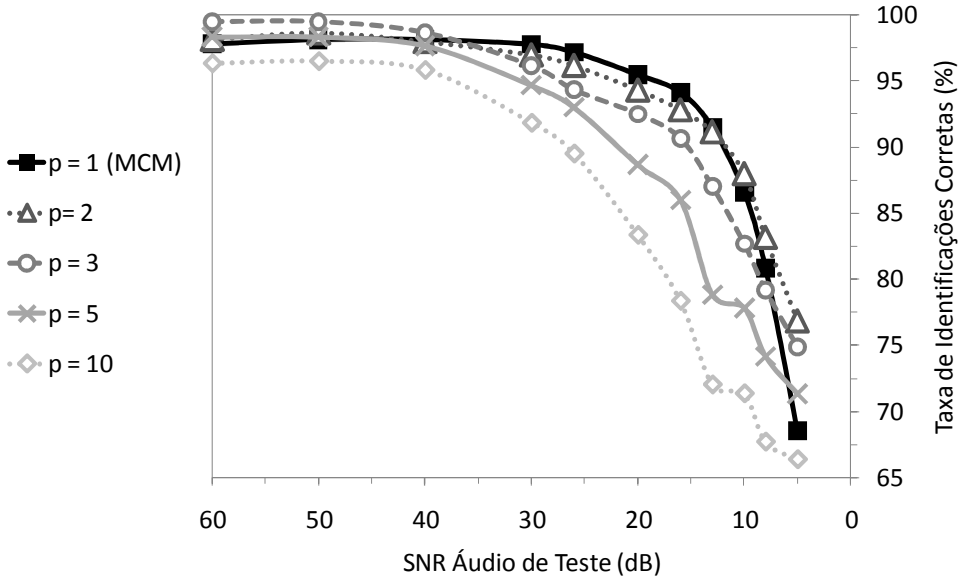

Figura 4.8: Comparação de desempenho de sistemas de RAL multicondicional utilizando MCP, para diferentes valores de persistência, *p*.

Como se pode verificar, a utilização da condição persistente, em geral (para $p \leq 5$), promove uma melhoria no desempenho com relação ao mesmo sistema sem persistência ($p = 1$) para as faixas de SNR extremas alta e baixa. Para a zona de SNR intermediário, o sistema com MCP apresentou sempre desempenho inferior ao sistema sem persistência (MCM).

* Novamente se destaca que o uso de $p = 1$ reduz o MCP ao MCM, não havendo qualquer vantagem computacional. Contudo, os valores estão aqui exibidos para fins de comparação de desempenho.

Observa-se ainda que a utilização da persistência $p = 10$ provocou uma degradação de desempenho significativa em todas as condições de ruído analisadas, de modo que tentativas de reduzir ainda mais o custo computacional pela utilização de valores mais elevados de p não foram realizadas.

Como a utilização da condição persistente promove, em alguns casos, melhora no desempenho do sistema e, em outros, degradação, a depender do nível de ruído utilizado, foram calculadas as médias das taxas de identificação corretas, a fim de observar o resultado global do uso da persistência no desempenho do sistema de RAL. As taxas de identificações corretas médias para cada valor da persistência, p, são apresentadas na Tabela 4.2[*].

Tabela 4.2: Taxa média de identificações corretas de sistemas de RAL multicondicional utilizando método da condição persistente (MCP).

Persistência (p)	**Taxa Média de Identificações Corretas (%)**
1	91,5
2	92,2
3	90,5
5	87,2
10	82,6

Para permitir uma melhor visualização, os resultados da Tabela 4.2 foram traçados na Figura 4.9.

[*] Deve-se destacar que, como a escolha dos valores de SNR para a realização das simulações foi arbitrária, o valor da média apresentada na Tabela 4.2 não contém um significado absoluto no sentido de se afirmar que o uso do MCP promove um "ganho" ou uma "perda" com relação ao sistema sem condição persistente. Como houve condições de ruído em que o sistema com o MCP teve desempenho superior ao tradicional e outros em que o MCP teve desempenho inferior, é possível, por meio de uma escolha dos valores de SNR a serem testados, fazer o valor médio dos testes ser superior ou inferior ao do modelo tradicional. Dessa forma, esses valores devem ser sempre observados considerando os resultados individuais exibidos na Tabela 4.1.

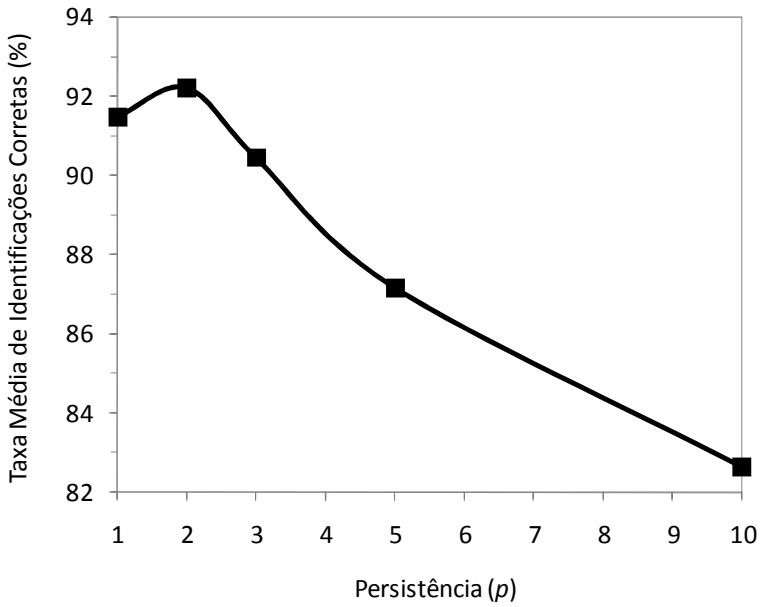

Figura 4.9: Taxa média de identificações corretas de sistemas de RAL multicondicional MCP, em função do valor de persistência, p.

Como se observa, o uso da condição persistente $p = 2$ promoveu um leve ganho na taxa média de identificações corretas, 0,7%[*], apesar de reduzir em mais de 45% o custo computacional do processo. O uso da condição persistente $p = 3$, resultou numa pequena diminuição da taxa média de identificações corretas, 1%, ao passo que reduziu o custo computacional em mais de 61%. A utilização da condição persistente $p = 5$ levou a perdas significativas no desempenho médio do sistema, 4,3%, para uma redução do custo computacional superior a 73%. O caso da condição persistente $p = 10$ provocou uma degradação ainda mais acentuada, 8,8%, para redução de custo computacional superior a 82%.

A análise dos resultados demonstra que é possível utilizar o MCP para reduzir o custo computacional em até 61% ($p = 3$) desde que se aceite uma pequena queda no desempenho do sistema de RAL, 1% em média. Se as exigências com relação ao desempenho do sistema forem mais rigorosas, ainda é possível reduzir o custo computacional em mais de 45% ($p = 2$) e obter um ganho na taxa de identificações corretas de 0,7%, na média.

4.2.1 Método da Condição Persistente Linearmente Variada

Um aspecto do MCP que pode ser aprimorado a fim de tentar melhorar os resultados obtidos é a evolução da condição de treinamento utilizada entre as janelas de análise em que os modelos de todas as condições são reavaliados. No MCP descrito na equação (4.3), a con-

[*] Ver nota anterior sobre os resultados da Tabela 4.2.

dição de treinamento utilizada permanece constante em todo o intervalo. Em princípio, o desempenho do sistema poderia ser melhorado se a condição de treinamento variasse de forma gradual (linear) entre os dois pontos em que é feita a análise da condição máxima, pois, dessa maneira, se estaria refinando a estimativa de evolução temporal da condição do áudio, passando de uma predição de ordem zero (constante), utilizada no MCP, para uma predição de primeira ordem. Nesse sentido, foi introduzido o Método da Condição Persistente Linearmente Variada (MCP-LV), que consiste numa alteração do MCP anteriormente definido de forma que a condição de treinamento utilizada a cada janela varia linearmente de acordo com:

$$p\left(\lambda_{s,n} \mid s, \vec{y}_{kp+1+q}\right) = 1, se\ n = \tilde{n}\left[kp+1+q\right] = \text{round}\left\{\left(1-\frac{q}{p}\right)\tilde{n}\left[kp+1\right]+\left(\frac{q}{p}\right)\tilde{n}\left[\left[k+1\right]p+1\right]\right\}$$

$$p\left(\lambda_{s,n} \mid s, \vec{y}_{kp+1+q}\right) = 0, se\ n \neq \tilde{n}\left[kp+1+q\right]$$

$$\forall k = 0,1,2,....$$

$$\forall q = 0,..., p-1$$

, (4.8)

sendo

$$\begin{aligned}\tilde{n}\left[kp+1\right] &= \arg\max_{n=0,...,N} p\left(\vec{y}_{kp+1} \mid \lambda_{s,n}\right)\\ \tilde{n}\left[kp+1+q\right] &= \arg\max_{n=0,...,N} p\left(\vec{y}_{(k+1)p+1} \mid \lambda_{s,n}\right)\end{aligned} \quad (4.9)$$

Deve-se destacar que, para a utilização do MCP-LV nas janelas de análise $kp + 1+q$, $q= 1,..., p - 1$, é necessário determinar o valor das condições de treinamento máximas nas janelas de análise $kp+1$ e $(k+1)p+1$.

Também é importante ressaltar que, do modo como definido na equação (4.8), a variação da condição de treinamento é linear com relação ao índice da condição utilizada, *n*, e não com relação ao valor da SNR em si. Dessa maneira, foi definida uma forma alternativa do MCP-LV em que a variação da condição de treinamento utilizada varia linearmente com o valor da SNR em dB das condições de treinamento máximas, que será denominada de MCP-LV$_{dB}$, e cuja expressão matemática é a seguinte:

$$
\begin{aligned}
& p\left(\lambda_{s,n} \mid s, \vec{y}_{kp+1+q}\right) = 1, se\ n = \tilde{n}\left[kp+1+q\right] \\
& = \arg\operatorname{round}_{\text{SNR\_List}}\left\{\left(1-\frac{q}{p}\right)\text{SNR\_List}\left[\tilde{n}\left[kp+1\right]\right]+\left(\frac{q}{p}\right)\text{SNR\_List}\left[\tilde{n}\left[\left[k+1\right]p+1\right]\right]\right\} \\
& p\left(\lambda_{s,n} \mid s, \vec{y}_{kp+1+q}\right) = 0, se\ n \neq \tilde{n}\left[kp+1+q\right] \\
& \forall k = 0,1,2,.... \\
& \forall q = 0,..., p-1
\end{aligned}
\quad , \tag{4.10}
$$

sendo SNR_List(n) a função que retorna o valor da SNR utilizada no treinamento do n-ésimo modelo GMM do modelo multicondicional e $\text{round}_{\text{SNR\_List}}(r)$ a função que retorna o valor de SNR mais próximo de r dentre os valores utilizados para o treinamento dos modelos que compõem o modelo multicondicional.

Com relação ao custo computacional, o MCP-LV e o MCP-LV_{dB} apresentam o mesmo custo do MCP, como detalhado entre a equação (4.4) e a equação (4.7).

O esquema de funcionamento do MCP-LV (ou do MCP-LV_{dB}) é ilustrado na Figura 4.10:

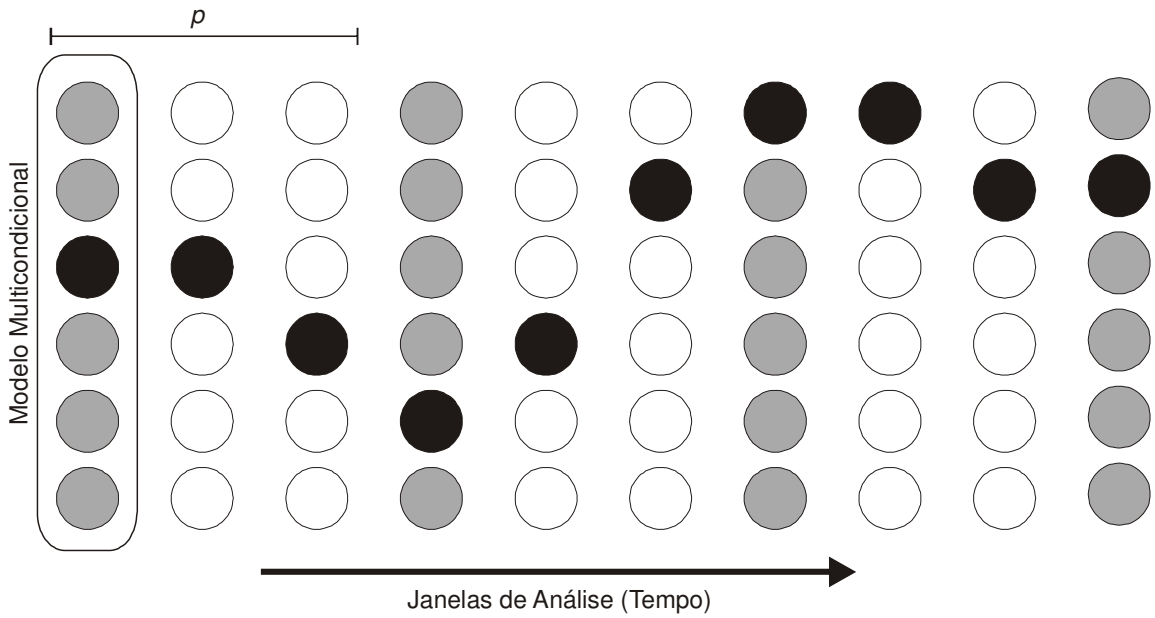

Figura 4.10: Ilustração do Método da Condição Persistente Linearmente Variado (MCP-LV).

Em complementação à ilustração da Figura 4.10, é inserida, no quadro da Figura 4.11, a programação básica do MCP-LV.

```
P(Y|λs) = 0 % verossimilhança do modelo multicondicional do
locutor s dado o conjunto de vetores de parâmetros Y
Para n de 1 a N % loop de condições de treinamento
     Calcule p(y[1]|λ[s,n]) % verossimilhança do modelo do
     locutor s para a condição n dado o vetor de parâmetros
     y[t]
Fim para n
n'[1] = argmax(p(y[1]|λ[s,n]) em relação a n) % determinar a
condição de treinamento que maximiza a verossimilhança
P(Y|λs) = P(Y|λs) + p(y[1]|λ[s,n'[1]])
Para t = (p + 1) a T, passo tamanho p % loop de janelas de
análise
   Para n de 1 a N % loop de condições de treinamento
     Calcule p(y[t]|λ[s,n])
   Fim para n
   n'[t] = argmax(p(y[t]|λ[s,n]) em relação a n)
   P(Y|λs) = P(Y|λs) + p(y[t]|λ[s,n'[t]])
   Para t2 = (- p + 1) a (- 1)
     n'[t+t2] = round(((p+t2)/p)*n'[t] + ((-t2)/p)*n'[t-p])
     Calcule p(y[t+t2]|λ[s,n'[t+t2]])
     P(Y|λs) = P(Y|λs) + p(y[t+t2]|λ[s,n'[t+t2]])
   Fim para t2
Fim para t
```

Figura 4.11: Programação básica do MCP-LV.

Para verificar o desempenho do MCP-LV e do MCP-LV_{dB} foram realizadas simulações com esses métodos utilizando o modelo usado na seção 3.3.1 (12 condições de treinamento: 60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB), para valores de *p* iguais a 2, 3, 5 e 10. Os resultados desses procedimentos são apresentados na Tabela 4.3 e na Tabela 4.4.

Tabela 4.3: Taxas de identificações corretas de sistemas de RAL multicondicional utilizando MCP-LV.

Taxas de Identificações Corretas (%)		SNR do Áudio de Teste (dB)										
		60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
Persistência (*p*) MCP-LV	**1**[*]	97,8	98,2	98,2	97,8	97,2	95,5	94,2	91,5	86,6	80,8	68,5
	2	97,8	98,3	98,2	96,3	96,3	94,5	93,0	91,5	88,2	84,0	77,3
	3	98,5	98,7	98,3	96,0	94,7	93,0	92,2	90,3	86,2	81,2	75,7
	5	98,7	98,7	98,2	95,3	95,3	91,3	90,3	88,7	83,8	80,8	74,5
	10	98,7	99,2	98,0	95,3	93,3	88,3	86,3	84,7	79,8	79,0	71,5

Tabela 4.4: Taxas de identificações corretas de sistemas de RAL multicondicional utilizando MCP-LV_{dB}.

Taxas de Identificações Corretas (%)		SNR do Áudio de Teste (dB)										
		60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
Persistência (*p*) MCP-LV_{dB}	**1**[†]	97,8	98,2	98,2	97,8	97,2	95,5	94,2	91,5	86,6	80,8	68,5
	2	98,7	99,0	98,7	97,3	97,0	95,5	93,5	91,2	88,2	85,5	76,2
	3	98,5	98,8	98,7	97,0	96,0	94,8	93,5	91,3	89,0	85,3	76,7
	5	98,7	98,8	99,0	96,0	96,0	93,2	91,8	89,3	86,8	83,3	77,2
	10	98,7	98,7	98,3	96,0	94,8	91,2	88,3	86,2	82,7	79,8	71,5

Para permitir uma melhor visualização, os resultados da Tabela 4.3 e da Tabela 4.4 foram traçados nos gráficos da Figura 4.12 e Figura 4.13.

[*] Novamente se destaca que o uso de $p = 1$ reduz o MCP ao MCM, não havendo qualquer vantagem computacional. Contudo, os valores estão aqui exibidos para fins de comparação de desempenho.

[†] Novamente se destaca que o uso de $p = 1$ reduz o MCP ao MCM, não havendo qualquer vantagem computacional. Contudo, os valores estão aqui exibidos para fins de comparação de desempenho.

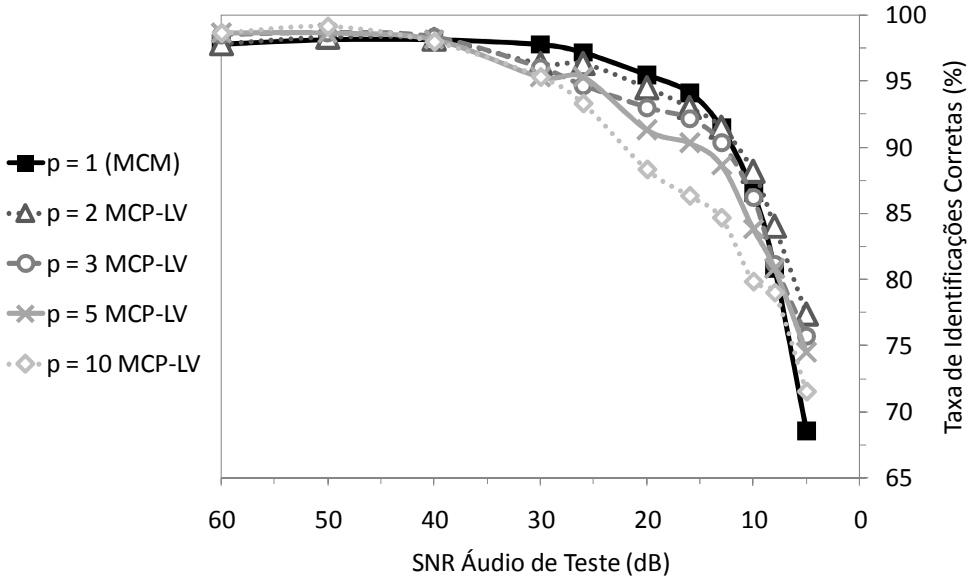

Figura 4.12: Desempenho de sistemas de RAL multicondicional utilizando MCP-LV, para diferentes valores de persistência, *p*.

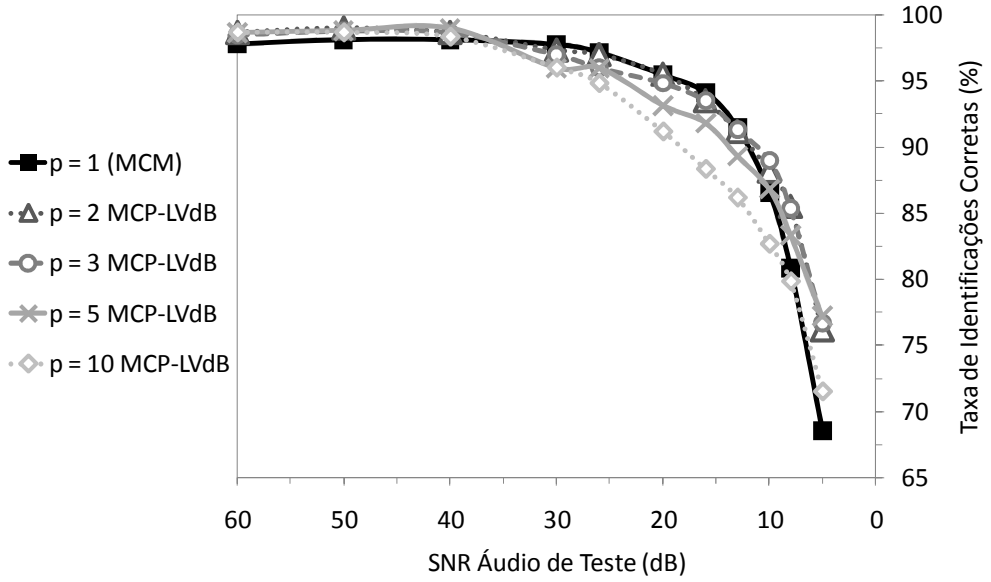

Figura 4.13: Desempenho de sistemas de RAL multicondicional utilizando MCP-$LV_{dB,}$ para diferentes valores de persistência, *p*.

Para uma comparação mais detalhada entre o MCP, o MCP-LV e o MCP-LV_{dB}, foram traçados gráficos comparando os desempenhos desses métodos quando utilizados com um mesmo valor de persistência, *p*.

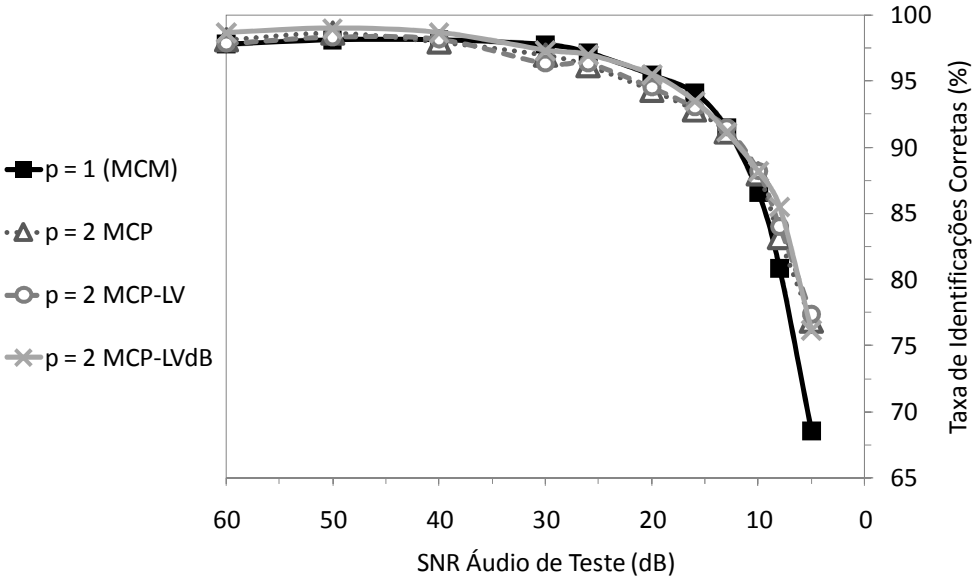

Figura 4.14: Comparação do desempenho de sistemas de RAL multicondicional utilizando MCM, MCP, MCP-LV e MCP-LV$_{dB}$, para $p = 2$.

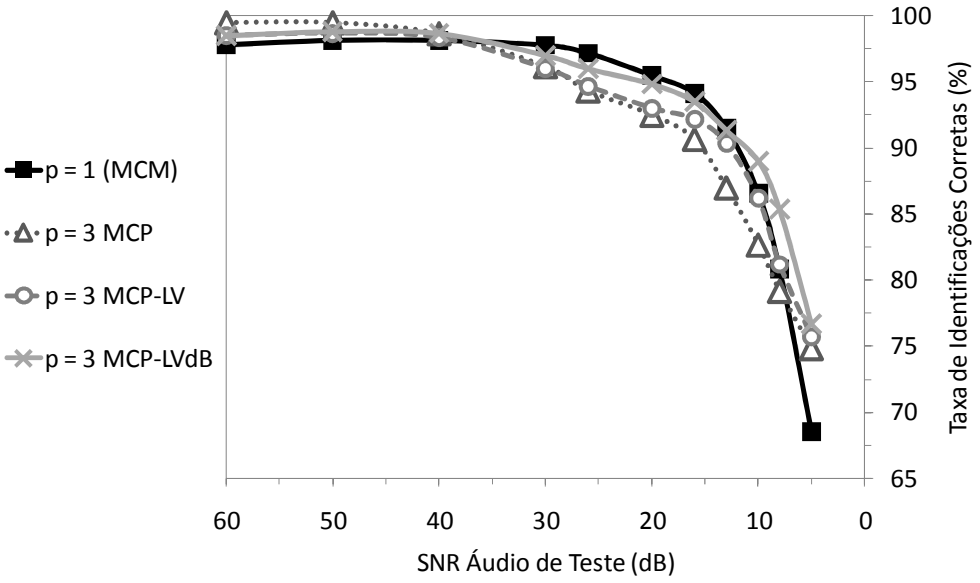

Figura 4.15: Comparação do desempenho de sistemas de RAL multicondicional utilizando MCM, MCP, MCP-LV e MCP-LV$_{dB}$, para $p = 3$.

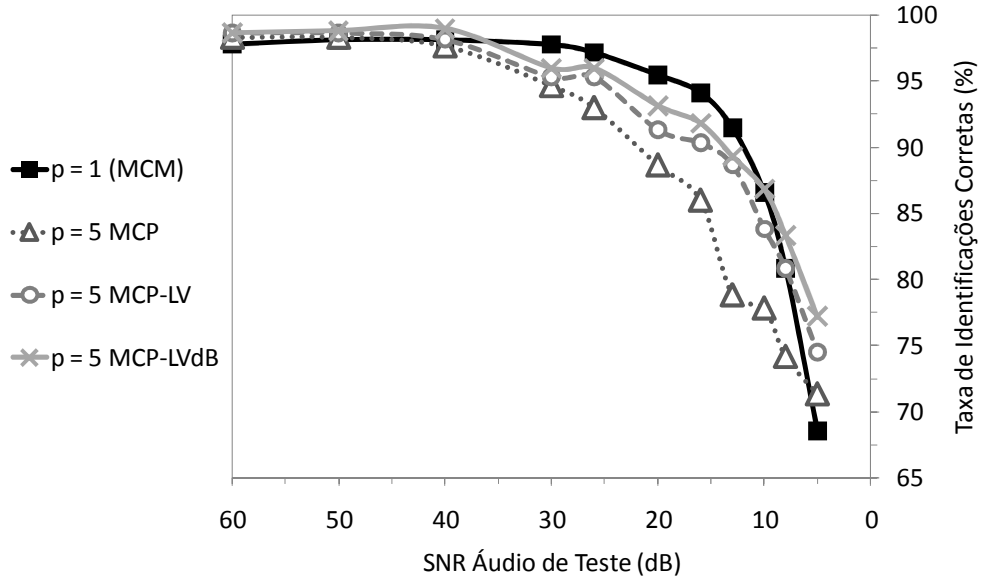

Figura 4.16: Comparação do desempenho de sistemas de RAL multicondicional utilizando MCM, MCP, MCP-LV e MCP-LV$_{dB}$, para $p = 5$.

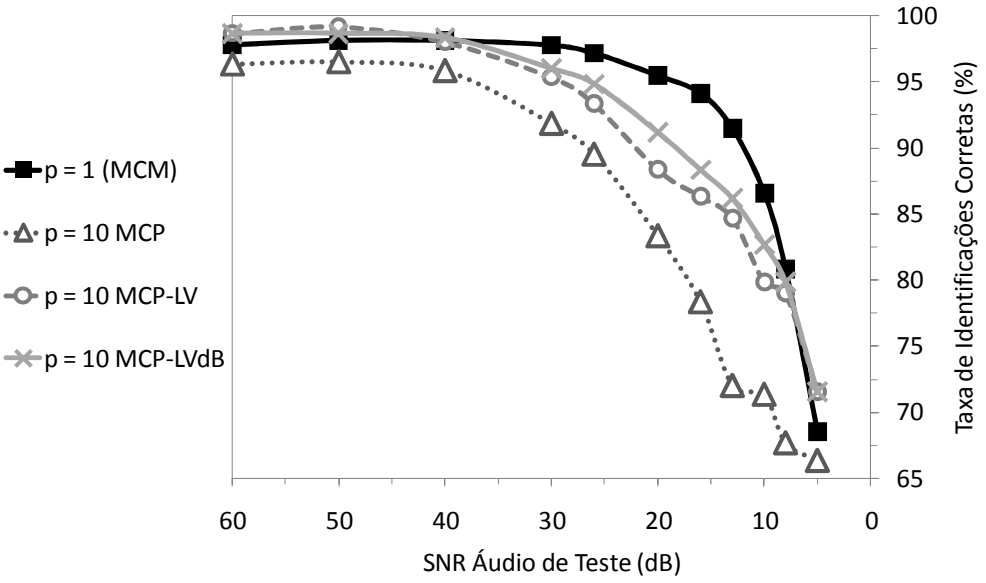

Figura 4.17: Comparação do desempenho de sistemas de RAL multicondicional utilizando, MCP, MCP-LV e MCP-LV$_{dB}$, para $p = 10$.

Como era de se esperar, as diferenças entre o MCP e o MCP-LV ou MCP-LV$_{dB}$ são mais acentuadas para valores mais elevados de p, pois, nesses casos, a diferença entre a predição de ordem zero do MCP e a predição linear do MCP-LV e MCP-LV$_{dB}$ torna-se mais acentuada. Em geral, observa-se que o MCP-LV apresenta desempenho superior ao MCP, confirmando que a estimação linear de evolução da condição do áudio proporciona melhoria

no desempenho do modelo multicondicional. Também, verifica-se que o MCP-LV_{dB} apresenta resultados superiores ao MCP-LV, demonstrando, como esperado, que a interpolação baseada no valor da SNR leva a resultados melhores que a interpolação baseada nos índices das condições de treino. Isso se justifica, particularmente, porque não foi adotada espaçamento uniforme entre os valores de SNR das condições de treino do modelo multicondicional.

Foram também calculadas as médias das taxas de identificação corretas, a fim de observar o resultado global do uso da persistência linearmente variada no desempenho do sistema de RAL. As taxas de identificações corretas médias para cada valor da persistência, *p*, são apresentadas na Tabela 4.5 a seguir[*], juntamente com as taxas para o sistema MCP, copiadas da Tabela 4.2, para comparações.

Tabela 4.5: Taxas médias de identificações corretas de sistemas de RAL multicondicional utilizando MCP, MCP-LV e MCP-LV_{dB}.

Persistência (*p*)	**Taxa Média de Identificações Corretas (%)**		
	MCP	**MCP-LV**	**MCP-LV_{dB}**
1	91,5	91,5	91,5
2	92,2	92,3	92,8
3	90,5	91,3	92,7
5	87,2	90,5	91,8
10	82,6	88,6	89,7

Para uma melhor visualização, os resultados da Tabela 4.5 foram traçados na Figura 4.26.

[*] Ver nota anterior sobre os resultados da Tabela 4.2.

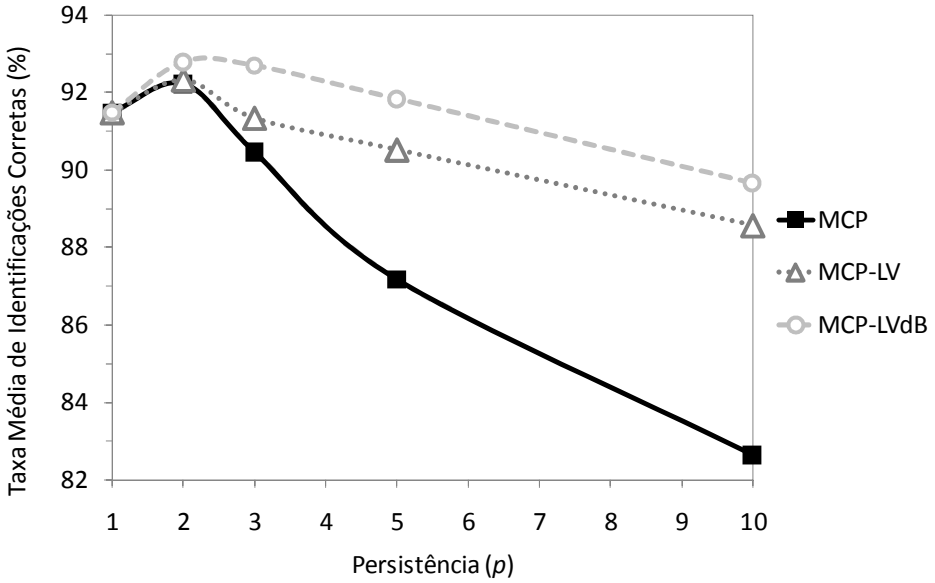

Figura 4.18: Taxa média de identificações corretas de sistemas de RAL multicondicional MCP, MCP-LV e MCP-LV_{dB}, em função do valor de persistência, p.

Como se observa no gráfico da Figura 4.26, o uso do MCP-LV ou do MCP-LV_{dB} reduz significativamente as perdas de desempenho associadas à restrição de variação temporal da condição de treinamento relativa ao MCP, permitindo a utilização de valores mais altos de persistência, p, e, consequentemente, permitindo uma redução maior do custo computacional. A utilização do MCP-LV_{dB} possibilitou reduções de custo computacional de 73,3% (p = 5) sem afetar o desempenho médio do sistema.

Um resultado relevante com relação aos sistemas de RAL com MCP é a melhoria de desempenho observada nas zonas extremas de SNR, em comparação com o mesmo sistema sem condição persistente, que chega a exceder 8%. Essa melhora nos resultados, pelas mesmas razões discutidas na seção 3.3.1, não pode ser atribuída a aumentos no valor da verossimilhança, $p(Y \mid s)$, do locutor correto, visto que a utilização da condição persistente não pode provocar tal aumento[*]. Dessa forma, a melhora dos resultados nas faixas de SNR extremas alta e baixa, observada nos sistemas com condição persistente, $p > 1$, somente pode decorrer de uma queda mais acentuada nas verossimilhanças, $p(Y \mid s)$, dos locutores incorretos que nas do locutor correto, quando é usado o MCP ou suas variantes.

[*] O uso da condição persistente torna o sistema subótimo com relação à maximização de $p(Y \mid s)$, pois não se calcula a condição que maximiza a probabilidade $p(\vec{y}_t \mid \lambda_{s,n})$ para cada vetor de parâmetros, mas se reutiliza a condição determinada para o vetor $\vec{y}_t$ nos vetores $\vec{y}_{t+1}$, $\vec{y}_{t+2}$,..., $\vec{y}_{t+p-1}$.

4.3 MODELOS MULTICONDICIONAIS ADAPTATIVOS

Outra técnica proposta para diminuir o custo computacional de modelos multicondicionais, enfocando a redução do *N* da equação (4.2) é a utilização dos Modelos Multicondicionais Adaptativos (MMA). Essa técnica, assim como o MCP, tira proveito da característica da variação gradual da SNR do áudio questionado para evitar o cálculo de todos os modelos do modelo multicondicional a cada janela de análise. Entretanto, ao contrário do que ser realiza no MCP, o princípio de funcionamento do MMA não reside na estimação direta da condição de treinamento que produz a máxima verossimilhança. A proposta do MMA apenas admite que as variações dessa condição ótima entre janelas subseqüentes não podem ser demasiadamente abruptas. Dessa maneira, dada a forte correlação entre as condições do áudio de janelas de análise subseqüentes admitida, o MMA propõe que não sejam calculadas as verossimilhanças para todos os modelos componentes do modelo multicondicional, mas somente para aqueles mais próximos (em termos de condições de treinamento) do modelo que resultou na máxima verossimilhança na janela anterior. Com isso, o valor efetivo de condições de treinamento do modelo, *N*, é reduzido, diminuindo o esforço computacional necessário para a identificação.

O método MMA é uma variação do MCM, descrito na equação (3.13). A fundamental alteração introduzida é a restrição da busca da condição máxima a uma vizinhança da condição determinada na janela de análise anterior. Dessa maneira, é necessário alterar a definição das probabilidades $P\left(\Phi_n \mid s, \vec{y}_t\right)$ para a forma:

$$\begin{aligned} &P\left(\Phi_n \mid s, \vec{y}_t\right) = 1, se\ n = \tilde{n}[t] = \underset{n \in V\left(\tilde{n}(t-1)\right)}{\arg\max}\ p\left(\vec{y}_t \mid \lambda_{s,n}\right) \\ &P\left(\Phi_n \mid s, \vec{y}_t\right) = 0, se\ n \neq \tilde{n}[t] \end{aligned} , \tag{4.11}$$

sendo $V\left(\tilde{n}\left(t-1\right)\right)$ uma vizinhança da condição máxima calculada para a janela de análise anterior definida por:

$$V\left(\tilde{n}[t-1]\right) = \left\{\tilde{n}[t-1] - a, ..., \tilde{n}[t-1], ..., \tilde{n}[t-1] + a\right\} , \tag{4.12}$$

onde o parâmetro *a* é denominado de adaptabilidade do modelo MMA.

Em termos práticos, o que o MMA faz é restringir a variação da condição de treinamento utilizada em janelas de análises sucessivas, diminuindo, consequentemente, o número de componentes gaussianas a serem calculadas para a determinação da verossimilhança do

modelo. Em princípio, essa restrição não deve comprometer o desempenho do sistema de identificação, visto que, dadas as características normalmente verificadas nos trechos áudio analisados, é de se esperar que em intervalos de 10 ms a 20 ms não haja variações muito grandes nas suas características. Desse modo, a restrição na variação das condições de treinamento não deve afetar significativamente o resultado final do processo.

Para facilitar a visualização, o esquema de funcionamento do MMA (para $a = 1$) é exibido na Figura 4.19.

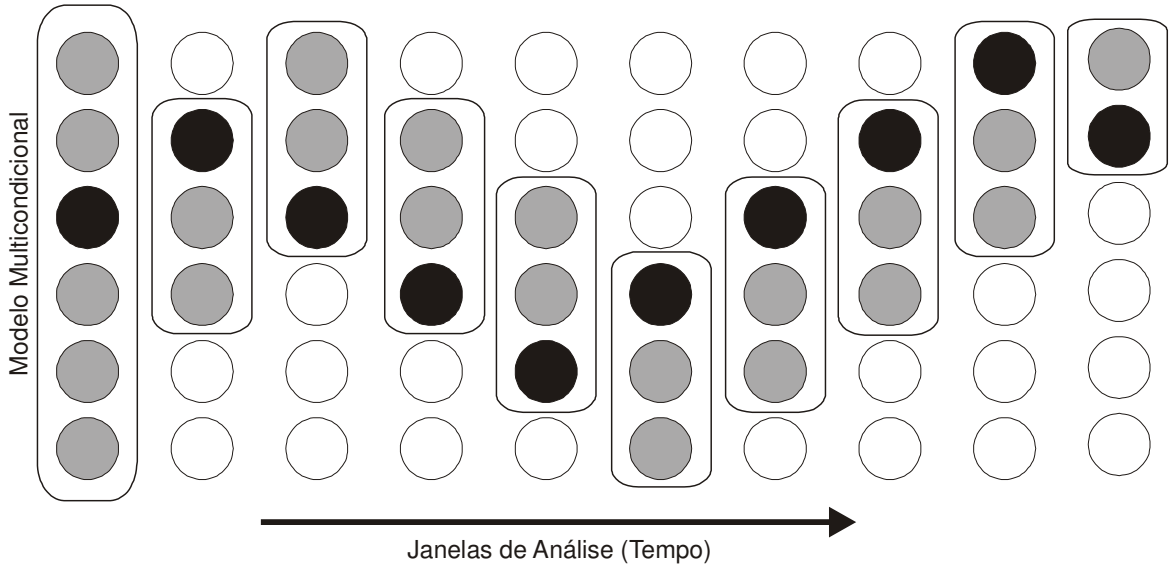

Figura 4.19: Ilustração dos Modelos Multicondicionais Adaptativos (MMA).

Além da ilustração da Figura 4.19, é inserida, no quadro da Figura 4.19, a programação básica do MMA.

```
P(Y|λs) = 0 % verossimilhança do modelo multicondicional do
locutor s dado o conjunto de vetores de parâmetros Y

Para n de 1 a N (condições de treinamento) % loop de condi-
ções de treinamento

     Calcule p(y[1]|λ[s,n]) % verossimilhança do modelo do
     locutor s para a condição n dado o vetor de parâmetros
     y[t]

Fim para n

n'[1] = argmax(p(y[1]|λ[s,n]) em relação a n) % determinar a
condição de treinamento que maximiza a verossimilhança

P(Y|λs) = P(Y|λs) + p(y[1]|λ[s,n'[1]])

Para t = 2 a T % loop de janelas de análise

   n_min = max(n'[t-1]-a,1) % limite inferior da vizinhança

   n_max = min(n'[t-1]+a,N) % limite superior da vizinhança

   Para n de n_min a n_max % loop de condições de treinamento
limitado à vizinhança

     Calcule p(y[t]|λ[s,n])

   Fim para n

   n'[t] = argmax(p(y[t]|λ[s,n]) em relação a n) % determinar
a condição de treinamento dentro da vizinhança que maximiza a
verossimilhança

   P(Y|λs) = P(Y|λs) + p(y[t]|λ[s,n'[t]])

Fim para t
```

Figura 4.20: Programação básica do MMA.

O custo computacional de uma identificação com o MMA é dado por:

$$W_{MMA}(a) \propto M(2a+1) , \tag{4.13}$$

pois, a cada janela de análise, são calculadas apenas as verossimilhanças para os GMM correspondentes a $2a+1$ condições de treinamento, e cada um desses GMM é formado por *M* gaussianas. Destaque-se que, dessa forma, o custo computacional do MMA independe do número de condições de treinamento do modelo multicondicional, *N*. Essa propriedade é bastante interessante, pois, como visto na seção 3.3.2.1, em geral, o aumento no número de condições de treinamento melhora o desempenho do sistema de identificação. Destarte, com o uso do MMA, pode-se aumentar livremente[*] o número de condições de treinamento do modelo multicondicional, melhorando seu desempenho, sem, contudo, afetar seu custo computacional.

[*] Há um limite prático para esse aumento. Deve-se perceber que, uma vez que as condições de treinamento sejam muito semelhantes entre si, será necessário aumentar a vizinhança das condições testadas, ou pode ocorrer de essa vizinhança não permitir a evolução da condição de treino utilizada tão rapidamente quanto necessário.

A redução do custo computacional, em relação ao modelo multicondicional tradicional, MCM, é dada por:

$$W_{MCM} - W_{MMA}(a) \propto M \cdot N - M(2a+1) = M(N-2a-1) \tag{4.14}$$

e diferença relativa entre os custos computacionais é calculada por:

$$\frac{\Delta W}{W} = \frac{W_{MCM} - W_{MMA}(a)}{W_{MCM}} \propto 1 - \frac{2a+1}{N} \tag{4.15}$$

Observa-se que a quantidade de componentes em cada modelo multicondicional, *M*, não influencia na diferença relativa de custo computacional; que é afetada apenas pelo número de condições de treinamento, *N*, e pela adaptabilidade, *a*. Os gráficos da Figura 4.21 e da Figura 4.22 ilustram a queda do custo computacional absoluto e relativo com o uso dos Modelos Multicondicionais Adaptativos para os casos de modelos multicondicionais de 10 e de 20 condições de treinamento (*N* = 10 e *N* = 20), de acordo com as equações (4.13) e (4.15). No caso do gráfico da Figura 4.21, a escala vertical utilizada é em número de vezes *M*, a quantidade de componentes gaussianas de cada modelo unidimensional. No caso do gráfico da Figura 4.22, a escala vertical utilizada é em relação ao custo computacional original de um modelo multicondicional sem o uso do MMA.

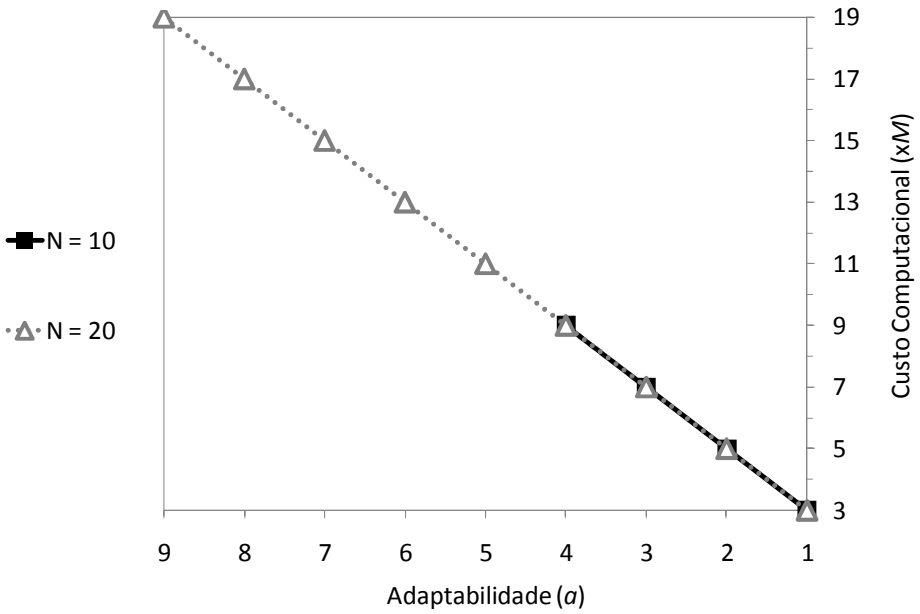

Figura 4.21: Custo computacional absoluto de sistemas com MMA, em função da adaptabilidade, *a*, para diferentes números de condições de treinamento, *N*.

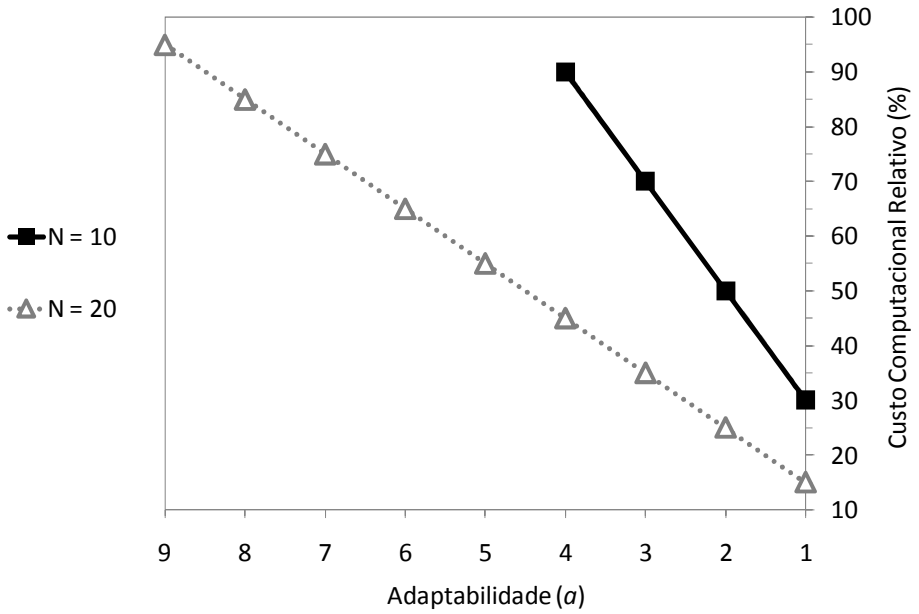

Figura 4.22: Custo computacional relativo de sistemas com MMA, em função da adaptabilidade, *a*, para diferentes números de condições de treinamento, *N*.

Observando o gráfico da Figura 4.21, verifica-se a independência do custo absoluto do MMA com relação ao número de condições de treinamento, *N*, anteriormente destacada na análise da equação (4.13). Também se observa, tanto no gráfico da Figura 4.21 como no da Figura 4.22, a linearidade do custo computacional com valor da adaptabilidade, *a*.

No limite da adaptabilidade, $a = 1$, tem-se que:

$$W_{MMA}(a=1) \propto 3M \tag{4.16}$$

Ou seja, independentemente do número de condições de treinamento do modelo multicondicional, é possível limitar o custo computacional ao equivalente a um modelo com apenas três condições. Nesse caso, a diferença relativa entre o custo computacional do modelo tradicional e o custo do MMA é:

$$\frac{\Delta W}{W} = \frac{W_{MCM} - W_{MMA}(a=1)}{W_{MCM}} \propto \frac{N-3}{N} = 1 - \frac{3}{N} \tag{4.17}$$

No gráfico da Figura 4.23, foi traçado o custo computacional relativo do MMA para o caso limite de $a = 1$ em função do número de condições de treinamento do modelo multicondicional, *N*.

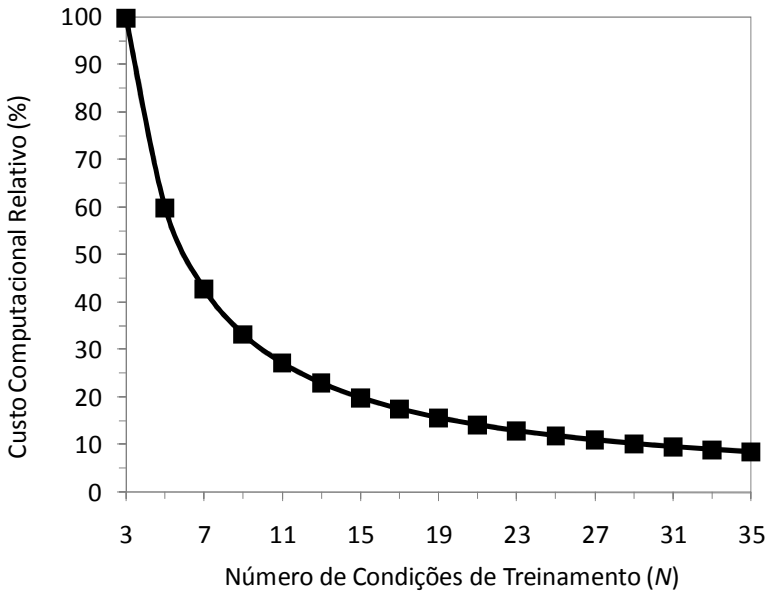

Figura 4.23: Custo computacional relativo de sistemas com MMA, em função do número de condições de treinamento, *N*, para *a = 1*.

Deve-se destacar que a equação (4.13) contém uma imprecisão, ao considerar que o modelo multicondicional adaptativo será composto sempre por $2a+1$ condições. Na realidade, em alguns casos, o MMA será composto por menos modelos porque a vizinhança definida em (4.12) pode ser truncada no caso de extrapolar os limites do modelo multicondicional[*]. Por exemplo, se $\tilde{n}[t-1]=2$ e se $a=2$, a vizinhança $V\left(\tilde{n}[t-1]\right)$ será:

$$V\left(\tilde{n}[t-1]\right)=\{1,2,3,4\} \tag{4.18}$$

e serão calculadas verossimilhanças para apenas quatro condições de treinamento, ao invés das cinco ($2a+1$) previstas por (4.13).

A correção da vizinhança, *V*, para incorporar essa limitação do modelo multicondicional pode ser estabelecida como:

$$V\left(\tilde{n}[t-1]\right)=\left\{\max\left(\tilde{n}[t-1]-d,1\right),...,\tilde{n}[t-1],\ ...,\min\left(\tilde{n}[t-1]+d,N\right)\right\} \tag{4.19}$$

Admitindo que todas as condições de treinamento ocorram com a mesma frequência, pode-se traçar os gráficos das reduções dos custos computacionais absolutos e relativos corrigidos, com relação ao tamanho da adaptabilidade, *a*, exibidos na Figura 4.24 e na Figura 4.25.

[*] Veja graficamente esse fenômeno na última janela de análise representada na ilustração da Figura 4.19, na qual são calculadas apenas os GMM de duas condições, quando o normal seria o cálculo de três condições (*a* = 1).

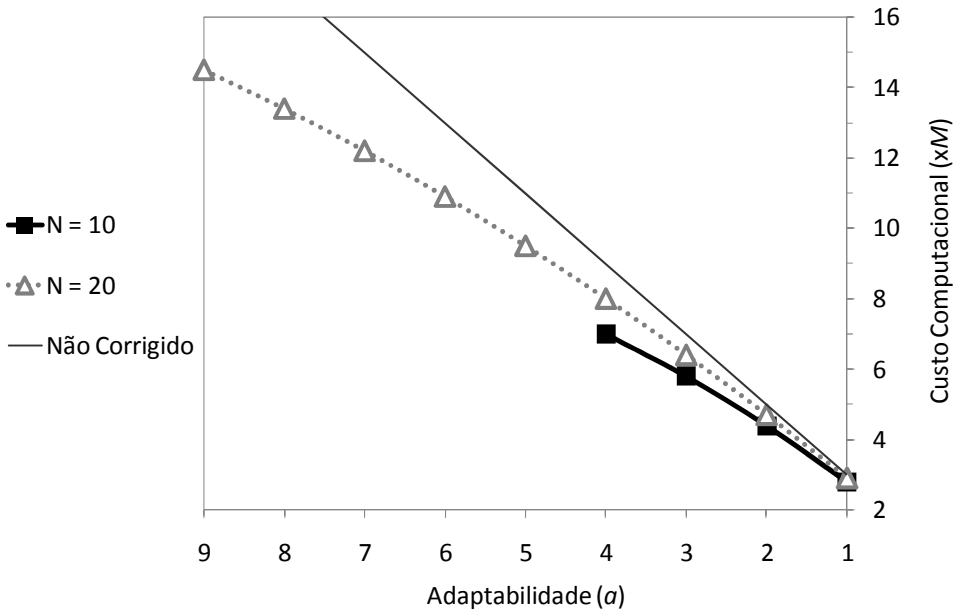

Figura 4.24: Custo computacional absoluto corrigido de sistemas com MMA, em função da adaptabilidade, *a*, para diferentes números de condições de treinamento, *N*.

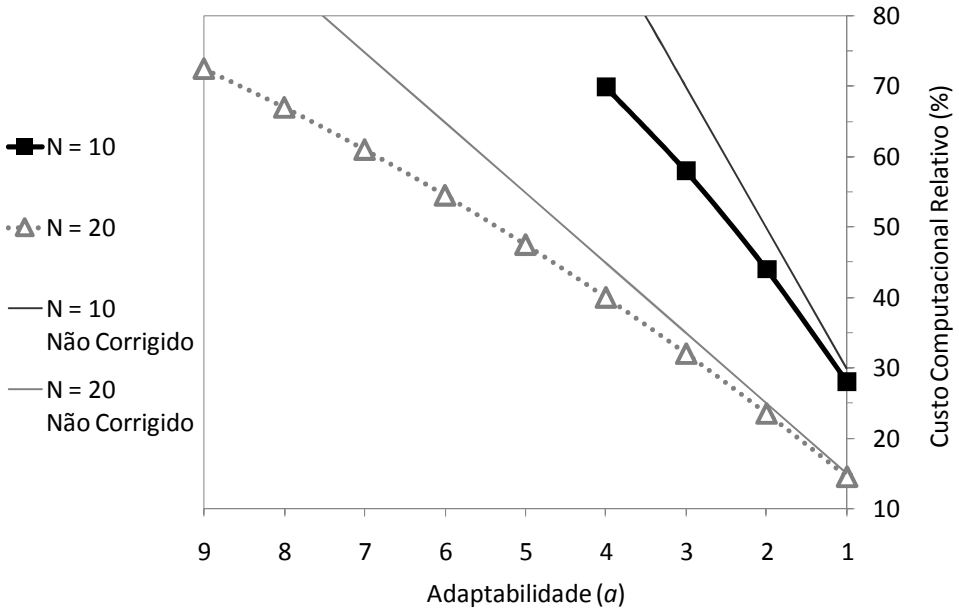

Figura 4.25: Custo computacional relativo corrigido de sistemas com MMA, em função da adaptabilidade, *a*, para diferentes números de condições de treinamento, *N*.

Para verificar o desempenho do MMA proposto, foram realizados procedimentos de identificação utilizando diferentes valores de adaptabilidade, *a*, e o modelo multicondicional de 12 condições de treinamento (60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB). As taxas de identificações corretas obtidas para $a = 3$, $a = 2$ e $a = 1$ são exibidas na Tabela 4.6, juntamente com as taxas para o caso do modelo multicondicional tradicional (MCM), para comparações.

Tabela 4.6: Taxas de identificações corretas de sistemas de RAL multicondicional utilizando MMA, para diferentes valores de *a*.

Taxas de Identificações Corretas (%)		SNR do Áudio de Teste (dB)										
		60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
Adaptabilidade (*a*)	**3**	97,8	98,3	98,5	98,5	98,0	95,8	94,5	91,5	88,0	83,8	73,8
	2	98,2	98,7	99,3	98,8	97,7	95,3	94,3	91,8	86,5	85,7	76,8
	1	99,7	99,7	99,3	98,7	98,2	96,8	94,3	92,8	89,5	87,3	80,7
MCM		97,8	98,2	98,2	97,2	96,8	95,5	94,2	91,5	86,6	80,8	68,5

Para permitir uma melhor visualização, os dados da Tabela 4.6 foram traçados no gráfico da Figura 4.26.

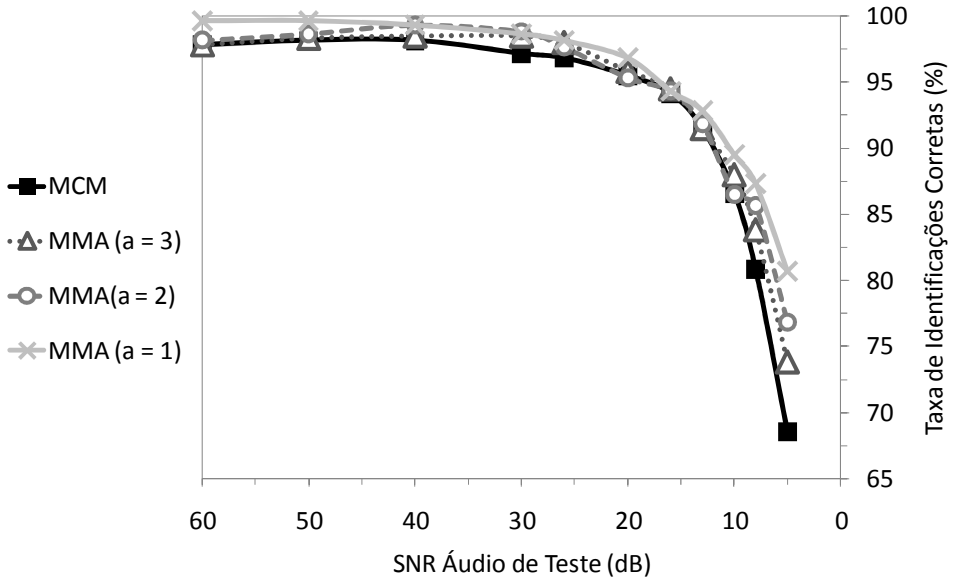

Figura 4.26: Desempenho de sistemas de RAL multicondicional utilizando MMA, para diferentes valores de adaptabilidade, *a*.

Como se verifica no gráfico da Figura 4.26, o uso do MMA, além da diminuição no custo computacional, promove uma melhoria sistemática no desempenho do sistema, especialmente nos casos de áudio com altos níveis de ruído.

A fim de analisar mais detalhadamente a melhoria do desempenho do sistema de RAL com o uso do MMA, foram traçados, no gráfico da Figura 4.27, as diferenças entre as taxas de

identificações corretas usando o MMA para diversos tamanhos de vizinhança e as taxas de identificações corretas usando o MCM.

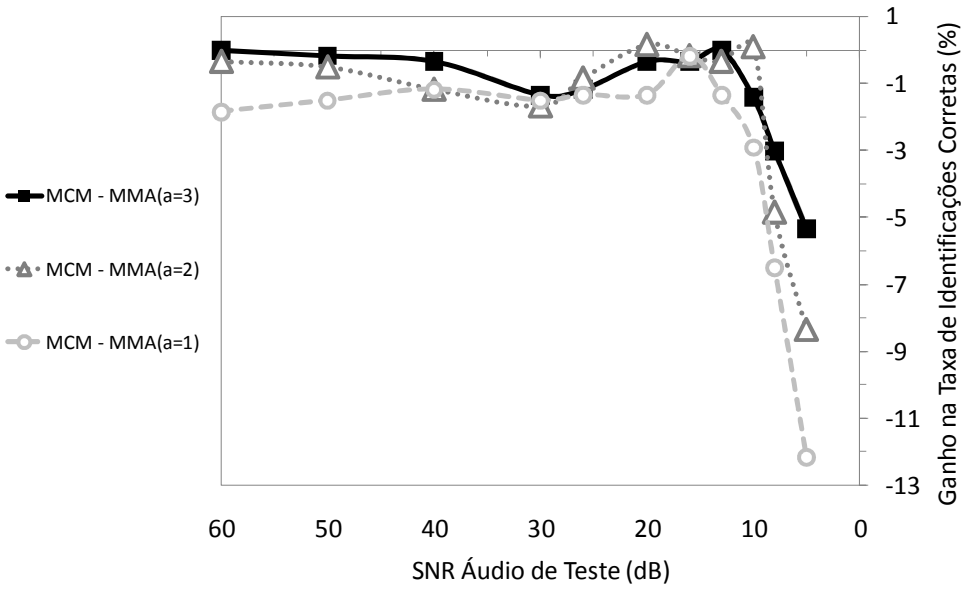

Figura 4.27: Diferença entre o desempenho de sistemas de RAL multicondicional MMA e MCM, para diferentes valores de adaptabilidade, *a*.

Observa-se, no gráfico da Figura 4.27, que o uso do MMA promove melhoria no desempenho do sistema em praticamente todas as condições de SNR e de tamanho de vizinhança testados. De modo geral, as melhorias se acentuam à medida que se diminui o tamanho da vizinhança testada (e que se diminui o custo computacional), e são mais elevadas para os valores extremos de SNR.

Para facilitar a comparação de desempenho do sistema para os diversos valores de adaptabilidade, *a*, foram calculados os valores médios[*] das taxas de identificação em cada condição, como exibido na Tabela 4.7.

Tabela 4.7: Taxa média de identificações corretas de sistemas de RAL multicondicional utilizando MMA.

Adaptabilidade (*a*)	**Taxa Média de Identificações Corretas (%)**
5,5[†]	91,4

[*] Deve-se destacar que, como os valores de SNR escolhidos para a realização das simulações foi arbitrário, o valor da média não contém realmente um significado no sentido absoluto, devendo ser sempre observados os resultados para cada valor de SNR.

[†] O valor de vizinhança 5,5 foi incluído na tabela para fins comparativos, representando o sistema sem utilização do MMA (usando MCM). O valor $a = 5{,}5$ utilizado decorre de o número condições de treinamento do modelo multicondicional, 12, e da equação (4.13).

Adaptabilidade (*a*)	Taxa Média de Identificações Corretas (%)
3	92,6
2	93,0
1	94,3

Os dados constantes da Tabela 4.7 foram traçados no gráfico da Figura 4.28.

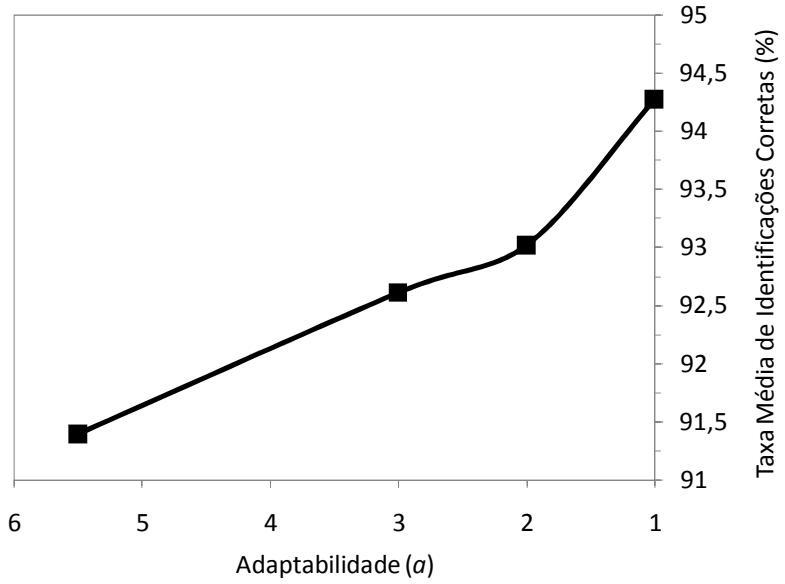

Figura 4.28: Taxa média de identificações corretas de sistemas de RAL multicondicional MMA, em função do valor adaptabilidade, *a*.

Verifica-se que a taxa média de identificações corretas cresce à medida que se diminui o tamanho da vizinhança, *a*, de modo que, com o uso do MMA, consegue-se simultaneamente diminuir significativamente o custo computacional da identificação e aumentar o desempenho do sistema de RAL.

Nas simulações realizadas, foi possível reduzir em 75% o custo computacional da identificação, conseguindo ainda ganho no desempenho do sistema em todas as condições de ruído testadas (ganho médio de 2,9%). A redução no custo computacional no MMA, como demonstrado, é fortemente influenciada pelo número de condições do modelo original, podendo variar entre 70% e 85% para modelos de 10 a 20 componentes.

As razões para se verificar um ganho de desempenho na identificação com o uso do MMA são semelhantes às discutidas no caso do Método da Condição Persistente (MCP), e, tanto

num caso quanto no outro, ocorrem de modo mais acentuado para os valores extremos de SNR, especialmente nos valores de SNR mais baixos. Basicamente, observa-se que, nessas situações, a existência de muitos modelos para um mesmo locutor beneficia mais intensamente os locutores incorretos que os locutores corretos. Dessa forma, a restrição no número de modelos realizada pelo MMA e pelo MCP proporciona uma perda maior aos locutores incorretos, levando a uma melhora na taxa de identificação correta.

4.3.1 Ganho de Desempenho com MMA

Um fenômeno interessante verificado nos resultados das análises do MMA foi o ganho de desempenho ocorrido para todos os valores de adaptabilidade, a, e para todas as condições de ruído examinadas, como visto na Figura 4.26 e na Figura 4.27. Esse fenômeno despertou interesse e provocou uma análise mais aprofundada de suas razões.

O MMA, como explicado, opera limitando a variação entre as condições de treinamento dos modelos utilizados em janelas sucessivas de análise do áudio questionado. Essa limitação é feita para diminuir o número de GMM que compõem o modelo multicondicional, reduzindo o esforço computacional necessário para a determinação da verossimilhança dos modelos e do processo de identificação como um todo.

Pelo mesmo raciocínio desenvolvido na seção 3.3.2, é impossível que o MMA provoque um aumento da verossimilhança dos modelos com relação aos valores obtidos com o MCM (método da condição máxima).

$$p(Y \mid s)_{MCM} \geq p(Y \mid s)_{MMA} \tag{4.20}$$

Portanto, para justificar a melhora das taxas de identificações corretas, é preciso que a restrição imposta pelo MMA afete os locutores incorretos mais intensamente que os locutores corretos, de forma que as quedas nos valores de verossimilhança para os locutores incorretos serão maiores que aquelas para os locutores corretos.

$$\begin{aligned} &p(Y \mid \tilde{s})_{MCM} - p(Y \mid \tilde{s})_{MMA} \leq p(Y \mid s')_{MCM} - p(Y \mid s')_{MMA} \\ &\Delta p(Y \mid \tilde{s}) \leq \Delta p(Y \mid s') \end{aligned} \tag{4.21}$$

Possíveis razões teóricas para esse maior impacto sobre os locutores incorretos seguem a mesma linha esboçada na seção 3.3.2, no sentido de que, para o locutor correto, como o modelo está mais bem ajustado aos dados, a possibilidade de escolha entre os N modelos

multicondicionais não promove ganhos tão expressivos no valor da verossimilhança quanto os observados para os locutores incorretos.

A fim de validar essa hipótese e justificar o ganho de desempenho observado com o MMA, foi analisado o valor do desvio padrão de *n*, o índice da condição de treinamento que proporcionou a máxima verossimilhança do locutor a cada janela de análise do áudio questionado. Foi comparado o valor médio desse desvio padrão para os locutores corretos com o valor médio para os locutores incorretos, para cada condição de ruído do áudio. Os resultados obtidos estão expostos no gráfico da Figura 4.29, que traça os valores médios dos desvios padrão de *n* para os locutores corretos (C) e incorretos (I) em função do nível de ruído do áudio questionado.

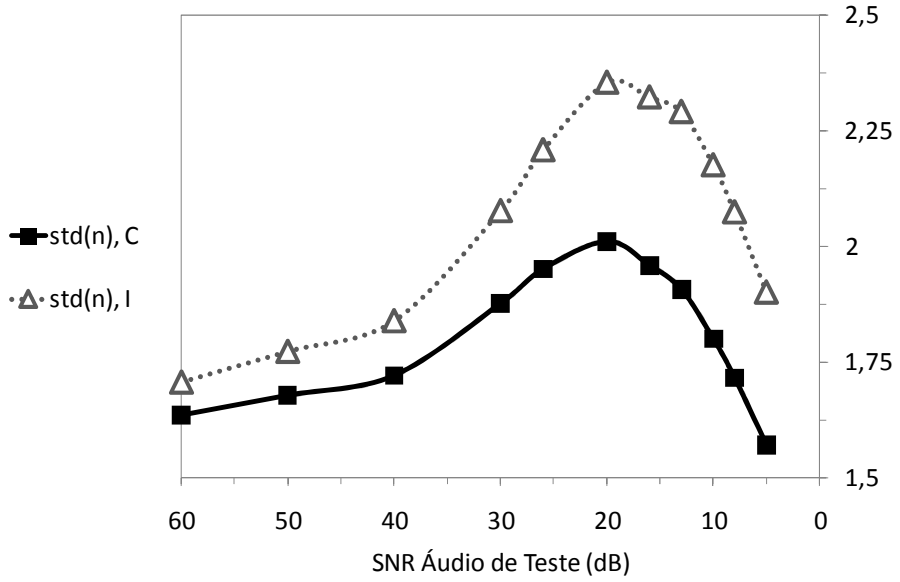

Figura 4.29: Desvio padrão médio do valor do índice da condição máxima, $\mathrm{std}(n)$, em função do nível de ruído no áudio de teste, para o locutor correto (C) e para os locutores incorretos (I).

Como se observa, em todas as situações de ruído, o desvio padrão médio de *n* para os locutores incorretos é significativamente superior ao valor obtido para os locutores corretos, o que ajuda a explicar os ganhos verificados com o MMA.

O MMA limita a variação entre os índices das condições de treinamento utilizadas em janelas consecutivas. Dessa maneira, os modelos dos locutores incorretos, que têm uma maior variação nesse índice, sofrem mais acentuadamente os efeitos da restrição imposta, ocasionando uma perda maior em seus valores de verossimilhança, quando comparadas com as perdas dos locutores corretos.

4.3.2 Comparação MMA x MCP

É importante destacar que tanto os Modelos Multicondicionais Adaptativos (MMA) como o Método da Condição Persistente (MCP) buscam reduzir o custo computacional da identificação do locutor pela restrição no número de condições de treinamento efetivamente utilizadas no modelo multicondicional, embora, para isso, utilizem abordagens distintas. Nesse sentido, vale à pena realizar uma comparação entre o desempenho de sistemas com esses dois métodos, a fim de observar qual deles proporciona maior redução de custo computacional conjugada a um bom desempenho de identificação.

O custo computacional do MMA, como observado na equação (4.13), e repetido a seguir por conveniência, é independente do número de condições de treinamento, *N*.

$$W_{MMA}(a) \propto M(2a+1) \ , \tag{4.22}$$

O custo do MCP, por outro lado, é fortemente influenciado por *N*, conforme a expressão (4.4), também repetida a seguir:

$$W_{MCP}(p) \propto M\left(\frac{N+(p-1)}{p}\right) = M\left(1+\frac{N-1}{p}\right) \tag{4.23}$$

Dessa maneira, não é possível estabelecer uma situação de equivalência entre os custos computacionais do MMA e do MCP se não for fixado o valor de *N*. Por questões práticas, será utilizado o valor $N=12$, visto que essa foi a situação efetivamente empregada em ambos os métodos. Nesse caso, para o MMA com $a=1$, tem-se um custo computacional $W_{MMA} \propto 3M$ (independente de *N*). Para o MCP, utilizando $p=5$, tem-se um custo computacional $W_{MCP} \propto 3,2M$. Com esses parâmetros, os custos computacionais dos dois métodos são bastante próximos, o que permite uma comparação direta de seus desempenhos.

Deve-se destacar que os valores de $p=5$, para o MCP, e de $a=1$, para o MMA, foram escolhidos por terem apresentado uma boa relação entre redução de custo computacional e desempenho na identificação. Destaque-se ainda que a comparação será realizada utilizando a variante MCP-LV_{dB} do MCP, visto que essa proporcionou os melhores resultados.

O gráfico comparativo do desempenho de identificação desses dois métodos está ilustrado na Figura 4.30.

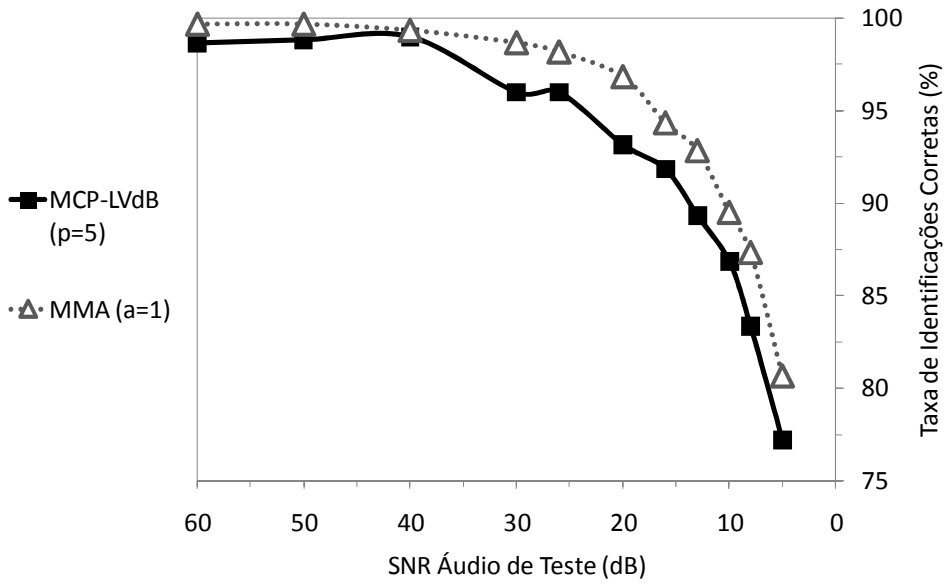

Figura 4.30: Comparação do desempenho de sistemas de RAL multicondicional utilizando MCP-LV_{dB}, $p = 5$, e MMA, $a = 1$.

Como se verifica, o MMA apresenta um desempenho sistematicamente superior ao MCP-LV_{dB}. Além disso, a redução de custo computacional obtida com o MMA (75%) é levemente superior àquela proporcionada pelo MCP-LV_{dB}, demonstrando a superioridade da técnica dos Modelos Multicondicionais Adaptativos.

4.4 MÉTODO DAS GAUSSIANAS DOMINANTES

O Método das Gaussianas Dominantes (MGD) procura reduzir o custo computacional dos modelos multicondicionais pela redução do número efetivo de gaussianas para cada condição de treinamento. Esse método objetiva minimizar a quantidade de componentes efetivamente calculadas em cada modelo GMM que compõe um modelo multicondicional, reduzindo, dessa maneira, o valor de *M*, na expressão (4.2).

A idéia fundamental do MGD é explorar o fato de que, a cada janela de análise, apenas uma ou algumas poucas componentes gaussianas do modelo são responsáveis pela maior parte do resultado final do valor da verossimilhança calculada, de modo que a maior parte das *M* gaussianas não contribui de modo significativo e pode ser descartada. Essa propriedade decorre do fato de a voz humana ser formada por sucessões de classes sonoras (fonemas) distintos. É por essa razão, de fato, que os GMM são uma forma eficaz de modelagem da voz humana, como discutido na seção 2.2. Então, como, nos GMM, cada compo-

nente gaussiana modela classes sonoras distintas (regiões distintas no espaço de parâmetros), não é possível quem ocorra simultaneamente valores elevados para muitas componentes gaussianas.

Para se determinar quais componentes gaussianas não contribuem de forma significativa para o resultado final da verossimilhança do GMM, entretanto, é preciso calcular todas as componentes e comparar seus valores, o que, em princípio, inviabilizaria a redução de custo desejada. Todavia, como se está trabalhando com modelos multicondicionais, que modelam dados razoavelmente semelhantes, considerando que exista uma correspondência entre as componentes gaussianas de cada um dos GMM[*], ou seja, que essas componentes correspondentes dos diversos modelos de um mesmo locutor modelem regiões próximas do espaço de parâmetros, pode-se determinar as componentes dominantes para uma única condição de treinamento e extrapolar essa informação para os modelos das demais condições multicondicionais.

O MGD, dessa forma, admite como hipótese para seu funcionamento que as componentes gaussianas dominantes para uma condição de treinamento serão as mesmas para as demais condições. Essa hipótese é razoável, considerando que exista uma correspondência entre as componentes gaussianas dos diversos GMM que compõem o modelo multicondicional (condição que será tratada detalhadamente na seção 4.4.1), e considerando que cada componente dos GMM modela uma classe sonora distinta.

Para tornar essa estimativa das componentes dominantes ainda mais segura, busca-se, no MGD, realizar a estimação numa condição de ruído próxima da condição que promoverá a máxima verossimilhança do modelo. Isso porque, mesmo considerando a correspondência entre as componentes dos modelos multicondicionais, se essa estimativa for realizada para uma condição muito diferente da condição do áudio da janela de análise em questão, pode haver distorções. Dessa forma, explorando as propriedades de correlação temporal da condição do áudio, demonstradas no Método da Condição Persistente, seção 4.2, e nos Modelos Multicondicionais Adaptativos, seção 4.3, a determinação das componentes dominantes é realizada para a condição de treinamento que promoveu a máxima verossimilhança na janela antecedente à atual.

[*] Esse tipo de correspondência não existe nos modelos multicondicionais tradicionalmente empregados nos sistemas de RAL, mas é possível alterar a forma de treinamento dos modelos para que se estabeleça essa correspondência, como será detalhado na seção 4.4.1.

O número de componentes dominantes a serem utilizadas, *g*, pode ser fixo, pré-definido, ou pode ser definido de forma dinâmica, com base na análise dos valores de cada componente gaussiana (por exemplo, definindo que serão utilizadas as componentes que respondem por um determinado percentual do valor total da verossimilhança). Como o objetivo desta tese é a redução de custo computacional e como se deseja obter uma medida determinada (fixa) da redução de custo possibilitada pela técnica dos MGD, optou-se, nesse momento, por limitar as análises a situações de número de gaussianas dominantes, *g*, pré-definido. Desse modo, a redução do custo computacional com o MGD também pode ser pré-definida. Sugere-se que, em trabalhos complementares, sejam realizadas simulações com a definição do número de gaussianas por meio da análise, janela a janela, dos valores das componentes individuais; pois, ao menos em tese, é possível obter resultados melhores (menor custo computacional e melhor desempenho de identificação) com esse procedimento.

Após a determinação das gaussianas dominantes e o descarte das demais componentes não dominantes, o MGD segue a mesma rotina do MCM, ou seja, determina-se qual a condição de treinamento do modelo que apresenta o máximo valor de verossimilhança dentre todas as condições do modelo multicondicional e toma-se essa verossimilhança máxima como verossimilhança do locutor em questão (equação (3.13)).

Num modelo multicondicional, cada modelo GMM de um locutor *s* tem a forma:

$$p\left(\vec{y}_t \mid \lambda_{s,n}\right) = \sum_{i=1}^{M} p_{n,i} b_{n,i}\left(\vec{y}_t\right) \quad , \tag{4.24}$$

Se, para o modelo treinado para uma determinada condição, n_1, apenas as gaussianas b_{n_1,i_1} e b_{n_1,i_2} contribuíram significativamente para o resultado final, $p\left(\vec{y}_t \mid \lambda_{s,n}\right)$, do modelo GMM na janela de análise *t*, não é necessário calcular todas as gaussianas, $b_{n,i}$, dos modelos GMM das demais $n-1$ condições para essa mesma janela de análise. Basta, nesse caso, calcular as componentes gaussianas correspondentes a b_{n_1,i_1} e b_{n_1,i_2}, ou seja, as componentes b_{n,i_1} e b_{n,i_2} para as demais $n-1$ condições.

Para facilitar o entendimento do princípio utilizado pelo MGD, a Figura 4.31, ilustra o funcionamento desse método proposto.

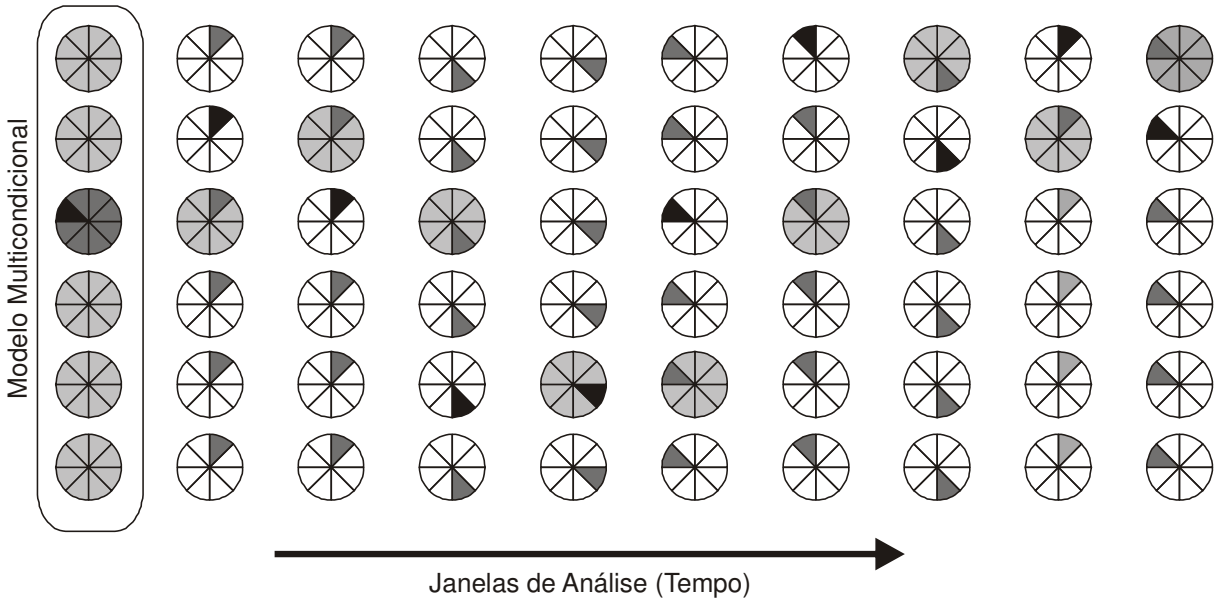

Figura 4.31: Ilustração do Método das Gaussianas Dominantes (MGD).

Deve-se ressaltar que a proposta do MGD é completamente distinta daquela do MCP ou do MMA, pois, enquanto o MCP e o MMA exploram correlação temporal da condição do áudio questionado, o MGD explora a correlação entre as componentes gaussianas correspondentes dos diferentes modelos GMM que compõem o modelo multicondicional (num mesmo instante de análise). Como são técnicas que exploram características completamente diferentes e não relacionadas, é possível utilizar conjuntamente o MGD com o MMA e, em princípio, também com o MCP[*].

Em complementação à ilustração da Figura 4.31, é inserida a programação básica do MGD no quadro da Figura 4.32.

[*] A utilização conjunta do MGD com o MCP fica restrita às janelas de análise em que há o cálculo do modelo multicondicional completo. Nas demais janelas de análise, como o MCP elimina a característica de multicondicionalidade do modelo, a utilização do MGD é inviável.

```
P(Y|λs) = 0 % verossimilhança do modelo multicondicional do
locutor s dado o conjunto de vetores de parâmetros Y

Para n de 1 a N % loop de condições de treinamento (para ini-
cialização)

     Calcule p(y[1]|λ[s,n]) utilizando m = 1 a M % verossimi-
     lhança do modelo do locutor s para a condição n dado o
     vetor de parâmetros y[t]

Fim para n

n'[1] = argmax(p(y[1]|λ[s,n]) em relação a n) % determinar a
condição de treinamento que maximiza a verossimilhança

P(Y|λs) = P(Y|λs) + p(y[1]|λ[s,n'[1]])

Para t = 2 a T % loop de janelas de análise

   Calcule p(y[t]|λ[s,n'[t-1]]), utilizando m = 1 a M % calcu-
   la todas as gaussianas da última condição máxima para de-
   terminar as componentes dominantes

   g' = índices das g componentes gaussianas que mais contri-
   buem para p(y[t]|λ[s,n'[t-1]])

   p'(y[t]|λ[s,n'[t-1]])= p(y[t]|λ[s,n'[t-1]], usando apenas
   componentes gaussianas de índices g') % verossimilhança u-
   sando apenas as componentes gaussianas de índices g'

   Para n de 1 a N, exceto n = n'[t-1] % loop de condições de
   treinamento

      Calcule p'(y[t]|λ[s,n]), usando apenas as componentes
      gaussianas de índices g'

   Fim para n

   n'[t] = argmax(p'(y[t]|λ[s,n]) em relação a n, n de 1 a N)

   P(Y|λs) = P(Y|λs) + p'(y[t]|λ[s,n'[t]])

Fim para t
```

Figura 4.32: Programação básica do MGD.

4.4.1 Método do Treinamento Progressivo

A primeira dificuldade a ser solucionada para possibilitar a utilização do MGD é o fato de que não há, nos modelos multicondicionais tradicionais, qualquer relação entre os índices das componentes gaussianas dos modelos treinados com diferentes níveis de ruído[*]. Assim, as componentes b_{n_α,i_k} e b_{n_β,i_k}, $\alpha \neq \beta$, $k = 1,..,M$, não modelam regiões relacionadas do espaço dos vetores de parâmetros, o que inviabiliza a técnica proposta.

[*] Dado que os áudios de treinamento serão semelhantes (diferindo apenas no nível de ruído aplicado), é possível que os modelos sejam razoavelmente semelhantes entre si. Entretanto, mesmo se isso ocorrer, não haverá qualquer relação entre os índices das componentes gaussianas que modelam as regiões correspondentes do espaço dos parâmetros, em decorrência das características do processo de inicialização e treinamento dos modelos.

Para resolver essa questão e criar uma relação de correspondência entre as componentes gaussianas correspondentes dos modelos GMM que compõem o modelo multicondicional, de modo a permitir a exploração dessa propriedade para redução do custo computacional, foi desenvolvido o Método do Treinamento Progressivo (MTP) dos modelos multicondicionais. No MTP, os modelos treinados para as diversas condições de ruído são derivados uns dos outros, ou seja, o modelo treinado na condição de ruído n_1 é utilizado como condição inicial para treinamento do modelo da condição n_2. Dessa maneira, como o nível de ruído do áudio de treinamento dos modelos GMM que compõem o modelo multicondicional varia progressivamente e lentamente, os modelos $\lambda_{s,n}$ gerados passam a ter uma forte semelhança, e as gaussianas b_{n,i_k}, $n = 1,...,N$ passam a modelar regiões próximas (ou correspondentes) do espaço dos vetores de parâmetros.

A Figura 4.33 ilustra a diferença entre o treinamento normal e o treinamento progressivo (MTP) de modelos multicondicionais.

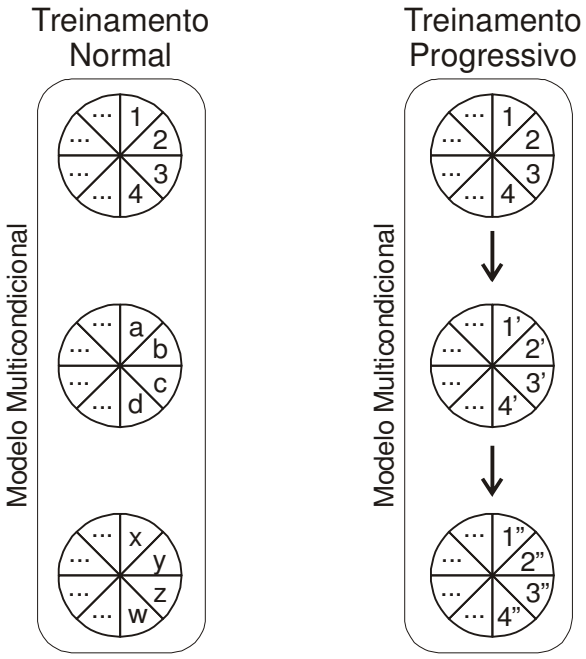

Figura 4.33: Ilustração do treinamento normal e do treinamento progressivo (MTP) de modelos multicondicionais.

No treinamento normal, não há qualquer relação entre os índices das componentes dos GMM que compõem o modelo multicondicional. Desse modo, se as componentes do GMM da primeira condição modelam determinadas regiões (1, 2, 3, ...) do espaço de vetores de parâmetros, as componentes do GMM da segunda condição modelam outras regiões

(a, b, c, ...) completamente diversas da primeiras, e assim por diante. Não há qualquer relação entre essas regiões.

Utilizando o método do treinamento progressivo (MTP), por outro lado, como os modelos são derivados uns dos outros (com exceção do primeiro), as regiões modeladas por componentes correspondentes são semelhantes ($1 \cong 1' \cong 1''$, $2 \cong 2' \cong 2''$, ...).

Uma análise que deve ser feita a fim de validar a utilização do MGD, é observar se o MTP não provoca degradação no desempenho individual dos modelos GMM que compõem o modelo multicondicional, se comparados com modelos treinados sem a utilização do MTP.

Experimentos utilizando modelos multicondicionais com 5 condições de treinamento: 50 dB, 40 dB, 30 dB, 20 dB, e 10 dB; demonstraram que, se o treinamento progressivo for realizado iniciando pela condição de maior SNR e evoluindo em direção às condições de SNR mais reduzido, não há qualquer prejuízo no desempenho dos modelos (esse tipo de treinamento progressivo será denominado de positivo e será representado por MTP^{+}). Por outro lado, se o treinamento MTP for iniciado pela condição de SNR mais baixo e evoluir em direção às condições de menos ruído, há intensa degradação do desempenho dos modelos treinados para os níveis mais baixos de ruído (esse tipo de treinamento progressivo será denominado de negativo e será representado por MTP^{-}).

Os resultados dos experimentos realizados com os diferentes tipos de treinamento estão expostos na Tabela 4.7.

Tabela 4.8: Taxas de identificações corretas de sistemas de RAL multicondicional (5 condições de treinamento:50 dB, 40 dB, 30 dB, 20 dB e 10 dB) utilizando treinamento normal, MTP^{+} e MTP^{-}.

Taxas de Identificações Corretas (%)		**SNR do Áudio de Teste (dB)**							
		60	**50**	**40**	**30**	**26**	**20**	**16**	**10**
Treinamento	**Normal**	97,0	97,5	98,2	97,8	97,2	94,5	89,7	66,2
	MTP^{+}	97,2	97,7	98,5	97,7	97,2	94,2	90,2	66,5
	MTP^{-}	6,8	7,0	8,5	15,3	18,5	35,3	48,3	71,2

Para uma melhor visualização, os resultados da Tabela 4.7 foram traçados no gráfico da Figura 4.34 a seguir.

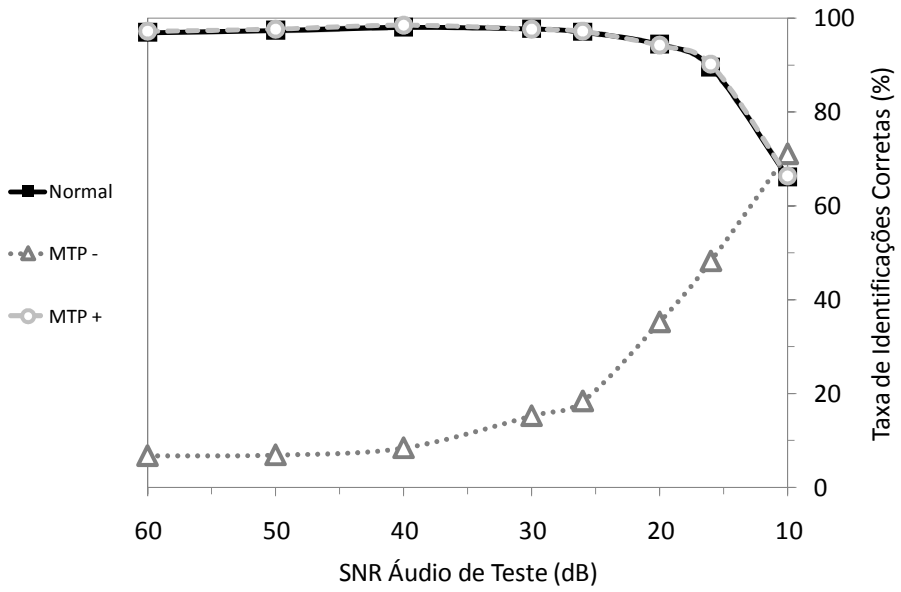

Figura 4.34: Comparação de desempenho de sistemas de RAL multicondicional (5 condições de treinamento: 50 dB, 40 dB, 30 dB, 20 dB, e 10 dB), utilizando treinamento normal, MTP^+ e MTP^-.

Como se observa, a utilização do MTP^+ praticamente não alterou o desempenho do sistema, de modo que esse tipo de treinamento pode ser utilizado para permitir a utilização do MGD sem causar prejuízos ao sistema de RAL.

O MTP^-, por sua vez, teve desempenho satisfatório apenas para a condição de ruído mais intenso utilizada. Em todos os demais casos, o modelo MTP^- teve desempenho insatisfatório.

Esse comportamento do MTP^-, de fato, não é esperado. Embora se pudesse supor que ocorreria alguma degradação no desempenho do sistema em decorrência desse método de treinamento, não se poderia prever tamanha perda. Tendo em vista os resultados obtidos, foram realizadas, seção 4.4.4, algumas análises a fim de estudar um pouco mais as limitações do MTP^-; embora, para os objetivos deste trabalho, seja suficiente o fato de o MTP^+ ter se mostrado um método de treinamento apropriado.

Uma avaliação complementar do MTP+ foi realizada, dessa voz utilizando o modelo com 12 condições de treinamento (60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB), comparando somente para os casos do treinamento normal (não progressivo) e do MTP^+, a fim de verificar eventuais alterações no desempenho do sistema na situação de avaliação mais completa. Os resultados dessa avaliação estão exibidos na Tabela 4.9.

Tabela 4.9: Taxas de identificações corretas de sistemas de RAL multicondicional (12 condições de treinamento: 60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB) utilizando treinamento normal e MTP$^+$.

Taxas de Identificações Corretas (%)		SNR do Áudio de Teste (dB)										
		60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
Treinamento	**Normal**	97,8	98,2	98,2	97,8	97,2	95,5	94,2	91,5	86,6	80,8	68,5
	MTP$^+$	99,3	99,3	98,8	98,2	96,8	94,2	93,7	91,0	86,2	81,2	67,8

Para facilitar a análise dos resultados da Tabela 4.9, os mesmos foram traçados no gráfico da Figura 4.35:

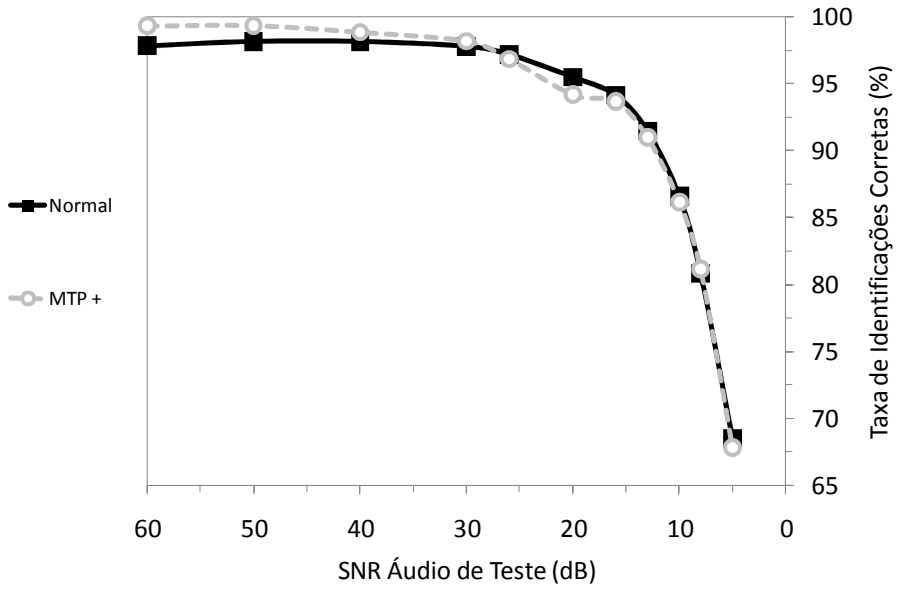

Figura 4.35: Comparação de desempenho de sistemas de RAL multicondicional (12 condições de treinamento: 60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB) utilizando treinamento normal e MTP$^+$.

Como se percebe, a utilização do MTP$^+$ proporcionou um discreto ganho para as condições de baixo ruído. Essa mesma tendência pode também ser observada no caso da simulação com o modelo multicondicional com 5 condições de treinamento, embora em intensidade menor. Para as demais situações, não houve uma tendência consistente de alteração nos resultados, de modo que as pequenas modificações observadas podem ser atribuídas simplesmente às variações dos modelos por conta do novo treinamento.

4.4.2 Vantagem Computacional

Solucionado o problema inicial, de estabelecer a correspondência entre as componentes gaussianas dos diferentes GMM que compõem o modelo multicondicional, e havendo agora um modelo multicondicional em que as componentes gaussianas b_{n,i_k}, $n = 1,...,N$, modelam regiões correspondentes do espaço de vetores de parâmetros, pode-se analisar a vantagem computacional que pode ser obtida com a utilização do MGD.

O custo computacional da identificação com o MGD depende, fundamentalmente, do número de gaussianas dominantes, *g*, utilizado. Deve-se lembrar que, para a determinação das gaussianas dominantes, é necessário calcular todas as gaussianas para ao menos uma condição de treinamento, de forma a se poder determinar quais as que contribuem de modo mais significativo para a verossimilhança do GMM, $p\left(\vec{y}_t \mid \lambda_{s,n}\right)$.

O custo computacional de uma identificação com o MGD, portanto, é dado pela soma do custo de um modelo GMM completo, com *M* componentes, mais o custo dos demais $N-1$ modelos GMM calculados com apenas *g* componentes. Dessa forma, podemos expressar o custo de uma identificação com o MGD por:

$$W_{MGD} \propto M + (N-1)\,g \tag{4.25}$$

A redução do custo computacional em comparação ao modelo multicondicional tradicional (MCM) é dada por:

$$W_{MCM} - W_{MGD}(g) \propto M \cdot N - \left[M + (N-1)\,g\right] = (M-g)(N-1)\ , \tag{4.26}$$

e diferença relativa entre os custos computacionais é calculada por:

$$\frac{\Delta W}{W} = \frac{W_{MCM} - W_{MGD}(g)}{W_{MCM}} \propto \frac{(M-g)(N-1)}{M \cdot N} \tag{4.27}$$

Observa-se, portanto, que a diferença relativa de custo computacional praticamente não depende do número de condições de treinamento, *N*, visto que a razão $(N-1)/N$ mantém-se basicamente constante para $N > 10$. Dessa forma, a expressão de (4.27) é dominada pelo termo $(M-g)/M$.

No gráfico da Figura 4.36, é ilustrada a variação do custo computacional absoluto de uma identificação com o MGD, em função do número de gaussianas dominantes utilizadas, *g*, e

do número de condições de treinamento do modelo, *N*. O custo é expresso em número de vezes *M*.

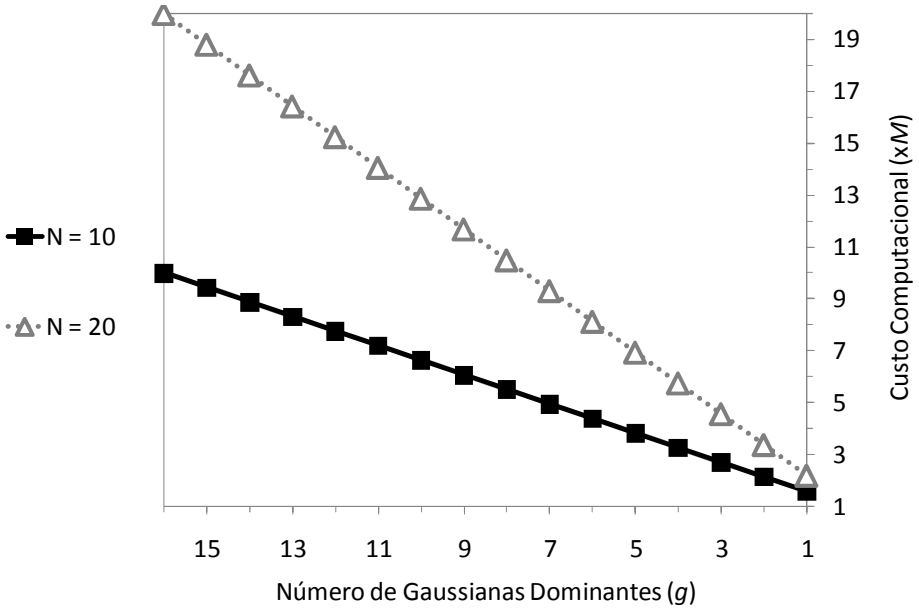

Figura 4.36: Custo computacional absoluto de sistemas com MGD, em função do número de gaussianas dominantes, *g*, para diferentes números de condições de treinamento, *N*.

O gráfico da Figura 4.47 exibe a variação do custo computacional relativo de uma identificação com o MGD, em função do número de gaussianas dominantes utilizadas, *g*, e do número de condições de treinamento do modelo, *N*. Nesse gráfico, pode-se observar a pequena influência do número de condições de treinamento do modelo, *N*, no custo computacional relativo da identificação.

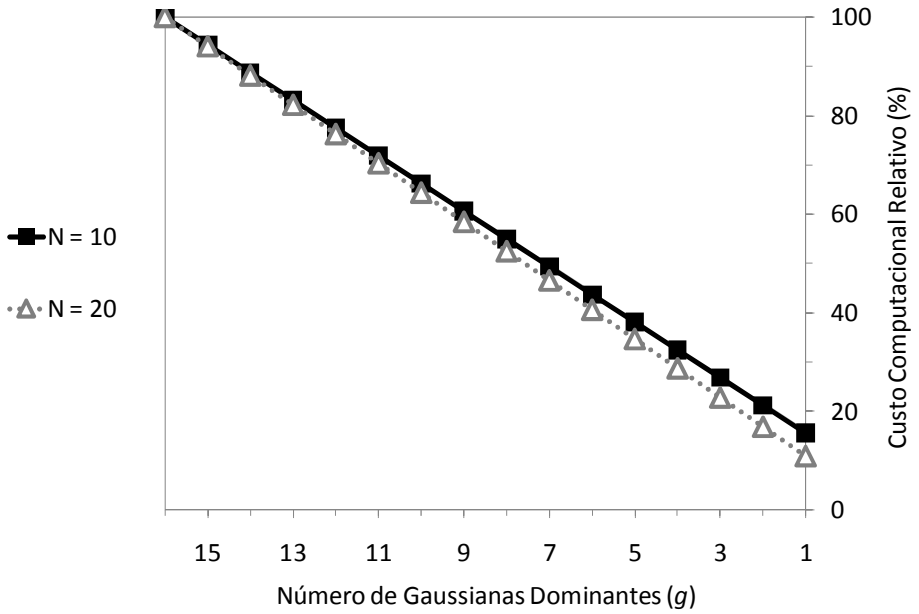

Figura 4.37: Custo computacional relativo de sistemas com MGD, em função do número de gaussianas dominantes, *g*, para diferentes números de condições de treinamento, *N*.

O custo computacional de uma identificação com o MGD pode ser ainda mais reduzido se esse método for combinado com o MMA ou com o MCP. Tendo em vista os resultados da comparação entre esses dois métodos, traçados no gráfico da Figura 4.30 (pág. 98), e diante das restrições de aplicação da combinação do MMA com o MCP[*], vê-se que é mais interessante a combinação do MGD com o MMA. Nesse caso, o custo computacional do sistema MMA-MGD é expresso por:

$$W_{MMA/MGD} \propto M + g \cdot 2a \tag{4.28}$$

A redução do custo computacional do sistemas MMA-MGD em comparação ao modelo multicondicional tradicional (MCM) é dada por:

$$W_{MCM} - W_{MMA/MGD}(a,g) \propto M \cdot N - [M + g \cdot 2a] = M(N-1) - 2g \cdot a \ , \tag{4.29}$$

e diferença relativa entre os custos computacionais é calculada por:

$$\frac{\Delta W}{W} = \frac{W_{MCM} - W_{MMA/MGD}(a,g)}{W_{MCM}} \propto \frac{M(N-1) - g \cdot 2a}{M \cdot N} = \frac{N-1}{N} - \frac{2g \cdot a}{M \cdot N} \tag{4.30}$$

Percebe-se que, com a combinação dos métodos MMA e MGD, usando $a = 1$ (o que se mostrou perfeitamente viável para o MMA) é possível restringir o custo computacional a[†]:

[*] Como discutido na nota de rodapé da página 101.
[†] Lembrando que o custo computacional mínimo obtido pela aplicação exclusiva do MMA é proporcional a *3M*, como disposto na equação (4.16).

$$W_{MMA/MGD}(1,g) \propto M + 2g \tag{4.31}$$

Dessa maneira, a diferença relativa entre os custos computacionais é dada por:

$$\frac{\Delta W}{W} = \frac{W_{MCM} - W_{MMA/MGD}(1,g)}{W_{MCM}} \propto \frac{N-1}{N} - \frac{2g}{M \cdot N} \tag{4.32}$$

O gráfico da Figura 4.38 ilustra a variação do custo computacional absoluto de sistemas com o MGD e com o MMA/MGD, usando $a = 1$, em função do número de gaussianas dominantes utilizado, *g*:

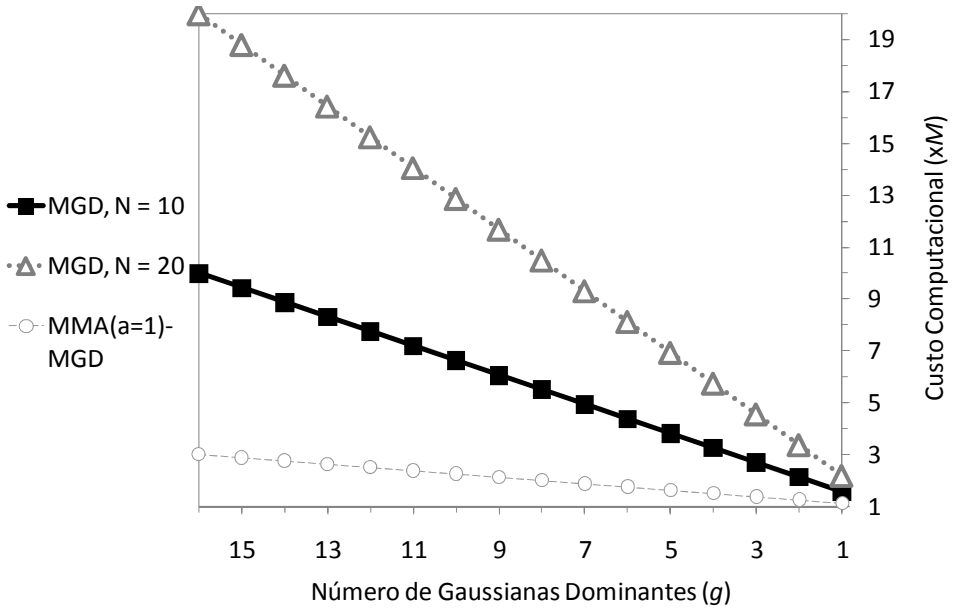

Figura 4.38: Custo computacional absoluto de sistemas com MGD e MMA(a = 1)/MGD, em função do número de gaussianas dominantes, *g*, para diferentes números de condições de treinamento, *N*.

Observa-se a significativa redução que se pode obter pela composição das duas técnicas.

A variação do custo computacional relativo em função de *g* está ilustrada no gráfico da Figura 4.39:

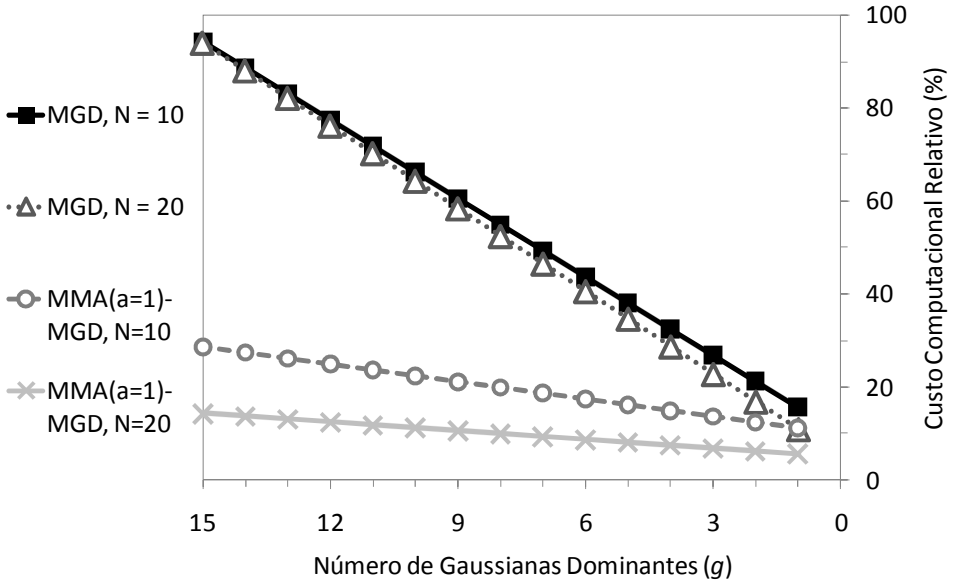

Figura 4.39: Custo computacional relativo de sistemas com MGD e MMA(a = 1)/MGD, em função do número de gaussianas dominantes, g, para diferentes números de condições de treinamento, N.

4.4.3 Resultados

Para verificar o desempenho do MGD proposto neste trabalho, foram realizados procedimentos de identificação utilizando diferentes números de gaussianas dominantes, g, e o modelo multicondicional de 12 condições de treinamento (60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB). As taxas de identificações corretas obtidas nesses procedimentos são exibidas na Tabela 4.10, juntamente com as taxas para o caso do modelo multicondicional tradicional (MCM), para comparações.

Tabela 4.10: Taxas de identificações corretas de sistemas de RAL multicondicional utilizando MGD, para diferentes valores de *g*.

Taxas de Identificações Corretas (%)		SNR do Áudio de Teste (dB)										
		60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
Gaussianas Dominantes (*g*)	**1**	99,3	99,2	98,8	96,8	94,8	93,8	91,8	89,8	86,2	81,3	72,7
	2	99,3	99,5	99,0	97,3	95,7	93,3	93,0	90,0	87,2	81,3	71,2
	4	99,3	99,3	98,8	97,8	96,8	93,8	93,2	90,3	86,3	81,5	70,0
	8	99,3	99,3	98,8	97,8	96,7	94,3	93,5	91,0	87,0	80,7	67,7
	16	99,3	99,3	98,8	97,8	96,7	94,3	93,5	91,0	87,0	80,7	67,7
MCM		97,8	98,2	98,2	97,2	96,8	95,5	94,2	91,5	86,6	80,8	68,5

Para permitir uma melhor visualização, os resultados da Tabela 4.10 foram traçados no gráfico da Figura 4.40.

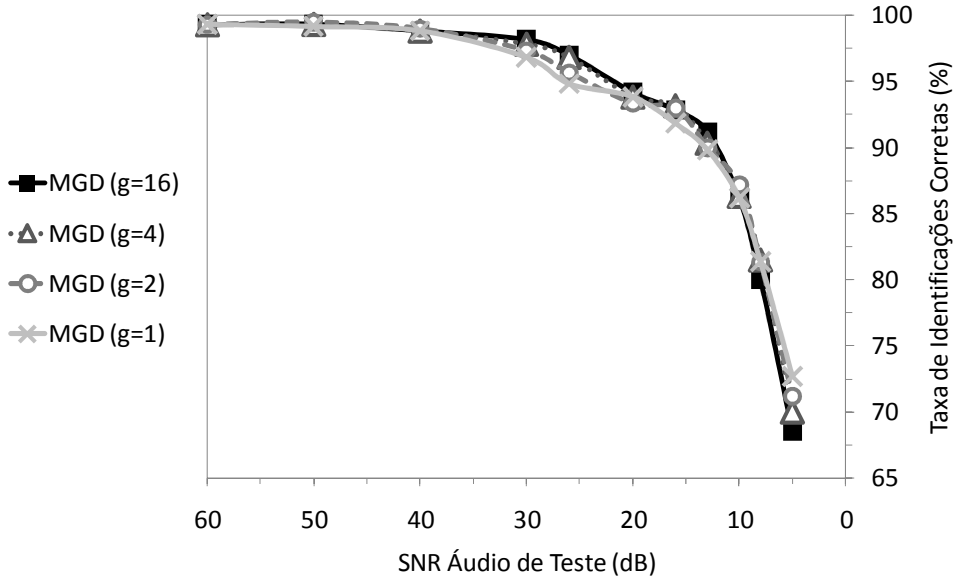

Figura 4.40: Desempenho de sistemas de RAL multicondicional utilizando MGD, para diferentes números de gaussianas dominantes, *g*.

Como se verifica, não há alterações significativas no desempenho do sistema mesmo para o caso extremo de utilizar apenas uma gaussiana dominante. Na realidade, à medida que se diminui o valor de *g*, se observa uma discreta diminuição de desempenho na zona de ruído

intermediário (entre 30 db e 20 db) e um pequeno aumento na zona de ruído extremo (abaixo de 10 db).

A fim de possibilitar uma avaliação do resultado global da utilização do MGD nas diversas condições de ruído avaliadas, foram computadas as médias das taxas de identificações corretas dispostas na Tabela 4.10. Essas médias foram traçadas em função do número de gaussianas dominantes utilizado, *g*, no gráfico da Figura 4.41.

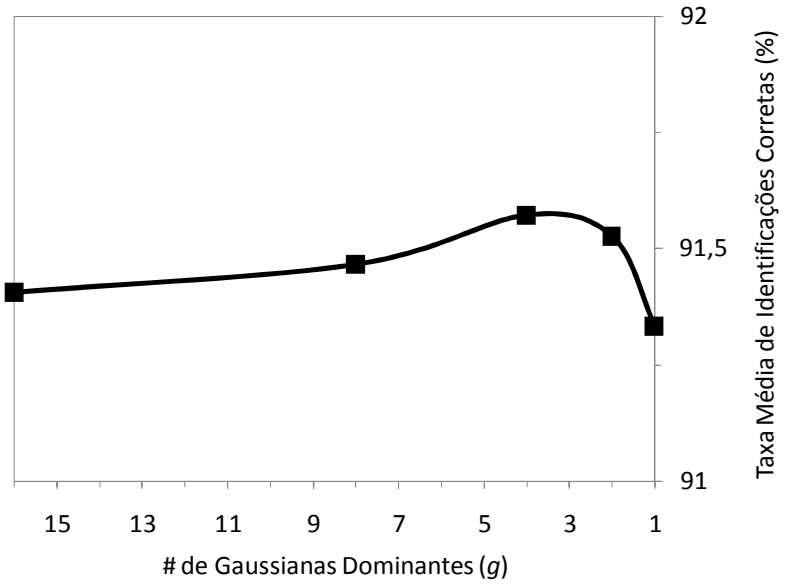

Figura 4.41: Taxa média de identificações corretas de sistemas de RAL multicondicional MGD, em função do número de gaussianas dominantes, *g*.

Como se observa no gráfico da Figura 4.41, as variações no valor de *g* provocam alterações pouco significativas no valor da média das taxas de identificações corretas. Pela análise do gráfico, mesmo a utilização de $g = 1$ seria viável, embora provoque uma pequena redução na taxa média de identificações corretas (0,1%) quando comparado com os resultados do sistema MCM tradicional. Contudo, em vista dos objetivos específicos deste trabalho, optou-se em limitar o uso do MGD ao caso de $g = 2$, pois, nessa situação, é possível obter uma significativa diminuição de custo computacional (aproximadamente 80%) enquanto ainda se obtém uma melhoria de 0,1% na taxa média de identificações corretas em comparação com os sistemas tradicionais baseados no MCM.

4.4.3.1 Resultados com a combinação MMA/MGD

Foram também realizadas avaliações com sistemas utilizando simultaneamente as técnicas de MMA e de MGD, pois, como discutido anteriormente, esses dois métodos propostos possibilitam reduções no custo computacional pela exploração de características distintas dos sistemas de RAL. Enquanto o MMA explora a correlação temporal da qualidade do áudio, o MGD explora a semelhança entre as componentes correspondentes de diferentes GMM de um modelo multicondicional.

Para verificar o desempenho do sistema combinando as técnicas MMA/MGD, foram realizados procedimentos de identificação utilizando diferentes números de gaussianas dominantes, g, e mantendo, em todos os casos, o valor de adaptabilidade $a = 1$. Destaque-se que foi fixado o valor da adaptabilidade $a = 1$ porque os resultados utilizando o MMA demonstraram que essa é a condição que proporciona melhor desempenho ao sistema. Nos procedimentos, foi utilizado o modelo multicondicional de 12 condições de treinamento (60 dB, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB).

As taxas de identificações corretas obtidas nesses procedimentos são exibidas na Tabela 4.11, juntamente com as taxas para o caso do modelo multicondicional tradicional (MCM) e do modelo MMA(a = 1), para comparações.

Tabela 4.11: Taxas de identificações corretas de sistemas de RAL multicondicional utilizando MMA(a = 1)/MGD, para diferentes valores de g.

Taxas de Identificações Corretas (%)		**SNR do Áudio de Teste (dB)**										
		60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
Gaussianas Dominantes (*g*)	**1**	99,8	100,0	99,3	98,0	97,7	95,0	92,7	90,2	90,8	86,8	80,2
	2	100,0	100,0	99,6	98,5	97,7	95,0	93,0	92,7	89,0	88,8	81,5
	4	100,0	100,0	99,7	98,5	98,0	95,8	93,2	92,7	90,3	88,2	81,3
	8	100,0	100,0	99,5	98,5	98,0	96,3	94,7	93,5	91,3	87,5	82,0
	16	100,0	100,0	99,7	98,3	97,8	96,0	95,2	93,5	91,2	89,5	81,5
MMA(a = 1)		99,7	99,7	99,3	98,7	98,2	96,8	94,3	92,8	89,5	87,3	80,7
MCM		97,8	98,2	98,2	97,2	96,8	95,5	94,2	91,5	86,6	80,8	68,5

Para permitir uma melhor visualização, os resultados da Tabela 4.11 foram traçados no gráfico da Figura 4.40.

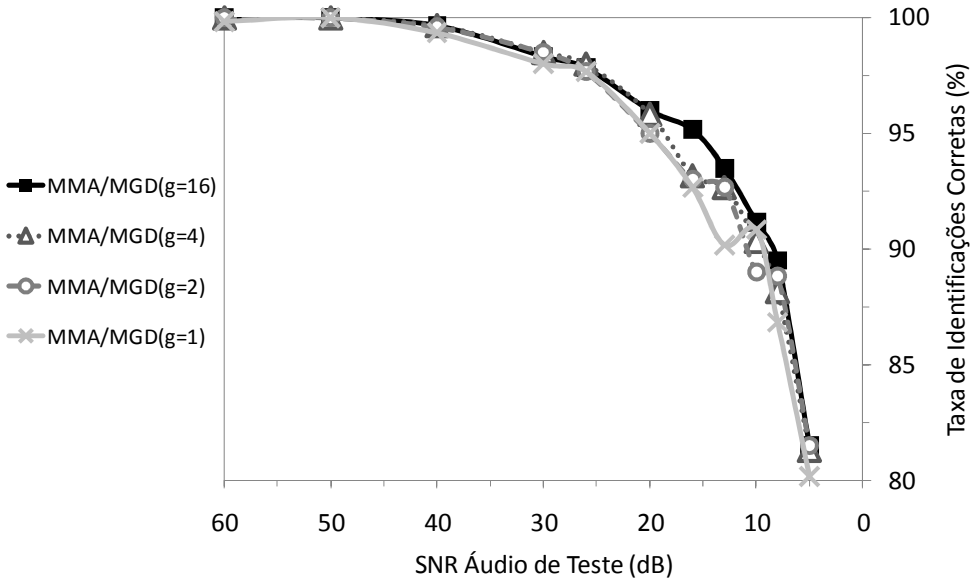

Figura 4.42: Desempenho de sistemas de RAL multicondicional utilizando MMA(a = 1)/MGD, para diferentes números de gaussianas dominantes, g.

Como se observa, assim como ocorreu no sistema que utilizava apenas o MGD, também no caso da associação MMA(a = 1)/MGD não houve grande sensibilidade ao número de gaussianas dominantes adotado, g. De fato, apenas houve alterações significativas no desempenho do sistema para as simulações com áudio de SNR entre 20 dB e 10 dB.

O gráfico da Figura 4.43, que traça a taxa média de identificações corretas dos sistemas, permite uma melhor visualização do efeito da variação do número de gaussianas dominantes, g, no desempenho do sistema.

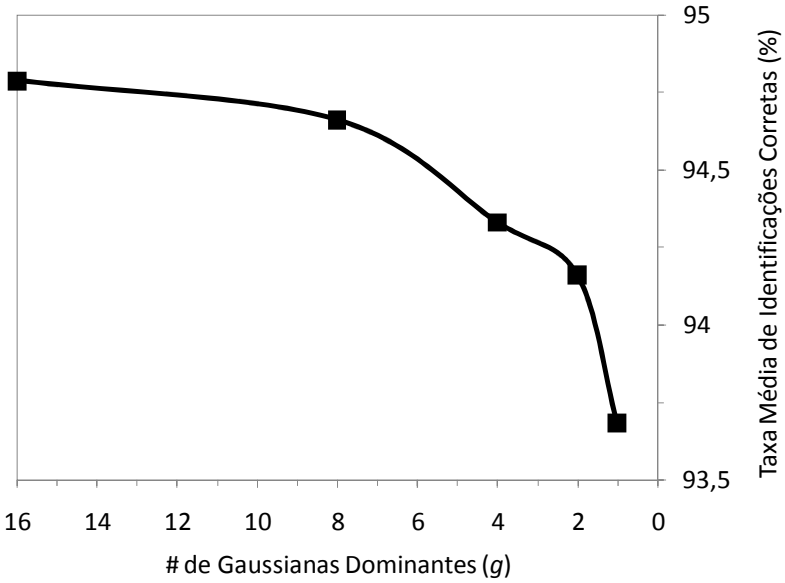

Figura 4.43: Taxa média de identificações corretas de sistemas de RAL multicondicional utilizando MMA(a = 1)/MGD, em função do números de gaussianas dominantes, g.

Apesar de haver uma progressiva degradação no desempenho, em todos os casos, o desempenho do sistema utilizando a combinação MMA(a = 1)/MGD é superior ao do sistema utilizando apenas o MGD e também é superior ao desempenho do sistema MCM tradicional, como se pode observar no gráfico da Figura 4.44. Na realidade, o sistema combinado MMA(a = 1)/MGD tem desempenho superior até mesmo ao sistema utilizando apenas o MMA(a = 1), desde que se utilize $g > 2$.

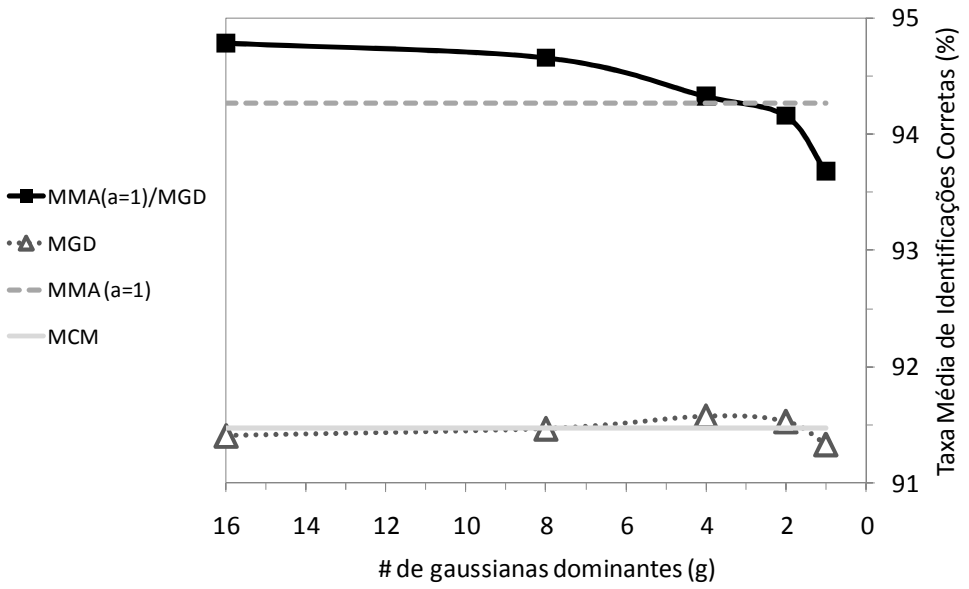

Figura 4.44: Comparação de taxas médias de identificações corretas de sistemas de RAL multicondicional utilizando MMA(a = 1)/MGD e MGD, em função do números de gaussianas dominantes, g; MMA(a = 1) e MCM[*].

Como se observa, a utilização de sistemas de RAL combinando a técnica dos modelos multicondicionais adaptativos (MMA) com o método das gaussianas dominantes (MGD) possibilita reduções de custo computacional da ordem de 90% enquanto promove ainda ganhos significativos na taxa média de identificações corretas do sistema (superiores a 2%) quando comparados com os resultados dos sistemas MCM tradicionais.

Deve-se ressaltar que a utilização de sistemas de RAL conjugando as técnicas MMA e MGD permite a construção de modelos multicondicionais com qualquer número de condições de treinamento, N, enquanto se mantém um custo computacional equivalente pouco superior ao de um único modelo GMM não multicondicional, como se observa na equação (4.28).

4.4.4 Análise do Método do Treinamento Progressivo Negativo (MTP$^-$)

Como visto, o MTP$^-$ promoveu uma intensa degradação do desempenho do sistema para todas as situações, exceto para o caso de máximo ruído (mínimo SNR). Embora fosse previsível algum prejuízo ao desempenho do sistema, o nível de degradação observado exce-

[*] Os traços correspondentes aos sistemas utilizando apenas MMA e utilizando MCM não estão em função do número de gaussianas dominantes, g, pois esse parâmetro não existe nessas modelagens. Esses traços são valores fixos do desempenho médio dos sistemas correspondentes, representados apenas para fins comparativos.

deu as expectativas, levando a se obter taxas de identificação corretas inferiores a 10% para áudio sem ruído.

A única diferença entre os métodos de treinamento tradicional (em que cada modelo é treinado separadamente), o MTP^{+} e o MTP^{-} é a forma de escolha das condições iniciais dos modelos. Assim, somente se pode atribuir o baixo desempenho do sistema MTP^{-} a uma escolha de condições iniciais tal que impossibilite a evolução do modelo para um conjunto de parâmetros mais adequado. Isso pode ocorrer se as condições iniciais estiverem próximas de um ponto de máximo local muito inferior ao máximo global[*].

Por ter ocorrido essa queda de desempenho além do esperado, foram feitas algumas análises adicionais a fim de tentar suas identificar as razões. A primeira análise baseou-se na observação dos valores de erro alcançados no treinamento dos modelos, comparando o valor do erro obtido para o modelo treinado normalmente com o valor do erro obtido no MTP^{-} e no MTP^{+}. Os resultados não demonstraram alterações significativas nos valores do erro para qualquer tipo de treinamento. As diferenças apuradas foram da ordem de 1% e não ocorreram de modo sistemático, sendo que o erro mínimo era alternadamente alcançado por cada um dos métodos de treinamento (normal, MTP^{+} ou MTP^{-}).

Uma segunda análise realizada consistiu na repetição do mesmo treinamento MTP^{-}, mas tomando como condição inicial modelos treinados diretamente (modo tradicional) para condições de SNR progressivamente mais altas[†]. Essa análise buscava verificar se a utilização de uma condição inicial menos ruidosa para o MTP^{-} teria algum efeito no resultado final da modelagem. O que se observou, nesse caso, foi que, à medida que o valor da SNR inicial do treinamento era elevada, o modelo melhorava seu desempenho. Contudo, esse desempenho era muito semelhante ao de um modelo GMM não multicondicional treinado para a condição inicial de ruído utilizado no MTP^{-}. Então, em suma, a adição de outros modelos treinados com o MTP^{-}, na modelagem multicondicional, praticamente não promoveu ganhos sistema de RAL.

O que se pode concluir das análises efetuadas é que o MTP^{-} não é uma forma de treinamento viável para modelos multicondicionais, pois não permite a adequada evolução dos

[*] No caso, não se deve falar exatamente de um "máximo global", pois não é possível determinar a existência desse ponto nem tão pouco localizá-lo. O termo, nesse contexto, está sendo utilizado para referenciar um ponto de máximo significativamente maior que o ponto a que o modelo chegou, visto o baixo desempenho do sistema.

[†] Nessa análise, não foram criados modelos treinados para condições SNR abaixo do valor da condição inicial estabelecida.

modelos para uma configuração adequada às condições de SNR mais elevadas que a condição inicial.

4.5 MODELOS DE MISTURA DE GAUSSIANAS MULTIRRESOLUÇÃO

Uma outra proposta apresentada para diminuir o custo computacional dos processos de identificação de locutores tem por objetivo diminuir o valor efetivo do número de componentes, *M*, de cada modelo GMM. Apesar dos experimentos realizados no sentido de minimizar o número de componentes dos modelos GMM sem comprometer a qualidade da representação dos locutores (Reynolds, 1995), um fato relevante não foi completamente explorado pela comunidade científica no sentido de otimizar o custo computacional associado às tarefas de identificação automática: a utilização de modelos com número de componentes inferior a 16 (para o caso de áudio sem ruído) realmente diminui significativamente o desempenho do sistema com relação ao número de locutores corretamente identificados, entretanto, nessa situação, espera-se que o modelo correto ainda seja classificado nas primeiras posições. Assim, a modelagem com um número menor de componentes pode não ser capaz de determinar exatamente o locutor correto, mas deve ser capaz de separar, dentro do universo, um subgrupo que conterá o locutor correto.

A fim de aproveitar essa idéia, se a condição de sucesso for flexibilizada, acredita-se ser possível obter taxas elevadas de acerto mesmo para modelos com apenas 2 ou 4 componentes. Essa flexibilização da condição de sucesso do modelo pode ser expressa matematicamente como:

$$\hat{s} \in \tilde{U} \,, \tag{4.33}$$

sendo $\tilde{U}$ um subconjunto do universo de locutores, *U*, dado por:

$$\tilde{U} = \left\{ \lambda_s \in U, \left[\frac{\sum_{t=1}^{T} \log p\left(\vec{y}_t \mid \lambda_s\right)}{T} \right] \geq \xi \right\}, \tag{4.34}$$

onde ξ é o *C*-ésimo maior valor da expressão entre colchetes de (4.34) calculado para todos os modelos do universo. O valor de *C* é definido a partir da ordem do modelo utilizado e do desempenho exigido.

Evidentemente, a flexibilização da condição de sucesso do sistema estabelecida em (4.33) e (4.34) suscita um novo problema: o resultado da identificação não é mais um locutor único, mas um conjunto com *C* locutores, o que impede uma determinação precisa do autor do trecho de voz. Entretanto, essa identificação precisa pode ser obtida por meio de um novo sistema de GMM, dessa vez com 16 componentes e utilizando a condição de identificação padrão, expressa em (2.12), aplicada sobre o conjunto restrito de locutores $\tilde{U}$.

A sistematização do procedimento de identificação de locutores por meio de modelos GMM de diferentes resoluções (números de componentes) em etapas sucessivas foi proposta por meio de uma nova forma de modelagem, os Modelos de Mistura de Gaussianas Multirresolução (MR-GMM) (D'Almeida *et al*, 2008 e 2008b).

Os MR-GMM são basicamente uma extensão dos modelos GMM utilizados tradicionalmente. A principal diferença que há entre essas duas técnicas de modelagem é que, num modelo MR-GMM, para um único locutor existem dois ou mais modelos GMM distintos com graus de complexidade (número de componentes) diferentes. Dessa maneira, o modelo MR-GMM pode ser escrito, de forma análoga à equação (2.4), como:

$$\Lambda = \{\lambda_k, k = 1, ..., K\}$$
$$\Lambda = \left\{ \begin{array}{c} \left\{p_{1,i_1}, \vec{\mu}_{1,i_1}, \Sigma_{1,i_1}\right\} \\ \left\{p_{2,i_2}, \vec{\mu}_{2,i_2}, \Sigma_{2,i_2}\right\} \\ \vdots \\ \left\{p_{K,i_K}, \vec{\mu}_{K,i_K}, \Sigma_{K,i_K}\right\} \end{array} \right\}, \begin{array}{l} i_1 = 1, ..., M_1; \\ i_2 = 1, ..., M_2; \\ \vdots \\ i_K = 1, ..., M_K \end{array} \quad (4.35)$$

sendo $M_k > M_{k-1}$. Destaca-se que, em (4.35), o subscrito *k* do modelo λ_k não indexa os diferentes locutores, o índice *k* refere-se aos submodelos componentes de um único MR-GMM; consequentemente, todos os modelos λ_k são de um único locutor.

O conjunto de todos os modelos MR-GMM, ou o universo de locutores modelados, *U*, é definido de forma semelhante à equação (2.5) como:

$$U = \{\Lambda_s, s = 1, ..., S\} \quad (4.36)$$

O treinamento de cada um dos submodelos, λ_k, de um modelo MR-GMM, Λ, é realizado como o treinamento de um modelo GMM normal e independente. Os submodelos distintos de um mesmo locutor podem ser treinados a partir de um mesmo trecho de áudio. Isso não traz nenhum prejuízo ao modelo global, pois o objetivo dos MR-GMM é o de utilizar mo-

delos de resoluções distintas para minimizar o custo computacional total das tarefas de identificação de locutor. Na realidade, o treinamento dos submodelos com um mesmo trecho de áudio é a alternativa mais natural a ser utilizada, pois evita a necessidade de alteração nos bancos de dados existentes.

Na fase de identificação do locutor (fase de testes) há uma alteração significativa entre os modelos MR-GMM e os modelos GMM. Nos modelos GMM, o modelo de cada locutor é avaliado com relação ao áudio questionado para que seja definido o modelo que maximiza a probabilidade de ocorrência do áudio questionado, de acordo com (2.12). Nos modelos MR-GMM, por outro lado, a identificação não é executada em uma única avaliação. São realizadas etapas de identificações sucessivas, cada uma utilizando um modelo de resolução maior que o precedente, e os locutores vão sendo selecionados gradualmente até que, na última etapa, chega-se ao melhor candidato.

A idéia fundamental dos MR-GMM é diminuir o custo computacional de uma tarefa de identificação de locutores pela redução da complexidade média dos modelos utilizados, mantendo o mesmo desempenho de identificação. A fim de obter essa melhoria sem afetar o desempenho do sistema, os MR-GMM utilizam uma seleção gradual dos locutores com modelos de complexidade crescente de modo a utilizar modelos de alta complexidade apenas para um número reduzido de locutores do universo.

Na primeira fase da identificação, são utilizados os modelos de menor complexidade de cada locutor, $\Lambda_{s,1}$, para selecionar, dentre todo o universo de modelos, U, os C_1 modelos que melhor se ajustam ao áudio. Indicaremos por $\Lambda_{s,k}$ o submodelo GMM λ_k associado modelo MR-GMM Λ_s. O resultado dessa primeira fase da identificação é um subconjunto do universo de locutores, U_1, contendo os C_1 modelos dos locutores que apresentaram os melhores resultados quando avaliados na resolução mais baixa.

$$U_1 = \left\{ \Lambda_s \in U, \left[\frac{\sum_{t=1}^{T} \log p\left(\vec{y}_t \mid \Lambda_{s,1}\right)}{T} \right] \geq \xi_1 \right\}, \tag{4.37}$$

sendo ξ_1 o C_1-ésimo maior valor da expressão entre colchetes em (4.37), avaliada para todos os modelos $\Lambda_{s,1}$ do Universo U.

Na segunda fase, repete-se o procedimento realizado na fase inicial, considerando desta vez os modelos de segunda menor resolução, $\Lambda_{s,2}$, e tomando como universo o conjunto U_1, resultado da etapa anterior. Nessa fase, tem-se por resultado um subconjunto $U_2 \subset U_1$, dado por:

$$U_2 = \left\{ \Lambda_s \in U_1, \left[\frac{\sum_{t=1}^{T} \log p\left(\vec{y}_t \mid \Lambda_{s,2}\right)}{T} \right] \geq \xi_2 \right\}, \tag{4.38}$$

sendo ξ_2 o C_2-ésimo maior valor da expressão entre colchetes em (4.38) avaliada para os modelos $\Lambda_{s,2}$, $s \in U_1$.

O processo continua diminuindo gradualmente o universo de locutores seguindo:

$$U_{k+1} = \left\{ \Lambda_s \in U_k, \left[\frac{\sum_{t=1}^{T} \log p\left(\vec{y}_t \mid \Lambda_{s,k+1}\right)}{T} \right] \geq \xi_{k+1} \right\}, \tag{4.39}$$

até que, na última etapa, K, é determinado o modelo que melhor se ajusta ao áudio questionado (portanto, sempre se tem $C_K = 1$) pela expressão análoga a (2.12):

$$\tilde{s} = U_K = \left\{ \Lambda_s \in U_{K-1}, \frac{\sum_{t=1}^{T} \log p\left(\vec{y}_t \mid \Lambda_{s,K}\right)}{T} \geq \xi_K \right\} = \arg \max_{s \in U_{K-1}} \frac{\sum_{t=1}^{T} \log p\left(\vec{y}_t \mid \Lambda_{s,K}\right)}{T} \tag{4.40}$$

4.5.1 Vantagem Computacional

A vantagem computacional da utilização dos MR-GMM decorre da possibilidade de se diminuir a complexidade média dos modelos utilizados no processo de identificação de um locutor. Nesse sentido, a definição dos parâmetros M_k, a quantidade de componentes de cada um dos submodelos $\Lambda_{s,k}$, e C_k, a quantidade de locutores classificados de uma etapa de identificação para a seguinte, é primordial.

O custo computacional total de uma identificação com modelos MR-GMM, $W_{MR\text{-}GMM}$, é dado por:

$$W_{MR-GMM} \propto M_1 + \sum_{k=2}^{K} M_k \frac{C_{k-1}}{S} = \sum_{k=1}^{K} M_k \frac{C_{k-1}}{S}, \tag{4.41}$$

sendo que, para simplificação, definiu-se $C_0 = S$ (indicando que na primeira avaliação são considerados todos os locutores)

Deve-se notar que, como é realizado mais de um cálculo de verossimilhança para um mesmo locutor (os locutores avaliados e classificados numa etapa k do processo de identificação, como modelos mais simples, são novamente testados na etapa $k + 1$, com modelos mais complexos), uma escolha inadequada dos M_k e dos C_k pode ocasionar até mesmo o aumento do custo computacional total do processo em comparação com os modelos GMM tradicionais.

Por exemplo, supondo um modelo MR-GMM com apenas duas resoluções $M_1 = 8$ e $M_2 = 16$ (a primeira fase será realizado com um modelo de 8 componentes e a segunda e última fase com um modelo de 16 componentes) e com $C_1 = 0{,}5.S$ (passam para a segunda fase do treinamento os 50% melhores modelos do universo), temos que o custo total do processo será

$$W_{MR-GMM} \propto M_1 + M_2 \frac{C_1}{S} = 16 = M_2 = W_{GMM}, \tag{4.42}$$

e não haverá qualquer diminuição no custo computacional com relação à utilização de um modelo GMM de M=16 componentes.

Destaque-se que, dada sua formulação, desde que os valores dos parâmetros M_k e C_k sejam adequadamente ajustados, espera-se que um modelo MR-GMM que efetue o último ciclo de descarte com modelos de número de componentes M_K tenha desempenho de identificação equivalente ao de um modelo GMM com $M \cdot K$ componentes. Por essa razão, as comparações de custo computacional devem ser tomadas com relação a um modelo GMM dessa complexidade.

Para que haja efetiva diminuição no custo computacional total do processo é necessário escolher os valores dos Mk e dos C_k de forma que:

$$W_{MR-GMM} \propto \sum_{k=1}^{K} M_k \frac{C_{k-1}}{S} < M_K = W_{GMM} \tag{4.43}$$

A diminuição relativa do custo computacional do processo de identificação pode ser calculada por:

$$G = 1 - \frac{W_{MR-GMM}}{W_{GMM}} = 1 - \frac{\sum_{k=1}^{K} M_k \frac{C_{k-1}}{S}}{M_K} = 1 - \sum_{k=1}^{K} \frac{M_k}{M_K} \frac{C_{k-1}}{S} \quad (4.44)$$

A fim de exemplificar como o uso dos MR-GMM pode diminuir o custo computacional de tarefas de identificação foram traçados, no gráfico da Figura 4.45, os custos computacionais de três modelos MR-GMM: o primeiro, usando $M_1 = 2$ e $M_2 = 16$, o segundo usando $M_1 = 4$ e $M_2 = 16$ e o terceiro usando $M_1 = 4$ e $M_2 = 24$. Os custos foram traçados como função da relação entre o parâmetro C_1, o número de locutores classificados para a segunda etapa da identificação, e do número total de locutores, S:

$$W = f\left(\frac{C_1}{S}\right)$$

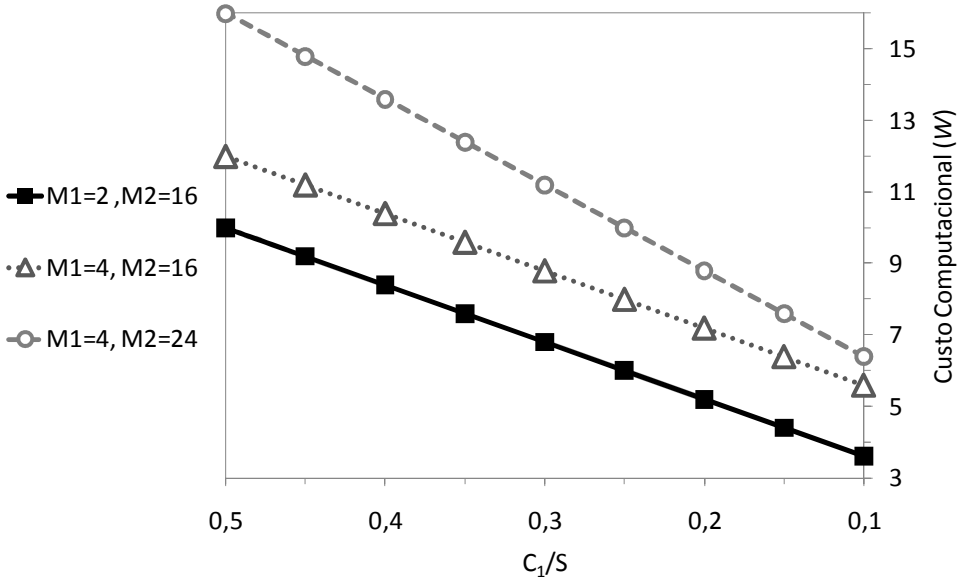

Figura 4.45: Custo computacional absoluto de sistemas de RAL utilizando MR-GMM, em função de C_1/S.

Como se verifica, é possível atingir custos muito reduzidos, utilizando MR-GMM com valor reduzido de M_1, o número de componentes do modelo utilizado na primeira seleção, e utilizando também uma baixa relação C_1/S, a relação entre o número de locutores a ser avaliado pelo modelo de alta complexidade e o número total de locutores do universo. Contudo, é necessário realizar experimentos práticos a fim de se determinar até que ponto é possível reduzir o custo computacional sem prejudicar o desempenho do sistema.

4.5.2 Resultados

A fim de verificar o desempenho dos sistemas MR-GMM, foram realizados experimentos com modelos treinados com áudio sem ruído e avaliados com áudio de vários níveis de ruído, de modo a analisar também a robustez da modelagem ao ruído. Nas avaliações, foram utilizados os modelos ilustrados na Figura 4.45, com diversos valores da relação C_1/S, de modo a permitir uma visão detalhada da redução de custo computacional que se pode obter com os modelos MR-GMM. Foram realizados experimentos com áudio de frequência de amostragem 22 kHz com quantização de 16 bits, e com frequência de amostragem 8 kHz com quantização de 8 bits μ-law, a fim de observar a influência dessas características na nova modelagem proposta.

O gráfico da Figura 4.46 representa as taxas de identificações corretas em função do custo computacional obtidas para simulações com frequência de amostragem de 22 kHz, sem ruído, para sistemas com modelos GMM e modelos MR-GMM. Nos casos dos modelos GMM, o custo computacional é proporcional ao número de gaussianas que compõem o modelo, *M*, conforme (4.1). Para os modelos MR-GMM, o custo computacional é calculado por (4.41).

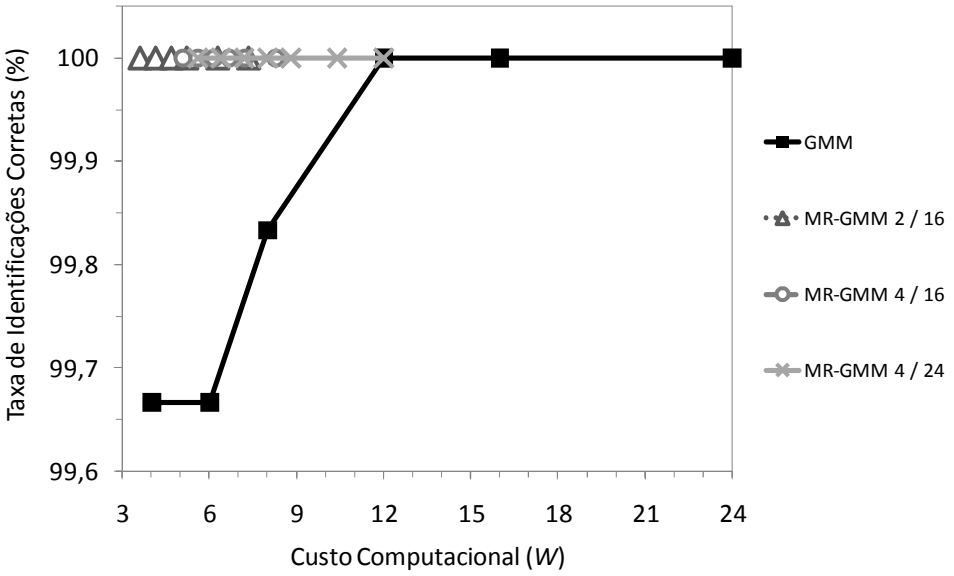

Figura 4.46: Comparação de taxas de identificações corretas de sistemas GMM e MR-GMM, em função do custo computacional, *W*, para frequência de amostragem 22 kHz, sem ruído.

Como se verifica pela observação do gráfico, o uso de modelos MR-GMM permite obter desempenho de 100% de acertos mesmo para custos computacionais $W < 4$; enquanto que essa mesma taxa de acertos só é conseguida com GMMs de 12 componentes, $W = 12$.

Com o objetivo de analisar também o desempenho dos sistemas em situações de áudio ruidoso, foram realizadas simulações com áudio de 12 condições de SNR: 60 dB*, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB, todas aplicadas aos modelos treinados com áudio sem ruído. O gráfico da Figura 4.47 representa as taxas médias de identificações corretas obtidas para os experimentos com frequência de amostragem de 22 kHz, nas diversas condições de ruído, em função do custo computacional, para sistemas com modelos GMM e modelos MR-GMM.

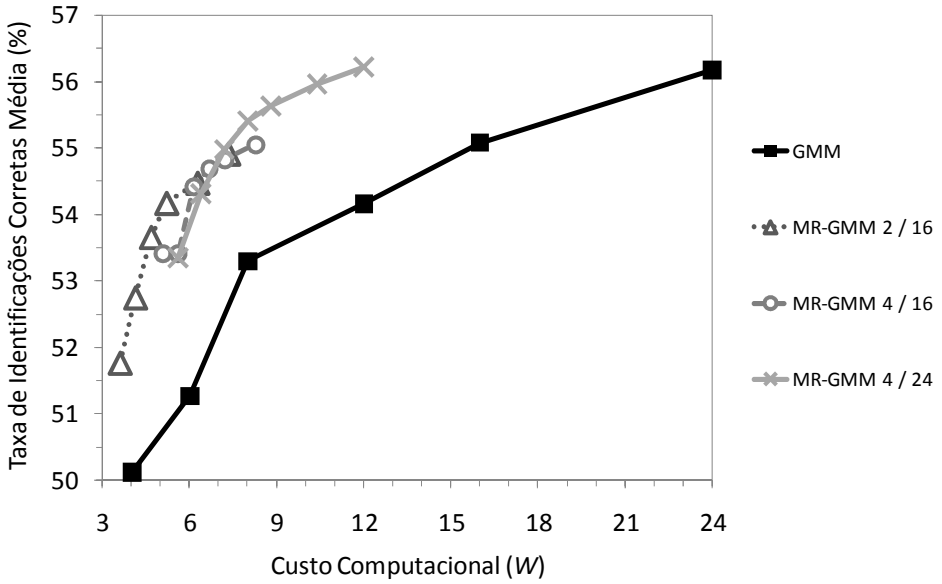

Figura 4.47: Comparação de taxas médias de identificações corretas de sistemas GMM e MR-GMM, em função do custo computacional, *W*, para frequência de amostragem 22 kHz.

O gráfico da Figura 4.48 exibe detalhe do gráfico da Figura 4.47, para melhor visualização.

* O áudio de SNR igual a 60 dB é, na realidade, o áudio originalmente capturado. Essa relação sinal-ruído corresponde aos ruídos intrínsecos do sistema de captura, conforme detalhado na seção 3.1.

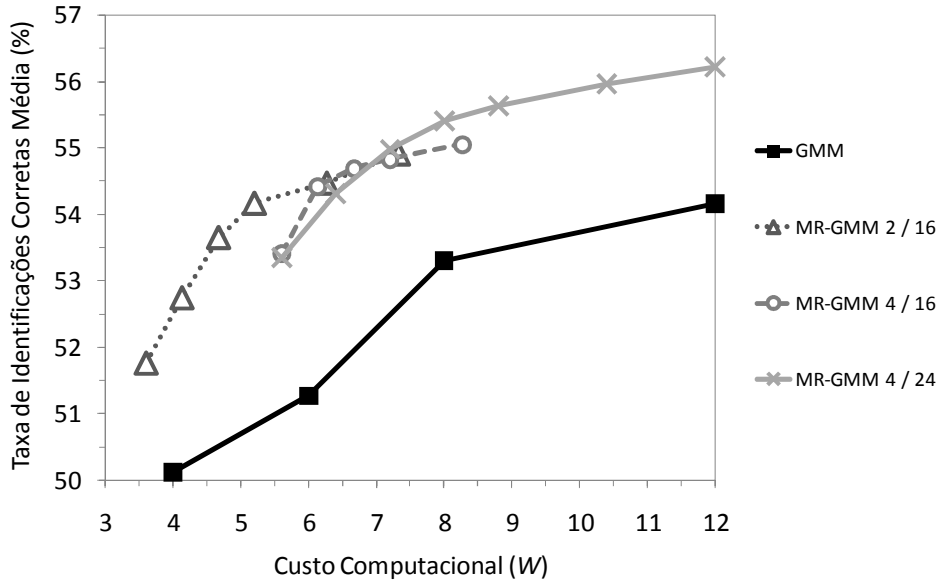

Figura 4.48: Comparação de taxas médias de identificações corretas de sistemas GMM e MR-GMM, em função do custo computacional, *W*, para frequência de amostragem 22 kHz.

Verifica-se novamente a significativa redução computacional que pode ser obtida pela utilização dos MR-GMM, que, mantendo o mesmo desempenho do GMM 16, atinge diminuições de custo da ordem de 50%. Assim, o desempenho obtido com o GMM de 16 componentes, $W = 16$, pode ser alcançado pelo MR-GMM 4 / 16 com C_1/S=0,27 (W=8,27), numa redução de 48% no custo computacional; ou pelo MR-GMM 4 / 24 com C_1/S=0,13 (W=7,20), reduzindo em 55% o custo. Por outro lado, com o MR-GMM 4 / 24 com C_1/S=0,33 (W=12,0) é possível obter desempenho médio 1,1% superior ao GMM 16 e ainda manter uma redução de 25% no custo computacional.

Para permitir uma análise mais detalhada do desempenho dos sistemas com modelos MR-GMM, foram traçadas, no gráfico da Figura 4.49 as diferenças entre a taxa de identificações corretas do GMM 16 e dos modelos MR-GMM citados no parágrafo anterior, para cada condição de áudio ruidoso testada.

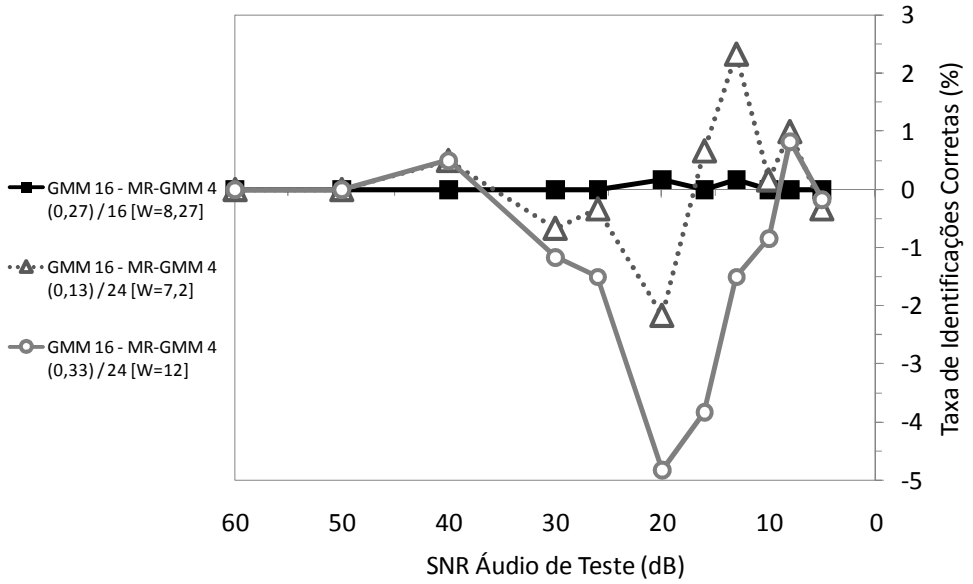

Figura 4.49: Diferença nas taxas de identificações corretas entre sistemas GMM e MR-GMM, em função do nível de ruído do áudio de teste, para frequência de amostragem 22 kHz.

Observa-se que o modelo MR-GMM 4 / 16 com C_1/S=0,27 (W=8,27), tem desempenho muito semelhante ao do GMM 16 para todas as condições de ruído testadas, o que é esperado, pois esse MR-GMM utiliza, na segunda etapa da identificação, exatamente o esse modelo GMM 16. O modelo MR-GMM 4 / 24 com C_1/S=0,13 (W=7,20), por outro lado, apresenta diferenças de desempenho, ora positivas ora negativas, embora essas diferenças não sejam muito acentuadas (limitadas a 2,3) e, no computo da média, o desempenho global seja equivalente ao do GMM 16. Já o modelo MR-GMM 4 / 24 com C_1/S=0,33 (W=12,0) tem desempenho muito superior ao GMM 16 para a maioria das condições de ruído testadas, embora, em duas das condições, tenha apresentado pequenas perdas.

Analisando agora os resultados para o áudio com frequência de amostragem de 8 kHz e quantização de 8 bits, μ-law, pode-se observar, no gráfico da Figura 4.50, as taxas de identificações corretas em função do custo computacional, em áudio sem ruído, para sistemas com modelos GMM e modelos MR-GMM.

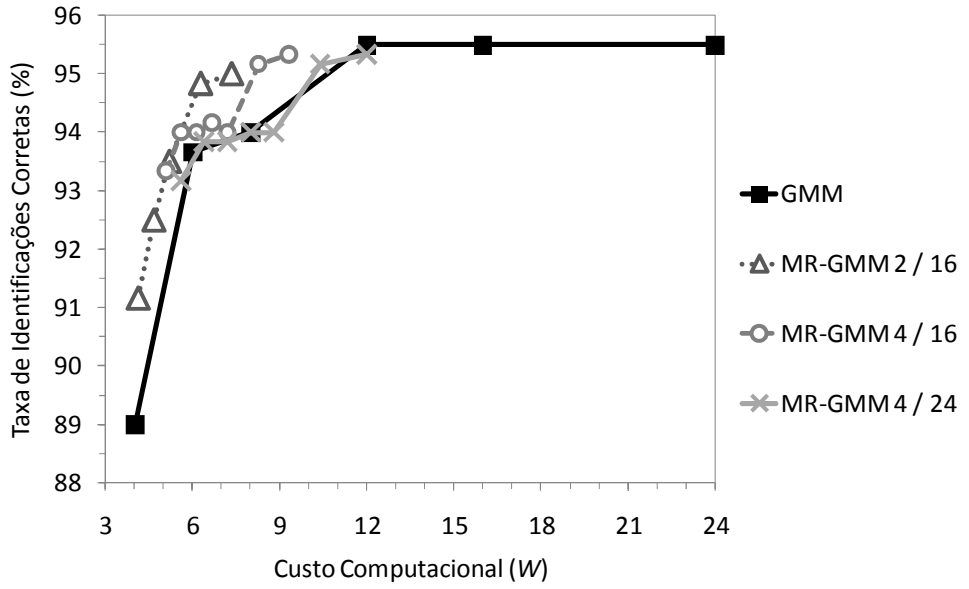

Figura 4.50: Comparação de taxas de identificações corretas de sistemas GMM e MR-GMM, em função do custo computacional, para frequência de amostragem 8 kHz, quantização de 8 bits μ-law, sem ruído.

Para permitir uma melhor visualização, o gráfico da Figura 4.51 apresenta detalhe do gráfico da Figura 4.50, na região de custo computacional limitada a $W = 12$.

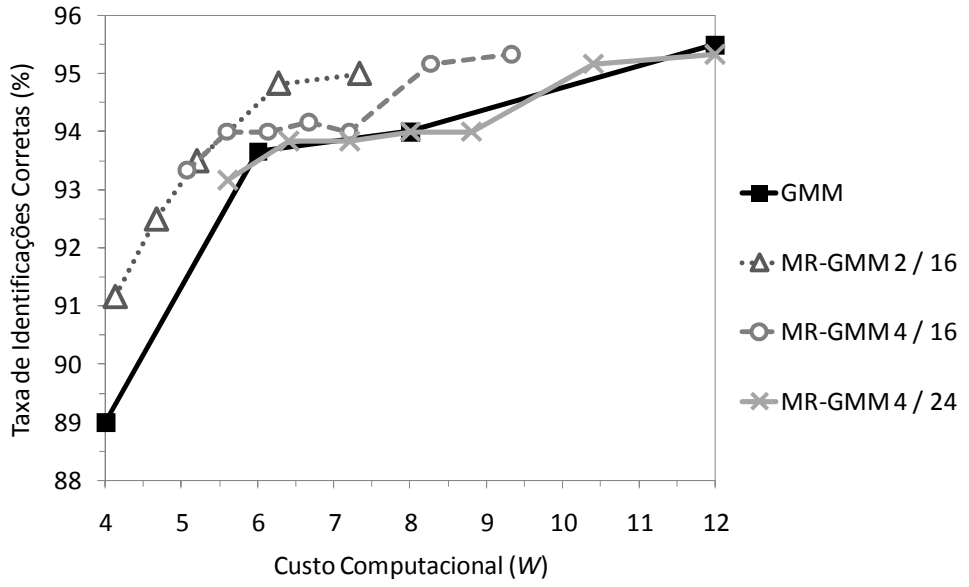

Figura 4.51: Comparação de taxas de identificações corretas de sistemas GMM e MR-GMM, em função do custo computacional, para frequência de amostragem 8 kHz, quantização de 8 bits μ-law, sem ruído.

Constata-se que, nessa situação, o modelo MR-GMM 4 / 24 não apresentou ganho com relação aos modelos GMM testados. Os outros modelos MR-GMM testados demonstraram ganhos consistentes em relação à modelagem GMM.

Do mesmo modo como realizado para a frequência de amostragem de 22 kHz, foi também analisado o desempenho dos sistemas em situações de áudio ruidoso, realizando simulações com áudio de 12 condições de SNR: 60 dB*, 50 dB, 40 dB, 30 dB, 26 dB, 20 dB, 16 dB, 13 dB, 10 dB, 8 dB, 5 dB e 3 dB, todas aplicadas aos modelos treinados com áudio sem ruído. O gráfico da Figura 4.52 representa as taxas médias de identificações corretas obtidas para os procedimentos com frequência de amostragem de 8 kHz, μ-law, nas diversas condições de ruído, em função do custo computacional, para sistemas com modelos GMM e modelos MR-GMM.

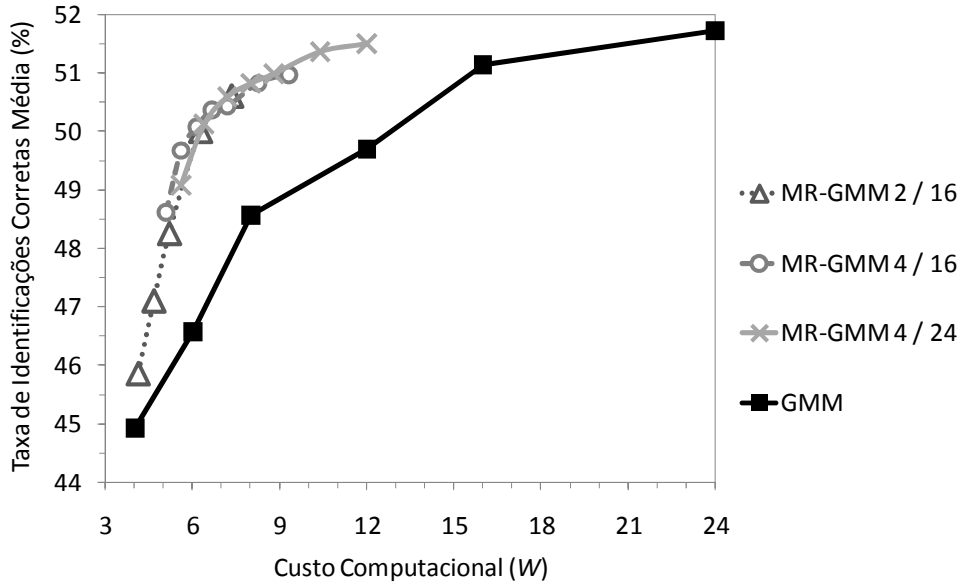

Figura 4.52: Comparação de taxas médias de identificações corretas de sistemas GMM e MR-GMM, em função do custo computacional, para frequência de amostragem 8 kHz, quantização de 8 bits μ-law.

O gráfico da Figura 4.53 exibe detalhe do gráfico da Figura 4.52, para melhor visualização.

* O áudio de SNR igual a 60 dB é, na realidade, o áudio originalmente capturado. Essa relação sinal-ruído corresponde aos ruídos intrínsecos do sistema de captura, conforme detalhado na seção 3.1.

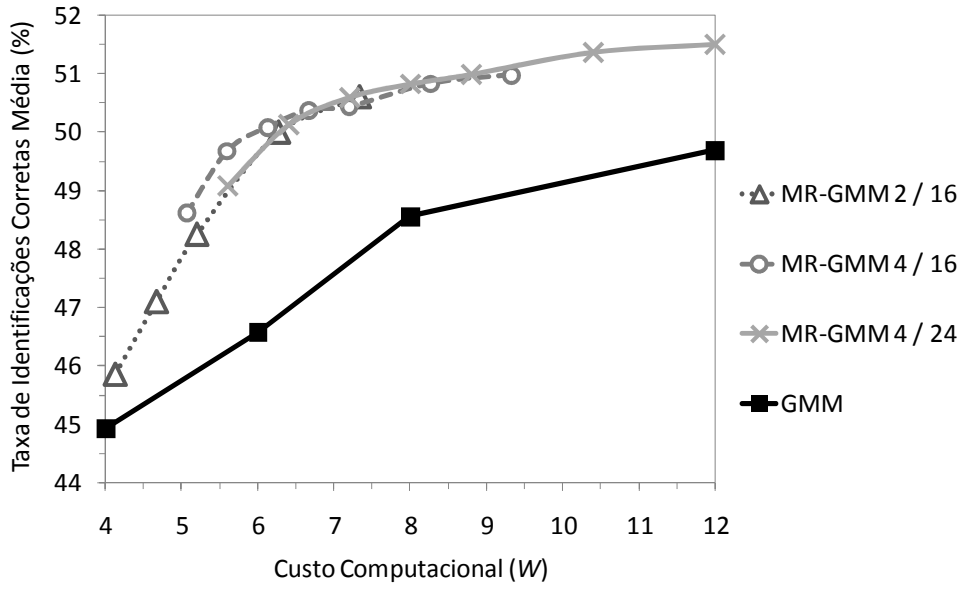

Figura 4.53: Comparação de taxas médias de identificações corretas de sistemas GMM e MR-GMM, em função do custo computacional, para frequência de amostragem 8 kHz, quantização de 8 bits μ-law.

Nos experimentos com ruído, observa-se a superioridade dos MR-GMM em todas as situações simuladas. O desempenho obtido com o GMM 16, com custo $W = 16$, é praticamente equiparado ao desempenho do MR-GMM 4 / 16 com C_1/S=0,33 ($W = 9{,}33$), promovendo redução de 42% no custo; e é equiparado ao desempenho do MR-GMM 4 / 24 com C_1/S=0,20 ($W = 8{,}8$), com redução de custo de 45%. O MR-GMM 4 / 24 com C_1/S=0,33 (W=12,0), por sua vez, tem desempenho médio 0,4% superior ao GMM 16 e ainda possibilita uma redução de 25% no custo computacional.

Para permitir uma análise mais detalhada do desempenho dos sistemas com modelos MR-GMM, foram traçadas, no gráfico da Figura 4.54, as diferenças entre a taxa de identificações corretas do GMM 16 e dos modelos MR-GMM citados no parágrafo anterior, para cada condição de áudio ruidoso testada.

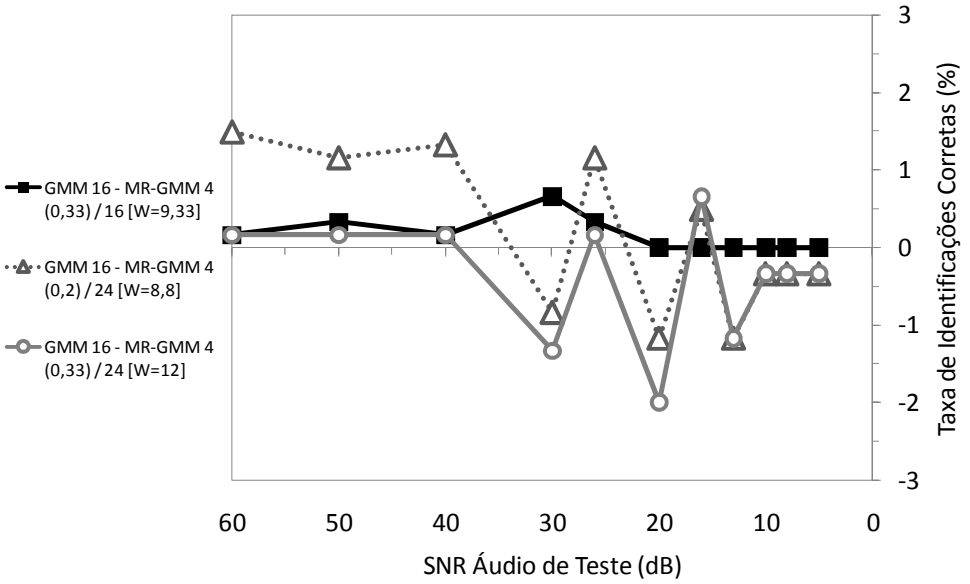

Figura 4.54: Diferença nas taxas de identificações corretas entre sistemas GMM e MR-GMM, em função do nível de ruído do áudio de teste, para frequência de amostragem 8 kHz, μ-law.

Assim como verificado nas simulações com frequência de amostragem de 22 kHz, observa-se que o modelo MR-GMM 4 / 16 com C_1/S=0,27 (W=8,27), tem desempenho muito semelhante ao do GMM 16 para todas as condições de ruído testadas, o que ocorre porque esse modelo MR-GMM utiliza, na fase final da identificação, esse mesmo modelo GMM 16. O modelo MR-GMM 4 / 24 com C_1/S=0,13 (W=7,20), por outro lado, apresenta diferenças de desempenho, ora positivas ora negativas, embora essas diferenças não sejam muito acentuadas (limitadas a 2,3) e, no computo da média, o desempenho global seja equivalente ao do GMM 16. Com relação ao modelo MR-GMM 4 / 24 com C_1/S=0,33 (W=12,0), verifica-se que, embora promova ganho de desempenho em boa parte das condições testadas, esse ganho é significativamente menor que o observado no caso da frequência de amostragem de 22 kHz.

A análise geral dos resultados permite concluir que a modelagem MR-GMM proposta é uma técnica válida para a redução do custo computacional de sistemas de ASR, sendo que os ganhos obtidos são maiores para situações de áudio com frequências de amostragem mais elevada e quantização de 16 bits.

4.5.2.1 Resultados com a combinação MMA/MGD/MR-GMM

O método dos MR-GMM, por explorar características da modelagem diversas das demais técnicas apresentadas nesta tese, pode ser utilizado em conjunto com qualquer um deles. Em particular, a combinação que se mostra mais interessantes é a tripla associação entre o MMA, o MGD e MR-GMM. Nesse caso, se estará utilizando simultaneamente, no sentido de reduzir o custo computacional total de uma tarefa de identificação, uma técnica que explora a coerência temporal entre as condições de ruído de janelas subsequentes do áudio questionado (MMA), uma técnica que explora a coerência entre os GMM componentes dos modelos multicondicionais (MGD), e uma técnica que explora a possibilidade de eliminação de locutores com custo mais baixo que o necessário para a identificação final (MR-GMM). A escolha dessa associação é decorrência dos bons resultados alcançados com a associação MMA/MGD, como demonstrado na seção 4.4.3.1.

Nesse sentido, foram construídos dois sistemas de RAL utilizando o MR-GMM 4 / 16; um com C_1/S=0,13 e outro com C_1/S=0,27. Esses parâmetros dos MR-GMM foram escolhidos pela observação dos resultados da utilização isolada dessa técnica. Esses MR-GMM foram associados com a técnica MMA, $a = 1$, e com o MGD. Para simplificar a notação, esses sistemas MR-GMM/MMA/MGD serão denominados de MR-GMM_1/MMA/MGD e de MR-GMM_2/MMA/MGD, respectivamente.

No caso da associação do MR-GMM com MGD, é necessário ainda destacar que, como o MR-GMM realizada a identificação em mais de uma etapa, com modelos de números de componentes crescentes, é razoável utilizar valores de g diferentes para cada uma dessas etapas[*]. Isso porque a utilização do MGD com $g = 2$ (definido como a melhor situação nos experimentos da seção 4.4.3) sobre um modelo GMM de dezesseis componentes significa que se estará utilizando 12,5% das componentes do modelo. Contudo, utilizar essa mesma configuração sobre um GMM de quatro componentes significa tomar 50% das componentes do modelo. Por essa razão, optou-se por utilizar, na primeira etapa de todos os MR-GMM, o parâmetro $g_1 = 1$, o que ainda leva à utilização de 25% das componentes do modelo. Na segunda etapa do MR-GMM, manteve-se o uso do MMA($a = 1$)/MGD($g = 2$).

Os custos de identificação associados aos modelos MR-GMM/MMA/MGD são dados por:

[*] Também no caso da associação MR-GMM/MMA pode-se optar por utilizar valores de a distintos em cada uma das etapas do MR-GMM, embora isso não faça tanto sentido como no caso da associação com o MGD.

$$W_{MR-GMM/MMA/MGD} \propto M_1^* + M_2^* \frac{C_1}{S} = \left(M_1 + g_1 \cdot 2a_1\right) + \left(M_2 + g_2 \cdot 2a_2\right)\left(\frac{C_1}{S}\right) \quad (4.45)$$

Sendo que M_1^* e M_2^* representam os custos equivalentes dos modelos da primeira e da segunda etapas da identificação do MR-GMM, respectivamente. Como, em cada uma dessas etapas, se utiliza um sistema MMA/MGD, seu custo equivalente é dado pela expressão de (4.28).

Dessa maneira, dadas as escolhas de parâmetros anteriormente descritas, os custos dos sistemas MR-GMM/MMA/MGD a serem utilizados são de:

$$\begin{aligned} W_1 &\propto \left(4 + 1 \cdot 2\right) + \left(16 + 2 \cdot 2\right)\left(0,13\right) = 8,6 \\ W_2 &\propto \left(4 + 1 \cdot 2\right) + \left(16 + 2 \cdot 2\right)\left(0,27\right) = 11,4 \end{aligned} \quad (4.46)$$

É importante destacar que os modelos multicondicionais com essas configurações apresenta custo computacional inferior ao de um único GMM, que tem $W \propto 16$.

A fim de observar o desempenho dos sistemas de RAL utilizando a combinação das três técnicas introduzidas neste trabalho, foram realizados procedimentos de identificação com diversos níveis de SNR no áudio questionado. Os resultados desses procedimentos para cada um dos modelos MR-GMM/MMA/MGD utilizados estão expostos na Tabela 4.12. Nessa mesma tabela, para efeito de comparações, foram incluídos os dados referentes ao sistema MMA(a = 1)/MGD(g = 2) e ao MCM.

Tabela 4.12: Taxas de identificações corretas de sistemas de RAL multicondicional utilizando MR-GMM/MMA(a = 1)/MGD(g_1 = 1,g_2 = 2).

Taxas de Identificações Corretas (%)	**SNR do Áudio de Teste (dB)**										
	60	**50**	**40**	**30**	**26**	**20**	**16**	**13**	**10**	**8**	**5**
MR-GMM$_1$/MMA/MGD	100,0	100,0	99,2	95,8	95,5	93,3	91,7	90,0	88,0	86,5	79,5
MR-GMM$_2$/MMA/MGD	100,0	100,0	99,5	98,0	96,5	93,7	93,0	91,2	90,0	87,3	81,3
MMA(a=1)/MGD(g=2)	100,0	100,0	99,6	98,5	97,7	95,0	93,0	92,7	89,0	88,8	81,5
MCM	97,8	98,2	98,3	97,8	97,2	95,5	94,2	91,5	86,6	80,8	68,5

Para permitir uma melhor visualização dos resultados, os valores das taxas médias de identificações corretas para cada sistema foram traçados no gráfico da Figura 4.55. Nessa figura, também foram incluídos os custos computacionais, *W*, correspondentes.

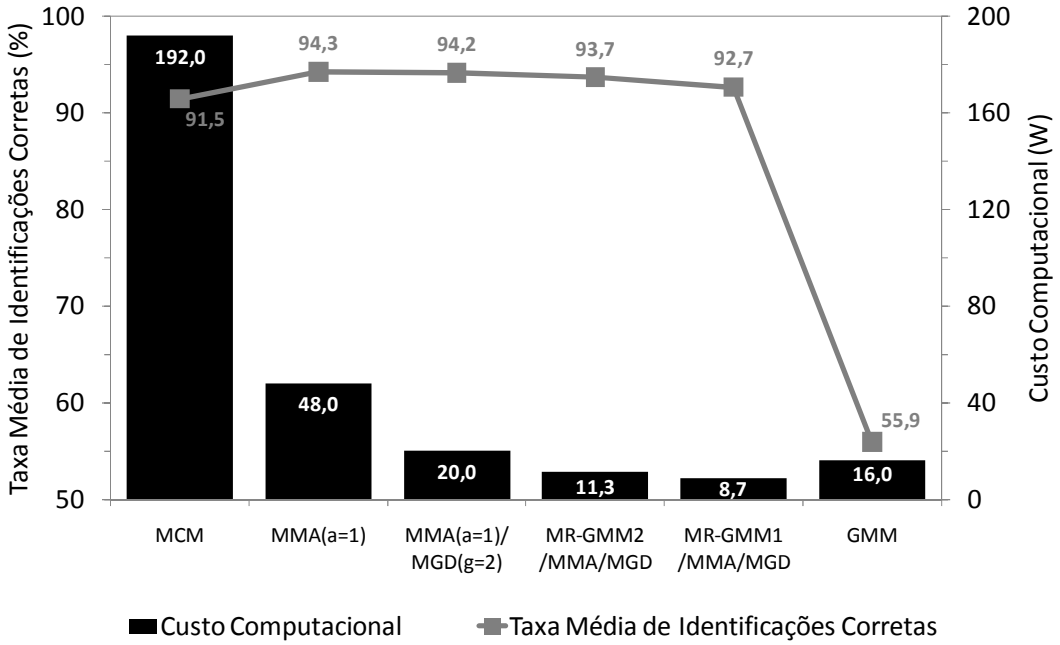

Figura 4.55: Comparação de taxas médias de identificações corretas e de custos computacionais para sistemas MR-GMM/MMA/MGD e outros.

Como se verifica, a utilização do MR-GMM em conjunto com o MMA/MGD permitiu uma redução da ordem de 50% sobre o custo computacional do sistema usando apenas MMA/MGD, totalizando uma redução de aproximadamente 94% sobre o custo inicial do MCM. Também se constata que o uso do MR-GMM em conjunto com o MMA/MGD provocou alguma perda de desempenho de identificação em comparação com os sistemas utilizando apenas MMA/MGD (entre 0,5% e 1,5%, a depender do modelo), embora, mesmo com essa perda, o desempenho ainda tenha se mantido acima do observado com os modelos MCM tradicionais.

5 CONCLUSÃO

No presente trabalho, foram propostas, implementadas, avaliadas e comparadas quatro novas técnicas para a redução do custo computacional de tarefas de identificação em sistemas de Reconhecimento Automático de Locutor (RAL) baseados em Modelos de Mistura de Gaussianas (GMM) multicondicionais: o Método da Condição Persistente (e suas variantes), os Modelos Multicondicionais Adaptativos, o Método das Gaussianas Dominantes e os Modelos de Mistura de Gaussianas Multirresolução. Essas quatro novas técnicas, as principais contribuições desta tese, permitem o desenvolvimento de sistemas de RAL que exploram todo o potencial da modelagem multicondicional, como sua robustez a variações no nível de ruído do áudio questionado, mas que demandam um esforço computacional semelhante ao de um sistema baseado em um GMM unicondicional.

O Método da Condição Persistente (MCP), proposto na seção 4.2, que se baseia na estimação da condição máxima de treinamento para novas janelas de análise a partir da informação de janelas passadas ou futuras, permitiu a redução do custo computacional em mais de 70%, sem afetar a taxa média de identificações corretas do sistema, principalmente quando utilizado na sua forma derivada, o método da condição persistente linearmente variada da seção 4.2.1.

A técnica dos Modelos Multicondicionais Adaptativos (MMA), introduzidos na seção 4.3, que também explora a coerência temporal da condição do áudio, mas que consegue a redução do custo computacional pela restrição da variação temporal das condições de treinamento utilizadas no modelo multicondicional, demonstrou ser possível reduzir em até 75% o custo computacional da identificação. Esse método, além disso, promoveu um aumento sistemático nas taxas de identificação corretas para todas as condições de ruído testadas, demonstrando, dessa forma, ser superior ao MCP ou às suas variantes.

Foi também apresentado, nesta tese, na seção 4.4, o Método das Gaussianas Dominantes (MGD), que busca reduzir a complexidade computacional pela exploração da coerência entre as componentes correspondentes dos diferentes GMM que compõem o modelo multicondicional. Na realidade, os modelos multicondicionais tradicionalmente empregados não têm qualquer coerência entre as componentes correspondentes dos GMM que o compõem, mas, como demonstrado, o uso do Método do Treinamento Progressivo (MTP) per-

mite a construção de modelos multicondicionais com essa propriedade sem, contudo, afetar o desempenho do modelo.

O MGD provou possibilitar uma redução de aproximadamente 80% no esforço computacional das identificações sem afetar a taxa média de identificações corretas do sistema. E, mais interessante que isso, o MGD pode ser diretamente aplicado em conjunto com o MMA, de forma a possibilitar a redução total de até 90% no esforço computacional em comparação com a modelagem tradicionalmente empregada (Método da Condição Máxima – MCM). A aplicação conjunta MMA/MGD ainda possibilita a obtenção de aumentos superiores a 2% na taxa média de identificações corretas.

Na seção 4.5, foi introduzido o Método das Gaussianas Multirresolução (MR-GMM), que obtém uma redução no esforço computacional da identificação pelo descarte de parte dos modelos de locutores numa classificação preliminar, realizada com modelos de baixa complexidade. O uso isolado do MR-GMM permite reduções da ordem de 50% nesse esforço sem prejudicar o desempenho do sistema de identificação, o que não é tão expressivo quanto o possibilitado pelos outros métodos. Entretanto, esse método também pode ser aplicado em conjunto tanto com o MMA como com o MGD, e mesmo com ambos simultaneamente. Dessa maneira, utilizando sistemas MR-GMM/MMA/MGD, pode-se chegar a reduções globais de até 95% no custo computacional, mantendo ainda a taxa média de identificações corretas mais de 1% acima da obtida com a modelagem MCM.

É relevante destacar que os sistemas MR-GMM/MMA/MGD propostos, desenvolvidos e validados durante esta tese, têm custo computacional inferior ao de um único GMM. Desse modo, com a aplicação dessas técnicas combinadas, o uso da modelagem multicondicional (que possibilita grande robustez aos sistemas de identificação) torna-se viável mesmo em aplicações que requerem baixos tempos de resposta para pesquisas em bancos de dados com elevado número de locutores.

Além dessas contribuições, diretamente relacionadas ao objetivo principal da tese, foram obtidos outros resultados relevantes.

Na seção 2.3, foram reavaliados diferentes parâmetros representativos da voz, sendo reafirmada a superioridade dos parâmetros cepstrais e mel-cepstrais. Esse estudo foi desenvolvido em vista dos resultados publicados por Souza e Souza (2001), que sugeriam a superioridade de parâmetros relacionados à biometria do trato vocal para a representação dos locutores em sistemas de identificação automática.

Na seção 3.3, foram avaliadas diferentes propostas de modelagem multicondicional, inclusive a recentemente proposta por Ming *et al* (2007), tendo-se concluído pela superioridade do Método da Condição Máxima (MCM) tradicionalmente empregado nos sistemas multicondicionais (Matsui, Kanno e Furui, 1996; Yang e Gong, 2006). Na seção 3.3.2.1, demonstrou-se ainda que é possível melhorar significativamente o desempenho dos modelos multicondicionais pela agregação de mais modelos GMM, treinados com condição de ruído mais intensa que a máxima condição a ser utilizada nas identificações. Esse aumento no número de condições de treinamento do modelo multicondicional, certamente, provoca um aumento no custo computacional da identificação, contudo, os Modelos Multicondicionais Adaptativos (MMA) propostos nesta tese, permitem eliminar esse efeito colateral, uma vez que o MMA possibilita obter um custo fixo, independentemente do número de condições de treinamento do modelo multicondicional, como explanado na seção 4.3.

Como sugestões de continuidade deste trabalho, visto que as limitações de tempo disponível restringiram os caminhos a serem percorridos, há alguns pontos que podem ser mais profundamente explorados.

Primeiramente, sugere-se a construção de um sistema de RAL baseado em MR-GMM (com ou sem combinação com outros métodos) que utilize, na segunda etapa de identificação (e nas subsequentes, se houver), a informação de condição de treinamento máxima para cada janela de análise processada com os modelos de baixa resolução. Dessa maneira, apenas a primeira etapa de identificação exigiria o cálculo da verossimilhança para mais de uma condição de treinamento. Nas etapas subsequentes, seria reutilizada a condição máxima determinada na primeira etapa. Esse procedimento pode promover reduções de custo computacional significativas para sistemas de RAL que não utilizem o MMA ou o MGD, cerca 50%[*]. Para sistemas que, além do MR-GMM, utilizam a combinação MMA/MGD, o potencial de redução de custo computacional é de aproximadamente 10%[†].

Em princípio, essa técnica, provisoriamente denominada de Reutilização da Condição Máxima (RCM), não deve afetar o desempenho do sistema de identificação, pois, embora não se recalcule a condição máxima na segunda etapa da identificação do MR-GMM, essa condição foi calculada na primeira etapa. Apesar de serem utilizados modelos mais simples na primeira etapa do MR-GMM, é de se esperar que a condição que promove a máxima

[*] Considerando um sistema MR-GMM de duas etapas com $M_1 = 4$, $M_2 = 16$ e $C_1/S = 0{,}25$.
[†] Considerando um sistema MR-GMM como descrito na nota anterior e mais a aplicação do MMA com $a = 1$ e do MGD com $g = 1$, na primeira etapa do MR-GMM, e com $g = 2$, na segunda etapa.

verossimilhança entre os modelos multicondicionais de baixa resolução seja a mesma que a produz para os modelos de alta resolução. Como, nas duas etapas, os modelos comparados entre si para a determinação da condição máxima são todos de um mesmo locutor e todos de uma mesma resolução, o simples incremento no número de componentes gaussianas dos modelos não deve, em tese, modificar a condição de treinamento que promove a máxima verossimilhança dos modelos.

Outra sugestão para reduzir ainda mais o custo computacional da identificação com uma variação do método MGD é a utilização de sistemas que utilizem a característica de coerência temporal dos parâmetros da voz no nível de componentes dos modelos GMM que compõem o modelo multicondicional. Como, no MGD, somente são computadas para a determinação da verossimilhança do modelo as g componentes gaussianas mais expressivas, pode-se tentar, após calcular todas as componentes gaussianas para uma determinada janela de analise, t, estimar, para a janela $t + 1$, as componentes $g´$ menos significativas. Implicitamente, o que essa técnica admite é que a variação do áudio entre duas janelas de análise sucessivas não pode ser tão pronunciada a ponto de que uma componente gaussiana que é das menos significativas para a verossimilhança do modelo na janela t se torne uma das mais significativas para a janela $t + 1$. Sobre essa idéia, podem ser experimentadas variações nos valores de g' e também na quantidade de janelas para quais serão estimadas as componentes g' (estendendo o processo para até a janela $t + \Delta$, não apenas para a janela $t + 1$). Também podem ser experimentadas variações calculando inicialmente todas as componentes do modelo para as janelas t e $t + \Pi$, $\Pi > 1$, e interpolando linearmente as janelas intermediárias. Essa técnica, provisoriamente denominada de Método das Gaussianas Dominantes Persistentes (MGDP), pode levar a reduções de aproximadamente 75%[*] no esforço computacional, para sistemas de RAL utilizando apenas o MGD. Para sistemas que usam a combinação MR-GMM/MMA/MGD, proposta e validade nesta tese, a adoção do MGD-P permite reduções estimados em cerca de 20%[†].

Outra linha de pesquisa que merece ser explorada é a utilização ou a adequação das técnicas propostas neste trabalho aos sistemas de RAL multicondicional sub-bandas recentemente introduzidos por Ming *et al* (2007). Esses sistemas, por sua característica de separa-

[*] Estimativa realizada tomando $g' = M - 2g$, para $M = 16$ e $g = 2$ (como utilizado no MGD deste trabalho), e recalculando o modelo completo a cada duas janelas de análise, ou seja, na janela t são calculadas todas as componentes gaussianas do GMM e na janela $t + 1$ são calculadas apenas as $2g$ componentes mais significativas da janela t. Optou-se por estimar a necessidade do uso de $2g$ componentes para deixar uma "folga" para incluir também as componentes razoavelmente significativas da janela t.

[†] Utilizando os mesmos parâmetros da nota anterior nas duas etapas de um sistema MR-GMM com $M_1 = 4$, $M_2 = 16$ e $C_1/S = 0{,}15$.

ção em sub-bandas da modelagem multicondicional, tornam-se mais aptos a lidar com situações de ruídos com perfis de frequência diferentes daqueles utilizados no treinamento, por exemplo, nos casos de ruídos com banda limitada (visto que o treinamento, em geral, é realizado com áudio corrompido por ruído branco). Destaque-se que essa alteração, o processamento em sub-bandas, torna esse tipo de sistema ainda mais exigente computacionalmente que os sistemas multicondicionais tradicionais, de forma que, para sua utilização prática, é ainda mais relevante o uso de técnicas capazes de eliminar os cálculos desnecessários e de, consequentemente, reduzir o custo total da identificação.

REFERÊNCIAS

Aronowitz, H. (2006). "Speaker recognition using dynamic time warp template spotting", *Journal of the Acoustical Society of America*, vol.120(5), p.2415

Atal, B.S. (1972). "Automatic Speaker Recognition Based on Pitch Contours", *Journal Acoustic Society of America*, vol.52, pp. 1687-1697.

Atal, B.S. (1974). "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification", *Journal Acoustic Society of America*, vol.55, no.6, pp. 1304-1312.

Atal, B.S. (1976). "Automatic Recognition of Speakers From Their Voices", *Proceedings IEEE*, vol.64, no.4, pp. 460-475.

Bricker, P. D.; Pruzansky, S. (1966). "Effects of Stimulus Content and Duration on Talker Identification," *Journal of the Acoustical Society of America*, vol. 40, pp.1441-1450.

Brookes, M. (2001). "Voicebox: Speech Processing Toolbox for Matlab", *Dept. Electrical & Electronic Eng., Imperial College, London*.

Buck, J.; Burton, D.; Shore, J. (1985). "Text-dependent speaker recognition using vector quantization", *IEEE International Conference on ICASSP 85*, vol.10, pp.391-394.

Bultheel, A. (1984). "Recursive relations for block Hankel and Toeplitz systems, Part II: Dual recursions". *Journal of Computer and Applied Math.*, vol.10, pp. 329-354.

Campbell Jr., J.P. (1997). "Speaker Recognition: A Tutorial", *Proceedings of the IEEE*, vol.85, no.9, pp. 1437-1462.

Champod, C. (1995). "Edmond Locard - Numerical Standards & 'Probable' Identifications", *Journal of Forensic Identification*, vol.45(2), pp.136-163.

Childers, D.G.; Skinner, D.P.; Kemerait, R.C. Kemerait (1977)., "The Cepstrum: A Guide to Processing", *Proceedings IEEE*, vol.65, no.10, pp. 1428-1443.

D'Almeida, F.Q.; Nascimento, F.A.O. (2006). "Comparação de Desempenho de Parâmetros da Fala em Sistemas de Reconhecimento Automático de Locutor", *Congresso Brasileiro de Automática – CBA 2006*.

D'Almeida, F.Q.; Nascimento, F.A.O, Berger, P.A.; da Silva, L. M. (2007). "Efeitos da Codificação MP3 em Sistemas de Reconhecimento Automático de Locutor via GMM", *Anais do XXV Simpósio Brasileiro de Telecomunicações SBrT 2007.*

D'Almeida, F.Q.; Nascimento, F.A.O, Berger, P.A.; da Silva, L. M. (2008). "Noise Robust Speaker Recognition using Reduced Multiconditional Gaussian Mixture Models", *International Journal of Forensic Computer Science*, vol.º3, no.º1, pp. 60-69.

D'Almeida, F.Q.; Nascimento, F.A.O, Berger, P.A.; da Silva, L. M. (2008). "Reconhecimento Automático de Locutor com Modelos de Mistura de Gaussianas Multirresolução (MR-GMM)", *Anais do Congresso Brasileiro de Automática CBA 2008*.

D'Almeida, F.Q.; Nascimento, F.A.O, Berger, P.A.; da Silva, L. M. (2008b). "Robustez ao Ruído em Sistemas de Reconhecimento Automático de Locutor com Modelos de Mistura de Gaussianas Multirresolução (MR-GMM)", *Anais do Congresso Brasileiro de Automática CBA 2008*.

Dempster, A.; Laird, N.; Rubin, D. (1977). "Maximum Likelihood From Incomplete Data Via the EM Algorithm", *Journal Royal Statistical Society*, vol.39, pp. 1-38.

Doddington, G.R. (1970). "A Method for Speaker Verification", *Ph.D. Thesis, University of Wisconsin, Madison*.

Doddington, G.R.; Flanagan, J.L.; Lummis, R.C. (1972). "Automatic Speaker Verification by Nonlinear Alignment of Acoustic Parameters", *U.S. Patent 3,700,815*.

Eriksson, A. (2005). "Tutorial on Forensic Speech Science", *Proceedings Interspeech*, Lisbon, Portugal.

Furui, S.; Itakura, F.; Saito, S. (1972). "Talker recognition by Long-Time Averaged Speech Spectrum", *Electronics Communications of Japan*, vol.55A, pp.54-61.

Furui, S. (1974). "An Analysis of Long-Term Variation of Feature Parameters of Speech and its Application to Talker Recognition", *Electronics Communications*, vol.57-A, pp.34-42.

Furui, S. (1981). “Cepstral analysis technique for automatic speaker verification”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol.29(2), pp.254-272.

Hollien, H.F. (2002). “Forensic Voice Identification”, *Academic Press*

Grey, G.; Kopp, G.A. (1944). “Voiceprint Identification”, *Bell Telephone Laboratories Report*, pp.1-14.

Kersta, L.G. (1962). “Voiceprint Identification”, *Nature*, 196, pp. 1253-1257.

Kopp, G.A.; Green, H.C. (1946). “Basic Phonetic Principles of Visible Speech”, *Journal of the Acoustical Society of America*, vol.18, pp.74-89.

Lindh, J. (2004). “Handling the ‘Voiceprint’ Issue”, *Proceedings FONETIK 2004*, Dept. of Linguistics, Stockholm University.

Luck, J.E. (1969). “Automatic Speaker Verification Using Cepstral Measurement”, *Journal Acoustic Society of America*, vol.46, pp.1026-1032.

Lummis, R. C. (1973). “Speaker Verification by Computer Using Speech Intensity for Temporal Registration”, IEEE Trans. Audio Electroacoust., vol.AU-21, pp.80-89.

Markel, J.D.; Oshika, B.; Gray Jr., A. (1977). “Long-term feature averaging for speaker recognition”, *IEEE Trans. On Acoust., Speech and Signal Proc., vol.25(4), pp.330-337.*

Markel, J.D.; Davis, S.B. (1979). “Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol.ASSP-27, no.1, pp.74-82.

Matsui, T.; Furui, S. (1992). “Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs”, *Proc. Intl. Conf. Acoustics, Speech, and Signal Processing*, vol.2, pp.157-160.

Matsui, T.; Kanno, T.; Furui, S. (1996). “Speaker recognition using HMM composition in noisy environments”, *Computer Speech and Language*, vol.10, pp.107-116.

McGehee, F. (1937). “The reliability of the identification of human voice”, *J.Gen. Psychol.*, vol.17, pp.249-271.

Ming, J.; Stewart, D.; Vaseghi, S. (2005). “Speaker identification in unknown noisy conditions - A universal compensation approach,” *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol.1, pp. 617–620.

Ming, J.; Hazen, T.; Glass, J.R.; Reynolds, D.A. (2007). “Robust Speaker Recognition in Noisy Conditions”, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.15, no.5, pp.1711-1723.

Park, A.; Hazen, T.J. (2002). “ASR dependent techniques for speaker identification”, *Proc. of Int. Conf. on Spoken Language Processing*, pp.1337-1340.

Poritz, A. (1982). “Linear predictive hidden Markov models and the speech signal”, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol.7, pp.1291-1294.

Reynolds, D.A. (1992)., “A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification”, *Ph. D. Thesis, Georgia Institute of Technology, Department of Electrical Engineering.*

Reynolds, D.A.; Rose, R.C. (1995). “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, *IEEE Transactions on Speech and Audio Processing*, vol.3, no.1, pp 72-83.

Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B. (2000). “Speaker Verification Using Adapted Gaussian Mixture Models”, *Digital Signal Processing*, vol.10, nos. 1-3, pp. 19-41.

Rosenberg, A.E. (1976). "Automatic speaker verification: a review", *Proceedings of the IEEE*, vol.64, no.4: 475-486.

Rosenberg, A.E.; Lee, C.-H.; Soong, F.K. (1990). “Sub-word unit talker verification using hidden Markov models”, *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol.1, pp.269-272.

Sambur, M.R. (1978). “Selection of Acoustic Features for Speaker Identification”, *IEEE Transactions on Accoustic, Speech and Signal Processing*, vol.23, no.2, pp. 176-182.

Shore, J.E.; Burton, D.K. (1983). “Discrete Utterance Speech Recognition Without Time Alignment”, *IEEE Trans. on Inform. Theory*, vol.IT-24, no.4, pp.473-491.

Soong, F.K.; Rosenberg, A.E.; Rabiner, L.R.; Juang, B.H. (1985). "A Vector Quantization Approach to Speaker Recognition", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp.387-390.

Souza, A.F.; Souza, M.N. (2001). "Comparative Analysis of Speech Parameters for the Design of Speaker Verification Systems", *Proceedings of the 23rd Annual EMBS International Conference*, Istanbul, Turkey.

Stevens, K. N.; Williams, C.E.; Carbonell, J.R.; Woods, B. (1968). "Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material" *Journal of the Acoustical Society of America*, vol.44, pp.1596-1607.

Tierney, J. (1980). "A Study of LPC Analysis of Speech in Additive Noise", *IEEE Transactions on Accoustic, Speech and Signal Processing*, vol.ASSP-28, pp. 389-397.

Tisby, N.Z (1991). "On the application of mixture AR hidden Markov models to text independent speaker recognition", *IEEE Transactions on Signal Processing*, vol.39(3), pp.563-570.

Tosi, O.; Oyer H.; *et al* (1972). "Experiment on Voice Identification" *Journal of the Acoustical Society of America*, vol.51, pp.: 2030-2043.

Brasil, Tribunal Regional Federal da 4ª Região (2006). "Apelação em Mandado de Segurança 2005.72.00.013027-1", *Diário da Justiça*, data 25/11/2006, p.1080.

Vanderslice, R.; Ladefoged, P (1967). "The Voiceprint Mystique", *UCLA Working Papers in Phonetics*, 7: 126-142.

Webb, J.J.; Rissanen, E.L. (1993). "Speaker identification experiments using HMMs", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol.2, pp.387-390.

Wolf, J.J. (1972). "Efficient Acoustic Parameters for Speaker Recognition", *Jounal Acoustic Society of America*, vol.51, pp. 2044-2056.

Xu,H.; Tan, Z-H.; Dalsgaard, P.; Lindberg, B. (2005). "Robust Speech Recognition Based on Noise and SNR Classification - a Multiple-Model Framework", *Proceeding of the 9th European Conference on Speech Communication and Technology – Interspeech 2005*.

Yang, L.; Gong, W. (2006). "Multi-SNR GMMs-Based Noise-Robust Speaker Verification Using 1/fa Noises", *Proc.* 18th *International Conference on Pattern Recognition - ICPR 2006*, vol.4, pp.241-244.

Young, M.A.; Campbell, R.A. (1967). "Effects of Context on Talker Identification" *Journal of the Acoustical Society of America*, vol.42, pp.1250-1254