Denis Pirttiaho Cardoso

*Identificação de locutor usando modelos de
mistura de gaussianas*

São Paulo

2009

Denis Pirttiahó Cardoso

*Identificação de locutor usando modelos de
mistura de gaussianas*

Dissertação apresentada à Escola Politécnica
da Universidade de São Paulo para obtenção
do título de Mestre em Engenharia.

Área de concentração:
Sistemas Eletrônicos

Orientador: Prof. Livre-Docente
Miguel Arjona Ramírez

São Paulo

2009

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 4 de maio de 2009

Assinatura do autor _____

Assinatura do orientador _____

FICHA CATALOGRÁFICA

Cardoso, Denis Pirttiah

Identificação de locutor empregando modelos de mistura de gaussianas / D.P. Cardoso. – São Paulo, 2009.

86p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1. Processamento de sinais acústicos 2. Processamento digital de voz 3. Reconhecimento de voz I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Eletrônica II.t

Dedico este trabalho à minha esposa Juliana por
suas palavras de motivação ao longo desta jornada.

Agradecimento

Ao meu orientador Prof. Livre-Docente Miguel Arjona Ramírez por suas sugestões e esclarecimentos que representaram iniciativas essenciais para a elaboração deste trabalho.

Resumo

A identificação de locutor está relacionada com a seleção de um locutor dentro de um conjunto de membros pré-definidos e neste trabalho os experimentos foram realizados utilizando um sistema de identificação de locutor independente de texto baseado em modelos de mistura de gaussianas. Para realizar os testes, foi empregado o banco de voz TIMIT e sua correspondente versão corrompida por ruído de canal telefônico, isto é, NTIMIT. O aparelho fonador pode ser representado por coeficientes mel-cepstrais obtidos por meio de banco de filtros ou, alternativamente, por coeficientes de predição linear. Adicionalmente, a técnica de subtração da média cepstral é aplicada quando o banco de voz NITMIT é utilizado com o intuito de minimizar a distorção de canal intrínseca a ele. A componente da locução para a qual os coeficientes mel-cepstrais são calculados é obtida através de um detector de atividade de voz (DAV). No entanto, os DAVs são em geral sensíveis à relação sinal-ruído da locução, sendo necessário adaptá-los para as condições de operação do sistema. É sugerida a integração no DAV de um estimador da relação sinal-ruído baseado no método *Minima Controlled Recursive Average* (MCRA), que é necessário para permitir o tratamento de sinais tanto limpos quanto ruidosos. É observado que em locuções de elevada relação sinal-ruído, como aquelas provenientes do banco de voz TIMIT, o método mais apropriado de extração dos coeficientes mel-cepstrais foi o padrão, isto é, baseado em banco de filtros, enquanto que para sinais de voz ruidosos a técnica de subtração da média cepstral aliada à extração dos coeficientes mel-cepstrais a partir de coeficientes de predição linear revelou os melhores resultados.

Abstract

Speaker identification is concerned with the selection of one speaker within a set of enrolled members and in this work the experiments were performed using a text-independent cohort Gaussian mixture model (GMM) speaker identification system. In order to perform the tests, TIMIT speech database is used and its corresponding version corrupted by a noisy telephone channel, i.e., NTIMIT. The vocal tract is represented by Mel-cepstral frequency coefficients with filter banks or, alternatively, by linear prediction cepstral coefficients. Additionally, the cepstral mean subtraction technique is applied when the NTIMIT database is used to minimize the channel distortion intrinsic to it. The utterance component for which the Mel-frequency cepstral coefficients is obtained using a voice activity detector (VAD). However, the VADs are generally sensitive to the signal-to-noise ratio of the utterance, making it necessary to adapt them to the system operating conditions. A signal-to-noise ratio estimator is included in the proposal VAD, which is based on Minima Controlled Recursive Average (MCRA), in order to be able to handle both clean and noisy speech. It is observed that in high signal-to-noise ratio utterances, such as those from the TIMIT database, the more appropriate extraction method for the Mel-frequency cepstral coefficients was the baseline one consisting of filter banks, while for noisy speech the technique of cepstral mean subtraction coupled with the extraction of Mel-frequency cepstral coefficients from linear prediction cepstral coefficients provided the best results.

Lista de Figuras

1	Componente sonora do sinal de voz.	p. 16
2	Componente surda do sinal de voz.	p. 17
3	Ganho do filtro de pré-ênfase.	p. 18
4	Segmentação do sinal de voz.	p. 19
5	Módulo do espectro do sinal de voz pré-enfatizado convoluído pelo espectro de uma janela retangular.	p. 20
6	Módulo do espectro do sinal de voz pré-enfatizado convoluído pelo espectro de uma janela de Hamming.	p. 21
7	Diagrama de cruzamento por zero para locução com elevada relação sinal-ruído.	p. 22
8	Diagrama de cruzamento por zero para locução com baixa relação sinal-ruído.	p. 22
9	Diagrama de energia para locução com elevada relação sinal-ruído.	p. 23
10	Diagrama de energia para locução com baixa relação sinal-ruído.	p. 24
11	Espectro de potência de um sinal de voz e sua versão suavizada.	p. 26
12	Algoritmo para encontrar S_{min}	p. 27
13	Módulo do espectro de vários quadros de sinal de voz pré-enfatizado.	p. 29
14	Valores de $S_m(l)$ para determinado sinal limpo e quando transmitido por linha telefônica.	p. 30
15	Sintetizador do sinal de voz baseado no modelo LPC.	p. 31
16	Espectro do sinal de voz e do filtro $H(z)$ com 10 coeficientes de predição linear.	p. 36
17	Espectro do sinal de voz e do filtro $H(z)$ com 16 coeficientes de predição linear.	p. 36

18	Espectro do sinal de voz e do filtro $H(z)$ com 128 coeficientes de predição linear.	p. 37
19	Cepstro da componente sonora do sinal de voz.	p. 40
20	Cepstro do ruído de fundo.	p. 41
21	Cepstro da componente surda do sinal de voz.	p. 41
22	Mapeamento de frequência da escala mel em linear.	p. 45
23	Pré-processamento do sinal de voz.	p. 46
24	Coeficientes mel-cepstrais obtidos a partir de banco de filtros.	p. 46
25	Espectro do banco de filtros triangulares de Slaney.	p. 48
26	Cálculo do logaritmo do espectro do sinal de voz obtido do banco de filtros.	p. 49
27	Comparação entre a fase do filtro passa-tudo com a frequência na escala mel.	p. 50
28	Modelo de mistura de gaussianas.	p. 55
29	Função de densidade de probabilidade de mistura de gaussianas.	p. 57
30	Conjuntos formados com a aplicação do algoritmo LBG sobre um conjunto de vetores bidimensionais.	p. 66
31	Sistema de identificação empregando banco de filtros para derivar os MFCCs.	p. 71
32	Sistema de identificação empregando predição linear para derivar os MFCCs.	p. 71
33	Visualização do desempenho do sistema de acordo com o número de locutores e da forma como os MFCCs são obtidos dos bancos de vozes TIMIT e NTIMIT.	p. 78

Lista de Tabelas

1	Níveis de decisão para segregação de componentes de uma locução. . .	p. 24
2	Intervalo de busca para os valores de S_{tmp}	p. 27
3	Estrutura de subdiretórios dos bancos de vozes TIMIT e NTIMIT. . . .	p. 69
4	Constantes adotadas durante as simulações.	p. 72
5	Resultados típicos de desempenho para modelos com 16 gaussianas utilizando amostras dos bancos TIMIT e NTIMIT.	p. 74
6	Simulação com NTIMIT, coeficientes derivados de predição linear com SMC e número variável de componentes.	p. 75
7	Efeito da variação do tempo de treinamento e de identificação para o desempenho do sistema utilizando-se amostras do banco NTIMIT. . . .	p. 76
8	Desempenho do sistema em cada diretório do banco TIMIT com MFCCs derivados de banco de filtros.	p. 77
9	Desempenho do sistema em cada diretório do banco TIMIT com MFCCs derivados de predição linear.	p. 77
10	Efeito no desempenho do sistema de acordo com o número de locutores e da forma como os MFCCs são obtidos dos bancos de vozes TIMIT e NTIMIT.	p. 77
11	Efeito da compatibilidade das locuções utilizadas nos treinamentos e nos testes de identificação com MFCCs derivados de banco de filtros.	p. 78
12	Efeito da compatibilidade das locuções utilizadas nos treinamentos e nos testes de identificação com MFCCs derivados de predição linear.	p. 79
13	Desempenho do sistema de acordo com o número de gaussianas e MFCCs adotados.	p. 79

LISTA DE ABREVIATURAS E SIGLAS

DAV		detector de atividade de voz
DCT		discrete cosine transform
DFT		discrete Fourier transform
EM		expectation maximization
ERR		equal error rate
FFT		fast Fourier transform
GMM		Gaussian mixture model
LPC		linear predictive coding
MCRA		minima controlled recursive averaging
MLE		maximum likelihood estimation
MMG		modelo de mistura de gaussiana
TCD		transformada do cosseno discreta
TFD		transformada de Fourier discreta
SMC		subtração da média cepstral
VAD		voice activity detector

Sumário

1	Introdução	p. 13
1.1	Verificação e identificação de locutor	p. 13
1.2	Dependência e independência de texto	p. 14
1.3	Verificação de desempenho	p. 14
1.4	Aparelho fonador	p. 15
2	Codificação do sinal de voz	p. 16
2.1	Amostragem	p. 17
2.2	Pré-ênfase	p. 18
2.3	Análise em tempo curto	p. 18
2.4	Janelamento	p. 19
2.5	Detector de sinal de voz	p. 21
2.5.1	Detector baseado no número de cruzamentos por zero e na energia	p. 21
2.5.2	Detector baseado no método Minima Controlled Recursive Averaging	p. 24
2.5.3	Parametrização do sinal de voz	p. 30
3	Predição linear do sinal de voz	p. 31
3.1	O modelo LPC do sinal de voz	p. 31
3.2	Estimação dos coeficiente do filtro LPC	p. 32
4	Cepstro do sinal de voz	p. 38
4.1	Estimação dos coeficientes cepstrais	p. 40

4.2	Escala Mel	p. 44
4.3	Estimação dos coeficientes mel-cepstrais	p. 45
4.3.1	Estimação utilizando banco de filtros	p. 46
4.3.2	Estimação utilizando coeficiente de predição linear	p. 49
4.3.3	Subtração da média cepstral	p. 52
5	Sistema de identificação de locutor	p. 54
5.1	Modelo de mistura de gaussianas	p. 54
5.2	Treinamento	p. 57
5.2.1	Algoritmo EM	p. 58
5.2.2	Inicialização	p. 64
5.3	Identificação	p. 66
5.3.1	Identificação por distâncias mínimas	p. 68
6	Avaliação experimental	p. 69
6.1	Considerações iniciais	p. 69
6.2	Procedimentos de simulação	p. 70
6.2.1	Modelo com número fixo de componentes	p. 72
6.2.2	Modelo com número de componentes variável	p. 74
6.2.3	Locuções com tamanho variável	p. 75
6.2.4	Número de locutores variável	p. 76
6.2.5	Alternando o banco de voz no treinamento e identificação	p. 78
6.2.6	Alternando o número de coeficientes mel-cepstrais	p. 79
7	Conclusão	p. 81
	Referências Bibliográficas	p. 83
	Apêndice A – Transformada de Fourier Discreta	p. 85

A.1 Transformada do Cosseno Discreta	p.85
--	------

1 *Introdução*

A identificação biométrica baseia-se em características individuais que são extraídas diretamente dos usuários de determinado serviço. Há inúmeras aplicações que empregam parâmetros biométricos, como, por exemplo, sistemas de reconhecimento de impressão digital, de íris, de face, de locutor, veias da mão, entre outros. A vantagem de se utilizar informações biométricas extraídas de uma locução deve-se à facilidade com que esta é capturada do ambiente e processada, não exigindo equipamentos de elevada complexidade ou custo. O termo *locutor* é empregado quando se deseja denotar o indivíduo que gerou determinada locução. Propriedades adequadamente extraídas de uma locução são relacionadas de forma mais provável com o seu respectivo locutor, permitindo criar sistemas de identificação e verificação que apresentam elevado grau de precisão. O índice de acertos no sistema de identificação é influenciado por uma série de fatores, sendo que a qualidade do sinal e diferenças no canal de comunicação são fatores que causam forte impacto no desempenho do sistema. Neste trabalho é utilizada a técnica denominada Gaussian Mixture Model (GMM)[1] ou modelos de mistura de gaussianas, através da qual é obtido um modelo para cada locutor do sistema de identificação, permitindo que ao ser fornecida uma locução de teste ao sistema, seja possível verificar o modelo de locutor mais correlacionado com as propriedades da locução inserida. Este locutor é então eleito pelo sistema de identificação como o mais correlacionado com as propriedades da locução de teste previamente disponibilizada.

1.1 Verificação e identificação de locutor

Os termos *verificação* e *identificação* servem para abordar situações distintas. Na verificação, um indivíduo afirma ser determinado locutor e o sistema deve confirmar ou não esta identidade. Na identificação, um único locutor é escolhido a partir de um conjunto limitado de locutores. Em termos de aplicações práticas, um sistema de verificação poderia ser empregado para confirmar a identidade de uma pessoa que tenta acessar a sua conta

num banco, utilizando a própria voz como chave de acesso. Já para o caso da identificação, pode-se imaginar uma situação em que a partir de um banco de vozes de criminosos e de uma locução emitida por um criminoso desconhecido, deseja-se descobrir a identidade deste locutor. Neste trabalho é abordada a problemática relacionada aos sistemas de identificação de locutor.

1.2 Dependência e independência de texto

Uma locução emitida a partir de um texto pré-fixado é dita dependente de texto e, em caso contrário, independente de texto. Vale observar que sistemas dependentes de texto tendem a apresentar performances superiores aos independentes de texto. Tal característica deve-se ao fato de que os sistemas dependentes de texto se apoiam inicialmente no reconhecimento da locução, para então a partir da divergência entre a locução de teste e um modelo selecionado identificar o locutor. [2]. O mesmo já não acontece para sistemas independentes de texto uma vez que não se sabe previamente as características da locução utilizada na fase de teste. O sistema de identificação descrito neste trabalho é um sistema independente de texto.

1.3 Verificação de desempenho

O desempenho de um sistema de identificação é influenciado por uma série de fatores, dentre os quais se pode destacar:

- número de locutores
- qualidade da locução
- canal de comunicação

Observa-se uma degradação de desempenho com a elevação do número de locutores. Isso acontece em virtude da menor variabilidade de características que distinguem os locutores, tornando mais difícil a tarefa do sistema em destacá-los de forma exclusiva.

Outro fator a ser observado é a relação sinal-ruído da locução. As características do sinal de voz, quando extraídas de um sinal contaminado por ruído, não refletem de forma fidedigna o locutor que as gerou e, portanto, acabam por comprometer a performance do sistema de identificação.

Deve ser empregado o mesmo canal de comunicação tanto na fase de treinamento quanto na de identificação. Modelos quando treinados a partir de locuções captadas do ambiente com o emprego de determinado microfone podem se mostrar incompatíveis com sinais amostrados utilizando-se de outros equipamentos com características distintas na fase de teste ou ambientes distintos quanto às suas propriedades acústicas. Uma alternativa quando se trabalha com sinais de baixa relação sinal ruído é o emprego de técnicas de melhoria de sinal. Finalmente, deve-se observar a forma de codificação do sinal de voz. Devido às suas propriedades de estacionaridade para curtos intervalos de tempo, o mesmo deve ser tratado de forma segmentada.

1.4 Aparelho fonador

O aparelho fonador de cada locutor apresenta uma estrutura peculiar que dá origem a um sinal de voz com características exclusivas. Da componente útil deste sinal, ou seja, sinal sonoro e surdo, podem ser extraídos os coeficientes mel-cepstrais que são utilizados na modelagem do correspondente aparelho fonador. Os sistemas de identificação usualmente empregam coeficientes extraídos de banco de filtros, contudo, o mesmo pode ser feito por meio de predição linear [3], o que pode se mostrar mais conveniente sob determinadas condições conforme resultados reportados na conclusão deste trabalho.

A componente útil do sinal de voz é segregada do ruído de fundo empregando-se um DAV. Os DAVs são sensíveis à relação sinal-ruído, sendo necessário realizar adaptações em sua estrutura de acordo com a qualidade do sinal de voz utilizado.

Para contornar esta limitação, uma vez que são empregados os bancos de vozes TIMIT e sua versão quando transmitida por um canal telefônico, NTIMIT, é proposta a inserção de um estimador da relação sinal-ruído baseado no método MCRA no DAV. Dessa forma, há maior compatibilidade dos trechos segregados pelo DAV quando utilizadas amostras correspondentes dos bancos TIMIT e NTIMIT, o que permite realizar comparações de desempenho do sistema de identificação variando-se a qualidade do sinal de voz empregado e a forma como os coeficientes mel-cepstrais são obtidos.

2 *Codificação do sinal de voz*

Uma locução é captada através de um processo de filtragem passa-baixas, amostragem e quantização. Após a sua captura e armazenamento, este sinal deve ser processado para que seja possível extrair informações úteis ao sistema de identificação. A locução apresenta duas componentes distintas, ou seja, a sonora composta por sinais de características periódicas e a surda, formada por sinais que se assemelham a um ruído colorido, cuja energia concentra-se em frequências mais elevadas do que a componente sonora. Na ausência destas componentes surge o ruído de fundo.

Na figura 1 é exibida a forma de onda de uma componente sonora extraída de uma locução. Como pode ser observado, esta componente apresenta um elevado grau de periodicidade em sua forma de onda.

Figura 1: Componente sonora do sinal de voz.

Nos momentos em que as variações de amplitude são menores e mais frequentes tem-se a predominância da componente surda. Pela figura 2 é exibido um exemplo desta componente e, como pode ser constatado, o sinal é quase aleatório não apresentando características periódicas em sua forma de onda. Deve-se ainda considerar o ruído de

fundo do ambiente que é modelado como um sinal randômico de baixa intensidade.

Figura 2: Componente surdo do sinal de voz.

No processo de identificação de locutor, as características do aparelho fonador precisam ser extraídas e convenientemente categorizadas. A extração desta informação é viabilizada ao se utilizar o cepstro da componente útil do sinal de voz, conforme será visto na seção 4. Os trechos em que está presente apenas o ruído de fundo devem ser descartados com o emprego de detectores de atividade de voz conforme detalhado na seção 2.5.

Devido às propriedades de estacionaridade do sinal de voz para curtos intervalos de tempo, o mesmo deve ser tratado de forma segmentada, ou seja, este sinal é dividido em segmentos superpostos de comprimento pré-estabelecido.

2.1 Amostragem

A energia do sinal de voz concentra-se abaixo dos 8 kHz. Para que seja possível capturar esta informação, é necessário amostrar o sinal a uma taxa de pelo menos 16kHz segundo o Teorema de Nyquist. Para evitar o fenômeno de *aliasing*, as componentes do sinal, presentes em frequências superiores à metade da frequência de Nyquist, devem ser atenuadas através de um filtro passa-baixas.

2.2 Pré-ênfase

A pré-ênfase consiste em filtrar um sinal de forma a enfatizar informações que estão presentes em frequências mais elevadas. Isto se faz necessário porque as informações em frequências mais elevadas da componente útil do sinal de voz apresentam menor energia que nas frequências baixas. O filtro de pré-ênfase no domínio da transformada Z apresenta função de transferência dada pela equação (2.1), onde o coeficiente a é escolhido no intervalo de 0,95 a 0,98.

$$H(z) = 1 - az^{-1} \quad (2.1)$$

Observando-se o módulo desta função de transferência para $a = 0,97$ e $0 \leq w < \pi$, onde $z = e^{-jw}$, obtém-se a figura 3 através do qual se confirma a ação do filtro em enfatizar as frequências mais elevadas.

Figura 3: Ganho do filtro de pré-ênfase.

2.3 Análise em tempo curto

Os parâmetros do sinal de voz podem ser considerados invariantes no tempo para curtos intervalos de tempo da ordem de 10 a 30 ms [2]. Aproveitando-se desta propriedade, o sinal de voz é dividido em quadros de tamanho fixo com um número de amostras obtidas num período de tempo escolhido dentro deste intervalo, o que permite caracterizar a cada instante o trato vocal como um filtro digital. Normalmente, busca-se escolher um número de amostras que seja da forma 2^n , n inteiro, para agilizar o cálculo de transformada de

Fourier em fases posteriores do processamento do sinal de voz.

Com a formação de quadros, podem surgir descontinuidades bruscas em suas extremidades, o que viria a prejudicar a avaliação de elementos localizados nestas posições. Para tornar o sistema mais robusto a estes erros de segmentação, os quadros adjacentes são formados de maneira que apresentem determinado número de amostras em comum. Assim, um quadro de comprimento M terá $M - N$ amostras sobrepostas aos quadros que lhe são adjacentes[4], conforme pode ser observado pela figura 4.

Figura 4: Segmentação do sinal de voz.

2.4 Janelamento

A partir dos quadros obtidos no processo de segmentação do sinal de voz, faz-se necessário reduzir o efeito das variações bruscas de amplitude presentes no início e término de cada quadro. Isto é feito atenuando-se o valor das amostras que se localizam nas extremidades do quadro, multiplicando-o por uma função janela. Dessa forma, se o quadro apresentar amostras $x(n)$ e a função janela for dada por $w(n)$, após a modulação do quadro ter-se-á $y(n)$ dado pela equação (2.2).

$$y(n) = x(n)w(n) \quad (2.2)$$

Há diversas funções de janelamento. No entanto, as mais frequentemente utilizadas no processamento de sinal de voz são as janelas de Hamming e Hanning, descritas pelas equações (2.3) e (2.4) respectivamente, onde M corresponde ao comprimento da janela.

$$w(n) = 0,54 - 0,46 \cos\left(\frac{2n\pi}{M-1}\right) \quad (2.3)$$

$$w(n) = 0,5 \left(1 - \cos \left(\frac{2n\pi}{M-1} \right) \right) \quad (2.4)$$

A transformada discreta de Fourier de um quadro similar ao da figura 4, quando multiplicado por uma janela $w(n)$, é dada por $Y(k, l)$ conforme a equação (2.5). O índice k representa a frequência considerada e l , o índice do quadro atual.

$$Y(k, l) = \sum_{n=0}^{M-1} y(n + lN)w(n)e^{-j(2\pi/M)nk} \quad (2.5)$$

A título de exemplo, ao se tomar o logaritmo do módulo do espectro obtido a partir de um quadro composto prioritariamente por componente sonora e multiplicado por uma janela retangular, obtém-se a figura 5. Multiplicando este mesmo quadro por uma janela de Hamming, obtém-se o espectro da figura 6. Como pode ser constatado ao se comparar estas figuras a janela de Hamming trouxe maior definição de amplitude ao espectro do sinal de voz e, em contrapartida, uma menor resolução em frequência. Isso deve-se à maior largura do lóbulo principal da janela de Hamming quando comparado ao de uma janela retangular.

Figura 5: Módulo do espectro do sinal de voz pré-enfatizado convoluído pelo espectro de uma janela retangular.

Figura 6: Módulo do espectro do sinal de voz pré-enfatizado convoluído pelo espectro de uma janela de Hamming.

2.5 Detector de sinal de voz

Não é possível extrair informações úteis para caracterizar o aparelho fonador dos trechos do sinal de voz em que está presente apenas o ruído de fundo. Assim, esta componente deve ser expurgada antes de inserida no sistema de identificação. Os DAVs realizam esta tarefa, ou seja, a de segregar o ruído de fundo da componente útil presente no sinal de voz. Isso é feito a partir de técnicas como a verificação do número de cruzamentos por zero, a detecção da diferença do nível de energia entre a componente surda e a sonora, entre outras. Independentemente da técnica empregada, devem-se separar os quadros que trazem informação útil daqueles que são compostos essencialmente por ruído de fundo, sendo aproveitados apenas os quadros com a componente útil.

2.5.1 Detector baseado no número de cruzamentos por zero e na energia

A técnica de cruzamento por zero consiste em verificar a frequência de alternância entre sinais positivos e negativos ao longo de uma janela de amostras. Como o ruído de fundo e a componente surda do sinal de voz apresentam um elevado grau de aleatoriedade,

suas taxas de cruzamento por zero tende a ser muito mais elevada do que para os trechos sonoros. Na figura 7 é possível comparar a taxa de cruzamento por zero (superior) com o correspondente sinal de voz (inferior) de uma locução extraída do banco TIMIT. Na figura 8 tem-se a correspondente locução de teste para o caso em que esta é transmitida por um canal telefônico, utilizando-se o banco NTIMIT. Observa-se nas figuras 7 e 8 uma linha horizontal que serve como referência para discriminar a componente sonora das demais, ou seja, quando um ponto no diagrama de cruzamento por zero está acima deste limiar, tem-se a predominância da componente surda ou ruído de fundo.

Nota-se através das figuras 7 e 8 que utilizando um mesmo limiar de segregação, representado pela linha horizontal, são obtidas diferentes amostras para cada uma das componentes de acordo com a qualidade do sinal de voz empregado.

Figura 7: Diagrama de cruzamento por zero para locução com elevada relação sinal-ruído.

Figura 8: Diagrama de cruzamento por zero para locução com baixa relação sinal-ruído.

Adicionalmente, observa-se que não é possível destacar o ruído de fundo da componente surda do sinal de voz uma vez que ambos apresentam um elevado índice de cruzamento por zero. Portanto, como o uso exclusivo da técnica de cruzamentos por zero não permite segregar adequadamente a componente útil, deve-se ainda considerar a energia do sinal de voz [5] para que esta tarefa seja realizada de forma apropriada.

A energia do sinal de voz é calculada de acordo com a equação (2.1) a partir de um intervalo de medições centrado na amostra atual.

$$E(n) = \sum_{i=-n_o}^{n_o} |s(n+i)|^2 \quad (2.6)$$

Na figura 9 é possível visualizar de que forma se distribui a energia de um sinal de voz extraído do banco TIMIT. Vale observar que amostras acima da linha horizontal presente no diagrama de energia indicam uma componente útil da locução. Na figura 10 tem-se a distribuição da energia ao ser utilizada a locução correspondente do banco NTIMIT.

Figura 9: Diagrama de energia para locução com elevada relação sinal-ruído.

Novamente verifica-se a necessidade de fixar limiares diferenciados de energia para compatibilizar as componentes segregadas de locuções correspondentes dos bancos TIMIT e NTIMIT. Tal fato evidencia que o emprego conjunto das técnicas de nível de cruzamento por zero e energia [5] exige adaptações no DAV para garantir uma decisão compatível de segregação quando se utilizam sinais de diferentes qualidades.

Na tabela 1 estão relacionadas as características de cruzamento por zero e de energia de forma comparativa para as diferentes componentes de um sinal de voz, evidenciando a necessidade de monitoramento de ambas as características apontadas.

Figura 10: Diagrama de energia para locução com baixa relação sinal-ruído.

decisão	nível de cruzamento por zero	nível de energia
componente sonora	baixo	alto
componente surda	alto	médio
componente surda	alto	médio
ruído de fundo	alto	baixo

Tabela 1: Níveis de decisão para segregação de componentes de uma locução.

2.5.2 Detector baseado no método Minima Controlled Recursive Averaging

O DAV apresentado na seção 2.5.1 precisa ser calibrado de acordo com a relação sinal-ruído do sinal de voz de forma a compatibilizar a componente útil extraída de locuções correspondentes dos bancos TIMIT e NTIMIT. Considerando que são entregues segmentos de sinal de voz ao DAV após o processo de janelamento discutido na seção 2.4, a necessidade de um conjunto de amostras para que se inicie o cálculo dos limiares de segregação supracitados faz com que as primeiras amostras destes segmentos não possam ser categorizadas adequadamente até que se processe um número mínimo delas.

Dessa forma, com o intuito de se adotar uma metodologia que permita extrair a componente útil da locução sem que seja necessário adaptar parâmetros no DAV sempre que a relação sinal-ruído se alterar e que não dependa de um grande número de amostras iniciais de cada quadro para iniciar o cálculo dos parâmetros de segregação, propõem-se o emprego do método *Minima Controlled Recursive Averaging (MCRA)*[6], o qual se baseia na observação da razão entre a energia local de determinado quadro pela energia mínima do ruído de fundo presente numa sequência de quadros que precedem o quadro atual.

A energia do ruído de fundo pode ser estimada a partir do espectro segmentado suavizado do sinal de voz. Este processo de suavização ocorre em duas etapas através das equações 2.7 e 2.8.

A energia local de um quadro, aqui denominada S_f , pode ser calculada pela equação (2.7), onde $b(i)$ corresponde à i -ésima amostra de uma janela de Hanning de comprimento 3 e $Y(k, l)$ corresponde à k -ésima amostra da transformada discreta de Fourier do quadro de índice l .

$$S_f(k, l) = \sum_{i=-1}^1 b(i) |Y(k - i, l)|^2 \quad (2.7)$$

Para cada quadro, a energia do ruído de fundo é obtida recursivamente pela equação (2.8) onde o valor do parâmetro α_s deve ser escolhido no intervalo $0,7 \leq \alpha_s \leq 0,9$ [6]. Vale observar que as variações no valor de $S_f(k, l)$ são sentidas de uma forma mais atenuada em $S(k, l)$. Assim, pode-se considerar que $S(k, l)$ corresponde a uma versão suavizada de $S_f(k, l)$ e, este último, a uma versão suavizada de $|Y(k, l)|^2$.

$$S(k, l) = \alpha_s S(k, l - 1) + (1 - \alpha_s) S_f(k, l) \quad (2.8)$$

Na figura 11 é possível comparar o espectro de potência original, ou seja, $|Y(k, l)|^2$ com sua versão suavizada dada por $S(k, l)$.

Dada uma frequência k , deve ser obtido o valor mínimo de $S(k, l)$ para uma sequência de D quadros que antecedem o quadro atual l conforme a equação (2.9):

$$S_{min}(k, l) = \min_{0 \leq j \leq D-1} S(k, l - j) \quad (2.9)$$

Estes valores mínimos de $S(k, l)$ para cada frequência k correspondem à uma estimativa da energia do ruído [7] denominada S_{min} .

A determinação de S_{min} a cada instante é muito onerosa se simplesmente for vasculhado o valor mínimo de $S(k, l)$ em D quadros anteriores ao atual. Dessa forma, procura-se adotar o procedimento descrito abaixo que reduz significativamente o número de operações necessárias para se encontrar S_{min} . Os valores das variáveis U e V são escolhidos de forma que $D = U \times V$. O valor da variável D deve ser tal que o intervalo de busca de $S_{min}(k, l)$ abranja um número de quadros presentes entre 0,5 e 1,5 segundos [6].

Figura 11: Espectro de potência de um sinal de voz e sua versão suavizada.

```

1:  TMP=0
2:  Para l=1 até Último quadro
3:      n = resto da divisão de TMP por U
4:      Para k=0 até Última frequência
5:          Smin(k,l) = min { Smin(k,l-1) , S(k,l) }
6:          Stmp(k,n) = min { Stmp(k,n) , S(k,l) }
7:      Se l for divisível por V
8:          TMP = TMP + 1
9:      Para k=0 até Última frequência
10:         Smin(k,l) = min { Stmp(k,J) }   para J=0,1,2,...,U-1
11:         n = resto da divisão de TMP por U
12:         Stmp(k,n) = S(k,l)

```

Este procedimento consiste em tomar para cada frequência k um segmento de D amostras formado por U segmentos de V amostras, conforme figura 12 onde $V = 4$, $U = 3$ e $D = 12$. A cada segmento de V amostras, o termo $S_{tmp}(k,n)$ assume o valor mínimo de $S(k,l)$ neste intervalo. O índice n , compreendido entre 1 e U , é atualizado de forma sequencial e circular sempre que a posição da amostra atual for divisível por V .

Considerando o caso de teste em que $U = 3$ e para uma frequência k_1 específica, os valores de $S_{tmp}(k_1,n)$ correspondem aos valores mínimos de $S(k_1,l)$ para os quadros de

Figura 12: Algoritmo para encontrar S_{min} .

índice l delimitados pela tabela 2.

quadros l	$S_{tmp}(k_1, 0)$	$S_{tmp}(k_1, 1)$	$S_{tmp}(k_1, 2)$
0 a $3V-1$	0 a $V-1$	V a $2V-1$	$2V$ a $3V-1$
V a $4V-1$	$3V$ a $4V-1$	V a $2V-1$	$2V$ a $3V-1$
$2V$ a $5V-1$	$3V$ a $4V-1$	$4V$ a $5V-1$	$2V$ a $3V-1$
...	...	...	...

Tabela 2: Intervalo de busca para os valores de S_{tmp} .

Através das linhas 5 e 10 do procedimento supracitado, o valor de $S_{min}(k_1, l)$ é constantemente atualizado de forma que corresponda ao valor mínimo entre $S_{tmp}(k_1, 0)$, $S_{tmp}(k_1, 1)$, $\dots$, $S_{tmp}(k_1, U-1)$. Como exemplo, tomando-se a linha 3 da tabela 2 ter-se-ia para $l = 5V - 1$:

$$S_{tmp}(k_1, 0) = \min \{S(k_1, 3V), S(k_1, 3V + 1), \dots, S(k_1, 4V - 1)\}$$

$$S_{tmp}(k_1, 1) = \min \{S(k_1, 4V), S(k_1, 4V + 1), \dots, S(k_1, 5V - 1)\}$$

$$S_{tmp}(k_1, 2) = \min \{S(k_1, 2V), S(k_1, 2V + 1), \dots, S(k_1, 3V - 1)\}$$

E para o intervalo de comprimento D , onde $2V \leq l \leq 5V - 1$, o valor de S_{min} corresponderia a:

$$S_{min}(k_1, 5V - 1) = \min \{S_{tmp}(k_1, 0), S_{tmp}(k_1, 1), S_{tmp}(k_1, 2)\}$$

Como pode ser constatado, o valor mínimo de $S(k, l)$ em uma janela de D quadros está sempre disponível, sendo realizadas buscas mais intensivas de mínimos apenas na linha 10 e a cada V amostras ao invés de a cada instante.

O número de quadros D a serem pesquisados considerando um sinal amostrado a F_s , num intervalo de busca correspondente a T_b , com quadros de dimensão M e sobreposição

N , conforme figura 4, é calculado de acordo com a equação (2.10):

$$(D - 1)N + M = F_s T_b \implies D = U.V = \frac{F_s T_b - M}{N} + 1 \quad (2.10)$$

A razão entre $S(k, l)$ e $S_{min}(k, l)$ permite estabelecer a grandeza $S_r(k, l)$, dada pela equação (2.11), que é menos suscetível às variações da relação sinal-ruído do sinal de voz. Como $S_{min}(k, l)$ está diretamente relacionado à energia do ruído presente no sinal de voz, $S_r(k, l)$ é tão mais elevado quando maior for a parcela de componente sonora carregada por $S(k, l)$.

$$S_r(k, l) = \frac{S(k, l)}{S_{min}(k, l)} \quad (2.11)$$

Calcula-se para cada quadro, através da equação (2.12), a média da soma logarítmica de $S_r(k, l)$ na faixa de frequência de $N1$ a $N2$. Esta operação corresponde à integração do logaritmo do espectro de $S(k, l)$ normalizado com relação a $S_{min}(k, l)$. Os valores de $N1$ e $N2$ são escolhidos de forma que $S_m(l)$ seja computado no intervalo de frequências F_{min} a F_{max} em que se concentra a energia da componente útil do sinal de voz.

$$S_m(l) = \frac{1}{N2 - N1 + 1} \sum_{i=N1}^{N2} \log S_r(i, l) \quad (2.12)$$

Os valores de $N1$ e $N2$, considerando quadros com M amostras, frequência de amostragem F_s , frequência mínima F_{min} e máxima F_{max} , são dados pelas equações (2.13) e (2.14) respectivamente.

$$N1 = \frac{F_{min}}{F_s} M \quad (2.13)$$

$$N2 = \frac{F_{max}}{F_s} M \quad (2.14)$$

Na figura 13, é visualizada a sobreposição do módulo do espectro do sinal de voz oriundo de vários quadros pré-enfatizados. Como pode ser constatado, a energia do espectro concentra-se em frequências superiores a aproximadamente 500Hz e inferiores a 3400Hz.

Os valores de $S_m(l)$ para os quadros de uma locução com elevada relação sinal-ruído e de sua versão quando transmitida por um canal telefônico, são observados pela figura

Figura 13: Módulo do espectro de vários quadros de sinal de voz pré-ênfatizado.

14. A linha contínua representa os valores de $S_m(l)$ oriundos do sinal limpo e a tracejada, deste mesmo sinal quando degradado pelo canal telefônico. Os quadros que apresentam valor de $S_m(l)$ inferior ao limiar δ representado pela linha horizontal são descartados uma vez que não são compostos predominantemente de componente útil do sinal de voz.

Observa-se que somente após um período de adaptação inicial há maior coincidência entre as linhas tracejada e contínua na figura 14, o que não compromete o desempenho do DAV já que esta adaptação ocorre rapidamente, garantindo ainda assim grande similaridade das parcelas segregadas de sinal útil.

Confirma-se, desta forma, que é possível utilizar um mesmo limiar δ de segregação de quadros para diversas relações sinal-ruído presentes no sinal a ser trabalhado, evitando a necessidade de adaptações ao DAV de acordo com as características do sinal disponibilizado ao sistema de identificação.

O limiar δ é calibrado experimentalmente. Esse processo de calibração consiste em reconstruir a locução utilizando técnicas de *overlap-add* [8] a partir dos quadros que foram submetidos ao DAV e cujos valores de $S_m(l)$ superaram δ , sendo que os quadros com predominância de ruído têm suas amostras substituídas por valores nulos. A partir de um valor elevado de δ , este é reduzido gradativamente até que seja obtida uma boa percepção acústica da locução assim reconstruída, quando se atinge o seu valor definitivo.

Figura 14: Valores de $S_m(l)$ para determinado sinal limpo e quando transmitido por linha telefônica.

2.5.3 Parametrização do sinal de voz

Uma locução precisa ser parametrizada adequadamente em vetores para que estes sejam processados pelo sistema de identificação. Esta parametrização implica em gerar uma representação cepstral da locução de forma que os vetores assim produzidos sejam formados por coeficientes cepstrais. Há duas abordagens clássicas para se obter estes vetores característicos, conforme será detalhado nos capítulos 3 e 4, que são:

- a partir de coeficientes de predição linear
- a partir de banco de filtros

3 *Predição linear do sinal de voz*

O processo de predição linear consiste em estimar valores futuros de um sinal a partir de amostras passadas deste mesmo sinal. O termo *linear predictive coding (LPC)* é correntemente utilizado para o processamento digital de sinal.

3.1 O modelo LPC do sinal de voz

O sistema de locução humano pode ser representado, em linhas gerais, como as cordas vocais sendo equivalentes a um gerador de pulsos, onde a intensidade do sinal sonoro é determinada por um ganho G . O aparelho fonador é modelado como um filtro linear $H(z)$, também denominado filtro de síntese. Quando um segmento de locução não é sonora, considera-se que o trato vocal está sendo excitado por uma fonte de ruído [9], o que equivale à chave presente na figura 15 estar recebendo informação exclusivamente do gerador de ruído.

Figura 15: Sintetizador do sinal de voz baseado no modelo LPC.

Os coeficientes de predição linear permitem recuperar a envoltória espectral de um sinal. Os parâmetros do filtro $H(z)$ são estimados a partir de um sinal sonoro ou surdo empregando a técnica denominada *linear predictive coding (LPC)*, onde o módulo do

espectro da resposta impulsiva de $H(z)$ corresponde à envoltória espectral do sinal. A predição linear baseia-se na aproximação de uma amostra em determinado instante por um número limitado de amostras deste mesmo sinal em p instantes anteriores de forma ponderada, conforme o somatório da equação (3.1).

Considerando o termo de excitação $Gu(n)$, o qual pode ser uma fonte de excitação sonora ou surda, conforme destacado na figura 15, e a formulação de predição linear supracitada, chega-se ao modelo LPC do sinal de voz da equação (3.1).

$$s(n) = \sum_{i=1}^p a(i)s(n-i) + Gu(n) \quad (3.1)$$

Tomando-se a transformada Z da equação (3.1), obtém-se:

$$S(z) = \sum_{i=1}^p a(i)z^{-i}S(z) + GU(z) \quad (3.2)$$

Dividindo a equação (3.2) por $S(z)$, chega-se à função de transferência $H(z)$ do filtro digital, conforme equação (3.3).

$$\begin{aligned} H(z) &= \frac{S(z)}{U(z)} \\ &= \frac{G}{1 - \sum_{i=1}^p a(i)z^{-1}} \end{aligned} \quad (3.3)$$

3.2 Estimação dos coeficiente do filtro LPC

Observando as p amostras anteriores de $s(n)$, o sinal $\hat{s}(n)$ corresponde a uma estimativa do valor presente de $s(n)$, conforme equação (3.4).

$$\hat{s}(n) = \sum_{i=1}^p a_i s(n-i) \quad (3.4)$$

O erro de predição $Gu(n)$, presente na equação (3.1), pode então ser representado como a diferença entre os sinais $s(n)$ e $\hat{s}(n)$, conforme equação (3.5).

$$\begin{aligned}
e(n) = s(n) - \hat{s}(n) &\implies e(n) = s(n) - \sum_{i=1}^p a_i s(n-i) \\
&= -\{-s(n) + \sum_{i=1}^p a_i s(n-i)\} \\
&= -(\sum_{i=0}^p a_i s(n-i))
\end{aligned} \tag{3.5}$$

Na determinação dos coeficientes de predição linear, é necessário definir um critério de otimização, o qual pode ser a minimização do erro quadrático médio definido na equação (3.6). Assim, escolhendo os coeficientes a_i apropriadamente, conforme equação (3.7), é possível minimizar, no sentido quadrático, o erro de predição.

$$\varepsilon = \sum_{n=-\infty}^{\infty} e^2(n) \tag{3.6}$$

$$\{a\}_{i=1}^p = \operatorname{argmin}\{\varepsilon(a_1, a_2, \dots, a_p)\} \tag{3.7}$$

Retomando o desenvolvimento em (3.5), tem-se que $a_0 = -1$. Substituindo na equação (3.6) o valor do erro de predição obtido na equação (3.5), chega-se a:

$$\varepsilon = \sum_{n=-\infty}^{\infty} \left(\sum_{i=0}^p a(i)s(n-i) \right)^2 = \sum_{i=0}^p \sum_{j=0}^p a(i)a(j) \sum_{n=-\infty}^{\infty} s(n-i)s(n-j) \tag{3.8}$$

Definindo os coeficientes de correlação φ_{ij} a partir da equação (3.9), observa-se que $\varphi_{ij} = \varphi_{ji}$. Assim, a expressão do erro quadrático de predição pode ser reformulada conforme a equação (3.10).

$$\varphi_{ij} = \sum_{n=-\infty}^{\infty} s(n-i)s(n-j) \tag{3.9}$$

$$\varepsilon = \sum_{i=0}^p \sum_{j=0}^p a_i \varphi_{ij} a_j \tag{3.10}$$

A minimização do erro quadrático é obtida tomando o gradiente de ε com relação aos coeficientes de predição a_i e igualando-o ao vetor zero, conforme equação (3.11).

$$\frac{\partial \varepsilon}{\partial a_j} = 0 \quad \text{para } j = 1, 2, \dots, p \tag{3.11}$$

Substituindo a expressão do erro ε , definido em (3.10), na equação (3.11) e separando os termos de forma apropriada, segue a equação (3.12).

$$\frac{\partial \varepsilon}{\partial a_k} = \frac{\partial \varepsilon}{\partial a_k} \left(\sum_{\substack{i=0 \\ i \neq j}}^p \sum_{j=0}^p a_i \varphi_{ij} a_j + \sum_{j=0}^p a_j^2 \varphi_{jj} \right) = \sum_{\substack{j=0 \\ j \neq k}}^p \varphi_{kj} a_j + \sum_{\substack{i=0 \\ i \neq k}}^p a_i \varphi_{ik} + 2a_k \varphi_{kk} = 0 \quad (3.12)$$

Considerando que $\varphi_{ij} = \varphi_{ji}$ e que $a_0 = -1$, a equação (3.12) pode ser simplificada para a equação (3.13).

$$2 \sum_{j=0}^p \varphi_{kj} a_j = -\varphi_{k0} + \sum_{j=1}^p \varphi_{kj} a_j = 0 \Rightarrow \sum_{j=1}^p \varphi_{kj} a_j = \varphi_{k0} \quad \text{para } k = 1, 2, \dots, p \quad (3.13)$$

Colocando o sistema da equação (3.13) em notação matricial, segue que $\Phi \vec{\mathbf{a}} = \vec{\Psi}$, onde:

$$\Phi = \begin{pmatrix} \varphi_{11} & \varphi_{12} & \dots & \varphi_{1p} \\ \varphi_{12} & \varphi_{22} & \dots & \varphi_{2p} \\ \dots & \dots & \dots & \dots \\ \varphi_{1p} & \varphi_{2p} & \dots & \varphi_{pp} \end{pmatrix} \quad \vec{\mathbf{a}} = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{pmatrix} \quad \vec{\Psi} = \begin{pmatrix} \varphi_{01} \\ \varphi_{02} \\ \dots \\ \varphi_{0p} \end{pmatrix} \quad (3.14)$$

Substituindo $m = n - i$ na equação (3.9) é obtida a equação (3.15), a qual corresponde a uma função de autocorrelação R_{ss} . Como é sabido, toda função de autocorrelação é simétrica em torno de sua origem para sinais reais e estacionários, isto é, $R_{ss}(k) = R_{ss}(-k) \Rightarrow \sum_{m=-\infty}^{\infty} s(m)s(m+k) = \sum_{m=-\infty}^{\infty} s(m)s(m-k)$, dando origem ao resultado da equação (3.15).

$$\varphi_{ij} = \sum_{m=-\infty}^{\infty} s(m)s(m+i-j) \Rightarrow \varphi_{ij} = \sum_{m=-\infty}^{\infty} s(m)s(m+|i-j|) \quad (3.15)$$

Para um quadro de M amostras, os valores de $s(m)$ são definidos apenas no intervalo $0 \leq m \leq M - 1$, assim, o limite inferior do somatório na equação (3.15) corresponde a $m \geq 0$.

As variáveis i e j da equação (3.15) podem assumir os valores $\{1, 2, \dots, p\}$. Portanto, $|i-j| \leq p$, fazendo com que o limite superior do somatório na equação (3.15) seja limitado de acordo com a relação $m + p \leq M - 1 \Rightarrow m \leq M - p - 1$, dando origem à equação

(3.16) onde $k = |i - j|$.

$$\varphi_{ij} = \sum_{m=0}^{M-k-1} s(m)s(m+k) = R(k) \quad \text{para} \quad k = 0, 1, \dots, p \quad (3.16)$$

A matriz Φ da equação (3.14) resultou ser a matriz de autocorrelação R_{ss} . Esta matriz apresenta uma simetria de Toeplitz, ou seja, os elementos da diagonal principal e suas paralelas são constantes. Além disso ela é uma matriz simétrica, conforme equação (3.17).

$$\Phi = \begin{pmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ R(2) & R(1) & R(0) & \dots & R(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R(p-1) & R(p-2) & \dots & R(1) & R(0) \end{pmatrix} \quad (3.17)$$

Sistemas lineares que apresentam simetria de Toeplitz simétrica podem ser resolvidos recursivamente a partir do algoritmo de Durbin [10] descrito pela equação (3.18), valendo lembrar que o coeficiente de predição linear a_0 apresenta valor unitário negativo.

$$\begin{aligned} E^{(0)} &= R(0) \\ \text{para } 1 \leq i \leq p \quad L_i &= \frac{\{R(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} R(|i-j|)\}}{E^{(i-1)}} \end{aligned} \quad (3.18)$$

$$\begin{aligned} \alpha_i^i &= L_i \\ \text{para } j = 1, 2, \dots, i-1 \quad \alpha_j^i &= \alpha_j^{i-1} - L_i \alpha_{i-j}^{i-1} \\ E^{(i)} &= (1 - L_i^2) E^{(i-1)} \end{aligned}$$

Para cada quadro considerado, os coeficientes de predição linear a_k são calculados através da equação (3.19).

$$a_k = \alpha_k^p \quad \text{para} \quad 1 \leq k \leq p \quad (3.19)$$

Substituindo os coeficientes a_k da equação (3.19) no filtro digital $H(z)$ da equação (3.3), considerando ganho unitário, é obtida a função de transferência do aparelho fonador. A resposta em frequência de $H(z)$ é uma versão suavizada da envoltória espectral do sinal

de voz e depende do número de coeficientes de predição empregados. Nas figuras 16 e 17, observa-se o módulo do espectro de $H(z)$ em linha pontilhada e o espectro do sinal de voz em linha contínua para 10 e 15 coeficientes de predição respectivamente. Como pode ser confirmado, elevando o número de coeficientes, o espectro de $H(z)$ se aproxima daquele apresentado pelo sinal de voz.

Figura 16: Espectro do sinal de voz e do filtro $H(z)$ com 10 coeficientes de predição linear.

Figura 17: Espectro do sinal de voz e do filtro $H(z)$ com 16 coeficientes de predição linear.

Figura 18: Espectro do sinal de voz e do filtro $H(z)$ com 128 coeficientes de predição linear.

Dessa forma, para cada quadro que contém informação útil do sinal de voz, são obtidos os coeficientes de predição linear pelo procedimento descrito anteriormente. Este conjunto de coeficientes é utilizado para compor os vetores que serão empregados nas fases posteriores do sistema de identificação.

4 *Cepstro do sinal de voz*

O aparelho fonador é modelado como um sistema variante no tempo cujas propriedades variam lentamente. Para fins de modelagem, entretanto, cada quadro do sinal de voz, obtido no processo de segmentação, tem seu correspondente filtro digital tido como linear e invariante no tempo, sendo excitado por um trem de pulsos periódicos ou ruído pseudoaleatório, dependendo de quando se está ou não tratando de segmento sonoro do sinal de voz. Sistemas lineares e invariantes no tempo apresentam resposta à excitação correspondente à convolução do sinal de entrada pela sua resposta impulsiva. Retomando os sinais presentes na figura 15, a convolução do sinal de entrada $u(n)$ pela resposta impulsiva do sistema $h(n)$ resultando no sinal $s(n)$, dá origem à formulação da equação (4.1), onde o ganho G é considerado unitário pois o seu valor não a forma do espectro do modelo preditivo.

$$s(n) = \sum_{k=-\infty}^{\infty} u(k)h(n-k) \quad (4.1)$$

No domínio da frequência, a representação deste processo é dada pela equação (4.2) onde $U(e^{jw})$ e $H(e^{jw})$ correspondem aos espectros dos sinais $u(n)$ e $h(n)$ respectivamente.

$$S(e^{jw}) = U(e^{jw})H(e^{jw}) \quad (4.2)$$

Tomando o logaritmo da equação (4.2), é possível separar em parcelas as componentes oriundas do espectro do sinal de excitação e da função de transferência do filtro digital, conforme a equação (4.3). Esta propriedade é encontrada em sistemas homomórficos.

$$\log[S(e^{jw})] = \log[U(e^{jw})] + \log[H(e^{jw})] \quad (4.3)$$

O cepstro real [11] de uma sequência $s(n)$ é definido como a transformada inversa de Fourier de $\log(|S(e^{jw})|)$. Alternativamente [12] pode-se adotar o espectro de potência

ao invés do módulo do espectro, conforme equação (4.4). O cálculo do cepstro complexo é análogo ao do cepstro, com exceção da aplicação do módulo sobre o sinal $S(e^{jw})$ na equação (4.4).

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(e^{jw})|^2 e^{jwn} dw \quad \text{para} \quad -\infty \leq n \leq \infty \quad (4.4)$$

Expandindo a equação (4.4) com os termos $U(e^{jw})$ e $H(e^{jw})$ é obtida a equação (4.5). Esta operação de superposição é típica de sistemas homomórficos [11], onde o sistema é linear.

$$\int_{-\pi}^{\pi} \log |S(e^{jw})| e^{jwn} dw = \int_{-\pi}^{\pi} \log |U(e^{jw})| e^{jwn} dw + \int_{-\pi}^{\pi} \log |H(e^{jw})| e^{jwn} dw \quad (4.5)$$

Considerando a segmentação do sinal de voz em quadros de comprimento M , uma aproximação para o cepstro [11] é obtida empregando-se a transformada discreta de Fourier, conforme a equação (4.6). O primeiro coeficiente cepstral revela o ganho G do modelo LPC, conforme resultado obtido em (4.15), não trazendo informação útil para as etapas posteriores do sistema de identificação. Dessa forma, o termo $c_p(0)$ é descartado [1], sendo considerados apenas os coeficientes subsequentes.

$$c_p(n) = \frac{1}{M} \sum_{k=0}^{M-1} \log |S(k)|^2 e^{\frac{2\pi j}{M} kn} \quad \text{para} \quad 0 \leq n \leq M-1 \quad (4.6)$$

O cepstro obtido a partir da transformada discreta de Fourier é relacionado com o cepstro verdadeiro pela equação (4.7).

$$c_p(n) = \sum_{r=-\infty}^{\infty} c(n + rM) \quad \text{para} \quad 0 \leq n \leq M-1 \quad (4.7)$$

O cepstro $c_p(n)$ para um quadro que contém informação sonora é visualizado pela figura 19 enquanto que na figura 20 observa-se o cepstro do ruído de fundo. Como pode ser notado, quando o quadro contém informação sonora, os primeiros coeficientes cepstrais apresentam uma variação de amplitude significativa e periodicamente surgem máximos locais. O mesmo já não acontece para os quadros que contemplam a componente surda, conforme figura 21. Nesta situação, os máximos locais deixam de existir, permanecendo apenas os primeiros coeficientes com amplitude mais significativa.

Analisando a figura 19, é possível notar que surgem dois picos no valor do cepstro em torno de 10ms e 22ms. Estes picos são devidos ao termo $\log(U(e^{jw}))$, presente na equação (4.3), e o intervalo de tempo entre eles, aproximadamente 12ms, representa o *pitch* do sinal de voz, ou seja, sua frequência fundamental percebida. As oscilações no valor do cepstro, devidas aos seus coeficientes iniciais, estão vinculadas às características do aparelho fonador, ou seja, $\log(H(e^{jw}))$ e são empregadas pelo sistema de identificação para destacar os diferentes locutores.

A componente de excitação que surge a partir do termo $\log(U(e^{jw}))$ não é necessária ao sistema de identificação, sendo descartada ao se desconsiderar os coeficientes cepstrais de ordem mais elevada. Os quadros que contemplam a componente surda não apresentam os coeficientes de ordem mais elevada devido à inexistência do sinal de excitação $U(e^{jw})$.

Figura 19: Cepstro da componente sonora do sinal de voz.

4.1 Estimação dos coeficientes cepstrais

Considerando a forma alternativa de cálculo do cepstro em que é adotado o espectro de potência ao invés do módulo do espectro, o cepstro da resposta impulsiva do filtro LPC definido em (3.3) corresponde ao resultado definido na equação (4.8).

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |H(e^{jw})|^2 e^{jwn} dw \quad \text{para} \quad -\infty \leq n \leq \infty \quad (4.8)$$

A transformada de Fourier do cepstro definido pela equação (4.8) dá origem ao espectro da equação (4.9).

Figura 20: Cepstro do ruído de fundo.

Figura 21: Cepstro da componente surda do sinal de voz.

$$C(e^{jw}) = \log |H(e^{jw})|^2 = \log H(e^{jw})H^*(e^{jw}) = \log H(e^{jw}) + \log H^*(e^{jw}) \quad (4.9)$$

Considerando a causalidade da resposta impulsiva $h(n)$ que modela um sistema físico, é possível formular a equação (4.10).

$$H(e^{jw}) = \sum_{n=0}^{\infty} h(n)e^{-jwn} \quad (4.10)$$

O complexo conjugado $H^*(e^{jw})$ é obtido conforme a equação (4.11).

$$H^*(e^{jw}) = \sum_{n=0}^{\infty} h(n)e^{jwn} \implies H^*(e^{jw}) = \sum_{m=-\infty}^0 h(-m)e^{-jwm} \quad (4.11)$$

Recorrendo à definição da transformada de Fourier do cepstro dada pela equação (4.9), tem-se o desenvolvimento da equação (4.12).

$$\begin{aligned} C(e^{jw}) &= \sum_{n=-\infty}^{\infty} c(n)e^{-jwn} \\ &= \sum_{n=-\infty}^0 c(n)e^{-jwn} + \sum_{n=0}^{\infty} c(n)e^{-jwn} \\ &= \log H^*(e^{jw}) + \log H(e^{jw}) \end{aligned} \quad (4.12)$$

Considerando o cepstro para $n \geq 0$ na equação (4.12) e calculando a sua transformada Z, obtém-se a igualdade definida pela equação (4.13).

$$\sum_{k=0}^{\infty} c(k)z^{-k} = \log H(z) \quad (4.13)$$

Definindo o filtro $A(z)$ como o inverso do filtro de síntese $H(z)$ multiplicado pelo ganho G do modelo LPC, conforme (3.3), tem-se o desenvolvimento da equação (4.14).

$$A(z) = 1 - \sum_{i=1}^p a_i z^{-i} \implies \sum_{k=0}^{\infty} c_k z^{-k} = \log G - \log A(z) \quad (4.14)$$

Considerando que $z = re^{jw}$, onde r é real, pode-se determinar o valor do coeficiente c_0 [12] observando o valor de z quando $r \rightarrow \infty$, conforme demonstrado no desenvolvimento da equação (4.15), partindo-se do resultado obtido na equação (4.14).

$$\lim_{r \rightarrow \infty} \sum_{k=0}^{\infty} c_k z^{-k} = c_0 + \lim_{r \rightarrow \infty} \sum_{k=1}^{\infty} c_k z^{-k} = c_0 \quad (4.15)$$

$$\lim_{r \rightarrow \infty} A(z) = 1 - \lim_{r \rightarrow \infty} \sum_{i=1}^p a_i z^{-i} = 1 \implies \log A(z) = 0$$

$$\sum_{k=0}^{\infty} c_k z^{-k} = \log G - \log A(z) \implies c_0 = \log G$$

Dessa forma, chega-se à relação definida pela equação (4.16).

$$\sum_{k=1}^{\infty} c_k z^{-k} = -\log A(z) \quad (4.16)$$

Derivando a equação (4.16) em relação a z^{-1} é obtido o resultado exibido pela equação (4.17).

$$\frac{d}{dz^{-1}}(\log(A(z))) = \frac{1}{A(z)} \frac{dA(z)}{dz^{-1}} = \frac{-1}{1 - \sum_{i=1}^p a_i z^{-i}} \sum_{i=1}^p i a_i z^{-(i-1)} \quad (4.17)$$

$$\frac{d}{dz^{-1}}\left(\sum_{i=1}^{\infty} c_i z^{-i}\right) = \sum_{i=1}^{\infty} i c_i z^{-(i-1)}$$

$$\sum_{i=1}^p i a_i z^{-(i-1)} = \left(1 - \sum_{i=1}^p a_i z^{-i}\right) \sum_{i=1}^{\infty} i c_i z^{-(i-1)}$$

Expandindo o resultado obtido na equação (4.17) e igualando os termos que apresentem mesmo coeficiente z^{-1} , é obtida uma relação direta entre os coeficientes cepstrais e os de predição linear. A título de exemplo, para $p = 3$, ter-se-ia o seguinte desenvolvimento:

$$\begin{aligned}
(a_1 + 2a_2z^{-1} + 3a_3z^{-2}) &= (1 - a_1z^{-1} - a_2z^{-1} - a_3z^{-1})(c_1 + 2c_2z^{-1} + 3c_3z^{-2} + 4c_4^{-3} + \dots) \\
a_1 &= c_1 \\
2a_2 &= 2c_2 - c_1a_1 \Rightarrow c_2 = a_2 + \frac{c_1a_1}{2} \\
3a_3 &= 3c_3 - 2c_2a_1 - c_1a_2 \Rightarrow c_3 = a_3 + \frac{c_1a_2 + 2c_2a_1}{3} \\
0 &= 4c_4 - 3c_3a_1 - 2c_2a_2 - c_1a_3 \Rightarrow c_4 = \frac{c_1a_3 + 2c_2a_2 + 3c_3a_1}{4} \\
0 &= 5c_5 - 4c_4a_1 - 3c_3a_2 - 2c_2a_3 \Rightarrow c_5 = \frac{2c_2a_3 + 3c_3a_2 + 4c_4a_1}{5} \\
&\dots
\end{aligned} \tag{4.18}$$

Observando o desenvolvimento da equação (4.18), é possível deduzir as relações definidas na equação (4.19).

$$\begin{aligned}
c_1 &= a_1 \\
c_n &= a_n + \frac{1}{n} \sum_{k=1}^{n-1} kc_k a_{n-k} \quad \text{para} \quad 2 \leq n \leq p \\
c_n &= \frac{1}{n} \sum_{k=1}^{n-1} (n-k)c_{n-k} a_k \quad \text{para} \quad n > p
\end{aligned} \tag{4.19}$$

4.2 Escala Mel

A escala Mel surgiu com o intuito de mapear a percepção de frequência de um tom, ou *pitch*, em uma escala não linear [13] e foi concebida a partir de experimentos, onde se mapearam os incrementos subjetivos constantes de *pitch* em suas correspondentes frequências. A escala Mel se relaciona de forma logarítmica com a frequência linear, conforme definido pela equação (4.20), onde A e B são constantes obtidas experimentalmente, F_{mel} , F_{linear} e F_a são as frequências mel, linear e de amostragem respectivamente, $\tilde{w}$ e w são as frequências mel e linear no domínio discreto do tempo. Empregando os valores usuais para os fatores $A=2595$ e $B=700$, é possível verificar visualmente a relação entre as escalas mel e linear a partir do gráfico da figura 22. Como pode ser notado, tem-se uma maior resolução para baixas frequências e menor para altas, o que condiz com o espectro do sinal de voz cuja energia se concentra em baixas frequências.

$$F_{mel} = A \log(1 + F_{linear}/B) \implies \tilde{w} = \frac{2\pi A}{F_a} \log \left(1 + \frac{w F_a}{2\pi B} \right) \quad (4.20)$$

Figura 22: Mapeamento de frequência da escala mel em linear.

Os coeficientes mel-cepstrais são os coeficientes cepstrais obtidos com base na escala mel ao invés da escala linear de frequências. Assim, deve-se considerar a relação $\tilde{w}=f(w)$, onde $\tilde{w}$ e w correspondem respectivamente às frequências na escala mel e linear. A função de mapeamento f permite relacionar a escala linear com a escala mel.

4.3 Estimação dos coeficientes mel-cepstrais

Na estimação dos coeficientes mel-cepstrais, os sinais devem ser submetidos a uma fase de pré-processamento, conforme definido no capítulo 2, para que seja disponibilizado um conjunto de dados com informação útil e sem a presença de ruído do sinal de voz. Em seguida, é possível utilizar diferentes abordagens na extração dos coeficientes mel-cepstrais, sendo mais comum obtê-los a partir de banco de filtros triangulares ou de coeficientes de predição linear.

Na fase de pré-processamento, esquematizada na figura 23, o sinal é submetido a um filtro de pré-ênfase para em seguida ser dividido em segmentos. Cada segmento é então ponderado espectralmente por uma janela de Hamming e submetido ao detector de atividade de voz para que os quadros que não contêm informação útil sejam descartados.

Figura 23: Pré-processamento do sinal de voz.

4.3.1 Estimação utilizando banco de filtros

Pela figura 24 é visualizado o diagrama esquemático para a extração dos coeficientes mel-cepstrais com o emprego de um banco de filtros. Como pode ser constatado, após a fase de pré-processamento do sinal de voz, é calculado o módulo da transformada de Fourier dos quadros com informação útil e os espectros destes segmentos são submetidos a um banco de filtros triangulares, cujas freqüências centrais, obtidas experimentalmente, correspondem àquelas em que se percebem as mudanças de tom. O formato triangular do filtro permite enfatizar as componentes presentes nas freqüências centrais, atenuando as demais. Em seguida, é tomado o logaritmo do espectro resultante e calculada a sua transformada do cosseno discreta (TCD). Seguindo o sugerido na seção 2.3, procura-se adotar quadros de comprimento da forma 2^n para agilizar o cálculo da transformada discreta de Fourier com métodos de transformação rápida. Conforme descrito no apêndice A.1, a TCD é aplicada para seqüências reais como as amostras do sinal de voz e a do tipo 2 concentra energia em seus primeiros coeficientes, compactando a informação a ser processada uma vez que coeficientes de ordem mais elevada podem ser descartados.

Figura 24: Coeficientes mel-cepstrais obtidos a partir de banco de filtros.

O uso de banco de filtros traz a vantagem de reduzir a dimensionalidade da informação extraída do sinal de voz, sendo esta informação tão mais compacta quanto menor for o número de filtros empregados, conforme será visto ao final desta seção.

As freqüências centrais dos filtros triangulares estão espaçadas idealmente de forma linear segundo a escala mel, permitindo que haja um número maior de filtros nas freqüências mais baixas, onde se concentra a energia do sinal de voz. Devido a considerações de ordem prática, como o fato de que há pouca concentração de energia no sinal de voz abaixo dos 100 Hz e de que há uma correspondência quase linear entre a escala mel e a linear de freqüências até aproximadamente 1000 Hz, são utilizados valores de freqüências centrais ligeiramente distintos dos ideais ou teóricos.

Um sinal quando submetido a um filtro triangular tem as componentes que estão próximas ao centro deste filtro enfatizadas e as demais, atenuadas. Dessa forma, ao se

empregar um banco de filtros o que se está fazendo é enfatizar as frequências onde ocorrem as mudanças perceptivas de tom conforme a escala mel.

Num sistema biométrico de segurança define-se o *Equal Error Rate (ERR)* como um limiar em que se obtém uma mesma taxa de falsa aceitação e falsa rejeição. Há diversas propostas de implementação de bancos de filtros triangulares, sendo que a apresentada por Slaney [14] resultou num *ERR* no sistema de verificação de locutor proposto em [15] ligeiramente menor que de outros bancos de filtros, além de não exigir grande esforço computacional para o seu cálculo.

Para o banco de filtros em questão, considera-se que o sinal é amostrado a uma taxa de 16kHz e é composto por 40 filtros, sendo que os centros dos 13 primeiros estão espaçados linearmente e os 27 restantes apresentam um espaçamento logarítmico entre si. A variável n , presente nas equações (4.21) e (4.22), corresponde à posição do filtro linear ou logarítmico e, como pode ser constatado, a faixa de frequência abrangida pelo banco de filtros vai de 133,33Hz a 6825,2Hz .

- filtros com espaçamento linear
 - espaçamento de 66,66Hz
 - inicia em 133,33Hz
 - frequências centrais dadas por

$$F_{linear} = 133,33 + 66,66n \quad \text{para} \quad 1 \leq n \leq 13 \quad (4.21)$$

- filtros com espaçamento logarítmico
 - inicia na frequência central do último filtro linear = 1000Hz
 - frequências centrais dadas por

$$F_{log} = 1000(1,0711703)^{n-13} \quad \text{para} \quad 14 \leq n \leq 40 \quad (4.22)$$

Cada filtro triangular é dimensionado de forma a apresentar uma mesma área, a qual pode ser unitária para fins de simplificação. Chamando a frequência inicial do filtro como F_{baixa} , a central como F_{centro} e a final como F_{alta} , a altura H do triângulo correspondente a este filtro é dada pela equação (4.23).

$$H = \frac{1}{2(F_{alta} - F_{baixa})} \quad (4.23)$$

Nota-se que os filtros lineares apresentam mesma altura uma vez que a diferença $F_{alta} - F_{baixa}$ é constante. O mesmo já não acontece para os filtros logarítmicos em que a diferença $F_{alta} - F_{baixa}$ é cada vez maior, fazendo com que a altura do filtro se reduza progressivamente. Na figura 25 é ilustrado o espectro do banco de filtros no intervalo de 0 a 6825,2Hz.

Figura 25: Espectro do banco de filtros triangulares de Slaney.

Considerando quadros dispostos conforme exemplificado pela figura 4 e frequência de amostragem F_a , o cálculo da TFD destes quadros dá origem às frequências discretas F_d , conforme definido pela equação (4.24). Se $X(k)$ for o resultado do cálculo da TFD de um quadro qualquer, tomando o seu módulo chega-se à seguinte relação $|X(k)| = |X(M - k - 1)|$ para $k = 0, 1, \dots, M - 1$ e M par. Observa-se assim uma redundância de informação que pode ser suprimida sem qualquer prejuízo, permitindo que se adote na equação (4.24) apenas o intervalo $0 \leq n \leq M/2 - 1$. Esta redução de informação, que abrange a faixa de $0 \leq w \leq \pi$ no domínio da frequência, traz maior agilidade ao sistema de identificação uma vez que se está trabalhando com apenas metade das amostras originalmente disponíveis, sem que se perca informação útil.

$$F_d(n) = \frac{n}{M} F_a \quad \text{para} \quad 0 \leq n \leq \frac{M}{2} - 1 \quad (4.24)$$

As primeiras $M/2$ amostras do módulo do espectro de cada quadro são submetidas ao banco de L filtros, conforme figura 26, dando origem às grandezas $E(i)$, conforme equação (4.25). A compatibilização dimensional do espectro dos quadros com o dos filtros, dados por $H_i(k)$, é possível tomando apenas as primeiras $M/2$ amostras do espectro de cada um

dos filtros.

$$E(i) = \log\left(\sum_{k=0}^{M-1} |X(k)|H_i(k)\right) \quad \text{para} \quad 0 \leq i \leq L-1 \quad (4.25)$$

Figura 26: Cálculo do logaritmo do espectro do sinal de voz obtido do banco de filtros.

Os coeficientes mel-cepstrais são obtidos com a aplicação da TCD normalizada do tipo 2, definida pela equação (A.3), sobre os valores de $E(i)$. Esta operação permite compactar um vetor de dimensão L para outro cuja dimensão corresponda ao número de coeficientes mel-cepstrais adotado. Considerando que são empregados C coeficientes mel-cepstrais, a aplicação da TCD normalizada dá origem à equação (4.26), onde L equivale ao número de filtros empregados.

$$c(k) = \sqrt{\frac{2}{L}} \beta(k) \sum_{n=0}^{L-1} E(n) \cos\left(\frac{\pi}{L}\left(n + \frac{1}{2}\right)k\right) \quad \text{para} \quad k = 0, 1, \dots, C-1 \quad (4.26)$$

4.3.2 Estimação utilizando coeficiente de predição linear

Um filtro passa-tudo tem a peculiar característica de modificar a fase de um sinal de acordo com a frequência, apresentando sempre ganho constante. Na equação (4.27) está definido um filtro passa-tudo [16] de primeira ordem, cuja constante α é real.

$$F(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad (4.27)$$

A linha pontilhada da figura 27 corresponde ao módulo da resposta em fase do filtro passa-tudo com $\alpha = 0,42$ para o intervalo de frequências $0 \leq w \leq \pi$ e sinal cuja frequência de amostragem é de 16KHz. A linha contínua corresponde ao mapeamento de frequência linear em logarítmica, conforme figura 22, porém utilizando $A=3175$ e $B=700$ na equação (4.20). Nota-se grande similaridade no comportamento de ambas as curvas, o que sugere que é possível adotar a resposta em fase de um filtro passa-tudo como uma forma de aproximação para o comportamento de mapeamento de frequências lineares em logarítmicas.

Figura 27: Comparação entre a fase do filtro passa-tudo com a frequência na escala mel.

O espectro de um sinal na escala linear de frequências é dado por sua transformada de Fourier. Este mesmo espectro pode ser mapeado para uma outra escala de frequências de forma que, para correspondentes frequências, a amplitude do espectro seja a mesma. No domínio da transformada Z , o que se busca é a correspondência $\tilde{z} = g(z)$, onde $z = e^{jw}$ e $\tilde{z} = e^{j\tilde{w}}$. Dessa forma, adotando a filtro passa-tudo visto anteriormente como meio para se obter esta correspondência, chega-se à relação da equação (4.28).

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \implies e^{-j\tilde{w}} = \frac{e^{-jw} - \alpha}{1 - \alpha e^{-jw}} \quad (4.28)$$

Igualando a fase para ambos os lados da igualdade expressa pela equação (4.28), chega-

se à relação definida pela equação (4.29), que corresponde ao mapeamento de frequências desejado.

$$\tilde{w} = \arctan \left[\frac{(1 - \alpha^2) \sin(w)}{(1 + \alpha^2) \cos(w) - 2\alpha} \right] \quad (4.29)$$

Uma seqüência $h(n)$ com transformada Z dada por $F(z)$ apresenta a transformada de Fourier de seu cepstro dada pela equação (4.30), conforme definido anteriormente na equação (4.9). De forma análoga, se for utilizada a escala mel de frequências, obtém-se o resultado expresso pela segunda igualdade da equação (4.30).

$$\log |F(z)|^2 = \sum_{m=-\infty}^{\infty} c(m) z^{-m} = \sum_{m=-\infty}^{\infty} \tilde{c}(m) \tilde{z}^{-m} \quad (4.30)$$

Aplicando procedimento análogo ao utilizado na formulação do resultado expresso pela equação (4.13), a equação (4.30) é simplificada para a equação (4.31).

$$\sum_{m=0}^{\infty} c(m) z^{-m} = \sum_{m=0}^{\infty} \tilde{c}(m) \tilde{z}^{-m} \quad (4.31)$$

Pela equação (4.18) foi exibido um método para calcular um número limitado de coeficientes cepstrais, enquanto que na equação (4.31) pressupõem-se a disponibilidade de infinitos coeficientes cepstrais. Como infinitos coeficientes não podem ser calculados, ter-se-ia um problema de truncagem conforme equação (4.32).

$$\sum_{m=0}^{\infty} \tilde{c}(m) \tilde{z}^{-m} \approx \sum_{m=0}^L c(m) z^{-m} \quad (4.32)$$

Para contornar o problema de truncagem, recorre-se à definição de $F(z)$ dada pela equação (3.3). Os coeficientes de predição linear na escala mel são obtidos a partir da relação definida na equação (4.33) pela fórmula de recursão para transformação em frequência [17].

$$\frac{1}{1 - \sum_{i=1}^p a_i z^{-i}} = \frac{1}{1 - \sum_{k=1}^{\infty} \tilde{a}_k \tilde{z}^{-k}} \quad (4.33)$$

O cálculo dos coeficientes $\tilde{a}_k$, que compreende a resolução da equação (4.33), é feito empregando a fórmula recursiva [3] definida pela equação (4.34). Vale observar que P corresponde ao número de coeficientes de predição da equação (4.33) e C , ao de coeficientes

mel-cepstrais que se deseja calcular.

Para $i = -P, \dots, -2, -1, 0$

$$\tilde{a}^{(i)}(m) = \begin{cases} a(-i) + \alpha \tilde{a}^{(i-1)}(0) & m = 0 \\ (1 - \alpha^2) \tilde{a}^{(i-1)}(0) + \alpha \tilde{a}^{(i-1)}(1) & m = 1 \\ \tilde{a}^{(i-1)}(m-1) + \alpha (\tilde{a}^{(i-1)}(m) - \tilde{a}^{(i)}(m-1)) & m = 2, 3, \dots, C \end{cases} \quad (4.34)$$

Dispondo dos valores dos coeficientes $\tilde{a}^{(0)}(m)$, os coeficientes $\tilde{a}_k$, definidos na equação (4.33), são obtidos a partir da equação (4.35).

$$\tilde{a}_k = \frac{\tilde{a}^{(0)}(k)}{\tilde{a}^{(0)}(0)} \quad \text{para} \quad 1 \leq k \leq C \quad (4.35)$$

Recorrendo às equações (3.3), (4.13), (4.31) e (4.33) chega-se à igualdade definida pela equação (4.36).

$$\log \frac{1}{1 - \sum_{k=1}^{\infty} \tilde{a}_k \tilde{z}^{-k}} = \sum_{m=0}^{\infty} \tilde{c}(m) \tilde{z}^{-m} \quad (4.36)$$

A resolução da equação (4.36) de forma recursiva [3] dá origem aos coeficientes mel-cepstrais. Vale observar que o coeficiente de energia $\tilde{c}(0)$ não é empregado, sendo descartado nas fases posteriores do sistema de identificação.

$$\tilde{c}(m) = \tilde{a}(m) + \sum_{k=1}^{m-1} \frac{k}{m} \tilde{c}(k) \tilde{a}(m-k) \quad \text{para} \quad 1 \leq m \leq C \quad (4.37)$$

4.3.3 Subtração da média cepstral

A técnica da subtração da média cepstral (SMC) é utilizada para a remoção da distorção de canal presente no sinal de voz [18], sendo possível implementá-la a um baixo custo computacional para sistemas em que esta distorção é constante. O resultado da aplicação de um sinal $s(n)$ a um canal cuja resposta impulsiva corresponda a $w(n)$ será a convolução destas duas seqüências. Utilizando a definição do cepstro da equação (4.4) e aplicando-o ao resultado desta convolução, chega-se ao equacionamento (4.38), em que uma operação de convolução dá origem a uma operação de soma no domínio cepstral.

$$y(n) = s(n) * w(n) \implies c_y(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(e^{jw})W(e^{jw})| e^{jwn} dw \implies \quad (4.38)$$

$$c_y(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(e^{jw})| e^{jwn} dw + \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |W(e^{jw})| e^{jwn} dw = c_s(n) + c_w(n)$$

Considerando que o canal é linear e invariante no tempo e que os coeficientes cepstrais oriundos do sinal de voz apresentam média nula, com exceção de $c_s(0)$, obtém-se a equação (4.39) ao ser aplicado o operador esperança sobre a equação (4.38), valendo lembrar que a constante C corresponde ao número de coeficientes cepstrais empregado.

$$c_y(n) = c_s(n) + c_w(n) \implies E\{c_y(n)\} = \overline{c_w} \implies \overline{c_w} = \frac{1}{C} \sum_{n=1}^C c_y(n) \quad (4.39)$$

Em seguida, $\overline{c_w}$ é subtraído de cada coeficiente cepstral observado, $c_y(n)$, para que sejam obtidos os novos coeficientes $c_s^{novo}(n)$, os quais carregam a informação do sinal de voz desejada, não contaminada pelo ruído de canal, conforme equação (4.40).

$$c_s^{novo}(n) = c_y(n) - \overline{c_w} \quad \text{para } n = 1, 2, \dots, C \quad (4.40)$$

5 *Sistema de identificação de locutor*

O sistema de identificação pode ser dividido em dois grandes blocos, ou seja, o primeiro sendo responsável pela extração de vetores característicos, os quais contêm informações parametrizadas de cada locutor, e o segundo bloco com a finalidade de realizar as tarefas de identificação de padrões nas atividades de identificação e treinamento dos modelos de cada locutor. Vale observar que se considera que as locuções sejam sempre proferidas por um único locutor.

Os parâmetros empregados no sistema de identificação são os vetores formados pelos coeficientes cepstrais ou mel-cepstrais obtidos nas fases preliminares de segmentação e processamento do sinal de voz, os quais são utilizados para modelar o aparelho fonador. Estes vetores são necessários para gerar os modelos de cada um dos locutores, bem como servir de suporte para que o sistema possa identificar o respectivo locutor a partir de uma locução qualquer que lhe seja disponibilizada.

O algoritmo empregado na fase de treinamento garante que a cada iteração sejam obtidos modelos cada vez mais correlacionados com os vetores de coeficientes inicialmente disponibilizados ao sistema. Na fase de identificação é calculada a probabilidade de que determinado conjunto de vetores possa ter sido gerado por cada um dos modelos presentes no sistema. O locutor cujo modelo for o mais provável é então identificado.

5.1 **Modelo de mistura de gaussianas**

Os locutores são modelados como uma combinação linear de funções de densidade de probabilidade gaussiana [1], dando origem aos modelos de mistura de gaussianas (MMG). Cada componente da mistura é representada por um função de densidade de probabilidade dada pela equação (5.1), onde se adotou um modelo com G componentes. Vale observar que a dimensão C da função de densidade de probabilidade é a mesma que a

do vetor característico $\vec{x}$ produzido nas fases preliminares do processamento do sinal de voz, já considerando o descarte do primeiro coeficiente cepstral. A função de densidade de probabilidade para cada componente na equação (5.1) apresenta média $\vec{\mu}_i$ e matriz de covariância Σ_i .

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{C/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) \right\} \quad \text{para } i = 1, 2, \dots, G \quad (5.1)$$

A composição do modelo de locutor λ se finaliza com a inserção de um peso p_i para cada uma de suas componentes. O modelo de cada um dos locutores pode então ser descrito completamente conforme equação (5.2).

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad \text{para } i = 1, 2, \dots, G \quad (5.2)$$

Este modelo estatístico pode ser melhor entendido pela figura 28, onde um vetor $\vec{x}$ é inserido num modelo de mistura de gaussianas, resultando na probabilidade $p(\vec{x}|\lambda)$, a qual corresponde à probabilidade de ocorrência do vetor $\vec{x}$ considerando o modelo λ .

Figura 28: Modelo de mistura de gaussianas.

Cada modelo é obtido a partir de informações extraída de um número elevado de

quadros, o que faz com que este modelo GMM não represente nenhum quadro em particular, mas sim as características comuns a todos eles. É razoável considerar que estas características abstraídas pelo modelo representem um locutor.

A probabilidade de que um vetor de coeficientes $\vec{x}$ tenha sido gerado por determinado modelo λ é obtida conforme a equação (5.3), que define uma função de densidade de probabilidade. Nota-se que todas as componentes contribuem na formulação desta probabilidade, com maior ou menor intensidade, dependendo do valor de seu peso p_i .

$$p(\vec{x}|\lambda) = \sum_{i=1}^G p_i b_i(\vec{x}) \quad (5.3)$$

As funções de densidade de probabilidade definidas pelas equações (5.1) e (5.3) quando integradas no intervalo $[-\infty, \infty]$ apresentam resultado unitário. Assim, é possível afirmar que a soma dos pesos presentes no modelo da equação (5.2) também é unitária, conforme desenvolvimento da equação (5.4).

$$1 = \int_{-\infty}^{\infty} p(\vec{x}|\lambda) dx = \int_{-\infty}^{\infty} \sum_{i=1}^G p_i b_i(\vec{x}) dx = \sum_{i=1}^G p_i \int_{-\infty}^{\infty} b_i(\vec{x}) dx = \sum_{i=1}^G p_i \quad (5.4)$$

Para exemplificar, considerou-se o emprego de vetores unidimensionais e um modelo gaussiano de duas componentes, conforme figura 29. Nota-se que a componente representada pela densidade de probabilidade da linha tracejada apresenta valores maiores que a componente da linha contínua, embora a componente da linha tracejada tenha maior variância. Tal fato permite afirmar que o peso p_i da componente da linha tracejada é maior que da linha contínua, sendo mais relevante para o modelo em questão.

Considerando o modelo da figura 29, a composição de $p(\vec{x}|\lambda)$ a partir de um vetor de coeficientes mel-cepstrais correspondente a $\vec{x} = [22]$ seria fortemente influenciada pela componente de linha tracejada, enquanto que a contribuição da componente de linha contínua seria praticamente nula uma vez que para esta função de densidade de probabilidade mais de 99% de seus valores mais prováveis estão compreendidos no intervalo de $[0, 15]$. Em situações reais, são empregados modelos de 8 a 64 componentes ou até mais, dependendo da base de dados disponível, e vetores com mais de 15 coeficientes.

Figura 29: Função de densidade de probabilidade de mistura de gaussianas.

5.2 Treinamento

Os coeficientes cepstrais ou mel-cepstrais extraídos nas fases preliminares do processamento do sinal de voz são empregados para compor o conjunto de vetores característicos $X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T\}$ que contém as propriedades intrínsecas de cada locutor. Considerando para simplificação matemática a independência entre os vetores característicos $\vec{x}_i$ [1], a probabilidade de observação do conjunto X dado o modelo λ é representado pela equação (5.5). Vale observar que o desempenho do sistema não é comprometido pela simplificação supracitada.

$$p(X|\lambda) = p(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T|\lambda) = p(\vec{x}_1|\lambda)p(\vec{x}_2|\lambda) \dots p(\vec{x}_T|\lambda) = \prod_{t=1}^T p(\vec{x}_t|\lambda) \quad (5.5)$$

Durante o treinamento, aplica-se o método conhecido como estimação de máxima verossimilhança ou *maximum likelihood estimation (MLE)* através do qual se estimam os parâmetros do modelo definidos pela equação (5.2) de forma a maximizar a probabilidade de observação do conjunto de dados X , representada pela equação (5.5), ou seja, busca-se o modelo ótimo $\hat{\lambda}$ da equação (5.6).

$$\hat{\lambda} = \arg \max_{\lambda} p(X|\lambda) \quad (5.6)$$

A princípio, a resolução da equação $\frac{\partial p(X|\lambda)}{\partial \lambda} = 0$ forneceria os subsídios para a deter-

minação do modelo de máxima verossimilhança $p(X|\lambda)$. Contudo, a sua resolução direta não é factível dada a não-linearidade dos parâmetros do modelo λ . Assim, um procedimento alternativo na determinação do modelo λ é ajustá-lo a cada iteração de forma que o conjunto X se torne mais verossímil. Isto é possível com o emprego do algoritmo expectation-maximization (EM), o qual garante que o modelo atual $\lambda^{(i+1)}$ esteja mais correlacionado com o conjunto de observações X em comparação ao modelo da iteração anterior $\lambda^{(i)}$, ou seja, a cada iteração observa-se a propriedade definida pela equação (5.7), onde o parâmetro i identifica a iteração.

$$p(X|\lambda^{(i+1)}) \geq p(X|\lambda^{(i)}) \quad (5.7)$$

5.2.1 Algoritmo EM

Definindo a probabilidade de observação de um vetor $\vec{x}_t$ em determinada componente k de um modelo λ dado, chega-se à equação (5.8). A função b_k , presente na equação (5.1), depende do vetor $\vec{x}_t$, da média $\vec{\mu}_k$ e da variância Σ_k , portanto, $b_k = g(\vec{x}_t, \vec{\mu}_k, \Sigma_k)$. Vale observar que apenas os elementos da diagonal da matriz de covariância Σ_k são considerados, pois tal aproximação, além de facilitar o cálculo da inversa de Σ_k , melhora o rendimento do sistema de identificação [1]. Esta simplificação será abordada mais detalhadamente ao final desta subsecção.

$$p(k, \vec{x}_t|\lambda) = p_k b_k(\vec{x}_t) = p_k g(\vec{x}_t, \vec{\mu}_k, \Sigma_k) \quad (5.8)$$

Manipulando a relação definida pela equação (5.8) através do Teorema de Bayes, chega-se à equação (5.9).

$$p(k|\vec{x}_t, \lambda) = \frac{p(k, \vec{x}_t|\lambda)}{p(\vec{x}_t)} = \frac{p_k g(\vec{x}_t, \vec{\mu}_k, \Sigma_k)}{\sum_{j=1}^G p_j g(\vec{x}_t, \vec{\mu}_j, \Sigma_j)} \quad (5.9)$$

Observa-se que cada componente contribui de forma diferenciada para a composição da probabilidade definida pela equação (5.9), conforme explicado anteriormente no exemplo da figura 29. Considerando o efeito de todas as G componentes disponíveis para a composição do modelo λ , chega-se à relação da equação (5.10).

$$\sum_{k=1}^G p(k|\vec{x}_t, \lambda) = 1 \quad (5.10)$$

Definindo a log-verossimilhança $\Gamma(X|\lambda)$, obtida ao se tomar o logaritmo da função de verossimilhança da equação (5.5) e posterior expansão de seus termos conforme equação (5.3), chega-se à equação (5.11). Vale lembrar que equivalentemente o que se busca é um modelo λ que maximize o valor de $\Gamma(X|\lambda)$ a partir do conjunto de vetores X disponível.

$$\Gamma(X|\lambda) = \log \left(\prod_{t=1}^T \sum_{k=1}^G p_k b_k(\vec{x}_t) \right) = \sum_{t=1}^T \log \sum_{k=1}^G p_k g(\vec{x}_t, \vec{\mu}_k, \Sigma_k) \quad (5.11)$$

Como a função logarítmica é côncava, a desigualdade de Jensen garante que a relação definida pela equação (5.12) é verdadeira [19].

$$\log \left(\sum_{k=1}^K c_k a_k \right) \geq \sum_{k=1}^K c_k \log(a_k) \quad \text{para} \quad c_1 + c_2 + \dots + c_K = 1 \quad (5.12)$$

Substituindo $a_k = d_k/c_k$ na equação (5.12), obtém-se a equação (5.13).

$$\log \sum_{k=1}^K d_k = \log \left(\sum_{k=1}^K c_k \frac{d_k}{c_k} \right) \geq \sum_{k=1}^K c_k \log \left(\frac{d_k}{c_k} \right) \quad (5.13)$$

Retomando a equação (5.11) e substituindo $d_k = p_k g(\vec{x}_t, \mu_k, \Sigma_k)$ e $c_k = p(k|\vec{x}_t, \lambda)$, chega-se à relação (5.14) ao se aplicar a desigualdade da equação (5.13).

$$\sum_{t=1}^T \log \sum_{k=1}^G p_k g(\vec{x}_t, \vec{\mu}_k, \Sigma_k) \geq \sum_{t=1}^T \sum_{k=1}^G p(k|\vec{x}_t, \lambda) \log \left(\frac{p_k g(\vec{x}_t, \vec{\mu}_k, \Sigma_k)}{p(k|\vec{x}_t, \lambda)} \right) \quad (5.14)$$

Chamando de $c(\lambda)$ a parcela correspondente ao limite inferior da equação (5.14), chega-se à relação da equação (5.15).

$$c(\lambda) = \sum_{t=1}^T \sum_{k=1}^G p(k|\vec{x}_t, \lambda) \log(p_k g(\vec{x}_t, \vec{\mu}_k, \Sigma_k)) - \sum_{t=1}^T \sum_{k=1}^G p(k|\vec{x}_t, \lambda) \log(p(k|\vec{x}_t, \lambda)) \quad (5.15)$$

Maximizando a equação (5.15), é obtido um limite inferior mais elevado para a equação (5.14) e, conseqüentemente, surge um novo modelo λ que incrementa o valor de $\Gamma(X|\lambda)$. Utilizando o índice i para representar a iteração, se um modelo $\lambda^{(i)}$ obtido por este processo for reutilizado numa próxima etapa de maximização, surge um novo modelo $\lambda^{(i+1)}$ con-

forme previsto pela equação (5.7). Este processo se repete até que (5.16) se verifique.

$$\frac{\Gamma(X|\lambda^{(i+1)}) - \Gamma(X|\lambda^{(i)})}{\Gamma(X|\lambda^{(i)})} < \eta \quad (5.16)$$

A componente à direita da igualdade presente na equação (5.15) apresenta um valor fixo que depende exclusivamente do modelo atual, pois o termo $p(k|\vec{x}_t, \lambda)$ é calculado a partir dos parâmetros $(p_k^i, \mu_k^i, \sigma_k^i)$ presentes no modelo atual $\lambda^{(i)}$, conforme equação (5.9). Dessa forma, apenas a parcela à esquerda, explicitada pela equação (5.17), pode ser alterada com o intuito de maximizar a equação (5.15) e, conseqüentemente, disponibilizar os parâmetros de um novo modelo $\lambda^{(i+1)}$ em que $\Gamma(X|\lambda^{(i+1)}) \geq \Gamma(X|\lambda^{(i)})$.

$$\beta^{(i)}(\lambda) = \sum_{t=1}^T \sum_{k=1}^G p^{(i)}(k|\vec{x}_t, \lambda) \log(p_k g(\vec{x}_t, \vec{\mu}_k, \Sigma_k)) \quad (5.17)$$

Derivando a equação (5.17) com relação a cada um dos parâmetros do modelo $\lambda^{(i)}$ e igualando o resultado a zero, têm-se como resultados os parâmetros do novo modelo $\lambda^{(i+1)}$. Na equação (5.18) é tomada a derivada parcial de $\beta^{(i)}(\lambda)$ no ponto de sua j -ésima iteração com relação a $\vec{\mu}_j$. Durante esta derivação o somatório em k na equação (5.17) se anula para $k \neq j$.

$$\frac{\partial \beta^{(i)}(\lambda)}{\partial \vec{\mu}_j} = \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \frac{\partial}{\partial \vec{\mu}_j} \log(p_j g(\vec{x}_t, \vec{\mu}_j, \Sigma_j)) \quad (5.18)$$

Como p_j não é função de $\vec{\mu}_j$, tem-se que:

$$\frac{\partial \beta^{(i)}(\lambda)}{\partial \vec{\mu}_j} = \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \frac{\partial}{\partial \vec{\mu}_j} \log(g(\vec{x}_t, \vec{\mu}_j, \Sigma_j)) \quad (5.19)$$

Utilizando a formulação definida em (5.1) para a função de densidade de probabilidade $g(\vec{x}_t, \vec{\mu}_j, \Sigma_j)$, obtém-se:

$$\frac{\partial \beta^{(i)}(\lambda)}{\partial \vec{\mu}_j} = \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \frac{\partial}{\partial \vec{\mu}_j} \left(-\frac{1}{2} \log(|\Sigma_j|) - \frac{1}{2} (\vec{x}_t - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}_t - \vec{\mu}_j) \right) \quad (5.20)$$

Como Σ_j não é função de $\vec{\mu}_j$, tem-se que:

$$\begin{aligned}
\frac{\partial \beta^{(i)}(\lambda)}{\partial \vec{\mu}_j} &= -\frac{1}{2} \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \frac{\partial}{\partial \vec{\mu}_j} ((\vec{x}_t - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}_t - \vec{\mu}_j)) \\
&= \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \Sigma_j^{-1} (\vec{x}_t - \vec{\mu}_j)
\end{aligned} \tag{5.21}$$

Igualando a equação (5.21) a zero, obtém-se a média $\vec{\mu}_j^{(i+1)}$, conforme:

$$\vec{\mu}_j^{(i+1)} = \frac{\sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \vec{x}_t}{p^{(i)}(j|\vec{x}_t, \lambda)} \tag{5.22}$$

Os multiplicadores de Lagrange são empregados para encontrar o valor extremo de uma função considerando um conjunto de restrições. Dado um modelo de locutor, o somatório dos pesos de suas G componentes é unitário, o que representa uma restrição. Portanto, se ε for um multiplicador de Lagrange, é possível encontrar um valor extremo para $\beta^{(i)}$ conforme [20]:

$$\beta^{(i)}(\lambda) + \varepsilon \left(\sum_{k=1}^G p_k - 1 \right) = 0 \tag{5.23}$$

Derivando a equação (5.23) com relação a p_j , obtém-se:

$$\begin{aligned}
\frac{\partial \beta^{(i)}(\lambda)}{\partial p_j} + \varepsilon &= \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \frac{\partial}{\partial p_j} \log(p_j g(\vec{x}_t, \vec{\mu}_j, \sigma_j)) + \varepsilon \\
&= \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \left(\frac{\partial}{\partial p_j} \log(p_j) + \frac{\partial}{\partial p_j} g(\vec{x}_t, \vec{\mu}_j, \sigma_j) \right) + \varepsilon \\
&= \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \frac{1}{p_j} + \varepsilon = 0
\end{aligned} \tag{5.24}$$

Somando em j a equação (5.24) e considerando os resultados obtidos pelas equações (5.10) e (5.4), o valor de ε pode ser calculado conforme:

$$\begin{aligned}
\sum_{j=1}^G \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) + \sum_{j=1}^G p_j \varepsilon &= \sum_{t=1}^T \sum_{j=1}^G p^{(i)}(j|\vec{x}_t, \lambda) + \varepsilon \sum_{j=1}^G p_j \\
&= T + \varepsilon \implies \varepsilon = -T
\end{aligned} \tag{5.25}$$

Substituindo o resultado da equação (5.25) em (5.24), chega-se ao resultado para o parâmetro $p_j^{(i+1)}$.

$$p_j^{(i+1)} = \frac{1}{T} \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \quad (5.26)$$

A derivada da equação (5.17) com relação à inversa da matriz de covariância Σ_j^{-1} é dada pela equação (5.27). Vale observar que o operador $tr()$, presente na equação (5.27), corresponde ao traço da matriz.

$$\begin{aligned} \frac{\partial \beta^{(i)}(\lambda)}{\partial \Sigma_j^{-1}} &= \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \frac{\partial}{\partial \Sigma_j^{-1}} \log(p_j g(\vec{x}_t, \vec{\mu}_j, \Sigma_j)) \\ &= \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \frac{\partial}{\partial \Sigma_j^{-1}} \left(\frac{1}{2} \log(|\Sigma_j^{-1}|) - \frac{1}{2} (\vec{x}_t - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}_t - \vec{\mu}_j) \right) \\ &= \frac{1}{2} \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \frac{\partial}{\partial \Sigma_j^{-1}} (\log(|\Sigma_j^{-1}|) - tr(\Sigma^{-1}(\vec{x}_t - \vec{\mu}_j)(\vec{x}_t - \vec{\mu}_j)^T)) \end{aligned} \quad (5.27)$$

Para uma matriz simétrica A , são válidas as propriedades [20]:

$$\frac{\partial \log |A|}{\partial A} = 2A^{-1} - diag(A^{-1}) \quad (5.28)$$

$$\frac{\partial tr(AB)}{\partial A} = B + B^T - diag(B) \quad (5.29)$$

Substituindo $A = \Sigma_j^{-1}$ nas equações (5.28) e (5.29), além de considerar que $B_{jt} = (\vec{x}_t - \vec{\mu}_j)(\vec{x}_t - \vec{\mu}_j)^T$, tem-se que $B_{jt} = B_{jt}^T$, resultando em:

$$\frac{\log |\Sigma_j^{-1}|}{\partial \Sigma_j^{-1}} = 2\Sigma_j - diag(\Sigma_j) \quad (5.30)$$

$$\frac{\partial tr(\Sigma^{-1}(\vec{x}_t - \vec{\mu}_j)(\vec{x}_t - \vec{\mu}_j)^T)}{\partial \Sigma_j^{-1}} = 2B_{jt} - diag(B_{jt}) \quad (5.31)$$

Substituindo os resultados obtidos nas equações (5.30) e (5.31) na equação (5.27) e chamando $S = \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \{\Sigma_j - B_{jt}\}$, obtém-se:

$$\frac{\partial \beta^{(i)}(\lambda)}{\partial \Sigma_j^{-1}} = \sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \{2(\Sigma_j - B_{jt}) - diag(\Sigma_j - B_{jt})\} = (2S - diag(S)) \quad (5.32)$$

Igualando a equação (5.32) a zero, resulta que S corresponde a uma matriz nula, o

que permite obter:

$$\sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) \{\Sigma_j - B_{jt}\} = 0 \implies \Sigma_j^{(i+1)} = \frac{\sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) (\vec{x}_t - \vec{\mu}_j) (\vec{x}_t - \vec{\mu}_j)^T}{\sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda)} \quad (5.33)$$

O cálculo da inversa da matriz Σ_j na equação (5.1) tem um elevado custo computacional uma vez que Σ_j é uma matriz cheia. Observa-se que anulando os elementos de Σ_j que não pertencem à sua diagonal principal, não há redução de desempenho no sistema de identificação quando comparado à situação em que se adota a matriz Σ_j cheia [21], além de simplificar o formato da equação (5.33). Para justificar a alteração da equação (5.33) conforme (5.39), será adotada a suposição de que $\vec{x}_t = [x_{t,1} \ x_{t,2}]^T$ e $\vec{\mu} = [\mu_{j,1} \ \mu_{j,2}]^T$ na equação (5.33) sem que se perca a generalidade dos resultados assim obtidos, conforme desenvolvimento que segue:

$$\begin{aligned} (\vec{x}_t - \vec{\mu}_j)(\vec{x}_t - \vec{\mu}_j)^T &= \begin{pmatrix} (x_{t,1} - \mu_{j,1})^2 & (x_{t,1} - \mu_{j,1})(x_{t,2} - \mu_{j,2}) \\ (x_{t,1} - \mu_{j,1})(x_{t,2} - \mu_{j,2}) & (x_{t,2} - \mu_{j,2})^2 \end{pmatrix} \\ &= \begin{pmatrix} x_{t,1}^2 + \mu_{j,1}^2 - 2x_{t,1}\mu_{j,1} & (x_{t,1} - \mu_{j,1})(x_{t,2} - \mu_{j,2}) \\ (x_{t,1} - \mu_{j,1})(x_{t,2} - \mu_{j,2}) & x_{t,2}^2 + \mu_{j,2}^2 - 2x_{t,2}\mu_{j,2} \end{pmatrix} \end{aligned} \quad (5.34)$$

Considerando a simplificação em que a média dos termos que não pertencem à diagonal principal da matriz da equação (5.34) são nulos, ou seja, $E(\vec{x}_t - \vec{\mu}_j) = 0$, obtém-se a equação (5.35) particularizada para os componentes de índice 1 e sem perda de generalidade. Multiplicando a equação (5.35) pela constante $\mu_{j,1}$ resulta em:

$$E(x_{t,1} - \mu_{j,1}) = 0 \Rightarrow E(x_{t,1}) = E(\mu_{j,1}) \quad (5.35)$$

$$\mu_{j,1}E(x_t) = \mu_{j,1}E(\mu_{j,1}) \Rightarrow E(\mu_{j,1}x_{t,1}) = E(\mu_{j,1}^2) \quad (5.36)$$

Partindo do primeiro termo da diagonal da equação (5.34) e utilizando o resultado obtido pela equação (5.36), chega-se a:

$$E\{(x_{t,1} - \mu_{j,1})^2\} = E(x_{t,1}^2 + \mu_{j,1}^2 - 2x_{t,1}\mu_{j,1}) = E(x_{t,1}^2 - \mu_{j,1}^2) \quad (5.37)$$

Considerando os resultados obtidos a partir da simplificação de se considerar nulos os termos da diagonal principal da matriz de covariância Σ_j , a equação (5.34) pode ser reescrita conforme:

$$(\vec{x}_{t,1} - \vec{\mu}_{j,1})(\vec{x}_{t,1} - \vec{\mu}_{j,1})^T = \begin{pmatrix} x_{t,1}^2 - \mu_{j,1}^2 & 0 \\ 0 & x_{t,2}^2 - \mu_{j,2}^2 \end{pmatrix} \quad (5.38)$$

Como os valores da média μ_j são os mesmos independentemente do valor assumido pela variável t na equação (5.33), chega-se à simplificação exibida na equação (5.39), onde x_t e μ_j referem-se aos elementos arbitrários dos vetores $\vec{x}_t$ e $\vec{\mu}_j$. O vetor $\vec{\sigma}_j^{(i+1)}$ corresponde aos elementos da diagonal principal da matriz $\Sigma_j^{(i+1)}$.

$$\vec{\sigma}_j^{(i+1)} = \frac{\sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda) x_t^2}{\sum_{t=1}^T p^{(i)}(j|\vec{x}_t, \lambda)} - \mu_j^2 \quad (5.39)$$

Resumidamente, o algoritmo de treinamento compreende o cálculo das grandezas explicitadas pelas equações (5.9), (5.22), (5.26), (5.39) e (5.11) respectivamente. A cada iteração é obtido um novo conjunto de parâmetros do modelo λ até que seja verdadeira a condição estabelecida pela equação (5.16), quando o algoritmo chega ao seu fim.

5.2.2 Inicialização

São consideradas duas formas de inicialização do modelo de mistura de gaussianas definido pela equação (5.2). Em uma delas as componentes da mistura são formadas a partir de vetores de coeficientes mel-cepstrais escolhidos aleatoriamente e, na outra, as componentes são construídas a partir dos centróides obtidos com a utilização do algoritmo LBG [22].

Na primeira forma, os parâmetros iniciais do modelo de mistura de gaussianas definido na equação (5.2) são obtidos escolhendo aleatoriamente G vetores $\vec{v}_{A_i}$ de coeficientes mel-cepstrais e verificadas as distâncias mínimas entre estes vetores e os remanescentes $\vec{v}_j$, extraídos do sinal de voz de determinado locutor que foi disponibilizado para o processo de treinamento, conforme a equação (5.40). Em seguida, os vetores remanescentes são agrupados com o vetor $\vec{v}_{A_i}$ que lhe for mais próximo, formando G conjuntos de vetores.

$$d(\vec{v}_j) = \arg \min_{1 \leq i \leq G} \|\vec{v}_{A_i} - \vec{v}_j\| \quad (5.40)$$

Utilizando o critério de proximidade, formam-se conjuntos que agrupam quantidades diferentes de vetores. Os parâmetros p_i da equação (5.2) são obtidos calculando-se a razão entre o número de vetores presentes em cada conjunto pelo número total de vetores. Os parâmetros $\vec{\mu}_i$ correspondem à média dos vetores presentes em cada conjunto e são obtidos pela equação (5.41), onde V_i é o número de vetores presentes no conjunto i .

$$\vec{\mu}_i = \frac{1}{V_i} \sum_{n=1}^{V_i} \vec{x}_{(i,n)} \quad \text{para} \quad i = 1, 2, \dots, G \quad (5.41)$$

Para determinada componente, o vetor de covariância $\vec{\sigma}_i^2$ é iniciado utilizando a média obtida pela equação (5.41) e os vetores $\vec{x}_i$ agrupados para esta componente, conforme equação (5.42).

$$\vec{\sigma}_i^2 = \frac{1}{V_i - 1} \sum_{n=1}^{V_i} (\vec{x}_{(i,n)} - \vec{\mu}_i)^2 \quad \text{para} \quad i = 1, 2, \dots, G \quad (5.42)$$

A outra forma de iniciar o modelo é utilizando o algoritmo LBG-VQ de quantização vetorial. A primeira etapa deste algoritmo consiste em localizar o centro de massa $\vec{c}_0$ dos vetores de coeficientes cepstrais conforme equação (5.43), onde T corresponde ao número total de vetores disponíveis.

$$\vec{c}_0 = \frac{1}{N} \sum_{t=1}^T \vec{x}_t \quad (5.43)$$

Sucessivamente são obtidos novos centros cujas coordenadas são fruto da multiplicação dos centros já disponíveis por uma constante η conforme equação (5.44). Vale observar que a constante η é arbitrária, sendo muito pequena com relação à unidade.

$$\begin{aligned} \vec{c}_1 &= (1 + \eta)\vec{c}_0 \\ \vec{c}_2 &= (1 - \eta)\vec{c}_0 \end{aligned} \quad (5.44)$$

A partir dos centros obtidos na equação 5.44 são formados conjuntos que agregam vetores de coeficientes menos distantes a cada um destes centros. Os centros de massas destes conjuntos recém-formados são localizados e divididos, conforme equação 5.44, reiniciando o processo de fragmentação dos agrupamentos. Assim que todos os centróides são obtidos, utiliza-se a equação (5.42) para o cálculo do vetor de variância de cada componente do

modelo.

A título de exemplo, considerando 1000 amostras bidimensionais linearmente distribuídas entre 0 e 1, após a 3ª iteração do algoritmo LBG seriam formados os 8 conjuntos identificados por cor na figura 30, cujos centróides aparecem destacados.

Figura 30: Conjuntos formados com a aplicação do algoritmo LBG sobre um conjunto de vetores bidimensionais.

5.3 Identificação

Quando uma locução é inserida no sistema de identificação, deseja-se descobrir o modelo de locutor mais verossímil. Para que isto seja possível, antes é necessário que esta locução seja submetida aos mesmos procedimentos de pré-processamento do sinal de voz que foram empregados na fase de treinamento, dando origem ao conjunto de vetores característicos $X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T\}$ composto por coeficientes cepstrais ou mel-cepstrais. Supondo que haja S modelos de locutores disponíveis, o que se deseja é identificar o modelo de locutor λ_k , onde $1 \leq k \leq S$, que tenha a máxima probabilidade *a posteriori* diante do conjunto de observações X , conforme a equação (5.45).

$$R = \arg \max_{1 \leq k \leq S} p(\lambda_k | X) \quad (5.45)$$

Aplicando a regra de Bayes à equação (5.45), é obtida a equação (5.46). Assumindo que todos os modelos são equiprováveis, pode-se considerar que $p(\lambda_k) = 1/S$. Outra simplificação surge do fato de que $p(X)$ apresenta o mesmo valor independentemente do modelo de locutor analisado. Portanto, a razão $p(\lambda_k)/P(X)$ é uma constante, não sendo

relevante para o cálculo de R na equação (5.46), dando origem à variável $\hat{R}$ na qual estes fatores constantes são desconsiderados.

$$R = \arg \max_{1 \leq k \leq S} p(\lambda_k | X) = \arg \max_{1 \leq k \leq S} p(X | \lambda_k) \frac{p(\lambda_k)}{p(X)} \Rightarrow \hat{R} = \arg \max_{1 \leq k \leq S} p(X | \lambda_k) \quad (5.46)$$

Considerando a independência das observações, a equação (5.46) se transforma em:

$$\hat{R} = \arg \max_{1 \leq k \leq S} \prod_{t=1}^W p(\vec{x}_t | \lambda_k) \quad (5.47)$$

Como na equação (5.47) estão sendo multiplicadas propabilidades cujos valores são necessariamente positivos e menores do que 1, o resultado deste produtório tende a se reduzir rapidamente com a elevação do número de observações W . Considerando limitações de natureza computacional para o tratamento de números muito reduzidos, pode-se, alternativamente, ser tomado o logaritmo da equação (5.47) [1], dando origem à equação (5.48), onde o valor de $p(\vec{x}_t | \lambda_k)$ é fornecido pela equação (5.3).

$$\hat{S} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^W \log p(\vec{x}_t | \lambda_k) \quad (5.48)$$

No processo de identificação, o conjunto de amostras é dividido em segmentos contendo W observações, sendo que segmentos consecutivos estão deslocados de uma observação. Como exemplo, na equação (5.49) estão exibidos os dois primeiros segmentos obtidos a partir de um conjunto de observações.

$$\begin{array}{c} \text{segmento 1} \\ \overbrace{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_W} \\ \vec{x}_1, \vec{x}_2, \dots, \vec{x}_W, \vec{x}_{W+1}, \vec{x}_{W+2}, \dots \\ \text{segmento 2} \\ \overbrace{\vec{x}_2, \dots, \vec{x}_{W+1}} \\ \vec{x}_1, \vec{x}_2, \dots, \vec{x}_{W+1}, \vec{x}_{W+2}, \vec{x}_{W+3}, \dots \end{array} \quad (5.49)$$

Para cada segmento é identificado o modelo de locutor que gera o valor de $\hat{S}$, definido pela equação (5.48). Executando este procedimento para todos os segmentos disponíveis, o modelo mais freqüentemente identificado é aquele apontado pelo sistema de identificação como o mais provável por produzir a locução inicialmente disponibilizada.

O número W de observações de cada segmento é obtido experimentalmente, ou seja,

disponibiliza-se um sinal de voz cujo locutor é conhecido e varia-se o valor de W de forma a maximizar o número de segmentos corretamente identificados. Este processo é repetido com todos os locutores até que se obtenha um valor médio de W otimizado para as condições de operação do sistema.

5.3.1 Identificação por distâncias mínimas

Alternativamente pode-se realizar a identificação considerando as distâncias mínimas entre os vetores de coeficientes cepstrais e os vetores de médias dos modelos gerados. Inicialmente são formados conjuntos em torno dos vetores de média que são populados com os vetores de coeficientes. O critério utilizado para povoar estes conjuntos é dado pela maior proximidade dos vetores de coeficientes com os vetores de média. A soma das distâncias entre os vetores média e os seus vetores de coeficientes agregados fornece um índice I_k , conforme equação (5.50), onde $\mu = \{\vec{\mu}_1, \vec{\mu}_2, \dots, \vec{\mu}_G\}$ corresponde aos vetores de média do modelo λ_k e $X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T\}$, aos vetores de coeficientes.

$$I_k = \sum_{\vec{x}_t \in X} \log \left(\min_{\vec{\mu}_i \in \mu} \|\vec{x}_t - \vec{\mu}_i\|^2 \right)_{\lambda_k} \quad (5.50)$$

Para cada um dos modelos $\lambda = \{\lambda_1, \dots, \lambda_k\}$ é calculado o índice I e aquele de menor valor identifica o modelo mais correlacionado com a locução de teste. Outro ponto a ser observado é que no cálculo de I não é utilizada a informação de variância e peso das componentes do modelo, sendo relevante apenas as médias.

6 *Avaliação experimental*

6.1 Considerações iniciais

Para realizar o treinamento dos modelos de locutor, bem como os testes de identificação, foram empregados os bancos de vozes na língua inglesa TIMIT e NTIMIT. Estes bancos são formados por 8 diretórios, “dr1” a “dr8”, e em cada um desses diretórios há uma série de subdiretórios para cada locutor, ou seja, em cada subdiretório há 10 locuções de um locutor específico. Os diretórios foram divididos de acordo com a região em que as locuções foram extraídas nos Estados Unidos da América. Na tabela 3 está destacado o número de subdiretórios disponíveis por diretório, perfazendo um total de 462 subdiretórios ou diferentes locutores dentro do diretório de treino destes bancos.

diretório	número de subdiretórios	região
dr1	38	New England
dr2	76	Northern
dr3	76	North Midland
dr4	68	South Midland
dr5	70	Southern
dr6	35	New York City
dr7	77	Western
dr8	22	Army Brat

Tabela 3: Estrutura de subdiretórios dos bancos de vozes TIMIT e NTIMIT.

Os arquivos presentes em cada um destes subdiretórios contêm informações de um único locutor e são identificados de uma forma lógica, ou seja, seus nomes se iniciam pelas siglas “sa”, “si” ou “sx”. Há sempre 2 arquivos que se iniciam por “sa”, 5 por “sx” e 3 por “si”, perfazendo um total de 10 arquivos por locutor ou subdiretório. Os arquivos iniciados por “sa” correspondem a locuções obtidas de frases sempre iguais independentemente do locutor, permitindo que se tenha um padrão ao adotá-las como frases de teste na fase de identificação. Os demais arquivos iniciados por “si” ou “sx” correspondem a frases diversas colhidas de cada locutor, sendo empregados prioritariamente para o treinamento dos modelos. O banco de voz TIMIT apresenta locuções com elevada relação sinal-ruído,

enquanto que no NTIMIT estão presentes as mesmas locuções do banco TIMIT quando transmitidas por um canal telefônico.

O número ideal de componentes utilizadas na composição dos modelos de cada locutor não pode ser determinado analiticamente. Contudo, quanto maior for a disponibilidade de locuções para a realização dos treinamentos, maior deve ser o número de componentes a ser empregado [1] para que se obtenham melhores índices de correta identificação.

Locuções mais extensas garantem um maior refinamento na modelagem do aparelho fonador uma vez que são disponibilizados mais coeficientes mel-cepstrais ao sistema de treinamento. Este, por sua vez, quando utiliza modelos mais complexos formado por um elevado número de componentes traz maior complexidade ao sistema para que este possa se adaptar mais adequadamente aos coeficientes mel-cepstrais que lhe deram origem.

A combinação de modelos com elevado número de componentes e pouco material de treinamento tende a reduzir o desempenho do sistema de identificação, pois os modelos assim gerados acabam por apresentar uma correspondência grosseira com o respectivo locutor. Neste caso é preferível reduzir o número de componentes dos modelos.

Quanto maior for o número de modelos de locutores, menor será o índice de correta identificação uma vez que o sistema visualizará uma gama maior de escolhas, elevando a probabilidade de falha.

Outro aspecto relevante a ser observado é que as condições de treinamento dos modelos devem ser replicadas para a situação de teste do sistema. Observa-se que se isto não acontecer, há uma significativa degradação dos índices de correta identificação.

6.2 Procedimentos de simulação

Para todos os testes realizados, convencionou-se adotar os arquivos de locução iniciados por “si” e “sx” para o treinamento dos modelos e aqueles iniciados por “sa”, para realizar os procedimentos de teste de identificação. Como as locuções têm curta duração, geralmente não ultrapassando 3 s, estas foram agregadas umas às outras para formar locuções mais longas de acordo com as condições de treinamento e teste previstos. A forma como se deu esta composição variou para cada uma das simulações realizadas, mas em cada um dos casos foi destacada a metodologia empregada.

Os coeficientes mel-cepstrais, necessários para compor os vetores característicos, são obtidos empregando banco de filtros ou a partir dos coeficientes de predição linear. Na

figura 31, é exibido o diagrama esquemático para os procedimentos de treinamento quando empregado banco de filtros. Já na figura 32, está esquematizado o procedimento de treinamento utilizando predição linear. A escolha pela aplicação de SMC é feita pelo usuário do sistema para a situação em que se está utilizando amostras do banco NTIMIT. Como pode ser notado, há uma série de processos comuns às duas metodologias, divergindo apenas os passos posteriores ao DAV e anteriores à SMC.

Figura 31: Sistema de identificação empregando banco de filtros para derivar os MFCCs.

Figura 32: Sistema de identificação empregando predição linear para derivar os MFCCs.

Na fase de identificação, uma locução de teste é inserida no sistema e verifica-se qual dos modelos de locutor apresenta a maior probabilidade de tê-la gerado. O algoritmo

de identificação deve ser totalmente compatível com o de treinamento, ou seja, se for empregada a técnica de banco de filtros com SMC no treinamento, exatamente o mesmo deve ser feito para a identificação. Outro fator relevante diz respeito à compatibilidade entre as características da base de dados utilizada na identificação e treinamento. Se, por exemplo, os modelos forem obtidos a partir do banco TIMIT e o teste de identificação se der a partir de amostras retiradas do banco NTIMIT, o desempenho do sistema será pior do que se o treinamento acontecer diretamente a partir do banco NTIMIT, embora este último apresente uma relação sinal-ruído de suas amostras menor do que aquela encontrada no banco TIMIT.

Na tabela 4 estão relacionados os valores das constantes utilizadas para os procedimentos de treinamento e identificação. Vale ressaltar que nesta tabela também estão assinaladas as correspondentes equações onde tais constantes surgem ao longo deste texto.

Geralmente, uma representação cepstral a partir parâmetros de predição é feita fazendo com o número de coeficientes LPC seja correspondente a aproximadamente $2/3$ do número de coeficientes cepstrais [23]. Assim, na tabela 4 procurou-se utilizar o número de coeficientes de acordo com essa premissa, ou seja, para 23 MFCCs foram empregados 15 LPCs.

constante	valor	equação	definição
a	0,97	(2.1)	constante de preênfase
α_s	0,8	(2.8)	constante do detector de voz
U	8	(2.10)	constante do detector de voz
V	15	(2.10)	constante do detector de voz
F_{min}	500	(2.13)	constante do detector de voz
F_{max}	3400	(2.14)	constante do detector de voz
F_s	16000	(2.10)	frequência de amostragem (Hz)
M	512	(2.10)	número de amostras por quadro
N	128	(2.10)	número de amostras de separação entre quadros adjacentes
p	15	(3.3)	número de coeficientes de predição
C	23	(4.34)	número de coeficientes mel-cepstrais
ε	0,0001	(5.16)	constante de parada de treinamento
η	0,01	(5.44)	constante de inicialização do LBG
W	160	(5.49)	observações por segmento

Tabela 4: Constantes adotadas durante as simulações.

6.2.1 Modelo com número fixo de componentes

Os dados para as simulações desta seção foram obtidos das locuções disponíveis nos 22 subdiretórios do diretório “dr8” dos bancos de vozes TIMIT e NTIMIT. Os arquivos

iniciados por “sx” e “si” de cada locutor foram compostos para formar um único arquivo de aproximadamente 23 s de amostras e, então, utilizados para o treinamento dos modelos. De forma análoga, aqueles iniciados por “sa” foram empregados como locução de teste com duração aproximada de 5,5 s. Adotou-se um número fixo de 16 componentes para todas as simulações desta seção, bem como o algoritmo de identificação da seção 5.3 para as simulações em que não foi explicitamente declarado o uso de algoritmo da seção 5.3.1.

Alterações nas condições ambientais foram simuladas empregando-se alternadamente os bancos TIMIT e NTIMIT, enquanto que o conjunto de vetores característicos, formado pelos coeficientes mel-cepstrais, foi obtido a partir de banco de filtros, conforme seção 4.3.1, ou diretamente dos coeficientes de predição linear, conforme seção 4.3.2. Paralelamente averiguou-se o impacto da subtração da média cepstral abordada na seção 4.3.3.

Os modelos foram iniciados utilizando LBG, conforme seção 5.2.2. Para cada treinamento o valor de η foi ligeiramente alterado, permanecendo entre 0,010 e 0,009. Para cada uma das simulações foi gerada uma tabela com os dados de teste, através da qual foi apontado o percentual de locutores identificados corretamente. Como há um grau de aleatoriedade na inicialização do processo de treinamento, notam-se diferenças não significativas nos níveis de desempenho do identificador quando se empregam modelos obtidos de diferentes simulações. Dessa forma, foram realizadas várias simulações para cada um dos casos de teste, sendo apontado na tabela 5 um valor característico de desempenho.

Na coluna “simulação” da tabela 5 adotou-se a seguinte codificação para melhor sintetizar os casos de teste:

- TIMIT : emprego do banco de voz TIMIT
- NTIMIT : emprego do banco de voz NTIMIT
- FB : coeficientes mel-cepstrais obtidos a partir de banco de filtros
- LP : coeficientes mel-cepstrais obtidos a partir de predição linear
- SMC : uso da técnica de subtração da média cepstral
- DM : uso do algoritmo de distâncias mínimas da seção 5.3.1

Como pode ser constatado pelos resultados exibidos na tabela 5, ao se utilizar o banco de voz TIMIT são obtidos melhores resultados quando os coeficientes mel-cepstrais são derivados de banco de filtros. Com o banco NTIMIT é mais indicado o uso de predição linear como meio de derivação desses coeficientes.

simulação	% acerto
TIMIT+FB	100 %
TIMIT+LP	88 %
NTIMIT+FB	61 %
NTIMIT+LP	76 %
NTIMIT+FB+SMC	68 %
NTIMIT+LP+SMC	77 %
NTIMIT+FB+SMC+DM	61 %

Tabela 5: Resultados típicos de desempenho para modelos com 16 gaussianas utilizando amostras dos bancos TIMIT e NTIMIT.

Outro fato observado é que a técnica de subtração da média cepstral, quando utilizada com amostras do banco NTIMIT, fez com que houvesse um melhora mais significativa de desempenho nos casos em que os coeficientes mel-cepstrais foram derivados de banco de filtros.

6.2.2 Modelo com número de componentes variável

Nas simulações desta e das subsequentes seções, os modelos de misturas de gaussianas foram iniciados utilizando vetores aleatórios, conforme definido na seção 5.2.2.

A partir do banco de voz NTIMIT diretório “dr8”, empregou-se a composição das locuções iniciadas por “sx” e “si” para realizar o treinamento dos modelos e a composição das iniciadas por “sa”, para a identificação. Foram adotados os coeficientes de predição linear para originar os coeficientes mel-cepstrais, além de ser utilizada a técnica de subtração da média cepstral. Os modelos foram iniciados de forma aleatória, conforme seção 5.2.2.

Como não é conhecido um método algébrico de determinação do número de componentes que melhor se ajuste às diversas condições de operação do sistema de identificação, este valor deve ser obtido experimentalmente a partir de simulações que se aproximem ao máximo das condições em que o sistema será mais rotineiramente empregado. Vale observar que deve ser priorizado o uso de modelos com o menor número de componentes possível, pois são mais compactos, exigindo menos recursos de armazenamento de informações, bem como de processamento para a execução dos algoritmos intrínsecos às tarefas de identificação e treinamento.

É possível confirmar o impacto do número excessivo ou reduzido de componentes sobre o desempenho do sistema diante do material disponibilizado para o treinamento e identificação a partir dos resultados exibidos pela tabela 6. Quando se utiliza um número

excessivo de componentes diante do material disponível para o treinamento, observa-se uma queda de performance de identificação por não haver informação suficiente para treinar adequadamente um modelo com tal grau de complexidade, sendo necessário reduzir o número de componentes para atingir o desempenho ótimo de identificação.

componentes	correto	incorreto	% acerto
2	4	18	18 %
4	13	9	59 %
8	16	6	73 %
16	17	5	77 %
24	20	2	91 %
32	20	2	91 %
48	20	2	91 %
64	19	3	86 %
80	12	10	55 %

Tabela 6: Simulação com NTIMIT, coeficientes derivados de predição linear com SMC e número variável de componentes.

Pelos resultados exibidos na tabela 6, o modelo a ser adotado seria o de 24 componentes por ser mais compacto e apresentar desempenho tão elevado quanto outros modelos que são compostos por um número maior de componentes.

6.2.3 Locuções com tamanho variável

Alterando-se a duração das locuções, há uma conseqüente modificação no desempenho do sistema. Pela tabela 7 é observado o desempenho do sistema quando se adota um modelo com 24 componentes, onde os coeficientes mel-cepstrais são derivados dos coeficientes de predição linear e uso de SMC.

Foram utilizadas as locuções oriundas do diretório “dr8” do banco de voz NTIMIT. No treinamento adotaram-se aquelas iniciadas por “si” e/ou “sx”, enquanto que na identificação foram empregadas as locuções iniciadas por “sa1” e/ou “sa2”. Os valores reportados na tabela 7 correspondem a valores médios de duração dos arquivos supracitados, os quais foram obtidos após análise de cada um dos locutores disponíveis. Como pode ser confirmado pela tabela 7, quanto maior for o tempo da locução inserida no sistema, tanto na fase de treinamento quanto de identificação, melhor será o percentual de acerto do sistema.

Como foi empregado um modelo com um número fixo de componentes na confecção da tabela 7, os resultados obtidos não são os melhores possíveis haja vista que uma melhor

locução	treinamento	locução	identificação	% acerto
si	9,28 s	sa1	3,11 s	32 %
sx	13,67 s	sa1	3,11 s	45 %
si+sx	22,96 s	sa1	3,11 s	55 %
si	9,28 s	sa1+sa2	5,76 s	50 %
sx	13,67 s	sa1+sa2	5,76 s	77 %
si+sx	22,96 s	sa1+sa2	5,76 s	91 %

Tabela 7: Efeito da variação do tempo de treinamento e de identificação para o desempenho do sistema utilizando-se amostras do banco NTIMIT.

otimização só seria possível se o número de componentes se adequasse às condições de operação do sistema de identificação, conforme citado na seção 6.2.2.

6.2.4 Número de locutores variável

Aumentando-se o número de locutores, a performance do sistema de identificação tende a se reduzir progressivamente diante da elevação de possibilidades de modelos a serem escolhidos. Com o intuito de verificar este impacto, foi empregado o banco TIMIT com um modelo de 24 componentes, sendo adotadas ambas as técnicas de obtenção dos coeficientes mel-cepstrais, ou seja, a partir de banco de filtros e dos coeficientes de predição linear. Vale observar que não há necessidade de se empregar a técnica da subtração cepstral neste caso, pois o banco TIMIT não apresenta distorção de canal como no caso do NTIMIT. As locuções iniciadas por “si” e “sx” foram compostas para formar uma única locução, sendo empregadas para o treinamento dos modelos. Aquelas iniciadas por “sa” também foram compostas para serem utilizadas nos testes de identificação.

Inicialmente foram obtidos os modelos para cada um dos locutores presentes nos diretórios “dr1” a “dr8” do banco TIMIT. Segregando estes modelos por diretório, foi executado o processo de identificação dando origem aos dados das tabelas 8 e 9 de acordo com a técnica de extração dos coeficientes mel-cepstrais empregada.

A figura 33, traçada a partir dos dados da tabela 10, permite observar comparativamente o efeito do número de locutores no percentual de acerto do sistema de identificação. Os modelos do banco TIMIT foram obtidos com a adoção de banco de filtros, enquanto que para o NTIMIT, utilizou-se banco de filtros e predição linear, ambos com subtração da média cepstral.

Em determinadas simulações observou-se uma inversão de desempenho esperado para o sistema, ou seja, para um número mais elevado de locutores foi obtido um resultado

diretório	correto	incorreto	% acerto
dr1	38	0	100 %
dr2	76	0	100 %
dr3	76	0	100 %
dr4	68	0	100 %
dr5	69	1	98 %
dr6	35	0	100 %
dr7	77	0	100 %
dr8	22	0	100 %

Tabela 8: Desempenho do sistema em cada diretório do banco TIMIT com MFCCs derivados de banco de filtros.

diretório	correto	incorreto	% acerto
dr1	28	10	73 %
dr2	54	22	71 %
dr3	39	37	51 %
dr4	46	22	68 %
dr5	49	21	70 %
dr6	31	4	88 %
dr7	51	26	66 %
dr8	21	1	95 %

Tabela 9: Desempenho do sistema em cada diretório do banco TIMIT com MFCCs derivados de predição linear.

modelos com 24 gaussianas		TIMIT	NTIMIT	NTIMIT
		filtro	predição + SMC	filtro + SMC
diretórios	locutores	% acerto	% acerto	% acerto
dr1	38	100 %	50 %	45 %
dr2	76	100 %	41 %	37 %
dr3	76	100 %	45 %	46 %
dr4	68	100 %	62 %	54 %
dr5	70	98 %	50 %	41 %
dr6	35	100 %	66 %	48 %
dr7	77	100 %	52 %	47 %
dr8	22	100 %	91 %	77 %
dr1+dr2	114	100 %	36 %	34 %
dr2+dr3	152	100 %	36 %	26 %
dr2+dr4	144	98 %	38 %	27 %
dr3+dr5	146	99 %	32 %	29 %
dr3+dr4+dr7	221	99 %	31 %	27 %
dr2+dr4+dr5	214	98 %	30 %	22 %
dr2+dr3+dr6+dr8	209	99 %	30 %	21 %
dr2+dr3+dr5+dr7	299	99 %	20 %	19 %
todos	472	97 %	17 %	13 %

Tabela 10: Efeito no desempenho do sistema de acordo com o número de locutores e da forma como os MFCCs são obtidos dos bancos de vozes TIMIT e NTIMIT.

melhor do que para um número menor de locutores. Isso acontece porque a inicialização dos modelos no processo de treinamento apresenta uma característica aleatória que dá origem a modelos com características distintas para as mesmas condições de simulação, conforme detalhado na seção 5.2.2.

Figura 33: Visualização do desempenho do sistema de acordo com o número de locutores e da forma como os MFCCs são obtidos dos bancos de vozes TIMIT e NTIMIT.

6.2.5 Alternando o banco de voz no treinamento e identificação

Para verificar o efeito de diferentes configurações nas etapas de treinamento e identificação, foram realizadas as simulações das tabelas 11 e 12, onde os coeficientes mel-cepstrais foram derivados respectivamente de banco de filtros e de predição linear. Foram utilizadas as locuções do diretório dr8 com 22 locutores e modelos de 24 componentes, sendo que os arquivos empregados para o treinamento foram aqueles iniciados por “si” e “sx”, enquanto que na identificação foram adotados os arquivos iniciados por “sa”.

treinamento → identificação ↓	TIMIT	NTIMIT	NTIMIT+SMC
TIMIT	100 %	4 %	0%
NTIMIT	14 %	68%	4%
NTIMIT+SMC	4 %	4%	77 %

Tabela 11: Efeito da compatibilidade das locuções utilizadas nos treinamentos e nos testes de identificação com MFCCs derivados de banco de filtros.

treinamento → identificação ↓	TIMIT	NTIMIT	NTIMIT+SMC
TIMIT	95%	4%	4%
NTIMIT	4%	86%	9%
NTIMIT+SMC	18%	9%	91%

Tabela 12: Efeito da compatibilidade das locuções utilizadas nos treinamentos e nos testes de identificação com MFCCs derivados de predição linear.

6.2.6 Alternando o número de coeficientes mel-cepstrais

Ao longo das simulações foram empregados invariavelmente 23 coeficientes mel-cepstrais. Para verificar o efeito de se alterar concomitantemente o número desses coeficientes e de componentes que compõem os modelos de cada locutor, foram realizadas simulações empregando MFCCs derivados de banco de filtros com locuções oriundas do banco TIMIT e do diretório “dr7” com 77 locutores.

Os resultados destas simulações encontram-se registrados na tabela 13, onde os arquivos empregados para o treinamento foram aqueles iniciados por “si” e “sx”, enquanto que na identificação foram adotados os arquivos iniciados por “sa”.

gaussianas → MFCC ↓	8	24	64
4	14%	18%	26%
8	70%	77%	82%
16	93%	97%	99%
24	97%	99%	100%
32	99%	100%	99%
40	97%	100%	100%

Tabela 13: Desempenho do sistema de acordo com o número de gaussianas e MFCCs adotados.

Como pode ser confirmado pelos resultados exibidos na tabela 13, a partir de 24 MFCCs não houve alteração significativa no desempenho do sistema de identificação. Considerando que a complexidade computacional para a elaboração dos modelos está diretamente relacionada ao número de MFCCs, conclui-se que não há vantagem em se utilizar mais do que 24 MFCCs.

O modelo mais complexo formado de 64 componentes se mostrou mais promissor do que o de 8 e 24 componentes enquanto se utilizou um número reduzido de MFCCs. A partir de 24 MFCCs houve grande similaridade de desempenho para os modelos de 24 e 64 componentes, confirmando que o número de locuções disponibilizadas foi suficiente

para que estes modelos mais complexos pudessem se adaptar adequadamente, permitindo, assim, que representassem o respectivo locutor de forma mais fidedigna do que o modelo mais simples composto de 8 componentes.

7 *Conclusão*

O sistema de identificação baseado em modelos de mistura de gaussianas é formado pelos grandes processos destacados abaixo:

- pré-processamento do sinal de voz
- cálculos dos coeficientes mel-cepstrais
- identificação dos locutores
- treinamento dos modelos

O desempenho de identificação do sistema é fortemente atrelado à qualidade do sinal de voz. Tal fato pôde ser confirmado pelos melhores resultados obtidos ao se empregar o banco TIMIT, em contraposição ao banco NTIMIT, em condições experimentais compatíveis.

A forma através da qual os coeficientes mel-cepstrais são calculados deve ser escolhida levando-se em consideração a qualidade do sinal de voz. Como pode ser constatado, a técnica empregando banco de filtros, comumente referenciada na literatura, mostrou-se mais promissora quando utilizado o banco de voz TIMIT, não contaminado pelo ruído de canal telefônico. Já com um sinal de menor qualidade, proveniente do banco de voz NTIMIT, os resultados mais promissores foram atingidos quando utilizada a predição linear e a subtração da média cepstral na determinação dos coeficientes mel-cepstrais.

A extensão das locuções para realizar o treinamento dos modelos serve de referência no dimensionamento do número de componentes a ser adotado. Modelos mais complexos podem se adaptar melhor para representar os coeficientes mel-cepstrais disponibilizados na fase de treinamento. No entanto, se o modelo dispuser de um número pequeno ou excessivo de componentes diante do material de treinamento disponível, haverá uma redução na performance de identificação, sendo necessário encontrar experimentalmente o número ideal de componentes diante das condições de operação do sistema.

O número de coeficientes mel-cepstrais impacta diretamente no desempenho do sistema de identificação. Isto acontece porque se utilizado um número insuficiente desses coeficientes a modelagem do aparelho fonador se torna precária, fornecendo modelos cujo grau de complexidade não é suficiente para representá-lo. O inverso, ou seja, o emprego de um número excessivo de coeficientes pode não trazer benefícios e apenas elevar o custo computacional para o treinamento dos modelos assim gerados, sendo verificado que 24 MFCCs representa um número adequado.

Elevando-se a diversidade de locutores, houve uma queda significativa na performance do sistema quando adotado o banco NTIMIT, enquanto que o mesmo já não aconteceu de forma tão impactante com o banco TIMIT. Dessa forma, sistemas que dispõem de um elevado número de locutores só são eficientes quando empregados com sinais de elevada relação sinal-ruído.

Não houve praticamente qualquer compatibilidade quando se testou o sistema de identificação com amostras do banco NTIMIT com modelos gerados através do banco TIMIT. A situação inversa tampouco. Tal fato permite concluir que deve haver um elevado grau de compatibilidade entre as condições de treinamento dos modelos e de utilização do sistema, ou seja, não é vantajoso obter modelos a partir de sinais de voz obtidos em laboratório com elevada relação sinal-ruído e utilizar o sistema num ambiente ruidoso.

Considerando que os processos de treinamento e identificação são eficientes uma vez que resultados de elevada performance foram obtidos ao se adotar o banco de voz TIMIT, aparentemente, o desenvolvimento de novas metodologias para a extração de informações que sejam menos susceptíveis à variabilidade das condições ambientais pode trazer um ganho de performance quando não se dispõem de sinais de elevada relação sinal-ruído.

Referências Bibliográficas

- [1] REYNOLDS, D. A. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, v. 3, n. 1, p. 1094–1105, Jan 1995.
- [2] CAMPBELL, J. P. Speaker recognition: a tutorial. *Proceedings of the IEEE*, v. 85, n. 9, p. 1437–1462, Set 1997.
- [3] TOKUDA, K. Recursive calculation of mel-cepstrum from LP coefficients. *Tokyo Institute of Technology*, p. 1–7, Abril 1994.
- [4] RABINER, L. *Fundamentals of speech recognition*. United States, New Jersey: Prentice Hall PTR, 1993. 113–114 p.
- [5] RABINER, L. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, v. 54, n. 2, p. 297–315, 1975.
- [6] COHEN, I. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Processing Letters*, v. 9, n. 1, Jan 2002.
- [7] MARTIN, R. Spectral subtraction based on minimum statistics. *Proceedings of the Seventh European Signal Processing Conference*, p. 1182–1185, Set 1994.
- [8] RABINER, L. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, v. 54, n. 2, p. 274–277, 1975.
- [9] RABINER, L. *Fundamentals of speech recognition*. United States, New Jersey: Prentice Hall PTR, 1993. 100–102 p.
- [10] RABINER, L. *Fundamentals of speech recognition*. United States, New Jersey: Prentice Hall PTR, 1993. 115 p.
- [11] RABINER, L. *Digital processing of speech signals*. United States, New Jersey: Prentice Hall, 1978. 364 p.
- [12] RAMÍREZ, M. A. *Codificador preditivo de voz por análise-mediante-síntese*. Dissertação (Mestrado) — Universidade de São Paulo, Brasil, São Paulo, 1992.
- [13] PICONE, J. W. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, v. 81, n. 9, Sep 1993.
- [14] SLANEY, M. *Auditory Tool Box Version 2*. 1998. Interval Research Corporation. Disponível em: <<http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/>>.

- [15] GANCHEV NIKOS FAKOTAKIS, G. K. T. Comparative evaluation of various MFCC implementations on the speaker verification task. *Proceedings of the 10th International Conference on Speech and Computer (SPECOM 2005)*, v. 1, p. 191–194, 2005.
- [16] OPPENHEIM, A. V. *Discrete-time signal processing*. 2nd edition. ed. United States, New Jersey: Prentice Hall, 1999. 274 p.
- [17] OPPENHEIM, A. V. Discrete representation of signals. *Proceedings of the IEEE*, v. 60, n. 7, p. 681–691, June 1972.
- [18] KERMORVANT, C. *A comparison of noise reduction techniques for robust speech recognition. IDIAP-RR 10*. 1999. Disponível em: <citeseer.ist.psu.edu/article/kermorvant99comparison.html>.
- [19] TOMASI, C. Estimating Gaussian mixture densities with EM - a tutorial. Disponível em: <<http://www.cs.duke.edu/courses/spring04/cps196.1/handouts/EM/tomasiEM.pdf>>.
- [20] BILMES, J. A. *A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov models*. 1997. Disponível em: <<http://crow.ee.washington.edu/people/bulyko/papers/em.pdf>>.
- [21] REYNOLDS, D. A. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, p. 430–451, 2004.
- [22] LINDE, A. B. Y.; GRAY, R. M. An algorithm for vector quantizer design,” *IEEE transactions on communications*. *IEEE Transactions on Communications*, n. 702-710, January 1980.
- [23] RABINER, L. *Fundamentals of speech recognition*. United States, New Jersey: Prentice Hall PTR, 1993. 115-116 p.
- [24] BURT, P. M. S. *Apostila de Processamento Digital de Sinais 2*. Universidade de São Paulo, Brasil, São Paulo, 2004.
- [25] OPPENHEIM, A. V. *Discrete-time signal processing*. 2nd edition. ed. United States, New Jersey: Prentice Hall, 1999. 589 p.

APÊNDICE A – Transformada de Fourier Discreta

O cálculo da transformada de Fourier discreta (TFD) ou *Discrete Fourier Transform (DFT)* de um quadro de comprimento M é definida pela equação (A.1). Considerando um sinal complexo $x(n)$, são necessárias M^2 operações de multiplicações complexas e $M(M - 1)$ somas complexas para o seu cálculo direto.

$$X(k) = \sum_{n=0}^{M-1} x(n)e^{-j\frac{2\pi k}{M}n} \quad \text{para} \quad k = 0, 1, \dots, M - 1 \quad (\text{A.1})$$

Com o intuito de reduzir este levado número de operações, em 1965 foram propostas por Cooley-Tukey métodos de transformada rápida de Fourier ou *Fast Fourier Transform (FFT)*. A título de exemplo, para um quadro de $M = 2^j$ amostras, seriam necessárias apenas $\frac{M}{2}\log_2 M$ multiplicações complexas e $M\log_2 M$ somas complexas [24] ao se aplicar métodos de transformação rápida na determinação da TFD, reduzindo de forma significativa o tempo de processamento para realizar esta operação.

A.1 Transformada do Cosseno Discreta

A transformada do cosseno discreta (TCD), ou *Discrete Cosine Transform (DCT)*, de forma análoga à TFD corresponde a uma transformada ortogonal, diferenciando-se desta última pelo fato de só ser aplicável a seqüências reais. Com a TCD busca-se formar uma seqüência periódica e simétrica a partir de uma seqüência finita de tal forma que esta última possa ser recuperada exclusivamente [25]. Como é possível realizar esta transformada de formas diversas, há várias formulações para a TCD, no entanto, para processamento de sinais de voz é empregada a TCD tipo 2 por concentrar mais energia nos seus primeiros coeficientes, permitindo uma maior compactação da informação presente no sinal. A TCD tipo 2 é definida pela equação (A.2).

$$X^{c2}(k) = \sum_{n=0}^{M-1} x(n) \cos\left(\frac{\pi}{M}\left(n + \frac{1}{2}\right)k\right) \quad \text{para} \quad k = 0, 1, \dots, M-1 \quad (\text{A.2})$$

Para tornar a transformação ortonormal, faz-se a inclusão do fator $\beta[k]$ na equação (A.2), dando origem à equação normalizada (A.3).

$$X_{ort}^{c2}(k) = \sqrt{\frac{2}{M}} \beta[k] \sum_{n=0}^{M-1} x(n) \cos\left(\frac{\pi}{M}\left(n + \frac{1}{2}\right)k\right) \quad \text{para} \quad k = 0, 1, \dots, M-1 \quad (\text{A.3})$$

$$\beta[k] = \begin{cases} \frac{1}{\sqrt{2}}, & k = 0 \\ 1, & k = 1, 2, \dots, M-1 \end{cases}$$

A TFD está relacionada com a TCD conforme a equação (A.4), permitindo que todo o conhecimento adquirido no cálculo rápido da TFD seja aproveitado para obter a TCD.

$$X^{c2}(k) = 2\text{Re}\{X(k)e^{-j\pi k/(2M)}\} \quad \text{para} \quad k = 0, 1, \dots, M-1 \quad (\text{A.4})$$