Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Programa de Pós-Graduação em Engenharia de Eletricidade

TESE DE DOUTORADO

Sistema de inferência genético-nebuloso para reconhecimento de voz: Uma abordagem em modelos preditivos de baixa ordem utilizando a transformada cosseno discreta

Washington Luis Santos Silva

São Luís, MA
2015

Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Programa de Pós-Graduação em Engenharia de Eletricidade

TESE DE DOUTORADO

Sistema de inferência genético-nebuloso para reconhecimento de voz: Uma abordagem em modelos preditivos de baixa ordem utilizando a transformada cosseno discreta

Washington Luis Santos Silva

Tese de Doutorado submetida à Coordenação do Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão, como parte dos requisitos para a obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Ginalber Luiz de Oliveira Serra

São Luis-MA
2015

SISTEMA DE INFERÊNCIA GENÉTICO-NEBULOSO PARA RECONHECIMENTO DE VOZ: UMA ABORDAGEM EM MODELOS PREDITIVOS DE BAIXA ORDEM UTILIZANDO A TRANSFORMADA COSSENO DISCRETA

Washington Luis Santos Silva

Tese aprovada em 20 de março de 2015.

Prof. Ginalber Luiz de Oliveira Serra, Dr.
(Orientador)

Prof. Jugurta Rosa Montalvão Filho, Dr.
(Membro da Banca Examinadora)

Prof. Fabio Lima, Dr.
(Membro da Banca Examinadora)

Prof. Ewaldo Eder Carvalho Santana, Dr.
(Membro da Banca Examinadora)

Prof. João Viana da Fonseca Neto, Dr.
(Membro da Banca Examinadora)

Resumo

Neste trabalho propõe-se uma metodologia que utiliza um sistema inteligente para reconhecimento de voz. Utiliza-se a definição de sistema inteligente, como o sistema que possui a capacidade de adaptar seu comportamento para atingir seus objetivos em uma variedade de ambientes (Fogel, 2006). Utiliza-se, também, a definição de Inteligência Computacional, como sendo a simulação de comportamentos inteligentes em termos de processo computacional (Schalkoff, 1990). Além do pré-processamento do sinal de voz com coeficientes mel-cepstrais, a transformada discreta cosseno (TCD) é utilizada para gerar uma matriz bidimensional para modelar cada padrão a ser reconhecido. Um sistema de inferências nebuloso Mamdani para reconhecimento de voz é otimizado por algoritmo genético para maximizar a quantidade de acertos na classificação dos padrões com um número reduzido de parâmetros. Os resultados experimentais alcançados no reconhecimento de voz com a metodologia proposta foram comparados com o *Hidden Markov Models*-HMM e com os classificadores *Gaussian Mixture Models*-GMM e máquina de vetor de suporte (*Support Vector Machine*-SVM) com intuito de avaliação de desempenho. O sistema de reconhecimento usado neste trabalho foi denominado *Intelligent Methodology for Speech Recognition*-IMSR.

Palavras-chave: Sistemas Nebulosos, Reconhecimento Automático de Voz, Algoritmo Genético, Transformada Cosseno Discreta, Sistemas Inteligentes.

Abstract

This thesis proposes a methodology that uses an intelligent system for voice recognition. It uses the definition of intelligent system, as the system has the ability to adapt their behavior to achieve their goals in a variety of environments. It is used also, the definition of Computational Intelligence, as the simulation of intelligent behavior in terms of computational process. In addition the speech signal pre-processing with mel-cepstral coefficients, the discrete cosine transform (DCT) is used to generate a two-dimensional array to model each pattern to be recognized. A Mamdani fuzzy inference system for speech recognition is optimized by genetic algorithm to maximize the amount of correct classification of standards with a reduced number of parameters. The experimental results achieved in speech recognition with the proposed methodology were compared with the Hidden Markov Models-HMM and the classifiers Gaussians Mixtures Models-GMM and Support Vector Machine-SVM. The recognition system used in this thesis was called Intelligent Methodology for Speech Recognition-IMSR.

Keywords: Fuzzy Systems; Automatic Speech Recognition; Genetic Algorithms; Discrete Cosine Transform; Intelligent System.

Agradecimentos

A Jesus, palavra viva, sabedoria feito carne, pelo dom da vida, da fé e da razão.

A minha esposa Janice Silva, aos meus filhos, Bárbara, Maria Luísa, Lucas, Pedro e Antônio, pela compreensão e pelos muitos momentos de afastamento do ambiente familiar para desenvolvimento deste trabalho, e principalmente pelo amor que me dá força e coragem para continuar caminhando sempre adiante.

Ao Prof. Dr. Ginalber Luiz de Oliveira Serra, por ter ensinado-me que orientação se faz com humildade, respeito, dedicação e sabedoria.

Ao Prof. José Raimundo Costa Borges, in memoriam, eterno Mestre.

Ao Prof. MSc. Nelson José Camelo, pela importante contribuição na minha formação profissional.

Aos meus alunos de iniciação científica, em especial a Amanda Abelardo e a Gracieth Cavalcanti, pela contribuição na elaboração deste trabalho.

A minha família pelo apoio durante esta jornada.

Ao Instituto Federal do Maranhão, Campus São Luis-Monte Castelo, pelo apoio material na execução deste trabalho.

Dedico esta obra a meus pais;
A Janice, minha inspiração;
Ao Professor José Raimundo Costa Borges, in memoriam,
por todo apoio que a mim dedicou ao longo de minha carreira.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

$IMSR$	-	Intelligent Methodology for Speech Recognition
RAV	-	Reconhecimento Automático de Voz
$SRAV$	-	Sistema de Reconhecimento Automático de Voz
$\bar{x}$	-	Vetor x
$\mathbb{R}^n$	-	Espaço Euclidiano de dimensão n
Φ_x	-	Matriz de Covariância
$P(.)$	-	Probabilidade
TFD	-	Transformada de Fourier Discreta
$TFTD$	-	Transformada de Fourier de Tempo Discreto
FFT	-	*Fast Fourier Transform*
DCT	-	*Discrete Cosine Transform*
TCD	-	Transformada Cosseno Discreta
$argmax$	-	Argumento máximo de um dado vetor ou função
$MFCC$	-	*Mel-Frequency Cepstrum Coefficient*
HMM	-	*Hidden Markov Models*
GMM	-	*Gaussian Mixtures Models*
SVM	-	*Support Vector Machine*
RBF	-	*Radial Basis Functions*
$\mathcal{N}$	-	Distribuição Normal
$w[.]$	-	Função Janela
mel	-	Escala Mel-Cepstral
f	-	Frequência Linear
$E[.]$	-	Energia de saída do banco de filtros
$\|\|A\|\|$	-	Norma da Matriz A
j	-	Modelo de palavra falada
λ^j	-	Conjunto de parâmetros do modelo de palavra j
$C_k(n,T)$	-	Matriz temporal bidimensional de ordem $k \times n$ calculados no segmento T
c_{kn}^j	-	Coeficientes da matriz temporal bidimensional de ordem $k \times n$ calculados no segmento T (variáveis linguísticas) do modelo de palavra j
$\tilde{c}_{kn}^j$	-	Valores linguísticos das variáveis linguísticas c_{kn} do modelo de palavra j
CM_{kn}^j	-	Matriz das médias do modelo j
CV_{kn}^j	-	Matriz das variâncias do modelo j
Ru	-	Base de regras
min	-	Argumento mínimo de um função
max	-	Argumento máximo de um função

Trabalhos Publicados pelo Autor

Revista:

1. Washington L.S. Silva, Ginalber L.O. Serra. "Intelligent Genetic Fuzzy Inference System for Speech Recognition: An Approach from Low Order Feature Based on Discrete Cosine Transform". *Journal of Control, Automation and Electrical Systems (JCAE)*, Springer, Vol.25(6), pp. 689-698, December 2014. ISSN: 2195-3880(Print), 2195-3899 (Online). DOI:10.1007/s40313-014-0148-0 .

Capítulos de Livros Publicados

1. Washington L.S. Silva, Ginalber L.O. Serra. "An Intelligent System Based on Discrete Cosine Transform For Speech Recognition". *Lecture Notes in Computer Science: Advances in Artificial Intelligence*, Springer-Verlag Berlin Heidlburg, pp. 320-329, November 2012. ISBN: 978-3-642-34653-8(Print) 978-3-642-34654-5(Online). DOI:10.1007/978-3-642-34654-5-33.

2. Washington L.S. Silva, Ginalber L.O. Serra. "A Hybrid Approach Based on DCT-Genetic-Fuzzy Inference System For Speech Recognition".*Lecture Notes in Computer Science:Intelligent Data Engineering and Automated Learning*, Springer-Verlag Berlin Heidlburg, pp. 52-59, August 2012. ISBN: 978-3-642-32638-7(Print) 978-3-642-32639-4(Online). DOI:10.1007/s40313-014-0148-0.

3. Amanda V. Abelardo, Washington L.S. Silva, Ginalber L. O. Serra. "CPSO Applied in the Optimization of a Speech Recognition System". *Lecture Notes in Computer Science:Intelligent Data Engineering and Automated Learning*, Springer-Verlag Berlin Heidlburg, pp. 134-141, September 2014. ISBN: 978-3-319-10839-1(Print) 978-3-319-10840-7(Online). DOI:10.1007/978-3-319-10840-7-17.

Artigos Aceitos para Publicação em Congressos

1. Washington L.S. Silva, Ginalber L.O. Serra. "Proposta de Metodologia TCD-Fuzzy para Reconhecimento de Voz". *X Simpósio Brasileiro de Automação Inteligente* (SBAI 2011), São João Del Rey, MG, Brasil, pp. 1054-1059, Setembro 2011.

2. Washington L.S. Silva, Ginalber L.O. Serra. "Sistema de Inferência Fuzzy Baseado na Transformada Cosseno Discreta para Reconhecimento de Voz". *X Congresso Brasileiro de Inteligência Computacional* (CBIC 2011), Fortaleza, CE, Brasil, pp. 1054-1059, Setembro 2011.

3. Washington L.S. Silva, Ginalber L.O. Serra. "Proposal of an Intelligent Speech Recognizer System". *IEEE Third Global Congress on Intelligent Systems* (GCIS 2012), Wuhan, China, pp.356-359, November 2012.

4. Washington L.S. Silva, Ginalber L.O. Serra. "GFIS:Genetic Fuzzy Inference System for Speech Recognition". IX International Conference on Informatics in Control, Automation and Robotics (ICINCO 2012), Rome, Italy, pp. 536-541, July 2012.

5. Washington L.S. Silva, Ginalber L.O. Serra. "Sistema de Inferência Genético-Fuzzy para Reconhecimento de Voz". XIX Congresso Brasileiro de Automática (CBA 2012), Campina Grande, PB, Brasil, pp. 536-541, July 2012.

6. Washington L.S. Silva, Ginalber L.O. Serra. "A Hibrid Method for Extraction of Low-Order Features for Speech Recognition Application". *The Sixth International Conference on Advanced Engineering Computing and Applications in Sciences*(ADVCOMP 2012), Barcelona, Spain, pp. 123-129, September 2012.

7. Washington L.S. Silva, Ginalber L.O. Serra. "Análise Comparativa entre as Implicações Lukasiewicz, Dienes-Rescher, Mamdani aplicadas ao Reconhecimento de Voz.". Segundo Congresso Brasileiro de Sistemas Fuzzy (II CBSF 2012), Natal, RN, Brasil, pp. 997-1012, Novembro 2012.

8. Washington L.S. Silva, Ginalber L.O. Serra. "Intelligent Genetic Fuzzy Inference System for Speech Recognition". 11th IEEE International Conference on Industrial Informatics (INDIN 2013), Bochum, Germany, July 2013.

9. Washington L.S. Silva, Ginalber L.O. Serra e Amanda A.V. Beserra. "Otimizacão Bio-inspirada Utilizando Enxame de Partículas com Aplicação em Reconhecimento de Voz". 11 Congresso Brasileiro de Inteligência Computacional (CBIC 2013), Porto de Galinhas-PE, Setembro de 2013.

10. Washington L.S. Silva, Ginalber L.O. Serra. "Intelligent Genetic Fuzzy Inference System for Speech Recognition". 14th Word Congress on Computational Intelligence (IEEE WCCI 2014), Beijing-China, July 2014. pp. 3599-3604. ISBN: 978-1-4799-6627-1. DOI:10.1109/IJCNN.2014.6889833.

11. Gracieth B. Cavalcanti, Washington L.S. Silva, Orlando R. Filho. "Classification of Pattern using Support Vector Machines: An Application for Automatic Speech Recognition". *The Eighth International Conference on Advanced Engineering Computing and Applications in Sciences, 2014*, Rome-Italy. IARIA Conference 2014. Vol. 5. pp. 91-96. ISBN: 978-1-61208-354-4.

Capítulo 1

Considerações Iniciais

O meio natural de comunicação usado por seres humanos é a voz [1]. Portanto, é comum que uma série ampla de pesquisas, geralmente denominada processamento de voz, seja dedicada a analisar e compreender a voz humana. Seguindo o modelo apresentado por (Bresolin, 2008), apresenta-se, na figura 1.1, um diagrama de um sistema de processamento de voz.

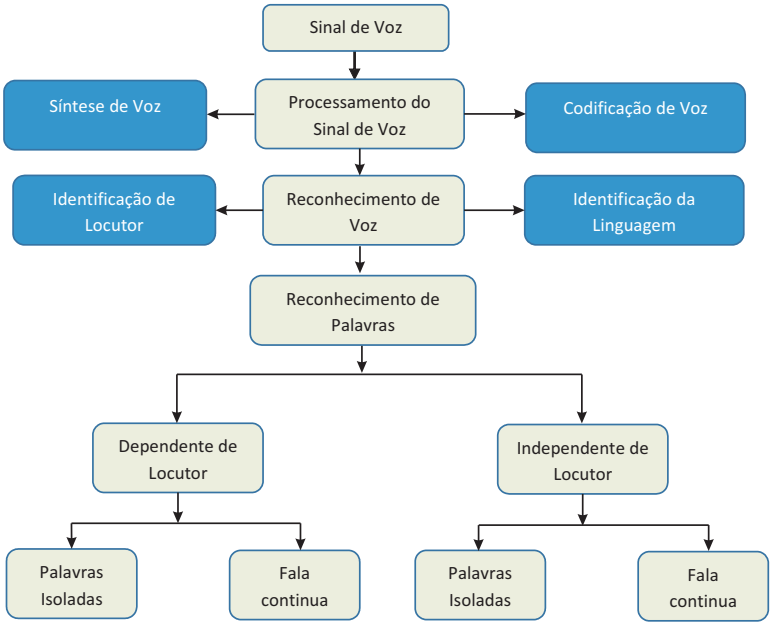


Fig. 1.1: Diagrama de um sistema de processamento de voz

Os sistemas desenvolvidos para processamento de voz podem ser divididos em:

1. Codificação de voz: a finalidade da codificação da fala é comprimir o sinal de voz com objetivo de transmissão e/ou processamento em meio digital de forma eficiente.

[1] Som produzido na laringe, pelo ar que sai dos pulmões e da boca do homem (Ferreira, 2010)

2. Síntese de voz: é o processo de produção artificial da voz humana. Um sistema digital utilizado para este propósito é denominado sintetizador de voz, e pode ser implementado em *software* ou *hardware*. Um sistema texto-voz converte texto em linguagem normal para voz; outros sistemas interpretam representações linguísticas simbólicas (como transcrição fonética) em voz.

3. Reconhecimento automático de voz (RAV): processo que tem por objetivo transcrever sons de voz em uma sequência correspondente de palavras[2], denominado reconhecimento de palavras, e/ou determinar a pessoa que está falando (locutor), denominado mais comumente de reconhecimento de locutor e/ou reconhecer a linguagem que está sendo utilizada e suas características fonéticas.

4. Este trabalho tem seu foco principal no reconhecimento de voz voltado para a identificação das palavras; isto é, transcrição dos sons pronunciados em forma de palavras.

1.1 Sistema de reconhecimento automático de voz

Um sistema básico de reconhecimento de voz compreende um conjunto de algoritmos elaborados a partir de conceitos desenvolvidos em áreas como reconhecimento estatístico de padrões, teoria de comunicações, processamento de sinal, matemática combinacional, linguística, entre outras. Embora cada uma dessas áreas possa contribuir em graus variados com diferentes reconhecedores, talvez o maior denominador comum entre todos os sistemas de reconhecimento seja o processamento da voz, o qual converte o sinal de voz para algum tipo de representação paramétrica para análise e posterior processamento.

O objetivo de um sistema de reconhecimento automático de voz (SRAV) é converter com precisão e eficiência um sinal de voz em uma representação paramétrica que possa ser processada para fins de identificação de voz e/ou de locutor, independente dos dispositivos usados para gravar tal sinal (transdutor ou microfone), sotaque, ou ambiente acústico em que os dispositivos estão localizados (escritório silencioso, sala barulhenta, ambientes externos). Ou seja, o objetivo final de um sistema de reconhecimento automático de voz é ter um desempenho tão bom como o do ouvido humano, no critério de separabilidade dos sons. A base para a maioria dos algoritmos de processamento digital do sinal de voz é um modelo de sistema no tempo discreto para a produção de amostras do sinal de voz que serão codificadas para manipulações posteriores (Rabiner, 1993; Sadaoki, 2000). Na figura 1.2 apresenta-se uma abordagem simples para um SRAV.

[2]Som ou conjuntos de sons, devidamente articulados, com significado (Ferreira, 2010)

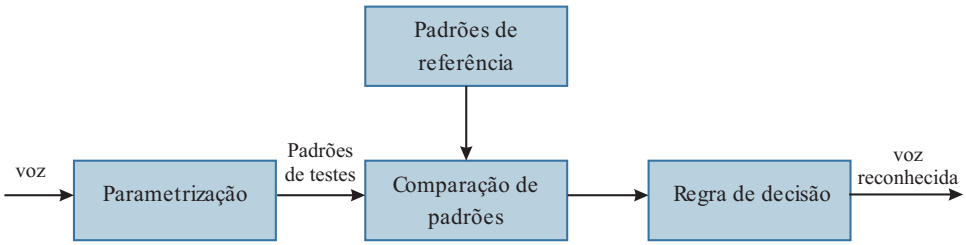


Fig. 1.2: Diagrama do SRAV simplificado.

As etapas básicas no modelo do reconhecedor de padrão apresentado na figura 1.2 são:

a) Parametrização: o sinal de voz a ser reconhecido é convertido em parâmetros adequados à análise acústica onde as características relevantes deste sinal são transformadas em representações paramétricas eficientes para fins de reconhecimento. A seleção das melhores representações paramétricas do sinal de voz é uma tarefa básica, contudo, não trivial, no desenvolvimento de qualquer sistema de reconhecimento de voz. O objetivo desta seleção é comprimir o sinal de voz, eliminando informações não pertencentes à análise fonética do sinal, e enfatizar aspectos que contribuam significativamente às detecções das diferenças fonéticas dos sons de voz. Dessa forma, o problema de reconhecimento de voz envolve a parametrização do sinal de voz, tal que as caracterísiticas acústicas observadas sejam relacionadas com os símbolos fonéticos do sinal e a geração dos padrões que serão utilizados como modelos no processo de reconhecimento (Azar, 2008; Picone, 1993; Revathi, 2011). Existe uma grande variedade de técnicas para representações paramétricas do sinal de voz, tais como, energia de curto prazo, razão de cruzamento por zeros, análise do envelope espectral de curto prazo - provavelmente a representação mais importante (Keshet, 2009; Picone, 1993; Rabiner, 1993).

b) Comparação de padrões: nesta etapa ocorre a comparação dos parâmetros do sinal de teste com os parâmetros dos padrões de referência com objetivo de determinar suas similaridades. Um fator muito importante na maioria dos algoritmos de comparações de padrões é a medida de distância (dissimilaridade) entre dois conjuntos de parâmetros a serem comparados. Essa medida de dissimilaridade pode ser tratada com rigor matemático se os conjuntos de parâmetros puderem ser observados em um espaço vetorial. Uma medida de dissimilaridade que pode ser utilizada na etapa de comparação é a medida de distância. Várias medidas de distância podem ser definidas baseadas em representações de vetores multivariáveis. Uma medida de distância $d(\bar{x}, \bar{y})$ entre dois vetores $\bar{x}$ e $\bar{y}$ pertecentes a um espaço vetorial χ deve satisfazer as seguintes propriedades (Ledermann, 1982):

$$0 \leq d\left(\bar{x}, \bar{y}\right) < \infty, \forall \left(\bar{x}, \bar{y}\right) \in \chi \quad e \quad d\left(\bar{x}, \bar{y}\right) = 0 \Leftrightarrow \bar{x} = \bar{y} \tag{1.1}$$

$$d\left(\bar{x}, \bar{y}\right) = d\left(\bar{y}, \bar{x}\right), \forall \left(\bar{x}, \bar{y}\right) \in \chi \tag{1.2}$$

$$d\left(\bar{x}, \bar{y}\right) \leq d\left(\bar{x}, \bar{z}\right) + d\left(\bar{y}, \bar{z}\right), \forall \left(\bar{x}, \bar{y}, \bar{z}\right) \in \chi \tag{1.3}$$

Contudo as propriedades de definição positiva, simetria e desigualdade triangular, respectivamente, obeservadas rigorosamente em uma medida de distância, não implicam bons resultados no processo de decisão. Desse modo, em reconhecimento de voz é muito importante formular algoritmos capazes de determinar $d(\bar{x}, \bar{y})$ eficientemente. Embora a distância euclidiana seja usada em muitos casos para a obtenção de $d(\bar{x}, \bar{y})$, diversas modificações para a distância euclidiana têm sido sugeridas na literatura especilizada. Entre elas estão as distâncias ponderadas baseadas na sensibilidade auditiva e as distâncias em espaços multidimensionais reduzidos obtidas através de análises estatísticas (Rabiner, 1993; Sadaoki, 2000).

Uma medida de distância útil para processamento de voz tem que possuir uma alta correlação entre valores numéricos e as hipóteses de medidas subjetivas avaliadas a partir do sinal de voz. As medidas subjetivas são baseadas em conjuntos de parâmetros extraídos a partir de características acústicas da voz, tais como compacidade, quantidade de graves e agudos, envelope espectral, entre outros. Para o problema de reconhecimento de voz, a consistência da medida de dissimilaridade implica necessariamente que a medida matemática de distância precisa estar de acordo com as características linguísticas conhecidas do sinal de voz (Rabiner, 1993; Sadaoki, 2000). Este requerimento de subjetividade inerente à lingüística, normalmente, não pode ser satisfeito com matemática determinística. Por exemplo, uma grande diferença no erro médio quadrático relativo à forma de onda do sinal de voz nem sempre implica grandes diferenças subjetivas da análise do sinal de voz (Grimm, 2007; Rabiner, 1993; Shabtai, 2010). Assim, para qualquer sistema de reconhecimento de voz, um ponto muito importante é a medida de dissimilaridade entres os padrões que se está analisando.

c) Decisão: Esta etapa está estritamente ligada à etapa de comparação. Responsável por determinar, através das medidas de dissimilaridade obtidas na etapa de comparação, entre os padrões representados, o que mais se aproxima do sinal a ser comparado e reconhecido.

1.1.1 Considerações sobre o SRAV

No campo de reconhecimento de padrões, os sinais são frequentemente considerados como produtos de fontes que atuam estatisticamente. Portanto, o principal objetivo da análise desses sinais é modelar as propriedades estatísticas das fontes e dos sinais o mais precisamente possível. Como base da construção de modelos dessas fontes e sinais estatísticos, normalmente, apenas as observações dos dados de exemplos e as hipóteses sobre as limitações de graus de liberdade dos modelos estão disponíveis. No entanto, dois aspectos sobre os modelos são muito importantes:

- Os modelos representativos dos padrões devem replicar a geração dos dados, através das observações, tão preciso quanto possível;

- Os modelos devem fornecer informações úteis para segmentar os sinais em unidades significativas.

Uma das técnicas mais difundidas para reconhecimento de padrões de voz é o *Hidden Markov Model* (HMM) (Abushariah et al., 2010; Gales, 2007; Rabiner, 1989; Shenouda, 2006; Tarihi et al., 2005). O HMM fornece uma estrutura simples e eficaz para a modelagem de sequências que variam no tempo e, além disso, atende aos dois aspectos citados anteriormente. Primeiro, o HMM gera um modelo estatístico com base em dados obtidos de uma dada sequência de acordo com distribuições de probabilidades bastante complexas e que podem ser usados para classificar os padrões sequenciais. Segundo, informações sobre a segmentação dos dados considerados podem ser deduzidas de processos estocásticos de dois estágios. Consequentemente, o HMM possui a capacidade notável para tratar a segmentação e classificação de padrões em uma estrutura integrada (Fink, 2014; Rabiner, 1989). Como resultado dessas características do HMM, uma grande parte dos sistemas de reconhecimento de voz utilizam o HMM em suas modelagens, seja para sistemas de reconhecimento de dígitos isolados ou para reconhecimento de fala contínua de grandes e pequenos vocabulários.

O HMM é um processo duplamente estocástico, com um processo estocástico não observável (oculto), mas que pode ser inferido através de outro processo estocástico que produz a sequência de observações. Os processos ocultos consistem de um conjunto de estados conectados por transições com probabilidades, enquanto que os processos observáveis consistem de um conjunto de saídas ou observações, cada qual pode ser emitido por cada estado de acordo com alguma saída da função de densidade de probabilidade (fdp) ou da distribuição de probabilidade dos processos em análise (Fink, 2014; Rabiner, 1989).

O comportamento de um processo que em um dado tempo t depende somente do estado anterior pode ser caracterizado como segue:

$$P\left(S_t/S_1, S_2, ..., S_{t-1}\right) = P\left(S_t/S_{t-1}\right) \tag{1.4}$$

Essa limitação da extensão temporal da dependência estatística dentro do modelo corresponde às chamadas propriedades de Markov. No segundo estágio do processo estatístico, define-se completamente o HMM; isto é, para cada ponto no tempo t uma saída ou observação O_t é gerada. A distribuição de probrabilidade conjunta depende somente do estado atual S_t e não dos estados anteriores ou observações:

$$P\left(O_t/O_1, O_2, \ldots, O_{t-1}, S_1, S_2, \ldots, S_{t-1}\right) = P\left(O_t/S_t\right) \tag{1.5}$$

Na literatura especializada, essa propriedade é denominada de hipóstese da independência da saída (Revathi, 2011; Shenouda, 2006). Somente a sequência das saídas pode ser observada do comportamento do modelo. Ao contrário, a sequência de estado utilizada durante a geração dos dados não pode ser observada.

Em reconhecimento de padrões, o comportamento do HMM normalmente é considerado sobre um intervalo de tempo finito. Para inicialização dos modelos no ínicio do período de observação, probabilidades iniciais são utilizadas para descrever a distribuição de probabilidades dos estados no tempo $t = 1$. Um "*hidden Markov model*" que usualmente é representado por λ pode ser descrito por:

1. Um conjunto finito de estados $\{s|1 \leq s \leq N\}$;

2. Uma matriz A de probabilidades de transição de estado,

$$A = \{a_{ij}|a_{ij} = P\left(S_t = j|S_{t-1} = i\right), \}, \tag{1.6}$$

   sendo a_{ij} a probabilidade de transição do estado i para o estado j.

3. Um vetor π de probabilidades iniciais,

$$\pi = \{\pi_i|\pi_i = P\left(S_1 = i\right)\} \tag{1.7}$$

4. Distribuições de probabilidades dos estados,

$$\{b_j(o_k)|b_j(o_k) = P(O_t = o_k|S_t = j)\} \text{ ou } \{b_j(\bar{x})|b_j(\bar{x}) = p(\bar{x}|S_t = j)\} \tag{1.8}$$

para todas as saídas do modelo.

Contudo, as distribuições das saídas precisam ser diferenciadas dependendo do tipo das observações das quais os modelos são gerados. No caso discreto, as saídas são geradas de observações discretas $O_1, O_2, \ldots, O_M$. Então, as quantidades $b_j(O_k)$ representam distribuições de probabilidades discretas que podem ser agrupadas em uma matriz de probabilidades de saídas, representadas por:

$$B = b_{jk}|b_{jk} = P\left(O_t = O_k|S_t = j\right) \tag{1.9}$$

Para este tipo de modelamento da saída, obtém-se o HMM discreto. Entretanto, se as observações são vetores cujos elementos possuem valores tal que, $\bar{x} \in \mathbb{R}^n$, as distribuições das saídas são descritas na base de funções densidades de probabilidades contínuas, representadas por:

$$b_j(\bar{x}) = p\left(\bar{x}|S_t = j\right) \tag{1.10}$$

Aplicações do HMM para solucionar problemas de análises de sinais usam, geralmente, o HMM contínuo, embora a necessidade para modelar distribuições contínuas aumentem significativamente a complexidade do formalismo matemático. Na literatura, é comum a introdução do HMM através de modelos de distribuição de probabilidades discretas. Esse tipo de HMM é muito utilizado na análise de sequências biológicas, principalmente, em estudos de DNA (Durbin et al., 1998). Para aplicações na área de processamento de sinais de voz, no entanto, modelos discretos dificilmente são considerados, uma vez que requerem a utilização de um quantizador vetorial que converte as representações características contínuas do sinal de voz em uma sequência de observações discretas antes da análise. A capacidade do HMM é aumentada consideravelmente se este passo de quantização é evitado ou incluído no processo de construção do modelo. Entretanto, para alcançar este objetivo, é necessário representar as distribuições de probabilidades contínuas sobre o $\mathbb{R}^n$ de uma maneira adequada.

Nos casos de modelagens discretas, distribuições de probabilidades empíricas que melhor representam os modelos em análise podem ser diretamente aplicadas. No caso contínuo, o uso de funções densidades de probabilidades empíricas ou aproximações das modelagens são mais complexas, pois, representações paramétricas de densidades de probabilidades são somente conhecidas para um pequeno número de famílias de distribuições; por exemplo, a Normal ou Gaussiana. Para aplicações em modelamento de sinais de voz, aplicações de distribuições unimodais podem não ser as mais adequadas, pois, elas não representam bem as características necessárias ao modelo, uma vez que, com apenas uma única região de alta densidade na área do valor esperado da função de densidade de probabilidade, somente dados com comportamento estatístico correspondente a estas distribuições podem ser descritos.

A fim de se poder lidar com distribuições contínuas arbitrárias com múltiplos modos ou regiões de alta densidade em geral, são aplicadas técnicas de aproximação. A técnica mais bem conhecida e mais largamente utilizada consiste no uso de misturas de densidades, utilizando-se densidades gaussianas. Pode ser demonstrado que cada distribuição de probabilidade contínua $p(\bar{x})$ pode ser aproximada com precisão arbitrária através da combinação linear de M componentes de distribuição normal, representada por $\mathcal{N}$ (Ferguson, 1983):

$$p(\bar{x}) \approx \sum_{k=1}^{M} c_k \, \mathcal{N}\left(\bar{x}|\bar{\mu}_k, C_k\right) \tag{1.11}$$

Se um número M finito de misturas é utilizado, observa-se um erro na aproximação; todavia, esse erro pode ser mantido pequeno, desde que se utilize um número adequado de misturas. Os pesos c_k da mistura devem ter as seguintes restrições:

$$\sum_{k} c_k = 1, \;\; 0 < c_k \leq 1, \;\; \forall k \tag{1.12}$$

Isso resulta em uma combinação convexa das densidades componentes da mistura e garante que o resultado das misturas seja novamente uma distribuição de probabilidade.

A forma geral do HMM contínuo utiliza uma mistura de densidades por estado j para a descrição da função densidade de probabilidade da saída, dada por:

$$b_j(\bar{x}) = \sum_{k=1}^{M_j} c_{jk} \, \mathcal{N}\left(\bar{x}|\bar{\mu}_{jk}, C_{jk}\right) = \sum_{k=1}^{M_j} c_{jk} g_{jk}(\bar{x}) \tag{1.13}$$

O número M_j de componentes da mistura utilizadas pode variar de estado para estado. Cada uma das densidades de probabilidade de saída $b_j(\bar{x})$ é parametrizada por um conjunto de pesos da mistura c_{jk} de um estado específico e um conjunto de densidades normais $\mathcal{N}\left(\bar{x}|\bar{\mu}_{jk}, C_{jk}\right)$. Além do mais, cada densidade componente, denotada por $g_{jk}(\bar{x})$, possui um conjunto individual de parâmetros que consiste de um vetor média $\bar{\mu}_{jk}$ e de uma matriz de covariância C_{jk}.

Um HMM contínuo pode ser considerado como um modelo discreto com estados especificados por estágio de quantização. A distribuição de saída discreta pode ser dada nos pesos da mistura e as densidades componentes da mistura utilizada definem a regra de quantização. A distribuição de saída discreta pode ser encontrada nos pesos das misturas e as densidades componentes usadas definem a regra de quantização, os valores da densidades de todas as distribuições normal usadas são incorporadas nos cálculos do HMM.

Para o HMM contínuo, o número de parâmetros é drasticamente aumentado em relação ao HMM discreto. Assim, na literatura especializada, técnicas foram e são apresentadas com o

intuito de reduzir o número de parâmetros utilizando características de ambos HMMs. Uma das mais conhecidas é denominada de HMM semicontínuo. Em tais modelos, somente um simples conjunto de densidades componentes é utilizado para a construção de todas as funções de densidade de probabilidade de saída:

$$b_j(\bar{x}) = \sum_{k=1}^{M} c_{jk}\, \mathcal{N}\left(\bar{x}|\bar{\mu}_k, C_k\right) = \sum_{k=1}^{M} c_{jk} g_k(\bar{x}) \tag{1.14}$$

O conjunto global das componentes $g_k(\bar{x})$ da mistura é frequentemente denominado de livro de códigos ("*codebook*"). A definição da densidade de probabilidade de saída é completamente análoga ao caso do HMM contínuo. Entretanto, não há mais uma dependência do estado do parâmetro das distribuições normais subjacentes. Adicionalmente, no HMM semicontínuo cada mistura consiste do mesmo número M de distribuições componentes da mistura.

Apesar de sua capacidade de reconhecimento, uma das principais deficiências do HMM convencional está relacionada com a modelagem inadequada da duração do estado associado ao evento acústico (Park et al., 1996; Rabiner, 1989; Smith et al., 1995). Desde que a probabilidade de recorrência para o mesmo estado é constante, a probabilidade de duração do evento acústico associado com o estado tem uma probabilidade exponencial decrescente com o tempo. A hipótese básica é que a voz é um sinal quase estacionário e a sua parte estacionária pode ser representada por um simples estado do HMM; este tipo de duração não representa a estrutura temporal da voz. Para diminuir o efeito negativo desse modelamento, várias técnicas têm sido propostas, tais como inclusão de densidade de duração explicita do estado (Rabiner, 1989), probabilidades de transição de estado dependente do tempo(Ramesh, 1992), entre outras.

Outra fragilidade do HMM é a suposição de que, dentro dos estados, os vetores de observação são não correlacionados, enquanto que, na realidade pode acontecer o oposto da hipótese admitida. Esta característica tem sido explorada para desenvolver sistemas de reconhecimento robustos. A ideia básica é incluir características dinâmicas nos vetores de observação. Para solucionar tais problemas, várias técnicas são apresentadas na literatura especializada (Wachter et al., 2007), e podem ser resumidas basicamente em quatro categorias referidas por (Milner, 1994):

1. Adição ao vetor de observação das primeiras e segundas derivadas de cada componente do vetor;

2. Uso de probabilidade condicional dentro de cada estado;

3. Modelamento explicito do vetor de correlação espectral por meio de modelos de predição linear associado a cada estado;

4. Uso de características cepstrais bidimensionais.

A primeira solução é mais popular, enquanto que a segunda e a terceira soluções apresentam problemas relacionados à precisão das estimações paramétricas e aumentam a complexidade de codificação. Como alternativa, a quarta solução apresenta características espectrais mais robustas (Milner, 1994). As três primeiras soluções apresentadas acima, contudo, não levam em consideração as variações espectrais globais. Assim, erros podem ocorrer devido ao modelamento inadequado das variações temporais, isto é, porque uma sequência de observação é decodificada por poucos estados, tipicamente absorvendo segmentos de pouca energia e com alta probabilidade de duração (Ariki et al., 1989; Milner, 1994). Os outros estados, em vez disso, são atravessados rapidamente por que sua distribuição não se adapta bem aos restantes das observações. Esses erros, portanto, não dependem da confusão intrínseca de palavras de acústica semelhantes, mas, principalmente, pela falta de boa modelagem da duração do evento acústico, o que produz hipótese fracamente relacionada à acústica da palavra correta (Fissore, 1997; Sadaoki, 1989).

Para justificar a estrutura dinâmica dos vetores de observação do sinal de voz, nesta tese, propõe-se um sistema de reconhecimento de dígitos isolados baseado nas variações (globais e locais) das características espectrais de cada palavra e suas correlações no tempo, duas importantes características que são exploradas parcialmente pelo HMM clássico (Deng et al., 2008; Revathi, 2011; Sadaoki, 1989). Nesta tese, um sinal de voz é parametrizado em uma matriz bidimensional temporal com o número reduzido de parâmetros gaussianos para o reconhecimento. Após a parametrização, os modelos obtidos são utilizados para gerar uma base de regras nebulosas de um sistema de inferência nebuloso Mamdani. Esses parâmetros são otimizados, utilizando-se algoritmo genético de modo a se obter o melhor desempenho do sistema de reconhecimento com um número reduzido de parâmetros. Para efeitos de validação da proposta, os modelos gerados serão parametrizados em gaussianas distribuídas em matrizes bidimensionais de ordem (2×2), (3×3) e (4×4).

Os coeficientes mel-cepstrais e a Tranformada Cosseno Discreta(TCD) (Ahmed, 1974; Zhou, 2009) são utilizados para gerar os parâmetros dos modelos de voz para representação dos padrões. O interesse no uso da TCD na compressão de dados e de classificação de padrões aumentou em anos recentes, principalmente devido ao fato de o seu desempenho ser muito próximo dos resultados obtidos pela Transformada de Karhunen-Loèvé, considerada ótima para uma variedade de critérios, tais como erro quadrático médio de truncamento e entropia (Effros et al., 2004; Fu, 1968; Hua, 1998). Este trabalho demonstra o potencial dos *mel-frequency cepstrais coefficients* (MFCCs), da TCD e de sistema de inferência nebuloso aplicados à solução do problema de reconhecimento de voz utilizando poucos parâmentros na representação dos

modelos (Azar, 2008; Zeng, 2006).

1.2 Motivação

Os problemas apresentados pelo HMM clássico diminuem sua capacidade discriminatória e erros podem ocorrer no processo de reconhecimento devido à modelagem inadequada dos padrões que serão utilizados no processo de reconhecimento. Para minimização deste problema, várias alternativas foram propostas (Ferguson, 1980; Levinson, 1986; Rabiner, 1989; Ramesh, 1992). Contudo, observa-se, pelos algoritmos apresentados nos trabalhos citados, que todos envolvem custo computacional, bem como considerável aumento da complexidade do formalismo matemático. Sistemas híbridos baseados na combinação de técnicas de inteligência computacional e HMM fornecem melhorias significativas no desempenho do reconhecimento (Zeng, 2011). Em anos recentes, várias metodologias para reconhecimento de voz foram propostas usando MFCC e redes neurais artificiais (RNA)[3] (Aggarwal, 2011; Hanchate et al., 2010), HMM e máquina de vetor de suporte (HMM-SVM) (Hejazi et al., 2008). Além das técnicas combinadas, outras propostas de sistemas inteligentes para reconhecimento de voz que não utilizam HMM também foram apresentadas (Ganesh et al., 2012; Urena et al., 2012); de forma específica, vários trabalhos em reconhecimento de voz no Brasil foram apresentados na literatura especilizada (Alencar, 2008; Bresolin et al., 2008; Montalvão, 2012; Silva et al., 2012).

A capacidade discriminatória das RNA foi logo reconhecida como uma característica que poderia contribuir para a melhoria dos sistemas de reconhecimento de voz. Contudo, o modelamento da variabilidade da duração dos sinais de voz torna complexa a aplicação direta nas RNA. Para resolver tal problema, uma variedade de técnicas híbridas com HMM e RNA foram propostas com o intuito de manter a capacidade discriminatória das RNA combinada à capacidade do HMM de modelar temporalmente o sinal de voz (Urena et al., 2012; Zeng, 2011). No entanto, a análise com RNA e HMM pode requerer uma grande quantidade de dados, que pode comprometer o desempenho computacional do sistema. Além disso é necessário uma atenção especial quanto à maldição da dimensionalidade em tais propostas (Haykin, 2009).

As contribuições da metodologia proposta neste trabalho podem ser enumeradas como segue:

1. Ela utiliza um número reduzido de parâmetros na representação do sinal de voz para reconhecimento de dígitos isolados. O sinal de voz é parametrizado em uma matriz bidimensional através de coeficientes mel-cepstrais(MFCC) e coeficientes TCD. A etapa

[3]Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamentos simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso (Haykin, 2009).

de parametrização é baseada na análise de segmentos, janelamento, análise de Fourier de curto prazo, sobreposição de janelas e geração dos coeficientes mel-cepstrais e TCD com o objetivo de reter características temporais e espectrais importantes para tornar eficiente o processo de reconhecimento com um número reduzido de parâmetros.

2. Ela utiliza um sistema inteligente nebuloso para classificar e reconhecer os dígitos. A média e a variância dos coeficientes MFCC e TCD de cada padrão dos dígitos a serem reconhecidos são variáveis de entradada de um sistema de inferência nebuloso otimizado pelo algoritmo genético. A base de regras do sistema nebuloso é gerada a partir do conhecimento do especialista com intuito de reduzir a complexidade computacional e evitar a maldição da dimensionalidade.

3. Apesar de o modelamento do sinal de voz utilizar Gaussianas, não se utilizam cálculos probabilísticos em nenhuma das estapas do processo de reconhecimento.

Dessa forma pretende-se demonstrar, com a metodologia proposta, uma solução viável para modelagem das variabilidades do sinal de voz, utilizando um número reduzido de parâmetros para codificação dos padrões a serem usados no processo de reconhecimento de voz.

1.3 Formulação do Problema

É bem difundido que o problema de reconhecimento de voz pode ser modelado como segue: supõe-se um vocabulário J de palavras a serem reconhecidas, um conjunto de treinamento com T observações de cada palavra $j \in J$ dadas por $O = \{O_1, O_2, ..., O_T\}$ e um conjunto independente de testes com m observações (Fink, 2014; Rabiner, 1993; Sadaoki, 2000). Ao se utilizarem as T observações do conjunto de treinamento para estimar os parâmetros de cada palavra j, dado por λ^j, deve-se responder uma questão relevante para um dado modelo: quão bem este modelo descreve o padrão? Para este fim a probabilidade a *posteriori*[4] $P(O/\lambda^j)$ das observações de uma dado modelo, ou pelo menos uma aproximação razoável deve ser calculada. Por outro lado, esta probabilidade indica quão bem algum modelo λ^j é capaz de representar as propriedades estatísticas das observações $O = \{O_1, O_2, ..., O_T\}$. Assim a função de verossimilhança de j pode ser calculada:

$$P\left(\lambda^j/O\right) = \frac{P\left(O/\lambda^j\right) P\left(\lambda^j\right)}{P\left(O\right)} \tag{1.15}$$

Analisando a equação (1.15), a probabilidade $P\left(O\right)$ de observações representa uma quantidade irrelevante para a classificação, pois independe dos parâmetros λ^j. Assim, pode-se utilizar

[4]Termo em Latim que significa posterior, e que no contexto refere-se à probabilidade calculada depois dos eventos observados.

a máxima verossimilhança para estimar o valor de j a partir λ^j, sendo suficiente considerar o numerador da equação (1.15):

$$j^* = \underset{j \in J}{\operatorname{argmax}} P\left(\lambda^j / O\right)$$

$$j^* = \underset{j \in J}{\operatorname{argmax}} \frac{P\left(O/\lambda^j\right) P\left(\lambda^j\right)}{P\left(O\right)}$$

$$j^* = \underset{j \in J}{\operatorname{argmax}} P\left(O/\lambda^j\right) P\left(\lambda^j\right) \tag{1.16}$$

Para que a equação (1.16) seja aplicada, a probabilidade a *priori*[5] $P\left(\lambda^j\right)$ dos parâmetros individuais precisa ser especificada. Como o cálculo desta probabilidade não é trivial, ela é frequentemente desprezada para simplificar-se a classificação e, assim, a decisão é feita somente com base na probabilidade $P\left(O/\lambda^j\right)$; considera-se também que os modelos que descrevem as diferentes classes são definidos independemente uns dos outros.

Vale ressaltar que, na metodologia proposta neste trabalho, não se utiliza o cálculo de probabilidade a *priori*, nem se aplica a teoria de representação de estados, fundamentações básicas do HMM, mas utiliza-se a modelagem temporal linguística, através de inferência nebulosa, baseada em um conjunto reduzido de parâmetros obtidos pelas codificação bidimensional ($MFCC \times TCD$). Apesar de se utilizar o HMM para efeito de comparação de desempenho com a metodologia proposta, não será abordado o modelamento do sinal de voz pelo HMM. Além da comparação com o HMM convencional, com o intuito de comparar o desempenho do classificador proposto nesta tese, os mesmos conjuntos de parâmetros utilizado pelo $IMSR$ serão apresentados a outras metodologias utilizadas para classificação, a saber: *Gaussian Mixtures Models-Expectation Maximization-GMM-EM* e máquina de vetor de suporte-SVM, técnica que também utiliza os princípios de inteligência computacional.

1.4 Revisão Bibliográfica

O HMM tornou-se a técnica mais aplicada em Sistemas Automáticos de Reconhecimento de Voz. Em contrapartida ao uso do HMM, o emprego de técnicas baseadas em inteligência computacional para reconhecimento de voz aumentou significativamente nas últimas décadas. Na literatura especializada, destacam-se os sistemas inteligentes e os sistemas híbridos baseados em combinações de técnicas inteligentes com HMM. Estes sistemas apresentam melhorias na capacidade de reconhecimento. Entretanto, o problema encontrado em tais técnicas é o

[5]Termo em Latim que significa anterior, e que no contexto refere-se à probabilidade que é pressuposta dos modelos observados.

aumento da carga computacional devido à complexidade dos reconhecedores, que em várias situações mostram efetivamente bom desempenho, mas com considerável carga computacional. O problema da carga computacional excessiva em sistemas inteligentes e/ou híbridos se dá principalmente pela quantidade de entradas e das complexidades paramétricas envolvidas no processamento dessas entradas. Recentemente vários trabalhos usando sistemas inteligentes e sistemas híbridos foram propostos na literatura especializada. A seguir, apresentam-se algumas dessas propostas.

Hanchate et al. propuseram um sistema de reconhecimento de dígitos vocais utilizando redes neurais perceptrons multicamadas. O sinal de voz foi codificado utilizando-se coeficientes mel-cepstrais, algoritmos de segmentação e de determinação dos níveis de energia por segmento para determinação dos pontos de silêncio com o objetivo de detectarem-se os pontos de início e término do sinal de voz. A rede neural com alimentação direta foi configurada com uma camada oculta com função de ativação sigmoidal, e a camada de saída contendo 10 neurônios com funções de ativação linear. A rede foi treinada com 10, 30, 50 a 70 neurônios na camada oculta, utilizando o algoritmo de retro-alimentação (*backpropagation*). O sinal de voz foi segmentado em 4 segmentos; cada segmento, com 256 amostras, foi codificado com 12 coeficientes mel-cepstrais (Hanchate et al., 2010). Os dez dígitos da língua inglesa foram gravados de pronúncias de locutores masculinos e femininos. O banco de voz consiste de 10 repetições de cada dígito produzido por cada locutor. Todas as amostras para um dado locutor foram gravadas em uma única seção. Subconjuntos do banco de voz foram utilizados nas etapas de treinamento e testes.

Azam et al. (Azam et al., 2007) apresentam um trabalho de reconhecimento de dígitos na linguagem urdu, paquistanesa, usando redes neurais como classificador. O pré-processamento do sinal de voz foi realizado com filtragem digital de pré-ênfase, janelamento de Hamming, com 256 amostras, segmentado em 67 segmentos e gerando 39 coeficientes mel-cepstrais, onde foram utilizados os coeficientes de maiores amplitudes no segmento. A rede foi treinada com algoritmo *backpropagation*, com várias arquiteturas, mas a que apresentou o melhor resultado foi a rede configurada com 39 neurônios na camada de entrada, 19 neurônios na camada oculta e 10 neurônios na camada de saída. O sistema de reconhecimento de dígitos isolados urdu é dependente de locutor. A base de dados contém 1000 pronúncias para os 10 dígitos, para cada dígito há 100 repetições. 500 pronúncias foram utilizadas para treinamento e 500 foram utilizadas para teste.

R.K. Aggarwall e M.Dave propõem um sistema de reconhecimento de voz para língua Hindi com redes neurais *perceptrons* de multi-camadas com otimização utilizando algoritmo genético. O pré-processamento do sinal foi feito através de segmentação, janelamento e análise com extração dos coeficientes mel-cepstrais. A detecção do silêncio no início e final do sinal foi

realizada através da análise da energia de curto prazo.

A rede utilizada foi configurada com 182 neurônios na camada de entrada, que dependem do número de coeficientes mel-cepstrais (MFCC) por segmento, sendo utilizados 13 coeficientes MFCC e 14 segmentos, com duas camadas ocultas, sendo a primeira configurada com 40 neurônios, a segunda configurada com 15 neurônios e a camada de saída configurada com 10 neurônios. Um Algoritmo genético foi aplicado, inicialmente para otimizar três características da rede, a saber: conexões dos pesos sinápticos, arquitetura da rede (topologia e função de transferência) e o algoritmo de aprendizagem. A base de dados consiste de pronuncias de 10 locutores masculinos e 10 locutores femininos. Cada um do 10 locutores participou com cinco pronuncias para cada palavra. A base de dados foi dividida em dois subconjuntos, um para a etapa de treinamento e outro para a etapa de teste. O primeiro subconjunto foi composto com os três primeiros exemplos de cada palavra pronunciada, enquanto o segundo subconjunto foi composto com os dois exemplos restantes (Aggarwal, 2011).

No trabalho de Hejazi e Ghaemmaghami (Hejazi et al., 2008), apresenta-se um sistema de reconhecimento baseado no modelo híbrido de HMM-SVM. O banco de vozes utilizado consiste de dígitos persas, foi gravado no "*Speech Laboratory of the Eletronic Research Center of Sharif University of Technology*"e foi composto por aproximadamente 400 amostras de dígitos pronunciados por homens e cinquenta amostras de dígitos pronunciadas por mulheres; em ambos os casos, os indivíduos eram de diferentes idades. Foram usadas 330 amostras para treinamente e 120 amostras para testes.

O pré-processamento do sinal de voz foi realizado com 13 coeficientes cepstrais distribuídos na escala Bark com 27 filtros triangulares. Foi realizada também a detecção ativa de voz para detecção de silêncio. O sinal de voz foi dividido em segmentos para alimentar a entrada do HMM. Este trabalho explora a característica de mudança de espaço do SVM, através da função kernel, e da área de busca restrita fornecida pelo HMM. O SVM utilizou a função Kernel ERBF (*Error Radial Basis Function*). Percebeu-se que a aplicação do sistema híbrido HMM-SVM apresentou melhores resultados quando comparados com o HMM clássico; contudo, a mudança de dimensionalidade introduzida pelo SVM e, por consequência, sua carga computacional não foi comentada pelos autores.

(Ganesh et al., 2012) apresentam um sistema híbrido para reconhecimento de dígitos da língua inglesa utilizando-se a Detecção Ativa de Voz (VAD- *Voice Activity Detection*) e um Algoritmo de Melhoramento de Voz (SEA - Speech Enhancement Algorithm). O sinal de voz com ruído foi decomposto em segmentos de 25 ms com janelas de 10 ms. Para a redução do ruído, utilizaram a teoria do filtro de Wiener, no qual a atenuação é uma função da relação sinal-ruído (SNR) do sinal de entrada; logo após, o sistema proposto realiza a análise de autocorrelação de

cada segmento. Então, os coeficientes de predição linear são calculados; em seguida, espaçados na escala mel, gerando 15 coeficientes mel-cepstrais. O banco de voz foi composto por palavras isoladas (10 dígitos da língua inglesa) pronunciadas por 100 locutores masculinos e femininos com idades entre 15 e 25 anos.

Para a etapa de treinamento foi utilizado o HMM com seis estados, gerando-se os padrões para o reconhecimento. O reconhecimento é realizado através de um processo de comparação entre o sinal de voz desconhecido e os padrões gerados no treinamento através de uma medida de similaridade através da estimativa de máxima verossimilhança.

James K. Tamgno et.al apresentam em (Tamgno et al., 2012), um sistema de reconhecimento de dígitos de vocabulário limitado da linguagem *Wolof*[6] baseado em uma variação do HMM elaborado com *software* livre denominado *HMMToolKit*. A extração de características foi realizada considerando o sinal de voz contínuo, quase estacionário em curto prazo. Além disso, foi utilizada a técnica de janelamento deslizante para dividir o sinal de voz em segmentos com tamanhos entre $20ms$ a $30ms$. De cada segmento foram calculados os coeficientes mel-cepstrais e suas derivadas de primeira e segunda ordem, que também foram consideradas na codificação do sinal de voz. A técnica utilizada considerou o HMM como uma rede de um conjunto de estados conectados, sendo que cada estado representa uma parte do padrão a ser obtido.

No trabalho de George E. Sakr e Imad H. Elhajj (Sakr and Elhajj, 2011), descreve-se um sistema de reconhecimento de dígitos da língua inglesa com medida de confidência baseado na teoria da dimensão de Vapnik e Chervonenskis do algoritmo de aprendizagem. A máquina de vetor de suporte utilizada neste trabalho foi configurada com uma função kernel Gaussiana, com objetivo de escolher o melhor parâmetro definido como o inverso da variância.

Rubén Solera-Ureña et al. (Urena et al., 2012) apresentaram um sistema robusto em tempo real de reconhecimento automático de voz na língua espanhola usando máquina de vetor de suporte compacta, que também foi baseado no sistema híbrido HMM-SVM e que usa o algoritmo dos mínimos quadráticos ponderado no processo de treinamento com objetivo de modelar o sinal de voz em um modelo semiparamétrico compacto para o SVM, com intuito de reduzir a quantidade de dados utilizados no processo de reconhecimento.

(Hassanzadeh et al., 2012) apresentaram um sistema de reconhecimento com título "*A Speech Recognition System Based on Structure Equivalent Fuzzy Neural Network Trained by Firefly Algorithm*". Neste trabalho é ressaltado o problema das esturuturas de redes neurais nebulosas na geração automática e adaptação das funções e regras nebulosas ao problema de reconhecimento de voz. Para sobrepor este problema os autores propõem um sistema denoninado estrutura equivalente para rede neural nebulosa (*Structure Equivalent Fuzzy Neural Network*-SFNN)

[6]Língua falada no Senegal, Gambia e Mauritânia, é a línguagem da etinia africana Wolof.

otimizada com o algorítmo inspirado em inteligência coletiva da população de vaga-lumes. Os parâmetros da SFNN, devidamente otimizados foram utilizados no processo de reconhecimento. O sistema de reconhecimento foi composto de três etapas, o pré-processamento, extração das características e treinamento da rede e reconhecimento da voz. Neste trabalho, foram utilizadas palavras isoladas, 50 palavras foram pronunciadas por 16 pessoas em ambientes de 15dB, 20dB, 25dB, 30dB e ambiente acústico. Cada locutor pronunciou o conjunto de palavras três vezes. Nove locutores foram utilizadas para o treinamento da rede e 7 locutores foram separados para os testes. A rede neural foi estruturada com vetores de características com 1024 elementos. O número de nós da camada de entrada da rede é igual ao número de elementos do vetor de características. A camada de decisão da rede é determinada pelo algoritmo de subtração de clusteres, através da extração de possíves centros de clusteres dos dados de entradas pela média de todos os centros dos dados de treinamento. Na etapa de treinamento, o número de regras para a rede neural é de 32. O número de saídas é igual ao número de padrões que devem ser reconhecidos.

No trabalho intitulado "*Spoken Digit Recognition in Portuguese Using Line Spectral Frequencies*"(Silva et al., 2012), Diego F. Silva et al. apresentam um algoritmo baseado em máquina de aprendizado para reconhecimento de dígitos. Nesse trabalho foi utilizado "*Line Spectral Frequencies*-(LSF)"para obter um conjunto de coeficientes preditivos para o reconhecimento proposto. O trabalho apresenta os resultados do reconhecimento de dígito, utilizando LSF e MFCC. A base de dados utilizada consiste de dígitos falados em português brasileiro, gravados durante um período de três meses, de oitenta e dois homens entre 18 e 42 anos de idade. A frequência de amostragem foi de $22.050Hz$ e a sobreposição entre as amostras dos segmentos do sinal de voz foi de 75%. A base de dados foi composta por 216 sequências de 10 dígitos (0-9) cada, com um total de 10 classes e 2.160 exemplos. Os parâmetros para entrada da máquina de aprendizado foram gerados, para efeito de comparação, nas seguintes quantidades: 13 MFCC contra LSF de ordem 24 e 48. Como os LSF são as raízes dos polinômios, então utilizou-se um total de 24 e 48 coeficientes LSF, respectivamente. Para ambas estratégias, MFCC e LSF, as características foram extraídas via janelamento dinâmico. A largura e o tamanho do passo foram de tal forma que um conjunto de vetores com 25 parâmetros por vetor adjacente foi gerado. Portanto, cada método utilizado para extração das características consistiu de $25 \times n$, onde n é o número de características extraídas. Desse modo, cada sinal foi transformado em uma instância com 325, 600 e 1.200 atributos, para 13-MFCC, 24-LSF e 48-LSF, respectivamente.

1.5 Organização do Trabalho

Este trabalho está organizado em capítulos a saber:

Capítulo 2: Consiste de apresentação sucinta das características da voz, sua fisiologia, bem como o modelamento matemático do sinal de voz e alguns comentários sobre os métodos mais utilizados. Aborda-se, também, neste capítulo o modelamento espectral do sinal de voz, ressaltando-se de forma simplificada os tópicos fundamentais ao reconhecimento de voz. Comentam-se os métodos e técnicas matemáticas mais utilizadas no processamento de voz dependente do tempo e na representação paramétrica do sinal de voz. Faz-se também uma análise sobre os bancos de filtros e análise cepstral.

Capítulo 3: Inicia-se este capítulo com a descrição das técnicas utilizadas no pré-processamento do sinal de voz utilizado nesta tese, bem como todas as etapas elaboradas no sistema de reconhecimento proposto. Descreve-se de modo detalhado a metodologia utilizada na geração da matriz bidimensional, modelo dos padrões e o sistema de inferência nebuloso utilizado. Descreve-se também a formulação do algoritmo genético utilizado.

Capítulo 4: Consideram-se os resultados experimentais obtidos bem como uma análise detalhada destes resultados através de gráficos e tabelas comparativas, entre o sistema proposto e outras técnicas apresentadas na literatura técnica para reconhecimento de voz.

Capítulo 5: Apresentam-se as conclusões obtidas, bem como melhorias no desenvolvimento desta tese de doutorado.

Apêndice A: Descrevem-se os conceitos básicos dos sistemas nebulosos que foram utilizados no desenvolvimento das regras do sistema do reconhecedor proposto na tese.

Apêndice B: Neste apêncice abordam-se os fundamentos das técnicas matemáticas de otimização utilizadas na elaboração do algoritmo genético utilizado para otimizar a base de regras do sistema nebuloso proposto com objetivo de melhorar o desempenho do sistema.

Apêndice C: Descrevem-se os conceitos básicos da metodologia Máquinas de Vetor de Suporte (*Support Vector Machine*-SVM) utilizada na comparação com o sistema de reconhecimento proposto.

Apêndice D: Descrevem-se os conceitos básicos da metodologia Modelos de Misturas Gaussianas (*Gaussian Mixtures Models*-GMM) utilizada também para efeito de comparação de desempenho de reconhecimento com a metodologia proposta.

Capítulo 2

Características Fundamentais da Voz

2.1 Fisiologia da Voz

A voz é uma onda de pressão acústica que se origina a partir dos movimentos fisiológicos voluntários dos órgãos vocais humanos. A região vocal, propriamente dita, é considerada um tubo acústico, não uniforme, com cerca de 17 cm de comprimento. Este tubo se inicia na região das cordas vocais e termina nos lábios, por onde o sinal é parcialmente irradiado. A área transversal desta região é determinada pela posição dos lábios, maxilar, língua e véu palatino, podendo variar desde zero até cerca de 20 cm^2, com o maxilar e lábios totalmente abertos (Fant, 2004; Rabiner, 1978; Sadaoki, 2000).

Durante a geração dos sons nasais, a região nasal é acoplada à região do trato vocal através da ação do véu palatino. Essa região adicional inicia-se no véu palatino e termina nas narinas; é constituída de uma cavidade com cerca de 12 cm de comprimento. Na geração de sons não nasais, o véu palatino bloqueia a cavidade nasal de modo que nenhum som seja irradiado pelo nariz. No processo de fonação, o ar armazenado nos pulmões é expelido e ao mesmo tempo, forçado através da traqueia para a abertura das cordas vocais (glote), constituindo-se, desse modo, a fonte de excitação para a produção dos sons da fala. A região do trato vocal modelada pelos órgãos articuladores forma um tubo acústico (ou cavidade) ressonante, de tal modo que o fluxo de ar que o atravessa seja modelado por tais cavidades produzindo os diferentes sons de voz (Campbell, 1997; Cegalla, 2008; Rabiner, 1993).

2.2 Tipos de Sons

Os sons produzidos por um sistema vocal podem ser classificados primariamente em três classes distintas de acordo com seu modo de excitação (Bechara, 2009; Cegalla, 2008; Rabiner,

2007):

Os sons sonoros são gerados elevando-se a pressão do ar nos pulmões e em seguida forçando-o através da abertura das cordas vocais (glote). Isso faz com que elas vibrem a intervalos aproximadamente regulares entre $3,5\ ms$ e $12\ ms$. Esta interrupção no fluxo de ar produz sons de natureza periódica. A taxa de vibração depende do locutor (sexo, idade, etc.), da pressão do ar nos pulmões, e também do comprimento, da espessura, e da tensão das cordas vocais. Quanto maior for a tensão, maior será a taxa de vibração, que é definida como a frequência fundamental da voz. Exemplos de sons sonoros: /a/, /e/, /i/, /o/,/u/, /z/, /v/, /r/.

Os sons surdos são gerados mantendo-se aberta a glote e formando uma constrição com auxílio dos órgãos articuladores em algum ponto da região do trato vocal. Em seguida uma quantidade de ar suficiente é forçada através desta oclusão produzindo uma turbulência que gera os sons fricativos, cuja forma de onda no domínio do tempo é semelhante a um ruído aleatório. Por exemplo: /s/- pronunciado como s̈h:

Os sons plosivos (ou oclusivos) são os resultados de se fazer, através dos órgãos articuladores, um fechamento completo da região do trato vocal, usualmente na direção frontal do mesmo (lábios). A pressão do ar é aumentada no ponto de oclusão e subitamente liberada. Desse modo estes sons são caracterizados por um instante de silêncio seguido de uma breve explosão; por exemplo, citam-se os sons /p/, /b/, /t/, /d/, /k/, /g/. As fontes vocais, para os sons classificados acima, apresentam um espectro relativamente plano. O sistema vocal atua como um filtro acústico, variante no tempo, no sentido de impor suas características ressonantes nas fontes. Assim, pode-se dizer que o mecanismo de produção da voz consiste de três partes principais: a fonte de excitação para a geração da fala, o trato vocal com seus articuladores e a radiação da voz. Estas partes são descritas na figura 2.1.

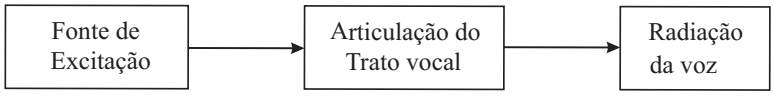


Fig. 2.1: Processo de produção da fala.

2.3 Forma de onda do sinal de voz

A fonte de excitação para a geração da voz é classificada em duas partes distintas: sonora (para sons sonoros), com um sinal periódico mostrado na figura 2.2, e surda (não-sonora), com um sinal semelhante a um ruído mostrada na figura 2.3.

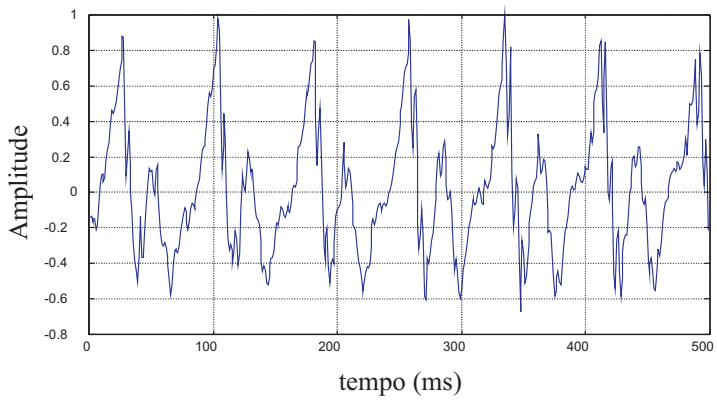


Fig. 2.2: Forma de onda de um sinal de voz sonoro.

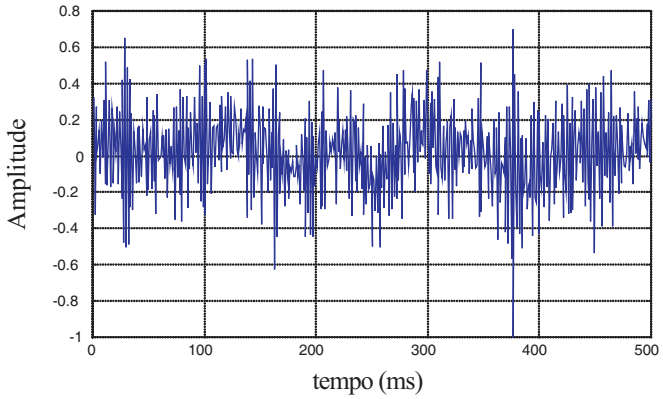


Fig. 2.3: Forma de onda de um sinal de voz surdo.

A intensidade relativa do sinal varia com o tempo de níveis relativamente baixos a níveis relativamente altos, podendo variar em uma faixa dinâmica de até 60dB, aproximadamente. Observando-se ambas as figuras citadas, verifica-se que as formas de ondas podem ser classificadas em dois grupos, a saber:

1. O sinal ilustrado na figura 2.2 apresenta características periódicas bem definidas, correspondendo a som sonoro (/a/, /e/, etc.). No caso dos trechos periódicos o período de repetição é chamado de período de tom (”pitch”) do sinal e seu inverso é chamado de

frequência fundamental f_0 e é gerada pela excitação pulsante, quase periódica, das cordas vocais (Picone, 1993; Sadaoki, 2000).

2. O sinal apresentado na figura 2.3 assemelha-se bastante a um sinal de ruído aleatório, correspondendo às consoantes surdas (/s/, /f/,... etc.) onde a amplitude do sinal varia aleatoriamente com o tempo (Rubin, 1998).

As vogais e outros sons sonoros são moldados pelas cavidades ressonantes na região do trato vocal. Como estes sons possuem formas de ondas periódicas ou quase periódicas, eles apresentam, consequentemente, um espectro harmonicamente relacionado, onde f_0 é o espaçamento entre os harmônicos e $1/f_0$ é o intervalo entre fluxos de ar sucessivos. A variação da frequência fundamental é a base física da entonação (Picone, 1993).

O trem de sucessivos pulsos de ar emergindo da vibração das cordas vocais é a fonte primária para os sons sonoros. As cavidades dentro da região do trato vocal atuam como filtros multi-ressonantes que ressaltam as frequências predominantes que caracterizam os sons vocálicos e por isso são chamados de frequências formantes.

As frequências das três formantes mais baixas são as principais responsáveis pela qualidade fonética[1] (timbre) de uma vogal. Elas variam continuamente sobre uma moderada faixa de frequências. Dependem do sexo do locutor e da entonação; às vezes desaparecem do espectro, indicando transição de sons ou deslocamento de sinal sonoro para ruidoso ou vice-versa.

As consoantes não sonoras (surdas) são produzidas pela turbulência do fluxo de ar direcionado para algum ponto de contração do trato vocal (o som da consoante /s/ é produzido por uma fresta entre a ponta da língua e as gengivas superiores). Nesses casos, a fonte primária gera, também, um sinal semelhante a um ruído aleatório sem periodicidade definida (Bresolin, 2008; Rabiner, 1993).

2.4 Modelo linear do trato vocal para a produção da voz

Muitos modelos para sistema de produção de voz são baseados na teoria de filtros-fontes de Fant (Fant, 1960, 1981). Seguindo este modelo, a voz pode ser considerada como o resultado da convolução entre a fonte de excitação e o sistema de filtros que modelam o trato vocal; isto é, a fonte representa o fluxo de ar no trato vocal e os filtros representam as ressonâncias do trato vocal que são variantes no tempo. Para sons sonoros, a excitação é modelada como uma série de pulsos, enquanto que, para sons surdos, a excitação tem características de um ruído

[1]Fonética: Estudo da produção, das características e da percepção dos sons e da fala isolados (Mesquita, 1998).

aleatório. A região do trato vocal atua como filtro variante no tempo, de modo a impor suas características ressonantes na fonte de excitação.

Na figura 2.4, ilustra-se o modelo linear de produção de sinais de voz desenvolvido por (Dias, 2012; Fant, 2004). Nota-se que esse modelo representa um caso especial no qual nenhuma previsão é feita para representar entradas misturadas (para sons fricativos) ou para acoplar um ramo de filtro para simular sons nasais (Fant, 1981). Este modelo consiste de três seções lineares em que se supõem a fonte e os filtros como sistemas independentes e separados, e que no domínio do tempo a voz pode ser representada pela convolução dos elementos que compõem o sistema, isto é:

$$s[n] = g[n] * v[n] * r[n] \tag{2.1}$$

onde, $g[n]$, $v[n]$ e $r[n]$ são, respectivamente, resposta ao impulso do sinal da fonte de excitação, resposta impulso do trato vocal e a radição dos lábios e nariz.

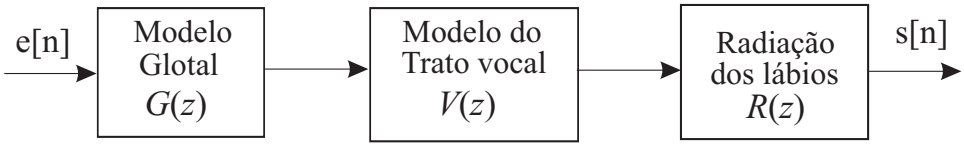


Fig. 2.4: Modelo linear de produção da voz.

A entrada do sistema é representada por um sinal que assume a forma de um trem de pulsos, separados pelo período de tom, quando o sinal de voz tem características sonoras, e assume a forma de um ruído aleatório, com espectro plano, quando o sinal não é sonoro. O fluxo de ar que passa pela região do trato vocal através da glote é modelado por um filtro passa-baixas com dois polos, cuja função de transferência é dada por (Markel, 1976):

$$G(z) = \frac{1}{(1 - e^{-cT} z^{-1})^2} \tag{2.2}$$

sendo que $G(z)$ representa a transformada z do sinal $g[n] = g(nT)$, sendo $g(nT) = g(t)$ amostrado a cada T segundos; c é a velocidade do som.

O Modelo aproximado da região do trato vocal é constituído por um filtro com apenas polos, com k frequências formantes que correspondem aos pólos da função $V(z)$ e representam as ressonâncias do trato vocal. O modelo do trato vocal é aproximado pela equação:

$$V(z) = \frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{2.3}$$

onde os coeficientes a_k dependem do formato da área transversal do trato vocal.

A radiação dos lábios e narinas pode ser modelada como uma derivada de primeria ordem no domínio do tempo, que pode ser aproximada por um filtro passa alta dado por:

$$R(z) = 1 - \alpha z^{-1} \tag{2.4}$$

sendo α o coeficiente de radiação e possui valor entre 0.95 e 0.99, de modo que o zero esteja dentro do círculo unitário no plano z (Dias, 2012; Javkin et al., 1987).

2.5 Processamento do sinal de voz no domínio do tempo

A voz é uma onda de pressão que deve ser convertida em valores numéricos para processamento por dispositivos digitais. para se converter a onda de pressão em valores numéricos, alguns dispositivos são necessários: Um microfone permite que uma onda de pressão acústica seja convertida em um sinal elétrico $s(t)$. Para que este sinal seja trabalhado através de processadores digitais de sinais, faz-se necessário que o mesmo seja convertido em um sinal digital.

Há várias formas de representar-se digitalmente o sinal de voz, que podem ser genericamente divididas em:

1. **Representação da forma de onda**: onde não há uma codificação da fonte, obtida fundamentalmente pela amostragem e quantificação, preservando a forma do sinal. Incluem-se, por exemplo, sistemas *Pulse Code Modulation*-PCM, *Differential Pulse Code Modulation*-DPCM, *Pulse Amplitude Modulation*-PAM (Lathi, 1998), etc.

2. **Representação paramétrica**: obtida pela codificação de parâmetros de um modelo de produção de voz, geralmente, mas não necessariamente, classificados em parâmetros de excitação e parâmetros da resposta da região vocal. Incluem-se aqui os vários métodos de análise/síntese.

Uma sequência típica de um sinal de voz é ilustrada na figura 2.5. Observa-se na figura que as propriedades do sinal de voz são variantes no tempo. Devido a essas mudanças nas caracteríticas temporais do sinal de voz, qualquer técnica de processamento de voz no domínio do tempo deve ser capaz de fornecer representações úteis das características dos sinais, tais como intensidade, modo de excitação, período de "*pitch*" e possivelmente, características do trato vocal, tais como frequências formantes[2].

[2]Frequências de ressonâncias do trato vocal (Rabiner, 1993).

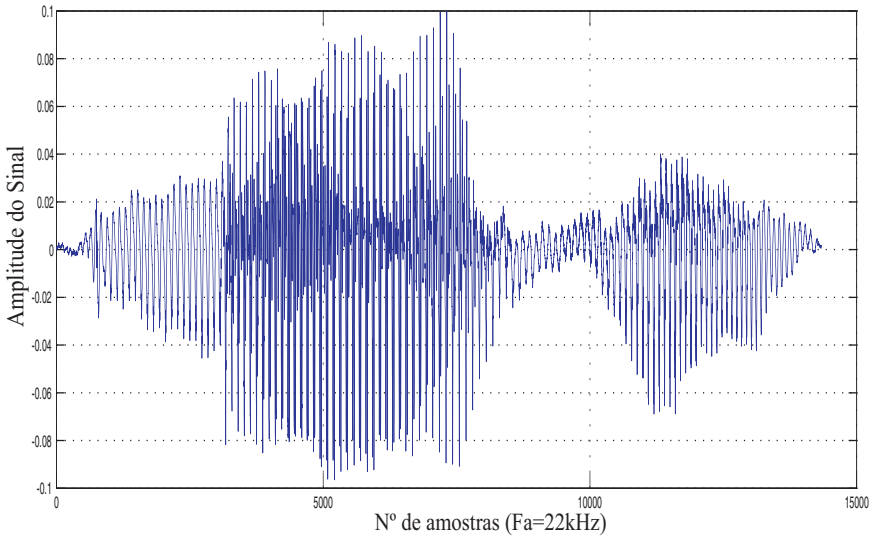


Fig. 2.5: Sinal de voz no domínio do tempo.

A hipóste básica para processamento de voz no domínio no tempo, para a maioria dos sistemas de processamento de voz, é que as propriedades do sinal de voz mudam de forma relativamente lenta com o tempo. Esta hipótese leva a uma variedade de metodologias de processamento de períodos curtos de tempo, onde pequenos segmentos do sinal de voz são isolados e processados como se fossem o sinal de voz contínuo e com propriedades fixas. Isto é, repetido tantas vezes quanto necessário. É comum que esta análise de segmentos sobreponha os segmentos, em algum nível, entre si. O resultado do processamento de cada segmento pode ser um número ou um conjunto de números. Desse modo, esta forma de análise resulta em uma nova sequência depedente do tempo, que pode ser utilizada como uma representação do sinal de voz.

2.5.1 Amostragem do sinal

Para que um sinal de voz seja adequadamente processado, após a conversão da onda de pressão em sinal elétrico, codifica-se o sinal elétrico em amostras quantificadas. Para este fim, normalmente utiliza-se um conversor Analógico-Digital (A/D) que converte um sinal analógico em uma sequência digital. A entrada do conversor, $s(t)$ é uma função de valores reais de uma variável contínua t. A saída do conversor A/D é uma cadeia de bits que corresponde a uma sequência de $s[n]$ de tempo discreto, com uma amplitude quantizada para cada valor de n dentro de um conjunto de números finitos. Um conversor A/D pode ser dividido em três etapas, amostragem, quantização e codificação. Serão tratadas aqui as etapas de amostragem e quantização. A etapa de codificação não será abordada, pois não foi utilizada na elaboração

da metodologia proposta.

Embora existam outras possibilidades, um método típico de obter-se uma representação de tempo discreto de um sinal de tempo contínuo é através de uma amostragem periódica, transformando o sinal contínuo $s(t)$ em uma sequência de amostras $s[n]$. Na amostragem, um sinal $s(t)$ de largura de banda limitada em frequência em $B\,(Hz)$, isto é, com espectro nulo fora da largura de banda B, é determinado univocamente por amostras uniformemente espaçadas no tempo de T_a desde que este espaçamento seja $T_a < \frac{1}{2B}$, onde $F_a = \frac{1}{T_a}$ é denominada frequência de amostragem. O sinal de tempo discreto, devidamente amostrado, é representado pelo produto do sinal $s(t)$ por um trem de impulsos:

$$s[n] = s(t) \cdot \sum_{n=-\infty}^{+\infty} \delta(t - nT_a) = \sum_{n=-\infty}^{+\infty} s(nT_a)\delta(t - nT_a) \tag{2.5}$$

Aplicando-se o teorema da convolução em frequência e suas propriedades, obtém-se,

$$S(f) = S(f) * \frac{1}{T_a} \sum_{n=-\infty}^{+\infty} \delta(f - nF_a) = \frac{1}{T_a} \sum_{n=-\infty}^{+\infty} S(f - nF_a) \tag{2.6}$$

Na figura 2.6 observa-se o espectro de um sinal filtrado com um filtro passa-baixa com banda passante de largura B, e este sinal será amostrado com uma frequência de amostragem F_a. Quando se realiza uma amostragem numa certa frequência obtém-se como resultado a repetição periódica do espectro do sinal $s(t)$, com período de repetição F_a, conforme ilustrado na figura 2.7.

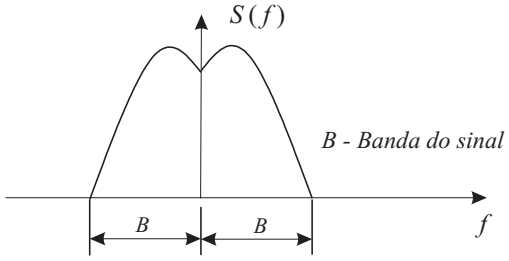


Fig. 2.6: Sinal de informação no domínio da frequência.

Se não houver superposição entre os ciclos do espectro periódico, recupera-se o sinal original $s(t)$, passando-se o sinal amostrado $s[n]$ por um filtro passa-baixa com frequência de corte $f_c = B$. Para que os espectros parciais não se superponham, é necessário que a frequência mais baixa do espectro centrado em F_a esteja acima da frequência mais alta do espectro centrado em zero, isto é:

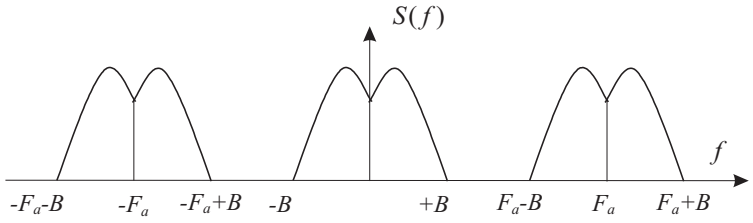


Fig. 2.7: Sinal amostrado no domínio da frequência.

$$F_a - B \geq B \Rightarrow F_a \geq 2B \tag{2.7}$$

Desta imposição surge o enuciado do cretério de Nyquist (Shannon, 1948), onde, se descreve que a frequência mínima de amostragem é maior ou igual ao dobro da maior frequência do sinal amostrado. Como na prática os filtros não são ideais, faz-se necessária uma banda de guarda para a transição dos filtros. Para sinais de voz é comum limitar-se a largura de banda em $4kHz$ e a frequência de amostragem igual a $F_a = 8kHz$, que corresponde a um espaçamento de $T_a = 125\mu s$ (Lathi, 1998; Oppenheim, 2013).

2.5.2 Filtragem do sinal de voz

O objetivo básico do processamento digital do sinal de voz é deduzir um conjunto de parâmetros denominados perceptualmente significantes, isto é, parâmetros que, de alguma forma, assemelham-se àqueles utilizados pelo sistema fonador e auditivo e que possam ser processados posteriormente, onde ambos os domínios, do tempo e da frequência, podem ser utilizados para análise acústica com fins de reconhecimento (Huang et al., 1990). Os tópicos abordados no domínio do tempo, tais como parâmetros de energia e taxa de cruzamento por zeros, trabalham diretamente com a forma de onda do sinal de voz e usualmente levam a algoritmos simples de implementar. A abordagem no domínio da frequência envolve alguma forma de análise espectral e normalmente inclui características que não são diretamente evidentes no domínio do tempo.

O modelamento espectral de um sinal envolve duas operações básicas: conversão A/D e filtragem digital, com objetivo de discriminar componentes de frequências importantes no sinal. A principal utilidade do processo de digitalização da voz é produzir uma representação de dados amostrados do sinal de voz com a relação sinal-ruído (SNR) tão elevada quanto possível. Em sistemas de telecomunicações e em sistemas de reconhecimento de voz, é desejado que esta relação seja maior que $30dB$; esta é uma relação bastante adequada para alcançar excelentes

resultados com esses sistemas. Uma vez realizada a conversão A/D do sinal, o último passo no processo de digitalização do sinal é realizar uma pós-filtragem utilizando-se, normalmente, um filtro de resposta finita ao impulso (*FIR-Finite Impulse Response*).

$$H_{pre}[z] = \sum_{k}^{N} a_{pre}[k] z^{-k} \tag{2.8}$$

Esta pós-filtragem é conhecida como filtragem de pré-ênfase, onde a_{pre} são os coeficientes do filtro de pré-ênfase. Normalmente é utilizado o seguinte filtro:

$$H_{pre}[z] = 1 + a_{pre} z^{-1} \tag{2.9}$$

Para a implementação do filtro da equação (2.9), utilizam-se, como prática, valores de a compreendidos entre $[-1$ a $-0,4]$; entretanto, valores próximos a -1 ou $-(1-1/16)$ são mais comuns em sistemas de reconhecimento de voz (Picone, 1993). A função básica do filtro de pré-ênfase é elevar o espectro do sinal aproximadamente $20dB$ por década uma vez que os sons abertos da voz têm como características fisiológicas uma atenuação espectral de $20dB$ por década. Assim o filtro servirá para compensar essa atenuação natural, melhorando a eficiência da análise.

Verifica-se também que a filtragem de pré-ênfase amplifica a região de frequências acima de $5kHz$, região na qual o sistema auditivo torna-se cada vez menos sensível. Contudo essa área é considerada de pouca importância para sistemas de reconhecimento de voz e é, naturalmente, atenuada pelo sistema que produz a voz.

Algoritmos mais sofisticados de pré-enfase, têm sido apresentados na literatura, sendo considerados mais notáveis aqueles baseados em filtragem de pré-ênfase adaptativa, nos quais a envoltória do espectro é suavizado antes da análise espectral (Markel, 1976). Outros algoritmos utilizam modelos de filtros que atenuam áreas do espectro conhecidas como ruidosas. Apesar das vantagens descritas acima, atualmente, muitos dos sistemas de reconhecimento de voz eliminam completamente o estágio de pré-ênfase do sinal (Picone, 1993).

2.6 Modelamento Espectral do Sinal de Voz

A análise em frequência de sinais de tempo discreto é geralmente realizada por processadores de sinais digitais, normalmente integrados em computadores, ou em hardware digital específico. Os métodos padrões para análises espectrais baseiam-se na transformada de Fourier de uma dada sequência amostrada $x\,[n] \leftrightarrow X(e^{j\omega})$, sendo ω a frequência dada em rad/s. Contudo, $X(e^{j\omega})$ é uma função de uma variável contínua, portanto, não é conveniente para processamento de $x[n]$ em processadores digitais. A forma mais conveniente para processamento digital do

espectro de $X(e^{j\omega})$ é amostrar esse espectro e limitar o número de amostras adequadamente.

A complexidade computacional é bastante reduzida se $X(e^{j\omega})$ é avaliada somente para valores discretos de ω. Se tais valores são igualmente espaçados, então, a Transformada de Fourier Discreta (TFD) é obtida.

$$X[k] = X\left(e^{j\omega}\right)|_{\omega=\frac{2k\pi}{N}} \tag{2.10}$$

Caso o número de amostras N seja potência de 2, $N = 2^p$, com p inteiro, a complexidade computacional pode ser reduzida para $(N)log_2(N)$, podendo ser utilizado, neste caso, o algoritmo FFT[3]. Nota-se também que, se o sinal for real a FFT pode ser calculada tendo uma complexidade computacional $(N/2)log_2(N/2)$ (Becchetti, 2000). Nos itens subsequentes, abordam-se os métodos clássicos utilizados no reconhecimento de voz. Ressalta-se, neste capítulo, a importância da análise do sinal de curto prazo em detrimento à análise do sinal como um todo, enfocando-se os métodos de codificação que serão utilizados no desenvolvimento deste trabalho.

2.7 Reconhecimento Automático de Voz

Para reconhecimento de voz, como em reconhecimento de locutor e reconhecimento de palavras isoladas, são aplicadas técnicas para se obter modelos parametrizados, na fase de teste, os quais serão comparados a padrões pré-definidos armazenados na fase de treinamento. No reconhecimento de locutor, o objetivo é identificar quem falou em um grupo de locutores. Já no reconhecimento de palavras, o objetivo é determinar que palavra, frase ou sentença foi pronunciada. Ao contrário do reconhecimento de locutor, o reconhecimento de voz voltado ao que foi falado possui um número muito grande de opções que devem ser especificadas antes de se começar efetivamente a trabalhar no reconhecimento (Rabiner, 1993). Como exemplo, abaixo, citam-se algumas dessas opções:

1. Tipo de fala, isto é, palavras isoladas ou fala contínua;

2. Número de locutores: único locutor, número determinado de locutores, população indeterminada de locutores (independência do locutor);

3. Tipo de locutores: casual, masculino, feminino, criança, idoso, jovem, etc;

[3] *Fast Fourier Transform* - Algoritmo eficiente para realizar a Transformada de Fourier Discreta (Oppenheim, 2013)

4. Ambiente de locução: com controle de ruído, de baixo ruído, sem controle de ruído, local público, etc;
5. Sistema de transmissão: microfone de alta qualidade, microfone comum, telefone, etc;
6. Tipo e quantidade de sistemas de treinamento: em treino, conjunto de treinamento fixo, treinamento contínuo;
7. Tamanho do vocabulário: vocabulário pequeno (1-100 palavras), vocabulário médio (101-5000 palavras), vocabulário grande (acima de 5000 palavras) (Picone, 1993);
8. Formato da palavra de entrada: texto restrito, formato livre de palavras.

Verifica-se, da listagem supracitada, uma grande variedade de opções disponíveis na especificação do sistema de reconhecimento. Dessa listagem, três sistemas são largamente utilizados e exemplificados na literatura especializada: a) Sistema de reconhecimento de dígitos isolados; b) Sistema de reconhecimento de dígito contínuo e, c) Sistema de reconhecimento de palavras de vocabulário grande.

2.7.1 Sistema de reconhecimento de dígitos isolados

A especificação para o sistema de reconhecimento de dígitos isolados desenvolvido neste trabalho é como segue:

1. Vocabulário de palavras isoladas;
2. População limitada de locutores;
3. Cooperativa de locutores sem restrição a sexo ou idade;
4. Ambiente de locução com nível de ruído acústico reduzido;
5. Transmissão com o locutor próximo ou microfone com $SNR = 60dB$;
6. Com sistema de treinamento;
7. Vocabulário de tamanho pequeno consistindo de dez dígitos (zero a nove);
8. Palavras de formatos simples com pausas entre cada palavra de entrada.

Na figura 2.8 mostra-se um diagrama de blocos do sistema de reconhecimento de dígitos isolados. A análise básica consiste da detecção de início e fim da locução, processamento da palavra para dar um padrão ou um conjunto de medidas, segmentação da locução em intervalos e, então, uma classe de decisão para escolher a palavra falada.

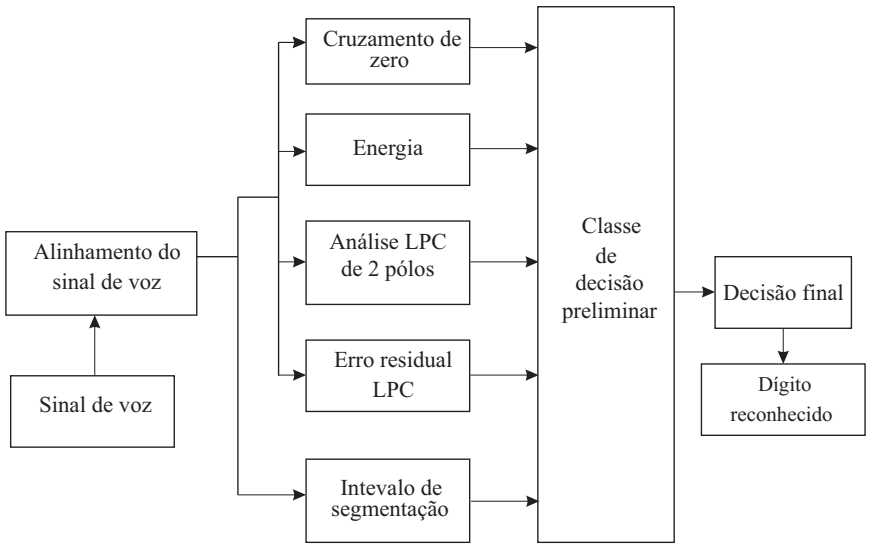


Fig. 2.8: Diagrama de blocos de um sistema de reconhecimento de dígitos isolados.

2.7.2 Processamento de voz dependente do tempo

O objetivo do processamento do sinal de voz é obter uma representação útil da informação nele contida. A precisão desta representação é determinada pelas informações particulares que devam ser preservadas ou, em alguns casos, que devam ser destacadas (Fant, 2004; Rabiner, 1993; Sadaoki, 2000). A hipótese básica na maioria dos sistemas de processamentos de voz é que as propriedades do sinal de voz mudam lentamente com o tempo (Fant, 2004; Rabiner, 2007). Esta hipótese induz a uma variedade de métodos de processamento nos quais pequenos segmentos do sinal de voz são isolados e processados como se fossem pequenos segmentos de um som contínuo com propriedades fixas.

O resultado do processamento desses pequenos segmentos pode ser um número ou um conjunto de números que produz uma nova sequência dependente do tempo que pode servir como uma representação do sinal de voz; desse modo, tem-se uma idéia intuitiva de que há um conceito de longo prazo que poderia fornecer a informação, se fosse possível generalizá-lo para o curto prazo. Por exemplo, supondo que se deseje saber se uma sequência de voz é sonora ou não-sonora no segmento adotado. Sabe-se que um sinal de voz sonoro possui, geralmente, maior energia (valor médio quadrático por amostras) que o sinal não-sonoro. A ideia, então, seria aplicar o conceito de energia média para ajudar na decisão. A energia média, contudo, é um conceito de longo prazo. A maioria das técnicas de processamento de curto-prazo pode ser representada matematicamente na forma abaixo:

$$Q_n = \sum_{m=-\infty}^{\infty} T\{s[m]\}\, w[n-m] \tag{2.11}$$

O sinal de voz, depois de submetido a uma filtragem para isolar a banda de frequência desejada, passa por uma transformação $T\{.\}$, que pode ser linear ou não-linear, e que pode depender de algum parâmetro ou conjunto de parâmetros. A sequência resultante é multiplicada por uma sequência $w[n]$ posicionada no tempo correspondente a amostra n, cujo objetivo é selecionar um pequeno intervalo do sinal de voz que será processado, no intuito de se conseguir parâmetros adequados ao modelamento do sinal. A função $w[n]$ é uma sequência real chamada de janela, cuja finalidade é determinar a porção do sinal de entrada que será enfatizado em um índice de tempo particular n. Os valores de Q_n são, portanto, uma sequência de valores médios ponderados da sequência $T\{s[m]\}$.

2.7.3 Janelamento

O processo pelo qual se seleciona um parte de um dado sinal é chamado de janelamento. A função utilizada para a seleção citada chama-se janela; em processamento de sinais, a janela mais comum é a janela retangular. A parte do sinal selecionada pela janela retangular denomina-se segmento. Na figura 2.9 ilustra-se um exemplo de janelamento retangular de um dado sinal. Na figura 2.9-a tem-se um dado sinal $s[n]$ e na figura 2.9-b tem-se um segmento de $s[n]$ obtido através da janela retangular. A duração do segmento é definida como a extensão de tempo na qual um conjunto de parâmetros do sinal é considerado válido. O período do segmento é utilizado para determinar a extensão de tempo entre os cálculos de sucessivos parâmetros. A escolha adequada da duração do segmento e da largura da janela depende dos parâmetros que se desejam analisar.

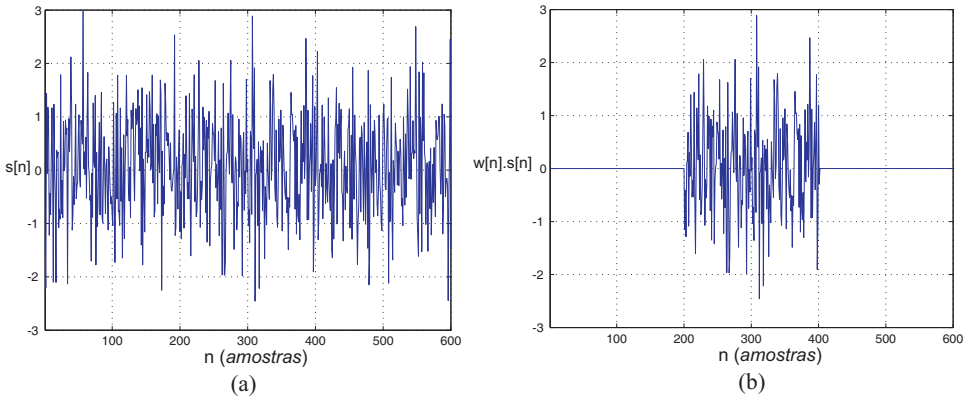


Fig. 2.9: Janelamento retangular de um sinal.

Devido ao fato de haver nas extremidades das janelas retangulares uma descontinuidade com taxa de variação de amplitude muito elevada, e por consequência, no espectro de frequências aparecerem ruídos excessivos, é comum o uso de janelas com as extremidades suaves. A multiplicação do sinal de voz por uma função janela com extremidades suaves tem dois efeitos. Primeiro, a amplitude resultante é gradualmente atenuada nas extremidades do intervalo do segmento considerado, fato que previne mudanças abruptas nas extremidades. Segundo, é produzida a convolução da tansformada de Fourier dos espectros da janela e do sinal de voz.

É desejável que a janela satisfaça duas características para reduzir as distorções espectrais causadas pelo janelamento: a) Resolução em alta frequência, fazendo-se o lóbulo principal da janela estreito; b) atenuação elevada nos lóbulos secundários para evitar efeitos indesejados de ruídos em alta frequência. Todavia, o segmento do sinal analisado sofre um amortecimento excessivo em suas amostras nas extremidades da janela, fato que pode eliminar características importantes do sinal. Desse modo, faz-se necessária então, a utilização da sobreposição entre sucessivas janelas para controlar quão rapidamente as características do sinal podem mudar de segmento para segmento. Assim, a cada novo segmento apenas uma fração do sinal irá mudar.

Em processamento de voz a janela mais utilizada é a de Hamming (Alam et al., 2012; Picone, 1993; Rabiner, 2007), que é um caso particular da janela generalizada de Hanning dada por:

$$w[n] = \frac{\alpha_w - (1 - \alpha_w)cos\,(2n\pi/(N-1))}{\beta_w} \tag{2.12}$$

com $\alpha_w = 0.54$, $0 \leq n \leq N$ e $w[n]=0$ para n fora do intervalo. De acordo com a equação (2.12) α_w é definida como uma constante no intervalo [0,1], N é o tempo de duração da janela e β_w é uma constante de normalização definida tal que o valor da raiz quadrada do valor médio quadrático (rms) da janela seja igual a unidade, como segue:

$$\beta_w = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} w^2[n]}. \tag{2.13}$$

A seguir, citam-se alguns exemplos de janelas.

Retangular:

$$w[n] = \begin{cases} 1, & 0 \leq n \leq N \\ 0, & \text{caso contrário} \end{cases} \tag{2.14}$$

Bartlett(triangular):

$$w[n] = \begin{cases} 2n/N, & 0 \leq n \leq N/2 \\ 2n - 2n/N, & N/2 \leq n \leq N \\ 0, & \text{caso contrário} \end{cases} \tag{2.15}$$

Hanning:

$$w[n] = \begin{cases} 0,5 - 0,5cos(2\pi n/N), & 0 \leq n \leq N \\ 0, & \text{caso contrário} \end{cases} \tag{2.16}$$

Hamming:

$$w[n] = \begin{cases} 0,54 - 0,46cos(2\pi n/N), & 0 \leq n \leq N \\ 0, & \text{caso contrário} \end{cases} \tag{2.17}$$

Blackman:

$$w[n] = \begin{cases} 0,42 - 0,5cos(2\pi n/N) + 0,08cos(4\pi n/N), & 0 \leq n \leq N \\ 0, & \text{caso contrário} \end{cases} \tag{2.18}$$

Na figura 2.10-a, mostra-se a janela de Hamming para $N = 100$ amostras; na figura 2.10-b, sua representação espectral.

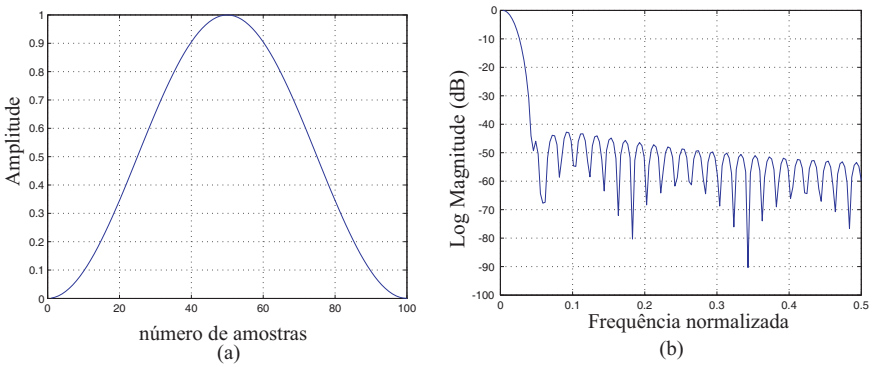


Fig. 2.10: a) Representação gráfica da janela de Hamming no domínio do tempo; b) sua representação espectral.

Na figura 2.11-a, mostra-se a representação espectral da janela retangular; na figura 2.11-b, mostra-se a representação espectral da janela de Hamming; na figura 2.11-c, ilustra-se o efeito da janela retangular em um segmento de voz no domínio espectral, e na figura 2.11-d, ilustra-se o efeito da janela de Hamming em um segmento de voz no domínio espectral.

Observa-se, comparando as figuras 2.11-a e 2.11-b, que a janela retangular é mais seletiva, fato visto através do seu lóbulo principal; contudo, observam-se os efeitos em altas frequên-

cias no sinal analisado. Na janela de Hamming observa-se que o envelope espectral do sinal resultante é mais suave; entretanto, o efeito nas extremidades da janela é bastante acentuado.

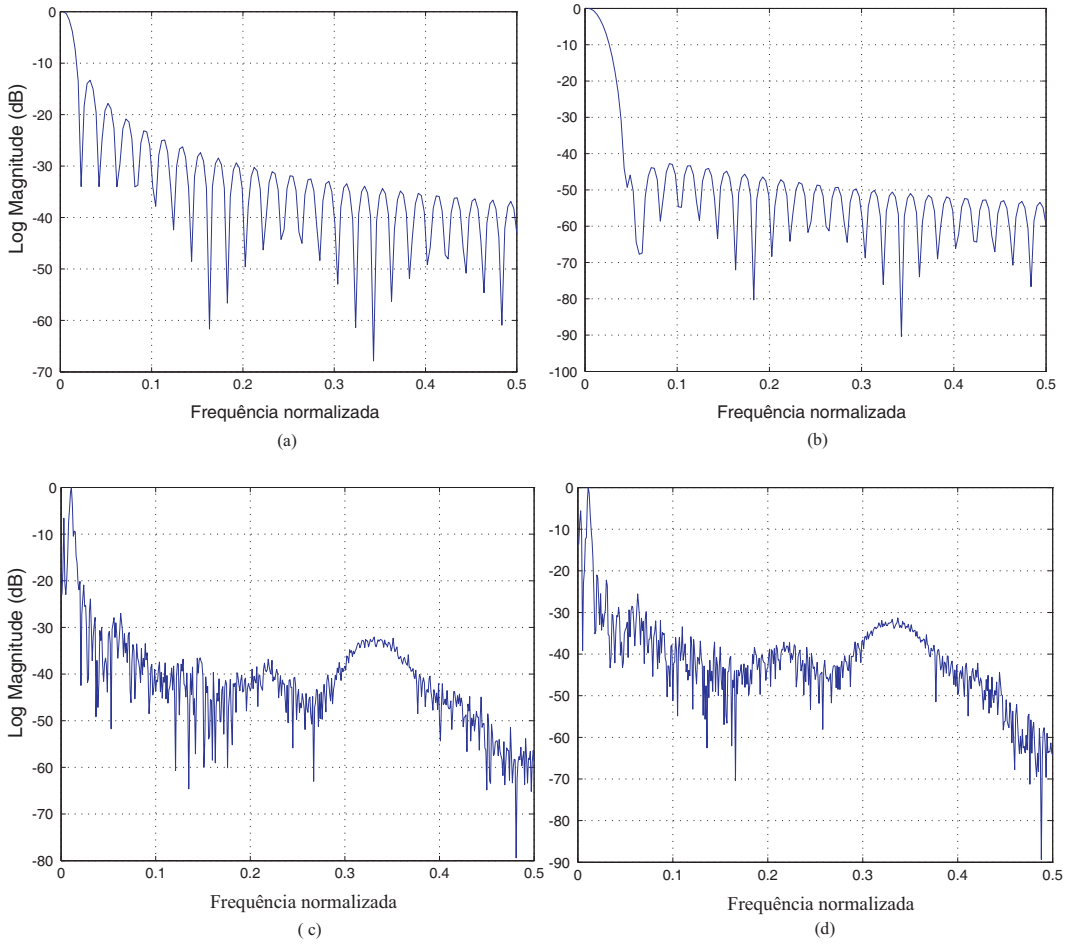


Fig. 2.11: a) Representação espectral da janela retangular; b) Representação espectral da janela de Hamming; c) Representação espectral do efeito da janela retangular em um segmento do sinal de voz; d) Representação espectral do efeito da janela de Hamming em um segmento do sinal de voz.

2.8 Análise do sinal de voz através de segmentos

Quando uma janela $w[n]$ é aplicada a um determinado sinal, ela seleciona um segmento deste sinal que será submetido à análise. A análise de Fourier de curto prazo é efetuada sobre esse segmento. Esta técnica é chamada análise de sinal segmento por segmento (curto prazo). O período do segmento T_f é definido como a extensão de tempo na qual um conjunto de parâmetros é considerado válido. O período do segmento é utilizado para determinar a

extensão de tempo entre os cálculos de sucessivos parâmetros. A razão de segmento determina o número de segmentos calculados por segundo.

Em processamento de sinais de voz, o período do segmento depende da velocidade dos sistemas articuladores que produzem a voz, considerando que informações importantes que caracterizam o sinal de voz, tais como o período de pitch que se estima próximo de 20 amostras para mulheres e criança e acima 250 amostras para homens, considerando a frequência de amostragem de 10 kHz (Picone, 1993; Rabiner, 2007), verifica-se que um tamanho adequado para o segmento deverá está entre $100 - 200$ amostras, para uma frequencia de 10 kHz, isto é, um período de segmento entre 10 ms e 20 ms (Rabiner, 2007). Esses valores são baseados na análise das mudanças das propriedades do sinal de voz com o tempo. Observa-se que uma duração de segmento muito grande pode enfatizar mudanças lentas no sinal de voz, o que pode não representar de forma adequada as propriedades do sinal de voz; por outro lado, a escolha de um valor muito pequeno para a duração do segmento pode deixar de fora da análise propriedades importantes do sinal de voz.

Outro fator muito importante na análise espectral é o intervalo no qual a potência do sinal é calculada. O número de amostras para calcular a potência é chamado de duração da janela T_w e normalmente é medido em unidades de tempo. O período do segmento e a duração da janela controlam a razão na qual o valor de potência representa a dinâmica do sinal. O período do segmento e a duração da janela são normalmente ajustados como par; uma duração de janela de 30 ms é muito comum com um período de segmento de 20 ms. Desde que um período curto do segmento seja usado para capturar dinâmicas rápidas do espectro, a duração da janela deve ser correspondentemente curta, para que detalhes do espectro não sejam excessivamente amortecidos (Trancoso, 1989).

Devido ao fato de nas extremidades das janelas o sinal analisado sofrer um amortecimento excessivo em suas amostras, faz-se necessária a utilização do processo denominado de sobreposição para controlar quão rapidamente os parâmetros do sinal podem mudar de segmento para segmento. Assim, a cada novo segmento apenas uma fração do sinal irá mudar. A porcentagem de sobreposição entre as janelas é dada por:

$$\text{sobreposição}(\%) = \frac{T_w - T_f}{T_w} \times 100 \tag{2.19}$$

onde T_w é o tempo de duração da janela e T_f é o tempo de duração do segmento. Assim, por exemplo, a combinação do período do segmento de $20ms$ e duração de janela de $30ms$ corresponde a aproximadamente 33% de sobreposição.

2.8.1 Discriminação de voz versus silêncio

Para sistemas de reconhecimento de voz em geral, é muito importante que os mesmos possam discriminar com bastante precisão o início de uma locução, isto é, quando realmente se inicia a palavra falada, a fim de que se possa eliminar, tanto quanto possível, o ruído inerente ao ambiente, inserido durante o processo de gravação do som. E ainda, com a mesma precisão, faz-se necessária a detecção do final da palavra falada para que se retire do sinal somente a parte que deva ser processada no reconhecimento. Atualmente, há vários algoritmos que realizam essa tarefa, entretanto, um dos mais populares é o que utiliza a energia do sinal e a taxa de cruzamento por zero.

Taxa de cruzamento por zero

No contexto de sinais de tempo discreto, diz-se ocorrer um cruzamento por zero se sucessivas amostras têm diferentes sinais algébricos. A razão na qual o cruzamento por zero ocorre é uma simples medida da característica de variação do sinal. Isto é particularmente verdade em sinais de banda estreita. Por exemplo, um sinal senoidal de frequência f_0, amostrado na razão F_a, tem-se $\frac{F_a}{f_0}$ amostras por ciclo da onda senoidal. Cada ciclo tem dois cruzamentos por zero tal que, a longo prazo, a média de cruzamentos por zero é dada por $Z = \frac{2f_0}{F_a}$ cruzamentos/amostras. Assim, a média de cruzamentos por zero fornece um razoável modo de se estimar a frequência da onda senoidal. Sinais de voz são sinais de banda larga e a interpretação da média da taxa de cruzamentos por zero é, portanto, menos precisa. Contudo, pode-se obter estimativas das propriedades espectrais usando a representação baseada na taxa de cruzamento por zero na análise de curto prazo.

Uma definição adequada para a taxa de cruzamento por zeros é dada na equação (2.20).

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[s[m]] - sgn[s[m-1]]|\, w[n-m] \tag{2.20}$$

onde

$$sgn[n] = \begin{cases} 1, & s[n] \geq 0 \\ -1, & s[n] < 0 \end{cases} \tag{2.21}$$

e

$$w[n] = \begin{cases} \frac{1}{2N}, & 0 \leq n \leq N \\ 0, & \text{caso contrário} \end{cases} \tag{2.22}$$

A energia de sons sonoros é concentrada em frequências inferiores a $3kHz$, devido ao decaimento espectral introduzido pelo pulso glotal, enquanto que a energia de sons não-sonoros é encontrada em frequências mais elevadas. Uma generalização razoável é que se a taxa de cruzamento por

zeros Z_n for alta, o sinal de voz é não-sonoro, enquanto que se Z_n for baixa, o sinal de voz pode ser considerado sonoro.

Energia do sinal de voz

A energia de um sinal de tempo discreto é dada por (Rabiner, 1993):

$$E = \sum_{m=-\infty}^{\infty} s^2[m] \tag{2.23}$$

Observa-se na equação (2.23) que o intervalo utilizado para cálculo da energia é infinito. Essa expressão tem pouca utilidade para análise do sinal de voz. Por outro lado, essa expressão pode ser modificada para análise de curto prazo em um intervalo finito de dimensão $N-1$.

$$E_n = \sum_{m=n-N+1}^{n} s^2[m] \tag{2.24}$$

Comparando-se a equação (2.23) com a equação (2.11), verifica-se que $T\{.\} \equiv (.)^2$ e que $w[n]$ é a janela retangular.

A faixa selecionada pela janela será efetivamente utilizada nos cálculos de energia. O objetivo desta análise é que a energia a curto prazo reflita as variações e a amplitude do sinal de voz. A janela deve ser curta para responder as variações rápidas de amplitude, mas não tão curta que não forneça uma tomada de segmento de voz razoável para resultar em uma função de energia suavizada. Um problema na utilização da equação (2.24) é que E_n é sensível às grandes variações nos valores das amostras (Picone, 1993; Rabiner, 1993, 2007); para evitar-se esse problema utiliza-se a expressão da função de magnitude média, dada na equação (2.25); neste caso, a soma dos valores absolutos é computada, em vez da soma dos valores quadrados das amostras.

$$M_n = \sum_{m=-\infty}^{\infty} |s[m]|\, w[n-m] \tag{2.25}$$

A estimativa de E_n fornece uma base razoável para a distinção entre sons sonoros, de energia elevada, e sons não-sonoros, de energia baixa. A despeito da dificuldade de se localizar o início e o final de locuções, energia e taxa de cruzamento por zero podem ser combinadas para servir como base para desenvolvimento de um algoritmo para o início e o fim da locução, uma vez que ambas as representações possuem informações importantes do sinal de voz. Uma abordagem mais detalhada de algoritmos de localização de início e fim de locução é dada por Becchetti e

Ricotti (Becchetti, 2000).

Análise espectral utilizando-se a Transformada de Fourier de Tempo Discreto (TFTD)

As representações de sinais através de senoides ou exponenciais complexas são muito úteis, principalmente por que podem simplificar a análise que poderia ser bastante complexa por outros métodos. Tais representações, geralmente chamadas de representações de Fourier, são bastante utilizadas por que descrevem de forma conveniente um dado sinal como uma superposição de senoides ou exponenciais complexas; também destacam propriedades do sinal que não são facilmente observadas no sinal original.

A representação de Fourier de um dado sinal via transformada de tempo discreto direta ou inversa é ponto chave na análise de sinais. As equações (2.26) e (2.27) são as equações de análise e síntese, respectivamente, conforme definido por Oppenheim e Schafer (Oppenheim, 2013).

$$X\left(e^{j\omega}\right) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n} \tag{2.26}$$

$$x[n] = \frac{1}{2\pi}\oint_{2\pi} X\left(e^{j\omega}\right)e^{j\omega n}d\omega \tag{2.27}$$

Similarmente, a resposta em freqüência, que é a TFTD da resposta impulso, fornece uma descrição concisa de um sistema linear invariante no tempo quando usado para filtragem. A TFTD $X\left(e^{j\omega}\right)$ é uma função complexa da frequência ω. O período é sempre 2π, e o período fundamental é usualmente escolhido no intervalo $[-\pi, \pi]$. O resultado da TFTD é a representação do sinal no domínio da frequência, e essa representação é muito importante pela quantidade de características inerentes ao sinal que facilitam seu processamento e sua análise.

Análise de Fourier de curto-prazo

O sinal de voz é estocástico e muito complexo. Assim, a representação de Fourier padrão, que é apropriada para sinais determinísticos, não seria adequada para aplicação direta no modelamento do sinal de voz. Entretanto, considerando-se a análise sobre pequenas parcelas do sinal, verifica-se que elas possuem características que podem ser representadas através da análise de Fourier, chamada de análise de Fourier de curto prazo, isto porque somente uma parcela do tempo total é analisada. Isso só é possível se às propriedades temporais do sinal de voz, tais como energia, cruzamento por zero e correlação, permanecerem estáveis em determinados períodos. Para o caso do sinal de voz, verificam-se essas propriedades em intervalos de 10 ms

a 30 ms (Picone, 1993; Rabiner, 1993; Trancoso, 1989).

A representação de Fourier de curto prazo, dada por, $X_n\left(e^{j\omega}\right)$ é uma representação bidimensional de um sinal $s[n]$ de uma dimensão. $X_n\left(e^{j\omega}\right)$ é uma função do tempo n e da frequência continua ω. Uma definição bastante utilizada é dada na equação (2.28) (Rabiner, 1993).

$$X_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} w[n-m]s[m]e^{-j\omega m} \tag{2.28}$$

onde $w[n-m]$ é uma sequência real, já definida anteriormente, que determina a porção do sinal que será enfatizada em um índice particular de tempo n.

Analisando-se a equação (2.28), nota-se que a transformada de Fourier dependente do tempo é uma função de duas variáveis: o índice de tempo n, que é discreto, e a variável da frequência ω, que é contínua. A equação alternativa à equação (2.28) pode ser obtida fazendo-se,

$$X_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} w[m]s[n-m]e^{-j\omega(n-m)} \tag{2.29}$$

$$X_n\left(e^{j\omega}\right) = e^{-j\omega n}\sum_{m=-\infty}^{\infty} s[n-m]w[m]e^{j\omega m} \tag{2.30}$$

definido-se

$$\tilde{X}_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} s[n-m]w[m]e^{j\omega m} \tag{2.31}$$

obtém-se:

$$X_n\left(e^{j\omega}\right) = e^{-j\omega n}\tilde{X}_n\left(e^{j\omega}\right) \tag{2.32}$$

Essas equações podem ser interpretadas de duas formas diferentes: se n é fixo, tem-se que $X_n\left(e^{j\omega}\right)$ é a transformada de Fourier normal da sequência $w[n-m]s[m]$, $-\infty < m < \infty$, e possui todas as propriedades da transformada de Fourier. Agora, se ω é fixo, $X_n\left(e^{j\omega}\right)$ como função do tempo de índice n, ambas as equações (2.28) e (2.29) apresentam-se como convolução. Esta interpretação leva naturalmente a considerar a análise de Fourier como um processo de filtragem linear.

Dado que $X_n\left(e^{j\omega}\right)$ é periódica em ω com período 2π, para se recuperar $s[n]$ com exatidão, é suficiente expecificar $X_n\left(e^{j\omega}\right)$ dentro de um conjunto finito adequado de frequências $\omega_k = \frac{2\pi k}{N}$, onde $k = 0, 1, 2, \cdots N-1$. Se a janela de análise de $X_n\left(e^{j\omega}\right)$ é limitada no tempo, então a transformada inversa de $X_n\left(e^{j\omega}\right)$ também é limitada no tempo. Sendo a transformada inversa de

$X_n\left(e^{j\omega}\right)$ dada por $s[n]w[n-m]$ com a duração do sinal dada por L amostras. Então, de acordo com o teorema da amostragem (Lathi, 1998; Oppenheim, 2013), $X_n\left(e^{j\omega}\right)$ deve ser amostrado, em frequência, em um conjunto de frequências dado por $\omega_k = \frac{2\pi k}{L}$, onde $k = 0, 1, 2, \cdots L-1$ para que se obtenha exatamente $s[n]$ a partir de $X_n\left(e^{j\omega_k}\right)$. Assim, para um sinal $s[n]$ de duração de L amostras, $X_n\left(e^{j\omega}\right)$ deve ser avaliada em pelo menos L amostras uniformemente espaçadas. Das equações (2.28) e (2.29), para uma dada frequência discreta ω_k, obtém-se:

$$X_n\left(e^{j\omega_k}\right) = \sum_{m=-\infty}^{\infty} w_k[n-m]s[m]e^{-j\omega_k m} \tag{2.33}$$

ou

$$X_n\left(e^{j\omega_k}\right) = e^{-j\omega_k n} \sum_{m=-\infty}^{\infty} w_k[m]s[n-m]e^{j\omega_k m} \tag{2.34}$$

onde $w_k[n-m]$ é a janela usada na frequência ω_k. Definindo-se

$$h_k[n] = w_k[n]e^{j\omega_k n} \tag{2.35}$$

então, a equação (2.34) pode ser expressa como:

$$X_n\left(e^{j\omega_k}\right) = e^{-j\omega_k n} \sum_{m=-\infty}^{\infty} s[n-m]h_k[m] \tag{2.36}$$

Desde que a janela $w_k[n]$ tenha as propriedades de um filtro passa-baixa, a equação (2.36) pode ser interpretada como um filtro passa-banda com resposta impulso $h_k[n]$ seguida de uma modulação exponencial complexa $e^{-j\omega_k n}$. Definido-se

$$y_k[n] = X_n\left(e^{j\omega_k}\right) e^{j\omega_k n} \tag{2.37}$$

obtém-se da equação (2.36) que

$$y_k[n] = \sum_{m=-\infty}^{\infty} s[n-m]h_k[m] \tag{2.38}$$

Assim, $y_k[n]$ é a saída de um banco de filtros com resposta impulso $h_k[n]$ como dado na equação (2.35).

2.8.2 Representação paramétrica do sinal de voz

O procedimento de extração de características consiste na transformação de um vetor de amostras do sinal voz em um vetor de observações apropriado, cujos componentes são chamados

de características do sinal. O objetivo da parametrização é representar o sinal de voz em um espaço mais apropriado em que as informações desnecessárias são descartadas. Em princípio, espera-se que o vetor de observações contenha características relevantes, capazes de representar adequadamente a palavra pronunciada.

A parametrização do sinal de voz deve reduzir o número total de dados a serem trabalhados no reconhecimento de voz. Isso significa que o tamanho do vetor de observações deve ser menor que o vetor de amostras. A parametrização pode ser idealizada como uma sequência de operações que mapeiam um vetor de entrada em um vetor de saída. Mais especificamente, o vetor de amostras é o vetor de entrada e o vetor de observações é o vetor de saída. Na figura 2.12 mostram-se seis algoritmos utilizados na análise espectral do sinal de voz.

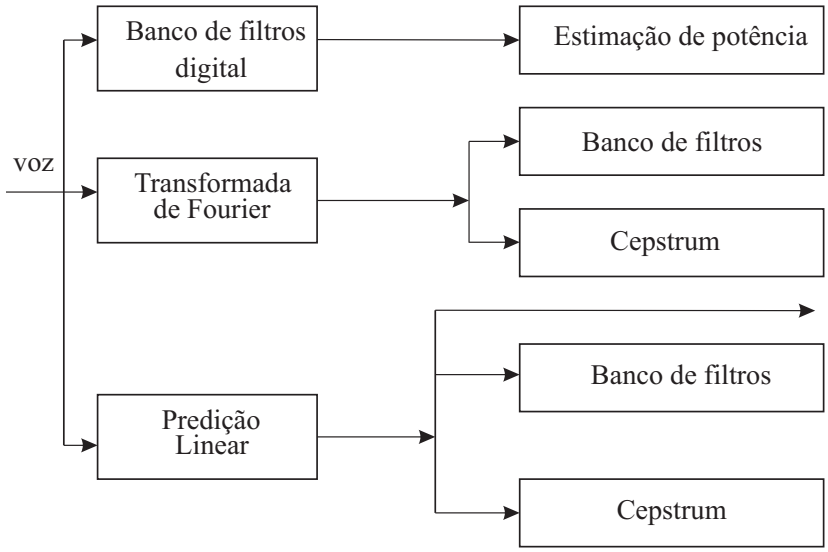


Fig. 2.12: Exemplos de algoritmos utilizados na representação paramétrica do sinal de voz.

Banco de filtros digitais

O banco de filtros digitais é um dos conceitos mais fundamentais em processamento de voz. Um banco de filtros pode ser considerado como um modelo das fases iniciais da transdução do sistema auditivo humano (Picone, 1993). Experimentos de percepção humana mostram que frequências de sons complexos dentro de uma certa largura de banda de uma dada frequência nominal não podem ser individualmente distinguidas. Quando uma das componentes desses sons está fora desta largura de banda ela pode ser distinguida individualmente. Esta largura de banda é chamada de largura de banda crítica e, geralmente, é de 10% a 20% da frequência central do som (Picone, 1993; Rabiner, 1993). A banda crítica, BW_{CR}, pode ser calculada

através da equação.

$$BW_{CR} = 25 + 75 \left[1 + 1.4 \left(\frac{f}{1000} \right)^2 \right]^{0.69} \tag{2.39}$$

onde f é a frequência linear dada em Hz. Esta transformação pode ser usada para calcular as larguras de bandas dos filtros em uma dada frequência e distribuídas em escalas não lineares, tais como a escala bark ou mel (Picone, 1993; Rabiner, 1993; Sadaoki, 2000).

Há evidências de que o sistema auditivo humano percebe sinais de voz ao longo de uma escala não linear no domínio da frequência. Uma abordagem para simular o espectro subjetivo é usar um banco de filtros, espaçados uniformemente em uma escala de frequência não linear, tal como a escala mel. A relação entre a escala mel de frequência e a escala linear de frequência é dada na equação (2.40) (Linkai, 2000):

$$mel = 2595 \cdot log_{10} \left(1 + \frac{f}{700} \right) \tag{2.40}$$

onde mel é a escala mel de frequências e f é a frequência linear dada em Hz. O banco de filtros é então implementado de acordo com a escala mel de frequência, como mostrado na figura 2.13, onde as 20 bandas dos filtros são desenvolvidos por 20 filtros triangulares passa-banda, $f[i,k]$, onde $(0 \leq i \leq 20; 0 \leq k \leq 63)$ (Davis, 1980; Picone, 1993; Wu, 2000), sendo i o índice da banda, e k um ponto em frequência na banda especificada sobre a faixa de frequência de 0 a aproximadamente $4.6kHz$. Assim, cada banda do filtro tem uma resposta em frequência passa-banda triangular. O espaçamento, bem como a largura de banda, são determinados por um intervalo constante da escala mel de frequência. O valor da função triangular $f[i,k]$ também representa o fator de ponderação da energia da banda de frequência no k-ésimo ponto da i-ésima banda.

Com o banco de filtros espaçados na escala mel de frequência, pode-se calcular a energia de cada banda de frequência para cada segmento de tempo do sinal de voz. Considerando um dado sinal ruidoso de voz $s[m,n]$, representando a magnitude do n-ésimo ponto do m-ésimo segmento, primeiro deve-se calcular o espectro $S[m,k]$, deste sinal através da transformada discreta de Fourier (TDF).

$$S[m,k] = \sum_{n=0}^{N-1} s[m,n] e^{-j\frac{2\pi kn}{N}}; 0 \leq k \leq N-1; 0 \leq m \leq M-1 \tag{2.41}$$

sendo,

$S[m,k]$: o espectro do k-ésimo ponto do m-ésimo segmento;

N: a largura da janela utilizada;

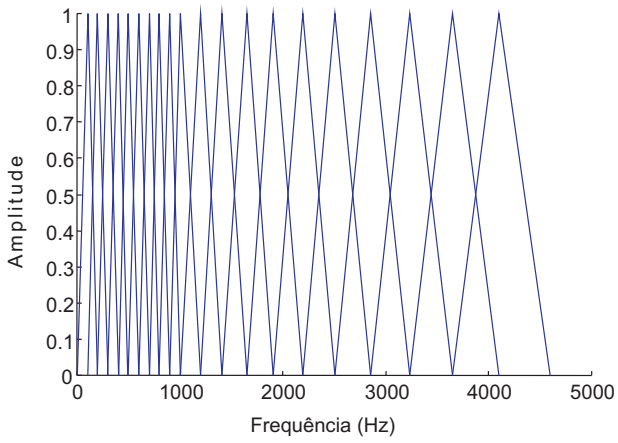


Fig. 2.13: Banco de filtros triangulares de 20 bandas distribuídos na escala mel de freqüências.

M: o número de segmentos do sinal de voz para análise.

Calculado o espectro, o seu módulo deve ser multiplicado pelos fatores de ponderação $f[i,k]$ no banco de filtros na escala mel de frequência e somados os produtos para todos os $k's$, para obter-se a energia $E\,[m,i]$ para cada banda i de frequência do m-ésimo segmento, utilizando-se a magnitude do espectro, conforme dado na equação,

$$\bar{E}[m,i] = \sum_{k=0}^{N-1} |S[m,k]|\, f[i,k]; 0 \leq m \leq M-1; 1 \leq i \leq 20 \tag{2.42}$$

em que,

i é o índice da banda do filtro;

k é o índice do espectro;

m é o número do segmento analisado;

M é o número total de segmentos para análise.

Na figura 2.14 ilustra-se a energia ponderada de um segmento de sinal de voz, apresentado na figura 2.5, devidamente filtrado pelo banco de filtros apresentado na figura 2.13.

Para a remoção de algum ruído impulsivo indesejado na equação (2.42), a energia deve ser suavizada. Um método eficiente é proposto em (Wu, 2000), em que se utiliza a média aritmética de três pontos do filtro dado por:

$$\hat{E}[m,i] = \frac{\bar{E}[m-1+i,i] + \bar{E}[m,i] + \bar{E}[m+1,i]}{3} \tag{2.43}$$

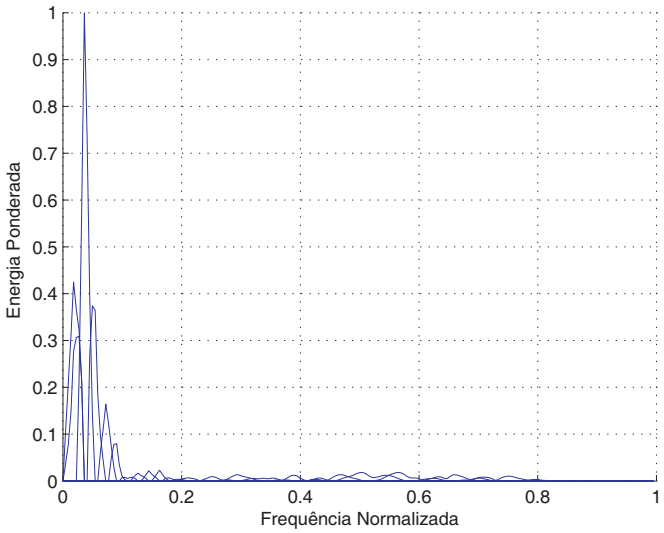


Fig. 2.14: Energia de um segmento de sinal de voz ponderada por um banco de filtro de 20 bandas.

Depois da energia suavizada $\hat{E}[m,i]$, remove-se a energia devida ao ruído de fundo do sinal de voz. No trabalho desenvolvido por Nasersharif (2007) são analisadas quatro estratégias para redução de ruído durante o cálculo dos coeficientes mel-cepstrais (Nasersharif, 2007). Uma estratégia mais simples é proposta por Wu (2000) em que a energia média quase pura do sinal de voz, $E[m,i]$, é obtida pela subtração da enegia do ruído e_{fd} como segue:

$$E[m,i] = \hat{E}[m,i] - e_{fd} \tag{2.44}$$

onde a energia do ruído de fundo pode ser estimada pela média aritmética da energia nos primeiros cinco segmentos do sinal de voz (Wu, 2000).

$$e_{fd} = \frac{\sum_{m=0}^{4} \hat{E}[m,i]}{5} \tag{2.45}$$

Com a energia suavizada da $i-$ésima banda do $m-$ésimo segmento, $E\,[m,i]$, pode-se calcular a energia total quase pura do sinal de voz na $i-$ésima banda como:

$$E[i] = \sum_{m=0}^{M-1} |E[m,i]| \tag{2.46}$$

Desde que o objetivo da análise da energia seja selecionar alguma banda útil tendo o máximo de informação, necessita-se de um parâmetro para representar a quantidade de informação de cada banda. De acordo com (2.46), $E[i]$ pode ser um bom parâmetro para indicar a quantidade

de informação do sinal de voz. Assumindo-se que a maior parte do sinal está coberto pelo ruído, obviamente, a energia $E[i]$ será pequena; por outro lado, quanto mais elevado for $E[i]$, mais informação estará contida na i-ésima banda (Wu, 2000).

Análise Cepstral - CEPSTRUM

Os métodos conhecidos, na literatura, como análise cepstral e desconvolução homomórfica têm-se mostrado extremamente efetivos e úteis em certas aplicações, tais como, análise sísmicas e codificação de sinal de voz. Foi observado, em 1963 por Bogert, Healy e Tukey (Bogert et al., 1963), que o logaritmo do espectro de potência de um sinal contendo um eco tem uma componente aditiva periódica devido ao eco. Assim, a transformada de Fourier do logaritmo do espectro de potência deve exibir um pico no atraso do eco. Eles chamaram esta função de cepstrum que é uma composição da palavra **"spectrum'**.

Oppenheim (Oppenheim, 2013) propôs uma nova classe de sistemas chamada sistemas homomórficos. Embora esses sistemas sejam não lineares no sentido clássico, tais sistemas satisfazem uma generalização do princípio da superposição, isto é, sinais de entrada e suas respostas correspondentes são superpostas por uma operação que possui as mesmas propriedades algébricas da adição, conforme dado abaixo:

- Aditividade

$$T\{s[n]\} = T\{s_1[n] + s_2[n]\} = T\{s_1[n]\} + T\{s_2[n]\} \tag{2.47}$$

e

- Homogeneidade

$$T\{a \cdot s[n]\} = aT\{s[n]\} \tag{2.48}$$

onde T representa um operador linear.

O conceito de filtragem homomórfica combina multiplicação e convolução, pois muitos modelos de sinais envolvem estas operações. A transformação de um sinal em seu cepstrum é uma transformação homomórfica. O modelo básico da produção da voz pode ser considerado como um filtro $h[n]$ excitado por uma função excitação $e[n]$ para sons sonoros ou ruído branco para sons não-sonoros. Portanto, o espectro de curto prazo compreende uma envoltória espectral correspondente ao filtro do trato vocal que varia lentamente e, no caso de sinais de voz sonoros, uma fina estrutura que varia rapidamente corresponde à frequência de excitação e seus harmônicos. A sequência de amostras de voz observada resulta em uma convolução da excitação e da resposta impulso do trato vocal no domínio do tempo, conforme a equação abaixo,

$$s[n] = h[n] * e[n] \tag{2.49}$$

onde o símbolo $(*)$ significa a operação de convolução. Uma vez que as partes componentes do sinal de voz não são combinadas linearmente, as técnicas lineares de análise não apresentariam resultados úteis. O espectro resultante desta operação é o produto dos espectros da excitação e do filtro do trato vocal no domínio da frequência, de acordo com a equação,

$$S(e^{j\omega}) = H(e^{j\omega}) \cdot E(e^{j\omega}) \tag{2.50}$$

Aplicando-se o logaritmo ($log = log_{10}$) em ambos os lados da equação (2.50) obtém-se,

$$log[S(e^{j\omega})] = log[H(e^{j\omega}) \cdot E(e^{j\omega})] = log[H(e^{j\omega})] + log[E(e^{j\omega})] \tag{2.51}$$

Da equação (2.51), observa-se que na escala do logaritmo o produto do espectro da excitação e do espectro do filtro do trato vocal é transformado em somatório destes dois espectros (operação de logaritmo). A transformação do $log[S(e^{j\omega})]$ pela transformada inversa de Fourier é o cepstrum que pode representar a excitação e o filtro do trato vocal separadamente. Observa-se que o cepstrum apresenta duas grandes vantagens em relação ao espectro quando se comparam as equações (2.50) e (2.51):

1. As representações das componentes da sequência gerada podem ser separadas no cepstrum;

2. As representações das componentes da sequência gerada podem ser combindas linearmente no cepstrum.

Na figura 2.15, é mostrada a operação para a obtenção do cepstrum.

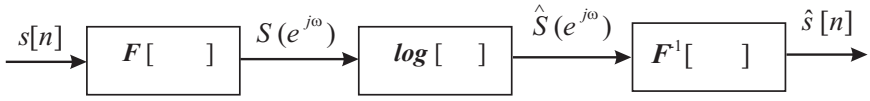


Fig. 2.15: Representação da característica de um sistema para a desconvolução homomórfica.

O parâmetro para o cepstrum, $\hat{s}[n]$, é chamado quefrency e é efetivamente um parâmetro no domínio do tempo já que o cepstrum foi obtido da inversão de uma função no domínio da frequência. Basicamente, a operação para obtenção dos coeficientes cepstrais $c[n]$ é dada por:

$$c[n] = F^{-1}\left[log\left|S(e^{j\omega})\right|\right] \tag{2.52}$$

Na equação (2.52), F^{-1} é a transformada inversa de Fourier. Há dois tipos de análises cepstral: análise *Fast Fourier Transform*$-FFT$ cepstral e análise *Linear Prediction Coefficients*$-LPC$ cepstral. Na análise FFT cepstral, a FFT é aplicada diretamente ao sinal de voz. Por outro lado, à análise LPC cepstral, a transformada z é aplicada ao sinal de voz modelado pela análise LPC.

Para calcular-se os coeficientes cepstrais pela FFT cepstral, primeiro deve-se calcular o logaritmo da magnitude espectral. Em seguida, calcula-se a transformada inversa de Fourier do logaritmo do espectro. Então, desenvolvendo a equação (2.52) através da TFD, tem-se,

$$c[n] = \frac{1}{N} \sum_{k=0}^{N-1} log\,|S[k]|\, e^{j\left(\frac{2\pi}{N} kn\right)} \tag{2.53}$$

Verifica-se na equação (2.53) que o coeficiente $c[0]$ representa o valor médio do log da magnitude do espectro, isto é, $c[0]$ é uma medida de potência. Inicialmente, nos trabalhos de parametrização do sinal de voz, este termo foi uma parte importante; contudo, verificou-se que o valor da medida de potência absoluta não é um parâmetro confiável e geralmente não utilizado (Picone, 1993).

Vale notar que o espectro do logaritmo da magnitude é uma função real e simétrica; então, fazendo-se as devidas simplificações, obtém-se:

$$c[n] = \frac{2}{N} \sum_{k=0}^{N-1} log\,(S[I[k]])\, cos\left(\frac{2\pi}{N} kn\right) \tag{2.54}$$

em que $I[k]$ é uma função de mapeamento adequada que relaciona k com as amostras de $S[k]$. Normalmente os coeficientes cepstrais $c[n]$ são limitados a uma ordem menor que N. O espectro $S[k]$ pode ser calculado usando-se a FFT. Os coeficientes cepstrais definidos em (2.53) podem ser facilmente modificados para ser espaçados em uma escala mel, bastando para isso amostrar a transformada de Fourier em frequências espaçadas apropriadamente.

Para analisar as propriedades do cepstrum LPC, a excitação $E(e^{j\omega})$ e o filtro do trato vocal $H(e^{j\omega})$ no espectro do sinal de voz $S(e^{j\omega})$ são separados linearmente por uma operação logaritmo complexa, vista na equação (2.52). Os coeficientes cepstrais $c[n]$ são definidos como a transformada de Fourier inversa sobre o logaritmo do espectro $S(e^{j\omega})$. Isto indica que as características do trato vocal e da excitação são bem representadas separadamente nos coeficientes cepstrais. Os coeficientes de ordem alta possuem as propriedades da excitação, e os coeficientes de ordem baixa possuem as propriedades do trato vocal.

Uma variedade grande de sistemas de reconhecimento de voz utiliza a análise cepstral. Uma vantagem de se utilizar essa análise é que a correlação entre os coeficientes é extremamente

pequena, possibilitando uma hipótese de modelamento simplificada. Entretanto, o cepstrum é calculado usando-se uma operação não linear, a função logaritmo. Parâmetros cepstrais deduzidos de estimadores espectrais de alta resolução ou adaptações paramétricas do espectro são preferidos para trabalhos em ambientes ruidosos (Ariki et al., 1989; Picone, 1993; Sadaoki, 2000).

Cepstrum de duas dimensões

A análise de sinais com cepstrum de uma dimensão aplica a transformada de Fourier de curto prazo de uma dimensão ao logaritmo do módulo do espectro. Por outro lado, a análise de sinais com o cepstrum de duas dimensões aplica a transformada de Fourier de duas dimensões de uma sequência no tempo ao logaritmo do módulo do espectro e converte-o em um cepstrum de duas dimensões. Na análise, portanto, vários segmentos consecutivos são agrupados como um bloco e assim são processados. É desejável que a duração do bloco aproxime-se daquela da percepção humana (Picone, 1993; Rabiner, 1993; Sadaoki, 2000).

O cepstrum de duas dimensões é calculado da seguinte forma: seja um dado segmento de N amostras, e M a quantidade de segmentos em um bloco. Dado um sinal de voz s_{nm}, amostrado no n-ésimo ponto e no m-ésimo segmento no bloco, representado da seguinte forma:

$$s_{nm} = s_{n+mT},\ 0 \leq n \leq N-1,\ 0 \leq m \leq M-1,\ ,0 < T < N, \tag{2.55}$$

onde T é o período do segmento. O logaritmo do espectro do m-ésimo segmento é expresso por:

$$S_{km} = 10\ log \left| \sum_{n=0}^{N-1} s_{nm} W_1^{-nk} \right|^2 \tag{2.56}$$

onde, $W_1 = e^{j\frac{2\pi}{N}}$, $0 \leq k \leq N-1$.

Os coeficientes do cepstrum de duas dimensões são obtidos aplicando-se novamente a transformada de Fourier no log do espectro S_{km},

$$c_{qp} = \frac{1}{NM} \sum_{k=0}^{N-1} \sum_{m=0}^{M-1} S_{km} W_1^{-kq} W_2^{-mp} \tag{2.57}$$

dado que, $W_2 = e^{j\frac{2\pi}{M}}$, $0 \leq q \leq N-1$ e $0 \leq p \leq M-1$. Os coeficientes c_{qp} são complexos. O eixo q corresponde à transformada do eixo de frequência, é chamado de "*quefrency*" e tem dimensão de tempo. O eixo p corresponde à transformada do eixo do tempo e tem dimensão

de frequência. Além disso, o cepstrum de duas dimensões tem as seguintes propriedades,

$$c_{qp} = c_{(N-q),p} = c^*(N-q)\,,(N-q, M-p),\ 0 \le q < N/2,\ 0 \le p < M/2 \tag{2.58}$$

em que c^*_{qp} é o complexo conjugado of c_{qp}. Portanto, é suficiente considerar somente um quarto dos coeficientes do cepstrum de duas dimensões (Ariki et al., 1989; Kitamura et al., 1991).

Os coeficientes de ordem elevada no eixo q dos coeficientes c_{tk} incluem informações de "pitch"e excitação do sinal, e os coeficientes de ordem baixa correspondem ao envelope espectral, fato equivalente ao cepstrum de uma dimensão. Por outro lado, os coeficientes de ordem elevada no eixo p correspondem a variações locais no tempo, e os coeficiente de ordem baixa correspondem às variações globais no tempo. O espaço do cepstrum de duas dimensões pode ser dividido conforme a figura 2.16 (Ariki et al., 1989; Milner, 1994).

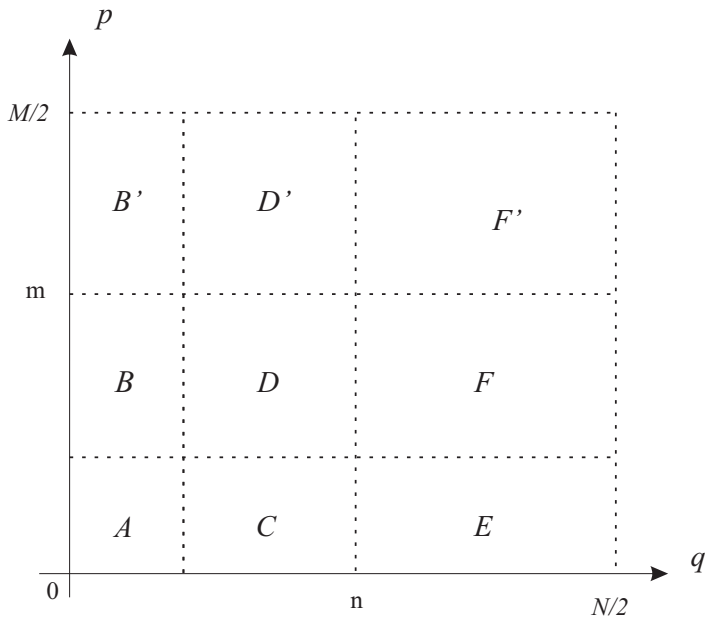


Fig. 2.16: Espaço dimensional do cepstrum de duas dimensões.

A : Valor médio do log espectral em um bloco;

B : Variações globais no tempo da média do log espectral;

B': Variações locais no tempo da média do log espectral;

C : Características do envelope espectral;

D : Variações globais do envelope espectral;

D': Variações locais do envelope espectral;

E : Variações de pitch e excitação;

F : Variações globais no tempo da estrutura fina do espectro devido a periodicidade da fonte;

F': Variações locais no tempo da estrutura fina do espectro devido a periodicidade da fonte.

As características do cepstrum de duas dimensões podem ser resumidas como segue:

- As características estáticas (A,C,E) e as características dinâmicas (B,D,F) são representadas simultaneamente;
- As variações globais no tempo (B,D,F) e as variações locais (B',D',F') são representadas simultaneamente;
- O envelope espectral (C) e a estrutura fina do espectro (E) são representados separadamente.

Os coeficientes de ordem superior (B',D',F', F, E) no cepstrum de duas dimensões têm pouca importância para o reconhecimento. A resolução de frequência e a resolução do tempo das características estáticas dependem do tamanho e do perído do segmento. Por outro lado, parâmetros das características dinâmicas dependem do tamanho e do período do bloco. (Ariki et al., 1989; Kitamura et al., 1991).

Neste capítulo abordou-se a fisiologia da voz, modelamento linear da produção da voz, bem como o modelamento espectral do sinal de voz normalmente utilizado em sistema de reconhecimento. Em continuação a análise paramétrica do sinal de voz, para efeito de reconhecimento, no próximo capítulo, tratar-se-á da da metodologia apresentada para reconhecimento de voz, utilizando-se técnicas computacionais inteligentes com o objetivo de solucionar o problema de reconhecimento de voz com o melhor desempenho possível com um número reduzido de parâmetros utilizado no modelamento do sinal de voz.

Capítulo 3

Metodologia Inteligente Híbrida para Reconhecimento de Voz

Vale ressaltar, que a ideia básica de um reconhecedor é selecionar e codificar características importantes de um dado padrão, que são consideradas modelos para o processo de reconhecimento, e armazená-las de alguma forma para que possam ser utilizados como parte do processo de reconhecimento. Posteriormente, o reconhecedor, através de alguma medida de similaridade pré-selecionada, compara os modelos armazenados com padrões a serem reconhecidos e, através da melhor medida de similaridade, seleciona o modelo que mais se aproxima do sinal a ser reconhecido. O objetivo do reconhecimento de voz é identificar que palavra, frase ou sentença foi falada através de sistemas de parametrização adequados a cada um dos tipos de identificação.

Neste trabalho apresenta-se uma metodologia inteligente híbrida para reconhecimento de voz, na qual são aplicadas técnicas computacionais inteligentes para solucionar o problema de reconhecimento de voz, sem a necessidade da utilização de cálculos de probabilidades, base da funcionalidade do HMM e da GMM. A metodologia proposta utiliza uma quantidade reduzida de parâmetros para solucionar o problema em questão. Para ser possível a utilização de uma quantidade reduzida de parâmetros, os mesmos devem conter informações fundamentais para o processo de reconhecimento. Para isso, optou-se, na etapa de pré-processamento do sinal de voz, pelo uso de uma matriz temporal bidimensional com capacidade de representar de forma confiável as variações globais e locias no tempo, bem como o envelope espectral do sinal de voz. Os elementos desta matriz serão utilizados na geração de uma base de regras para o modelamento linguístico de inferição nebulosa que será utilizada no processo de decisão através do sistema de inferência nebuloso Mamdani. Com o intuito de se obter o melhor desempenho possível no processo de reconhecimento, a base de regras do sistema de inferência nebuloso será otimizada por algoritmo genético. Outra contribuição importante da metodologia proposta é a

de que, na elaboração da base de regras para inferição nebulosa, utiliza-se o conhecimento do especialista para evitar a maldição da dimensionalidade. Na figura 3.1 apresenta-se o diagrama de blocos do sistema proposto.

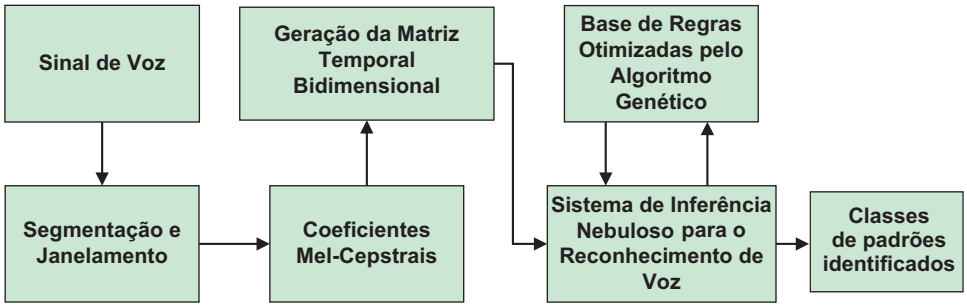


Fig. 3.1: Diagrama de Blocos do sistema híbrido proposto.

3.1 Pré-processamento do sinal de voz

As combinações de acústica, trato vocal, e características auditivas são utilizadas em grande parte dos sistemas de reconhecimento de voz. As características acústicas mais populares são os coeficientes mel-cepstrais (*Mel-Frequency Cepstrum Coefficients*-MFCC). Inicialmente o sinal de voz é digitalizado e, então dividido em segmentos. Estes segmentos passam por um processo de janelamento e, então, são codificados em um conjunto de parâmetros mel-cepstrais. Em seguida a transformada cosseno discreta (TCD) (Ahmed, 1974) é calculada e, então, uma matriz temporal bidimensional é gerada. Nela as linhas são determinadas pela quantidade de coeficientes mel-ceptrais, e as colunas, pela ordem da TCD aplicada.

3.1.1 Segmentação e Janelamento

Na figura 3.2 é ilustrado um processo de segmentação e janelamento no qual são tomados N segmentos de K amostras do sinal. No trabalho proposto, optou-se pela janela de Hamming e sobreposição entre as janelas de $66,67\%$, de modo que $\frac{T_\omega - T_f}{T_\omega} \times 100 = \frac{2}{3}$. O Tamanho da janela em amostras foi determinado através da multiplicação da duração da janela $T_\omega = 20\ ms\times$ frequência de amostragem do sinal $f_a = 22.050Hz$.

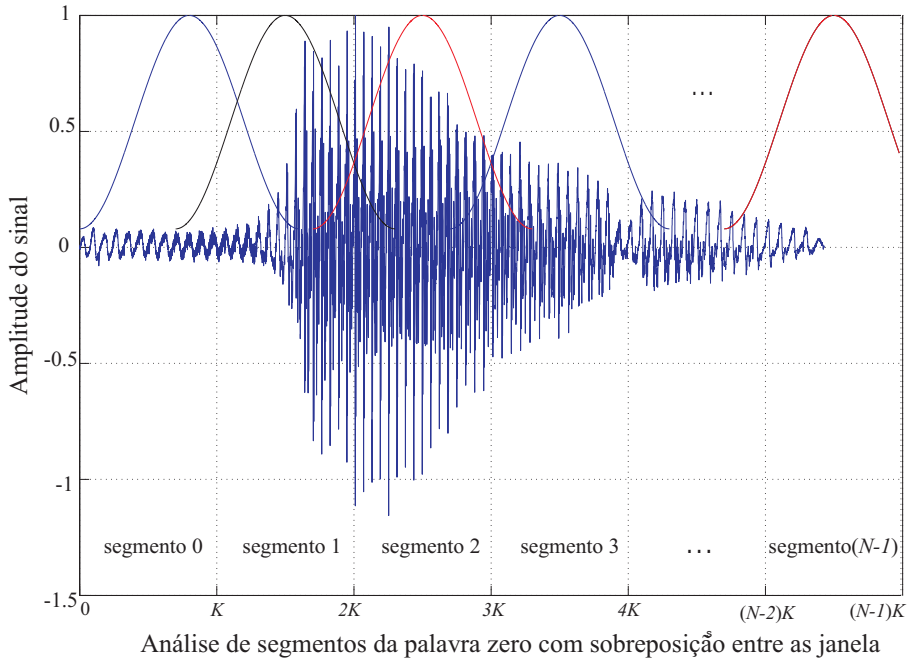


Fig. 3.2: Análise de segmentos da palavra com sobreposição entre as janelas.

3.1.2 Codificação do sinal de voz

Transformada Cosseno Discreta-TCD

A transformada cosseno discreta tornou-se especialmente útil e importante em aplicações de processamento de sinais, sobretudo devido a sua propriedade de compactação de energia (Azar, 2008; Jain, 1979; Martuci, 1994). A maior parte das características importantes do sinal tende a se concentrar em poucas componentes em baixas frequências para a TCD. A TCD está estritamente relacionada com a transformada de Fourier discreta (Pearl, 1973). Em relação à transformação de um sinal no domínio do tempo em componentes de frequências, da mesma forma que a TFD, a TCD pode ser desenvolvida através de algoritmos rápidos de conversão de domínio (Chow, 1992; Hou, 1987). A TCD é uma transformada cuja sequência resultante possui coeficientes reais; além disso, é periódica e possui simetria par. Em comparação com a TFD, a TCD apresenta melhor aproximação de um dado sinal com poucos coeficientes. A TCD tem obtido bastante interesse para aplicações em vários tipos de processamento de sinais (Haykin, 2002). Para sinais não estacionários, como, por exemplo, o sinal de voz, A TCD apresenta também boa aproximação com poucos coeficientes (Oppenheim, 2013; Sunitha, 2000).

De modo similar à TFD, a TCD corresponde a uma sequência periódica com extensão finita, de forma que o sinal original, obtido a partir da transformada inversa, possa ser recuperado. Existem muitas maneiras de se realizar esta operação; consequentemente, também existem

muitas definições para a TCD (Ahmed, 1974; Sunitha, 2000; Zhou, 2007, 2009). Neste trabalho, optou-se pelo uso da TCD-II, que possui período $2N$ e cujos pontos extremos não se sobrepõem e nenhuma modificação é necessária para garantir que o sinal seja unicamente recuperado a partir de sua transformada (Oppenheim, 2013; Zhou, 2009). A TCD-II é dada na equação (3.1), cujos coeficientes são dados por:

$$X(k) = \sum_{n=0}^{N-1} \alpha(n) x(n) cos \frac{(2k+1)n\pi}{2N} \tag{3.1}$$

$k = 0, 1, 2, ..., N-1$, e

$$\alpha(n) = \begin{cases} \sqrt{1/N}, & \text{se } n = 0 \\ \sqrt{2/N}, & \text{caso contrário} \end{cases}$$

Codificação em coeficientes mel-cepstrais

Para determinação dos MFCC's utilizou-se um banco de filtros com uma frequência limite para segmentação uniforme $F_u = 1kHz$, uma distribuição em 10 intervalos uniformes, uma frequência de amostragem mínima de 8kHz e a escala mel (Rabiner, 1993) dada na equação (2.40). A largura de banda total do filtro abrange a faixa de 0 a $4600Hz$, sendo distribuído em 20 filtros, e, através da transformada rápida de Fourier (FFT), gera-se a saída log-energia $E[i]$, dada na equação (2.46) já devidamente espaçada na escala mel. Os MFCC's são calculados através da equação:

$$mfcc[k] = \sum_{i=1}^{N_F} E[i] cos \left[\frac{i\,(k - 0.5) \cdot \pi}{N_F} \right] \tag{3.2}$$

sendo $k = 1, 2, \cdots, K$ é o número de coeficientes mel-cepstrais, N_F é o número de filtros utilizados e $E[i]$ é a saída log energia da $i-$ésima banda.

3.1.3 Geração da matriz temporal bidimensional

A matriz temporal bidimensional (Ariki et al., 1989; Milner, 1994), que é resultado da TCD realizada em uma sequência de K MFCC's, calculados em uma sequência de T vetores de observação no eixo do tempo, é obtida pela equação:

$$C_k[n, T] = \frac{1}{T} \sum_{t=1}^{T} mfcc_k[t] cos \left[\frac{(2t+1)n\pi}{2T} \right] \tag{3.3}$$

onde $k, 1 \leq k \leq K$, refere-se à k-ésima (linha) componente do $t-$ésimo segmento da observação e $n, 1 \leq n \leq N$(coluna), refere-se à ordem da TCD da matriz. Dessa forma, obtém-se a matriz de duas dimensões, em que o interesse está nos coeficientes de baixa ordem de k e n que codificam as variações locais e globais no tempo e as variações de longo prazo do envelope espectral do sinal de voz (Fissore, 1997). Esse procedimento é realizado para cada palavra falada. Desse modo, tem-se uma matriz bidimensional $C_k(n, T)$ para cada sinal de entrada. Os elementos da matriz são obtidos da seguinte forma:

1. Para uma dada palavra falada j, assumindo-se que $j \in J$, onde J é um conjunto limitado de palavras do vocabulário português brasileiro, e j^m é a *m-ésima* observação da palavra j, cada j^m é apropriadamente dividido em T_m segmentos e então codificado.

2. A partir de cada segmento das j^m observações, é gerado um total de K MFCC's e características significantes são retidas para cada segmento ao longo do tempo; isto é, c_1 do segmento t_1^m, c_1 do segmento t_2^m,..., c_1 do segmento $t_{T_m}^m$, c_2 do segmento t_1^m, c_2 do segmento t_2^m,..., c_2 do segmento $t_{T_m}^m$,..., c_K do segmento t_1^m, c_K do segmento t_2^m,..., c_K do segmento $t_{T_m}^m$. A função de mapeamento dada na equação (3.2) efetua o mapeamento do conjunto $J \subset \mathbb{R}^{j \times m}$ no espaço $\Theta \subset \mathbb{R}^{K \times T_m}$, onde $g : J \longrightarrow \Theta$, K é a quantidade de MFCC's e T_m é o número de segmentos da $m-$ésima observação. Na figura 3.3 ilustra-se a transformação dos sinais de vozes de um banco de voz (dígitos) $J = \{'0', '1', '2', '3', '4', '5', '6', '7', '8', '9'\}$ e $m = \{1, 2, ..., 10\}$ observações para cada dígito, com $k = 2$ por segmento t^m.

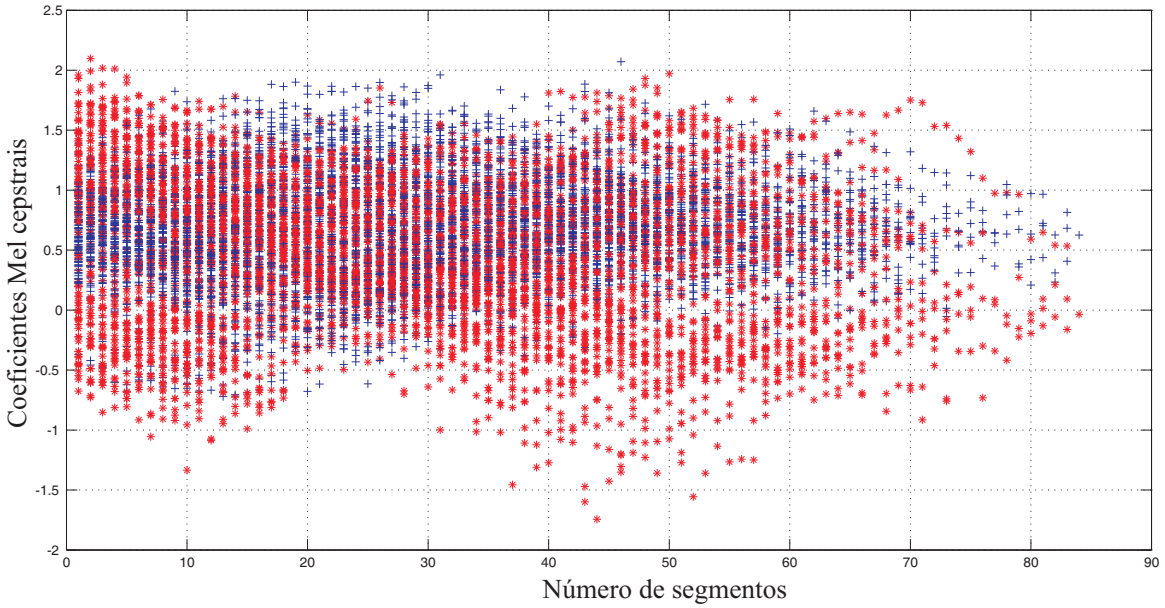


Fig. 3.3: Espaço $\Theta(K \times T_m$ segmentos).

3. Calculado os K MFCC's de uma dada observação j^m para cada segmento t^m ao longo do tempo, determina-se a TCD de ordem N para todos os coeficientes MFCC's de mesma ordem dentro dos T_m segmentos distribuídos ao longo do tempo, assim obtém-se os elementos da matriz dada na equação (3.4). Desse modo, o espaço Θ é convertido em um novo espaço de observações $\Omega \subset \mathbb{R}^{MFCC \times TCD}$, onde $f : \Theta \longrightarrow \Omega$.

$$C_{kn}^{jm} = \begin{pmatrix} c_{11}^{jm} & c_{12}^{jm} & \dots & c_{1n}^{jm} \\ c_{21}^{jm} & c_{22}^{jm} & \dots & c_{2n}^{jm} \\ \vdots & \dots & \ddots & \vdots \\ c_{k1}^{jm} & c_{k2}^{jm} & \dots & c_{kn}^{jm} \end{pmatrix} \tag{3.4}$$

Portanto, uma matriz temporal bidimensional TCD é gerada para cada observação j^m. Neste trabalho, para efeito de validação, gerou-se o espaço de observação Ω com $k = \{2, 3, 4\}$, $n = \{2, 3, 4\}$ e $m = \{1, 2, ..., 10\}$. Apresenta-se na figura 3.4 o espaço bidimensional Ω de parâmetros gerado a partir dos MFCC's e TCD, para $k = 2$ e $n = 2$ e $m = \{1, 2, ..., 10\}$ observações do dígito $j \in J$.

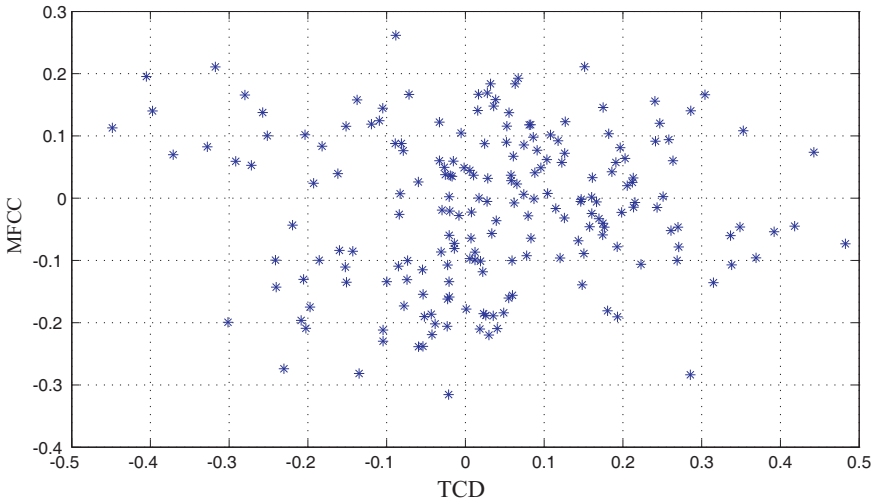


Fig. 3.4: Espaço Ω(MFCC x TCD).

4. Finalmente, matrizes de média CM_{kn}^{j} (equação 3.5) e variância CV_{kn}^{j}(equação 3.6) são geradas. Os parâmetros de CM_{kn}^{j} e CV_{kn}^{j} são utilizados para gerar matrizes com parâmetros gaussianos C_{kn}^{j} que serão utilizados como funções de pertinências para a realização do sistema de inferência nebuloso para o reconhecimento. Esses parâmetros serão otimizados pelo algoritmo genético.

$$CM_{kn}^{j} = \frac{1}{M} \sum_{m=0}^{M-1} C_{kn}^{jm} \tag{3.5}$$

$$CV_{kn}^{j} = \frac{1}{M-1}\sum_{m=0}^{M-1}\left[C_{kn}^{jm} - \left(\frac{1}{M}\sum_{m=0}^{M-1} C_{kn}^{jm}\right)\right]^2 \tag{3.6}$$

Dado o espaço Ω, calcula-se, através das equações (3.5) e (3.6), as médias e as variâncias, gerando-se um novo espaço $\Omega' \subset \Omega$, conforme descrito na figura 3.5.

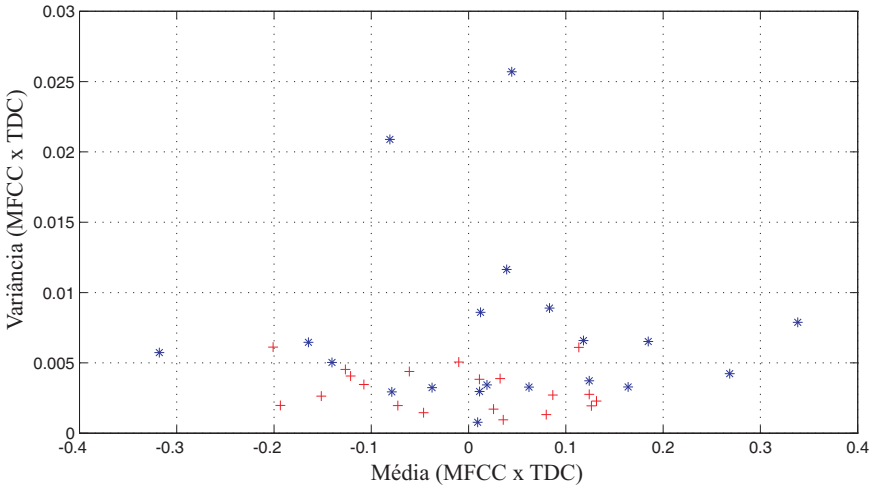


Fig. 3.5: Espaço Ω'(Média x Variância).

3.2 Sistema de Inferência Nebuloso

Na lógica Aristotélica, há dois valores verdade possíveis: proposições são verdadeiras ou falsas. Sistemas baseados nesta lógica são denominados de sistemas de lógicas bivalentes, porque consistem de dois valores lógicos. A lógica empregada no racioncínio bayesiano e em outros modelos probabilísticos também são bivalentes. A probabilidade é usada para expressar a plausibilidade de uma proposição específica ser verdadeira ou falsa.

Uma lógica polivalente pioneira foi usada para raciocinar sobre o princípio da incerteza, usado em física quântica. Essa lógica possuía três valores para representar graus de verdade. Isso é diferente de probabilidade. Por exemplo, se um dado fato tiver valor de probabilidade $(0,5)$, então, é tão provável que seja verdadeiro ou falso. Se em uma lógica polivalente uma proposição tenha valor lógico $(0,5)$, isso significa o grau até o qual esta proposição é verdadeira. Na teoria da probabilidade, lida-se com incertezas. Com o exemplo citado, não se sabe se a proposição é falsa ou verdadeira, não é ambas, nem nenhuma das duas, nem algo a meio caminho. Na lógica polivalente, não se está certo do valor verdade da proposição, ele é apenas vago, não é nem verdadeiro nem falso, ou é tanto verdadeiro como falso. Apesar de parecer

absurdo, este tipo de lógica, em especial a lógica nebulosa, tornou-se uma parte extremamente importante de sistemas inteligentes (Cox, 1999; Engelbrecht, 2003).

A lógica nebulosa é uma técnica que incorpora o raciocínio humano em modelamento de sistemas físicos. Um sistema modelado com lógica nebulosa pode ser projetado para comportar-se de acordo com o raciocínio dedutivo, isto é, inferir conclusões baseadas em informações já conhecidas. Especialistas humanos podem modelar sistemas com características não lineares e até com comportamentos dinâmicos poucos conhecidos. A lógica nebulosa pode ser implementada baseada nesse conhecimento do especialista, possibilitando o desenvolvimento de sistemas computacionais com desempenho próximo ao raciocínio humano (Coppin, 2004; Cox, 1999). A lógica nebulosa é utilizada para racionar em conjuntos nebulosos, que contrastam com os conjuntos usados na teoria tradicional de conjuntos, que para efeito de diferenciação, são chamados de conjuntos crisp[1]. Uma descrição mais detalhada dos fundamentos dos sistemas nebulosos é dada no Anexo (A).

3.3 Sistema de inferência nebuloso utilizado no reconhecimento de voz

Na figura 3.6 ilustra-se a estrutura geral do sistema de inferência nebuloso utilizado neste trabalho, cuja função é mapear todos os elementos de entradas, oriundos do processo de parametrização e codificação, dados na matriz bidimensional C_{kn}, em saídas cujos os valores são os dígitos $j \in J$. O sistema de inferência consiste, basicamente, de quatro componentes: fuzificador, base de conhecimento, máquina de inferência e defuzificador. Uma vez estabelecidas as regras, o sistema de inferência mapeia as entradas c_{kn} em saídas, $y = f(c_{kn})$, onde f corresponde à representação quantitativa deste mapeamento.

A base de conhecimento é constituída pela base de dados e pela base de regras de maneira a caracterizar o funcionamento completo do sistema de inferência. Na base de dados estão armazenadas as variáveis línguísticas c_{kn} as definições do respectivo universo de discurso e as funções de pertinências, caracterizando os termos linguísticos utilizados para cada variável linguística. Na Base de regras estão as declarações linguísticas do tipo *SE-ENTÃO* definidas pelo conhecimento do especialista e extraídas de dados numéricos que caracterizam os padrões a serem reconhecidos. A Interface de fuzificação (ou Fuzificador) mapeia os números das variáveis de entrada em conjuntos nebulosos. A máquina de inferência mapeia os conjuntos nebulosos da entrada em conjuntos nebulosos na saída, de acordo com as características da base de dados

[1]Adota-se neste trabalho a definição de crisp, como sentido bem definido em oposição a nebuloso. Outra opção seria adotar conjunto clássico.

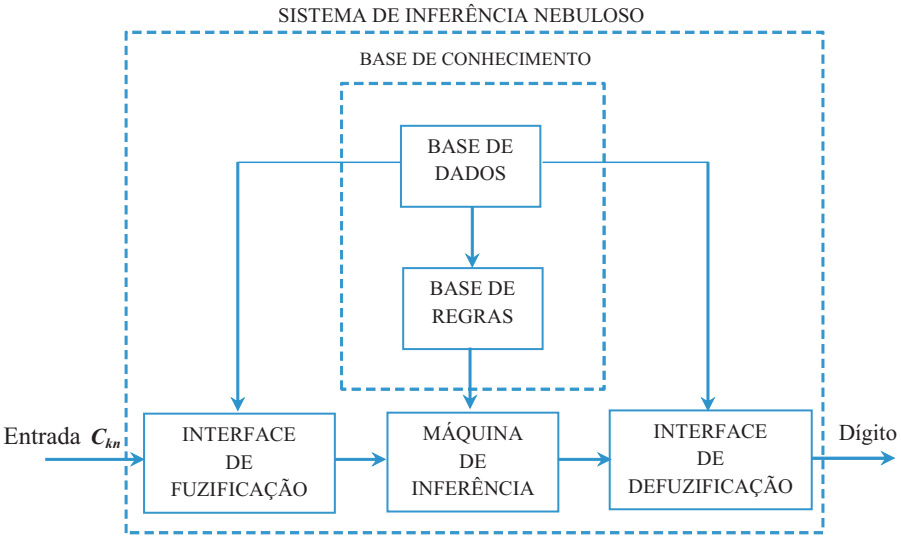


Fig. 3.6: Sistema de inferência nebuloso.

e da base de regras, combinando as regras. A Interface de defuzificação (ou Defuzificador) mapeia os conjuntos nebulosos de saída nos dígitos reconhecidos.

O sistema de inferência nebuloso Mamdani, proposto por E. H. Mamdani (Mamdani, 1977), foi utilizado como forma de capturar o conhecimento qualitativo disponível na base de regras elaborada para o sistema nebuloso de reconhecimento. O sistema proposto apresenta $k \times n$ valores lingüísticos relacionados a todos os elementos $\{c_{11},\ldots, c_{kn}\}$ da matriz bidimensional C_{kn}; e cada variável de entrada é qualificada por um valor linguístico j associado ao padrão j a ser reconhecido, resultando em um total de $j^{k\times n}$ regras. Verifica-se que para um sistema de reconhecimento para dez dígitos, proposto neste trabalho, considerando uma matriz bidimensional quadrada de ordem 2, de ordem 3 e de ordem 4, o sistema nebuloso necessitaria de 10^4, 10^9, 10^{16} regras, respectivamente. Observa-se que esse sistema sofre do problema da maldição da dimensionalidade de Bellman (Bellman, 1961), que diz que a complexidade intríseca de uma classe de funções aproximativas aumenta exponencialmente na razão $\left(\frac{m_0}{s}\right)$, onde m_0 é a dimensionalidade de entrada e s é um índice de suavidade que mede o número de restrições imposta à função aproximativa daquela classe particular.

Uma vez que, independentemente da técnica de aproximação utilizada, se o índice de suavidade s for mantido constante, o número de parâmetros necessários para a função aproximativa manter um determinado grau de precisão aumenta exponencialmente com a dimensionalidade de entrada m_0. O único modo de se conseguir uma taxa adequada de aproximação independente da dimensionalidade de entrada m_0, e dessa forma ser imune à maldição da dimensionalidade, é fazer com que o índice de suavidade aumente com o número de parâmetros da função aproxima-

tiva, de forma a compensar o aumento de complexidade. A proposta deste trabalho é utilizar o conhecimento do especilista para a elaboração da base de regras, de forma que o parâmetro de suavidade seja estabelecido de acordo com as regras, isto é, propõem-se utilizar somente regras que fazem sentido ao problema de reconhecimento especificado. Deste modo, a $l-$ésima regra da base de regras para o problema de reconhecimento é dada por:

$$\begin{aligned} Ru^l: \quad & SE \;\; c_{11} \;\; \text{é} \;\; \tilde{c}^j_{11} \;\; E \;\; c_{12} \;\; \text{é} \;\; \tilde{c}^j_{12} \;\; E \cdots E \;\; c_{1N} \;\; \text{é} \;\; \tilde{c}^j_{1N} \\ & E \;\; c_{21} \;\; \text{é} \;\; \tilde{c}^j_{21} \;\; E \;\; c_{22} \;\; \text{é} \;\; \tilde{c}^j_{22} \;\; E \cdots E \;\; c_{2N} \;\; \text{é} \;\; \tilde{c}^j_{2N} \;\; \cdots \\ & \cdots \;\; E \;\; c_{K1} \;\; \text{é} \;\; \tilde{c}^j_{K1} \cdots E \;\; c_{Kn} \;\; \text{é} \;\; \tilde{c}^j_{KN} \;\; \text{ENTÃO} \;\; y \;\; \text{é} \;\; \tilde{y}^j \end{aligned} \tag{3.7}$$

onde, $l^{|l=1,2,\cdots,M}$ é o número de regras, $c_{11}, c_{12}, \cdots c_{KN}$, são as variáveis línguisticas do antecedente (entrada) e y é a variável linguística do consequente (saída). Os conjuntos nebulosos $\tilde{c}^j_{11}$, $\tilde{c}^j_{12}$,$\cdots$, $\tilde{c}^j_{KN}$ e $\tilde{y}^j$ são os valores línguísticos utilizados para particionar os universos de discursos das variáveis linguísticas do antecedente e do consequente, onde j representa os padrões a serem reconhecidos. De uma forma geral, a variável c_{kn} pertence ao conjunto nebuloso $\tilde{c}^j_{kn}$ com um valor $\mu_{\tilde{c}^j_{kn}}$ definido por uma função de pertinência $\mu_{c_{kn}} : \mathbb{R} \rightarrow [0,1]$ e a variável y pertence ao conjunto nebuloso $\tilde{y}^j$ com um valor $\mu_{\tilde{y}^j}$ definido por uma função de pertinência $\mu_y : \mathbb{R} \rightarrow [0,1]$. Vale ressaltar que, em cada regra, todas as variáveis linguística de entrada c_{11}, $\cdots$, c_{kn} são avaliadas somente para o mesmo padrão j, suprimindo-se termos cruzados, tais como, $SE \;\; c_{11} \;\; \text{É} \;\; \tilde{c}^0_{11} \;\; E \;\; c_{12} \;\; \text{É} \;\; \tilde{c}^1_{11}$, etc. Assim, como o problema de reconhecimento é definido para o padrão j, o total de regras é de j regras, em vez de $j^{k \times n}$ regras, reduzindo-se a complexidade computacional e superando a maldição da dimensionalidade.

3.3.1 Fuzificação

O mapeamento dos valores precisos da entrada em conjuntos nebulosos é realizado por um fuzificador gaussiano. A escolha do tipo de fuzificador foi baseada nas características dos parâmetros obtidos na etapa de codificação do sinal de voz. Os elementos das matrizes CM^j_{kn} e CV^j_{kn}, calculados durante o processo de treinamento, são utilizados para gerar os valores linguísticos de entrada $\tilde{c}^j_{kn}$ para o padrão j, associada à variável linguística de entrada c_{kn} do reconhecedor nebuloso, dados por:

$$\tilde{c}^j_{kn} = \exp^{-\frac{\left(c_{kn} - cm^j_{kn}\right)^2}{2 \times cv^j_{kn}}} \tag{3.8}$$

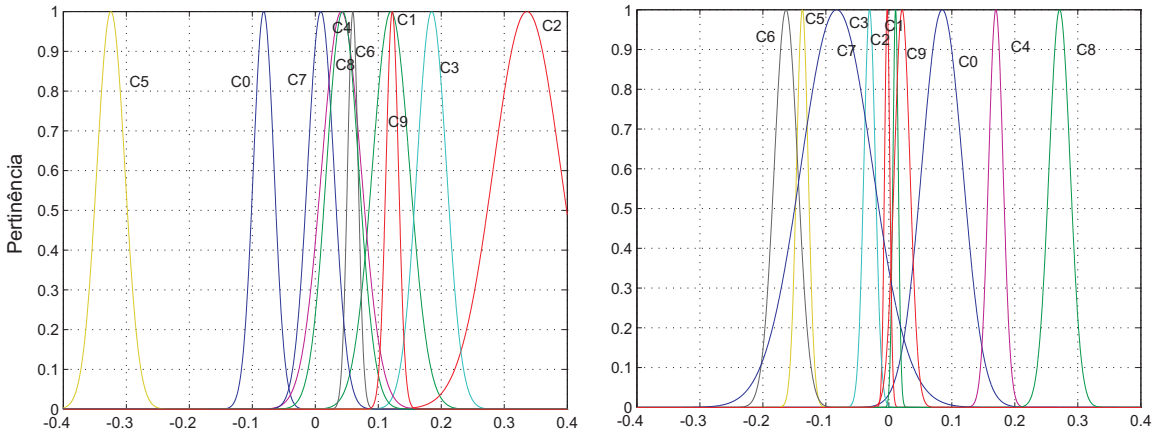
onde $j \in J$ é o padrão a ser reconhecido, $k\,\big|^{k=1,2,\cdots,K}$ é o número de MFCC's, $n\,\big|^{n=1,2,\cdots,N}$ é a ordem da TCD, c_{kn} é qualquer elemento da matriz bidimensional C_{kn} dada na equação (3.3), cm_{kn}^{j} é o elemento da matriz de médias CM_{kn}^{j} dada na equação (3.5) e cv_{kn}^{j} é o elemento da matriz de variâncias CV_{kn}^{j} dada na equação (3.6).

O valor linguístico $\tilde{y}^{j}$ para o padrão j, associado à variável linguística de saída y do reconhecedor nebuloso é dado por:

$$\tilde{y}^{j} = \exp^{-\frac{(y-j)^2}{2\times\left(\sigma_y^j\right)^2}} \tag{3.9}$$

onde $j \in J$ é o padrão a ser reconhecido, y é qualquer valor do universo de discurso J para a saída do reconhecedor nebuloso, e $\left(\sigma_y^j\right)^2$ é o valor da variância atribuída a j.

Na figura 3.7 ilustram-se as funções de pertinências das entrada c_{kn} considerando $k = n = 2$.



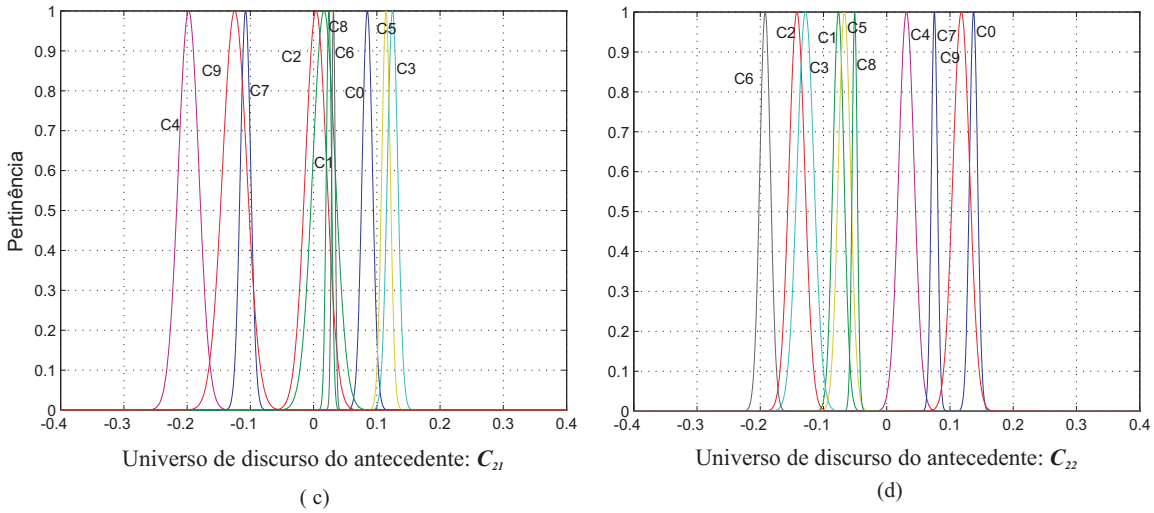


Fig. 3.7: Representação das funções de pertinências devidamente particionadas com $k = n = 2$. (a) Partição associada a c_{11}; (b) Partição associada a c_{12}; (c) Partição associada a c_{21} e (d) Partição associada a c_{22}.

Na figura 3.8 mostram-se as funções de pertinências associadas aos elemento de saída y.

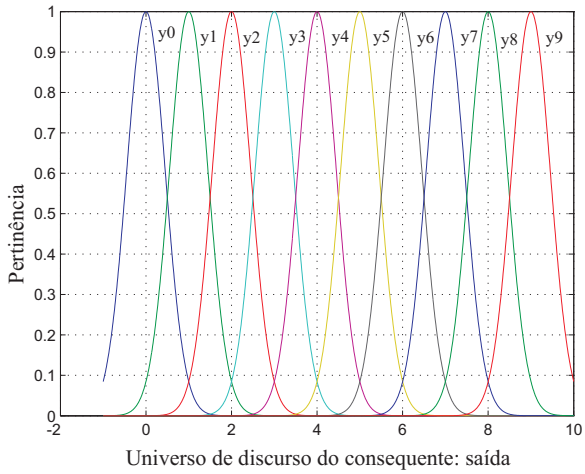


Fig. 3.8: Representação das funções de pertinências do consequente associadas a saída.

3.3.2 Máquina de inferência para o problema de reconhecimento de voz

De acordo com a figura 3.9, assumindo-se que $\tilde{c}_{kn}^{j}$ seja um conjunto nebuloso de entrada, $\tilde{y}^{j}$ seja um conjunto nebuloso de saída e Q seja uma relação nebulosa em $(\Omega \times J)$ e projetando-se $\tilde{c}_{kn}^{j}$ em Q, obtém-se $\tilde{c}_{Ekn}^{j}$. A interseção desta extensão com a relação nebulosa Q resulta em

$\tilde{c}^{j}_{Ekn} \cap Q$, cuja projeção no eixo y gera o conjunto nebuloso $\tilde{y}^{j}$, conforme mostrado na figura 3.9.

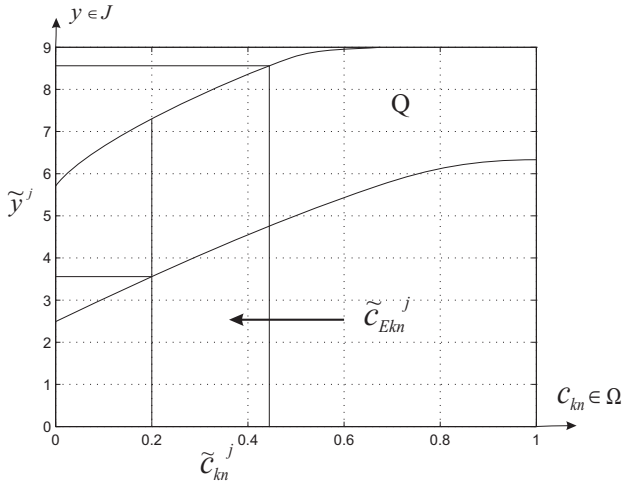


Fig. 3.9: Conjunto de saída nebuloso $\tilde{y}$ da inferência nebulosa do conjunto de entrada nebuloso $\tilde{c}m_{kn}$

De forma mais específica, dadas as pertinências $\mu_{\tilde{c}^{j}_{kn}}(c_{kn})$ e $\mu_{\tilde{c}^{j}_{Ekn}}(c_{kn}, y)$, tem-se:

$$\mu_{\tilde{c}^{j}_{Ekn}}(c_{kn}, y) = \mu_{\tilde{c}^{j}_{kn}}(c_{kn}) \tag{3.10}$$

e consequentemente,

$$\mu_{\tilde{c}^{j}_{Ekn} \cap Q}(c_{kn}, y) = t\left[\mu_{\tilde{c}^{j}_{Ekn}}(c_{kn}, y), \mu_{Q}(c_{kn}, y)\right] \tag{3.11}$$

Fazendo-se a projeção de $\tilde{c}^{j}_{Ekn} \cap Q$ em y, tem-se que,

$$\mu_{\tilde{y}^{j}}(y) = \sup_{x \in J} t\left[\mu_{\tilde{c}^{j}_{Ekn}}(c_{kn}), \mu_{Q}(c_{kn}, y)\right] \tag{3.12}$$

Das equações (A.10), (A.12) e (3.12) pode-se determinar a projeção em y dada por:

$$\mu_{\tilde{y}^{j}}(y) = \sup_{x \in J} t\left[\mu_{\tilde{c}^{j}_{kn}}(c_{kn}), \mu_{c_{kn} \to y}(c_{kn}, y)\right] \tag{3.13}$$

Portanto, de acordo com as etapas descritas na seção (A.1.2) e a equação (3.7), tem-se que:

$$\mu_{Ru^{l}}(\bar{\mathbf{c}}, y) = \mu^{l}_{\bar{\mathbf{c}} \to y}(\bar{\mathbf{c}}, y) \tag{3.14}$$

onde, $\bar{\mathbf{c}} = [c_{11}, c_{12}, \ldots, c_{1N}, \ldots, c_{k1}, \cdots, c_{KN}]$ é o vetor de entradas linguísticas dos padrões que gerarão os modelos. Utilizando-se a implicação Mamdani baseada no mínimo dada na equação (A.17), e o produto para a $norma - t$ relativa a proposição E , obtém-se a relação nebulosa

Mamdani para a $l-$ésima regra, dada por:

$$\mu_{Ru^l}(\bar{\mathbf{c}}, y) = min \left[\prod_{k=1}^{K} \prod_{n=1}^{N} \exp^{-\frac{\left(c_{kn} - cm_{kn}^{j}\right)^2}{2 \times cv_{kn}^{j}}}, \exp^{-\frac{(y-j)^2}{2 \times \left(\sigma_y^j\right)^2}} \right] \tag{3.15}$$

Por consequência, utilizando-se a inferência definida na equação (A.9), a saída nebulosa para o dígito a ser reconhecido é dada por:

$$\mu_{\tilde{y}'}(y) = \max_{l=1}^{L} \left\{ \sup_{\bar{\mathbf{c}}^* \in \Omega} min \left[\mu^l \left(\bar{\mathbf{c}}^* \right), \mu_{Ru^l}(\bar{\mathbf{c}}, y) \right] \right\} \tag{3.16}$$

onde, $\bar{\mathbf{c}}^* = [c_{11}^*, c_{12}^*, \ldots, c_{1N}^*, \ldots, c_{k1}^*, \cdots, c_{KN}^*]$, cujos os valores fuzificados são dados no vetor abaixo:

$$\mu^l \left(\bar{\mathbf{c}}^* \right) = \left[\exp^{-\frac{\left(c_{11}^* - cm_{11}^{j}\right)^2}{2 \times cv_{11}^{j}}}, \exp^{-\frac{\left(c_{12}^* - cm_{12}^{j}\right)^2}{2 \times cv_{12}^{j}}}, \cdots, \exp^{-\frac{\left(c_{KN}^* - cm_{KN}^{j}\right)^2}{2 \times cv_{KN}^{j}}} \right] \tag{3.17}$$

3.3.3 Defuzificação

Para realizar o mapeamento do conjunto nebuloso $\tilde{y}\,' \in J\,' = [-2, 12] \subset \mathbb{R}$ em um valor preciso $y^* \in J$ utilizou-se o processo de defuzificação para especificar um ponto em J' que melhor representa o conjunto nebuloso $\tilde{y}\,'$. A saída defuzificada do reconhecedor nebuloso é baseada na média dos máximos, cuja operação é escolher o valor de y^* como um ponto em J, no qual $\mu_{\tilde{y}'}(y)$ alcança seu valor máximo. Define-se, assim, um $hgt\left(\tilde{y}\,'\right)$ conjunto, conforme abaixo,

$$hgt\left(\tilde{y}\,'\right) = \left\{ y \in J | \mu_{\tilde{y}'}(y) = \sup_{y \in J} \mu_{\tilde{y}'}(y) \right\} \tag{3.18}$$

Isto é, $hgt\left(\tilde{y}\,'\right)$ é o conjunto de todos os pontos em J nos quais $\mu_{\tilde{y}'}(y)$ alcança seu valor máximo. Especificamente, se $hgt\left(\tilde{y}\,'\right)$ contém um único ponto, então, y^* é um valor relativo a esse ponto; caso contrário, então, pode-se determinar o valor de y^* escolhendo-se uma das técnicas a saber: o bisetor, centróide, o menor dos máximos, o maior dos máximos, ou a média dos máximos. Para o problema proposto, o sistema de inferência nebuloso tem J saídas, uma vez que para cada saída j tem-se um máximo, o defuzificador que apresentou os resultados mais coerentes com o problema de reconhecimento de dígitos foi o da média dos máximos, que é dado por:

$$y^* = \frac{\int_{hgt(\tilde{y}\,')} y dy}{\int_{hgt(\tilde{y}\,')} dy} \tag{3.19}$$

onde, $\int_{hgt(\tilde{y}\,')}$ é a integração usual para $hgt\,(\tilde{y}\,')$ contínuo e o somatório para $hgt\,(\tilde{y}\,')$ discreto.

3.4 Otimização do reconhecedor nebuloso com algoritmo genético

Otimização é o processo pelo qual se ajustam as entradas ou características de um dispositivo, modelo matemático, ou qualquer outro experimento cujo objetivo seja encontrar um resultado mínimo ou máximo para a saída. O processo ou função a ser otimizada chama-se função custo, função objetivo ou função aptidão. Considerando que o custo é algo que se deseja minimizar, então, otimização torna-se um problema de minimização. Caso o processo necessite da maximização, basta que se inverta o sinal da função custo.

A otimização pode ser aplicada para modelos matemáticos existentes ou na elaboração de modelos. Neste caso, o objetivo é encontrar um ponto extremo de saída de um dado modelo, alterando-se os parâmetros do modelo. A razão usual para se encontrar parâmetros adequados é encontrar a melhor saída para o modelo. Neste trabalho, utilizou-se otimização não linear bio-inspirada denominada Algoritmo Genético (Tang et al., 1998; Weihong et al., 2010; Zhang et al., 2010).

O algoritmo genético(AG) é uma técnica de busca baseada no princípio de reprodução genética e seleção natural, a qual permite que uma população composta de muitos indivíduos possa ser envolvida sobre certas regras de seleção na obtenção do melhor indivíduo que otimize uma função aptidão ou função custo. Esse método foi desenvolvido por John Holland (Holland, 1975). A mais simples representação para o algoritmo genético é considerar um vetor, denominado cromossomo, e seus elementos denominados de genes. Estes dois termos foram apropriados diretamente da genética.

Uma população inicial é gerada de forma pseudo-aleatória, devido às restrições imposta na otimização, e consiste de um conjunto de cromossomos compostos de genes. Um cromossomo representa um indivíduo dentro da população e os genes representam as características desse indivíduo. Esse indivíduo é a representação completa de uma solução (Tang et al., 1997). No processo de busca da melhor solução, os cromossomos da população são combinados para formar um novo indivíduo com potencial de ser uma solução melhor que a anterior para o problema a ser resolvido. Para ilustrar a possibilidade de se maximizar os acertos do reconhecedor nebuloso, definiu-se, neste trabalho, a seguinte função custo:

$$f\,(j, y^*) = \frac{1}{M} \sum_{m=1}^{M} E\,(j^m, y_m^*) \tag{3.20}$$

sendo j^m a $m-$ésima observação do padrão j, y_m^* é a saída defuzificada do reconhecedor nebuloso na $m-$ésima observação, dada na equações (3.19) e $E\left(j^m, y_m^*\right)$ é o erro, definido por:

$$E(j^m, y_m^*) = \begin{cases} 1, & \text{if } y_m^* \neq j^m \\ 0, & \text{caso contrário} \end{cases} \tag{3.21}$$

Os elementos cm_{kn}^j e cv_{kn}^j, de acordo com a equação (3.8), são parâmetros dos valores linguísticos $\tilde{c}_{kn}^j$ a serem calculados geneticamente no procedimento de treino do reconhecedor nebuloso, minimizando-se a função custo. Portanto, há um valor mínimo para a função custo $f\left(j, y^*\right)$ para os valores ótimos cm_{kn}^j e cv_{kn}^j que ajustam a relação nebulosa Mamdani dada na equação (3.15), tal que a função de classificação correta $H(f)$, dada por

$$H\left(f\right) = \left[1 - f\left(j, y^*\right)\right] \times 100\% \tag{3.22}$$

é máxima.

O procedimento computacional para a otimização genética (Haupt, 2004; Zhang et al., 2010) do reconhecedor nebuloso é resumido como segue:

Passo 1 - Definem-se a função custo, as variáveis e os parâmetros do AG:

- Os genes dos cromossomos são os elementos cm_{kn}^j and cv_{kn}^j das matrizes CM_{kn}^j e CV_{kn}^j, respectivamente:
  $crom \triangleq \left[cm_{11}^j, \cdots, cm_{1N}^j, \cdots, cm_{K1}^j, \cdots, cm_{KN}^j, cv_{11}^j, \ \cdots, cv_{1N}^j, \cdots, cv_{K1}^j, \cdots, cv_{KN}^j\right]$
- Determinam-se o número de gerações, o tamanho da população e o critério de parada;
- Gera-se a população inicial.

Passo 2 - Encontra-se a função custo para cada cromossomo da população:

- A função custo é dada por $f(j, y^*)$ definida nas equações (3.20) e (3.21);
- A população é organizada em ordem decrescente do cromossomo de mais alto custo para o cromossomo de mais baixo custo.

Passo 3 - Seleciona-se para o cruzamento entre os indíviduos:

- 50% dos cromossomoss de mais alto custo são selecionado para o cruzamento.

Passo 4 - Cruzamento:

– O cruzamento aritmético é usado para combinar linearmente dois cromossomos. Se os pais $[crom]_a^t$ e $[crom]_b^t$ são selecionados no processo de cruzamento, na $t-$ésima geração, os descendentes são obtidos como segue:

$$[crom]_a^{(t+1)} = \alpha\,[crom]_b^t + (1-\alpha)\,[crom]_a^t \tag{3.23}$$

$$[crom]_b^{(t+1)} = (1-\alpha)\,[crom]_b^t + \alpha\,[crom]_a^t \tag{3.24}$$

onde $\alpha \in [0,1]$ é um número aleatório;

Passo 5 - Mutação: aleatória normalmente distribuída (Haupt, 2004) e não uniforme (Michalewicz, 1994):

– Se o gene γ_k^t é selecionado para mutação, na $t-$ésima geração, sendo $t \leq \frac{\text{geração}}{2}$, na posição k no cromossomo, então, o gene mutante será dado por:

$$\gamma_k^{(t+1)} = \gamma_k^t + \sigma\mathcal{N}^t\,(0,1) \tag{3.25}$$

com a restrição $l_k \leq \gamma^{(t)} \leq u_k$, onde l_k e u_k são os limites de restrição para o gene γ_k, σ é o desvio padrão da distribuição normal, $\mathcal{N}\,(0,1)$ é a distribuição normal padrão (media $= 0$ and variância $= 1$).

– Se o gene γ^t é selecionado para mutação, na $t-$ésima geração, sendo $t >$ geração$/2$, o gene mutante será dado por:

$$\gamma_k^{t+1} = \begin{cases} \gamma_k^t + \Delta(t, u_k - \gamma_k^t), & \text{se } \delta \leq 0,5 \\ \gamma_k^t - \Delta(t, \gamma_k^t - l_k), & \text{caso contrário} \end{cases} \tag{3.26}$$

onde $\delta \in [0,1]$ é um número aleatório, l_k e u_k são os limites inferior e superior, respectivamente do gene γ_k. A função $\Delta(t,y)$ retorna um valor na faixa $[0,y]$, tal que $\Delta(t,y)$ aproxima-se de zero quando t aumenta. Esta propriedade proporciona a este operador uma busca uniformemente espaçada inicialmente, e uma busca mais localizada nas etapas finais do algoritmo. A função Δ é dada por:

$$\Delta(t,y) = y \cdot (1 - r^{(1-\frac{t}{T})^b}) \tag{3.27}$$

sendo r um número aleatório uniformemente distribuído entre $[0,1]$, T é o número máximo de gerações e b é o parâmetro do sistema que determina o grau de não uniformidade.

Passo 6 - Verificação da Convergência:

- Encontrar o custo para cada descendente;
- A população é ordenada do cromossomo de mais alto custo para o mais baixo custo;
- Manter o melhor cromossomo ("Melhor solução");
- O critério de convergência foi atendido? Não: voltar ao **Passo 3**.

Passo 7 - Fim.

O fluxograma do algoritmo genético usado para otmização do sitema de inferência nebuloso proposto é mostrado na figura 3.10 (Haupt, 2004; Zhou, 2007). Com o intuito de se obter o melhor desempenho tanto em velocidade de convergência quanto em quantidade de digítos reconhecidos corretamente, o algoritmo genético foi configurado com uma população inicial de 100 indivíduos, 500 gerações, com probabilidade de mutação igual a 20%, um cromossomo com oitenta genes para otimizar a função custo com 80 variáveis (40 variáveis de médias e 40 variáveis de variâncias de cada padrão) para o processo de reconhecimento do sistema nebuloso proposto com ($K = N = 2$). Para o caso de ($K = N = 3$) foram otmizadas 180 variáveis (90 variáveis de médias e 90 variáveis de variâncias de cada padrão) e para ($K = N = 4$) foram otimizadas 320 variáveis (160 variáveis de médias e 160 variáveis de variâncias de cada padrão). Com a solução final dada pelo algoritmo genético, tem-se todos os parâmetros otimizados dos padrões que serão utilizados no processo de reconhecimento.

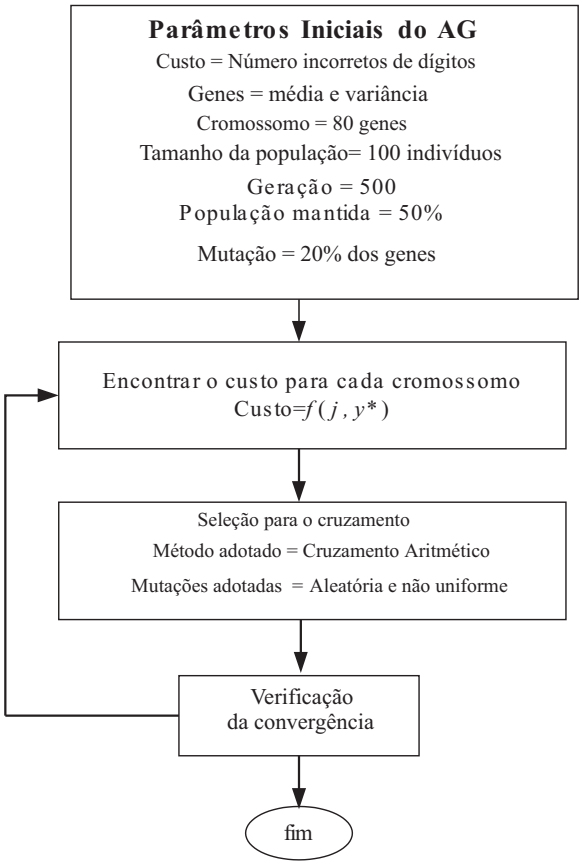


Fig. 3.10: Fluxograma do algoritmo genético utilizado.

No próximo capítulo serão apresentados os resultados obtidos com a metodologia proposta, juntamente com uma análise desses resultados quando comparados com o *Hidden Markov Models-HMM*, *Suport Vector Machine-SVM* e *Gaussian Mixture Models-GMM*.

Capítulo 4

Resultados Experimentais

4.1 Processo de Treinamento

Neste trabalho, para efeito de procedimentos práticos, utilizou-se o programa Matlab, na versão R2013a(8.1.0.604), 64 bits (win64). As funções e programa principal foram desenvolvidos através de linhas de comandos. Foram utilizados os dez dígitos, zero (0) a nove (9), pronunciados na língua portuguesa brasileira, cujas pronuncias em portugês e IPA são apresentadas na Tabela (4.1), conforme o sistema IPA ("*International Phonetic Alphabet symbols*').

Tab. 4.1: Dígitos utilizados no sistema de reconhecimento de voz

Dígito	Escrita em Português	Pronúncia	IPA
'0'	zero	zeh-ro	['zɛru]
'1'	um	oom	[ũ]
'2'	dois	doy-z	['doiʃ]
'3'	três	treh-z	['trejʃ]
'4'	quatro	kwah-trouh	['kwatru]
'5'	cinco	seen-coh	['s̃inku]
'6'	seis	say-z	['sejʃ]
'7'	sete	seh-chee	['sɛtʃi]
'8'	oito	oy-too	['ojtu]
'9'	nove	noh-vee	['nɔvi]

Para estes procedimentos práticos, foram utilizados os seguintes banco de voz:

1. Banco de Voz do Laboratório de Processamento de Sinais-LPS, da Escola Politécnica da Universidade de São Paulo (EPUSP)[1]: banco de voz gravado em ambiente de laboratório, em sala acústica, com baixo nível de ruído, composto de vozes de cinco locutores do sexo

[1]http://www.bv.fapesp.br/en/auxilios/58789/analysis-of-audio-and-speech-signals-for-reconstruction-and-recognition/

masculino e cinco locutores do sexo feminino, todos na faixa etária de 18 a 30 anos de idade. Cada um dos locutores pronunciou os exemplos dos dígitos duas vezes, num total de duzentas locuções, com pausa entre as pronúncias de cada dígito;

2. Banco de voz do Instituto Nacional de Telecomunicações (Inatel) apresentado no trabalho (Ynoguti, 2008). Deste banco, foram tomadas pronúncias dos dígitos de cinco locutores do sexo masculino e cinco do sexo feminino, todos na faixa etária de 18 a 50 anos de idade. Cada um dos locutores pronunciou os exemplos dos dígitos uma vez, num total de cem locuções. Vale ressaltar que este banco é composto de exemplos de dígitos pronunciados de forma contínua, sem pausa entre as pronúncias dos dígitos.

3. Banco de voz gravado no Instituto Federal do Maranhão (IFMA): Banco de voz gravado em ambiente sem controle de ruído, composto de vozes de doze locutores do sexo masculino e doze locutores do sexo feminino, todos na faixa etária de 18 a 50 anos de idade. Cada um dos locutores pronunciou os exemplos dos dígitos dez vezes, num total de duas mil e quatrocentas locuções, com pausa entre as pronúncias de cada dígito.

Para o procedimento de geração dos modelos no processo de treinamento foram utilizadas 100 locuções do banco da EPUSP, sendo 50 locuções masculinas e 50 locuções femininas. Após o pré-processamento e a codificação do sinal de voz, uma matriz bidimensional C_{kn} foi gerada para todas as $(m = 10)$ observações do dígito j, para serem utilizadas nos procedimentos de teste. Os parâmetros cm_{kn}^{j} e cv_{kn}^{j}, dados na equação (3.8), correspondentes ao valor lingüístico $\tilde{c}_{kn}^{j}$ para o dígito $j\,|^{j=0,1,2,\cdots,9}$, são otimizados pelo AG, tal que, a função custo $f\,(j, y^{*})$ seja minimizada. No processo de otimização, foram feitas 15 realizações do AG, cujos resultados são mostrados na figura 4.1. Obteve-se o resultado estatístico dado por $H(f) = 96,3750 \pm 0,7188$.

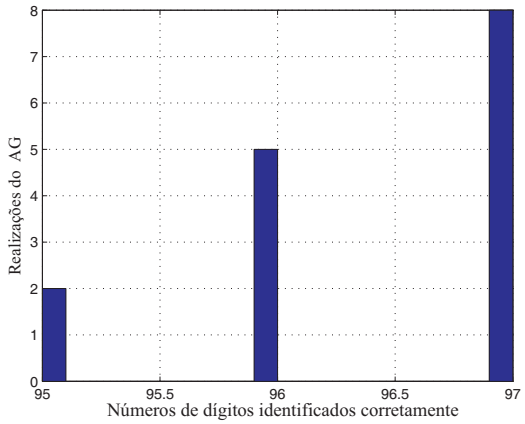


Fig. 4.1: Histograma dos resultados para 15 realizações do processo de treinamento.

A atuação do AG na otimização da função custo $f\,(j, y^*)$, no treinamento, onde a ordem da matriz C_{kn} é $K = N = 2$, é ilustrada na figura 4.2. O resultado obtido foi $f(j, y^*) = 3\%$, e o número de acertos $H(f) = 97\%$.

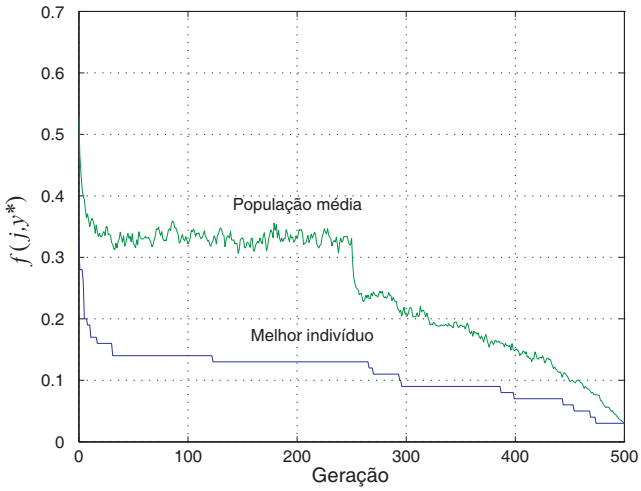


Fig. 4.2: Melhor resultado obtido no processo de treinamento com a matriz C_{kn} de ordem $K = N = 2$.

Ilustra-se, na figura 4.3, o processo de otimização com AG no treinamento, com a matriz C_{kn} de ordem $K = N = 3$. O resultado obtido foi $f(j, y^*) = 2\%$, e o número de acertos $H(f) = 98\%$.

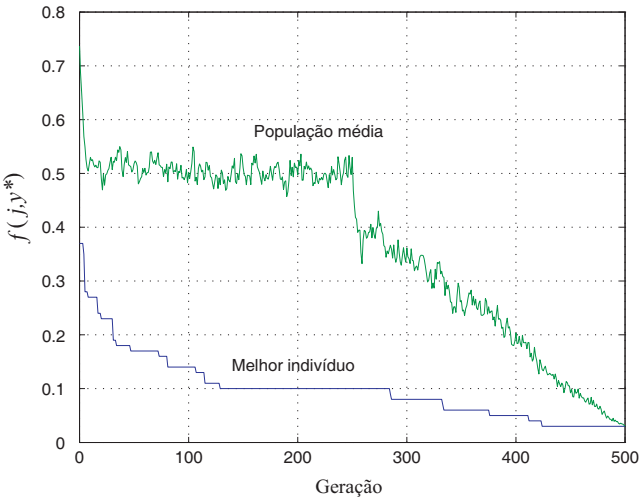


Fig. 4.3: Melhor resultado obtido no processo de treinamento com a matriz C_{kn} de ordem $K = N = 3$.

Na figura 4.4 mostra-se o resultado da otimização com AG para a matriz C_{kn} de ordem $K = N = 4$ no processo de treinamento. O resultado obtido foi $f(j, y^*) = 0\%$, e o número de acertos $H(f) = 100\%$.

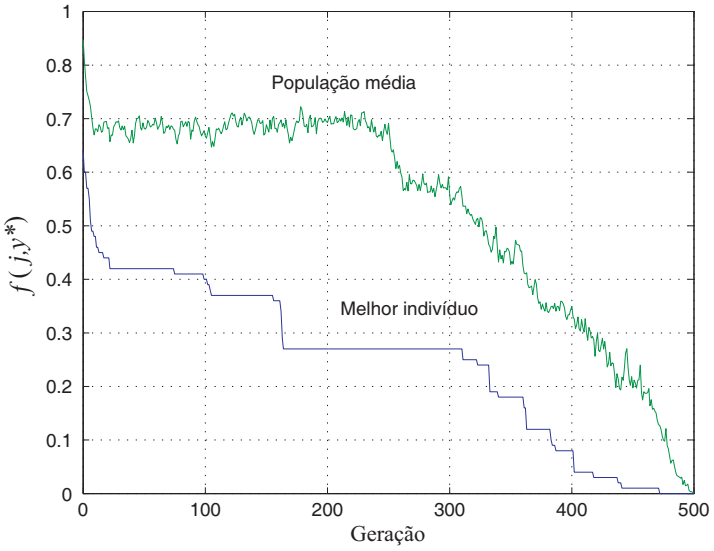


Fig. 4.4: Melhor resultado obtido no processo de treinamento com a matriz C_{kn} de ordem $K = N = 4$.

Para efeito de observação da atuação do algoritmo genético nas funções de pertinências geradas a partir da matriz C_{kn} de ordem $K = N = 2$, apresentam-se na figura 4.5 as funções de pertinências do melhor indivíduo para a primeira geração; neste caso, o total de dígitos identificados corretamente foi 64%. Na figura 4.6 mostram-se as funções de pertinências otimizadas pelo AG para a obtenção do melhor indivíduo com um total de 97% de acertos.

Comparando as duas figuras perecebe-se que na figura 4.5 as funções de pertinência estão mais misturadas e mais condensadas em seus universos de discursos. Contudo, após a atuação do AG, observa-se na figura 4.6 uma melhor distribuição das funções de pertinências em seus universos de discursos. Isso se deve, por que o AG altera a posição da funções de pertinência localizadas através das médias dos parâmetros de entrada, e também aumenta ou diminui a variância de cada uma das funções de pertinências. Essa melhor localização e melhor ajuste das variâncias das funções de pertências objetiva aumentar o número de acertos no processo de reconhecimento.

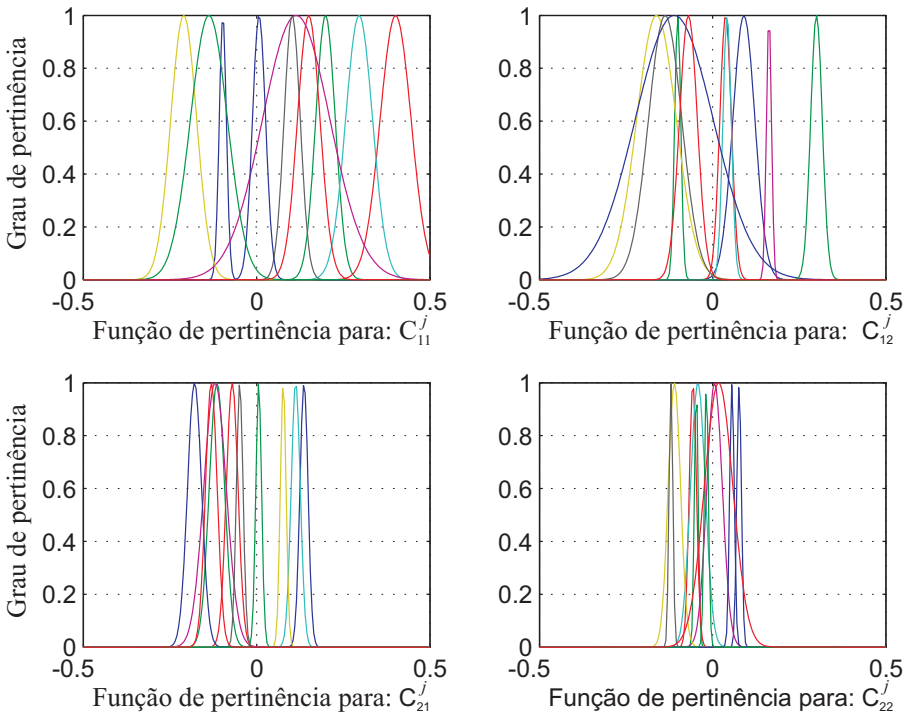


Fig. 4.5: Funções de Pertinências para c_{kn}^j na primeira geração.

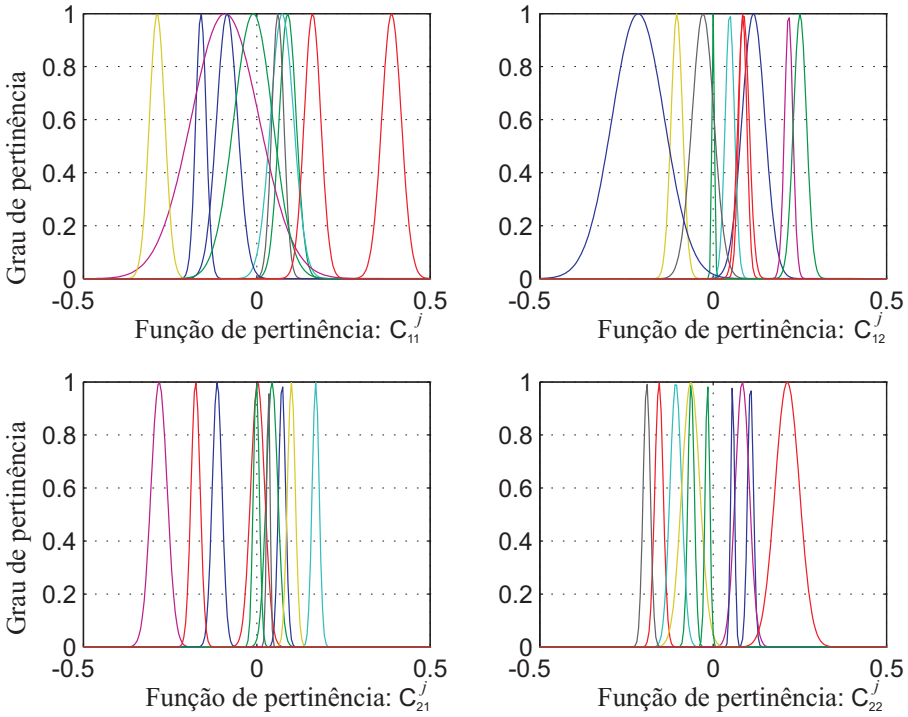


Fig. 4.6: Funções de Pertinências para c_{kn}^j otimizadas pelo AG.

A superfície relacional gerada pela base de regras obtida na equação (3.7) é mostrada na figura 4.7: (a) Superfície relacional de c_{11}^j E c_{12}^j ENTÃO y; (b) Superfície relacional de c_{12}^j E c_{21}^j ENTÃO y; (c) Superfície relacional de c_{21}^j E c_{22}^j ENTÃO y; (d) Superfície relacional de c_{22}^j E c_{11}^j ENTÃO y; (e) Superfície relacional de c_{11}^j E c_{21}^j ENTÃO y; (f) Superfície relacional de c_{12}^j E c_{22}^j ENTÃO y.

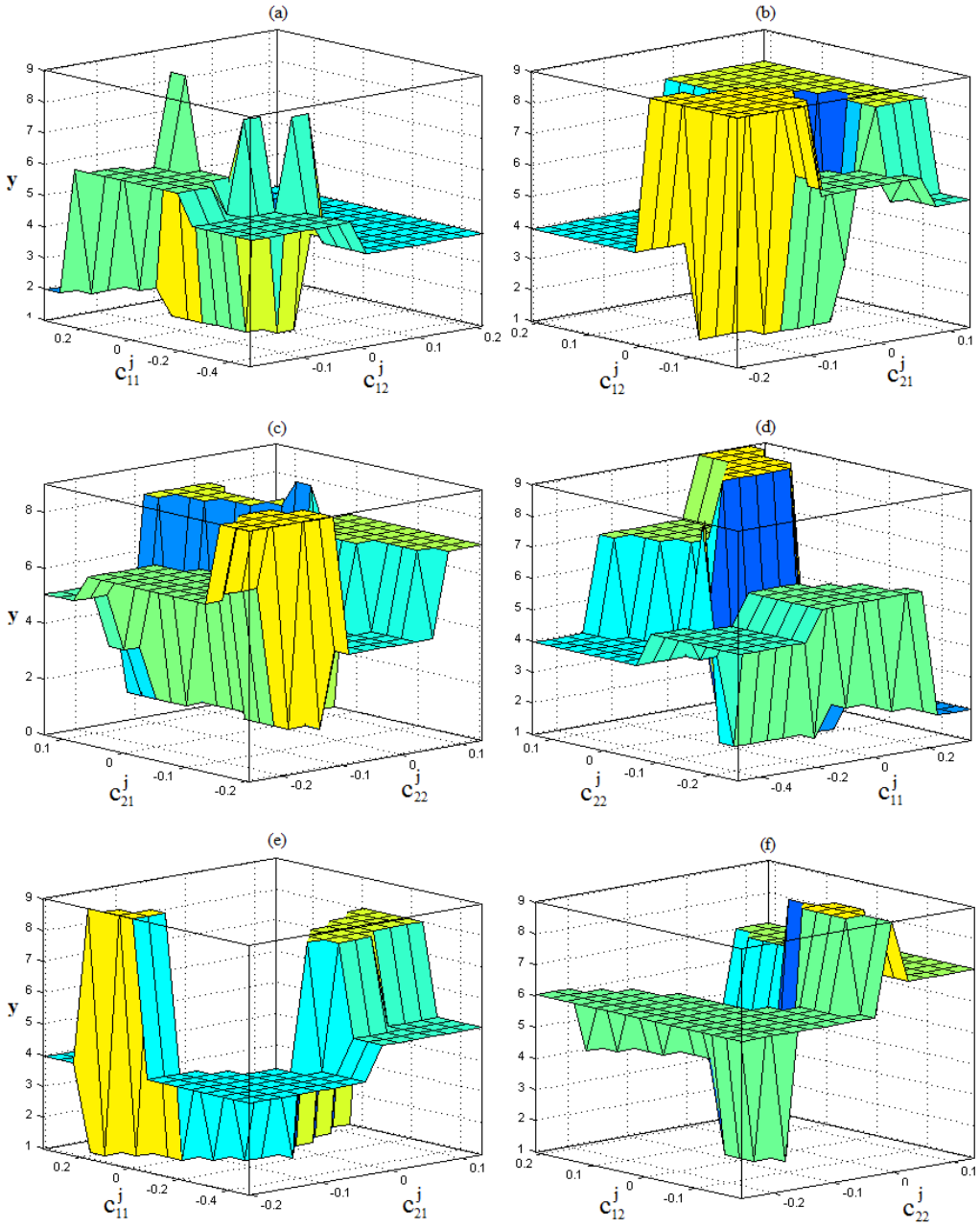


Fig. 4.7: Superfície relacional dada na equação (3.7) otimizada pelo AG.

4.2 Sistema de Teste: Validação

Nesta etapa, foram tomadas 100 locuções do banco da EPUSP, que foram utilizadas para a geração dos padrões, sendo 50 masculinas e 50 femininas. Outras 100 locuções pronunciadas pelos mesmos locutores do banco da EPUSP foram utilizadas nos testes com dependência extrita de locutores. Em seguida, foram tomadas 400 locuções do banco IFMA. Para todas as $m = 10$ observações de cada dígito pronunciado, uma matriz bidimensional C_{kn}, com $k = n = \{2, 3, 4\}$, foi gerada e utilizada para os procedimentos de testes, a saber:

Treinamento: Reconhecimento otimizado pelo AG (5 mulheres e 5 homens- SNR=40dB). Para efeito de análise de desempenho, os testes de validação foram feitos em aproximadamente $32,4\,ms$ para o IMSR e 25 ms para o HMM, para cada dígito pronunciado.

Teste 1: Validação - Reconhecimento estritamente dependente de locutor, em que as palavras usadas na validação foram pronunciadas pelos mesmos locutores que pronunciaram as palavras utilizadas para geração dos padrões (5 mulheres e 5 homens-SNR=40dB).

Teste 2: Validação - Reconhecimento com dependência parcial de locutores, em que dois exemplos das 10 palavras utilizadas no treinamento foram pronunciadas por esse locutor(Locutor feminino-SNR=40dB).

Teste 3: Validação - Reconhecimento com dependência parcial de locutores, em que dois exemplos das 10 palavras utilizadas no treinamento foram pronunciadas por esse locutor(Locutor masculino-SNR=40dB).

Teste 4: Validação - Reconhecimento independente de locutor. O locutor não participou do processo de treinamento(Locutor feminino-SNR=20dB).

Teste 5: Validação - Reconhecimento independente de locutor. O locutor não participou do processo de treinamento(Locutor masculino-SNR=20dB).

Nas figuras 4.8 a 4.13 apresentam-se as análises comparativas do HMM com matriz quadrada de transição de estado de ordem dois, três e quatro, respectivamente, e ainda, com duas, três e quatro componentes gaussianas na mistura, respectivamente, e parametrização com 12 MFCCs. Para efeito de comparação, o método proposto (IMSR) utilizou a matriz bidimensional C_{kn} com ordem $(K = N = 2)$, $(K = N = 3)$ e $(K = N = 4)$. Com os resultados de acertos, obtidos experimentalmente, representando pontos de dados, traçaram-se os gráficos comparativos de desempenho.

Na figura 4.8, na qual se ilustram os resultados obtidos no treinamento, observa-se que, com o IMSR com a matriz bidimensional C_{22}, obteve-se um total de $H = 97\%$ de acertos; com a matriz bidimensional C_{33}, um total de $H = 98\%$, e com a matriz bidimensional C_{44},$H = 100\%$. Os resultados apresentados pelo HMM foram $H = 86\%$, $H = 91\%$ e $H = 95\%$ de acertos, respectivamente.

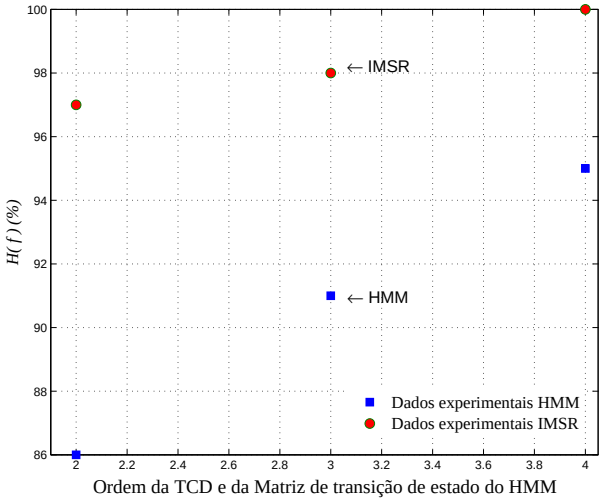


Fig. 4.8: Resultados do Treinamento.

Na figura 4.9 apresentam-se os resultados dos testes extritamente dependente de locutor, tanto para o IMSR quanto para o HMM.

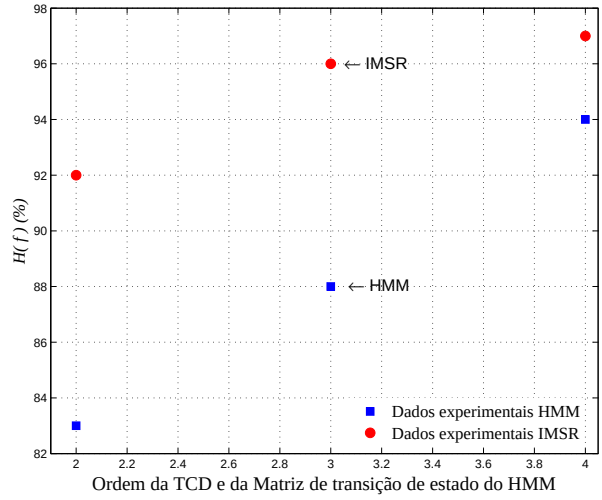


Fig. 4.9: Validação: Teste 1.

Os resultados obtidos nos testes com dependência parcial de locutor, com locutores masculinos e femininos, são ilustrados nas figuras 4.10 e 4.11, respectivamente.

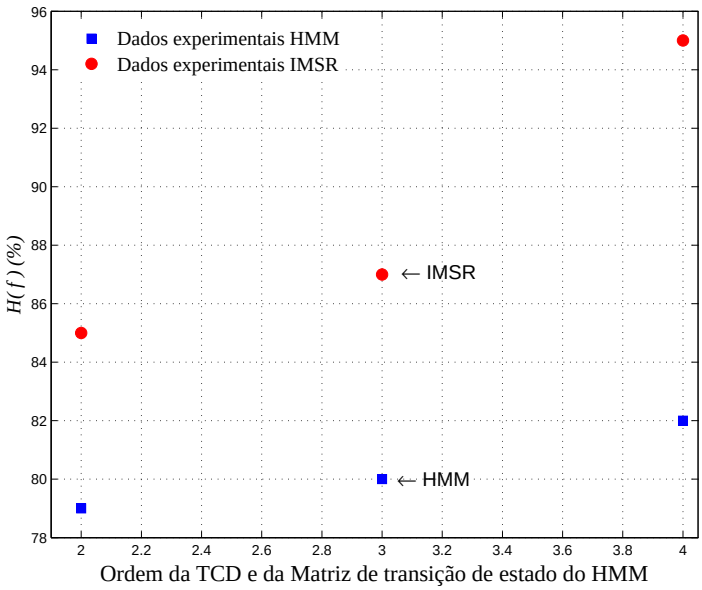


Fig. 4.10: Validação: Teste 2.

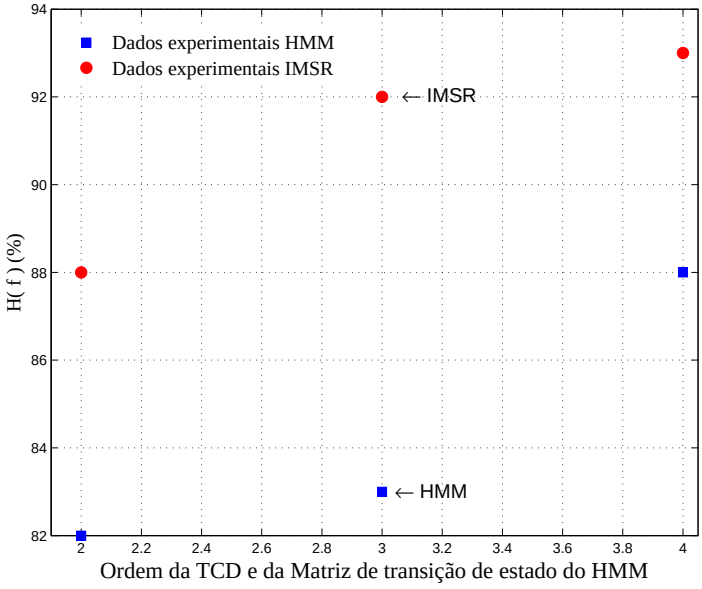


Fig. 4.11: Validação: Teste 3.

Na figuras 4.12 e 4.13 saõ apresentados os resultados alcançados nos testes independentes de locutror, para locutores masculino e feminino, respectivamente.

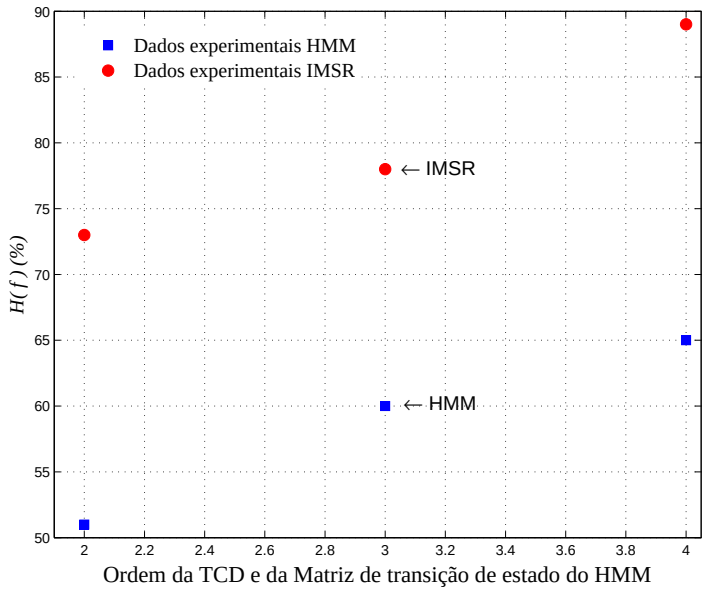


Fig. 4.12: Validação: Teste 4.

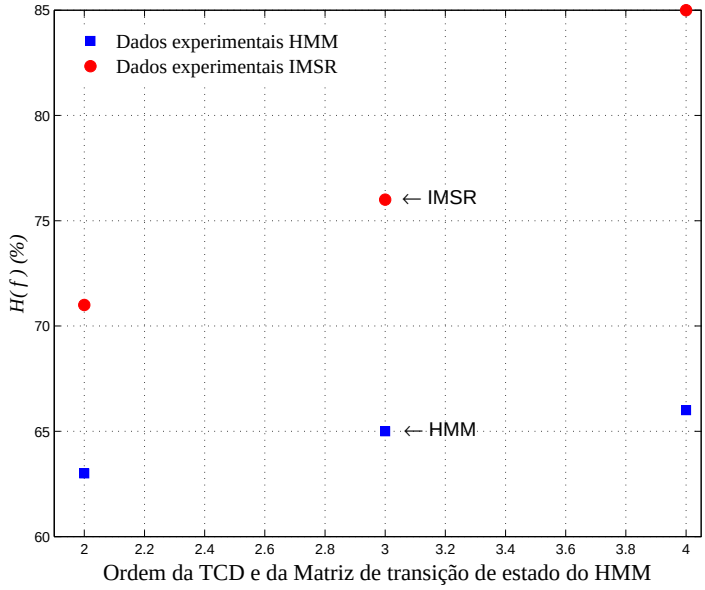


Fig. 4.13: Validação: Teste 5.

Os resultados experimentais dos testes realizados também são apresentados na Tabela (4.2).

Tab. 4.2: Resultados (%): Metodologia Proposta × HMM

		Ordem da Matriz=2	Ordem da Matriz=3	Ordem da Matriz=4
TREINAMENTO	IMSR	97	98	100
	HMM	86	93	95
TESTE 01	IMSR	92	96	97
	HMM	83	88	94
TESTE 02	IMSR	85	87	95
	HMM	79	80	85
TESTE 03	IMSR	88	92	93
	HMM	82	83	88
TESTE 04	IMSR	73	78	89
	HMM	56	60	65
TESTE 05	IMSR	71	76	85
	HMM	63	65	70

Com o objetivo de melhorar o desempenho do HMM, alterou-se os parâmetros do HMM para Teste 4 e Teste 5 da validação. O reconhecimento foi feito de modo independente de locutor. Os parâmetros do HMM utilizado foram: número de estados= 4, número de misturas= 8, quantidade de coefeicientes cepstrais utilizados= 12, obteve-se como resultado: Teste 4, locutor feminino, 91% de acertos; e para o Teste 5, locutor masculino, 89% de acertos.

4.2.1 Comparação com outras metodologias utilizadas em reconhecimento de voz

Para validação estatística do modelo proposto, montou-se um banco de vozes para testes independentes de locutores. Com este objetivo realizaram-se as seguintes etapas

1. Treinamento: Foram selecionados dez (10) locutores: cinco (5) masculinos e cinco (5) femininos do Banco IFMA; cada um pronunciou um exemplo de cada dígito, num total de cem (100) dígitos pronunciados. Foram selecionados também, seis (6) locutores: três (3) masculinos e três (3) femininos do Banco EPUSP; cada um pronunciou um exemplo de cada dígito, num total de sessenta (60) dígitos pronunciados. Selecionaram-se ainda, quatro (4) locutores do Banco INATEL: dois (2) masculinos e dois (2) femininos, num total de quarenta (40) dígitos pronunciados. Dessa forma, o banco de treinamento para a etapa de validação foi composto com ($m = 20$) observações para cada dígito j, num total de 200 observações.

2. Validação: Para a validação, selecionaram-se, dez locutores do sexo masculino e dez locutores do sexo feminino que não partiparam da etapa de treinamento; cada um pronunciando dez (10) exemplos para cada dígito, isto é, cada locutor participou do processo de teste com 100 pronúncias dos dígitos, num total de 1000 locuções masculinas e 1000 locuções femininas. Todos os locutores foram escolhidos do Banco IFMA e não participaram da etapa de treinamento.

3. Os testes também foram realizados com os reconhecedores baseados em SVM e GMM. Para efeito de comparação, os parâmetros de entradas para os três reconhecedores foram os mesmos, isto é, os elementos da matriz C_{kn} para cada padrão j. Para a etapa de treinamento do SVM, foram feitas 100 realizações do algoritmo elaborado e utilizaram-se como modelos as dez melhores máquinas que apresentaram os melhores resultados de reconhecimento. Neste trabalho o reconhecedor baseado no SVM utilizou os *kernel* polinomial de ordem 2 e a *Radial Basis Function*-RBF com $\sigma = 0.03$. Os procedimentos de treino e testes para o SVM e GMM são tratados nos apêndices C e D, respectivamente.

Após a geração das matrizes C_{kn} para as 20 observações de cada dígito, os parâmetros cm_{kn}^{j} e cv_{kn}^{j}, dados na equação (3.8), associados ao valor lingüístico $\tilde{c}_{kn}^{j}$ para o dígito $j\,|^{j=0,1,2,\cdots,9}$, são otimizados pelo AG, tal que a função custo $f\,(j, y^{*})$ seja minimizada. Na figura 4.14 apresenta-se o melhor resultado da otimização da função custo, cujos os valores obtidos foram $f\,(j, y^{*}) = 14,5$, $H(f) = 85,5\%$, com 171 dígitos reconhecidos corretamente e matriz C_{kn} de ordem $K = N = 2$.

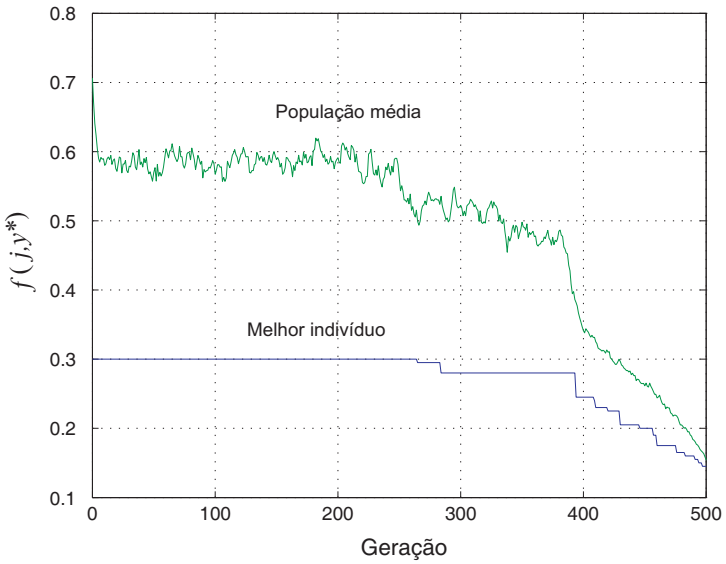


Fig. 4.14: Melhor resultado obtido no processo de treinamento com a matriz C_{kn} de ordem $K = N = 2$.

Ilustra-se na figura 4.15 o resultado da otimização da função custo $f\,(j, y^{\,*})$, onde a ordem da matriz C_{kn} é $K = N = 3$. O resultado obtido foi $f(j, y^*) = 8,5\%$, e o número de acertos $H(f) = 91,5\%$, com 183 dígitos reconhecidos corretamente.

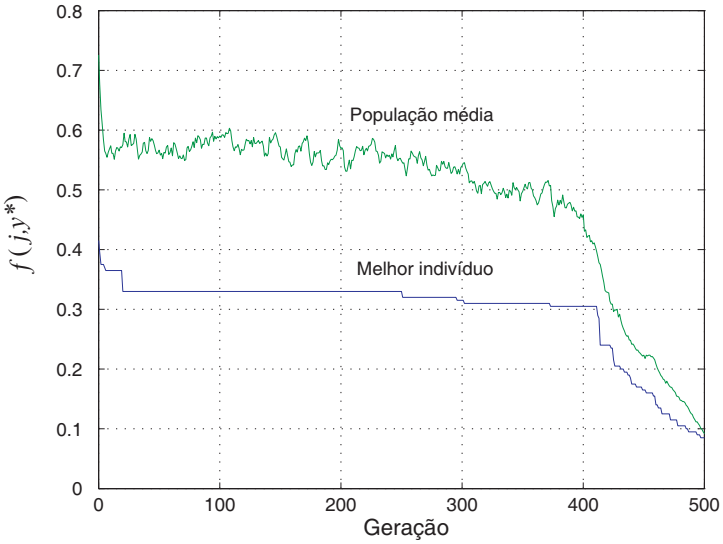


Fig. 4.15: Melhor resultado obtido no processo de treinamento com a matriz C_{kn} de ordem $K = N = 3$.

O resultado para a otimização da função custo $f\,(j, y^{\,*})$, onde a ordem da matriz C_{kn} é $K = N = 4$, é apresentado na figura 4.16. O resultado obtido foi $f(j, y^*) = 4,5\%$, e o número de acertos $H(f) = 95,5\%$, com 190 dígitos reconhecidos corretamente.

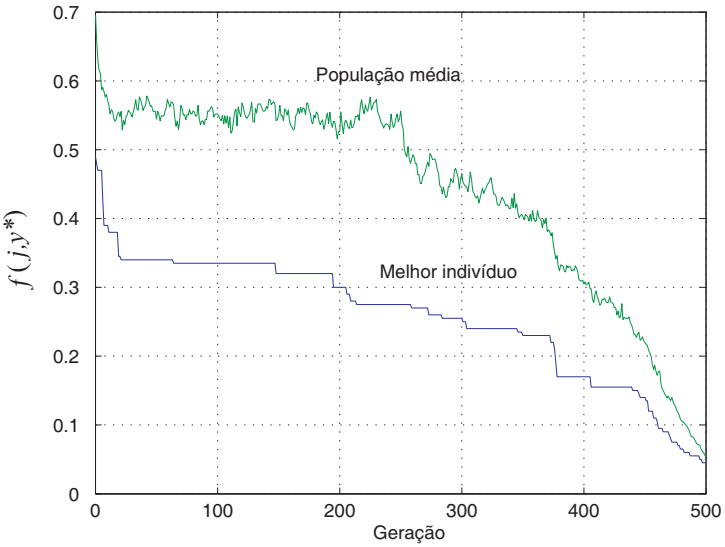


Fig. 4.16: Melhor resultado obtido no processo de treinamento com a matriz C_{kn} de ordem $K = N = 4$.

Na Tabela (4.3) são apresentados os resultados (em porcentagem) dos testes independentes de locutor, obtidos com a metodologia proposta, com o SVM-polinomial de ordem ($p = 2$), com o SVM-RBF com $\sigma = 0,03$ e com o GMM, para locutores masculinos.

Tab. 4.3: Resultados (%): [Metodologia Proposta] × [SVM-Polinomial de ordem $p = 2$] × [SVM-RBF com $\sigma = 0,03$] × [GMM] para locutores masculinos

		Ordem da Matriz=2	Ordem da Matriz=3	Ordem da Matriz=4
TREINAMENTO	IMSR	85,5	91,5	95,5
	SVM-Poli	90	94	98
	SVM-RBF	80	82	90
	GMM	75	85,5	89
Loc_M 01	IMSR	91	95	92
	SVM-Poli	68	58	70
	SVM-RBF	76	78	80
	GMM	82	80	86
Loc_M 02	IMSR	85	94	95
	SVM-Poli	61	70	77
	SVM-RBF	78	80	80
	GMM	83	79	79
Loc_M 03	IMSR	80	88	99
	SVM-Poli	70	66	73
	SVM-RBF	76	80	80
	GMM	57	64	72
Loc_M 04	IMSR	80	96	99
	SVM-Poli	67	63	71
	SVM-RBF	76	63	81
	GMM	80	87	91
Loc_M 05	IMSR	80	79	86
	SVM-Poli	62	63	70
	SVM-RBF	78	80	78
	GMM	52	67	77
Loc_M 06	IMSR	76	73	86
	SVM-Poli	68	63	69
	SVM-RBF	76	80	80
	GMM	66	70	71
Loc_M 07	IMSR	75	93	87
	SVM-Poli	62	66	72
	SVM-RBF	70	80	78
	GMM	62	71	78
Loc_M 08	IMSR	73	82	94
	SVM-Poli	70	66	74
	SVM-RBF	78	78	80
	GMM	62	83	85
Loc_M 09	IMSR	72	93	87
	SVM-Poli	66	72	75
	SVM-RBF	76	80	80
	GMM	90	84	85
Loc_M 10	IMSR	71	80	99
	SVM-Poli	66	66	74
	SVM-RBF	76	80	82
	GMM	72	74	86

Na figura 4.17 ilustram-se os resultados apresentados na Tabela 4.3 com a matriz C_{kn} de ordem $K = N = 2$.

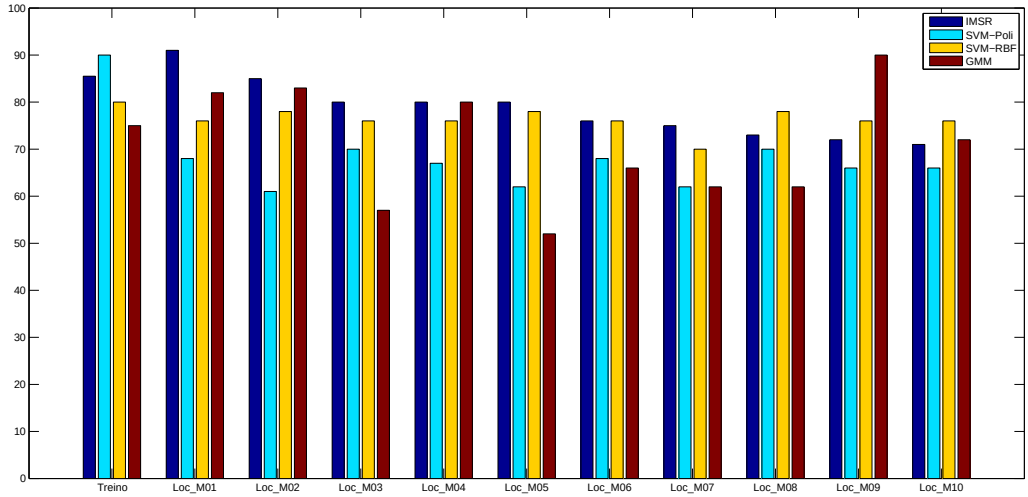


Fig. 4.17: Resultados da validação para locutores masculinos com C_{kn} de ordem $K = N = 2$.

Apresentam-se, na figura 4.18, os resultados da Tabela 4.3 com a matriz C_{kn} de ordem $K = N = 3$.

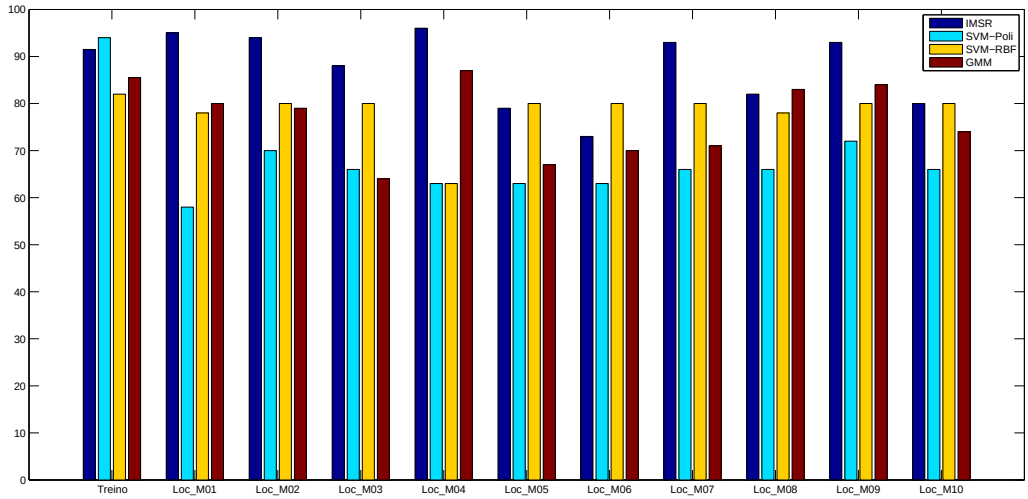


Fig. 4.18: Resultados da validação para locutores masculinos com C_{kn} de ordem $K = N = 3$.

Na figura 4.19 são apresentados os resultados da Tabela 4.3 com a matriz C_{kn} de ordem $K = N = 4$.

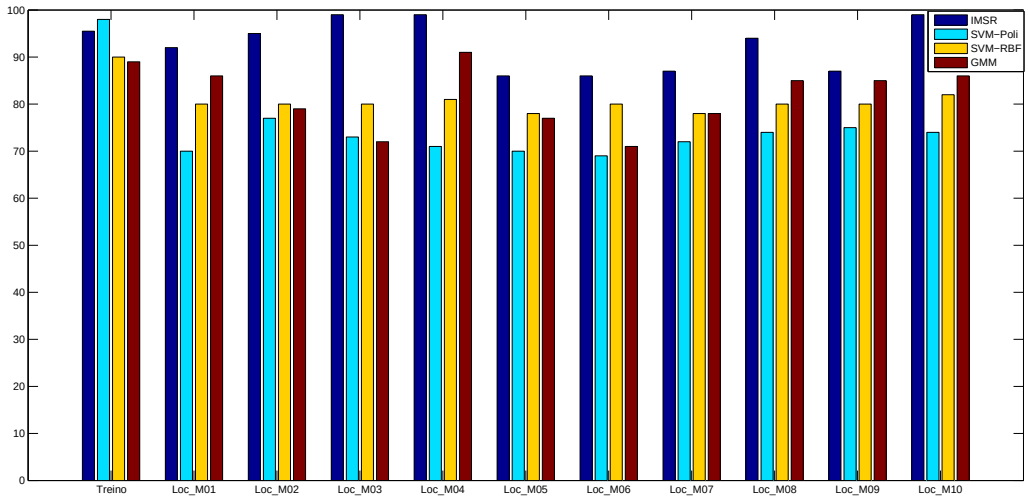


Fig. 4.19: Resultados da validação para locutores masculinos com C_{kn} de ordem $K = N = 4$.

Na Tabela (4.4) são apresentados os resultados (em porcentagem) dos testes independentes de locutor, obtidos com a metodologia proposta, com o SVM-polinomial de ordem ($p = 2$), com o SVM-RBF com $\sigma = 0,03$ e com o GMM, para locutores femininos.

Tab. 4.4: Resultados (%): [Metodologia Proposta] × [SVM-Polinomial de ordem $p = 2$] × [SVM-RBF com $\sigma = 0,03$] × [GMM] para locutores femininos

		Ordem da Matriz=2	Ordem da Matriz=3	Ordem da Matriz=4
TREINAMENTO	IMSR	85,5	91,5	95,5
	SVM-Poli	90	94	98
	SVM-RBF	80	82	90
	GMM	75	85,5	89
Loc_F 01	IMSR	90	98	98
	SVM-Poli	68	62	65
	SVM-RBF	74	76	78
	GMM	92	84	88
Loc_F 02	IMSR	88	94	99
	SVM-Poli	65	65	66
	SVM-RBF	80	80	80
	GMM	94	89	82
Loc_F 03	IMSR	86	94	94
	SVM-Poli	60	60	77
	SVM-RBF	78	78	80
	GMM	88	88	95
Loc_F 04	IMSR	82	88	87
	SVM-Poli	67	68	75
	SVM-RBF	78	80	82
	GMM	80	78	89
Loc_F 05	IMSR	79	96	97
	SVM-Poli	66	67	64
	SVM-RBF	80	80	80
	GMM	82	78	90
Loc_F 06	IMSR	77	82	75
	SVM-Poli	66	72	72
	SVM-RBF	78	72	82
	GMM	70	72	83
Loc_F 07	IMSR	74	84	81
	SVM-Poli	63	64	62
	SVM-RBF	74	80	80
	GMM	68	66	74
Loc_F 08	IMSR	70	83	82
	SVM-Poli	54	64	72
	SVM-RBF	72	74	78
	GMM	75	65	71
Loc_F 09	IMSR	66	92	83
	SVM-Poli	64	68	75
	SVM-RBF	72	72	78
	GMM	78	77	81
Loc_F 10	IMSR	65	82	93
	SVM-Poli	76	70	80
	SVM-RBF	67	80	82
	GMM	59	74	92

Apresenta-se na figura 4.20 os resultados da Tabela 4.4, em gráficos de barras, com a matriz C_{kn} de ordem $K = N = 2$.

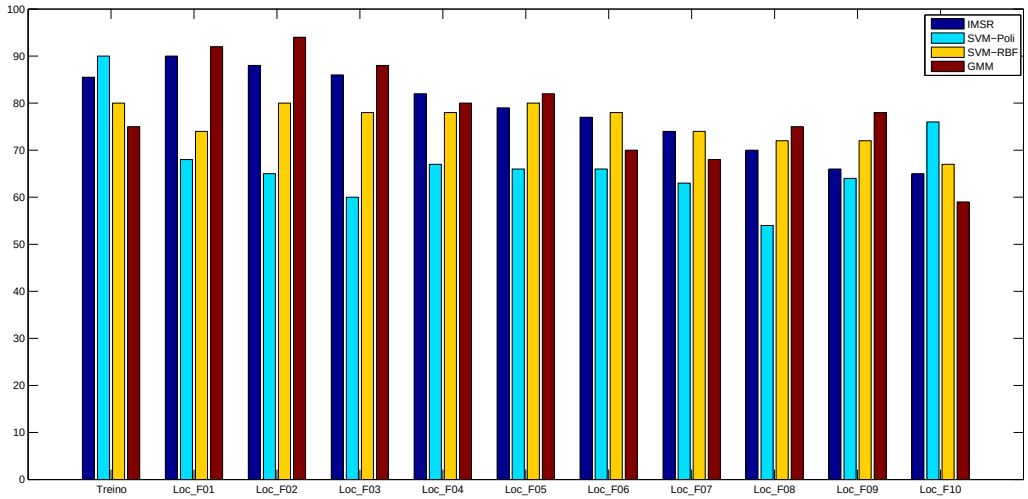


Fig. 4.20: Resultados da validação para locutores feminino com C_{kn} de ordem $K = N = 2$.

Na figura 4.21 ilustram-se os resultados da Tabela 4.4 com a matriz C_{kn} de ordem $K = N = 3$.

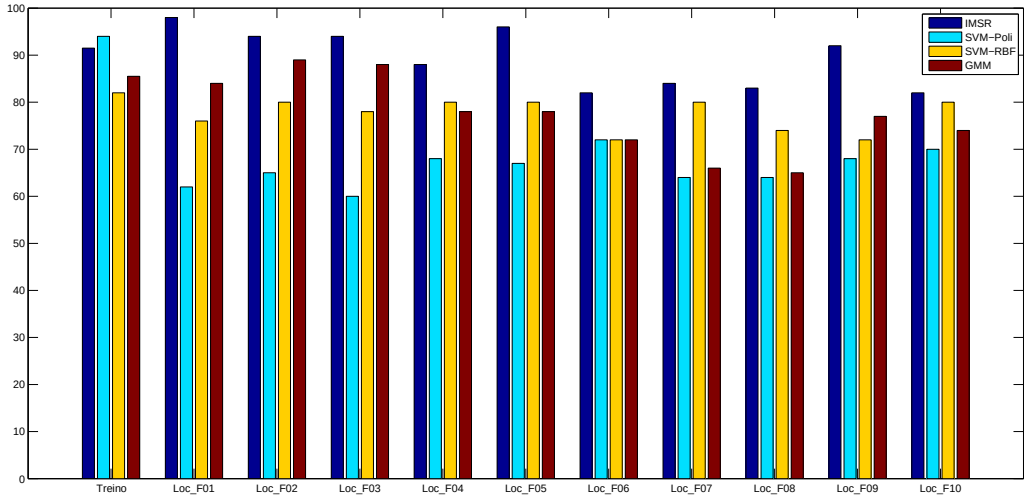


Fig. 4.21: Resultados da validação para locutores feminino com C_{kn} de ordem $K = N = 3$.

Na figura 4.22 ilustram-se os resultados da Tabela 4.4 com a matriz C_{kn} de ordem $K = N = 4$.

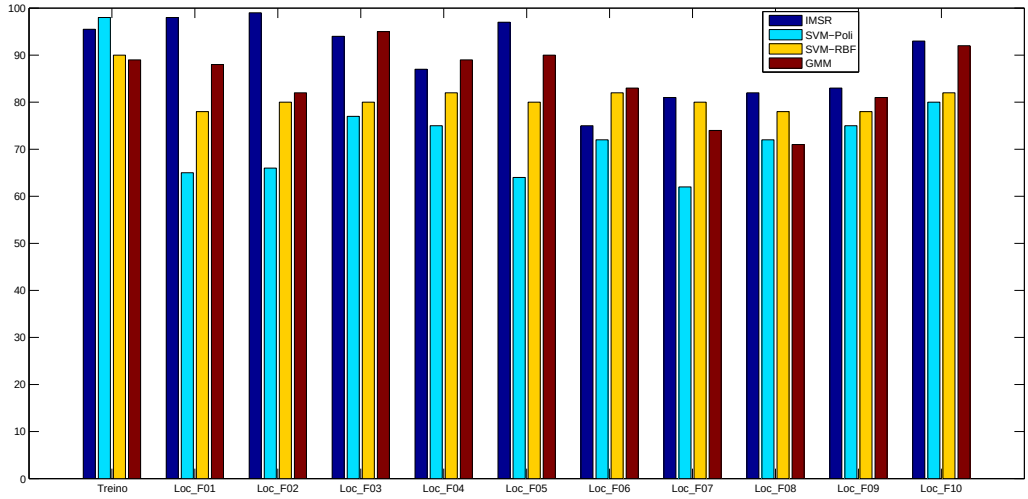


Fig. 4.22: Resultados da validação para locutores feminino com C_{kn} de ordem $K = N = 4$.

4.2.2 Análise dos dados experimentais

Após a realização dos experimentos propostos e análises dos resultados obitidos apresentados neste trabalho, ressaltam-se as seguintes considerações:

1. Inicialmente, para escolha das técnicas de extração de características analisaram-se aquelas mais largamente utilizadas na parametrização do sinal de voz, os MFCC e os coeficientes LPC. Os melhores resultados preliminares foram obtidos com a técnica MFCC. Além disso, optou-se pela parametrização bidimensional utilizando a TCD devido aos excelentes resultados apresentados em trabalhos de reconhecimento de voz com conjunto reduzido de informações. Fez-se uma comparação da capacidade de compactação da informação com sinal de voz entre a transformada de Fourier discreta e a transformada cosseno discreta, em que a última apresentou melhor desempenho. Os resultados desta comparação foram publicados no artigo (Lima et al., 2012).

2. O desempenho dos sistemas de proposições foram analisados também, e os melhores resultados obtidos foram aqueles provenientes da operação produto, em comparação com o mínimo. A implicação utilizada no sistema de inferência nebuloso foi a Mamdani, pois esta obteve melhor desempenho quando comparada com a Dienes-Rescher e Luckasiewicz (Silva, 2012a).

3. Para a otimização, optou-se por algoritmos bio-inspirados devido a características do problema de otimização apresentado. A função custo com um número elevado de variá-

veis a serem otimizadas, e os valores das variáveis, bem como suas restrições de variação muito pequenas. Após comparação entre as técnicas tentativa e erro, *Particle Swarm Optimization-PSO*, *Chaotic-PSO* (Abelardo et al., 2014) e algoritmo genético, optou-se pelo AG, devido ao seu melhor desempenho. Utilizaram-se como metodologias de cruzamento e de mutação para o AG, o cruzamento uniformemente aleatório e o cruzamento aritmético, sendo obtidos os melhores resultados com o cruzamento aritmético.

Incialmente, para a mutação, optou-se pela mutação uniformemente aleatória. Todavia, observou-se que o algoritmo chegava rapidamente a um mínimo local e não conseguia sair desse ponto. Os resultados do uso da mutação uniformemente aleatório foram apresentados nos trabalhos (Silva, 2012b, 2014a). Diante desse resultado, e pela quantidade de genes e pelos seus valores serem muito pequenos, optou-se também pela mutação não uniforme, após as iterações estarem na metade das gerações, uma vez que a medida que o número de gerações aumenta, os valores dos genes são alterados em uma faixa cada vez mais estreita. Isso permitiu ao AG sair do mínimo local alcançado quando da mutação aletaória.

4. Através dos resultados obtidos, percebe-se que, para um número reduzido de parâmetros codificados, o desempenho do reconhecedor proposto é melhor quando da participação dos locutores no processo de treinamento, demonstrando assim uma certa dependência de locutor. Contudo, com o aumento desses parâmetros, o sistema fica menos dependente da participação do locutor no processo de treinamento. Os resultados preliminares desta tese foram publicados no artigo (Silva, 2014b).

Capítulo 5

Considerações Finais

Há uma gama muito grande de metodologias para reconhecimento de voz, a maior parte utiliza o HMM, técnica amplamente difundida e eficiente para resolução do problema de reconhecimento de voz. Mais recentemente, técnicas híbridas utilizando classificadores não lineares, tais como as redes neurais e SVM, foram aplicadas aos reconhecedores baseados em HMM com a finalidade de melhorar o desempenho de classificação. Além disso, técnicas baseadas em inteligência computacional também têm encontrado espaço na solução do problema de reconhecimento de voz, devido às suas capacidades de classificação de padrões complexos, como é o caso da voz. Contudo, apesar do poder de classificação, os classificadores dependem muito dos tipos de codificadores utilizados para a geração dos parâmetros que representam os modelos a serem classificados. Isso é recorrente não só em reconhecimento de voz, como em todos os tipos de padrões. A voz em particular, devido a suas particularidades, tais como variações no tempo, tanto em relação ao tempo de fala como em relação às características intrínsecas a voz, é um padrão de díficil parametrização. Contudo, técnicas de parametrização eficientes, tais como os coeficientes de predição linear, ou, ainda, os coefientes mel-cepstrais, têm sido utilizadas ao longo de décadas na parametrização do sinal de voz com a finalidade de reconhecimento.

Neste trabalho propôs-se uma metodologia inteligente para reconhecimento de voz usando coeficientes mel-cepstrais e a transformada cosseno discreta para gerar uma matriz bidimensional contendo as características temporais local e global de cada padrão a ser reconhecido. O reconhecedor nebuloso proposto foi otimizado por algoritmo genético para maximizar a quantidade de acertos com um número reduzido de parâmetros no processo de reconhecimento. A abordagem nebulosa utilizada é conceitualmente intuitiva devido à facilidade de interpretação, quando se observam as inferições linguísticas utilizadas na composição da base de regras para geração do modelo otimizado. A metodologia nebulosa utilizada pode ser vista como uma abordagem local, uma vez que ela particiona o domínio do sistema de reconhecimento em um

certo número de regiões nebulosas, associadas ao espaço de entradas para definição da saída do modelo. A natureza exata dessas regiões, bem como o modo como elas são combinadas, dependem do tipo de regra e do mecanismo de inferência envolvidos. Os modelos nebulosos foram obtidos a partir do conhecimento do especialista, fato observado na elaboração da base de regras.

5.1 Conclusões

Nesta tese, atenção especial tem sido dada à modelagem adequada dos padrões para utilização na etapa de reconhecimento, usando modelos nebulosos de inferência Mamdani, cujas entradas foram obtidas com uma codificação eficiente do sinal de voz. Entre as várias estruturas apresentadas, e que poderiam ser usadas tanto na modelagem dos parâmetros de entrada como na elaboração do sistema de inferência, fez-se opção por aquelas que se mostraram mais eficientes, fato verificado através de experimentos. Com este objetivo em mente, os assuntos essenciais relacionados aos passos que podem ser desenvolvidos no procedimento de reconhecimento de voz, utilizando sistemas nebulosos, desde o modelamento eficiente dos padrões de entrada do modelo, até o procedimento de validação do modelo e a utilização de modelos nebulosos, foram apresentados para ilustrar a teoria matemática envolvida. Os conhecimentos do modelamento do sinal de voz através de parâmetros estatísticos de média e variância e da estruturação de inferências nebulosas foi o ponto de partida para a proposta e para o desenvolvimento do algoritmo nebuloso. Além disso, observou-se que os parâmetros estatísticos obtidos no início do modelamento não apresentavam os melhores resultados no processo de reconhecimento. Esta limitação nos modelos motivou à utilização de técnicas de otimização baseada em algoritmos bio-inspirados, que melhor caracterizavam os modelos dos padrões a serem utilizados no reconhecimento. Nesse contexto, uma vez definida a forma dos parâmetros de entrada do sistema de inferência nebuloso, elaborou-se a base de regras mais adequada aos modelos estatísticos obtidos. Problemas de convergência em mínimos locais motivou a utilização de técnicas, dentro do algoritimo de otimização, para fugas destes mínimos para possibilitar o melhor desempenho. As propostas de técnicas de otimização, bem como suas alterações, foram discutidas nos resultados experimentais apresentados. Dentre as principais conclusões, pode-se destacar:

- A eficiência da parametrização bidimensional através de MFCCs e TCD no modelamento das variações locais e globais do sinal de voz;

- O particionamento dos domínios e o conhecimento do especilista na composição da base de regra e no modelo nebuloso, além de simplificar a forma de modelamento dos padrões

para fins de reconhecimento, eliminam a maldição da dimensionalidade para o sistema inteligente proposto;

- O conjunto de características que compõem a metodologia proposta permite obter um bom desempenho no processo de reconhecimento;
- O tempo de reconhecimento é apropriado para aplicações práticas.

A busca por sistema de reconhecimento de voz tem aumentado consideravelmente, não só pelo avanço da tecnologia, principalmente ná área de comunicação e entretenimento, mas também devido às necessidades tanto discutidas na atualidade, que dizem respeito à inclusão social de pessoas limitadas fisicamente ou por limitações congênitas, ou por limitações adquiridas. Os sistemas de parametrização têm sido uma constante ao longo de décadas; contudo, os classificadores utilizados com objetivo de reconhecimento de voz apresentados na literatura especilizada baseiam-se, geralmente em HMM com suas variações e técnicas de melhoramento de desempenho. Ao longo das últimas duas décadas, as redes neurais também têm sido muito utilizadas com este propósito; além disso, mais recentemente técnicas que utilizam as máquinas de vetor de suporte também têm sido uma forte proposta para esse tipo de reconhecimento. Os sistemas nebulosos também vêm obtendo êxito em sistema de reconhecimento de voz. Baseado no desempenho de sistemas nebulosos em identificação de sistemas, bem como em classificações de padrões, apresentou-se uma metodologia híbrida inteligente para reconhecimento de voz. Uma base de regras foi elaborada a partir dos parâmetros estatísticos obtidos e do conhecimento do especialista, baseada em sistema nebuloso de inferência Mamdani. Para fins de aplicação da metodologia proposta, foram realizados testes em três bancos de vozes, bem como na composição de outros bancos a partir dos três bancos originais. Para análise de desempenho, a metodologia proposta foi comparada a quatro técnicas utilizadas em reconhecimento de voz. Entre as principais conclusões sobre a metodologia híbrida inteligente para reconhecimento de voz, podemos ressaltar que ela:

- Permite obtrer um bom desempenho mesmo utilizando uma quantidade reduzida de parâmetros para os modelos;
- Trabalha com um sistema inteligente de ajuste de parâmetros para aumento de desempenho do reconhecimento;
- Apresenta boa capacidade de generalização;
- Aprensenta tempo de treinamento médio de $24h$; contudo, o tempo de teste foi da ordem de $32ms$, coerente com as técnicas já consolidadas na literatura especializada.

5.2 Propostas futuras

Algumas possíveis extensões e problemas ainda em aberto associados à metodologia de reconhecimento de voz com o sistema inteligente híbrido proposto nesta tese são apresentadas como seguem:

1. Robustez ao ruído, assunto muito importante quando da aplicação em reconhecimento de sinais práticos, em ambientes variados. A aplicação de técnicas de cancelamento ou de diminuição de ruído, levando-se em conta que o melhor desempenho do reconhecedor é de grande importância;

2. Geração de um banco de voz com uma quantidade razoável de sotaques diferentes para os diversos grupos de palavras a serem reconhecidas, uma vez que esta caractarística mostrou-se importante no desempenho da metodologia proposta, pois quanto maior a variablidade da forma apresentada de uma mesma palavra, pior foi o desempenho observado;

3. Outro passo importante sugerido é que, para o aumento de vocabulário, utilize-se clusterização para se dividir as palavras em classes antes do reconhecimento. Recomenda-se, para isso, o uso de sistemas nebulosos hierárquicos e/ou sistemas fuzzy tipo-2 que trabalham com incertezas de incertezas;

4. Além disso, percebeu-se ao longo dos testes que a metodologia apresentada tem potencial para aplicação em reconhecimento de voz contínua; dessa forma, sugere-se também aplicação de análise em sistema em tempo real, para aplicação em reconhecimento contínuo de voz;

5. Desenvolvimento em hardware com processador digital de sinais.

Referências Bibliográficas

Abelardo, A., Silva, W. and Serra, G. (2014). CPSO Applied in the optimization of a speech recognition system, *Intelligent data engineering and automated learning-IDEAL.* pp. 134–141.

Abushariah, A., Gunawan, T., Khalifa, O. and Abushariah, M. (2010). English digits speech recognition system based on hidden Markov models, *International Conference on Computer and Communication Engineer-ICCCE.* pp. 1–5.

Aggarwal, R.K. ; Dave, M. (2011). Application of genetically optimized neural networks for Hind speech recognition system, *Word Congress on Informatic and Communication Technologies.* pp. 512–517.

Ahmed, T.N.N. ; Rao, K. (1974). Discrete cosine transform, *IEEE Transaction on Computers.* **C-24**(1): 90–93.

Alam, J., Kinnunem, T., Kenny, P., Ouellet, P. and O'Shaughnessy (2012). Multitaper MFCC and PLP features for speaker verification using i-vectors, *Elsevier- Speech Communication.* **vol.55**: 237–251.

Alencar, V. F. S.; Alcaim, A. (2008). LSF and LPC - Derived features for large vocabulary distributed continuous speech recognition in brazilian portuguese, *Conference on Signals, Systems and Computers.* pp. 1237–1241.

Ariki, Y., Mizuta, S., Nagata, M. and Sakai, T. (1989). Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum, *IEEE Proceedings on Communications, Speech and Vision.* **vol.136**(2): 133–140.

Azam, S., Mansor, Z., Mughal, M. and Moshin, S. (2007). Urdu persian digit recognition using a hybrid HMM-SVM, *International Symposium on Intelligent Signal Processing and Communications Systems* .

Azar, M.Y. ; Razzazi, F. (2008). A dct based nonlinear predictive coding for feature extraction in speech recognition systems, *IEEE International Conference on Computational Intelligente for Measurement Systems and Applications.* pp. 19–22.

Bazaraa, M. S., Sherali, H. D. and Shetty, C. M. (1993). *Nonlinear programming-theory and algorithms*, 2 edn, John Wiley & Sons, Inc.

Becchetti, C. ;Ricotti, L. P. (2000). *Speech recognition theory and C++ implementation*, 1 edn, Wiley.

Bechara, E. (2009). *Moderna gramática portuguesa*, 37 edn, Nova Fronteira.

Bellman, R. (1961). *Adaptive control process: A guide tour*, Princeton University Press.

Bishop, C. M. (2007). *Pattern recognition and machine learning*, Springer.

Bogert, B. P., Healy, M. J. R. and Tukey, J. W. (1963). The quefrency analysis of time series for echoes: Cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking, *Proceedings of the Symposium on Time Series Analysis.* (15): 209–243.

Bresolin, A. A. (2008). *Tese de Doutorado: Reconhecimento de voz através de unidades menores do que a palavra, utilizando Wavelet Packet e SVM, em uma nova estrutura hierárquica de decisão*, Universidade Federal do Rio Grande do Norte.

Bresolin, A. A., Doria Neto, A. D. and Alsina, P. J. (2008). Digit Recognition using wavelet and SVM in brazilian portuguese, *IEEE - International Conference on Acoustics, Speech and Signal Processing.* pp. 1545–1548.

Cai, K. (2002). Robustness of fuzzy reasoning and delta-equalities of fuzzy sets, *IEEE Transactions on Fuzzy Systems.* **vol.09**: 738–750.

Campbell, J. (1997). Speaker recognition: A tutorial, *Proceedings of the IEEE.* (9): 1437–1462.

Cegalla, D. P. (2008). *Novíssima gramática da Língua portugesa*, 48 edn, Companhia Editora Nacional.

Chow, S.C. ; Ka-Leung, H. (1992). Fast algorithms for computing the discrete cosine transforms, *IEEE Transactions on circuits and systems-II: Analog and digital signal processing.* **vol.39**(3): 185–190.

Coppin, B. (2004). *Inteligência artificial*, LTC.

Cox, E. (1999). *The fuzzy systems handbook: A practitioner's guide to building, using and maintaining fuzzy systems*, Morgan Kaufmann.

Davis, S.B.; Mermelstein, P. (1980). Comparison of parametric representation for monosylabic word recognition in continuosly spoken sentences, *IEEE Transactions on acustics, speech and signal processing.* **vol.28**(4): 357–366.

De Gang, C., Heng, Y. and Tsang, E. (2008). Generalized Mercer theorem and its application to feature space related to indefinite kernels, *International Conference Machine Learning and Cybernetics.* **vol.02**: 774–777.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statiscal Society.* **vol.39**: 1–38.

Deng, J., Bouchard, M. and Yeap, T. (2008). Feature enhancement for noisy speech recognition with a time-variant linear predictive HMM structure, *IEEE Transactions on Audio, Speech, and Language Processing.* **vol.16**(5): 891–899.

Dias, S. (2012). *Master Thesis: Estimation of the glottal pulse from speech or singing voice*, School of Engineering of the University of Porto.

Ding, C.H ; Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks, *Oxford University Press.* **vol.17**(4): 349–358.

Dubois, D., Fargier, H. and Prade, H. (1997). Beyond min aggregation in multicriteria decision:(ordered) weighted min, discri-min, leximin, *Springer.* pp. 181–192.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological sequence analysis: Probabilistc models of proteins and nucleic acids*, Cambridge University Press.

Effros, M., Feng, H. and Zeger, K. (2004). Suboptimality of the Karhunen Loeve transform for transform coding, *IEEE Transactions on Information Theory.* **vol.50**: 1605–1619.

Elloumi, S., Jaam, J., Hasnah, A., Jaoua, A. and Nafkha, I. (2004). A multi-level conceptual data reduction approach based on the Lukasiewicz implication, *Eselvier-Information Sciences.* **vol.163**: 253–262.

Engelbrecht, A. P. (2003). *Computational intteligence: An introduction*, John Wiley & Sons.

Fant, G. (1960). *Acoustic theory of speech production*, Mouton, The Hague.

Fant, G. (1981). The source filter concept in voice production, **vol.22**(1): 21–37.

Fant, G. (2004). *Speech acoustics and phonetics*, 1 edn, Kluwer Academic Publishers.

Ferguson, J. (1980). Variable duration models for speech, *Proc. Symposium on the Application of Hidden Markov Models to Text and Speech.* pp. 143–179.

Ferguson, T. (1983). *Bayesian density estimation by mixtures of normal distribution*, Academic Press.

Ferreira, A. B. H. (2010). *Dicionário Aurélio da Língua portuguesa*, 5 edn, Positivo.

Filho, S., Drew-Jr, P. and Marcolino, L. (2013). Mistura de Gaussianas: uma abordagem rápida para modelar nuvens de pontos, *Simpósio Brasileiro de Automação Inteligente* .

Fink, G. (2014). *Markov models for pattern recognition: From theory to applications*, 2 edn, Springer.

Fissore, P.L.L.; Rivera, E. (1997). Using word temporal structure in HMM Speech recongnition, *IEEE International Conference on Acoustics, Speech, and Signal Processing.* **vol.02**: 975–978.

Fogel, B. (2006). *Evolutionary Computation:Toward a New Philosophy of Machine Intelligence*, 3 edn, John Wiley and Sons.

Fu, K. (1968). *Sequential methods in pattern recognition and machine learning*, Acadmic Press., New York.

Gales, M. (2007). Discriminative models for speech recognition, *IEEE Information Theory and Applications Workshop.* pp. 170–176.

Ganesh, C., Kumar, H. and Vanathi, P. (2012). Performance analysis of hybrid robust automatic speech recognition, *IEEE International Conference on Signal Processing, Computing and Control.* pp. 1–4.

Gang, C. (2010). Discussion of approximation properties of minimum inference fuzzy system, *Proceedings of the 29th Chinese Control Conference.* pp. 2540–2546.

Garcia, V., Sanchez, J. and Mollineda, R. (2007). An empirical study of the behavior of class classification on inbalanced and overlapped data sets, *Spring-Verlag.* **vol.4756**.

Grimm, M.; Kroschel, K. (2007). *Robust speech recognition and understanding*, 1 edn, Tech education and publishing.

Hanchate, D., Nalawade, M., Pawar, M., Pohale, V. and Maurya, P. (2010). Vocal digit recognition using artificial neural network, *2nd International Conference on Computer Engineering and Technology* **vol.06**: 88–91.

Hassanzadeh, T., Faez, K. and Syfi, G. (2012). A speech recognition system based on structure equivalent fuzzy neural network trained by firefly algoritm., *International Conference on Biomedical Engineering.* pp. 63–67.

Haupt, R.L. ; Haupt, S. (2004). *Pratical genetic algorithms*, John Wiley & Sons, Inc.

Haykin, S. (2002). *Adaptive filter theory*, 4 edn, Pearson.

Haykin, S. (2009). *Neural networks and learning machines*, 3 edn, Pearson Prentice Hall.

Hearst, M., Dumais, S., Osman, E., Platt, J. and Scholkopf, B. (1998). Support vector machine, *IEEE Intelligent Systems.* **vol.13**(4): 18–28.

Hejazi, S., Kazemi, R. and Ghaemmaghami, S. (2008). Isolated persian digit recognition using a hybrid HMM-SVM, *International Symposium on Intelligent Signal Processing and Communication Systems.* pp. 1–4.

Holland, J. (1975). *Adaption in natura and artificial systems*, The University of Michigan Press, Ann Arbor.

Hou, H. S. (1987). A fast recursive algorithm for computing the discrete cosine transform, *IEEE Transactions on acustics, speech and signal processing.* **35**(10): 1455–1461.

Hua, Y. ; Liu, W. (1998). Generalized Karhunen Loeve transform, *IEEE Signal Processing Letters.* **vol.05**(6): 141–142.

Huang, X., Ariki, Y. and Jack, M. A. (1990). *Hidden Markov models for speech recognition*, 1 edn, Edinburgh, Edinburgh University Press.

Jain, A. K. (1979). A sinusoidal family of unitary transforms, *Transactions on pattern analysis and machine intelligence.* **vol.01**(4): 356–365.

Javkin, H., Barroso, N. and Maddieson, I. (1987). Digital inverse filtering for linguistic research, **vol.30**: 122–129.

Keshet, J.; Bengio, S. (2009). *Automatic speech and speaker recognition:Large margin and kernel methods*, 1 edn, John Wiley and Sons Ltd.

Kitamura, T., Nishioka, K., A., I. and Hayahara, E. (1991). Speaker-dependent 100 word recognition using dynamic spectral features of speech and neural networks., *IEEE-Proceedings of the 34th Midwest Symposium on Circuits and Systems.* **1**: 83–86.

Lathi, B. P. (1998). *Modern digital and analog communication system*, 3 edn, New York, Oxford University press.

Ledermann, W.; Vajda, S. (1982). *Handbook of applicabel mathematics*, 1 edn, Jonh Wiley & Sons.

Levinson, S. (1986). Continuously variable duration hidden Markov models for automatic speech recognition, *Compute, Speech and Language.* **vol.1**(1): 29–45.

Lima, P., Silva, W. and Serra, G. (2012). Análise comparativa entre as transformada de Fourier discreta e a transformada cosseno discreta na compressão e recuperação espectral de sinais de voz., *XXX-Simpósio brasileiro de telecomunicações.* pp. 1545–1548.

Linkai, B.U. ; Chiueh, T. (2000). Perceptual speech processing and phonetic feature mapping for robust vowel recognition, *IEEE Transactions on speech and audio processing.* **vol.08**(2): 105–114.

Mamdani, E. (1977). Application of fuzzy logic to approximate reasoning using linguistic synthesis, *IEEE Transactions on Computers.* **C-26**: 1182–1191.

Markel, J.; Gray, A. H. J. (1976). *Linear predction of speech*, Spring-Verlag.

Martuci, S. (1994). Symetric convolution and the discrete sine and cosine tranforms, *IEEE Transactions on signal processing.* **vol.42**(5): 1038–1051.

Mas, M., Monserrat, M., Torrens, J. and Trillas, E. (2007). A survey on fuzzy implication functions, *IEEE Transactions on Fuzzy Systems.* **vol.15**: 1107–1121.

Mercer, J. (1909). Functions of positive and negative type, and their connections with theory of integral equations, *Transactions of the London Philosophical Society.* **vol.209**: 415–446.

Mesquita, R. (1998). *Gramática da Língua portuguesa*, 7 edn, Editora Saraiva.

Michalewicz, Z. (1994). *Genetic algorithms+Data structures=Evolution programs*, Spring-Verlag-New York.

Milner, B.P. ; Vaseghi, S. (1994). Speech modeling using cepstral-time feature matrices and hidden Markov models, *IEEE Proceedings of Conference on Acustic Speech and Signal Processing.* **vol.140**(5): 317–320.

Montalvão, J.; Araujo, M. R. R. (2012). Is masking a relevant aspect lacking in MFCC ? A speaker verification perspective, *Pattern Recognition Letters. Elsevier.* **vol.33**(16): 2156–2165.

Nasersharif, B.; Akbari, A. (2007). SNR-dependent compression of enhanced Mel subband energies for compensation of noise effects on MFCC features, *Pattern Recognition Letter.* pp. 1320–1326.

Oppenheim, A. V.; Schafer, R. W. (2013). *Processamento em tempo discreto de sinais*, 3 edn, Pearson.

Park, Y., Un, C. K. and Kwon, O. (1996). Modeling acoustic transitions in speech by modified hidden Markov models with state duration and state duration-dependent observation probabilities , *IEEE Transactions on Speech and Audio Processing.* **vol.04**(5): 389–392.

Pearl, J. (1973). Asymptotic equivalence of spectral representations, *IEEE Transactions on acustics, speech and signal processing.* **vol.23**(6): 229–232.

Picone, J. (1993). Signal modeling techiniques in speech recognition, *IEEE Transactions on Computer.* **vol.81**(1): 1215–1247.

Rabiner, L. (1989). A tutorial on hidden Markov models and selective applications in speech recognition, *IEEE Proceedings.* **vol.77**: 257–286.

Rabiner, L. ; Biing-Hwang, J. (1993). *Fundamentals of speech recognition*, Prentice Hall.

Rabiner, L. R.; Schafer, R. W. (2007). *Introduction to digital speech processing*, 1 edn, now Publishers Inc.

Rabiner, L.R. ; Schafer, R. (1978). *Digital processing of speech recognition*, Prentice Hall.

Ramesh, P.; Wilpon, J. (1992). Modeling state durations in hidden Markov models for automatic speech recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing.* **vol.01**: 381–384.

Revathi, A. ; Venkataramani, Y. (2011). Speaker independent continuous speech and isolated digit recognition using VQ and HMM, *International Conference on Communications and Signal Processing-ICCSP.* pp. 198–202.

Reynolds, D. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models, *IEE Transactions on speech and audio processing.* **vol.3**: 72–83.

Rubin, P.; Vatikiotis, E. (1998). *Animal acustic communications*, 1 edn, Spring-Verlag.

Sadaoki, F. (1989). On the role of spectral transition for speech perception, *Journal of Acoustic Society of America.* **vol.136**(2): 133–140.

Sadaoki, F. (2000). *Digital speech processing, synthesis, and recognition*, 2 edn, Marcel Dekker, Inc.

Sakr, G. and Elhajj, I. (2011). Digit recognition with confidence, *IEEE Workshop on Signal Processing Systems.* pp. 299–304.

Schalkoff, R. I. (1990). *Artificial Intelligence:An Engineering Approach*, McGraw-Hill, New York.

Seki, H., Ishii, K. and Mizumoto, M. (2010). On the monotonicity of fuzzy inference methods related to TS inference method, *IEEE Transactions on Fuzzy Systems.* **vol.18**: 629–634.

Shabtai, N. R. (2010). *Advances in speech recognition*, 1 edn, Sciyo.

Shannon, C. (1948). A Mathematical Theory of Communication, *The Bell System Technical Journal.* pp. 379–423.

Shenouda, D.; Goneid, D. (2006). Hybrid fuzzy HMM system for Arabic connectionist speech recognition, *The 23rd National U.Jio Science Conference.* pp. 1–8.

Silva, D., de Souza, V., Batista, G. and Giusti, R. (2012). *Spoken gigit recognition in Portuguese using Line Spectral Frequencies*, Lectures Notes in Artificial Intelligence-LNAI 7637.

Silva, W.L.S.; Serra, G. (2012a). Análise Comparativa entre as implicações Lukasiewicz, Dienes-Rescher, Mamdani aplicadas ao reconhecimento de voz, *II CBSF-Proceedings.* pp. 997–1012.

Silva, W.L.S.; Serra, G. (2012b). Proposal of an intelligent speech recognition system., *Third Global Congress on Intelligent Systems.* pp. 356–359.

Silva, W.L.S.; Serra, G. (2014a). A novel intelligent system for speech recognition, *International Joint Conference on Neural Networks-IJJCNN.* pp. 3599–3604.

Silva, W.L.S.; Serra, G. (2014b). Intelligent genetic fuzzy inference system for speech recognition: An approach from low order feature based on discrete cosine transform, *Journal of Control, Automation and Eletrical systems-Springer.* **vol.25**(6): 689–698.

Smith, F., Ming, J., O'Boyle, P. and Irvine, A. (1995). A hidden Markov model with optmized intr-frame dependence, *IEEE International Conference on Acoustics, Speech and Signal Processing.* **vol.01**: 209–212.

Smola, A., Bartlett, P., Scholkopf, B. and Schuurmans, D. (2000). *Advances in large margin classifiers*, Massachusetts Institute of Technology.

Song, S., Feng, C. and Lee, E. (2002). Triple I method of fuzzy reasoning, *Eselvier-Computers and Mathematics with Applications.* **vol.44**: 1567–1579.

Sunitha, S. U. V. (2000). Fast recursive DCT-LMS speech enhancement for performance enhancement of digial hearin aid, *Acadmic Open Journal.* **vol.18**.

Tamgno, J., Bernard, E., Lishou, C. and Richome, M. (2012). Wolof speech recognition model of digits and limited-vocabulary based on HMM and ToolKit, *14th International Conference on Computer Modelling and Simulation.* pp. 389–395.

Tang, C., Lai, E. and Wang, Y. (1997). Distributed fuzzy rules for preprocessing of speech segmentation with genetic algorithm, *IEEE Fuzzy Conference.* **vol.01**.

Tang, K., Man, K., Zhi, Z. and Kwong, S. (1998). Minimal fuzzy memberships and rules using hierarchical genetic algorithms, *IEEE Transactions on Industrial Eletronics.* **vol.45**(1): 162–169.

Tarihi, M., Taheri, A. and Bababeyk, H. (2005). *A new method for fuzzy hidden Markov models in speech recognition*, IEEE-International Conference on Emnerging Technologies.

Trancoso, I. M.; Tribolet, J. M. (1989). Harmonic post-processing of speech synthesised by stochastic coders, *IEEE Proceedings.* **vol.136**(1): 141–144.

Urena, R., Moral, A. I. G., Moreno, C., Ramon, M. and Maria, F. (2012). Real-time robust automatic speech recognition using compact suporte vector machine, *IEEE Transaction on Audio, Speech and Language Processing.* **vol.20**: 1347–1361.

Vapnik, V. (1995). *The nature of statical learning theory*, Springer-Verlag.

Vapnik, V.N.; Chervonenkis, A. Y. (1968). *On the uniform convergence of relative frequencies of events to their probabilities*, Dokl.

Wachter, M., Matton, M., Demuynck, K., Wambacq, P., Cools, R. and Compernolle, D. (2007). Template-based continuous speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing.* **vol.15**: 1377–1390.

Wang, L. (1994). *A course in fuzzy systems and control*, Prentice Hall.

Weihong, Z., Shunqing, X. and Ting, M. (2010). A fuzzy classifier based on Mamdani fuzzy logic system and genetic algorithm, *IEEE Youth Conference on Information Computing and Telecommunications.* pp. 198–201.

Wu, Gin-Der; Lin, C.-T. (2000). Word boundary detection with mel-scale frequency bank in noisy enviroment, *IEEE Transactions on speech and audio processing.* **vol.08**(5): 541–554.

Ynoguti, C.A.; Violaro, F. (2008). A brazilian portuguese speech database, *Simpósio Brasileiro de Telecomunicações.* .

Zeng, F.F. ; Shi, P. (2011). Neural network design based on isolated words, *IEEE Intenational Conference on Machine Learning and Cybernetics.* **vol.02**: 769–772.

Zeng, J.; Liu, Z. (2006). Type-2 fuzzy hidden Markov models and their application to speech recognition, *IEEE Transactions on Fuzzy Systems.* **vol.14**(3): 454–467.

Zhang, X., Wang, X., Zhang, S. and Yu, F. (2010). *Approximating the true domain of fuzzy inference sentence with genetic algorithm*, Vol. vol.01, Seventh International Conference on Fuzzy Systems and Knowledge Discovery.

Zhou, E. ; Khotand, A. (2007). *Fuzzy classifier design genetic algorithms*, Vol. vol.40, Journal of the Pattern Recognition Society-Elsevier.

Zhou, J. ; Chen, P. (2009). Generalized discrete cosine transform, *Pacific-Asia Conference on Circuits,Communications and System.* pp. 449–452.

Apêndice A

Noções básicas sobre sistemas nebulosos

A.1 Introdução

Uma noção básica da teoria de conjuntos é a pertinência de um dado elemento x em um conjunto A, indicado pelo símbolo $\in$, tal que, $x \in A$. Uma forma de se indicar essa pertinência é através de uma função de pertinência $\mu_A(x)$, onde o valor indica se o elemento x pertence ou não ao conjunto A. Isto é, seja x um elemento de um dado conjunto U, então, o subconjunto A de U é um conjunto de pares ordenados dado por:

$$\{[x, \mu_A(x)]\,,\ \forall x \in U\} \tag{A.1}$$

sendo $\mu_A(x)$ definido como o grau de pertinência de x em A. Na lógica aristotélica, essa função é bivalente, isto é, ou $x \in A$ ou $x \notin A$. Na lógica nebulosa todos os valores de $\mu_A(x)$ estão dentro do intevalo $[0, 1]$, isto é, uma pertinência igual a 0 significa que $x \notin A$; por outro lado, uma pertinência igual a 1 significa que $x \in A$ e valores de pertinência entre 0 e 1 equivalem a dizer que a pertinência do elemento x ao conjunto A pode ser parcial. Portanto, um conjunto nebuloso é uma generalização de um conjuto clássico. Quando U é um conjunto contínuo, por exemplo, $U = \mathbb{R}$, A é dado por:

$$A = \int_U \mu_A(x)/x \tag{A.2}$$

em que o símbolo de integral não denota a integral, mas sim a coleção de todos os pontos $x \in U$ associados à função $\mu_A(x)$. E quando U é discreto, tem-se que:

$$A = \sum_{U} \mu_A(x)/x \tag{A.3}$$

onde o símbolo do somatório não representa a operação aritmética de soma, mas sim, a coleção de todos os pontos de $x \in U$ associados à função $\mu_A(x)$.

Outra definição importante em lógica nebulosa são as variáveis linguísticas que são expressas em linguagem natural (termos linguísticos), porém, interpretadas com valores numéricos sobre o conjunto no qual ela está definida. Denomina-se universo de discurso o conjunto de definição da variável linguística (Wang, 1994). Uma variável linguística é caracterizada por (X, T, U, M), onde,

1. X é o nome da variável linguística;
2. T é o conjunto de valores linguísticos que X pode assumir;
3. U é o universo de discurso da variável linguística;
4. M é uma regra semântica que relaciona cada valor linguístico em T com um conjunto nebuloso em U.

No que se refere a variável linguística X, há algumas restrições, a saber:

1. X deve ser normal:

$$max\left[\mu_X(u)\right] = 1 \ , u \in U \tag{A.4}$$

2. X deve ser convexo:

$$\mu_X\left[\lambda u_1 + (1-\lambda)u_2\right] \geq min\left[\mu_X(u_1), \mu_X(u_2)\right], \ u_1, u_2 \in U, \ \lambda \in [0,1] \tag{A.5}$$

   A convexidade de X garante a unicidade da avaliação numérica do valor no universo de discurso (Bazaraa et al., 1993).

Sistemas estáticos ou dinâmicos que fazem uso de conjuntos nebulosos ou de lógica nebulosa em seu modelamento matemático, chama-se de Sistemas Nebulosos. Há um número muito grande de modos em que os conjutos e lógicas nebulosos podem ser aplicados ao modelamento de sistemas, tais como: na descrição de sistema, na especificação dos parâmetros dos sistemas, e a relação de variáveis de estado de entrada e saída também pode ser modelada por sistemas nebulosos.

Os sistemas mais empregados são aqueles relacionados por meio de regras ***SE-ENTÃO***, que são chamados de sistemas nebulosos baseados em um conjunto de regras. Os sistemas nebulosos podem ser aplicados em modelagem, identificação de sistemas, análise de dados, classificação, predição e controle entre outros. Nos sistemas baseados em regras, as relações entrada-saída são dadas na forma:

SE proposição *antecedente* ***ENTÃO*** proposição *consequente*.

A.1.1 Proposições Nebulosas

Há dois tipos de proposições nebulosas: a proposição atômica, e a proposição composta. Uma proposição atômica é uma simples afirmação do tipo "x é A", onde x é uma variável linguística e A é um valor linguístico de x; isto é, A é um conjunto nebuloso definido no domínio físico de x. O valor verdadeiro da proposição (número real entre zero e um) depende do grau pertinência (similaridade) entre x e A. Uma proposição composta é uma composição de proposições atômicas que utilizam conectivos lógicos "E", "OU" e "$NÃO$", que representam as operações de interseção, união e complemento, respectivamente. Vale ressaltar que, em uma proposição composta, as proposições atômicas são independentes. Uma composição nebulosa é compreendida com relações nebulosas, a saber:

Para o conectivo "E" usa-se a relação de interseção, por exemplo: se x e y são variáveis línguísticas no domínio físico de U e V, e A e B são conjuntos de valores línguísticos em U e V, respectivamente; então, a proposição nebulosa composta dada por:

$$SE\ x \text{ é } A\ E\ y \text{ é } B$$

é interpretada como uma relação nebulosa $A \cap B \in (U \times V)$ com função de pertinência dada por:

$$\mu_{A \cap B}(x, y) = t\left[\mu_A(x), \mu_B(y)\right] \tag{A.6}$$

onde $t : [0, 1] \times [0, 1] \rightarrow [0, 1]$ é qualquer $norma - t$ (Mas et al., 2007; Seki et al., 2010; Wang, 1994)

Para o conetivo "OU" utiliza-se a relação de união. Especificamente esta relação é dada por:

$$SE\ x \text{ é } A\ OU\ y \text{ é } B$$

e é interpretada como uma relação nebulosa de $A \cup B \in (U \times V)$ com função de pertinência dada por:

$$\mu_{A \cup B}(x, y) = s\left[\mu_A(x), \mu_B(y)\right] \tag{A.7}$$

em que $s : [0,1] \times [0,1] \rightarrow [0,1]$ é qualquer $norma - s$ (Mas et al., 2007; Seki et al., 2010; Wang, 1994)

Para o conectivo "$N\tilde{A}O$"utiliza-se a relação complemento. Isto é, troca-se A pelo seu valor negado, esta relação é dada por:

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x) \tag{A.8}$$

A.1.2 Base de Regras Nebulosas

Uma base de regras nebulosas Ru consite de um conjunto de regras nebulosas $SE - ENT\tilde{A}O$. Por exemplo: $SE\ x$ é $A\ ENT\tilde{A}O\ y$ é B. Isso pode ser interpretado como uma relação nebulosa no produto cartesiano dos domínios de x e y. Inferências em um sistema nebuloso baseado em regras é um processo no qual se obtém um conjunto nebuloso de saída dado um conjunto nebuloso de entrada. Em um sistema de inferência nebuloso, o princípio da lógica nebulosa relaciona as regras $SE - ENT\tilde{A}O$ para uma base de regras Ru em um mapeamento de um conjunto nebuloso $A' \subset U$ para um conjunto nebuloso $B' \subset V$, e as relações das regras $SE - ENT\tilde{A}O$ são interpretadas como um produto dentro do espaço $(U \times V)$. Se a base de regras consiste de uma simples regra, então pode-se utilizar o *Modus Ponens* generalizado para especificar um mapeamento do conjunto $A' \subset U$ para o conjunto $B' \subset V$ com a seguinte forma (Wang, 1994):

$$\begin{aligned} &\text{SE}\ \ x\ \ \text{é}\ \ A\ \ ENT\tilde{A}O\ \ y\ \ \text{é}\ \ B \\ &\text{SE}\ \ x\ \ \text{é}\ \ A'\ \ ENT\tilde{A}O\ \ y\ \ \text{é}\ \ B' \end{aligned} \tag{A.9}$$

Dados um conjunto de entrada nebuloso $A' \subset U$, um conjunto de saída nebuloso $B' \subset V$ e ainda, que $Q(U, V)$ é uma realção de mapeamento de U em V, pode-se obter a regra composicional de inferência (Mas et al., 2007), dada por:

$$\mu_{B'}(y) = \sup_{x \in U} \mathrm{t}\left[\mu_{A'}(x), \mu_Q(x, y)\right] \tag{A.10}$$

Esta equação é denominada regra de composição de inferência. Na literatura especializada em sistemas nebulosos, o símbolo "$\star$" é, normalmente, utilizado para representar uma norma-t;

assim, pode-se representar a equação (A.10) por

$$\mu_{B'}(y) = \sup_{x \in U} [\mu_{A'}(x) \star \mu_Q(x, y)] \tag{A.11}$$

Desse modo, tem-se a chamada composição "*super* − *star*", onde o "*sup*" representa o valor máximo da função. Das equações (A.9) e (A.10) obtém-se a equação para o *Modus Ponens Generalizado*, dado o conjunto nebuloso $A^{'} \subset U$ (que representa a proposição antecedente x é $A^{'}$) e a relação nebulosa $[A \text{ ENTÃO } B] \subset (U \times V)$ (que representa a implicação a SE x é A ENTÃO y é B), o conjunto nebuloso $B^{'} \subset V$ (que representa a proposição consequente y é $B^{'}$), infere-se que:

$$\mu_{B'}(y) = \sup_{x \in U} \mathrm{t}\left[\mu_{A'}(x), \mu_{A \to B}(x, y)\right] \tag{A.12}$$

onde $(\to)$ é o operador de implicação.

A base de regras de sistemas práticos, usualmente, consiste de mais de uma regra. Há dois modos para inferir um conjunto de regras: Inferência baseada em Composição e Inferência baseada em regras individuais(Gang, 2010; Wang, 1994). Neste trabalho utilizou-se a Inferência baseada em regras individuais. De modo geral, uma base de regras nebulosas é dada por:

$$R^l: \; SE \;\; x_1 \;\text{é}\; A_1^l \;\; E \;\; x_2 \;\text{é}\; A_2^l \;\; E \cdots E \;\; x_n \;\text{é}\; A_n^l \;\text{ENTÃO}\; y \;\text{é}\; B_l \tag{A.13}$$

onde A_i^l e B_l são conjuntos nebulosos em $U_i \subset \mathbb{R}$ e $V \subset \mathbb{R}$, e $\bar{\mathbf{x}} = [x_1, x_2, ..., x_n]^T \in U$ e $y \in V$ são variáveis linguísticas de entrada e saída do sistema nebuloso, respectivamente, e M é o número de regras em uma base de regras nebulosas, isto é, $l = 1, 2, ...M$.

Em sistema de inferências nebuloso baseado em regras individuais, cada regra na base de regras determina uma saída nebulosa, e a saída do sistema como um todo é a combinação das M regras individuais. Esta combinação pode ser obtida pelas relações de união ou interseção, através dos seguintes passos:

Passo 1: Para a $M-$ésima regra nebulosa SE-ENTÃO na forma da equação (A.13), determina-se as funções de pertinências dadas por:

$$\mu_{A_1^l \times \cdots \times A_n^l}(x_1, \ldots, x_n) = \mu_{A_1^l}(x_1) \star \cdots \star \mu_{A_n^l}(x_n) \tag{A.14}$$

Passo 2: Notando-se que $A_1^l \times \cdots \times A_n^l$ é uma proposição e que B_l também é uma proposição, determina-se:

$$\mu_{R^l}(x_1, \ldots, x_n, y) = \mu_{A_1^l \times \cdots \times A_n^l \to B_l}(x_1, \ldots, x_n, y) \tag{A.15}$$

para $l = 1, 2, \cdots, M$, a implicação pode ser qualquer implicação nebulosa, tais como Dienes-Rescher (Cai, 2002), Lukasiewicz (Elloumi et al., 2004), Zadeh (Song et al., 2002), Godel (Dubois et al., 1997) e Mamdani (Mamdani, 1977). A implicação Mamdani baseada no mínimo é dada por:

$$\mu_{Q_{MM}}(\bar{\mathbf{x}}, y) = min\left[\mu_{A_1^l \times \cdots \times A_n^l}(\bar{\mathbf{x}}), \mu_{B_l}(y)\right] \tag{A.16}$$

Passo 3: Para um dado conjunto nebuloso de entrada $A^{'} \subset U$, calcula-se o conjunto nebuloso de saída $B_l^{'} \subset V$ para cada regra individual R^l de acordo com o Modus Ponens Generalizado, dado na equação (A.9), e a equação (A.15)

$$\mu_{B_l^{'}}(y) = \sup_{\bar{\mathbf{x}} \in U} \mathrm{t}\left[\mu_{A^{'}}(\bar{\mathbf{x}}), \mu_{R^l}(\bar{\mathbf{x}}, y)\right] \tag{A.17}$$

para $l = 1, 2, \cdots, M$.

Passo 4: A saída do sistema de inferência nebuloso é a combinação dos M conjuntos nebulosos $\left\{B_1^{'} \ldots B_M^{'}\right\}$, obtida através das operações nebulosas de $norma-s$ e $norma-t$, respectivamente, conforme abaixo:

$$\mu_{B^{'}}(y) = \mu_{B_1^{'}}(y) \dot{+} \cdots \dot{+} \mu_{B_M^{'}}(y) \tag{A.18}$$

onde "$\dot{+}$" representa qualquer $norma - s$.

$$\mu_{B^{'}}(y) = \mu_{B_1^{'}}(y) \star \cdots \star \mu_{B_M^{'}}(y) \tag{A.19}$$

onde "$\star$" representa qualquer $norma - t$ (Zhou, 2007).

Em sistemas nebulosos, observa-se que se pode fazer várias escolhas para caracterização dos mesmos, tais como: que tipos de inferências de regras que serão utilizadas, os tipos de implicações e as operações $norma - s$ ou $norma - t$ nas suas várias formulações. Em geral três critérios devem ser utilizados nesta escolha:

1. Recursos intuitivos: As regras devem ser escolhidas de forma a fazerem um mínimo de sentido na resolução do problema;

2. Eficiência Computacional: A escolha deve resultar em uma forma simples de se calcular B' e A' e suas relações de domínios.

3. Propriedades Especiais: As escolhas devem gerar um sistema de inferência que tem propriedades especiais relacionadas aos problemas a serem resolvidos.

Apêndice B

Considerações sobre o Algoritmo Genético

B.1 O Algoritmo Genético e a Otimização

"*As primeiras descrições e definições técnicas de adaptação vêm da biologia. Naquele contexto, adaptação designa qualquer processo pelo qual uma estrutura seja modificada progressivamente para melhor desempenho no seu ambiente. As estruturas podem variar de uma molécula de uma proteína à pata de um cavalo ou a um cérebro humano ou mesmo um grupo de organismos interagindo tal como a vida selvagem da selva africana*" (Holland, 1975).

A solução de problemas que envolvem, de alguma forma, método de busca local promovem pequenas alterações em possíveis soluções até que uma solução ótima seja identificada. Algoritmos genéticos são uma forma de busca local que usa métodos baseados em evolução para fazer pequenas alterações em uma população de cromossomos, na tentativa de identificar uma solução ótima. A solução ótima, geralmente, está associada a processos de otimização, nos quais procura-se ajustar as caracteríesticas de um dispositivo, processo matemático, ou experimento para encontrar o valor mínimo ou máximo para a saída. As características ou entradas consistem de variáveis; o processo ou função a ser otimizada e denominada função custo ou função de adaptação (*fitness*). Se o processo é um experimento, então as variáveis são entradas físicas para o experimento.

A otimização, enquanto método de busca, consiste em obter variações partindo de um conceito inicial e, através das informações adquiridas, encontrar uma solução que introduza melhorias em determinado processo (Coppin, 2004; Cox, 1999).

A otimização busca encontrar a "melhor solução". Esta terminologia implica que existe mais de uma solução e as mesmas não possuem valores iguais, por consequência, a definição de

"melhor" é relativo ao problema, ao método de solução e a tolerância permitida na aplicação. A busca pela melhor solução, nem sempre é uma tarefa trivial e, muitas vezes, requer um grande número de avaliações da função de custo a fim de encontrar o valor ótimo.

A otimização e o conjunto de técnicas de buscas baseadas nos princípios da genética e da seleção natural é denominado Algoritmo Genético (AG). Um AG permite que uma população composta de muitos indivíduos desenvolva um estado de máxima aptidão de acordo com certas regras de selação. O método foi desenvolvido por John Holland (Holland, 1975). Algumas das vantagens do AG podem ser enumeradas com abaixo:

1. Otimiza processos com variáveis continuas ou variáveis discretas;
2. Não requer uso de informações de derivadas;
3. Busca simultânea para uma larga quantidade de variáveis de uma dada função custo;
4. Trabalho com computação paralela;
5. Otimiza variáveis com superfícies de custo extremamente complexas;
6. Fornece uma lista de variáveis ótimas e não uma simples solução;
7. Trabalha com dados gerados numericamente, dados experimentais ou funções analíticas;

O AG não é a melhor solução para resolver todos os tipos de problema de otimização. Os métodos tradicionais encontram soluções ótimas, normalmente, de forma mais rápida para funções analíticas convexas bem comportadas e de poucas variáveis.

Exemplo: Deseja-se minimizar a função definida por:

$$f(x, y) = xsen(4x) + 1.1ysen(2y) \tag{B.1}$$

sujeito a $0 \leq x \leq 10$ e $0 \leq y \leq 10$.

As figuras B.1 e B.2, mostram os gráficos tridimensional e de contorno de $f(x, y)$, respectivamente. Estes gráficos auxiliam na análise dos pontos de mínimo da função e as regiões onde estão localizados.

Nesta análise, verifica-se um grande número de soluções, porém em um espaço finito de busca, com um número de combinações de valores diferentes para as variáveis, dado por:

$$V = \prod_{i=1}^{N_{var}} Q_i \tag{B.2}$$

onde V é o número das diferentes combinações entre as variáveis; N_{var} é o número total de variáveis; Q_i é o número de valores que a variável i pode assumir.

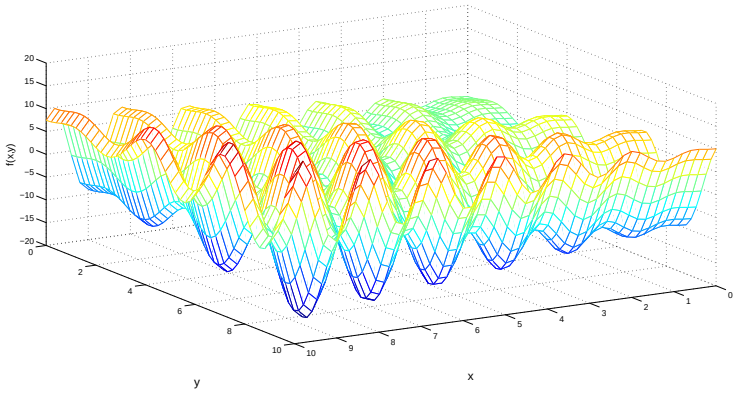


Fig. B.1: Gráfico tridimensional de f(x,y).

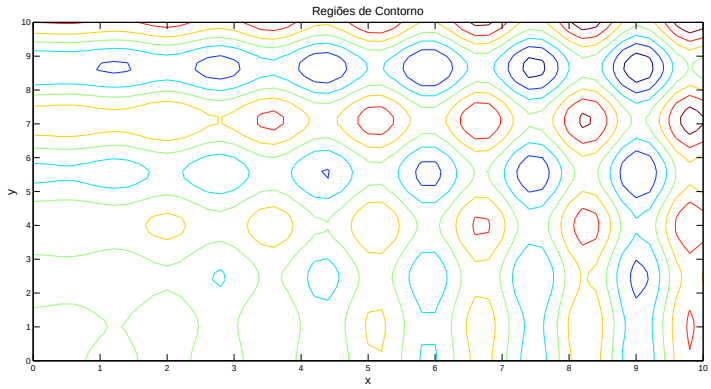


Fig. B.2: Gráfico de contorno de f(x,y).

B.1.1 Otimização Analítica

Nos métodos analíticos clássicos, para encontrar o ponto extremo (mínimo ou máximo) de uma função, com uma única variável, deve-se igualar a zero a derivada a primeira da função de custo e, então, calcular o valor da variável. Caso a derivada a segunda seja maior que zero, o extremo é um ponto de mínimo; se for menor que zero, é ponto de máximo.

Já para uma função com mais de uma variável, para o cálculo do ponto extremo, deve-se igualar o gradiente da função a zero, $\nabla f(x, y) = 0$. Como exemplo, utiliza-se a função B.1, considerada anteriormente, a qual possui o gradiente dado por:

$$\frac{\partial f}{\partial x} = sen(4x_m) + 4xcos(4x_m) = 0, \qquad 0 \leq x \leq 10 \tag{B.3}$$

e

$$\frac{\partial f}{\partial y} = 1.1sen(2y_m) + 2.2y_m cos(2y_m) = 0, \qquad 0 \leq y \leq 10 \tag{B.4}$$

Estas equações são resolvidas através das raízes, x_m e y_m, que correspondem a uma família de linhas. O extremo ocorre na intersecção destas linhas, porém estas equações nem sempre aparecem separadamente, o que torna ainda mais difícil o cálculo destas raízes. O Laplaciano da função é dado por:

$$\frac{\partial^2 f}{\partial x^2} = 8cos(4x) - 16xsen(4x) = 0, \qquad 0 \leq x \leq 10 \tag{B.5}$$

e

$$\frac{\partial^2 f}{\partial y^2} = 4.4cos(2y) - 4.4ysen(2y) = 0, \qquad 0 \leq x \leq 10 \tag{B.6}$$

onde as raízes representam pontos de mínimo quando $\nabla^2 f(x_m, y_m) > 0$. Porém, este processo não deixa claro se o mínimo encontrado é o mínimo global, o qual é o objetivo em um processo de Otimização.

No século XVIII, Lagrange introduziu uma técnica, que incorpora as restrições de igualdade dentro da função de custo, denominada *Multiplicadores de Lagrange*. Para demonstrar esta técnica, considera-se novamente a equação B.1, com restrição de $x + y = 0$. Com a restrição aplicada à equação B.1 (função de custo), tem-se:

$$f_\lambda = xsen(4x) + 1.1ysen(2y) + K(x + y) = 0 \tag{B.7}$$

Aplicando o gradiente na equação B.7, resulta em:

$$\frac{\partial f}{\partial x} = sen(4x_m) + 4x_m cos(4x_m) + K = 0 \tag{B.8}$$

$$\frac{\partial f}{\partial y} = 1.1sen(2y_m) + 2.2y_m cos(2y_m) + K = 0 \tag{B.9}$$

$$\frac{\partial f}{\partial K} = x_m + y_m = 0 \tag{B.10}$$

Subtraindo (B.9) de (B.8) e utilizando-se a equação (B.10), tem-se:

$$4x_m cos(4x_m) + sen(4x_m) + 1.1sen(2x_m) + 2.2x_m cos(2x_m) = 0 \tag{B.11}$$

onde $(x_m, -x_m)$ representam os mínimos da equação (B.11). A solução apresenta-se novamente

como uma família de linhas que atravessam o domínio. Uma das desvantagens da utilização de abordagens como essa, que envolvem análises analíticas, é que o ponto extremo (ponto de mínimo ou ponto de máximo) encontrado pode não representar uma solução viável na resolução de alguns problemas práticos que utilizem este procedimento. Embora muitas vezes impraticável, as abordagens numéricas baseiam-se em cálculos de derivadas da função de custo e os algoritmos de busca utilizados, geralmente, começam em algum ponto aleatório do espaço de busca, calcula-se o gradiente, e então segue em uma única direção, induzindo ao erro de encontrar um mínimo local, por exemplo, em vez de um mínimo global, além de ter o funcionamento comprometido quando envolvem variáveis discretas.

O Algoritmos Evolucionário (AE) surge como uma alternativa diante dessas limitações que os métodos analíticos de otimização apresentam. Através de um método de busca estatístico, baseado nas teorias biológicas evolucionistas, permite que uma população possa evoluir sob regras de seleção especificadas, para um estado que miniminize a função de custo. Em um AE, os pontos no espaço de busca representam indivíduos que interagem entre si, um conjunto de possíveis soluções da função de custo (população) é manipulado a cada iteração (geração).

Um AG é um tipo de abordagem utilizada em AE, desenvolvida conforme a teoria evolucionista do neo-Darwinismo, a qual descreve que os quatro procedimentos essenciais na evolução biológica das espécies, são: competição, reprodução, mutação e seleção. E são justamente esses procedimentos que são utilizados na manipulação dos indivíduos de cada população, a cada geração, em um AG. Além de não requerer propriedades de convexidade e diferenciabilidade, os AGs, fornecem uma lista de possiveis soluções (não uma única solução) e podem convergir com variáveis contínuas e discretas.

B.2 Processo de Implementação de um Algoritmo Genético

Tanto o algoritmo binário quanto o contínuo utilizam a modelagem genética de recombinação e seleção natural. No início de um algoritmo genético, alguns parâmetros devem ser informados pelo usuário, tais como: número de indivíduos da população inicial (cromossomos), número de genes, função de custo, número de gerações, taxa de cruzamento, taxa de mutação e parâmetro de convergência.

Um AG inicia definindo o número de cromossomos, com seu respectivo número de genes, representado como segue:

$$cromossomo = [g_1, g_2, \ldots, g_n] \tag{B.12}$$

onde g representa cada gene associado ao cromossomo, e n é um número inteiro.

Cada cromossomo tem um custo associado à função de custo f, definida por:

$$Custo = f(cromossomo) = f(g_1, g_2, \ldots, g_n) \tag{B.13}$$

A população inicial é definida através de indivíduos gerados aleatoriamente, com soluções factíveis dentro do espaço de busca considerado. Logo na primeira iteração (geração), os custos associados a cada indivíduo são obtidos através da função de custo do AG. Os indivíduos que possuem as melhores soluções são selecionados para permanecer e interagir com os demais indivíduos (através dos procedimentos de cruzamento e mutação). Dentre estes indivíduos que foram selecionados, uma parcela é utilizada para cruzamento. O número de indivíduos selecionados para cruzamento é definido pela seguinte função:

$$I_c = T_c N \tag{B.14}$$

onde T_c é a taxa de seleção para cruzamento e N corresponde ao número de indivíduos que permanecem na população.

A partir daí é realizado o processo de cruzamento com os indivíduos selecionados para este fim. Dentre os descendentes provenientes do cruzamento, são selecionados alguns indivíduos para sofrer a mutação. O número de indivíduos selecionados para mutação é definido como segue:

$$I_M = T_M * descendentes \tag{B.15}$$

onde T_M é a taxa de seleção de indivíduos para o processo de mutação.

Após as mutações, os custos associados aos descendentes e os indivíduos que sofreram mutações são calculados. O processo descrito é iterativo, descrevendo cada geração, como uma iteração. A cada nova geração, nova população de descendentes é avaliada, sendo tal resultante de operações de cruzamento e mutação da geração anterior. A convergência está relacionada ao número de gerações que evoluem, a qual é definida a partir de que a solução aceitável seja encontrada ou determinado número de iterações seja alcançado.

B.2.1 Parâmetros de um AG

Alguns parâmetros, descritos anteriormente, que compõem um AG, merecem atenção no processo de implementação, pois influenciam diretamente no desempenho do algoritmo, são eles:

Tamanho da população: determina o número de cromossomos na população, o que afeta diretamente o desempenho e a eficiência de um AG. Com uma população pequena, o desempenho do AG pode não ser satisfatório, pois a população limitada diminui o espaço de busca do problema. Uma grande população geralmente abrange um espaço maior de busca, além de prevenir convergências prematuras para soluções locais ao invés de globais, no entanto, um maior número de cromossomos requer maior recurso computacional ou um período de tempo maior de processamento;

Taxa de Cruzamento: determina a porcentagem de indivíduos a serem selecionados para o processo de cruzamento. Quando o valor desta taxa é muito alto, a maior parte da população pode ser substituída, o que pode acarretar na perda de pontenciais indivíduos com valor ótimo, porém o valor baixo desta taxa, pode aumentar o tempo de convergência do algoritmo;

Taxa de Mutação: trata-se da probabilidade de ocorrer o processo de mutação. A mutação é utilizada para inserir novo material genético à população, ou seja, fornecer novas informações aos cromossomos, além de prevenir que a população se sature com cromossomos semelhantes (Convergência Prematura). Um baixa taxa de mutação previne que um valor permaneça como melhor solução durante várias gerações, além disso evita que se chegue em qualquer ponto do espaço de busca. Porém, quando esta taxa tem um valor alto a busca se torna excessivamente aleatória, o que pode acarretar na perda de soluções ótimas. Caso não houvesse as mutações, depois de um tempo todos os cromossomos e os custos associados aos indivíduos seriam os mesmos.

B.2.2 Operadores Genéticos

São os operadores genéticos que proporcionam a evolução de um AG nas sucessivas gerações, fazendo com que o mecanismos de busca alcance um resultado satisfatório. Um algoritmo genético padrão evolui, mediante o uso de três operadores básicos:

Seleção: este operador permite que os melhores indivíduos permaneçam, ou seja, é-lhes dada a preferência no processo de cruzamento. A caracterização deste indíviduo, enquanto solução ótima, é dado através da função de custo. É importante ressaltar que este operador não cria nenhuma nova solução, apenas enfatiza as melhores soluções que constituem uma população. Dentre os operadores de seleção, podemos citar os seguintes:

Seleção Proporcional: utiliza uma distribuição de probabilidade de tal forma que a seleção de um dado indivíduo para reprodução é proporcional à função de custo do indivíduo.

Assim, dada a função de custo de cada indivíduo em uma dada geração, a representação do somatório dos custos totais da população de uma dada geração é dada por:

$$F_T = \sum_{s=1}^{N_{ind}} f_s(x) \tag{B.16}$$

onde N_{ind} é o número total de indivíduos da população considerada e f_s é a função de custo relacionada a cada indivíduo desta população. Assim, a propabilidade de seleção, p_s é atribuída para cada indivíduo através da seguinte equação:

$$p_s(x) = \frac{f_s(x)}{F_T} \tag{B.17}$$

A probabilidade acumulada para cada indivíduo é obtida através da soma das funções de custo dos membros da população com classificação inferior à sua:

$$c_s = \sum_{V=1}^{s} p_V, \qquad s = 1, 2, \ldots, N_{ind} \tag{B.18}$$

ou seja, um número R uniformemente distribuído em $[0, 1]$ é obtido N_{ind} vezes e a cada tempo o s-ésimo valor de p é selecionado tal que $c_{s-1} < r \leq< c_s$. Caso $r < c_1$, o primeiro indivíduo é selecionado. Este procedimento pode ser visualizado por meio de uma roleta com N_{ind} partes, onde cada parte tem tamanho proporcional ao custo do indivíduo.

Uma variante da seleção proporcional é através do *ranking*, ou seja, o enfileiramento ou organização dos indivíduos em ordem crescente, em relação aos custos que apresentam, a cada geração. Os métodos de *ranking* requerem somente o valor da função de custo, para mapear as soluções em um conjunto parcialmente ordenado. Um exemplo desta forma de seleção é a seleção por *ranking* geométrico normalizado, dado por:

$$p(x_s) = q'(1-q)^{R_a - 1} \tag{B.19}$$

$$q^{'} = \frac{q}{1-(1-q)^{N_{ind}}}, \qquad s = 1, 2, \ldots, N_{ind} \tag{B.20}$$

onde q é a probabilidade de selecionar o melhor indivíduo e $R_a(s)$ é a posição que o indivíduo ocupa no enfileiramento, ou seja, $R_a(s) = 1$ é a melhor posição.

Seleção elitista com truncamento: consiste na seleção, dos melhores indivíduos da população através de um coeficiente de truncamento, cujo valor encontra-se no intervalo

entre $[0, 1]$. Este coeficiente determina os melhores indivíduos, através do valor de custo que apresentam, que serão mantidos na próxima geração, e dentre estes serão selecionados os indivíduos que passarão pela operação de cruzamento.

Seleção por torneio: nesta seleção um grupo de indivíduos é escolhido aleatoriamente. O grupo de indivíduos ocupa parte de um torneio, onde o indivíduo vencedor é determinado mediante a função de custo que apresenta. O melhor indivíduo pode ser escolhido deterministicamente ou através de processos estocásticos. Somente o vencedor da cada grupo deve ser inserido na população da próxima geração. O número de indivíduos de cada grupo a participarem do torneio, bem como o número de vezes que o procedimento será realizado, devem ser definidos pelo usuário no início do AG.

Cruzamento: O operador de cruzamento é o operador do AG que efetua a troca de partes dos cromossomos entre os indivíduos. O/os ponto/pontos de cruzamento são escolhidos aleatoriamente. A parte de dois cromossomos ancestrais, à direita do ponto de cruzamento são trocadas, para a formação de dois cromossomos descendentes. A frequência do cruzamento entre os indivíduos de uma dada geração é determinada a partir da taxa de cruzamento indicada pelo usuário do AG. O operador de cruzamento com um ou mais pontos de corte são representados na figura B.3.

O cruzamento uniforme é um outro tipo de operador muito utilizado, onde seleciona-se genes de um cromossomo para a troca de material genético, a quantidade de genes a serem trocados é determinada pelo usuário através de porcentagem. A figura B.4 mostra a realização de um operador de cruzamento uniforme entre dois indivíduos, onde há a troca de dois genes no cromossomo.

O operador de cruzamento também pode ser realizado através de uma soma ponderada entre dois indivíduos selecionados, um operador de cruzamento simples com fator de ponderação é dado como segue:

$$\begin{aligned} descendente_1 &= \eta * cromossomo_1 + (1 - \eta) * cromossomo_2 \\ descendente_2 &= \eta * cromossomo_2 + (1 - \eta) * cromossomo_1 \end{aligned} \tag{B.21}$$

onde o $(cromossomo_1)$ e o $(cromossomo_2)$, representam os indivíduos selecionados e utilizados no operador de cruzamento, $descendente_1$ e $descendente_2$ correspondem aos indivíduos obtidos a partir da operação de cruzamento realizada entre os dois cromossomos selecionados e η é o fator de ponderação, cujo valor aleatório está compreendido entre 0 e 1.

Mutação: este operador permite a introdução de material genético à população, evitando convergência prematura do algoritmo a ótimos locais, porém não pode ser muito alto, pois

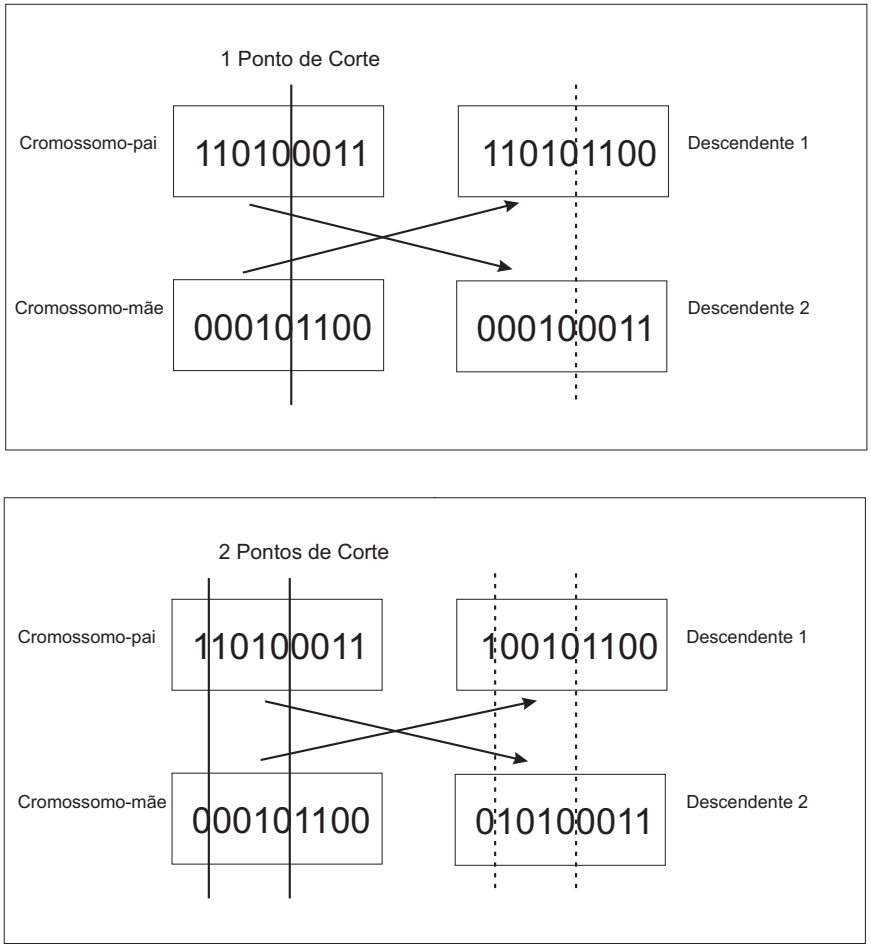


Fig. B.3: Operadores de cruzamento com pontos de corte.

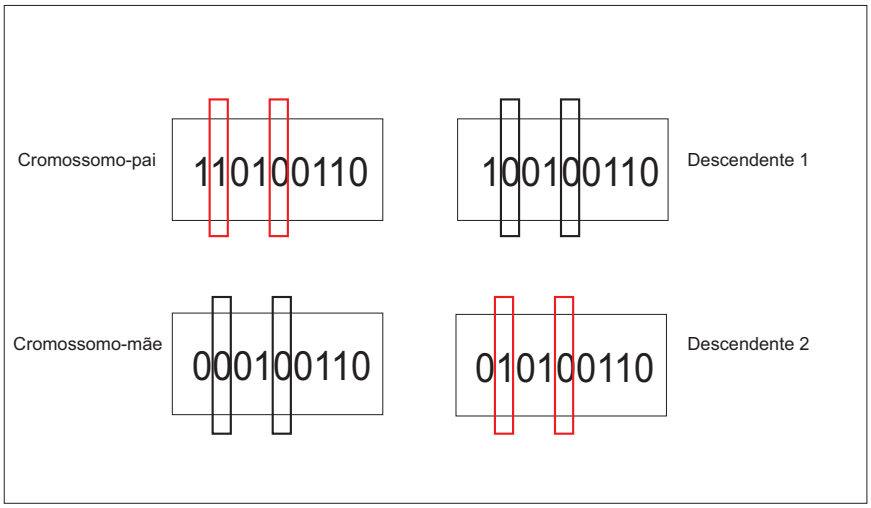


Fig. B.4: Operador de cruzamento uniforme.

pode provocar a perda de boas soluções dentro da população considerada. A definição do grau de mutação em um AG é definido através da taxa de mutação. A posição, bem como o número de genes a ser modificado em um cromossomo é definida pelo usuário de um AG. Na figura B.5 pode ser observado um operador de mutação em um cromossomo, onde dois genes sofrem a mudança de material genético.

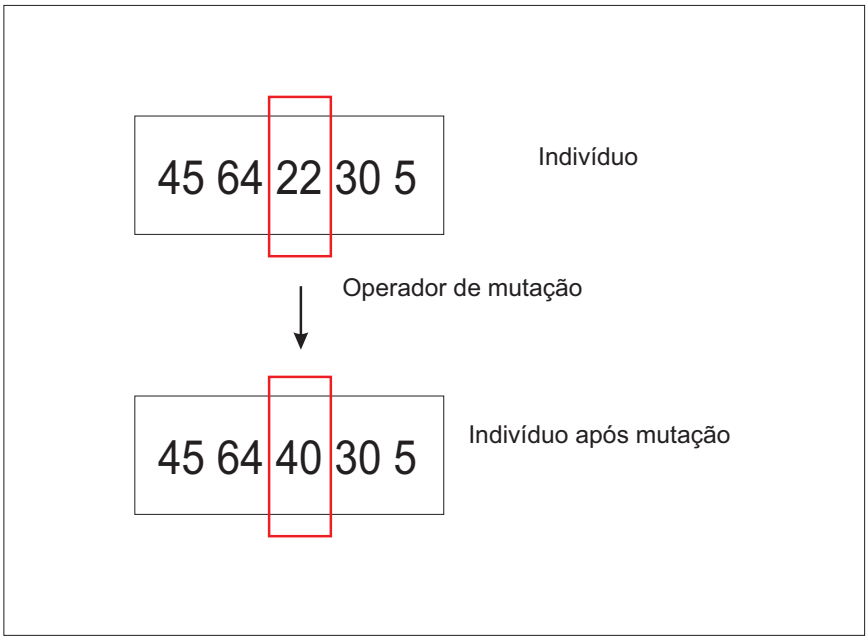


Fig. B.5: Operador de mutação.

Em alguns AGs o gene selecionado par mutação é definido através de uma poderação definida pela seguinte equação:

$$g_n^{'} = g_n + \sigma N_n\left(0,1\right) \tag{B.22}$$

onde g_n é a variável selecionada para a mutação, σ corresponde ao desvio padrão da distribuição normal e $N_n\left(0,1\right)$ é a distribuição padrão normal, com média igual a zero e variância igual a 1.

Apêndice C

Máquina de Vetor de Suporte: uma análise qualitativa

C.1 Introdução

A máquina de vetor de suporte foi desenvolvida por (Vapnik, 1995), com o intuito de resolver problemas de classificação de padrões, a partir de estudos iniciados no trabalho "*On the uniform convergence of relative frequencies of events to their probabilities*"(Vapnik, 1968). O classificador SVM é uma outra categoria das redes neurais *feed-forward*, ou seja, redes cujas saídas dos neurônios de uma camada alimentam os neurônios da camada posterior, não ocorrendo a realimentação (Haykin, 2009). Essa técnica, originalmente desenvolvida para classificação binária, busca a construção de hiperplanos como superfícies de decisão, de tal forma que a separação entre classes seja máxima, considerando-se que os padrões sejam linearmente separáveis. Já para padrões não-linearmente separáveis, busca-se uma função de mapeamento apropriada para tornar o conjunto mapeado linearmente separável. Devido a sua eficiência em trabalhar com dados de alta dimensionalidade, é citada na literatura como uma técnica altamente robusta, muitas vezes comparada as redes neurais (Ding, 2001). Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado, como as redes neurais artificiais (RNAs). Exemplos de aplicações de sucesso podem ser encontrados em diversos áreas, como na classificação de voz (Urena et al., 2012), no processamento de imagens (Garcia et al., 2007) e em Bioengenharia (Ding, 2001). Os SVMs são baseadas na Teoria de Aprendizagem Estatística (Vapnik, 1995), que estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu.

C.2 Teoria da Aprendizagem Estatísitca

Um modelo de aprendizagem supervisionada baseada na Toeria da Aprendizagem Estatística é dado na figura C.1 (Haykin, 2009).

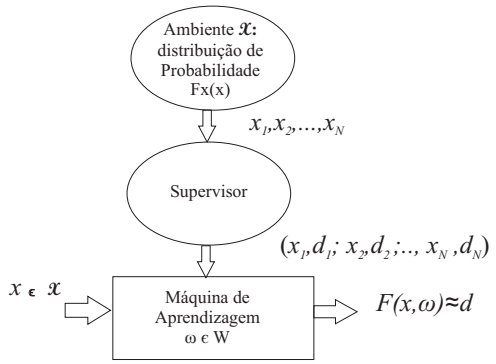


Fig. C.1: Fluxograma de um modelo de aprendizagem supervisionada

Ambiente: É estacionário e fornece um vetor de entrada $\bar{x}$ com uma função de distribuição de probabilidade cumulativa fixa, mas desconhecida $F_x(\bar{x})$.

Supervisor: Apresenta uma resposta desejada d para cada vetor de entrada $\bar{x}$ fornecido pelo ambiente, de acordo com uma função de distribuição cumulativa condicional $F_x(\bar{x}/d)$ que é também fixa, mas desconhecida. A resposta desejada d e o vetor de entrada x estão relacionados pela equação (C.1):

$$d = f(\bar{x}, v) \tag{C.1}$$

em que v é o ruído.

Máquina de aprendizagem: Algoritmo capaz de desenvolver um conjunto de funções de mapeamento do tipo entrada-saída que é dado por:

$$y = F(\bar{x}, \bar{w}) \tag{C.2}$$

em que y é a resposta real produzida pela máquina de aprendizagem associada à entrada $\bar{x}$, e $\bar{w}$ é um conjunto de parâmetros livres, chamados pesos de ponderação, selecionados do espaço de parâmetros W. O problema da aprendizagem supervisionada é escolher uma função particular $F(\bar{x}, w)$ que melhor aproxima a resposta desejada d. A seleção da função é baseada no conjunto dos N exemplos de treinamento *independente, identicamente distribuídos (iid)* descrito por:

$$\Gamma = \{(\bar{x}_i, d_i)\}_{i=1}^{N} \tag{C.3}$$

em que $\bar{x}_i$ é o padrão de entrada para o $i-$ésima observação e d_i é a resposta desejada cor-

respondente. Cada par de exemplo é retirado de Γ pela máquina de aprendizagem com uma função de distribuição cumulativa conjunta $F_{X,D}(\bar{x}, d)$. A aprendizagem é viável se os exemplos de treinamento contém informações suficientes para construir uma máquina de aprendizagem capaz de ter bom desempenho de generalização. O trabalho de Vapnik e Chervonenkis fornecem ferramentas matemáticas para se trabalhar com a generalização das máquinas de aprendizagem. A aprendizagem supervisionada é um problema de aproximação que consiste em encontrar a função $F(\bar{x}, w)$ que seja a melhor aproximação da função desejada $f(\bar{x})$ (Vapnik, 1968). No problema de aprendizagem supervisionada, deseja-se desempenho de generalização adequado disponível dos dados de treinamento. O método de minimização do risco estrutural fornece um procedimento indutivo, no qual, utilizando-se a dimensão VC como uma variável de controle, pode-se obter tal desempenho.

C.2.1 Funcional de Risco

O desempenho desejado de um classificador f é que o mesmo obtenha o menor erro durante o treinamento, sendo o erro mensurado pelo número de predições incorretas de f. Assim, define-se como Risco Empírico $R_{emp}(f)$, como a medida de perda entre a resposta desejada e a resposta real. A definição do Risco Empírico é descrita como segue,

$$R_{emp}(\bar{w}) = \frac{1}{N} \sum_{i=1}^{N} c(f_i(\bar{x}, y_i)) \tag{C.4}$$

onde $c(.)$ é a função custo relacionada à previsão de $f(\bar{x}_i)$, com saída desejada y_i, onde um tipo de função custo é a "perda 0/1" definida pela equação (C.5). O processo de busca por uma equação $f(\bar{x})$ que represente um menor valor de R_{emp} é denominado de Minimização do Risco Empírico.

$$c(f(\bar{x}_i, y_i)) = \begin{cases} 1, & \text{se } y_i f(\bar{x}_i) < 0 \\ 0, & \text{caso contrário} \end{cases} \tag{C.5}$$

Supõe-se que os padrões utilizados para treinamento $(\bar{x}_i, y_i)$ são gerados por uma distribuição *iid* de probabilidade $P(\bar{x}, y)$ em $\mathbb{R}^N \subset \{-1, +1\}$ sendo P desconhecida. A probabilidade de classificação incorreta do classificador f é denominada de Risco Funcional, que quantifica a capacidade de generalização, dada por (Urena et al., 2012):

$$R(f) = \int c(f(\bar{x}_i, y_i)) dP(\bar{x}_i, y_i) \tag{C.6}$$

Durante o processo de treinamento, $R_{emp}(f)$ pode ser facilmente obtida, ao passo que $R(f)$ não, uma vez que a probabilidade P é desconhecida. Então, dado um conjunto de dados de treinamento $(\bar{x}_i, y_i)$ com $\bar{x}_i \in \mathbb{R}^N$ e $y_i \in \{-1, +1\}, i = 1, 2, ..., n$, sendo $\bar{x}_i$ os vetor de entrada e y_i a saída referente ao vetor $\bar{x}_i$. O objetivo, então, é estimar uma função $f : \mathbb{R}^N \to \{-1, +1\}$; caso nenhuma restrição seja imposta na classe de funções em que se escolhe a estimativa f, pode ocorrer que a função obtenha um bom desempenho no conjunto de treinamento; porém, não terá o mesmo desempenho em padrões desconhecidos, sendo este fenômeno chamado de "*overfitting*". Dessa forma, a minimização do risco empírico não garante uma boa capacidade de generalização, sendo desejado um classificador f^* tal que $R(f^*) = \min_{f \in F} R(f)$, onde F é o conjunto de funções possíveis de f. A Teoria da Aprendizagem Estatística fornece formas de limitar a classe de funções (hiperplanos), com o intuito de excluir modelos ruins, ou seja, que levem ao "*overfitting*", implementando uma função com a capacidade adequada para o conjunto de dados de treinamento. As restrições ao Risco Funcional utilizam o conceito de dimensão VC (Vapnik, 1968).

C.2.2 Limites às classes: Dimensão VC

Dado um conjunto de funções sinal G, sua dimensão VC (Vapnik-Chervonenkis) é definida como o tamanho do maior conjunto de pontos que pode ser particionado arbitrariamente pelas funções contidas em G. Em outras palavras, a dimensão VC do conjunto de funções de classificação G é o número máximo de exemplos de treinamento que pode ser aprendido pela máquina sem erro, para todas as saídas possíveis das funções de classificação (Haykin, 2009). De forma genérica, para funções lineares no $\mathbb{R}^N$ para $n > 2$ a dimensão VC é dada abaixo,

$$VC(n) = n + 1 \tag{C.7}$$

C.3 Máquina de vetor de suporte (*Support Vector Machine* - SVM)

O SVM é proposto como uma extensão do método generalizado para o desenvolvimento de classificadores lineares, não-lineares e regressores. O SVM realiza a minimização do Risco Estrutural, um princípio que limita *overfitting* definindo um equilíbrio entre a complexidade do modelo e seu risco empírico. Isto resulta na solução de margem máxima, o que faz com que o SVM tenha uma capacidade maior de generalização e robustez melhorada na presença de ruído, comparado com outros métodos de aprendizagem de máquina (Smola et al., 2000).

O classificador SVM atibui a saída $y \in \{+1, -1\}$ para um vetor de entrada $\bar{x}$ de acordo com,

$$f(x) = \bar{w}^T \phi(\bar{x}) + b \tag{C.8}$$

em que $\phi(\bar{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^H$ é uma função não-linear que mapeia o vetor de entrada $\bar{x}$ em um espaço de característica de uma dimensionalidade maior. O vetor $\bar{w}$ representa hiperplanos de separação em tal espaço e b é um viés em relação a origem. A razão do SVM ter boa capacidade de generalização é que a sua formulação envolve a minimização conjunta de ambos os riscos empíricos e estrutural. A Minimização do risco estrutural é equivalente à minimização da norma do vetor $\bar{w}$. Assim, a solução para o SVM é dada pela minimização do seguinte problema quadrático:

$$\min_{\bar{w},b,\xi_i} (\frac{1}{2}||\bar{w}||^2 + C \sum_{i=1}^{n} \xi_i) \tag{C.9}$$

sujeito a:

$$y_i(\bar{w}^T \phi(\bar{x}_i) + b) \geq 1 - \xi_i; \forall i = 1, 2, ..., n$$

$$\xi_i \geq 0; \forall i = 1, 2, ..., n. \tag{C.10}$$

onde $\bar{x}_i \in \mathbb{R}^d (i = 1, 2, ..., n)$ são os vetores de treinamento com saídas $y_i \in \{+1, -1\}$. A variável ξ_i, denominada variável livre Haykin (2009), mede o desvio de um ponto dado da condição ideal de separabilidade e o conjunto C representa o compromisso entre a minimização dos Riscos Empírico e Estrutural (Hearst et al., 1998). Este problema pode ser resolvido usando-se os multiplicadores de Lagrange α_i dados por:

$$\max_{\alpha_i} \left(\sum_{i=i}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \phi^T(\bar{x}_i) \phi(\bar{x}_j) \right) \tag{C.11}$$

O Limite ótimo de decisão w é dado por:

$$w = \sum_{i=1}^{n} \alpha_i y_i \phi(\bar{x}_i) \tag{C.12}$$

Somente os vetores de treinamento com multiplicadores de Lagrange $\alpha_i \neq 0$ associados irão contribuir para determinar o limite de decisão. Dessa forma, eles recebem no nome de *vetores de suporte*. A função de mapeamento $\phi(\bar{x})$ é raramente conhecida explicitamente. Entretanto, o problema de otimização na equação (C.11) é definido em termos do produto interno

$(\phi^T(\bar{x}_i)\phi(\bar{x}_i))$, que pode ser analisado utilizando-se a função Kernel de Mercer $K(.,.)$. O teorema de Mercer (Mercer, 1909) afirma que a função de mapeamento ϕ e a função $K(\bar{x}_i, \bar{x}_j)$ é positiva semidefinida. Por meio do então chamado "truque"do Kernel, a saída do SVM segue a expressão dada abaixo,

$$f(\bar{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\bar{x}_i, \bar{x}) + b \tag{C.13}$$

C.3.1 Hiperplano ótimo para padrões linearmente separávies

Supondo-se um problema com duas classes de padrões linearmente separáveis, o objetivo é demonstrar a existência de um hiperplano capaz de separar tais classes de forma maximizada. Dada uma amostra de treinamento $\Gamma = \{(\bar{x}_i, d_i)\}_{i=1}^{N}$ - na qual $\bar{x}_i$ é o padrão de entrada para o $i-$ésimo exemplo, d_i é a resposta desejada correspondente, assumindo-se ainda que esses padrões representam duas classes distintas lineamente separáveis - a superfície de decisão na forma de um hiperplano que realiza esta separação é dada por:

$$\bar{w}^T \bar{x} + b = 0 \tag{C.14}$$

sendo $\bar{x}$ o vetor de entrada; $\bar{w}$ é um vetor de pesos ajustáveis, e b é o viés em relação à origem.

$$\begin{aligned} \bar{w}^T \bar{x}_i + b \geq 0 \text{ para } d_i = +1 \\ \bar{w}^T \bar{x}_i + b < 0 \text{ para } d_i = -1 \end{aligned} \tag{C.15}$$

A margem de separação representada por ρ é a distância entre o hiperplano definido na equação (C.14) e o ponto de dado mais próximo, isto para um vetor de peso $\bar{w}$ e viés b específicos. O objetivo do SVM é encontrar o hiperplano particular para o qual a margem de separação ρ é máxima. Sob essas condições, a superfície encontrada é chamada de ótima (Urena et al., 2012). Na figura C.2 ilustra-se a geometria de um hiperplano ótimo para espaço bidimensional. Considerando ainda que $\bar{w}_0$ e b_0 representam os valores ótimos do vetor peso e do viés, respectivamente. O hiperplano ótimo representa uma superfície de decisão linear multidimensional no espaço de entrada e é definido abaixo,

$$w_0^T x + b_0 = 0 \tag{C.16}$$

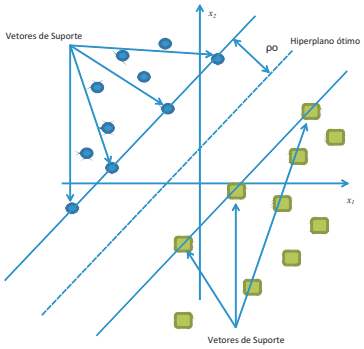


Fig. C.2: Hiperplano ótimo para padrões linearmente separáveis.

Dessa forma, a distorção algébrica da distância de $\bar{x}$ até o hiperplano é dada por:

$$g(x) = w_0^T x + b_0 \tag{C.17}$$

assim, pode-se expressar $\bar{x}$ de uma outra maneira, conforme abaixo,

$$\bar{x} = \bar{x}_p + r\frac{\bar{w}_0}{\|\bar{w}_0\|} \tag{C.18}$$

onde $\bar{x}_p$ é a projeção normal de $\bar{x}$ sobre o hiperplano ótimo, e r é a distância algébrica desejada; r é positivo, se $\bar{x}$ estiver no lado positivo do hiperplano ótimo, e negativo, se $\bar{x}$ estiver no lado negativo. Uma vez que por definição $g(\bar{x}) = 0$, então resulta que:

$$g(\bar{x}) = \bar{w}_0^T \bar{x} + b_0 = r\,\|w_0\| \tag{C.19}$$

ou

$$r = \frac{g(\bar{x})}{\|w_0\|} \tag{C.20}$$

Assim, o hiperplano ótimo definido na equação (C.14) é único no sentido de que o vetor peso w_0 fornece a máxima separação possível entre exemplos positivos e negativos. Esta condição é alcançada minimizando-se a norma euclidiana do vetor de peso w. A margem de separação ρ é, então, definida pela equação (C.21). Na figura C.3 mostra-se a distância algébrica de um ponto até o hiperplano ótimo para um caso bidimensional.

$$\rho = \frac{2}{\|w_0\|} \tag{C.21}$$

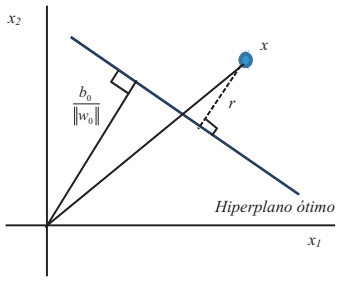


Fig. C.3: Distância algébrica de um ponto até o hiperplano ótimo para um caso bidimensional.

C.3.2 Hiperplano ótimo para padrões não separávies linearmente

Considerando agora que os padrões são não-separáveis linermente, então, para este novo conjunto de treinamento, não é possível elaborar um hiperplano de separação sem se defrontar com erros de classificação. Contudo, pode-se encontrar um hiperplano ótimo que minimize a probabilidade de erro de classificação; neste caso, a probabilidade é calculada como a média sobre o conjunto de treinamento. Para efeito de generalização, uma variável escalar e não negativa $\bar{x}_i = \{\bar{x}_{i1}, \bar{x}_{i2}, ...\bar{x}_{il}\}$ é inserida na equação que define o hiperplano de separação, dado por:

$$(\bar{w}^T \bar{x}_i + b) \geq 1 - \xi_i \tag{C.22}$$

sendo $i = 1, 2, ..., N$.

As variáveis livres ξ medem o desvio de cada amostra de sua condição ideal de separabilidade de padrões. Para $0 < \xi \leq 1$, o ponto de dado encontra-se dentro da região de separação, mas no lado correto da superfície de decisão. Para $\xi > 1$, a amostra está localizada no lado incorreto do hiperplano de separação. Os vetores de suporte são as amostras que estão mais próximas do hiperplano (Vapnik, 1968). Ressalta-se que, se um exemplo $\xi_i > 0$ for deixado fora do conjunto de treinamento, a superfície de decisão não muda. Desse modo, os vetores de suporte são definidos do mesmo modo tanto para o caso linearmente separável, como para o caso não-separável linearmente. Desta forma, o objetivo é encontrar um hiperplano de separação para o qual o erro de classificação, como média sobre o conjunto de treinamento, é minimizado. Isto pode ser feito minimizando-se o funcional dado por:

$$\Phi(\xi) = \sum_{i=1}^{N} I(\xi_i) - 1 \tag{C.23}$$

Em relação ao vetor peso w, a restrição da equação (C.22) é a restrição em relação a $\|\bar{w}\|^2$,

dado por:

$$\|\bar{w}\|^2 \leq \frac{1}{\rho} \tag{C.24}$$

A função $I(\xi)$ é uma função indicadora e definida por:

$$I(\xi) = \begin{cases} 0, \text{ se } \xi \leq 0 \\ 1, \text{caso contrário} \end{cases} \tag{C.25}$$

A minimização de $\Phi(\xi)$ em relação a w é um problema de otimização que pertence a uma classe de problemas *NP completos*. Há muitos problemas computacionais que aparecem na prática, para os quais nenhum algoritmo eficiente pode ser encontrado. Diz-se que muitos, se não todos esses problemas aparentemente intratáveis, pertencem a uma classe de problemas *NP completos*, onde o termo *NP* significa *Não deterministicamente Polinomial* (Haykin, 2009). Para tratar essa questão deve ser feita uma aproximação, a qual é dada por:

$$\Phi(\xi) = \sum_{i=1}^{N} \xi_i \tag{C.26}$$

Além disso, deve-se fazer com que o funcional seja minimizado em relação ao vetor peso w, conforme abaixo,

$$\Phi(\bar{w}, \xi) = \frac{1}{2}\bar{w}^T\bar{w} + C\sum_{i=1}^{N} \xi_i \tag{C.27}$$

A minimização de $\bar{w}$ está relacionada à minimização da dimensão VC. Já o segundo termo da equação (C.27) é equivalente ao limite superior para o erro de classificação. O parâmetro C pode ser considerado como um parâmetro de regulação, isto é, controla o compromisso entre a complexidade da máquina e o número de erros de treinamento.

C.3.3 Funções Kernel

A superfície de decisão do SVM, que no espaço de caracterísitcas é sempre linear, normalmente, é não linear no espaço de entrada. Como visto antriormente, a ideia de uma Máquina de Vetor de Suporte depende de duas operações matemáticas: 1. Mapeamento não-linear de um vetor de entrada para um espaço de características de alta dimensionalidade, que é oculto da entrada e da saída; 2. É necessário construir um hiperplano ótimo para separar as características descobertas no primeiro passo. Para a elaboração deste hiperplano ótimo, necessita-se de uma função Kernel, ou núcleo do produto interno. Um Kernel é uma função que recebe dois pontos x_i e x_j do espaço de entradas e calcula o produto escalar desses dados no espaço de

características, como segue,

$$k(\bar{x}_i, \bar{x}_j) = \Phi^T(\bar{x}_i) \cdot \Phi(\bar{x}_j) \tag{C.28}$$

Para garantir a convexidade do problema de otimização, de modo que o Kernel apresente mapeamento no qual seja possível o cálculo de produto escalares, deve-se utilizar uma função Kernel que siga as condições estabelecidas pelo teorema de Mercer (De Gang et al., 2008; Mercer, 1909). Os Kernels que satisfazem as condições de Mercer são caracterizados por darem origem a matrizes positivas semi-definidas k, em que cada elemento k_{ij} é definido por $k_{ij} = k(\bar{x}_i, \bar{x}_j)$, $\forall i, j = 1, 2, ..., n$. Uma vez que o mapeamento do SVM é realizado por uma função Kernel, e não diretamente por $\Phi(x)$, nem sempre é possível saber exatamente qual mapeamento é efetivamente realizado, pois as funções Kernel realizam um mapeamento implicito. Na tabela C.1 mostram-se algumas funções comumente utilizadas como funções Kernel. A expansão do núcleo do produto interno $K(\bar{x}_i, \bar{x}_j)$, na equação (C.28), permite encontrar uma superfície de decisão que é não-linear no espaço de entrada, mas cuja imagem no espaço de característica é linear (Haykin, 2009).

Tab. C.1: Funções Kernel do SVM

Kernel	Função
Polinomial	$(\bar{x}^T\bar{x}_i + 1)^p$
RBF Kernel	$exp(-\frac{1}{2\sigma^2}\left\|\bar{x} - \bar{x}_i\right\|^2)$
Perceptron	$tanh(\beta_0\bar{x}^T\bar{x}_i + \beta_1)$

Para as funções Kernel utilizadas, tem-se as seguintes restrições:

- No kernel Polinomial, o parâmetro p é especificada pelo usuário a priori.
- No kernel RBF, o parâmetro σ^2 é comum a todos os núcleos.
- No perceptron, o teorema de Mercer é satisfeito apenas para alguns valores de β_0, β_1.

C.4 Sistema de reconhecimento de voz utilizando o SVM

O diagrama de blocos do sistema de reconhecimento de voz com SVM utilizado neste trabalho é apresentado na figura C.4. Vale ressaltar que o pré-processamento do sinal de voz, a geração dos parâmetros devidamente codificados seguem rigorosamente os procedimentos apresentados nos Capítulos (3) e (4) desta tese.

Foram utilizadas as equações (3.2) e (3.3) para a geração das matrizes bidimensionais C_{kn}, a partir das quais foram calculadas as matrizes de média e variâncias dadas nas equações (3.5) e (3.6).

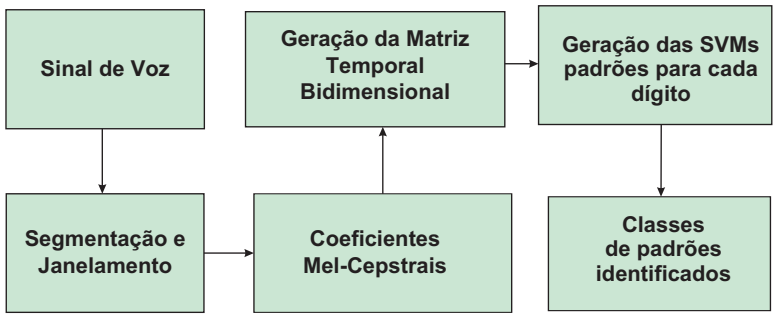


Fig. C.4: Diagrama de blocos do sistema de reconhecimento de voz com SVM.

C.4.1 Geração das Máquinas

As matrizes de média, dada na equação (3.5), e de variância, dada na equação (3.6), foram transformadas em dois vetores, denominados de $CMed$ e $CVar$, dados por:

$$\begin{aligned} CMed_i^j &= \langle CM_{11}^0, CM_{12}^0, ..., CM_{1N}^0, CM_{21}^0, CM_{22}^0, ..., CM_{2N}^0, ..., CM_{KN}^0, CM_{11}^1, CM_{12}^1, ..., CM_{1N}^1, \\ &\quad CM_{21}^1, CM_{22}^1, CM_{2N}^1, ..., CM_{KN}^1, ..., CM_{11}^J, CM_{12}^J, ..., CM_{1N}^J, CM_{21}^J, CM_{22}^J, ..., CM_{2N}^J, ..., CM_{KN}^J \rangle \end{aligned} \tag{C.29}$$

$$\begin{aligned} CVar_i^j &= \langle CV_{11}^0, CV_{12}^0, ..., CV_{1N}^0, CV_{21}^0, CV_{22}^0, ..., CV_{2N}^0, ..., CV_{KN}^0, CV_{11}^1, CV_{12}^1, ..., CV_{1N}^1, \\ &\quad CV_{21}^1, CV_{22}^1, CV_{2N}^1, ..., CV_{KN}^1, ..., CV_{11}^J, CV_{12}^J, ..., CV_{1N}^J, CV_{21}^J, CV_{22}^J, ..., CV_{2N}^J, ..., CV_{KN}^J \rangle \end{aligned} \tag{C.30}$$

em que j é a classe (padrão) a ser reconhecida, e i é a posição do elemento da matriz bidimensional no vetor.

Cada classe j é representada por $(K \times N)$ elementos no vetor de médias e $(K \times N)$ elementos no vetor de variâncias, de acordo com as equações (C.29) e (C.30). Os elementos dos vetores de média e variâncias são distribuídos ao longo do espaço euclidiano, sendo adotados neste trabalho, os elementos do vetor de médias como o eixo das abscissas e os elementos do vetor de variâncias como eixo das ordenadas. Desse modo cada par de elementos forma um ponto no espaço euclidiano. A relação da função euclidiana é dada por:

$$\Omega = f\left(\left[CMed_i^j; CVar_i^j\right], \omega\right) \tag{C.31}$$

sendo Ω a resposta real produzida pela máquina de aprendizagem associada com os pares de entradas de médias e variâncias, e ω é um conjunto de parâmetros livres chamados de pesos de ponderação, selecionados dos parâmetros relacionados as classes a serem reconhecidas. Na figura C.5 mostra-se o espaço euclidiano com os pontos representativos das j classes devidamente

distribuídos para a classificação do SVM.

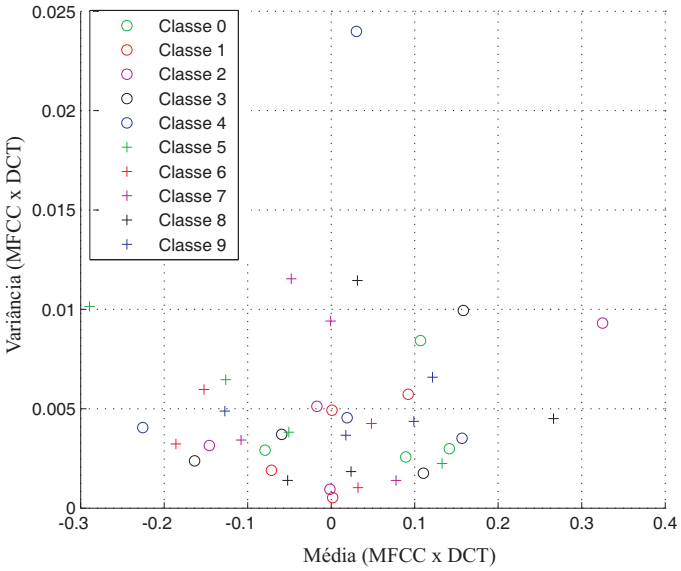


Fig. C.5: Distribuição das classes no espaço euclidiano.

C.4.2 Treinamento

Após a realização da codificação dos parâmetros de voz e a geração das matrizes bidimensionais dadas nas equações (3.5) e (3.6), os modelos j foram divididos em dez Máquinas especialistas SVM, treinadas no modo um contra todos com kernel polinomial de ordem 2 e kernel RBF com $\sigma = 0.03$. Nessa etapa, foram realizados os mesmos procedimentos descritos na seção (4.2), subitem (4.2.1). Os resultados dos testes de validação utilizando SVM foram apresentados nas Tabelas 4.3 e 4.4, respectivamente.

Para efeito de ilustração da obtenção das máquinas especialistas, nas figuras C.6 e C.7 mostram-se as máquinas especialistas da classe zero para a função polinomial de ordem $p = 2$ e função RBF com $\sigma = 0,03$, respectivamente.

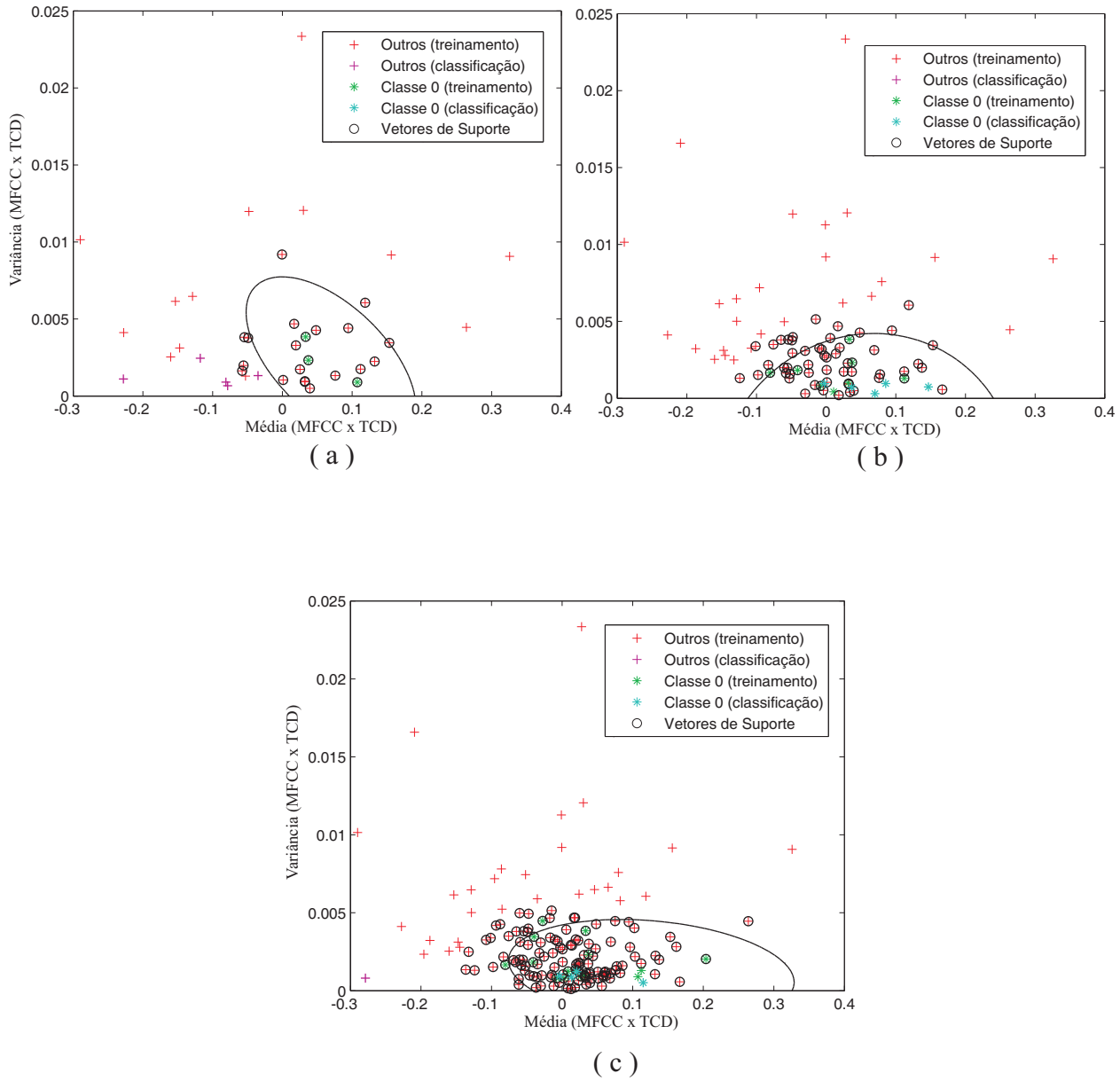


Fig. C.6: Classe 0 $\times$ todos, com função polinomial de ordem $p = 2$: (a) Ordem da matriz $K = N = 2$, (b) Ordem da matriz $K = N = 3$ e (c) Ordem da matriz $K = N = 4$.

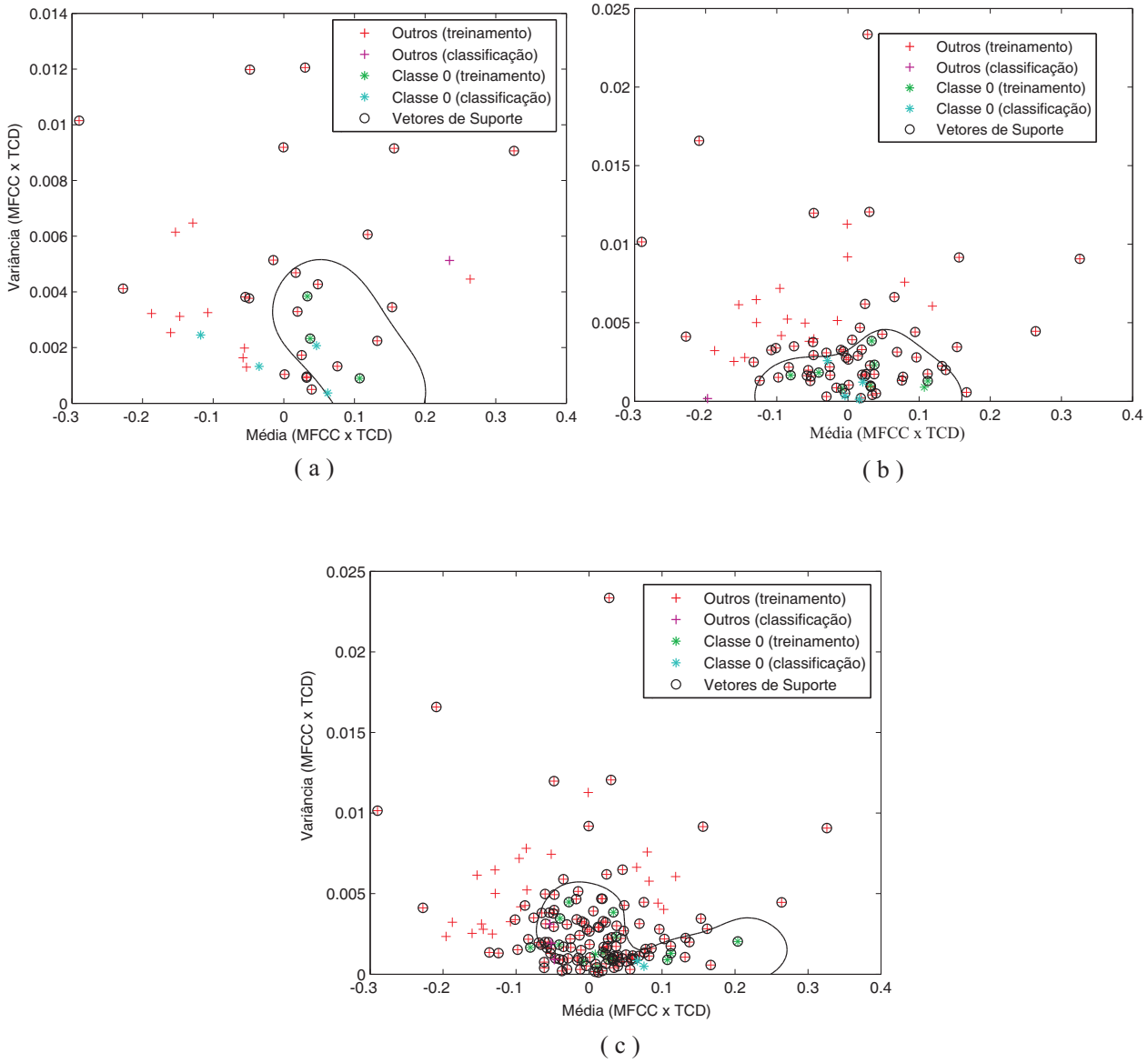


Fig. C.7: Classe 0 × todos, com função RBF com $\sigma = 0,03$: (a) Ordem da matriz $K = N = 2$, (b) Ordem da matriz $K = N = 3$ e (c) Ordem da matriz $K = N = 4$.

Apêndice D

Gaussian Mixture Models-GMM

D.1 Introdução

Um Modelo de misturas gaussianas (*Gaussian Mixture Model*- GMM) é uma função densidade de probabilidade paramétrica representada como uma soma de componentes de densidades gaussianas ponderadas (Filho et al., 2013; Reynolds, 1995), denominada mistura. As GMMs são comumente utilizadas como modelos paramétricos da distribuição de probabilidade de medidas contínuas ou características de sistemas biométricos, tais como o trato vocal relacionado às características espectrais em sistema de reconhecimento de voz. A equação (D.1) é uma GMM com K densidades Gaussianas componentes da mistura dada por:

$$p\left(\bar{\mathbf{x}}/\Theta\right) = \sum_{i=1}^{K} \pi_i \, p\left(\bar{\mathbf{x}}/\bar{\mu}_i, \Sigma_i\right) \tag{D.1}$$

em que $\bar{\mathbf{x}}$ é um vetor de dados, π_i são pesos das misturas, e $p\left(\bar{\mathbf{x}}/\mu_i, \Sigma_i\right)$, $i = 1, 2, \ldots, K$, são as densidades gaussianas componentes da mistura. Cada densidade componente da mistura é uma função densidade de probabilidade gaussiana de dimensão D com a seguinte representação,

$$p\left(\bar{\mathbf{x}}/\mu_i, \Sigma_i\right) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} exp\left\{-\frac{1}{2}\left(\bar{\mathbf{x}} - \bar{\mu}_i\right)' \Sigma_i^{-1}\left(\bar{\mathbf{x}} - \bar{\mu}_i\right)\right\} \tag{D.2}$$

com vetor média μ_i, matriz de covariâncica Σ_i. Os pesos da mistura satisfazem as restrições de probabilidades, $\pi_i \in [0, 1]$ e $\sum_{i=1}^{K} \pi_i = 1$. O modelo completo de misturas de Gaussianas é parametrizado por vetores médias, matrizes de covariância e pesos de misturas para todas as densidades componentes da mistura. Estes parâmetros são representados por:

$$\Theta = \left\{\pi_i, \mu_i, \Sigma_i\right\}, \; i = 1, 2, ..., K. \tag{D.3}$$

Há diversas variantes da GMM mostrada na equação (D.1). A matriz de covariância pode ser de posto completo ou uma matriz diagonal. Parâmetros podem ser compartilhados ou vinculados entre as componentes gausssianas, de tal forma que tenham uma matriz de covariância comum para todas as componentes. A escolha da configuração do modelo (número de componentes, matriz de covariância de posto completo ou diagonal e parâmetros compartilhados) é, geralmente, determinada pela quantidade total de dados disponíveis para estimar os parâmetros da GMM e do tipo de aplicação (Bishop, 2007; Reynolds, 1995).

Também é importante notar que, como as componentes Gaussianas agem em conjunto para modelar as densidades características, matrizes de covariâncias de postos completos não são necessárias, mesmo se as características não são estatisticamente independentes. A combinação linear da diagonal da matriz de covariância, base da Gaussiana, é capaz de modelar as correlações entre os elementos do vetor de característica. O efeito de se usar um conjunto de M Gaussianas de matrizes de covariância de posto completo pode ser igualmente obtido pelo uso de Gaussianas com grandes conjuntos de matrizes de covariância diagonal.

GMMs são aplicadas em sistema de reconhecimento de voz, principalmente, devido a sua capacidade de representar uma grande classe de distribuições. Um dos principais atributos da GMM é sua habilidade para formar aproximações suaves para formas de densidades de probabilidades arbitrárias. O modelo de gaussiana unimodal clássico representa distribuições de características por uma posição (vetor média) e uma forma elíptica (matriz de covariância), e uma quantização de vetor ou "*nearest neighbor model*"que representa distribuições de características por um conjunto discreto de um modelo de características. Uma GMM atua como híbrido entre estes dois modelos usando um conjuto discreto de funções gaussianas, cada um com sua própria média e matriz de covariância, para permitir a melhor capacidade de modelamento.

O uso da GMM para representar distribuições de caracteríticas em sistema de reconhecimento de voz também deve ser motivado pela noção intuitiva que a componente densidade individual deve modelar algum conjunto subjacente de classes ocultas. Por exemplo, no reconhecimento de locutor, é razoável assumir que o espaço acústico das características espectrais relacionadas correspondem a eventos fonéticos do locutor a ser reconhecido, tais como vogais, nasalização ou fricativos. Essas classes acústicas refletem alguma depedência geral do trato vocal do locutor, que pode ser útil para identificá-lo (Reynolds, 1995). A forma espectral da $i-$ésima classe acústica pode, por sua vez, ser representada pelas médias μ_i da $i-$ésima densidade componente, e as variações da forma espectral média podem ser representadas pela matriz de covariância Σ_i.

D.2 Estimação de Parâmetros por Máxima Verossimilhança

Dados um vetor de treinamento e uma configuração GMM, deseja-se estimar os parâmetros de uma GMM, λ, que, em algum sentido, melhor aproxima-se da distribuição dos vetores de características de treinamento. Há diversas técnicas disponíveis para a estimação dos parâmetros de uma GMM. Contudo, o método mais popular e estabelecido é a estimação por máxima verossimilhança. A contribuição do método citado na estimação é encontrar os parâmetros do modelo que maximizam a verossimilhança da GMM com os dados de treinamento. Para uma sequência de T vetores de treinamento $\mathbf{X} = \{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_T\}$, a verossimilhança, assumindo a independência entre os vetores, pode ser dada por:

$$p\left(\mathbf{X}/\Theta\right) = \prod_{t=1}^{T} p\left(\bar{\mathbf{x}}_T/\Theta\right) \tag{D.4}$$

Infelizmente esta expressão é uma função não linear dos parâmetros λ e a sua maximização direta não é possível. Todavia, estimação dos parâmetros por máxima verossimilhança pode ser obtida iterativamente, utilizando-se algoritmo de maximização da esperança (*Expectation-maximization* - EM) (Dempster et al., 1977; Reynolds, 1995). A idéia básica do algoritmo EM é começar com um modelo inicial λ, para estimar um novo modelo $\bar{\lambda}$, tal que $p\left(\mathbf{X}/\bar{\lambda}\right) \geq p\left(\mathbf{X}/\lambda\right)$. O novo modelo, então, torna-se o modelo inicial para o novo modelo da próxima iteração. Esse procedimento se repete até que um critério de parada previamente estabelecido seja alcançado. Em cada iteração do EM, as seguintes re-estimações são utilizadas para garantir o aumento monotônico no valor do modelo de verossimilhança:

Peso das misturas

$$\bar{\pi}_i = \frac{1}{T} \sum_{t=1}^{T} Pr\left(\pi_i/\bar{\mathbf{x}}_t, \Theta\right) \tag{D.5}$$

Médias

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} Pr\left(\pi_i/\bar{\mathbf{x}}_t, \Theta\right) \bar{\mathbf{x}}_t}{\sum_{t=1}^{T} Pr\left(\pi_i/\bar{\mathbf{x}}_t, \Theta\right)} \tag{D.6}$$

Variâncias-Diagonal da matriz de covariâncias

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} Pr\left(\pi_i/\bar{\mathbf{x}}_t, \Theta\right) x_t^2}{\sum_{t=1}^{T} Pr\left(\pi_i/\bar{\mathbf{x}}_t, \Theta\right)} - \bar{\mu}_i^2 \tag{D.7}$$

onde σ_i, x_t e μ_i referem-se a elementos arbitrários dos vetores $\bar{\sigma}_i^2$, $\bar{\mathbf{x}}_t$ e $\bar{\mu}_i$, respectivamente.

A probabilidade a posteriori para a componente i é dada por:

$$Pr\left(\pi_i/\bar{\mathbf{x}}_t, \Theta\right) = \frac{\pi_i \; g\left(\bar{\mathbf{x}}_t/\bar{\mu}_i, \Sigma_i\right)}{\sum_{k=1}^{M} \pi_k \; g\left(\bar{\mathbf{x}}_t/\bar{\mu}_k, \Sigma_k\right)} \tag{D.8}$$

Dois fatores críticos no treinamento de modelos para reconhecimento de voz utilizando GMM são o número de componentes da mistura e a inicialização dos parâmetros dos modelos para o algoritmo EM (Dempster et al., 1977; Reynolds, 1995).

D.3 Aplicação em reconhecimento de voz

Para a classificação de padrões de voz, utilizados neste trabalho, adotaram-se os seguintes procedimentos; dado o padrão $j \in J$, tomadas m observações de cada padrão j, foi determinado um conjunto de parâmetros $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_K\}$. O objetivo é encontrar o modelo de cada padrão que tem a máxima probabilidade a *posteriori* para uma dada sequência de observações. Assim,

$$j^* = \arg\max_{1 \le k \le K} Pr\left(\lambda_k/\mathbf{X}\right) = \arg\max_{1 \le k \le K} \frac{p\left(\mathbf{X}/\lambda_k\right) Pr\left(\lambda_k\right)}{p\left(x\right)} \tag{D.9}$$

onde a segunda equação é devida à regra de Bayes. Assumindo-se que os padrões são equiprováveis, isto é, $Pr\left(\lambda_k\right) = \frac{1}{J}$ e observando-se que $p(\mathbf{X})$ é a mesma para todos os padrões, a regra de classificação simplificada é dada por:

$$j^* = \arg\max_{1 \le k \le K} p\left(\mathbf{X}/\lambda_k\right) \tag{D.10}$$

Utilizando-se o logaritmo e considerando-se a independência entre as observações, tem-se que:

$$j^* = \arg\max_{1 \le k \le K} \sum_{t=1}^{T} log \; p\left(\bar{\mathbf{x}}_t/\lambda_k\right) \tag{D.11}$$

em que $p\left(\bar{\mathbf{x}}_t/\lambda_k\right)$ é dado na equação (D.1).

O diagrama de blocos do sistema de reconhecimento de voz com GMM-EM utilizado para comparação com a metodologia proposta é apresentado na figura D.1.

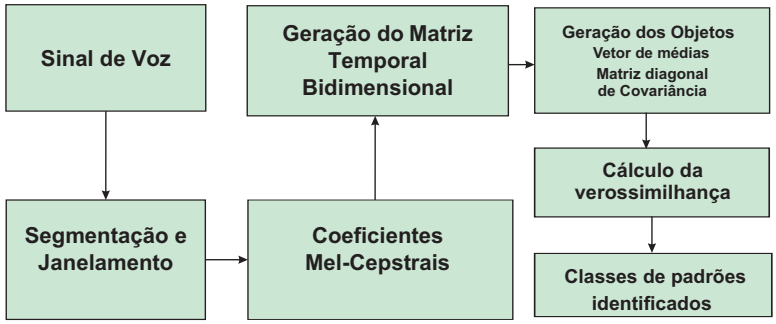


Fig. D.1: Diagrama de blocos do sistema de reconhecimento com GMM-EM.

As matrizes de média e variâncias, dadas nas equações (3.5) e (3.6), respectivamente, foram transformadas, para cada padrão j, em um vetor denominado $CMed^j$ e uma matriz de diagonal de covariância, denominada $CVar^j$ dadas por:

$$CMed^j = \left\langle CM_{11}^j, CM_{12}^j, \ldots, CM_{1N}^j, CM_{21}^j, CM_{22}^j, \ldots, CM_{2N}^j, \ldots, CM_{KN}^j \right\rangle \tag{D.12}$$

$$CVar^j = \begin{pmatrix} CV_{11}^j & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & CV_{12}^j & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ddots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & CV_{1N}^j & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & \ldots & CV_{21}^j & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & 0 & \ldots & CV_{22}^j & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & \ddots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & CV_{2N}^j & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & CV_{KN}^j \end{pmatrix} \tag{D.13}$$

Após a codificação, utilizou-se uma mistura com a quatidade de parâmetros calculados; assim, na matriz (2×2), utilizou-se uma mistura com quatro gaussianas; na matriz (3×3), nove gaussianas; e na matriz (4×4), dezesseis gaussianas. A GMM foi determinada com $k = \{1, 2, \ldots, K\}$ componentes com probabilidade a priori dada por π_k, usando-se amostras da matriz bidimencional C_{kn} determinada pela equação (3.3) com a $k-$ésima distribuição Gaussiana dada por:

$$p\left(\lambda^j / CMed_k^j, CVar_k^j\right) \tag{D.14}$$

em que k representa tanto o índice do elemento do vetor $CMed^j$, quanto o elemento da diagonal da matriz $CVar^j$. Assim, a GMM é completamente especificada pelos parâmetros $\Theta = \{w_k, CMed^j, CVar^j; k = 1, 2, \ldots, K\}$. Dado o conjunto de parâmetros determinado pela matriz C_{kn}, dada na equação (3.3), definiu-se λ^j como um vetor formado pelo elementos desta matriz, distribuídos da seguinte forma:

$$\lambda^j = \left\langle c_{11}^j, c_{12}^j, \ldots, c_{1N}^j, c_{21}^j, c_{22}^j, \ldots, c_{2N}^j, \ldots, c_{KN}^j \right\rangle \tag{D.15}$$

logo a verossimilhança $P(\lambda^j/\Theta)$ é obtida por:

$$P\left(\lambda^j/\Theta\right) = \sum_{k=1}^{K} \pi_k \, p\left(\lambda_k^j/CMed_k^j, CVar_k^j\right) \tag{D.16}$$

sendo

$$p\left(\lambda_k^j/CMed_k^j, CVar_k^j\right) = \frac{1}{\sqrt{(2\pi)^2 \, det\left(CVar^j\right)}} \left[-\frac{1}{2}\left(\lambda_k^j - CMed_k^j\right)' CVar_k^j \left(\lambda_k^j - CMed_k^j\right)\right] \tag{D.17}$$

O treinamento da GMM para a base de dados $j \in J$ foi feito através da maximização do log verossimilhança dado pela equação,

$$j^* = \frac{1}{(K \times N)} \sum_{i=1}^{(K \times N)} log \sum_{k=1}^{(K \times N)} \pi_k \, p\left(\lambda_i/CMed_k^j, CVar_k^j\right) \tag{D.18}$$

Para o cálculo do log-verossimilhança, utilizou-se o algoritmo de maximização da esperança (*Expectation Maximization*-EM). Para efeito de comparação de dempenho entre a metodologia proposta(IMSR) e o GMM-EM, foram apresentados ao GMM-EM os mesmos parâmetros na etapa de validação, cujo banco de dados para o treinamento e o banco de testes são descritos no item (4.2.1). Os parâmetros de médias de variâncias foram calculados conforme as equações (3.5) e (3.6). Nessa etapa, foram realizados os mesmos procedimentos descritos na seção (4.2), subitem (4.2.1). Os resultados dos testes de validação utilizando GMM-EM foram apresentados nas tabelas 4.3 e 4.4, respectivamente.