Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Faculdade de Engenharia

Diana Alejandra Bonilla Cardona

Reconhecimento Automático dos Fonemas da Língua Portuguesa Usando Modelos de Composição de Redes Neurais Especialistas

Rio de Janeiro
2016

Diana Alejandra Bonilla Cardona

Reconhecimento Automático dos Fonemas da Língua Portuguesa Usando Modelos de Composição de Redes Neurais Especialistas

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Engenharia Eletrônica, da Universidade do Estado do Rio de Janeiro. Área de concentração: Sistemas Inteligentes e Automação.

Orientadora: Profª. Drª. Nadia Nedjah

Coorientadora: Profª; Drª. Luiza de Macedo Mourelle

Rio de Janeiro

2016

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

B715 Bonilla Cardona, Diana Alejandra.
Reconhecimento automático dos fonemas da língua portuguesa usando modelos de composição de redes neurais especialistas / Diana Alejandra Bonilla Cardona. – 2016.
148f.

Orientador: Nadia Nedjah.
Coorientador: Luiza de Macedo Mourelle.
Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia.

1. Engenharia Eletrônica. 2. Redes neurais (Computação) - Dissertações. 3. Reconhecimento automático da voz - Dissertações. I. Nedjah, Nadia. II. Mourelle, Luiza de Macedo. III. Universidade do Estado do Rio de Janeiro. IV. Título.

CDU 004.032.26

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta dissertação.

_____ _____

Assinatura Data

Diana Alejandra Bonilla Cardona

Reconhecimento Automático dos Fonemas da Língua Portuguesa Usando Modelos de Composição de Redes Neurais Especialistas

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Engenharia Eletrônica, da Universidade do Estado do Rio de Janeiro. Área de concentração: Sistemas Inteligentes e Automação.

Aprovado em: 30 de agosto de 2016.

Banca Examinadora:

Prof.ª Dr.ª Nadia Nedjah (Orientadora)
Faculdade de Engenharia - UERJ

Prof.ª Dr.ª Luiza de Macedo Mourelle (Coorientadora)
Faculdade de Engenharia - UERJ

Prof. Dr. Rogério Andrade Flauzino
Escola de Engenharia de São Carlos, USP

Prof. Dr. José Gabriel Rodriguez Carneiro Gomes
Programa de Engenharia Elétrica, COPPE/UFRJ

Prof. Dr. Jayme Luiz Szwarcfiter
Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ

Rio de Janeiro
2016

DEDICATÓRIA

Dedico este trabalho a meu amado esposo Nicolás por me incentivar a cursar o mestrado, por sua motivação e apoio constante neste processo, sem você nada disto teria acontecido, a meus pais Yolanda e Germán por acreditar em mim e ter me formado corretamente en todos os aspectos da vida. A meu irmão Eduardo por apoiar à família durante minha ausência e ser aquele anjo da guarda que cuida de cada detalhe. A meu irmãozinho, Juan Pablo, por criar em mim a motivação de te dar um bom exemplo. A minha família política Solange, Gabino e Vladimir, agradeço seu apoio constante e por cuidar sempre de Nicolás e eu. A minhas tias, avós e primos, por acreditar em minhas capacidades e sempre me levar nas suas orações. Sou abençoada por ter a uma família tão amorosa!

AGRADECIMENTOS

Agradeço a toda minha família pelo apoio, ánimo e o carinho durante esta etapa da minha vida.

Agradeço a minha orientadora Nadia Nedjah por sua paciência, compreensão e animo, além da disponibilidade e valiosas sugestões para realizar este trabalho. Agradeço à professora Luiza de Macedo Mourelle pelas aulas e pelas conversas para contextualizar-me no Brasil. Aos professores do PEL-UERJ pelas lições aprendidas nestes dois anos de curso.

Agradeço aos colegas de mestrado Alan, Alexandre, Joelmir, Luneque, Tarso e Yuri. Obrigada pelas conversas relacionadas com o projeto, com a língua portuguesa e com o Brasil.

Agradeço a CAPES pelo apoio econômico brindado durante todo o curso de mestrado, sem o qual não poderia ter desenvolvido este trabalho em tempo integral.

"... A linguagem é uma fonte de mal-entendidos..."

O pequeno príncipe, Antoine de Saint-Exupéry.

RESUMO

BONILLA C, D A. *Reconhecimento Automático dos Fonemas do Língua Portuguesa Usando Modelos de Composição de Redes Neurais Especialistas.* 2016. 130f. Dissertação (Mestrado em Engenharia Eletrônica) – Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2016.

Os sistemas de reconhecimento automático de fala para a detecção de fonemas proporcionam vantagens para o reconhecimento *online* de fala representada por um sinal de som. Os estudos na área de inteligência computacional têm permitido associar as vantagens do processamento paralelo de informação à distribuição da carga computacional visando simplificar o uso de modelos complexos de sistemas. O desenvolvimento de um sistema de reconhecimento automático de fala envolve vários processos que abrangem várias áreas da pesquisa, tais como a linguística, o processamento de sinais e a inteligência computacional. Nesta dissertação, o processo inicia-se com o pré-processamento do sinal de fala fornecido na entrada do sistema, visando extrair e representar de modo mais sucinto as características principais relacionadas ao sinal em um dado instante de tempo. Inspirado pelo preceito que recomenda "dividir para conquistar", um modelo de composição de redes neurais especialistas é explorado, permitindo dividir o espaço de decisão do problema complexo de reconhecimento de fala para que cada especialista cuide somente de uma área bem delimitada deste espaço de decisão. Vale ressaltar que cada especialista incluído no modelo composto precisa tratar e tomar uma decisão a respeito de cada uma das amostras pré-processadas. O conjunto de decisões assim obtidas pelos especialistas são ponderadas. Desse modo, o sistema especialista que estiver com maior peso na saída acaba determinando o resultado final da classificação a respeito da amostra considerada. Em seguida, uma etapa de pós-processamento dinâmico, implementado através de uma rede recorrente é realizada. Esta permite mitigar os efeitos de oscilação que ocorre durante o reconhecimento de classes com características semelhantes. Nesta dissertação são investigados dois modelos de composição de especialistas. O primeiro é baseado no agrupamento de classes associadas a fonéticas semelhantes enquanto o segundo leva em conta a distribuição desequilibrada das amostras apresentadas nos dados de treinamento. A comparação do modelo proposto nesta dissertação com os trabalhos relacionados ao reconhecimento automático de fala indica um ganho de 7,62% em termos de acurácia.

Palavras-chave: Reconhecimento automático de fala. Reconhecimento fonético.
Redes neurais artificiais. Composição de especialistas.
Redes neurais recorrentes.

ABSTRACT

BONILLA C, D A. *Automatic Speech Recognition of Phonemes in Portuguese Using Mixtures of Neural Networks Experts.* 2016. 130f. Dissertação (Mestrado em Engenharia Eletrônica) – Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2016.

The automatic speech recognition based on detection of phonemes provides advantages for online recognition of a speech represented by a sound signal. Studies in computational intelligence have allowed associating the advantages of information parallel processing to the computational workload distribution, aiming at simplifying the use of complex models. The development of a system for automatic speech recognition involves various processes, covering several areas of research, such as linguistics, signal processing and computational intelligence. In this dissertation, the process starts with a speech signal pre-processing to extract and represent more succinctly, the main characteristics of the speech signal at a given instant of time. Inspired by the principle that recommends to "divide and conquer", a model of mixture of neural network experts is explored, allowing to divide the decision space of the complex problem of speech recognition so that each expert takes care only of a delimited area of this decision space. It is noteworthy to point out that each expert included in the composed model takes handles in the decision regarding each one of the preprocessed input sample. This set of decisions thus obtained is weighted. So, the expert system with the highest weight for the output will determine the final classification of the considered sample. After that, a dynamic post-processing step, implemented as a recurrent network, is executed. It aims at mitigating the oscillatory effect that occurs during the recognition of classes with similar characteristics. In this dissertation, two models of mixture of experts are investigated. The first is based on the grouping of similar phonetics classes while the second considers the imbalanced distribution of samples in the training set. The comparison of the model proposed in this dissertation with the work related to automatic speech recognition indicates a gain of 7.62 % in terms of accuracy.

Keywords: Automatic speech recognition. Phonetic recognition. Artificial neural networks. Mixture of experts. Recorrent neural networks.

LISTA DE FIGURAS

1 Sinal de fala no tempo 22
2 Classificação da fala em um sinal dependente do tempo 23
3 Caracterização do silêncio em tempo e frequência. 23
4 Caracterização da fala não articulada em tempo e frequência 24
5 Caracterização da fala em tempo e frequência 24
6 Comparação entre os espectros de frequências do fonema /ɔ/ correspondentes a diferentes falantes. 25
7 Elementos externos à fala que interferem na qualidade do sinal 27
8 Características acústicas das vogais. 29
9 Características acústicas dos ditongos 30
10 Características acústicas das oclusivas 32
11 Espectrograma da silaba /no/ 32
12 Características acústicas das consoantes fricativas 33

13 Esquema básico de um sistema RAF 45
14 Etapas do preprocessamento sobre o sinal de entrada. 46
15 Esquema de conversão de um sinal analógico para um sinal digital. 47
16 Filtro de pré-ênfase aplicado ao sinal de entrada 49
17 Processo de janelamento sobre um sinal de fala 50
18 Janela de Hamming no sinal de entrada segmentado. 51
19 MFCC: Método de extração de características para o sistema RAF proposto 52
20 Densidade Espectral de Potência para um sinal de fala. 53
21 Relação entre a escala referente as frequências de Mel e a escala de frequências absolutas 54
22 Banco de filtros de Mel filtrando o sinal com representação espectral 54
23 Representação da DCT aplicada a um sinal de fala. 55

24 Neurônio biológico 59
25 Modelo de um neurônio artificial 60
26 Partes de uma rede neural artificial 62
27 Diagrama de uma rede neural recorrente comum 63
28 Esquema de um aprendizado supervisionado aplicado em uma RNA 65
29 Componentes de uma rede multicamada 67
30 Comparação entre as faixas de representação para um banco de dados sem normalização e as mesmas observações normalizadas usando *z-score*. 73
31 Exemplos de padrões atípicos do conjunto de treinamento usado na implementação 74
32 Exemplos de padrões atípicos contidos dentro do conjunto de treinamento . 75

LISTA DE FIGURAS

33 Quantidade de amostras por fonema para um grupo de treinamento do banco de dados de fala. 76
34 Tipos de comitê de máquinas 79
35 Decomposição do espaço de entrada dado um sistema de composição de especialistas. 79
36 Representação do desdobramento da rede recorrente em uma rede com propagação reversa de erro. 82

37 Diagrama de operação do sistema de captura de som de SIMULINK®. . . . 87
38 Estágios do pré-processamento implementados em SIMULINK. 88
39 Componentes de especialistas distribuídos de acordo com os grupos fonéticos 92
40 Componentes de especialistas distribuídos de acordo com a taxa de representatividade fonética e suas características espectrais 94
41 Deslocamento temporal da amostras dos ditongos para a rede especialista . 95
42 Estratégia de combinação de especialistas para níveis hierárquicos 98

43 Média do erro de validação para hipóteses apresentadas pelos especialistas. 104
44 Resultados de classificação usando o modelo MEF 107
45 Resultados de classificação usando o modelo MEF e a rede recorrente de pós-processamento 108
46 Exemplo do comportamento da saída do modelo MEF com e sem pós-processamento vs a saída desejada 109
47 Comparação entre modelos MEF com e MEF+RNN 110
48 Resultados do processo de classificação usando o modelo MEHC 111
49 Resultados do processo de classificação usando o modelo MEHC com pós-processamento da saída 112
50 Comparação entre modelos MEHC com e sem RNN no pós-processamento 114
51 Comparação de desempenho entre modelos de mistura MEF e MEHC 115
52 Desempenho global de acurácia e PER para os modelos propostos com e sem pós-processamento 115
53 Comparação entre os diferentes modelos de classificação dos trabalhos relacionados e os modelos MEHC 116

54 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em vogais. 132
55 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em fricativas. . . 133
56 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em fonemas plosivos.. 134
57 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista dos fonemas Nasais.. 135
58 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista dos fonemas vogais nasais.. 136
59 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em ditongos sem atraso. 137

LISTA DE FIGURAS

60 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em pausas. 138
61 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /o//a/ do modelo MEHC. 139
62 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ẽ/ /e/ do modelo MEHC. 140
63 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ĩ/ /i/ do modelo MEHC. 141
64 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ɔ/ do modelo MEHC. 142
65 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ʧ//v/ do modelo MEHC. 143
66 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /r/ do modelo MEHC. 144
67 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ũ//u/ do modelo MEHC. 145
68 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ũ//u/ do modelo MEHC. 146
69 Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista ɛdo modelo MEHC. 147

LISTA DE TABELAS

1 Comparação da média de frequências dos primeiros formantes entre homens e mulheres . . . 25
2 Alfabeto fonético internacional IPA- Vogais . . . 28
3 Alfabeto fonético internacional IPA- Consoantes . . . 31

4 Fonemas presentes no Português Brasileiro usando a notação Worldbet - Vogais . . . 85
5 Consoantes presentes no Português Brasileiro e outros tipos de sons usando a notação Worldbet . . . 86
6 Distribuição das classes fonéticas dentro de cada sistema especialista . . . 91
7 Distribuição de amostras das classes fonéticas por cada especialista e em cada nível hierárquico . . . 93
8 Configuração topológica das redes de passagem . . . 96
9 Matriz de confusão para classificação com múltiplas classes . . . 101

10 Resultados do Teste T para as topologias candidatas dos especialistas pertencentes ao modelo MEF . . . 105
11 Resultados do Teste T para as topologias candidatas dos especialistas pertencentes ao modelo MEHC . . . 105
12 Comparação de desempenhos entre os modelos MEF e MEF+RNN . . . 109
13 Comparação de desempenhos entre os modelos MEHC e MEHC+RNN . . . 113
14 Comparação de configuração e acurácia para trabalhos relacionados ao reconhecimento de fala . . . 116

15 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em vogais . . . 132
16 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em fricativas . . . 133
17 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em plosivos . . . 134
18 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em fonemas nasais . . . 135
19 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em vogais nasais . . . 136
20 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em ditongos sem atraso . . . 137
21 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em pausas . . . 138
22 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /o//a/ . . . 139

LISTA DE TABELAS

23 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ẽ/ /e/ 140
24 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ĩ/ /i/ 141
25 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ɔ/ 142
26 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ʧ//v/ 143
27 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /r/ 144
28 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ũ//u/ 145
29 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ũ//u/ 146
30 Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em ɛ 147
31 Matriz de confusão correspondente ao modelo MEF usando o banco de dados *Birthdate* 148
32 Matriz de confusão correspondente ao modelo MEF usando o banco de dados *Age* 148
33 Matriz de confusão correspondente ao modelo MEF com pós-processamento dinâmico usando o banco de dados *Birthdate* 149
34 Matriz de confusão correspondente ao modelo MEF com pós-processamento dinâmico usando o banco de dados *Age* 149
35 Matriz de confusão correspondente ao modelo MEHC usando o banco de dados *Birthdate* 150
36 Matriz de confusão correspondente ao modelo MEHC usando o banco de dados *Age* 150
37 Matriz de confusão correspondente ao modelo MEHC com pós-processamento dinâmico usando o banco de dados *Birthdate* 151
38 Matriz de confusão correspondente ao modelo MEHC com pós-processamento dinâmico usando o banco de dados *Age* 151

LISTA DE ALGORITMOS

1 Propagação reversa de erro para uma rede neural multicamada 69
2 Atualização dos pesos usando a técnica propagação resiliente. 71
3 Distância de Levenshtein 100

LISTA DE SIGLAS

ASR	*Automatic Speech Recognition*
CNN	*Convolutional Neural Network*
DCL	*Differential Competitive Learning*
DCT	*Discrete Cosine Transform*
DFT	*Discrete Fourier Transform*
DL	Distância de Levenshtein
DNN	*Deep Neural Network*
DP	*Dynamic Programming*
DTAK	*Dynamic Time Alignment Kernel*
DTW	*Dynamic Time Wraping*
FFT	*Fast Fourier Transform*
FIR	*Finite Impulse Response*
GPD	*Generalized Probabilistic Descent*
HMM	*Hidden Markov Model*
IPA	*International Phonetic Alphabet*
LDC	*Linguistic Data Consortium*
MEF	Modelos Especialsitas Fonéticos
MEHC	Modelos Especialistas Hierarquizados por Classes
MFCC	*Mel-Frequency Cepstrum Coefficients*
MFSC	*Mel-Frequency Spectral Coeffients*
MLP	*Multilayer Perceptron*
PAR	*Phoneme Accuracy Rate*
PCM	*Pulse Code Modulation*
PER	*Phoneme Error Rate*

PSD	*Power Spectral Density*
RAF	Reconhecimento Automático de Fala
RBF	*Radial Basis Function*
RBM	*Restricted Boltzmann Machine*
RNA	Rede Neural Artificial
RNN	*Recurrent Neural Network*
RP	Rede de Passagem
SMLP	*Sequence of Multilayer Perceptron*
SOM	*Self-Organization Map*
SQNR	*Signal to Quantization Noise Ratio*
SVM	*Support Vector Machine*
TDNN	*Time Delay Neural Network*
TL	*Telecomunication Laboratories*
VSR	*Very Smart Recognizer*

SUMÁRIO

INTRODUÇÃO **16**

1 RECONHECIMENTO AUTOMÁTICO DE FALA **20**
1.1 Aspectos básicos da fala 21
1.1.1 Interpretação biológica da comunicação oral 21
1.1.2 Representação da fala em domínios de tempo e frequência 21
1.2 Características de complexidade de um sistema de fala 24
1.2.1 O falante 25
1.2.2 O modo de fala 26
1.2.3 A gramática 26
1.2.4 O ambiente físico externo 26
1.3 O vocabulário: Sons do Português Brasileiro e suas características 27
1.3.1 Vogais 28
1.3.2 Consoantes 30
1.4 Considerações finais do capítulo 34

2 TRABALHOS RELACIONADOS **35**
2.1 Classificação usando redes neurais artificiais 35
2.1.1 Classificação de vogais 35
2.1.2 Classificação de fonemas 37
2.1.3 Classificação de palavras 38
2.1.4 Classificação de outras características da fala 39
2.2 Novas arquiteturas 40
2.3 Implementações em dispositivos inteligentes 43
2.4 Considerações finais do capítulo 44

3 PRÉ-PROCESSAMENTO E EXTRAÇÃO DE CARACTERÍSTICAS DO SINAL DE FALA **45**
3.1 Pré-processamento 46
3.1.1 Conversor analógico-digital 46
3.1.2 Filtro de pré-ênfase 48
3.1.3 Análise em tempo curto e janela de Hamming 49
3.2 Extração de características 51
3.2.1 Transformada discreta de Fourier e densidade de potência espectral 52
3.2.2 Banco de filtros de mel 53
3.2.3 Transformada discreta de cosseno 55
3.3 Considerações finais do capítulo 56

SUMÁRIO

4 Redes Neurais Artificiais no reconhecimento automático de fala . 57
4.1 Redes neurais artificiais . . . 58
4.1.1 Modelo biológico de um neurônio . . . 58
4.1.2 Funcionamento do neurônio artificial . . . 60
4.1.3 Topologia de uma RNA . . . 61
4.1.4 Aprendizagem . . . 64
4.1.5 Treinamento . . . 66
4.2 Aperfeiçoamento dos dados de treinamento . . . 71
4.2.1 Normalização usando técnicas estatísticas . . . 72
4.2.2 Padrões atípicos ou ruidosos . . . 73
4.2.3 Desequilíbrio de classes . . . 76
4.3 Comitê de máquinas . . . 78
4.4 Função dinâmica de pós-processamento . . . 80
4.4.1 Algoritmo de aprendizado . . . 81
4.5 Considerações finais do capítulo . . . 82

5 ASPECTOS DE IMPLEMENTAÇÃO . . . 84
5.1 Características do conjunto de treinamento . . . 84
5.2 Implementação do pré-processamento . . . 86
5.3 Configuração e treinamento da composição de especialistas . . . 89
5.3.1 Número de especialistas e profundidade hierárquica . . . 89
5.3.1.1 Topologia baseada em agrupamentos fonéticos MEF . . . 91
5.3.1.2 Topologia baseada no desequilíbrio de classes MEHC . . . 92
5.3.2 Definição topológica dos sistemas especialistas . . . 93
5.3.3 Aspectos de implementação das redes de passagem . . . 96
5.3.4 Função combinadora . . . 96
5.3.5 Pós-processamento usando uma rede recorrente . . . 97
5.4 Medidas de desempenho um sistema de RAF . . . 99
5.4.1 Taxa de acurácia fonética . . . 99
5.4.2 Medidas de desempenho para classificação . . . 100
5.5 Considerações finais do capítulo . . . 102

6 RESULTADOS . . . 103
6.1 Topologia das redes especialistas. . . . 103
6.2 Resultados do método de mistura do tipo MEF . . . 106
6.2.1 Resultados usando o modelo MEF sem pós-processamento . . . 106
6.2.2 Resultados do modelo MEF com pós-processamento dinâmico . . . 107
6.3 Resultados do método de mistura do tipo MEHC . . . 110
6.3.1 Resultados usando o modelo MEHC sem pós-processamento . . . 110
6.3.2 Resultados do modelo MEHC com pós-processamento dinâmico . . . 111
6.4 Comparações dos modelos de mistura propostos . . . 113
6.5 Comparação dos resultados relacionados . . . 116
6.6 Considerações finais do capítulo . . . 117

7 CONCLUSÕES . . . 118
7.1 Conclusões . . . 118
7.2 Trabalhos Futuros . . . 121

REFERÊNCIAS . . . 122

SUMÁRIO

APÊNDICE A – Seleção dos modelos topológicos para cada especialista . 131

INTRODUÇÃO

CONSTANTE evolução tecnológica faz do reconhecimento automático de fala um campo de estudo fascinante mas ao mesmo tempo desafiador, uma vez que é bastante grande a gama de aplicações em que a voz tem o papel de agilizar a realização de tarefas cotidianas, buscando extrair da fala as informações relevantes para a realização do reconhecimento. Ao mesmo tempo, buscam-se sempre metodologias inovadoras que permitam um reconhecimento mais ágil e preciso para fazer com que a máquina compreenda a linguagem falada.

O reconhecimento automático de fala foi introduzido por Petterson e Barney em 1952 (PETERSON; BARNEY, 1952). É uma área de estudo multidisciplinar, que abrange conhecimentos de linguística computacional, processamento de sinais e inteligência artificial. Mediante a extração da informação de um sinal acústico que contém fala, um dispositivo inteligente, com métodos de reconhecimento de fala embarcado, pode gerar transcrições para um usuário final. Também pode interpretar um conjunto de comandos que orientem e controlem as ações e atividades a serem feitas pelo dispositivo.

O funcionamento do cérebro humano na interpretação da fala inspirou as pesquisas relacionadas à adaptação de técnicas da inteligência artificial no reconhecimento e interpretação de fala (BABER, 1991). O uso das redes neurais artificiais para o reconhecimento automático de fala iniciou-se na década de 80, devido ao potencial destas arquiteturas para o processamento paralelo e distribuído assim como o potencial dos métodos de aprendizado utilizados para a incorporação de novos padrões acústicos.

Um sistema que utiliza o reconhecimento automático de fala pode tratar da identificação de padrões que representem palavras completas e/ou identificar fonemas. Os sistemas baseados na detecção de palavras lidam com um grande conjunto de classes, proporcional ao tamanho do vocabulário a ser interpretado. Para um sistema que tem como finalidade a identificação de um determinado número de comandos, esta solução é adequada. Porém, um sistema que requer a interpretação de fala contínua, baseada

na estrutura linguística de um determinado idioma, precisa de um enorme conjunto de classes, requerendo por tanto uma implementação eficiente de um sistema complexo de reconhecimento para aumentar seu desempenho.

Em alternativa ao reconhecimento de palavras, um sistema baseado em reconhecimento de fonemas lida com um número de classes limitado. Desse modo, o sistema de reconhecimento automático de fala pode proporcionar um tempo de resposta mais curto para a classificação. Tais sistemas poderão ser implementados usando modelos mais simples.

O objetivo desta dissertação é propor um sistema de reconhecimento automático de fala, baseado na identificação de fonemas, que possa classificar em modo *online* cada amostra de som capturado em tempo real. Tratamento *online* consiste em fornecer o resultado de classificação após um atraso reduzido devido ao processamento, de ordem de milissegundos. Mediante um conjunto de características espectrais extraídas do sinal de fala, o sistema deveria identificar e classificar as amostras, uma a uma, de acordo com o conjunto de classes fonéticas presentes no Português.

O sinal de fala, capturado na entrada, deve ser tratado, para mitigar alguns problemas gerados na aquisição. Também é preciso extrair as características frequenciais presentes na fala. Por último, as características assim obtidas, são transformadas de modo a diminuir a cardinalidade das características representativas, eliminando qualquer correlação entre estas.

De acordo com o a identificação online pretendida, foi usado um modelo de composição de especialistas, baseados em redes neurais artificiais que atuem de modo paralelo e distribuído durante a fase de operação. Contudo o treinamento das redes é feito de uma maneira sequencial e *offline*. Vale ressaltar que tendo em vista a divisão do espaço de decisão, o treinamento das redes que compõem o sistema especialista fica mais simples.

Finalmente, para dar conta do comportamento dinâmico da fala, que muda de características durante o tempo, é proposto um pós-processamento que possa aproveitar o comportamento anterior do sinal de fala que está sendo classificado na decisão final de reconhecimento.

A principal contribuição desta dissertação é a modelagem e a implementação de um sistema de classificação baseado em uma composição de especialistas, cuja função é dividir o espaço de decisão, assim, o processo de treinamento de cada especialista fica mais

simples. A definição da distribuição destes especialistas está baseada nos agrupamentos fonéticos, usando o modelo MEF, propostos em (SCANLON; ELLIS; REILLY, 2007) e no desequilíbrio das classes fonéticas inerente aos dados de treinamento, usando o modelo MEHC. Utilizando como estratégia o uso de redes neurais recorrentes no pós-processamento da classificação, foi mitigada a oscilação gerada pela classificação estática de amostras.

Os resultados obtidos dos experimentos são satisfatórios, obtendo uma acurácia de 86,525% e uma taxa de erro fonético de 13,128% no caso do modelo MEHC. O pós-processamento implementado minimizou a oscilação presente na identificação fonética por amostra, incrementando, de modo geral, os desempenhos dos modelos propostos. Vale ressaltar que o sistema de reconhecimento baseado em MEHC proporcionou uma melhoria na taxa de acurácia quando comparado com o estado da arte no reconhecimento de fonemas.

O restante desta dissertação está organizado em seis capítulos, cujos conteúdos são resumidamente descritos a seguir. Inicialmente, o Capítulo 1 apresenta uma introdução teórica do problema de reconhecimento automático de fala, onde são expostas as variáveis que influenciam este reconhecimento, incluindo as características acústicas dos fonemas relacionados à língua Portuguesa.

O Capítulo 2 apresenta de maneira sucinta, os trabalhos relacionados ao reconhecimento automático de fala. São apresentadas diferentes abordagens do problema e as soluções propostas.

O Capítulo 3 descreve detalhadamente as técnicas de pré-processamento e extração de características usadas para sinais de fala entrante, onde também é visado melhorar as condições acústicas do sinal amostrado.

O Capítulo 4 apresenta a fundamentação teórica relacionada à redes neurais artificiais, as técnicas de tratamento de sinais pré-processados considerando os problemas de ruído. O capítulo também define as redes de composição de especialistas, baseadas em redes neurais. Por último, é descrita a rede recorrente usada durante o pós-processamento dinâmico, usado para lidar com o comportamento dinâmico da fala no tempo.

O Capítulo 5 descreve as características e considerações técnicas adotadas na implementação de um sistema de RAF, tendo em vista as etapas de pré-processamento, classificação e pós-processamento, assim como as condições técnicas referentes ao banco de dados usado na fase de treinamento e teste.

Finalmente, o Capítulo 6 completa esta dissertação, apresentando as principais conclusões obtidas do desenvolvimento do presente trabalho. São também apontadas possíveis melhorias do modelo proposto e as direções para futuros trabalhos.

Capítulo 1

RECONHECIMENTO AUTOMÁTICO DE FALA

RECONHECIMENTO automático de fala ou RAF (em inglês é conhecido como ASR, *Automatic Speech Recognition*) é um campo de estudo da inteligência artificial que tem como finalidade a implementação de sistemas que transformam características acústicas de um sinal de fala em uma sequência digital representada por fonemas ou palavras escritas (DANIEL; JAMES, 2009).

Os principais objetivos do RAF são o desenvolvimento e a implementação de métodos eficientes para a comunicação homem-máquina. Mesmo que as comunicações verbais demandem mais recursos computacionais que outros métodos de interfaceamento tais como teclados ou telas do tipo *touchscreen*, as aplicações de RAF oferecem maior velocidade de resposta computacional para operações de comando. Por outro lado, é uma ferramenta de muita utilidade para pessoas que possuem problemas físico-motores (DAMPER, 1982), pessoas que têm trabalhos onde suas mãos estão ocupadas (BABER, 1991) e automação de espaços residenciais e comerciais (PRINCIPI et al., 2015), entre outros.

Por muito tempo, a implementação do RAF tem sido afetada por vários elementos que comprometem o sinal a ser interpretado pelo sistema. O trabalho de reconhecimento de fala torna-se desafiante devido às diferentes variações nos sinais de fala tais como as características próprias ao ruído externo, dependência de reconhecimento a um locutor, a fonética e fonoaudiologia de cada idioma, além das regras de semântica que também influenciam o processo de reconhecimento de fala (VIRTANEN; SINGH; RAJ, 2012).

1.1 Aspectos básicos da fala

Nesta seção, serão apresentados alguns conceitos básicos, frequentemente encontrados na área de processamento e reconhecimento de sinais de fala, e que serão utilizados no decorrer deste trabalho.

1.1.1 Interpretação biológica da comunicação oral

Em um processo de comunicação oral, o transmissor da mensagem inicia com um processo neurolinguístico que é considerada como a primeira etapa responsável pela produção de fala. Em seguida, um processo neuromuscular aciona em conjunto a respiração, as cordas vocais, a língua e os lábios para gerar as ondas acústicas que canalizam a mensagem até o receptor (FRATTALI; ASSOCIATION et al., 1995).

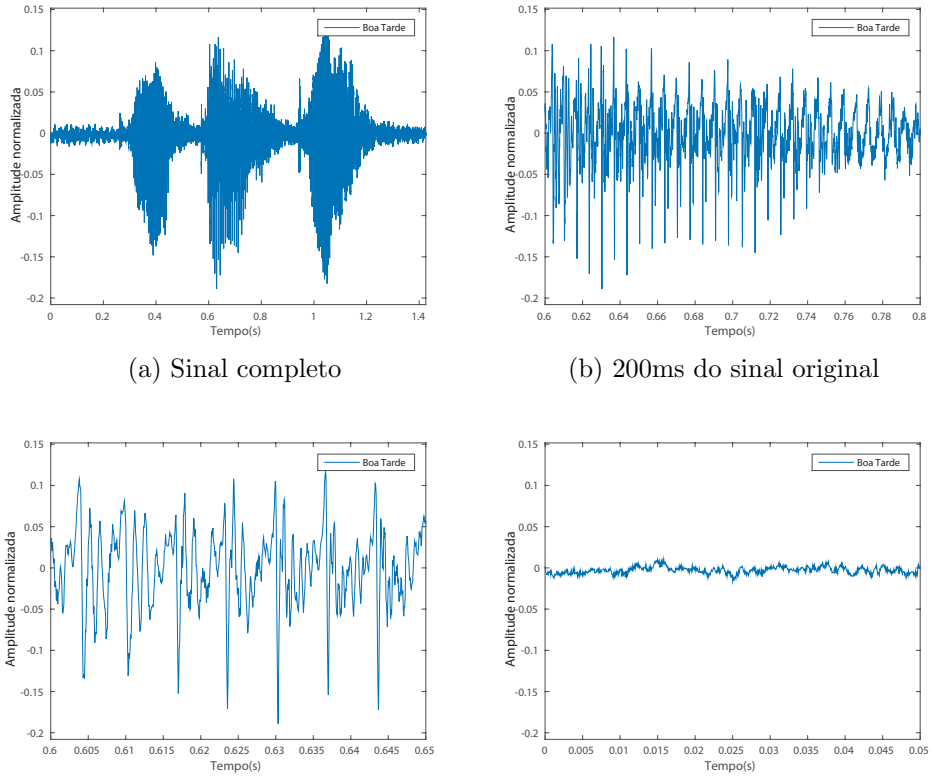
Depois desta construção da fala, o receptor captura o sinal e o processo de reconhecimento começa. O sinal acústico é captado pela membrana basilar no ouvido interno do receptor. Assim, o sistema biológico fornece uma análise de espectro do sinal de entrada. Posteriormente, é realizado um processo de extração das características usando uma transdução neural dentro do nervo auditivo. Finalmente, os centros de processamento dentro do cérebro geram o código de linguagem que é interpretado subsequentemente como a mensagem original (RABINER; JUANG, 1999).

1.1.2 Representação da fala em domínios de tempo e frequência

A fala é um sinal unidimensional que depende da variável associada ao tempo, das frequências presentes na produção da fala e da potência com que esta é gerada. De acordo com estas características físicas associadas à fala é possível obter diferentes representações que ofereçam maior informação sobre a voz (HUANG et al., 2001). Inicialmente, a fala pode ser interpretada dependendo da faixa de tempo em que esta é analisada.

A Figura 1(a) apresenta um sinal normalizado em termos de tempo e amplitude que carrega a fala “Boa tarde”. Se o sinal é analisado usando uma faixa de tempo igual ou superior a 200ms, conforme ilustra a Figura 1(b), as características mudarão de acordo com os sons incluídos, gerando um sinal aleatório não estacionário, onde suas características estatísticas variaram com o tempo, e portanto, acrescenta a complexidade de análise do sinal (HUANG et al., 2001). Porém, se a faixa de análise é menor do que 100ms,

conforme mostrado na Figura 1(c), o sinal apresentará características quase-estacionarias, facilitando o análise de cada elemento falado.



(a) Sinal completo

(b) 200ms do sinal original

(c) 50ms de fala do sinal original

(d) 50ms de silêncio do sinal original

Figura 1: Sinal de fala no tempo

Mesmo assim, a característica quase-estacionaria em períodos de análise curtos não é aplicável em todos os casos. A Figura 1(d) apresenta 50ms de silêncio do sinal original, onde permanecem as caraterísticas aleatórias e não estacionárias.

Portanto, considerando os atributos de amplitude e tempo, os sinais que contém fala são divididos em 3 respostas dependentes do funcionamento do trato vocal. A Figura 2 apresenta as componentes de um sinal com fala, onde V representa os sinais com uma excitação de entrada quase estacionária e M os sinais com excitação de entrada aleatória, semelhante ao ruído. Usualmente, estes últimos são produzidos no início e/ou final das palavras faladas. No terceiro caso, diz respeito à inexistência de excitação e por conseguinte, as faixas identificadas como S representam o silêncio somado com a presença de ruído externo ao trato vocal.

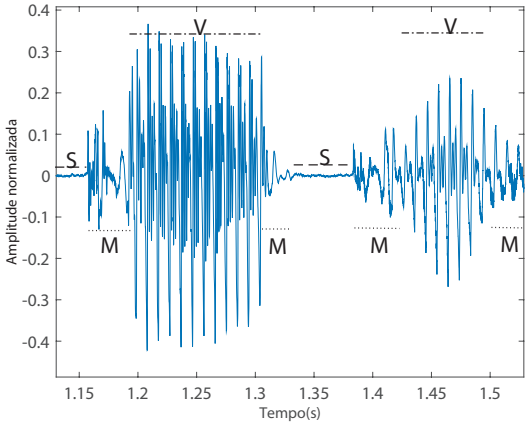
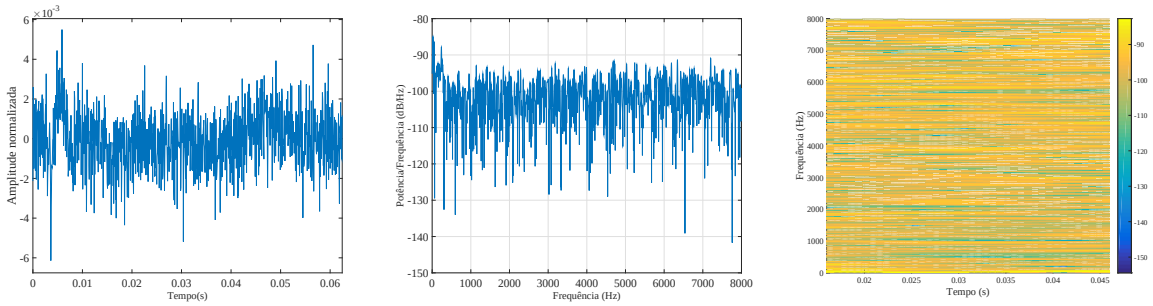


Figura 2: Classificação da fala em um sinal dependente do tempo

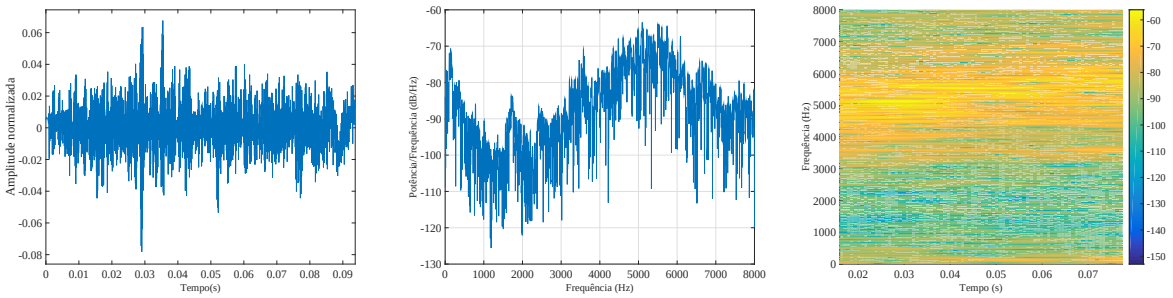
O silêncio é um sinal aleatório e não estacionário com características espectrais semelhantes ao ruído branco. Apesar de não conter informação de fala, o silêncio é parte integral de um sinal de voz. Sem a presença do silêncio, a fala não é inteligível. A Figura 3 expõe as características temporais e espectrais de um sinal catalogado como silêncio, onde é constatada sua natureza aleatória e externa ao sistema fonador.



(a) Silêncio em tempo (b) Espectro do silêncio (c) Espectrograma do silêncio

Figura 3: Caracterização do silêncio em tempo e frequência

Os sinais com excitação de fala aleatória, M, são conhecidos como fala não articulada ou não vozeada, onde o aparelho fonador humano gera um sinal aleatório semelhante a ruído e sem qualquer carácter periódico. O processo de produção deste tipo de sinal não é bloqueado pelas cordas vogais. No entanto, em algum lugar ao longo do trato vocal é feito algum encerramento total ou parcial que obstrui o fluxo de ar proveniente dos pulmões. A natureza típica deste sinal é apresentado na Figura 4. A representação temporal mostrada na Figura 4(a) não possui natureza estacionária. Por outro lado, a representação espectral apresenta sons com componentes de alta frequência e baixa energia, com pouca estabilidade em períodos pequenos.



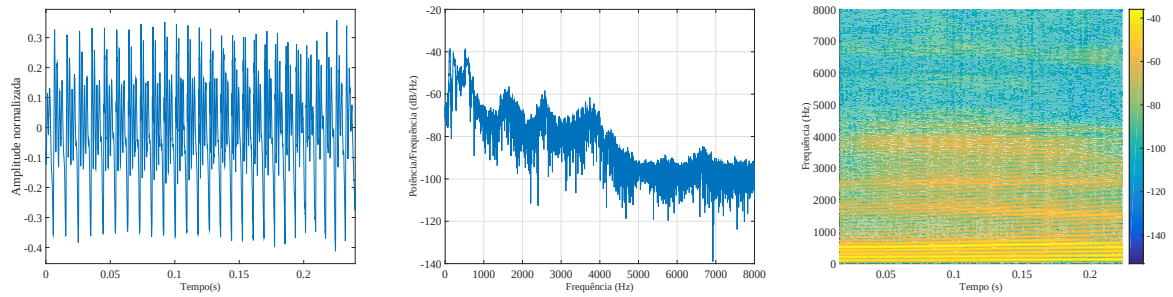
(a) Fala não articulada em tempo.

(b) Espectro da fala não articulada.

(c) Espectrograma da fala não articulada.

Figura 4: Caracterização da fala não articulada em tempo e frequência

Finalmente, as faixas de sinais com excitação periódica de entrada, V, são conhecidos como sinais vozeados. Neste caso, as cordas vogais estão tensas e vibram periodicamente com o fluxo de ar vindo dos pulmões, gerando ondas quase-estacionárias, onde a frequência fundamental de vibração das cordas vogais representa o tom. São sons de alta energia e estabilidade numa faixa estreita. A Figura 5(b) apresenta claramente a frequência fundamental do sinal e seus formantes.



(a) Fala em tempo.

(b) Espectro da fala.

(c) Espectrograma da fala.

Figura 5: Caracterização da fala em tempo e frequência

1.2 Características de complexidade de um sistema de fala

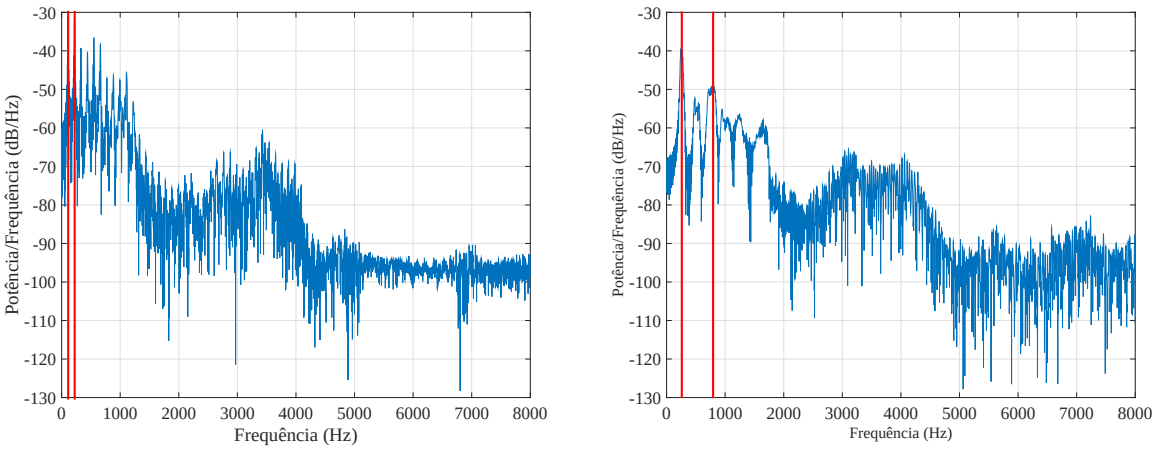
Nesta seção, são detalhados os diferentes fatores implícitos e explícitos próprios à produção da fala que afetam diretamente a complexidade de desenvolvimento de um sistema computacional para o reconhecimento automático de fala.

1.2.1 O falante

Este é o aspecto que apresenta maior variabilidade na onda de entrada e por conseguinte, requer um sistema altamente robusto. Uma pessoa não pronuncia sempre do mesmo modo, devido às diferentes situações físicas e psicológicas. Isto é conhecido como fator intra-locutor. Por outro lado, existem diferentes tipos de locutores, homens, mulheres e crianças, com diferenças de idade e distintas origens geográficas que geram mais uma fonte de variabilidade interlocutora. Um reconhecedor automático de fala independente do locutor é um sistema robusto e generalizável entre um grupo diverso de falantes (OSHAUGHNESSY, 2008). A Tabela 1 apresenta uma comparação entre as primeiras frequências fundamentais em Hz correspondentes a falantes masculinos e femininos usando como exemplo as vogais /i/, /e/, /a/ e /ɔ/. Note-se a diferença do cumprimento da faixa entre os formantes F_1 e F_2 na Figura 6, sendo mais ampla para um fonema falado por uma mulher.

Tabela 1: Comparação da média de frequências dos primeiros formantes entre homens e mulheres

Vogal	**Homem**			**Mulher**		
	F_1	F_2	F_3	F_1	F_2	F_3
/i/	400	2000	2550	430	2500	3100
/e/	530	1850	2500	600	2350	3000
/a/	660	1700	2400	860	2050	2850
/ɔ/	570	850	2400	590	900	2700



(a) Falado por um homem.

(b) Falado por uma mulher.

Figura 6: Comparação entre os espectros de frequências do fonema /ɔ/ correspondentes a diferentes falantes.

1.2.2 O modo de fala

A fala natural humana não contém pausas entre o início e fim de cada palavra, elas aparecem principalmente ao nível sintático, como em frases ou sentenças. Ademais as variações tônicas introduzidas como fenômenos de entonação, conhecidos como eventos prosódicos, geram palavras que têm características acústicas e semânticas diferentes dependendo dos vocábulos predecessores ou sucessores. Um reconhecedor automático de fala é relativamente simples quando se trata de reconhecer uma palavra isolada. No entanto esta complexidade aumenta quando se trata de reconhecer as palavras de uma frase produzida durante uma fala natural contínua (OSHAUGHNESSY, 2008).

1.2.3 A gramática

É definida como o conjunto de regras que limita o número de combinações permitidas considerando o conjunto das palavras de um vocabulário. Em geral, a existência de uma gramática contida em um sistema RAF ajuda a melhorar a taxa de reconhecimento, já que elimina ambiguidades e pode otimizar a necessidade de cálculo mediante a limitação do número de palavras numa determinada fase de reconhecimento. Num sistema de reconhecimento de palavras isoladas, como por exemplo, o reconhecimento de números falados ou grupos de comandos, onde não existe uma gramática no sentido formal da definição, a regulamentação gramatical pode ser implementada como um número limitado de palavras a serem reconhecidas (MARIANI, 2013).

1.2.4 O ambiente físico externo

O ambiente físico implica o espaço de gravação da fala, os meios eletrônicos usados para a obtenção dos sons e os tipos de canais de comunicação usados para transmitir o sinal capturado, se for necessário.

A Figura 7 apresenta os diferentes elementos que prejudicam o sinal de fala a ser gravado (RABINER; JUANG, 1999). Começando pelo tipo de ambiente onde está o falante, este pode ser um espaço controlado e adequado para a gravação, facilitando a tarefa de reconhecimento. No entanto, este pode ser um ambiente não controlado, onde podem ser gravadas diferentes interferências de elementos externos e outras vozes diferentes ao falante objetivo, dificultando a tarefa de reconhecimento.

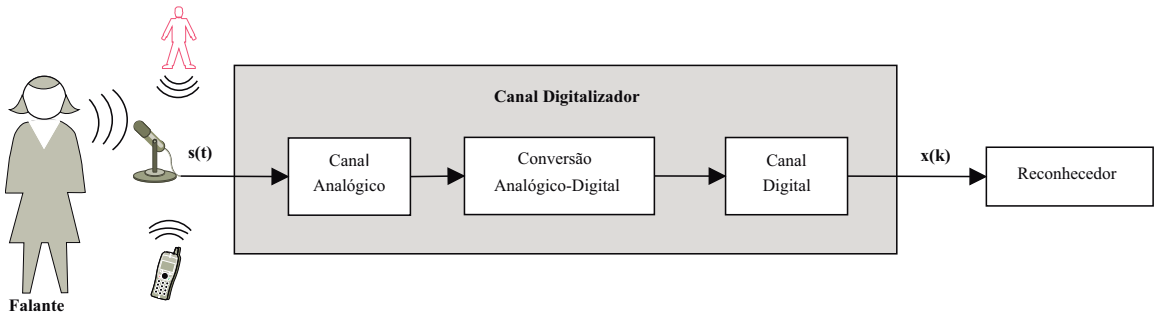


Figura 7: Elementos externos à fala que interferem na qualidade do sinal

Um outro elemento consiste no tipo de microfone usado. Cada microfone possui características físicas distintas, incluindo sua resposta de frequência. Um microfone adequado deve garantir certas propriedades para uma gravação adequada, dado que a faixa média de frequências da fala humana está entre 100Hz e 8000Hz (VIRTANEN; SINGH; RAJ, 2012).

Finalmente, estão os meios e métodos de transmissão e de digitalização, onde as modulações de frequências e amplitudes, as restrições de largura de banda nos canais de comunicação e os métodos de amostra e profundidade durante a conversão analógica-digital podem interferir no processo de reconhecimento de fala.

1.3 O vocabulário: Sons do Português Brasileiro e suas características

O vocabulário é o número de palavras diferentes que o sistema de RAF deve reconhecer (RABINER; JUANG, 1999). A dificuldade de implementação de um sistema RAF é proporcional ao número de palavras que devem ser reconhecidas. Em particular, isto é devido ao fato que ter um maior banco de dados de palavras aumenta a probabilidade de obter palavras semelhantes, Ademais, o tempo de processamento aumenta com o número de palavras que devem ser avaliadas. Uma possível solução é a utilização de unidades linguísticas menores a palavra, tais como sílabas e fonemas. Assim, limita-se o número de elementos a serem identificados (TORRE, 2012).

A fonologia é o ramo da linguística que estuda o sistema sonoro de um idioma, onde a menor unidade sonora é conhecida como fonema. Um fonema é a representação simbólica de um som da fala, não obstante, este tipo de representação universal da fala não deve ser confundido com as letras ou grafemas, já que estas são representações próprias dependentes de cada idioma (CASTELLVÍ, 2010).

A língua portuguesa contém uma das fonologias mais ricas dos idiomas românicos (BISOL, 2005). Na atualidade, o Português Brasileiro é falado aproximadamente por 158 milhões de pessoas e abrange numerosos dialetos derivados da língua portuguesa (BORGES, 2000). A fala do Português Brasileiro é descrita por 36 fonemas, onde 26 deles são classificados como consoantes e 10 são catalogados como vogais. Por outro lado, a fonologia do Português Brasileiro integra representações fonéticas para dígrafos vogais que estão presentes em de 5 vogais nasais e 21 ditongos orais (SILVA; YEHIA, 2015).

As Tabelas 2 e 3 apresentam os fonemas descritos no alfabeto fonético internacional IPA (*International Phonetic Alphabet*) (ASSOCIATION, 1999). O propósito deste alfabeto é dar uma representação única e precisa dos sons de qualquer linguagem oral. Os fonemas destacados na tabela correspondem a sons falados dentro do Português Brasileiro (SILVA; YEHIA, 2015), incluindo todos os dialetos usados no do território Brasileiro.

1.3.1 Vogais

A Tabela 2 apresenta os símbolos fonéticos das vogais de acordo com o método de articulação dentro do aparelho fonador. As vogais são produzidas por uma excitação do trato vocal, formando ondas quase-periódicas que dependem da vibração das cordas vocais. O modo em que estas vibram determina as frequências de formantes. Geralmente, as vogais têm uma longa duração e um espectro de frequência bem definido.

Tabela 2: Alfabeto fonético internacional IPA- Vogais

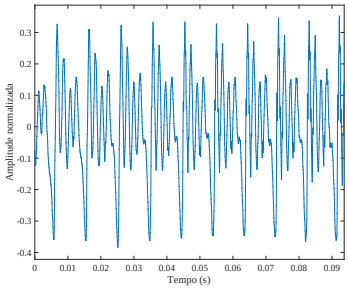
	Anterior	Quase-Anterior	Central	Quase-Posterior	Posterior
Fechada	**i**, ʏ		ɨ, ʉ		ɯ,**u**
Quase-Fechada		**I**, Y		ʊ	
Meia Fechada	**e**, ø		ɘ, ɵ		ɤ, **o**
Meia			ə		
Meia Aberta	ɛ, œ		ɜ, ɞ		ʌ, ɔ
Quase-Aberta	æ		ɐ		
Aberta	**a**, ɶ				ɑ, ɒ

Nasal	**ã, ẽ, ĩ, õ, ũ**
Ditongos	a͡I, e͡I, ɛ͡I, o͡I, ɔ͡I, u͡I a͡ʊ, e͡ʊ, ɛ͡ʊ, o͡ʊ, i͡ʊ I͡a, I͡i, I͡e, I͡ʊ, I͡o ʊ͡a, ʊ͡I, ʊ͡e, ʊ͡o, ʊ͡u

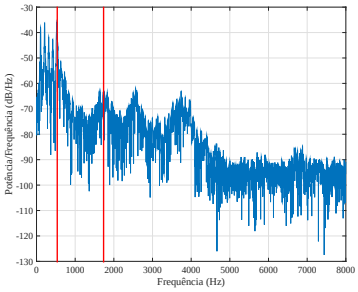
*Fonemas presentes no Português Brasileiro

Existem diferentes modos de caracterizar as vogais: a configuração articulatória requerida para produzir o som, a forma de onda e a gráfica de espectro. A Figura 8

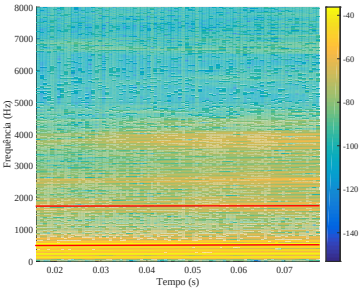
apresenta um exemplo de forma de onda, análise de frequências e espectrograma das vogais /ɛ/ e /ɔ/.



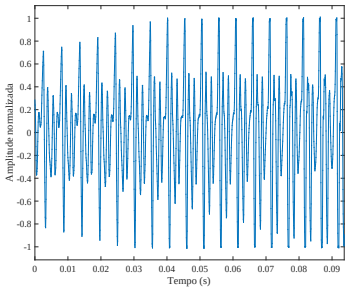
(a) Forma de onda /ɛ/



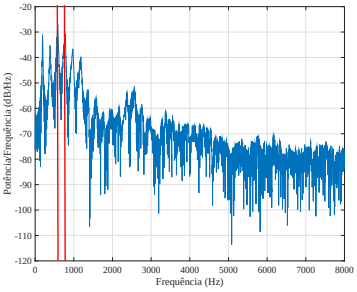
(b) Frequência fundamental em 526Hz, frequência de resonancia em 1727Hz



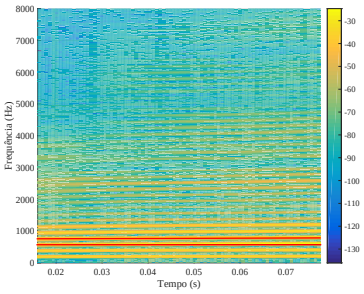
(c) Espectrograma do fonema /ɛ/



(d) Forma de onda /ɔ/



(e) Frequência fundamental em 583Hz, frequência de resonancia em 775Hz



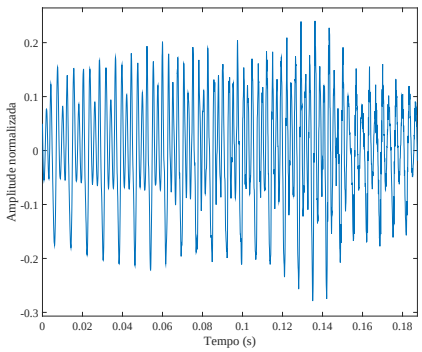
(f) Espectrograma do fonema /ɔ/

Figura 8: Características acústicas das vogais.

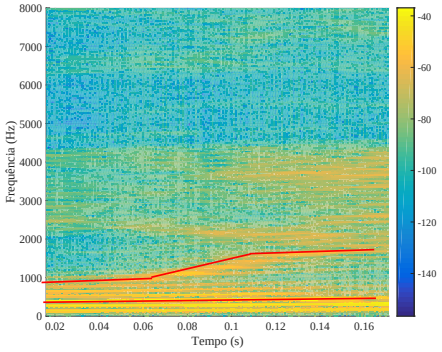
As vogais conhecidas como frontais ou anteriores: /i/, /e/, /ɛ/ e /a/ apresentam formantes bem pronunciadas em altas frequências, conforme mostra a Figura 8(c), enquanto o grupo das vogais posteriores ou traseiras /u/, /o/ e /ɔ/ contém uma predominância de informação nas baixas frequências, conforme mostra a Figura 8(f).

Dentro do grupo dos fonemas vocálicos estão as vogais nasais. Para a produção de uma vogal nasal, a cavidade oral encontra-se aberta, o véu palatino se abaixa e assim, o ar que fez vibrar as cordas vocais, passa tanto pela cavidade nasal como pela cavidade oral (BISOL, 2005). O resultado acústico é produto das ressonâncias produzidas pelas cavidades nasais e orais. Comparadas com as vogais orais, as vogais nasais acrescentam mais informação dentro das frequências baixas, o que gera uma percepção de nasalidade e gravidade no som. Além disso, as vogais nasais são mais longas que as vogais orais, onde uma faixa inicial rapidamente evolui para um som nasal, e aí, a energia correspondente é inferior, afetando sobretudo as primeiras frequências de ressonância (SOUSA, 1994).

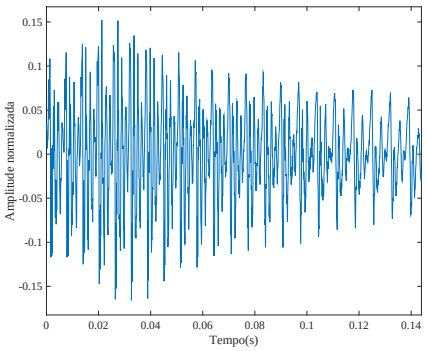
Finalizando os grupos vocálicos estão os ditongos, conhecidos como sons de fala monossilábicos. Eles podem ser crescentes ou decrescentes. Um ditongo crescente se inicia com uma semivogal, ou seja, as vogais /i/ e /u/ quando acompanha outra vogal na mesma sílaba, e depois cresce para tornar-se uma vogal. Esta característica é apreciada na Figura 9(b). Por outro lado, um ditongo decrescente é formado por uma vogal que decresce para se tornar uma semivogal. A Figura 9(d) apresenta o comportamento espectral de um ditongo decrescente.



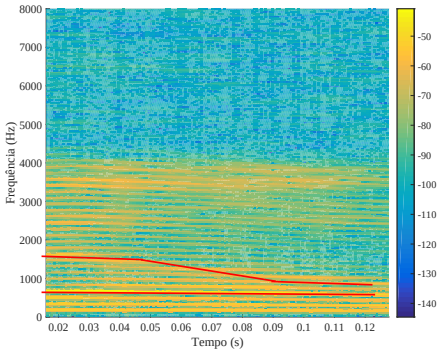
(a) Forma de onda do ditongo /oI̯/



(b) Espectrograma do ditongo /oI̯/



(c) Forma de onda do ditongo /ɛʊ̯/



(d) Espectrograma do ditongo /ɛʊ̯/

Figura 9: Características acústicas dos ditongos

1.3.2 Consoantes

As consoantes são fonemas produzidos com uma total ou parcial obstrução da passagem de corrente de ar gerado pelos pulmões (SILVA; YEHIA, 2015). De acordo com a Tabela 3 as consoantes são classificadas conforme o modo de articulação, que destaca a disposição dos órgãos articulatórios, principalmente seu grau de constrição e lugar onde o fonema é articulado.

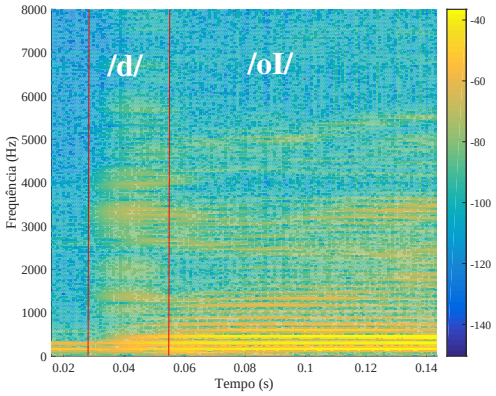
Tabela 3: Alfabeto fonético internacional IPA- Consoantes

	Bilabial	Labio-dental	Dental	Alveolar	Pós-alveolar	Retro-flexa	Paletal	Velar	Uvular	Faringal	Epiglotal	Glotal
Oclusiva	**p**,**b**			**t**,**d**		ʈ,ɖ	c,ɟ	**k**,**g**	q,ɢ		ʡ	ʔ
Nasal	**m**	ɱ		**n**		ɳ	ɲ	ŋ	ɴ			
Fricativa	ɸ,β	**f**,**v**	θ,ð	**s**,**z**	ʃ,ʒ	ʂ,ʐ	ç,ʝ	**x**,ɣ	χ,ʁ	ħ,ʕ	ʢ,ʜ	**h**,ɦ
Africada					ʧ**,**ʤ							
Fricativa lateral				ɮ,ɬ								
Vibrante	ʙ			**r**					ʀ			
Tepe				ɾ		ɽ						
Flepe lateral				ɺ								
Aproximante		ʋ		ɹ		ɻ	j	ɰ				
Aprox. lateral				**l**		ɭ	ʎ	ʟ				

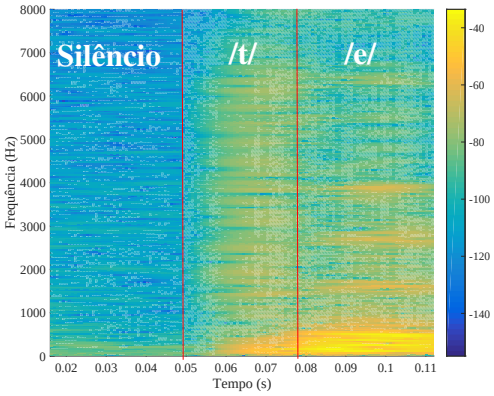
* Consoantes não vozeadas correspondentes ao Português Brasileiro
** Consoantes vozeadas correspondentes ao Português Brasileiro

Dentro do primeiro grupo classificatório de consoantes estão as oclusivas. São consoantes caracterizadas por um bloqueio inicial do fluxo de ar devido à obstrução completa da cavidade vocal, que em seguida é liberado por uma explosão de ar. Suas propriedades são altamente influenciadas pelas vogais que as precedem, portanto, as consoantes oclusivas são por natureza dinâmicas. Os fonemas oclusivos vozeados /b/, /d/ e /g/ são transitórios e produzidos por uma pressão posterior à constrição. Este fechamento pode ser labial, dental ou velar. Durante o período de obstrução, uma pequena quantidade de energia de baixa frequência gerada nas cordas vocais é irradiada através das paredes da garganta. As características acústicas são representadas na Figura 10(a). Existem também, os fonemas oclusivos não vozeados /p/, /t/ e /k/ não apresentam vibração nas cordas vocais durante o período de obstrução, ou seja, há uma ausência de energia acústica. A Figura 10(b) exemplifica este comportamento. Em geral, as consoantes oclusivas têm uma duração média de 50ms a 150ms durante a oclusão e uma duração menor de 50ms durante a etapa de explosão de ar.

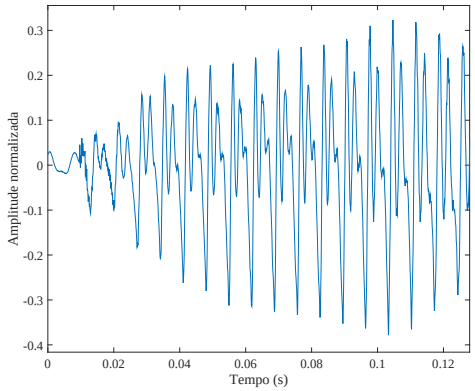
Continuando com o grupo de consoantes estão as nasais /m/, /n/ e /ɲ/. Estas são produzidas por uma excitação na área das cordas vocais e uma total constrição em algum ponto da cavidade oral, gerando uma circulação do ar dentro da cavidade nasal. Do mesmo modo que as vogais nasais, as ressonâncias vogais e nasais geram resultados espectralmente amplos, com altas concentrações de energia nas frequências baixas. A Figura 11 ilustra estas características.



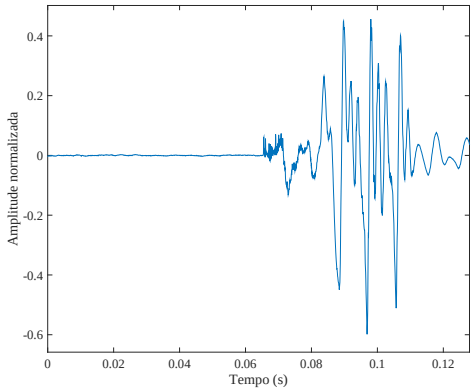
(a) Oclusiva não vozeada /t/



(b) Oclusiva vozeada /d/



(c) Forma de onda da consoante oclusiva não vozeada /t/



(d) Forma de onda da consoante vozeada /d/

Figura 10: Características acústicas das oclusivas

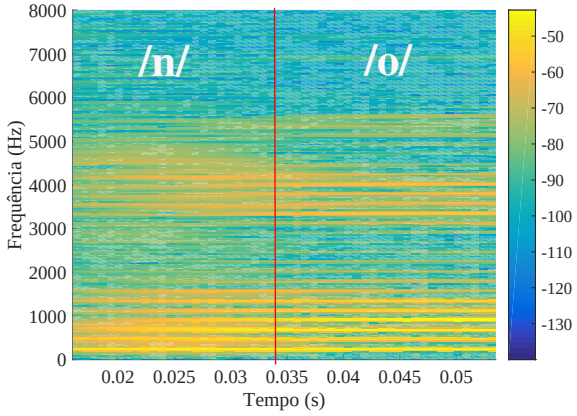
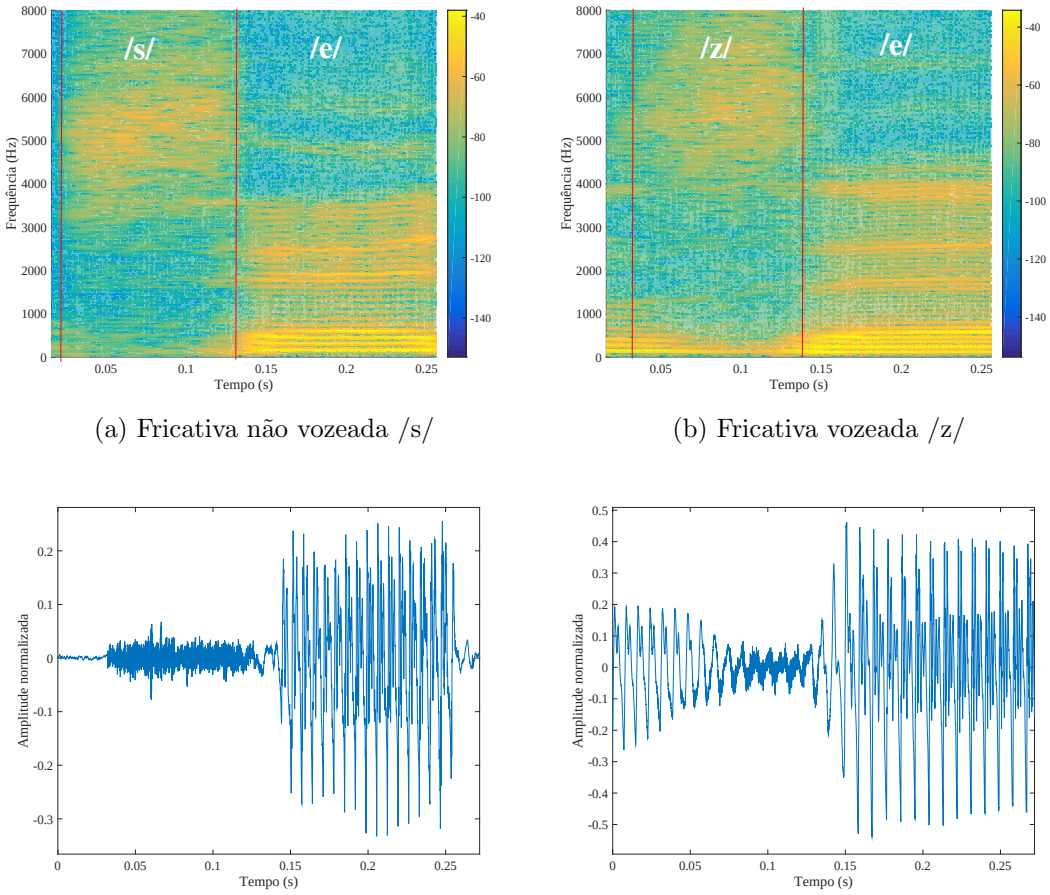


Figura 11: Espectrograma da silaba /no/

As consoantes fricativas se articulam forçando o ar através de um espaço estreito criado pela fricção entre dois articuladores, mas sem interromper o fluxo de ar. As fricativas não vozeadas /f/, /s/, /ʃ/, /x/ e /h/ são produzidas por um fluxo de ar constante

que gera uma excitação do trato vocal. Este fluxo de ar torna-se turbulento devido a uma constrição na cavidade vocal. As consoantes fricativas vozeadas /v/, /z/, /ʒ/, /ɣ/ e /ɦ/ têm os mesmos locais de constrição que as não vozeadas, porém, na produção das fricativas vozeadas, as cordas vocais vibram, gerando duas componentes diferentes, onde há presença de elementos turbulentos e energias distribuídas no espectrograma de acordo com o fonema falado. A Figura 12 exemplifica o comportamento acústico deste tipo de consoantes.



(a) Fricativa não vozeada /s/

(b) Fricativa vozeada /z/

(c) Forma de onda da consoante fricativa não vozeada /s/

(d) Forma de onda da consoante vozeada /z/

Figura 12: Características acústicas das consoantes fricativas

Por último, têm-se as consoantes pequenas que representam as variabilidades expostas entre os diferentes dialetos do Português Brasileiro. Entre estas estão as consoantes africadas /ʧ/ e /tʒ/, caracterizadas por ser sons da fala que mistura as propriedades dos grupos oclusivos e fricativos, onde a articulação inclui uma fase de obstrução total seguida de uma fase de fricção.

As consoantes vibrantes são caracterizadas por um movimento vibratório do articulador ativo sem interromper o fluxo de ar. No caso do fonema /r/, esta vibração é feita na zona alveolar. Este tipo de fonema é amplamente usado na zona Amazônica Brasileira.

A consoante catalogada como tepe, /ɾ/ tem um tipo de obstrução leve e alveolar. É um tipo de representação sonora da letra r amplamente usada nos dialectos derivados do Português e falados dentro do território Brasileiro.

A consoante aproximante /ɹ/ é um fonema cuja emissão de ar é obstruída mas não o suficiente ao ponto de constituir uma fricativa. É amplamente usada nos estados do Sudeste Brasileiro como uma representação sonora da letra r. É vinculado principalmente com o Português Paulista.

Finalmente, as consoantes laterais aproximantes /l/ e /ʎ/ são geradas quando a língua produz um bloqueio central e desce nas laterais para formar canais onde o ar tem um fluxo permanente. Podem ser geradas na zona alveolar para o caso do fonema /l/ ou na zona palatal, para o fonema /ʎ/.

1.4 Considerações finais do capítulo

Neste capítulo foram apresentados os conceitos básicos e desafios correspondentes ao reconhecimento automático de fala. O reconhecimento de fala proporciona uma interação apropriada entre o homem e a máquina através de ordens faladas. Também foram explicados de forma geral, alguns problemas e tópicos que abrangem os sistemas de reconhecimento de fala para o Português Brasileiro. Vale destacar que o desenvolvimento desta dissertação está centrado no reconhecimento e classificação dos sons do Português Brasileiro usando um sistema de RAF independente do locutor e sem interpretação gramatical. O capítulo seguinte apresenta um estudo sobre os principais trabalhos realizados com técnicas de reconhecimento de fala usando redes neurais artificiais.

Capítulo 2

TRABALHOS RELACIONADOS

ESTE capítulo, organizado em três seções, apresenta alguns trabalhos relacionados ao reconhecimento automático de fala usando redes neurais artificiais e sua implementação em dispositivos inteligentes. Na Seção 2.1 são apresentadas as arquiteturas clássicas de redes neurais artificiais ou RNA, usadas como sistemas de reconhecimento automático de fala ou ASR (*Automatic speech recognition*), de acordo com as características da linguagem a serem classificadas. Na Seção 2.2 são expostas as novas arquiteturas de RNAs, desenvolvidas como reconhecedores de fala, que proporcionam um maior abarcamento do problema de ASR. Por último, na Seção 2.3 são expostos os trabalhos que incluem desenvolvimento de sistemas ASR utilizando dispositivos inteligentes.

2.1 Classificação usando redes neurais artificiais

Redes neurais artificiais vêm sendo usadas como um método popular em aplicações de reconhecimento automático de fala. As redes neurais artificiais dedicadas ao ASR usam uma aproximação discriminativa por meio de aprendizagem de máquina. Os trabalhos relacionados discutidos nesta seção estão organizados cronologicamente e considerando os elementos de fala a serem classificados: vogais, fonemas e palavras.

2.1.1 Classificação de vogais

As vogais são os primeiros elementos linguísticos que foram classificados, devido à facilidade de obter amostras de fala para alimentar processos de treinamento. De acordo com a abrangência do problema de ASR, a importância das vogais na classificação e representação de palavras escritas é baixa (RABINER; JUANG, 1993), já que não geram uma

informação compreensível e completa do vocabulário de um idioma. Porém, estas são usadas para extrair informações tais como a identificação do falante e os estados anímicos.

Em (KNAGENHJELM; BRAUER, 1990) os desempenhos de dois sistemas neuronais artificiais para a classificação de vogais são comparados. Os classificadores empregados são:

- Uma rede neural artificial do tipo MLP (*Multilayer Perceptron*) de uma camada escondida com 20 neurônios. A rede usa 40 entradas referentes à representação dos sinais de fala e fornece 12 saídas referentes às classes classificatórias.
- Um mapa bidimensional auto-organizável de Kohonen ou SOM (*Self-Organization Map*) de 16×16 neurônios.

Um banco de dados próprio referente a um único locutor profissional masculino é usado nas etapas de treinamento e teste. Doze vogais são extraídas e segmentadas manualmente para gerar um vetor de coeficientes cepstrais com janelas de 16 ms. Neste caso, os resultados foram mais favoráveis para a rede MLP que obteve uma taxa de acerto de 83.7% enquanto a rede SOM obteve uma taxa de acerto de 82.5%. Porém, o fato de usar somente amostras obtidas de um falante profissional faz que os resultados sejam menos generalizáveis e tenham uma aplicação reduzida em ASR.

Em (COSI; BENGIO; MORI, 1990) é descrito um método de classificação de 10 vogais do inglês usando uma arquitetura MLP. O sistema consiste em dois módulos principais: O primeiro é um modelo matemático do funcionamento do ouvido usado para extrair as características do sinal de entrada; enquanto o segundo módulo é um MLP usado como reconhecedor de vogais. O sistema é independente do falante, mas tem um ambiente de gravação fixo. O algoritmo de treinamento é de retro-propagação clássica. Bons resultados na classificação das vogais são obtidos usando uma codificação binária das saídas desejadas. Uma particularidade deste trabalho é que a rede é treinada até atingir 0% de erro. Isto pode levar a um problema de sobre-treinamento (*overfitting*) e consequentemente uma redução do desempenho da rede usando um conjunto de teste.

Em (ZAHORIAN; NOSSAIR, 1999) é descrito um método de partição do processo de classificação dividindo a classificação de várias classes em múltiplos classificadores de duas classes. Neste trabalho são classificadas as vogais da base de dados TIMIT, onde é aplicada uma nova técnica de extração de características do sinal de voz usando um

marco matemático unificado que permite escalas de frequência não lineares em termos de amplitude e tempo. Assim, o sistema gera uma representação compacta das características espectrais e temporais do sinal de voz. Neste caso, o sistema inteligente é composto de múltiplas redes do tipo MLP treinados usando o algoritmo de retro-propagação. Os resultados obtidos são muito bem descritos e a análise é ampla e adequada. Porém a abrangência do problema do ASR é baixa devido à pouca generalização para outros fonemas e componentes do ASR.

2.1.2 Classificação de fonemas

O fonema é a menor unidade sonora na linguística. Os fonemas incluem vogais, semivogais e consoantes, os quais configuram um grupo amplo de classes. Portanto, as configurações das redes neurais artificiais usadas no reconhecimento de fonemas são mais complexas e representam um grande desafio de classificação, já que, dentro do novo grupo de classes, há elementos fonéticos com características similares que podem afetar a acurácia (RABINER; JUANG, 1993).

Em (KONG; KOSKO, 1991) são propostas novas arquiteturas de redes neurais artificiais utilizando aprendizagem competitivo diferencial DCL (*Differential Competitive Learning*) para obter uma estimativa dos centroides de classificação no ASR. A estrutura proposta é similar aos mapas SOM lineares com conexão total e inibição lateral. A contribuição desta proposta é o método de aprendizagem que consiste em uma combinação de método competitivo e o Hebbiano diferencial. Contudo, só são usados somente 9 fonemas facilmente classificáveis (/a/, /e/, /i/, /o/, /u/, /f/,/s/, /n/, /t/). A base de dados usada para realizar o treinamento é própria e se refere a um único falante masculino. Em quanto aos resultados, foi possível concluir que todas as vogais são bem reconhecidas, mas ocasionou uma taxa de erro considerável para a classificação dos fonemas /f/ e /t/. Além disso, é um sistema de classificação estático que não usa elementos temporários na classificação.

Com uma melhor visão sobre a dinâmica da voz, em (HANES; AHALT; KRISHNAMURTHY, 1994) é realizado um reconhecimento de fonemas usando uma rede neural artificial recorrente de Elman ou RNN (*Recurrent Neural Network*) (HAYKIN et al., 2009). São usadas três RNNs, cada uma permitindo três entradas e três saídas, representando as possíveis vogais e consoantes. Cada rede inclui uma camada escondida. A taxa de

acerto apresentada nos resultados é satisfatória para uma rede de camada escondida de 3 neurônios.

2.1.3 Classificação de palavras

A palavra é definida como um conjunto de fonemas. É reconhecida como a unidade básica de linguagens naturais, já que representa partes do pensamento humano. Assim, a classificação de palavras além de depender das características de cada janela do sinal de entrada, terá uma dependência do tempo que indica a mudança de fonema formante da palavra.

Seguindo esta linha de pesquisa, em (GRAMSS; STRUBE, 1990) propõe-se um reconhecimento de dígitos em alemão, usando um sistema de pré-processamento baseado na psicoacústica, onde o reconhecedor é uma rede do tipo MLP de 730 neurônios sem camada escondida. As dez saídas permitidas correspondem aos dígitos. As porcentagens de acerto são elevadas, mas o alcance do problema de ASR é limitada, já que devido ao número de neurônios usados neste trabalho é pouco prático implementar este sistema sobre uma quantidade considerável de palavras.

Em (WOODLAND; SMYTH, 1990) comparam-se 4 métodos de classificação estática para o reconhecimento de duas palavras em Inglês *yes* e *not*. Os classificadores são baseados no método *k-means*, misturas Gaussianas, MLP e SOM. São analisadas várias configurações juntamente com os respetivos resultados e o custo computacional de cada técnica. Os melhores resultados foram obtidos por uma rede MLP que tem uma estrutura simples.

Em (CHEN; CHEN; LIN, 1996) é usada uma sequência de MLP ou SMLP (*Sequence of Multilayer Perceptron*) como reconhecedor de palavras. Cada rede MLP é empregada como um reconhecedor que identifica um determinado evento acústico. Em seguida, as redes MLP são organizadas como uma cascata serial de todos os estados para configurar o reconhecedor de palavras. Para o treinamento de cada rede é usada uma combinação de descida probabilística generalizada GPD (*Generalized Probabilistic Descent*) e programação dinâmica DP (*Dynamic Programming*) . O banco de dados usado é fornecido pela TL (*Telecomunication Laboratories*) e consiste em 100 pessoas que falam repetidamente os dez dígitos em mandarim. A extração de características do sinal de voz é feita usando uma análise temporal DTW (*Dynamic Time Wraping*) . Os resultados desta implementa-

ção são comparados com as saídas de uma implementação do processo baseado em HMM, (*Hidden Markov Models*) , onde é concluído que a implementação SMLP requer menos parâmetros e poder computacional que a implementação usando HMMs.

2.1.4 Classificação de outras características da fala

Em (WANG; LIAO; CHEN, 2002) é proposto um sistema reconhecedor de características prosódicas da fala em Mandarim usando uma rede neural recorrente ou RNN *(Recurrent Neural Network).* Este modelo é usado como um pós-processo do reconhecimento de sílabas e visa ajudar na transcrição de fala para texto, identificando palavras monossilábicas ou a posição da sílaba estudada dentro de uma palavra.

O método de modelagem prosódico abrange duas etapas: A etapa de treinamento e a de teste. Durante a etapa de treinamento, cada frase usada é processada por um decodificador acústico para segmenta-la em sílabas contidas nas palavras incluídas na frase. Assim, as características prosódicas são subtraídas baseando-se na melhor segmentação silábica. Com esta informação, a rede neural recorrente é treinada para aprender a relação entre as características prosódicas de entrada e os limiares da palavra associada no texto. Durante a etapa de teste, a rede recorrente é usada para indicar os limiares da palavra estudada que finalmente é tratada por uma máquina de estados usada para discriminar entre indicações confiáveis e indicações incertas.

As experiências feitas durante este trabalho utilizaram um banco de dados de 652 gravações usando um vocabulário amplo e um falante masculino. Os testes usam um modelador acústico baseado em um processo de HMM (*Hidden Markov Model*) que remite as características silábicas do sinal de entrada. Enquanto à rede neural recorrente, é usado um modelo de uma única camada escondida que contém 30 neurônios e uma retroalimentação entre a camada escondida e a de entrada. A quantidade de neurônios foi determinada empiricamente.

Os resultados obtidos indicam uma melhoria da acurácia usando a máquina de estados como validador do reconhecimento feito pela RNN, onde os limiares impostos pela máquina de estados incrementam em um 17% a acurácia da rede neural recorrente.

2.2 Novas arquiteturas

Na atualidade, a disponibilidade de sistemas computacionalmente mais poderosos tem permitido o desenvolvimento e a implementação de reconhecedores de fala para um vocabulário amplo usando algoritmos mais complexos e mais precisos. Isto implica projetar redes neurais artificiais que possam manipular dinamicamente grandes quantidades de informação para obter uma resposta de classificação satisfatória.

Em (CECCARELLI; HOUNSOU, 1996), os autores usam uma rede com funções de ativação de base radial ou RBF (*Radial Basis Function*) para o reconhecimento de palavras isoladas. É feita uma alteração sobre a RBF que consiste em acrescentar algumas conexões com atraso e uma camada de integração baseada no tempo para capturar a natureza dinâmica da voz. Este novo modelo é identificado como RBF com atraso no tempo.

Além da camada de distribuição, a rede tem outras três camadas. Na primeira camada inclui os neurônios com ativação RBF. A segunda camada é ativada com funções sigmoides e finalmente na última camada é feita uma integração temporal das saídas da segunda camada. Entre a primeira e segunda camada, além das conexões normais, também existem dois conjuntos de conexões com atraso no tempo. Durante o processo de classificação, todos os vetores de entrada são processados pela rede e a saída com maior peso determina a palavra classificada.

As experiências realizadas consistem em comparar a rede com várias modalidades de treinamento e uma rede neuronal com atraso no tempo ou TDNN (*Time Delay Neural Network*). Para estes ensaios, foi usado um banco de dados que contém gravações com dez dígitos pronunciados várias vezes por 60 falantes Italianos. Cada palavra é transformada em 10 diferentes vetores de características.

Durante a primeira experiência são calculados os centroides do núcleo radial usando um treinamento não supervisionado de mapas SOM, onde os pesos da camada com sigmoides são ajustados usando uma minimização da média de erro quadrático. A segunda experiência substitui a técnica SOM por um treinamento supervisionado com aprendizagem Levenberg-Marquardt. Durante o terceiro teste, é usado um algoritmo generalizado para redes de base radial com diferentes combinações de neurônios na camada de RBF. Por último é usada uma rede TDNN.

As médias das melhores acurácias das experiências são de 94.8, 95.5, 95.1 e 89.8 respetivamente.

Os sistemas de aprendizagem clássicos como as redes neurais artificiais podem ter problemas de múltiplos mínimos locais. Isto gera uma dificuldade na classificação de elementos linguísticos presentes em um ambiente ruidoso. Em (SOLERA-UREÑA et al., 2007), é usado um modelo baseado em máquinas de vetores de suporte ou SVM (*Support Vector Machine*) para o reconhecimento de dígitos isolados em Espanhol falados em um ambiente caracterizado por diferentes ruídos. A arquitetura do sistema SVM é inspirada em uma rede do tipo MLP e procura maximizar a margem de separação entre amostras de diferentes classes em um espaço de decisão de múltiplas dimensões. Assim, o sistema é mais robusto na presença de ruídos de fundo e consegue uma melhor generalização durante a etapa de validação. O sistema SVM emprega uma função de ativação conhecida como *kernel* que modifica a dimensionalidade dos dados de entrada, procurando um espaço que permita uma margem de separação máxima entre as amostras de diferentes classes. Porém, neste caso a fala produz diferentes dimensões de entrada o que dificulta o uso de um classificador SVM.

Para mitigar este problema, os autores empregam uma função de normalização conhecida como *kernel* de alinhamento automático de tempo ou DTAK (*Dynamic Time Alignment Kernel*) que resolve o problema de sequências de longitudes diferentes para que o SVM possa operar com amostras de dimensionalidades diferentes.

As experiências basearam-se em dois bancos de dados: O *Spanish SpeechDat* e um banco de dados próprio que contém 72 falantes que repetem 11 vezes cada um dos dígitos de zero até nove. O banco de dados próprio é usado para o treinamento e os sinais do *Spanish SpeechDat* são usados para validação.

A implementação do sistema inclui diferentes ruídos de fundo e os resultados comparam o desempenho de um sistema SVM de kernel DTAK com um sistema hibrido de HMM com SVM, onde o HMM normaliza os vetores de entrada do classificador SVM. Assim, o sistema SVM-DTAK obteve maior acurácia que o sistema HMM-SVM dentro de ambientes com relação sinal-ruído ou RSR baixa, porém, para ambientes com alto RSR, a diferencia de acurácia destes dois sistemas é mínima.

Em (SINISCALCHI et al., 2013) é empregada uma rede neural profunda ou DNN (*Deep Neural Network*) para aumentar a acurácia de classificação das unidades básicas da fala como os atributos fonéticos e fonemas. Para medir o desempenho desta rede foram avaliadas várias arquiteturas de redes neurais profundas que variam desde 5 até 7 camadas

escondidas, onde cada camada contém 2048 neurônios. Também foi implementada uma rede do tipo MLP de uma camada escondida para fins de comparação com os resultados obtidos a partir das redes profundas.

As experiências foram aplicadas sobre o banco de dados WSJ0 que contém 5000 palavras com falantes independentes. As características destas gravações foram obtidas usando análise espectral para gerar 12 coeficientes cepstrais, também conhecidos como MFCCs (*Mel-Frequency Cepstrum Coefficients*).

Para o treinamento da rede neural profunda, primeiro foi usado um treinamento não supervisionado conhecido como máquina restrita de Boltzmann ou RBM (*Restricted Boltzmann Machine*). Esse processo permite gerar uma matriz com os pesos iniciais da rede neural profunda evitando um sobre-treinamento da rede, devido à quantidade alta de camadas escondidas. Em seguida, a rede é treinada usando o algoritmo de retro-propagação, comparando a entropia cruzada entre a saída da rede que é dado pela função *softmax* e a distribuição de probabilidade predeterminada para cada saída.

Os resultados demonstram melhor desempenho das redes neurais profundas na classificação de atributos fonéticos proporcionando uma taxa de acurácia superior à 90% para todas as classes de saída do sistema. Da mesma maneira, é verificada a diferencia de acurácia entre uma rede de cinco camadas ocultas e uma de sete camadas com o fim de constatar a influência do número de camadas e o desempenho da rede. Assim, é demonstrado que a diferença entre suas acurácias é mínima, de tal forma que a rede com 5 camadas ocultas é considerada melhor tendo em vista sua simplicidade.

Em (ABDEL-HAMID et al., 2014), os autores usam uma adaptação das redes neurais convolucionais ou CNN (*Convolutional Neural Network*) para o reconhecimento automático de números telefónicos em pequena escala e reconhecimento de palavras em um amplo vocabulário.

As redes convolucionais são uma variante das redes neurais profundas que em vez de uma conexão de camadas escondidas, introduz uma estrutura de rede especial, que consiste em alternar dois tipos de camadas conhecidas como camada de convolução e camada de agrupamento. Os dados de entrada precisam ser organizados em mapas característicos compostos por arranjos de duas dimensões que representam os elementos espectrais do sinal de entrada. Deste modo, a arquitetura da rede neural convolucional contém um número de atributos localizados em um mapa de características, onde a resolução destes

mapas fica menor em camadas superiores à medida que são aplicadas mais operações de convolução e agrupamento. Normalmente uma ou mais camadas ocultas totalmente ligadas são adicionadas no topo da camada final da rede convolucional, a fim de combinar as características em todas as bandas de frequência antes da propagação para a camada de saída.

As experiências utilizaram o banco de dados TIMIT, onde cada sinal é pre-processado usando um método de extração espectral de características conhecido como MFSC (*Mel Frequency Spectral Coefficents*) para depois implementar dois tipos de redes neurais: Uma rede neural profunda com três camadas escondidas e uma outra rede neural convolucional com uma par de camadas de convolução e agrupamento, acrescidas de mais duas camadas escondidas totalmente conectadas. Os resultados obtidos demonstram uma melhoria na taxa de erro de 8% na comparação com uma rede profunda. Contudo, a implementação das redes convolucionais aumenta o custo computacional para a análise e classificação de bancos de dados com um amplo vocabulário.

2.3 Implementações em dispositivos inteligentes

A principal dificuldade na implementação de reconhecimento automático de fala em dispositivos inteligentes é trabalhar eficientemente com as restrições de uma plataforma embutida. Os dispositivos embutidos têm recursos computacionais modestos como a frequência de operação do processador e o tamanho de sua memória, enquanto que sistemas ASR normalmente envolvem uma computação exaustiva. Apesar destas dificuldades, muitas empresas têm desenvolvido ferramentas de reconhecimento de fala que usam algoritmos eficientes. Estas implementações reduzem o tamanho do modelo e a complexidade da codificação para oferecer um desempenho aceitável para reconhecimento de fala.

Em (VARGA et al., 2002) é apresentado um sistema de reconhecimento automático de fala baseado em HMM, para dispositivos celulares, conhecido como VSR (*Very Smart Recognizer*). Devido ao fato que celulares modernos integram um processador digital de sinais ou DSP e um microcontrolador, os autores sugerem dividir o sistema para que a parte de extração de atributos do reconhecedor seja feita sobre o processador de sinais enquanto a decodificação das características seja tratada pelo microcontrolador. Os procedimentos de optimização são relacionados com à modelagem acústica do sinal de entrada e a emissão de probabilidades computacionais requisitadas pelo módulo de decodificação.

O sistema VSR obteve uma importante redução de consumo de memória e uso de recursos de processamento, já que este usa modelos fonéticos dependentes do contexto e um algoritmo de distribuição clusterizado do subespaço em HMM. Para uma tarefa de reconhecimento de palavras isoladas com vocabulário reduzido, os modelos acústicos de HMM necessitaram 12Kbytes de memória flash e geraram uma taxa de erro de palavra de 7%.

Uma aproximação similar de implementação de ASR é apresentado em (WANG et al., 2004), onde um sistema de reconhecimento de fala referente a um vocabulário reduzido é implementado em um microcontrolador S3CB519 da Samsung. O sistema usa um procedimento híbrido de HMM e redes neurais artificiais, onde as probabilidades de distribuição dadas pelo sistema HMM são post-processados por uma rede neural classificadora. A taxa de reconhecimento é de 97% para uma classificação de 300 palavras isoladas.

2.4 Considerações finais do capítulo

Neste capítulo foram apresentados resumidamente diversos trabalhos sobre reconhecimento automático de fala usando redes neurais artificiais, sendo estes utilizados como classificadores de diferentes características e elementos que compõem a linguagem falada. Também foram apresentados alguns trabalhos relacionados com as últimas pesquisas que visam melhorar a acurácia no reconhecimento de palavras quando o vocabulário é extenso. Finalmente, foram apresentados alguns trabalhos relacionados com as considerações que devem ser feitas durante a implementação do sistema reconhecedor de fala em um sistema inteligente. O capítulo seguinte apresenta detalhadamente a estrutura do pré-processamento do sinal e extração de características propostos nesta dissertação.

Capítulo 3

PRÉ-PROCESSAMENTO E EXTRAÇÃO DE CARACTERÍSTICAS DO SINAL DE FALA

ESTE capítulo apresenta as técnicas de pré-processamento e de extração de características usadas em sistemas RAF propostos nesta dissertação. Os procedimentos aqui expostos visam melhorar as propriedades acústicas do sinal de fala entrante e condicionar os dados, a fim de extrair os atributos característicos necessários para a classificação de padrões, optimizando a identificação dos fonemas falados.

Um sistema RAF clássico é composto principalmente por três etapas principais apresentadas na Figura 13, onde a entrada é um sinal contínuo no tempo, acústico e com conteúdo de fala. Por outro lado, a saída é um texto representativo da fala configurada no sinal de entrada. Ambas caracterizações estão em função do tempo (MPORAS et al., 2007).

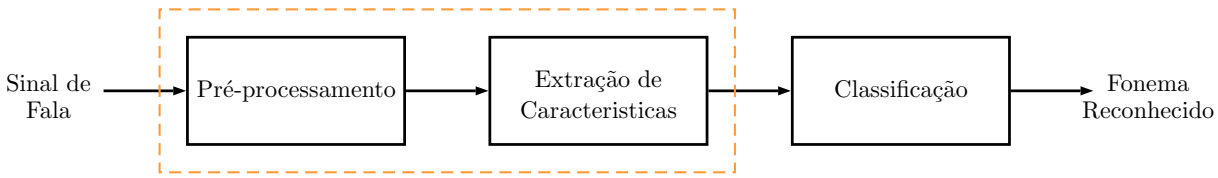


Figura 13: Esquema básico de um sistema RAF

As etapas de pré-processamento e extração de características condicionam o sinal de entrada para a etapa de reconhecimento e classificação. Os coeficientes obtidos na etapa de extração de características devem conter toda a informação pertinente que ajude ao reconhecimento dos fonemas, mas também deve-se considerar a redução máxima das

informações redundantes para evitar um incremento no tamanho da arquitetura de reconhecimento. Logo, estes processos influenciarão diretamente na qualidade dos resultados obtidos na etapa de classificação, assim como na eficiência do processo de reconhecimento.

3.1 Pré-processamento

A etapa de pré-processamento de um sinal é extremamente importante para um reconhecimento correto da fala. O pré-processamento melhora as propriedades acústicas do sinal de fala entrante, aprimorando as componentes mais significativas do sinal de fala antes de executar a etapa de extração de características. As funções implementadas neste estágio dependem da aproximação usada na etapa de extração de características (DELCROIX; NAKATANI; WATANABE, 2009). Usualmente são usados removedores de ruído, detectores de início e fim de palavras, filtros de pré-ênfase, janelamento e normalização da amplitude da onda entrante (SAHA; CHAKROBORTY; SENAPATI, 2005). A Figura 14 mostra as funções incluídas no pré-processamento do sistema RAF proposto.

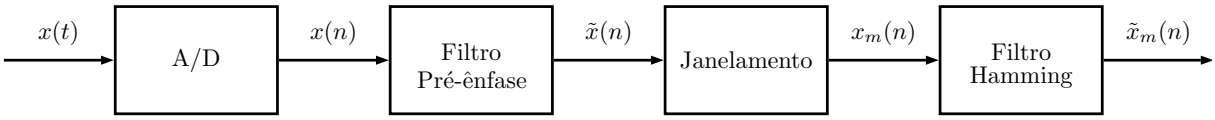


Figura 14: Etapas do preprocessamento sobre o sinal de entrada.

onde é apresentada a transformação do sinal analógica $x(t)$ em cada etapa do pré-processamento inicial. A componente de conversão analógico-digital amostra o sinal, gerando um sinal discreto $x(n)$, logo é filtrada mediante um filtro pré-ênfase. O sinal de saída $\tilde{x}(n)$ é janelada, produzindo um vetores $x_m(n)$, sendo m índice de cada vetor. Por último, o cada vetor de entrada m é filtrado novamente, gerando uma saída $\tilde{x}_m(n)$.

3.1.1 Conversor analógico-digital

Devido à necessidade de interpretação do sinal de fala entrante por parte de dispositivos digitais, primeiramente, o sinal análogo é transformado em um sinal digital usando métodos de conversão analógico-digital ou A/D. A fala digitalizada é uma sequência de bits com uma frequência de amostragem constante, onde cada elemento é quantizado de acordo com uma resolução de bits pré-estabelecida e uma codificação sem perda.

Para começar, é necessário obter amostras periódicas da amplitude de onda do sinal $x(t)$. A Figura 15 apresenta o processo de digitalização de um sinal analógico $x(t)$. A

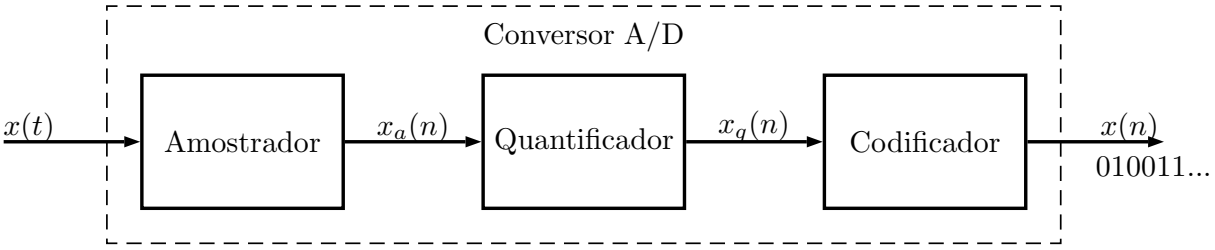


Figura 15: Esquema de conversão de um sinal analógico para um sinal digital

velocidade em que são obtidas as amostras em um segundo é conhecida como frequência de amostragem. De acordo com o teorema de Nyquist, a frequência de amostragem mínima necessária para reconstruir o sinal com um mínimo de perda deve ser igual o maior à duas vezes a largura de banda do sinal de entrada (SHANNON, 1949), conforme a Equação 1:

$$f_s \geq 2B, \tag{1}$$

onde f_s é a frequência de amostragem e B é a largura de banda do sinal $x(t)$. A largura de banda da fala humana abrange as frequências entre 80Hz e 14KHz. Porém, os sistemas de telefonia atuais operam em uma faixa de frequências entre 50Hz e 7KHz (MARTIN; HEUTE; ANTWEILER, 2008) A condição estabelecida pelos sistemas de telefonia e a Equação 1 infere uma frequência de amostragem $f_s = 16$KHz, que é compatível com a maioria dos dispositivos inteligentes atuais.

O processo de quantização mede o nível de tensão de cada sinal amostrado $x_a(n)$, onde n é a amostra discreta equivalente ao tempo t. Para que os valores de amplitude das amostras possam ser representados por uma quantidade finita de bits, é necessário determinar um número finito de representações possíveis das amostras (GIBSON, 2012). Assim, a faixa de valores possíveis do sinal é subdividida em intervalos de quantização. Cada intervalo é definido por um nível de quantização correspondente a um valor de tensão médio do intervalo. A quantidade de dados discretos usados na representação do sinal amostrado $x_a(n)$ é dependente da memória do dispositivo reconhecedor e do erro de quantização tolerado. Note que maior é o número de níveis de quantização, menor será a distância entre eles, diminuindo assim o erro de quantização. A Equação 2 expõe a relação entre o sinal original e o ruído de quantização do sinal de saída ou SQNR (*Signal to Quantization Noise Ratio*):

$$\text{SQNR} = 20\log(2^q), \tag{2}$$

onde q representa o número de bits usados na quantização para um SQNR medido em decibéis ou dB.

O sinal quantizado $x_q(n)$ é então codificado como uma sequência de bits usando uma modulação por codificação de pulso ou PCM (*Pulse Code Modulation*), onde cada pulso de amplitude variável é transformado em uma sequência de bits com amplitude fixa e representação binária (GIBSON, 2012). Desta forma é obtido o sinal digital $x(n)$, possibilitando seu tratamento na etapa seguinte de filtragem.

3.1.2 Filtro de pré-ênfase

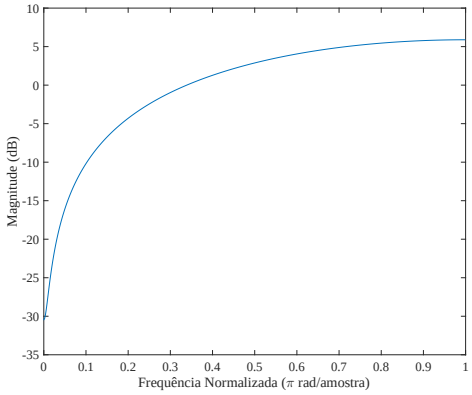
Um sinal que é transmitido usando diferentes canais de comunicação e posteriormente quantizado, está sujeito à atenuações da informação de altas frequências (MARTIN; HEUTE; ANTWEILER, 2008). A falta de dados causada pode alterar a eficiência do sistema reconhecedor, principalmente, na detecção daqueles fonemas com presença espectral nas frequências altas como as fricativas, conforme exposto no Capitulo 1.

O propósito de um filtro de pré-ênfase é incrementar a magnitude das altas frequências de acordo com o comportamento das componentes espectrais baixas, mediante a aplicação de um atraso e uma inversão no sinal digitalizado $x(n)$ para adicioná-lo ao sinal de entrada como é definido nas Equações 3 e 4:

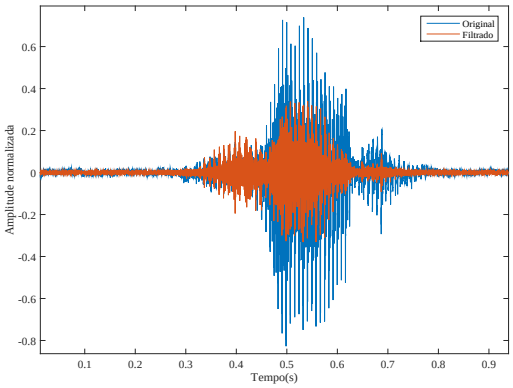
$$H(z) = 1 - az^{-1}, \qquad 0,9 \leq a \leq 1; \tag{3}$$

$$\tilde{x}(n) = x(n) - ax(n-1), \qquad 0,9 \leq a \leq 1; \tag{4}$$

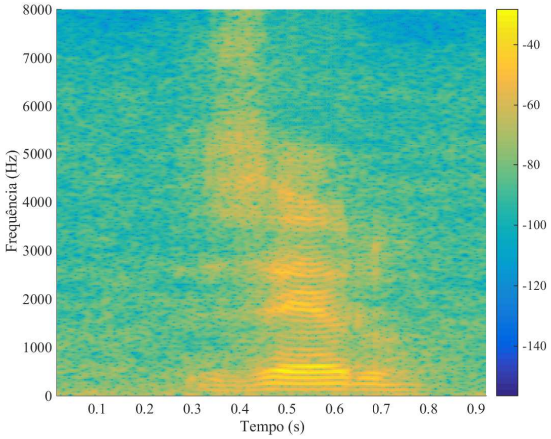
onde a variável a indica um peso de ajuste da correção do estado atual e $H(z)$ representa a função de transferência do sistema, caracterizado por ser um filtro de resposta ao impulso finito ou FIR (*Finite Response Impulse*) de primeira ordem, representado na Figura 16(a). Por outro lado, o sistema discreto exposto na Equação 4, é equivalente à função de transferência da Equação 3, onde $\tilde{x}(n)$ é o sinal de saída do filtro de pré-ênfase. Em suma, o sinal de saída $\tilde{x}(n)$ é nivelado espectralmente. Isto é evidenciado na comparação dos espectrogramas das Figuras 16(c) e 16(d), que representam o espectro do sinal de entrada e o sinal de saída, respectivamente. Com esta filtragem, o sinal $\tilde{x}(n)$ torna-se menos susceptível aos efeitos de precisão finita que serão gerados pelos processamentos posteriores no sinal de fala (RABINER; JUANG, 1999).



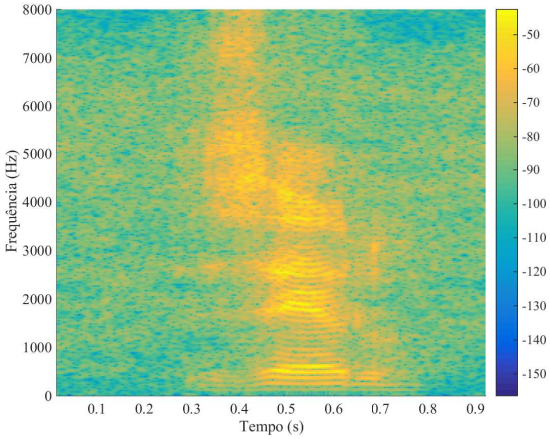
(a) Magnitude da resposta de frequência do filtro de pré-ênfase.



(b) Sinal original e filtrado que contém a palavra falada “cinco”



(c) Espectrograma do sinal original $x(n)$



(d) Espectrograma do sinal filtrado $\tilde{x}(n)$

Figura 16: Filtro de pré-ênfase aplicado ao sinal de entrada

3.1.3 Análise em tempo curto e janela de Hamming

Como já foi indicado no Capítulo 1, a fala é um sinal acústico aleatório. Porém, para usar funções de aproximações e análise é necessário obter sinais estacionários. Logo, o sinal deve ser dividido em pequenos blocos que possam ser tratados e analisados. Nesta etapa do pré-processamento, o sinal pré-enfatizado $\tilde{x}(n)$ é segmentado usando janelas de L amostras de comprimento. O tamanho das janelas deve obedecer à média do tempo usado para a articulação de um fonema, permitindo que cada vetor de saída m do sinal $x_m(n)$, tenha características quase-estacionários. Normalmente, a variável L representa o comprimento da janela segmentadora, configurada entre 320 a 400 amostras, sendo equivalente a 20ms e 25ms em termos de tempo para uma frequência de amostra de $f_s = 16$KHz. (RABINER; JUANG, 1999).

Cada segmento de sinal janelado deve manter uma correlação com seus blocos vizinhos. Portanto, é usada uma técnica de sobreposição de informação entre os blocos janelados. Para mudanças rápidas no sinal é usada uma sobreposição curta. Esta taxa de sobreposição é útil para sistemas com pouca memória e processadores com baixo desempenho. Porém, a pouca correlação entre os dados dos blocos vizinhos pode alterar as análises espectrais do sinal. Na média, de acordo com (RABINER; JUANG, 1999), cada bloco deve ter uma sobreposição de pelo menos 30% do tamanho da bloco anterior. A Figura 17, mostra o processo de segmentação de um sinal de fala, onde a quantidade de amostras superpostas esta denotada por S. Esta superposição corresponde a uma deslocação no eixo temporal de $L_{\text{shift}} = L - S$.

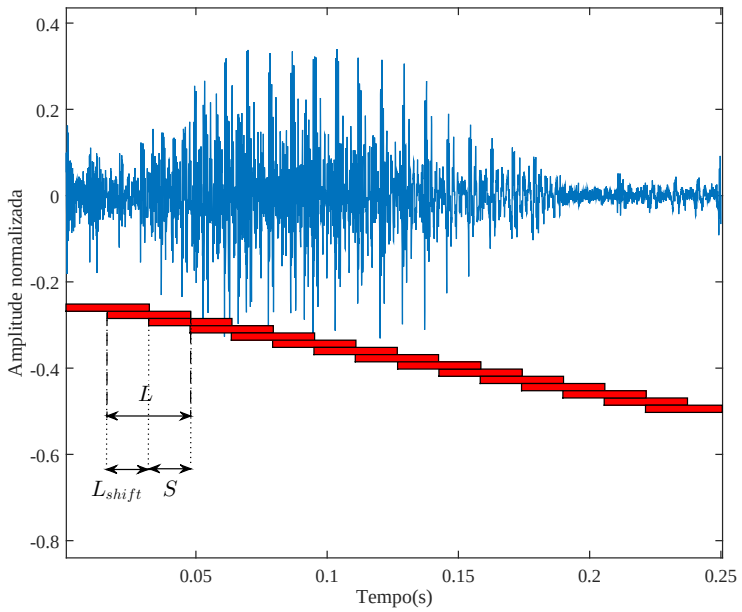


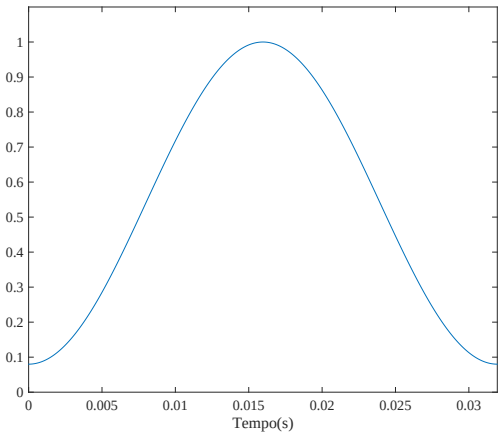
Figura 17: Processo de janelamento sobre um sinal de fala

O tipo de janela usada, conhecida como de *Hamming*, conforme mostra a Figura 18(a), filtra cada um dos sinais segmentados para evitar descontinuidades no início de cada bloco. Esta função é descrita pela Equação 5:

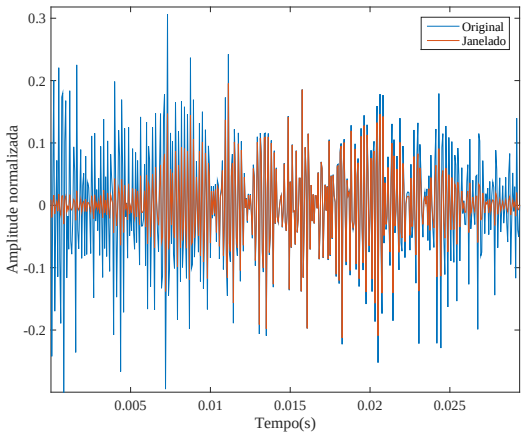
$$w(n) = \begin{cases} 0,54 - 0,46\cos\frac{2\pi n}{L-1} & 0 \leq n \leq L-1, \\ 0 & \text{senão.} \end{cases} \tag{5}$$

onde $w(n)$ é a função da janela, n é a amostra tratada e L é o comprimento do bloco. Cada segmento de $x_m(n)$ é filtrado pela função $w(n)$. A Figura 18(b) mostra o bloco original

comparado com o bloco filtrado, onde são notadas as atenuações feitas nas laterais do sinal segmentado.



(a) Representação da janela de Hamming



(b) Bloco de um sinal de fala original e o filtrado

Figura 18: Janela de Hamming no sinal de entrada segmentado.

3.2 Extração de características

A partir do sinal pré-processado, a etapa de extração de características deve escolher um número limitado de atributos que represente adequadamente o sinal de fala. As características extraídas devem permitir a discriminação dos diferentes elementos fonéticos e ser robustas frente aos sinais que apresentam perturbações externas como o ruído. Portanto, o desempenho de um reconhecedor automático de fala é altamente dependente do método de extração de características escolhido (ZAMANI et al., 2011).

Numerosas pesquisas relacionadas ao reconhecimento de fala utilizam o método de coeficientes cepstrais nas frequências de Mel ou MFCC (*Mel Frequency Cepstral Coefficients*) (CUTAJAR et al., 2013). O método MFCC imita a dinâmica do ouvido humano, onde as frequências são distribuídas de forma não linear no espectro de som. A Figura 19 representa todas as etapas envolvidas para obter as características necessárias para o treinamento e reconhecimento dos fonemas falados em tempo *online*. Cada uma destas etapas transforma os blocos que contém o sinal de fala, a fim de procurar as maiores concentrações de energia dentro do espectro de som. O sinal janelado correspondente ao bloco m, $\tilde{x}_m(n)$, é transformado para obter suas características frequenciais. Logo, é achada a densidade de potência, visando facilitar o filtrado usando um banco de filtros de

Mel. Assim a transformada logarítmica dos vetores filtrados é novamente transformado, de modo que possam ser diminuídos os vetores de representação de saída. Os detalhes destes processos são expostos nas próximas seções.

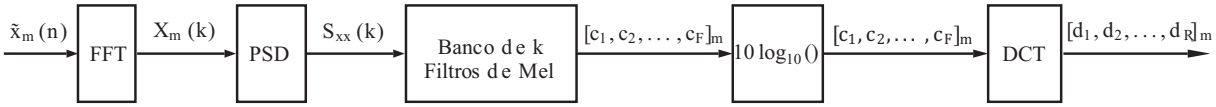


Figura 19: MFCC: Método de extração de características para o sistema RAF proposto

3.2.1 Transformada discreta de Fourier e densidade de potência espectral

A transformada discreta de Fourier ou DFT (*Discrete Fourier Transform*) é uma ferramenta utilizada para analisar as frequências de um sinal limitado, mediante a transformação de uma função matemática em outra. Desse modo é obtida uma representação espectral de um sinal discreto. A Equação 6 descreve a DFT para um sinal finito:

$$X_m(k) = \sum_{k=0}^{L-1} \tilde{x}_m(n) W_L^{nk}, \quad \text{com} \quad W_L = e^{-j2\pi/L} \tag{6}$$

onde $X_m(k)$ corresponde à DFT do bloco $\tilde{x}_m(n)$ e L é o comprimento do sinal de entrada.

A transformada rápida de Fourier FFT (*Fast Fourier Transform*) representa uma família de algoritmos usados para o cálculo rápido da DFT. O método *Radix-2* é o algoritmo mais simples usado para o cálculo da FFT. Este método calcula a DFT das amostras de entrada pares $f(n) = \tilde{x}_m(2n)$ e das amostras impares $g(n) = \tilde{x}_m(2n+1)$. Depois, a função combina os resultados para gerar a DFT de toda a sequência de amostras. A fim de realizar o cálculo recursivamente, se assume um comprimento do bloco L igual a uma potência de 2. A Equação 7 apresenta esta operação:

$$X_m(k) = \sum_{k=0}^{L/2-1} f(n) W_{L/2}^{nk} + W_L^k \sum_{k=0}^{L/2-1} g(n) W_{L/2}^{nk}, \quad 0 \leq n < L/2 \tag{7}$$

onde W_L é definida na Equação 6, n indica a amostra do bloco m processado e k representa a amostra de saída do vetor $X_m(k)$ de comprimento $K = L/2$. Normalmente, o cálculo da DFT requer L^2 operações iterativas, enquanto que um algoritmo FFT precisa somente de $L \log_2 L$ iterações (OPPENHEIM et al., 1989).

Finalmente, é calculada a densidade espectral de Potência ou PSD (*Power Spectral Density*) aplicada a cada vetor $X_m(k)$. Esta função representa a distribuição de

potência sobre as diferentes frequências, ou seja, seu espectro. A equação 8 expõe este procedimento:

$$S_{xx}(k) = |X_m(k)|^2, \tag{8}$$

onde $S_{xx}(k)$ representa o sinal de saída da função PSD sobre cada vetor $X_m(k)$.

Para uma melhor representação, cada vetor $S_{xx}(k)$ é transformado usando uma logarítmica por meio da função $10\log_{10}(S_{xx}(k))$. A Figura 20, apresenta as características espectrais obtidas pela PSD, onde é apresentado o comportamento das energias do sinal de fala no espectro.

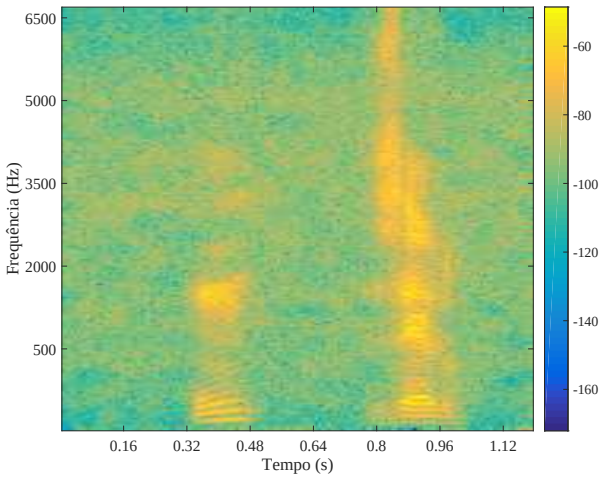


Figura 20: Densidade Espectral de Potência para um sinal de fala.

3.2.2 Banco de filtros de mel

O sistema de audição humana não interpreta as frequências de uma maneira linear. Para emular o funcionamento do ouvido, foi desenvolvida a escala de Mel. Esta representação visa descrever a escala de audição do ouvido em um sistema linear. Uma forma de relacionar a frequência linear com a frequência aparente é apresentada nas Equações 9 e 10:

$$M(f) = 1125\ln(1 + f/700), \tag{9}$$

$$M^{-1}(m) = 700e^{m/1125} - 1, \tag{10}$$

onde $M(f)$ é a frequência resultante em escala de Mel dada uma entrada f medida em Hz. A relação entre estas escalas de frequências é apresentada na Figura 21.

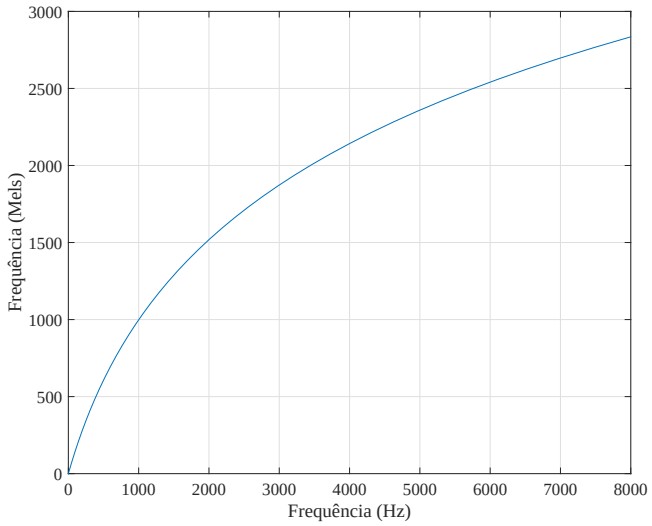
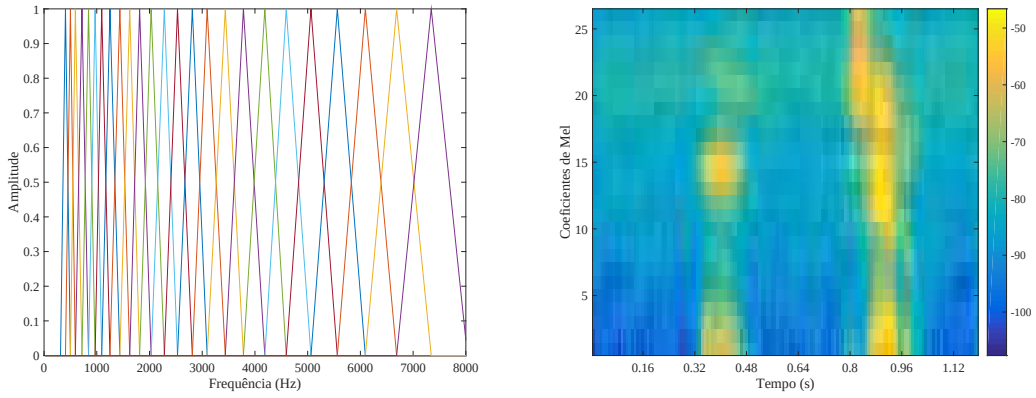


Figura 21: Relação entre a escala referente as frequências de Mel e a escala de frequências absolutas

Como já foi exposto na Seção 3.2.1, um vetor de saída da PSD tem um comprimento de $K = L/2$. Em particular, se a janela segmentadora tivesse um comprimento de $L = 512$ amostras, o sinal de saída da PSD teria um comprimento de $K = 256$. Isto representa vetores característicos com informação redundante, que incrementará o tamanho da arquitetura do reconhecedor de fala. Assim, para reduzir a quantidade de informação desnecessária, são determinados um grupo F de filtros passa-banda, normalmente 25 filtros (RABINER; JUANG, 1999), distribuídos uniformemente na escala de Mel. Note que isso na escala de frequências absolutas equivale a uma distribuição logarítmica, conforme mostrado na Figura 22(a).



(a) Banco de 25 Filtros de Mel distribuídos entre 300Hz e 8000Hz.

(b) Espectro do sinal filtrado com 25 filtros de Mel.

Figura 22: Banco de filtros de Mel filtrando o sinal com representação espectral

Finalmente, cada filtro concentra as energias contidas em um só valor, assim, de acordo com a quantidade F de filtros do banco, serão obtidos vetores que contém os coeficientes representativos de acordo à distribuição de Mel, $C_m = [c_1, c_2, c_3, ..., c_F]$. A Figura 22(b) apresenta um grupo de vetores distribuídos no tempo, cada um com 25 coeficientes de saída correspondentes a um sinal de fala entrante variante no tempo.

3.2.3 Transformada discreta de cosseno

A etapa final da extração de características consiste em aplicar a transformada discreta de cosseno ou DCT (*Discrete Cosine Transform*) a cada vector C_m, visando obter uma representação vectorial $D_m = [d_1, d_2, d_3, ..., d_R]$ de R coeficientes para cada vetor de entrada. Esta transformação está descrita pela Equação 11:

$$D_m(r) = \sum_{f=1}^{F} C_m(f) cos\left(\frac{r\pi(f-0.5)}{F}\right) \tag{11}$$

onde $D_m(r)$ representa o vetor dos coeficientes de saída da DCT para $0 \leq r \leq R$, F é o comprimento do vetor de coeficientes de entrada que é determinado de acordo à quantidade de filtros de Mel aplicados ao o sinal espectral, e R representa o número de coeficientes obtidos usando a DCT. De acordo com (RABINER; JUANG, 1999) é recomendado usar $R = 13$ coeficientes cepstrais, para garantir uma distribuição de energias com o mínimo de auto-correlação entre os coeficientes calculados. A Figura 23 apresenta a distribuição de energia para os coeficientes da DCT presentes em uma sequência temporal de vetores.

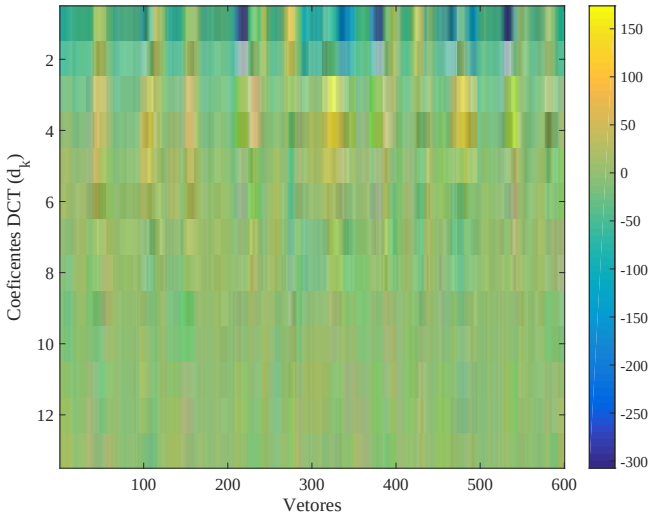


Figura 23: Representação da DCT aplicada a um sinal de fala.

3.3 Considerações finais do capítulo

Neste capítulo foram apresentadas as etapas de pré-processamento e extração de características referentes a um sinal de fala entrante. Cada uma das funções usadas durante estas etapas tem como objetivo melhorar a qualidade do sinal de fala e permitir uma melhor extração das características representativas da fala com a mínima correlação possível entre as características extraídas. A configuração de cada etapa deve garantir um tratamento adequado do sinal, onde cada coeficiente extraído deve possibilitar a máxima eficiência e acurácia do sistema reconhecedor de fala. No próximo capítulo serão descritos os detalhes da arquitetura do reconhecedor proposto.

Capítulo 4

REDES NEURAIS ARTIFICIAIS NO RECONHECIMENTO AUTOMÁTICO DE FALA

DENTRO dos estudos que abrangem os problemas presentes nos sistemas de RAF diferentes métodos de análise de eventos acústicos têm sido usados, incluindo técnicas topológicas, baseadas no cálculo e comparação de distâncias (WACHTER et al., 2007), métodos probabilísticos generativos de palavras contidas considerando um certo vocabulário (ZHANG; TOGNERI; ALDER, 1997) e sistemas baseados no conhecimento, onde são aprendidos determinados conjuntos de padrões de fala que determinam a classificação posterior.

As redes neurais artificiais (RNA), são sistemas computacionais poderosos usados para modelar relações complexas geralmente não lineares entre entradas e saídas, permitindo uma maior compreensão de eventos físicos complexos e variantes no tempo, tais como as ondas acústicas.

Os sistemas de reconhecimento baseados em redes neurais artificiais visam, mediante a interconexão de um conjunto de unidades de processo atuando em paralelo, obter resultados de reconhecimento similares ao desempenho obtido pelas interconexões biológicas presentes no cérebro humano. Este tipo de interconexão é especialmente necessário em aplicações que demandam grande poder computacional para calcular e avaliar simultaneamente um conjunto de hipóteses, as quais são características presentes dentro dos sistemas de reconhecimento e classificação de padrões acústicos da fala.

Neste capítulo são apresentadas as técnicas computacionais baseadas em RNAs e usadas no reconhecimento e classificação de elementos acústicos presentes no idioma Português. Na Seção 4.1 é introduzida a fundamentação teórica necessária para a compreensão

do funcionamento de uma rede neural, considerando seus elementos básicos, conhecidos como neurônios, as diferentes topologias e tipos de conexão, os modelos de aprendizado e os algoritmos utilizados durante a etapa de treinamento do sistema reconhecedor fonético. Na Seção 4.2, são descritos os problemas decorrentes dos dados pré-processados, obtidos mediante as técnicas apresentadas no Capitulo 3, e pertencentes ao conjunto de treinamento, sendo propostas várias metodologias que visam melhorar a classificação de eventos que descrevem eventos de natureza física. Na Seção 4.3, é exposta a base teórica da técnica de comitê de especialistas, usada na implementação do sistema reconhecedor e classificador de eventos acústicos. Na Seção 4.4, é apresentado o método proposto, modificando o sistema combinador de especialistas para o reconhecimento online dos fonemas. As considerações finais deste capítulo são feitas na Seção 4.5

4.1 Redes neurais artificiais

Uma rede neural é um processador maciço e paralelamente distribuído, constituído por unidades de processamento simples que têm uma capacidade de armazenamento, conhecimento experimental e disponibilizá-lo para um uso posterior (HAYKIN et al., 2009).

Geralmente os neurônios artificiais, usados nos modelos de RNAs são não lineares. Fornecem saídas tipicamente contínuas, realizando funções simples, tais como coletar os sinais existentes em suas entradas, somar-lhes de acordo a uma função operacional e produzir uma resposta. Esta resposta é dada de acordo com a função de ativação usada.

O projeto de uma rede neural é motivado pela analogia com o cérebro humano, onde o conhecimento é adquirido pela rede a partir de estímulos de ambientes externos e mediante um processo de aprendizagem. As forças de conexão entre neurônios, conhecidas como *pesos sinápticos*, são utilizadas para armazenar o conhecimento adquirido.

Esta seção resume a fundamentação teórica, abrangendo os elementos que compõem uma RNA, os diferentes tipos de interconexão entre os elementos processadores, os métodos de aprendizado e finalmente e técnicas de treinamento usadas nas redes propostas para esta dissertação.

4.1.1 Modelo biológico de um neurônio

A teoria e o modelo das redes neurais artificiais são inspirados da estrutura e funcionamento do sistema nervoso humano, onde o elemento fundamental é a célula conhecida

como neurônio. Este componente biológico conduz impulsos elétricos gerados por ações físico-químicas em determinadas condições de operação. A célula é dividida em quatro partes principais: os dendritos, o corpo celular, o axônio e as terminações sinápticas (KANDEL et al., 2000). A Figura 24 apresenta a configuração de uma célula nervosa.

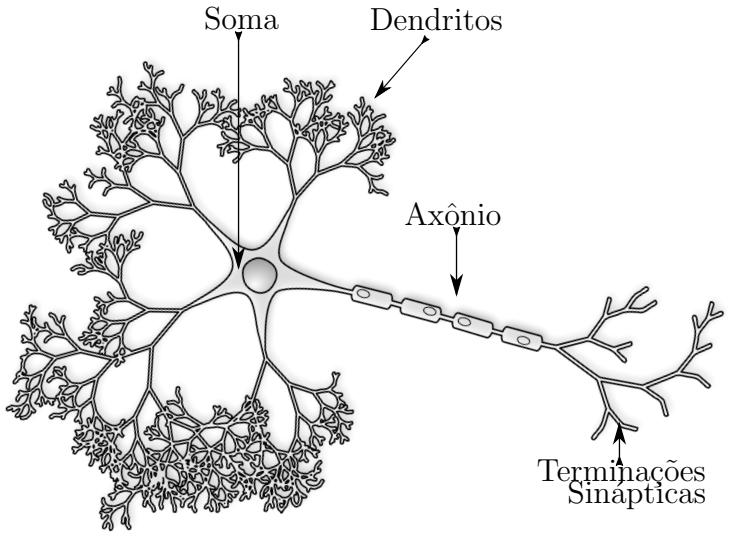


Figura 24: Neurônio biológico

Os dendritos iniciam-se no corpo celular e são constituídos por vários finos prolongamentos que formam a árvore dendrítica. Estes capturam os sinais eletroquímicos das células vizinhas usando as *conexões sinápticas*.

O corpo celular, também conhecido como *soma*, contém o núcleo. Este controla todas as atividades metabólicas do neurônio baseado na informação dos neurônios vizinhos usando as conexões sinápticas. Esta informação é combinada e integrada para emitir os sinais de saída.

O axônio é a saída do neurônio. É usado para enviar sinais aos terminais axônicos de neurônios vizinhos. Quando o axônio está perto dos neurônios destino, este divide-se em numerosas ramificações que criam as *sinapses* com o soma ou axônios de outras células. A sinapse é um processo químico que permite a transmissão de um sinal entre os neurônios. Estes sinais são de dois tipos: Elétricos e químicos. O sinal gerado pelo neurônio e transportado através do axônio é um impulso elétrico. Por outro lado, o sinal transmitido entre as terminações sinápticas e os dendritos de outro neurônio é de origem química. Estima-se que cada neurônio recebe de 10,000 a 100,000 sinapses (KANDEL et al., 2000).

Algumas estruturas neurais são determinadas no nascimento, enquanto outras são estabelecidas mediante o processo de aprendizagem, onde são geradas novas conexões neurais. Efetivamente, as estruturas neurais continuam evoluindo durante toda a vida, gerando reforços ou debilitações entre as conexões sinápticas. Este processo de aprendizagem é uma das características que pretende-se emular com as RNAs, descritas nas próximas 5 seções.

4.1.2 Funcionamento do neurônio artificial

Os elementos computacionais, denominados neurônios artificiais, são modelos simplificados dos neurônios biológicos. O protótipo de neurônio mais simples e que engloba as principais características de uma rede neural biológica, ou seja, paralelismo e alta conectividade, foi proposto por McCulloch-Pitts. É o modelo mais utilizado nas arquiteturas de RNAs (HAYKIN, 2001).

A Figura 25 ilustra a estrutura de um neurônio, constituído por m sinais de entrada e somente uma saída.

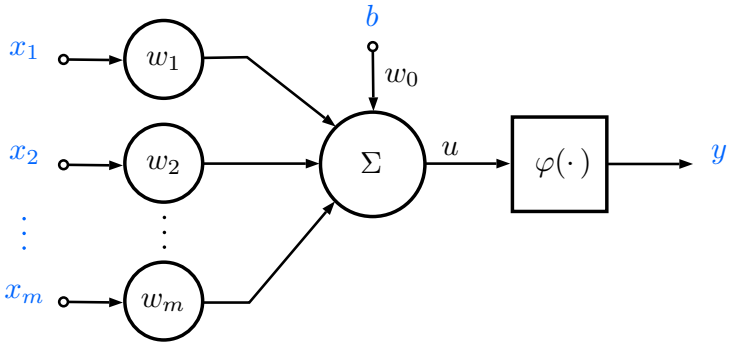


Figura 25: Modelo de um neurônio artificial

O processo interno realizado por este tipo de neurônio artificial é sintetizado matematicamente usando as seguintes expressões:

$$u = \sum_{j=1}^{m} w_j x_j + w_0 b, \tag{12}$$

$$y = \varphi(u), \tag{13}$$

onde cada elemento do neurônio artificial é descrito a seguir:

- Os sinais de entrada x_j, para $1 \leq j \leq m$, representam as componentes advindas do meio externo e são equivalentes aos sinais electro-químicos processados pelos neurônios biológicos.

- Os pesos sinápticos do neurônio, w_j, onde $1 \leq j \leq m$, ponderam cada uma das variáveis de entrada da rede, permitindo quantificar as relevâncias de cada entrada em relação à funcionalidade do respetivo neurônio.

- Combinador linear Σ, cuja função é agregar todos os sinais de entrada que foram ponderados pelos respetivos pesos sinápticos a fim de produzir um valor de potencial de ativação.

- O bias ou limiar de ativação b, indica o patamar apropriado para que o resultado produzido pelo combinador linear possa gerar um valor significativo em direção à saída do neurônio. Esta entrada também é ponderada por um peso sináptico w_0. Na literatura, visando simplificar as expressões matemáticas, o limiar de ativação b, também é representado como x_0.

- A função de ativação $\varphi(\cdot)$, induz os limiares de saída do neurônio dentro de um intervalo de valores assumidos pela função escolhida.

- O sinal de saída y, é o valor final produzido pelos processos internos do neurônio artificial para o conjunto de valores das variáveis de entrada x_j, $1 \leq j \leq m$.

Efetivamente, as comparações entre as características apresentadas na Seção 4.1.1 e o comportamento de um neurônio artificial demostram que o modelo sintetiza o funcionamento básico de um neurônio biológico. Ademais, os atributos deste tipo de neurônios artificiais permitem sua interconexão, podendo gerar estruturas com maior poder de processamento do sistema com relações entrada saída altamente complexas. Este tipo de topologias é tratado a seguir.

4.1.3 Topologia de uma RNA

A arquitetura de uma RNA toma como base a organização e disposição dos neurônios em relação aos neurônios vizinhos, formando níveis ou camadas distribuídas entre as entradas e saídas da rede. Uma camada ou nível é um conjunto de neurônios cujas entradas são originadas das variáveis externas, ou da camada anterior, e suas saídas são destinadas à próxima camada ou à saída. Os parâmetros fundamentais de uma rede neural artificial básica são: O número de camadas, a quantidade de neurônios por camada e o grau de conectividade entre os neurônios.

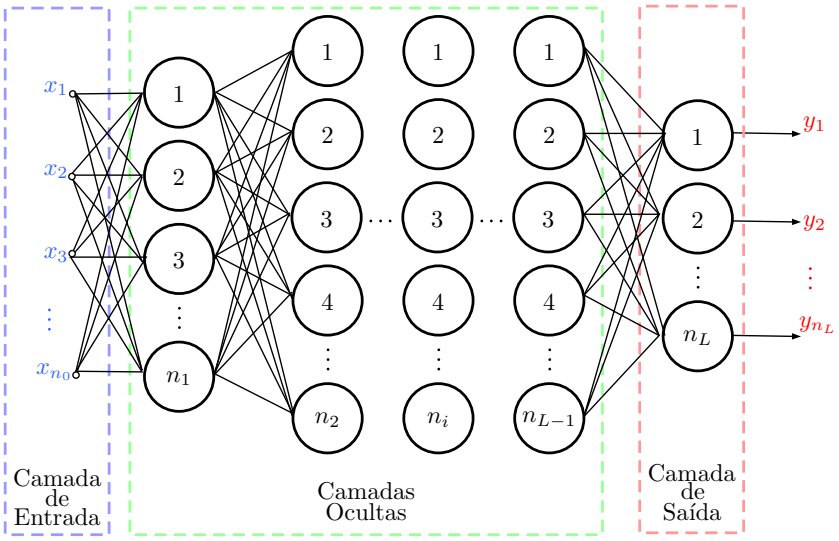


Figura 26: Partes de uma rede neural artificial

A Figura 26 ilustra uma RNA genérica de L camadas sendo $0 \leq l \leq L$. A camada $l = 0$ é conhecida como *camada de entrada*, cuja função é receber um vetor de características correspondentes às variáveis externas $\mathbf{x} = [x_1, x_2, \ldots, x_{n_0}]$. As *camadas ocultas*, distribuídas entre os índices de camada $l = 1$ e $l = L - 1$, recebem a informação de uma camada posterior, gerando uma saída destinada à camada seguinte. Finalmente, a *camada de saída*, $(l = L)$, prepara e apresenta o vetor de respostas $\mathbf{y} = [y_1, y_2, \ldots, y_{n_L}]$ calculadas pelas camadas anteriores da RNA. Note-se que cada camada da rede contém um número independente de neurônios n_l.

Considerando os parâmetros da topologia de uma RNA, podem-se distinguir três topologias básicas:

1. Redes a monocamada: Estas redes só contém uma camada de neurônios, onde o vetor de entrada $\mathbf{x}$ é projetado diretamente à camada de saída, sendo uma rede propagada em uma direção. Quando os neurônios estão completamente interligados, a topologia da RNA é conhecida como *rede Hopfield* (HAYKIN et al., 2009).

2. Redes a multicamada: Este tipo de topologia apresenta uma ou mais camadas ocultas, onde os neurônios das camadas são conhecidos como *neurônios ocultos*. A função das camadas ocultas é melhorar a representação e aprendizagem da rede. Uma rede composta por mais de uma camada oculta adquire uma perspectiva global devido a um conjunto extra de conexões sinápticas e uma dimensão extra de interações neurais (CHURCHLAND; SEJNOWSKI, 1992).

3. Redes recorrentes: Este tipo de estrutura tem pelo menos um ciclo de retroalimentação. A presença deste tipo de ciclo produz um profundo impacto nas capacidades de aprendizagem da RNA e seu desempenho. Além disso, os ciclos de retroalimentação que abrangem o uso de atrasos unitários de tempo geram um comportamento dinâmico não linear (HAYKIN et al., 2009). A Figura 27 ilustra este tipo de topologia, onde z^{-1} representa um atraso temporal sobre o sinal, x_1 e x_2 são as entradas da rede, além disso, y_1 e y_2 representam as saídas do sistema. Note a retroalimentação dada entre as saídas e os neurônios.

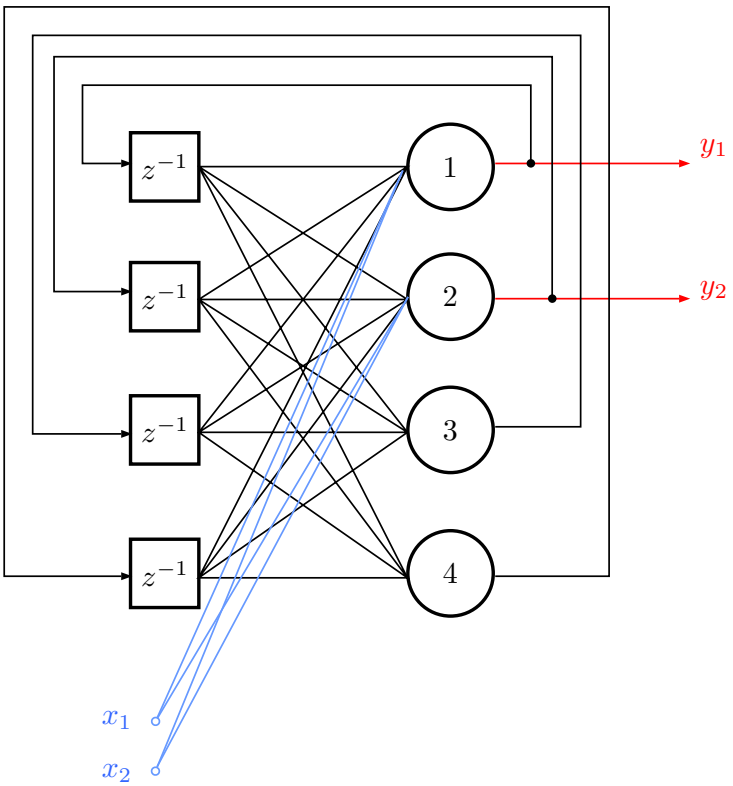


Figura 27: Diagrama de uma rede neural recorrente comum

Considerando a configuração da topologia de uma rede neural artificial, é relevante notar a eleição de uma função de ativação dos neurônios. Logo, considerando os métodos usados em de alguns algoritmos de treinamento apresentados na Seção 4.1.4, é necessário usar funções de ativação $\varphi(\cdot)$ não lineares, principalmente, nas camadas ocultas da RNA. No entanto, a escolha da função de ativação da camada de saída depende do objetivo de uso da RNA: classificação ou regressão. Basicamente, existem três tipos de funções de ativação no caso de classificação.

Na *função limiar*, a saída de um neurônio assume o valor 1 se o campo local induzido daquele neurônio é não negativo e 0 caso contrário. O comportamento de uma

rede neural com função limiar na camada de saída descreve a propriedade de *tudo ou nada* (HAYKIN et al., 2009). A Equação 14 define este comportamento para uma entrada u, conforme definido na Equação 13:

$$y = \begin{cases} 1 \text{ se } u \geq 0, \\ 0 \text{ se } u < 0. \end{cases} \tag{14}$$

Por outro lado, na *função linear por partes*, assume-se que o fator de amplificação dentro da região linear de operação é a unidade. Acrescentando suas aplicações, uma função de ativação não limitada gera uma função linear, amplamente usada na predição de comportamentos simples. A Equação 15 define o funcionamento de uma função linear por partes:

$$y = \begin{cases} 1, & u \geq 0,5 \\ u, & -0,5 < u < 0,5 \\ 0, & u \leq -0,5. \end{cases} \tag{15}$$

Finalmente, a *função sigmoide* é a mais usada. É definida como uma função estritamente crescente que exibe um balanceamento adequado entre os comportamentos linear e não linear. Este tipo de função de ativação deve ser diferenciável, uma característica importante dentro da teoria das redes neurais artificiais, já que permite o cálculo do gradiente de erro conforme apresentado na Seção 4.1.4. A Equação 16 define a função sigmoide logística:

$$y = \frac{1}{1 + e^{-au}}, \tag{16}$$

onde a é o parâmetro de inclinação da função sigmoide. Devido às facilidades analíticas (MASTERS, 1993), é preferível usar uma função sigmoide com limiares entre -1 e 1, conforme definida na Equação 17:

$$y = \tanh(u) = \frac{1 - e^{-au}}{1 + e^{-au}}, \tag{17}$$

onde a novamente representa o parâmetro de inclinação da função. Este tipo de função de ativação é conhecido como *tangente hiperbólica*, principalmente usada nos neurônios das camadas ocultas e a camada de saída para as redes propostas nesta dissertação.

4.1.4 Aprendizagem

A aprendizagem é um processo onde os pesos sinápticos de uma RNA são ajustados mediante uma ação contínua de estimulação gerada por um ambiente externo. Assim, o conhecimento adquirido é configurado e armazenado através dos pesos sinápticos. O tipo

de aprendizagem está determinado de acordo com o comportamento desejado, usando os ajustes dos parâmetros (MENDEL; FU et al., 1970).

Depois de anos de pesquisa, foram desenvolvidos diversos métodos para treinamento de RNAs, podendo ser agrupados em dois paradigmas principais: aprendizagem supervisionada e não supervisionada.

1. *Aprendizagem supervisionada:* Este tipo de aprendizagem, também conhecida como *aprendizagem com professor*, envolve a modificação dos pesos sinápticos da RNA usando um conjunto de amostras de treinamento etiquetadas. Cada conjunto está representado por um vetor de características e um outro de respostas desejadas correspondentes:

$$\mathcal{F} = [\mathbf{x}^p, \mathbf{d}^p]_{p=1}^{P}, \tag{18}$$

onde $\mathbf{x} = [x_1, x_2, \ldots, x_{n_0}]$ representa o vetor de entradas do conjunto de treinamento $\mathcal{F}$ e $\mathbf{d} = [d_1, d_2, \ldots, d_{n_L}]$ corresponde às saídas desejadas, sendo p o índice da amostra que pertence ao banco de dados de P amostras.

O treinamento inicia-se apresentado exemplos etiquetados de forma aleatória. Com cada amostra apresentada, os pesos sinápticos da RNA são modificados para minimizar a diferença entre a resposta da RNA e saída desejada. Este processo é repetido até atingir um estado estável, onde não há mais modificações significativas dos pesos sinápticos (HAYKIN et al., 2009).

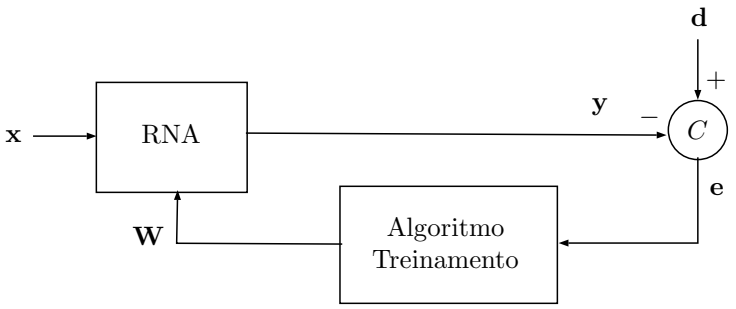


Figura 28: Esquema de um aprendizado supervisionado aplicado em uma RNA

A Figura 28 ilustra o mecanismo de aprendizado supervisionado, baseado nos vetores de características $\mathbf{x}$ e as etiquetas de saída correspondentes, $\mathbf{d}$. Adicionalmente, a rede tem sua saída corrente $\mathbf{y}$, que é comparada com a saída desejada $\mathbf{d}$. Desse modo, mediante o algoritmo de treinamento, a matriz de pesos $\mathbf{W}$ é modificada de acordo com o erro $\mathbf{e}$.

2. *Aprendizado não supervisionado:* Este tipo de aprendizado não conta com um supervisor para acompanhar o processo de aprendizagem. Para os algoritmos de aprendizado não supervisionado só estão disponíveis os padrões de entrada da RNA, ao contrário do aprendizado supervisionado, onde o conjunto de treinamento contém pares de entrada saída. No momento em que a rede estabelece uma harmonia com as regularidades estatísticas da entrada de dados, desenvolve-se nela uma habilidade de formar representações internas para codificar características de entrada e criar novas classes ou grupos automaticamente. Este tipo de aprendizado só se torna possível quando existe redundância nos dados de entrada permitindo encontrar padrões ou características semelhantes nos dados (BRAGA; CARVALHO; LUDERMIR, 2000).

De acordo com as caraterísticas dos métodos de aprendizado expostos, esta dissertação usa o aprendizado supervisionado, dado que na apresentação do problema de RAF, no Capítulo 1, já são estabelecidos os diferentes grupos classificatórios, conhecidos como fonemas. É de notar que, é difícil padronizar os componentes de entrada em grupos semelhantes usando métodos de aprendizagem não supervisionados, dado que as características frequenciais de entrada da RNA mudam de acordo com o tipo de microfone, o ambiente, o tempo e o locutor.

4.1.5 Treinamento

O algoritmo de treinamento supervisionado mais usado para topologias de múltiplas camadas é a propagação reversa de erro ou *backpropagation*. Este método usa pares de treinamento, entrada e saída desejada, para ajustar os pesos da RNA mediante um mecanismo de correção de erros. O treinamento ocorre em duas fases, conhecidas como propagação adiante ou *forward* e propagação reversa ou *backward*. A fase *forward* é usada para definir a saída da rede dado um vetor de características de entrada. A fase *backward* usa a saída desejada, correspondente ao conjunto de treinamento, e a saída atual da rede para atualizar os pesos das conexões da RNA. A Figura 29 mostra o conjunto de variáveis que são usadas na descrição do algoritmo de treinamento e a relação destas com a topologia de uma RNA multicamada.

Nos algoritmos de aprendizado supervisionado, os pesos sinápticos são continuamente ajustados mediante a aplicação de ações comparativas. Na *adaptação por correção*

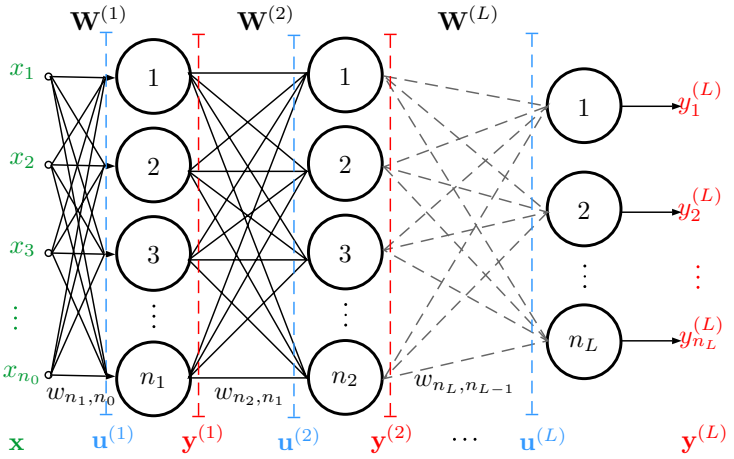


Figura 29: Componentes de uma rede multicamada

de erros, procura-se minimizar o erro global da RNA, E_T, definido como:

$$E_T = \frac{1}{P}\sum_{p=1}^{P} E(p), \tag{19}$$

onde P é o número de amostras de treinamento, e E_p é o erro de treinamento da amostra p definido como:

$$E(p) = \frac{1}{2}\sum_{i=1}^{n_L}\left(d_i(p) - y_i^{(L)}(p)\right)^2, \tag{20}$$

onde $y_i^{(L)}(p)$ é o valor produzido pelo i-ésimo neurônio da camada de saída da RNA, considerando o p-ésima amostra de treinamento, enquanto $d_i(p)$ representa a respectiva saída desejada.

A primeira parte do algoritmo de propagação reversa de erro está fundamentada no método de ajuste de pesos conhecido como *Adaline*, cujo objetivo é obter uma matriz $\mathbf{W}$ ótima tal que o erro quadrático $E(\mathbf{W})$ do conjunto de treinamento seja mínimo possível. Desse modo, é aplicado o operador de gradiente em relação à matriz $\mathbf{W}^{(l)}$, onde $1 \leq l \leq L$ é o número da camada da RNA. A Equação 21 determina a operação gradiente sobre o erro quadrático médio dada uma matriz de pesos:

$$\nabla E(\mathbf{W}) = \frac{\partial E(\mathbf{W})}{\partial \mathbf{W}}. \tag{21}$$

Em primeira instância, é calculado o erro da rede para matriz de pesos $\mathbf{W}^{(L)}$, sendo que o gradiente de erro para a camada de saída da Figura 29, definido conforme Equação 22:

$$\nabla E\left(\mathbf{W}^{(L)}\right) = \frac{\partial E}{\partial w_{i,j}^{(L)}} = -\left(d_i - y_i^{(L)}\right)\varphi'\left(u_i^{(L)}\right)y_j^{(L-1)} = \delta_i^{(L)} y_j^{(L-1)}, \tag{22}$$

onde $u_i^{(L)}$ denota a entrada ponderada em relação com o i-ésimo neurônio da camada L, definida na Equação 12. Por outro lado, $y_i^{(L)}$ e $y_j^{(L-1)}$ representam a saída da função de ativação do neurônio i e j respetivamente, e $\varphi'(\cdot)$ indica a derivada de primeira ordem da função de ativação considerada na RNA. É importante apontar que o termo $-(d_i - y_i^{(L)})\varphi'(u_i^{(L)})$ da Equação 22 mede o erro do neurônio i da última camada L, representada como $\delta_i^{(L)}$. A demonstração completa do gradiente de erro é descrita em (HAYKIN et al., 2009).

Entretanto, diferente dos neurônios da última camada, os neurônios ocultos não têm um acesso direto ao conjunto de saídas desejadas $\mathbf{d}$. Nesta situação, os ajustes de seus pesos são sujeitos às estimativas dos erros de saída produzidos pelos neurônios da camada posterior, os quais foram previamente ajustados. Assim, o gradiente de erro para um neurônio oculto é obtido usando a Equação 23:

$$\nabla E\left(\mathbf{W}^{(l)}\right) = \frac{\partial E}{\partial w_{i,j}^{(l)}} = -\left(\sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} w_{k,i}^{(l+1)}\right) \varphi'(u_i^{(l)}) y_j^{(l-1)} = \delta_i^{(l)} y_j^{(l-1)}, \tag{23}$$

onde $\delta_k^{(l+1)}$ representa a propagação reversa do erro originado nos neurônios das camadas posteriores e $w_{k,i}^{(l+1)}$ representa o peso do neurônio i na camada posterior. Novamente, a expressão que mede o erro do neurônio i da camada atual, composto por $-\left(\sum_{k=1}^{n_{l+1}} \delta_k^{(l+1)} w_{k,i}^{(l+1)}\right) \varphi'(u_i^{(l)})$ é representado por $\delta_i^{(l)}$.

O algoritmo de propagação reversa de erro além de calcular o erro mínimo de uma camada de neurônios, deve ajustar os pesos conforme o gradiente de erro calculado. Desse modo, é feito um ajuste sobre a matriz de pesos $\mathbf{W}^{(l)}$ em direção oposta ao gradiente, pois o objetivo da otimização é minimizar o erro quadrático médio. Nesta condição, a variação dos pesos $\Delta\mathbf{W}$ é dada pela Equação 24:

$$\Delta\mathbf{W} = -\eta \nabla E(\mathbf{W}), \tag{24}$$

onde η é uma constante que define a taxa de aprendizagem da rede, indicando a velocidade de processamento de treinamento da rede até chegar ao mínimo da função de erro quadrático médio.

Assim, a correção dos erros é um processo iterativo que modifica os pesos de acordo ao número de amostras de treinamento utilizadas:

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \Delta\mathbf{W}, \tag{25}$$

onde t representa o instante de tempo em que é apresentado um novo par entrada saída desejada do conjunto de treinamento definida na Equação 18.

A sequência de procedimentos computacionais usados para o treinamento de uma rede de topologia multicamada e que resume o processo matemático usado é apresentada no Algoritmo 1.

Algoritmo 1 Propagação reversa de erro para uma rede neural multicamada

Entrada $\mathcal{F} = [\mathbf{x}^p, \mathbf{d}^p]_{p=1}^P$, L: Número de camadas, n_l: Número de neurônios na camada, η: taxa de aprendizagem, ϵ: precisão requerida, P: número de amostras
Saída Matrizes de pesos: $\mathbf{W} = \mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \ldots, \mathbf{W}^{(L)}$.

1: $epoca \leftarrow 0$
2: inicializar as matrizes $\mathbf{W}$ com valores aleatórios.
3: Calcular E_T usando o vetor de matrizes $\mathbf{W}$.
4: **enquanto** $E_T \geq \epsilon$ **faça**
5: **para todo** $p = 1 \rightarrow P$ **faça**
6: **para todo** $l = 1 \rightarrow L$ **faça**
7: Calcular $\mathbf{u}^{(l)}(p)$, de acordo com Equação 12
8: Calcular $\mathbf{y}^{(l)}(p)$, de acordo com Equação 13
9: **fim para**;
10: **para todo** $l = L \rightarrow 1$ **faça**
11: **se** $l = L$ **então**
12: Calcular $\nabla E\left(\mathbf{W}^{(L)}\right)$, de acordo com Equação 22
13: **senão**
14: Calcular $\nabla E\left(\mathbf{W}^{(l)}\right)$, de acordo com Equação 23
15: **fim se**;
16: Ajustar $\mathbf{W}^{(l)}(p)$, de acordo com Equação 24
17: **fim para**;
18: **fim para**;
19: Calcular $\mathbf{y}^{(L)}(p)$ levando em conta os pesos ajustados.
20: Calcular E_T; $epoca \leftarrow epoca + 1$
21: **fim enquanto**;

A estrutura básica de propagação reversa de erro, apresentada no Algoritmo 1, é finalizada quando o erro global da rede E_T atinge a precisão ϵ, geralmente configurada como 0 ou, de acordo com as características próprias do problema, admitindo certo grau de liberdade na precisão.

Neste processo, é importante configurar cuidadosamente o valor da taxa de aprendizado η, já que uma taxa pequena afeta o tempo de localização do mínimo erro. Por outro lado, uma taxa grande pode gerar oscilações, evitando que o algoritmo de propagação reversa ache um mínimo local. Um modo básico de resolver este problema é introduzindo o *momentum*, μ, conforme a Equação 26:

$$\Delta \mathbf{W}(t) = -\eta \nabla E(\mathbf{W}) + \mu \Delta \mathbf{W}(t-1), \tag{26}$$

onde $\Delta \mathbf{W}(t-1)$ representa a variação de ajuste passada e ponderada pelo *momentum*. Usando este parâmetro na variação de ajuste dos pesos, o processo de convergência da rede se torna bem mais eficiente, pois se leva em consideração o critério de quão afastada está a solução atual da solução ótima.

Porém, a variação de ajuste definida na Equação 26 ainda depende do *momentum* μ fixo. Além disso, o algoritmo de treinamento pode apresentar problemas em regiões onde as saídas são saturadas devido aos valores elevados dos potenciais de ativação u, produzindo lentidão no processo de treinamento (SILVA; SPATTI; FLAUZINO, 2010).

Assim, diferente à variação proposta na Equação 26, o algoritmo conhecido como *propagação resiliente*, ao invés de considerar as magnitudes do gradiente da função de erro, leva em conta somente a variação do sinal (RIEDMILLER; BRAUN, 1993). Desse modo, a taxa de aprendizado do método se torna dinâmica, onde é incrementada nos casos em que os sinais do gradiente são os mesmos em duas iterações sucessivas. Caso contrário, se os sinais do gradiente são diferentes, a taxa de aprendizado é reduzida, devido ao fato que o ponto mínimo da função foi ultrapassado. Em termos matemáticos, a avaliação do sinal dos gradientes é dada pela Equação 27:

$$\Lambda_{i,j}(t) = \begin{cases} \eta^{+}\Lambda_{i,j}(t-1), & \text{se } \left(\frac{\partial E(t-1)}{\partial w_{i,j}} \frac{\partial E(t)}{\partial w_{i,j}}\right) > 0, \\ \eta^{-}\Lambda_{i,j}(t-1), & \text{se } \left(\frac{\partial E(t-1)}{\partial w_{i,j}} \frac{\partial E(t)}{\partial w_{i,j}}\right) < 0, \\ \Lambda_{i,j}(t-1) & \text{caso contrário,} \end{cases} \tag{27}$$

onde $\Lambda_{i,j}(t-1)$ representa a taxa de aprendizagem individual associada a cada peso $w_{i,j}^{(l)}$ de acordo com as taxas de aprendizagem definidas entre $0 < \eta^{-} < 1$ e $\eta^{+} > 1$.

Finalmente, as matrizes de pesos da RNA são ajustadas seguindo a Equação 28:

$$\Delta w_{i,j}(t) = \begin{cases} -\Lambda_{i,j}(t), & \text{se } \frac{\partial E(t)}{\partial w_{i,j}} > 0, \\ \Lambda_{i,j}(t-1), & \text{se } \frac{\partial E(t)}{\partial w_{i,j}} < 0, \\ 0 & \text{caso contrário.} \end{cases} \tag{28}$$

Desse modo, a alteração das matrizes de pesos da RNA depende das variações dos sinais das derivadas parciais, desconsiderando os valores de suas magnitudes (SILVA; SPATTI; FLAUZINO, 2010). O Algoritmo 2 descreve a o processo de adaptação e aprendizagem correspondente a técnica de propagação resiliente:

Algoritmo 2 Atualização dos pesos usando a técnica propagação resiliente.

Entrada Matrizes de pesos: $\mathbf{W}(t) = \mathbf{W}^{(1)}(t), \mathbf{W}^{(2)}(t), \ldots, \mathbf{W}^{(L)}(t)$.
Saída Matrizes de pesos atualizadas:
$\mathbf{W}(t+1) = \mathbf{W}^{(1)}(t+1), \mathbf{W}^{(2)}(t+1), \ldots, \mathbf{W}^{(L)}(t+1)$.

$\Lambda_0 \leftarrow 0,1$; $\Lambda_{max} \leftarrow 50$; $\Lambda_{min} \leftarrow 10^{-7}$
$\eta^- \leftarrow 0,5$; $\eta^+ \leftarrow 1,2$
para todo $w_{i,j}(t)$ **faça**
  se $\left(\frac{\partial E(t-1)}{\partial w_{i,j}} \frac{\partial E(t)}{\partial w_{i,j}}\right) > 0$ **então**
    $\Lambda_{i,j}(t) = \min(\Lambda_{i,j}(t-1)\eta^+, \Lambda_{max})$
    $\Delta w_{i,j}(t) = \text{sign}\left(\frac{\partial E(t)}{\partial w_{i,j}}\right) \Lambda_{i,j}(t)$
    $w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t)$
  senão se $\left(\frac{\partial E(t-1)}{\partial w_{i,j}} \frac{\partial E(t)}{\partial w_{i,j}}\right) < 0$ **então**
    $\Lambda_{i,j}(t) = \max(\Lambda_{i,j}(t-1)\eta^-, \Lambda_{min})$
    $w_{i,j}(t+1) = w_{i,j}(t) - \Delta w_{i,j}(t-1)$
    $\left(\frac{\partial E(t)}{\partial w_{i,j}}\right) = 0$
  senão
    $\Delta w_{i,j}(t) = -\text{sign}\left(\frac{\partial E(t)}{\partial w_{i,j}}\right) \Lambda_{i,j}(t)$
    $w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j}(t)$
  fim se;
fim para;

No Algoritmo 2, os operadores `max()` e `min()` permitem achar o valor máximo ou mínimo entre dois valores, enquanto o operador `sign()` retorna $+1$ se o argumento é positivo e -1 caso contrário. No início do processo, todos os dados de atualização $\Lambda_{i,j}$ são inicializados como $\Lambda_0 = 0,1$ (RIEDMILLER; BRAUN, 1993). Os valores das atualizações são limitados por $\Lambda_{min} = 10^{-7}$ e $\Lambda_{max} = 50$, para evitar problemas de memória com valores em ponto flutuante.

As redes neurais usadas nesta dissertação são treinadas usando a técnica de propagação resiliente, já que satisfaz os requerimentos de velocidade e memória para problemas de reconhecimento de acordo com a comparação de algoritmos de treinamento feito em (HAGAN et al., 2014).

4.2 Aperfeiçoamento dos dados de treinamento

Um conjunto de dados $\mathcal{F}$, apresentado na Equação 18, pode ser constituído por observações de natureza física, sujeitas à condições ambientais e técnicas de captura de dados. Estas podem inserir elementos não compatíveis com a maioria das observações, alterando o comportamento do sistema reconhecedor se não fossem tratadas corretamente.

Um conjunto de treinamento para um sistema classificatório de múltiplas classes, está exposto a alguns problemas que obstruem e dificultam o processo de aprendizagem, diminuindo os índices de precisão e acurácia na classificação de novas amostras de entrada. Idealmente, estes conjuntos de treinamento contêm uma distribuição de padrões, permitindo diferenciar cada conjunto de classes no espaço de representação.

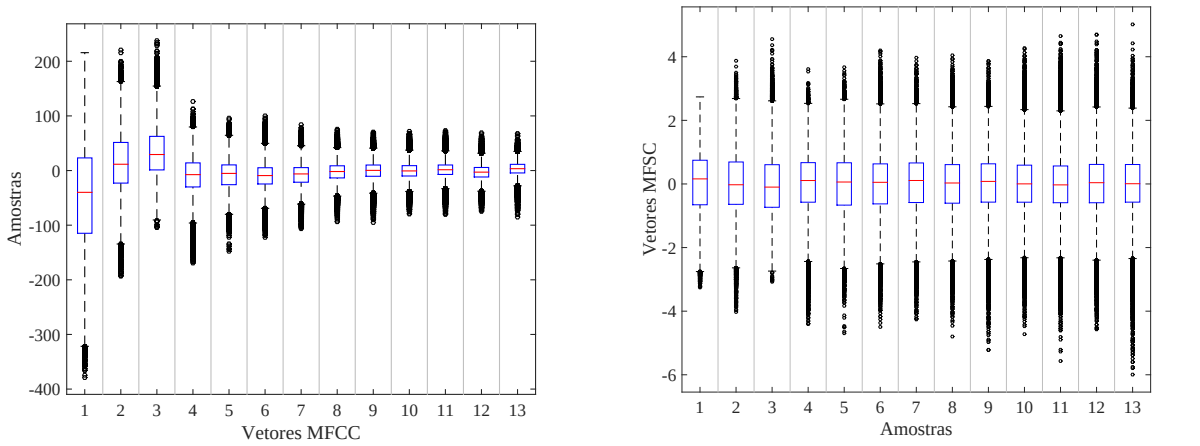
Os principais fatores que diminuem a acurácia da classificação, relacionados com o conjunto de treinamento são: vetores de representação com diferentes ordens de magnitude, presença de padrões atípicos ou ruidosos, assim como conjuntos de treinamento cujo tamanho requer grandes quantidades de memória para seu processamento, sendo que em algumas situações tais requerimentos ultrapassam a capacidade do equipamento usado. Por último, conjuntos de treinamento com classes não balanceadas. Nas próximas seções serão apresentadas as características dos problemas referentes às características de som que contém fala e suas respetivas estratégias de mitigação.

4.2.1 Normalização usando técnicas estatísticas

A normalização de dados é relacionada com o escalamento das amostras pertencentes ao conjunto de vetores de características. Alguns destes atributos podem ter faixas de representação amplas que consequentemente terão maior peso no processo de aprendizagem enquanto outros atributos de faixas de representação estreitas terão menor peso. Desse modo, um processo de normalização prévio ajudará na aceleração do estágio de aprendizagem.

A distribuição de dados do conjunto de treinamento apresentado na Figura 30(a), demonstra a diferença de escalas de representação entre os vetores de características presentes. Neste caso, uma operação de normalização do tipo *minmax* (HAYKIN, 2001) não é conveniente, dada a quantidade de observações incidentes fora da função de distribuição para cada atributo. Por outro lado, uma normalização estatística do tipo *z-score*(HAYKIN, 2001) evita que as observações atípicas alterem os parâmetros de distribuição para novas entradas. Note que a transformação de dados é feita tanto para o conjunto de dados de treinamento, como para os dados apresentados na etapa de teste e operação da RNA.

A Figura 30 compara a distribuição das amostras por atributo de entrada no conjunto de treinamento usado nesta dissertação. A transformação feita sobre o banco de dados neste exemplo mantém a correspondência das amostras dentro de suas respectivas



(a) Vetores de características MFCC (b) Vetores de características MFCC normalizados

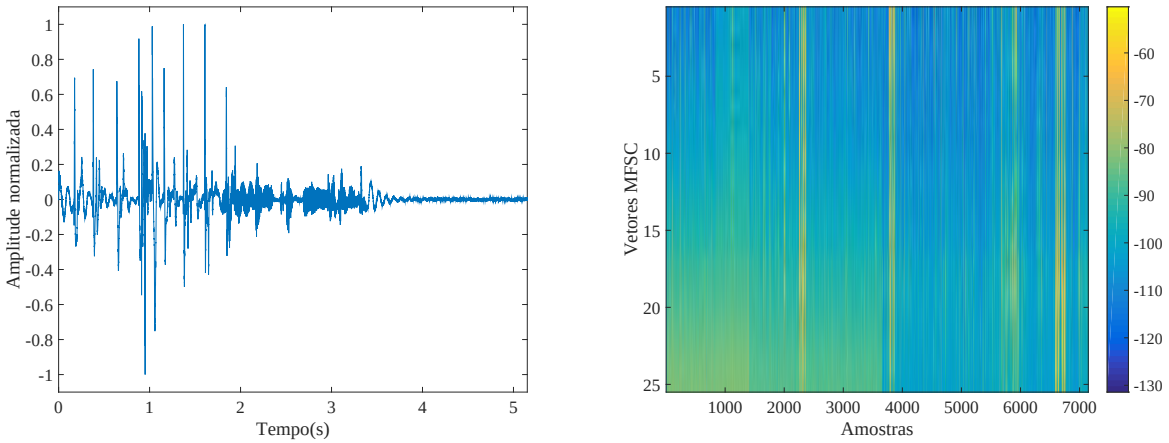
Figura 30: Comparação entre as faixas de representação para um banco de dados sem normalização e as mesmas observações normalizadas usando *z-score*.

distribuições estatísticas sem ser alterados por observações atípicas. Este comportamento é apresentado na Figura 30(b).

4.2.2 Padrões atípicos ou ruidosos

Os padrões atípicos são aqueles que, apesar de pertencer a uma classe determinada, são significativamente diferentes do resto das amostras da mesma classe. Em geral, este problema é derivado da construção do conjunto de treinamento e da natureza física dos sinais de fala. Uma observação atípica pode ser produzida por erros de captura e processamento, novos padrões que não pertencem a nenhuma classe representada do conjunto de treinamento, ou simplesmente por elementos de entrada mal etiquetados. A Figura 31 apresenta um exemplo dos casos de padrões atípicos apresentados dentro do conjunto de treinamento usado no processo de aprendizagem de um sistema RAF e usado nesta dissertação.

O sinal de fala da Figura 31(a) apresenta uma sequência de ruídos produzidos pela manipulação incorreta do microfone, ocasionando explosões de energia sonoras. Este tipo de eventos podem confundir o classificador, sendo este comportamento sonoro similar com os fonemas plosivos expostos no Capítulo 1. Por outro lado, foram encontrados parâmetros mal etiquetados dentro do conjunto de treinamento. A Figura 31(b) exemplifica este caso usando uma representação espectral para um conjunto de amostras etiquetadas como 'Pausa'. As faixas de alta energia presentes no espectro representam erros de etiqueta-



(a) Erros de captura

(b) Erros de etiquetado dentro da classe 'Pausa'

Figura 31: Exemplos de padrões atípicos do conjunto de treinamento usado na implementação

gem. Particularmente, estas observações atípicas presentes no espectro da Figura 31(b) correspondem a segmentos com conteúdo de fala. Por último, existe um grupo de fonemas que têm uma representatividade inferior a 0,01% do total de amostras usadas. Estes tipos de fonemas são excluídos do conjunto de classes, devido à pouca informação que podem oferecer durante a para uma etapa de treinamento satisfatória.

As observações detectadas como padrões atípicos podem ser agrupadas em dois conjuntos distintos, de acordo com o conhecimento obtido daquela amostra: Os elementos que conhecemos, neste caso, as amostras pertencentes as classes com pouca representatividade dentro do conjunto de treinamento. Por outro lado, os elementos desconhecidos, os quais são elementos atípicos gerados por erros de etiquetado.

A estratégia de tratamento para o primeiro tipo de observações é configurar uma nova classe que abranja todas as amostras de comportamento esporádico. Esta nova classe é identificada como 'indeterminado' (TWYCROSS, 2007). Entretanto, as observações desconhecidas devem ser detectadas e separadas do conjunto de treinamento, evitando uma possível alteração de processo de aprendizagem diminuindo a acurácia de classificação. As amostras eliminadas são incluídas no conjunto de teste para medir os níveis de generalização do sistema RAF.

Para detectar os padrões atípicos presentes em um conjunto de observações multivariadas é usada a distância *Mahalanobis*, (JIMENEZ-MARQUEZ; LACROIX; THIBAULT, 2002). Esta técnica identifica as anomalias, medindo a distância de um conjunto de amos-

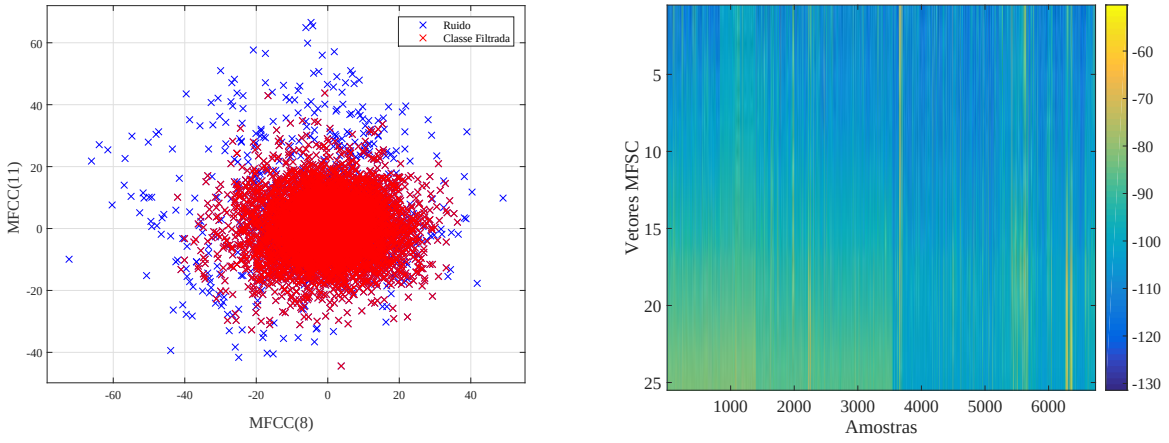
tras usando a média de cada vetor de características, conforme definido nas Equações 29 e 30:

$$\mathbf{D}_m(\mathbf{x}_p, \bar{\mathbf{x}}) = \sqrt{(\mathbf{x}_p - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_p - \bar{\mathbf{x}})}, \tag{29}$$

$$\mathbf{S} = \frac{1}{P-1} \sum_{p=1}^{P} (\mathbf{x}_p - \bar{\mathbf{x}})(\mathbf{x}_p - \bar{\mathbf{x}})^T. \tag{30}$$

onde $\mathbf{x}_p$ representa o vetor de características da observação p, $\bar{\mathbf{x}}$ a média estatística correspondente ao vetor MFCC e $\mathbf{S}$ é a matriz de covariância amostral do conjunto de dados $\mathbf{X}$ de n características e P amostras.

Logo, usando o vetor de distancias $\mathbf{d}_m$, equivalente à diagonal da matriz resultante da Equação 29, é calculado o limiar de valores críticos de acordo com a *distribuição de Fisher*, $F_{0,05}(n, P-n-1)$, onde n corresponde ao número de características das amostras, com 5% para o nível de significância (RENCHER, 2003). Todos os elementos cujas distâncias de *Mahalanobis* estejam acima do limiar calculado serão catalogados como observações atípicas e consequentemente eliminados do conjunto de treinamento e incluídas na nova classe 'indeterminado'.



(a) Erros de captura (b) Classe 'Pausa' sem as observações atípicas

Figura 32: Exemplos de padrões atípicos contidos dentro do conjunto de treinamento

A Figura 32a apresenta um exemplo de detecção dos elementos atípicos dentro da classe 'pausa' usando os vetores MFCC(8) e MFCC(11). Evidentemente, aquelas amostras de alta energia apresentadas na Figura 31(b), identificadas como elementos erroneamente etiquetados, foram removidos usando a técnica de distância de Mahalanobis. Os resultados desta filtragem são mostrados na Figura 32(b).

4.2.3 Desequilíbrio de classes

A maioria dos sistemas que usam a aprendizagem supervisionada foram elaborados assumindo um conjunto de treinamento balanceado, ou seja, um banco de dados onde a quantidade de observações por classe é aproximadamente a mesma. Contudo, uma grande parte dos problemas reais não se encaixam dentro desta suposição, existindo casos nos quais uma ou várias classes contém menos amostras que a média do conjunto de classificação. Estas classes são conhecidas como *classes minoritárias*. Em contra partida, existe outro conjunto de classes com maior representatividade, conhecidas como *classes majoritárias*. A Figura 33 apresenta a distribuição das observações por fonema, sendo que estes constituem o conjunto de classes usado no treinamento do sistema de RAF.

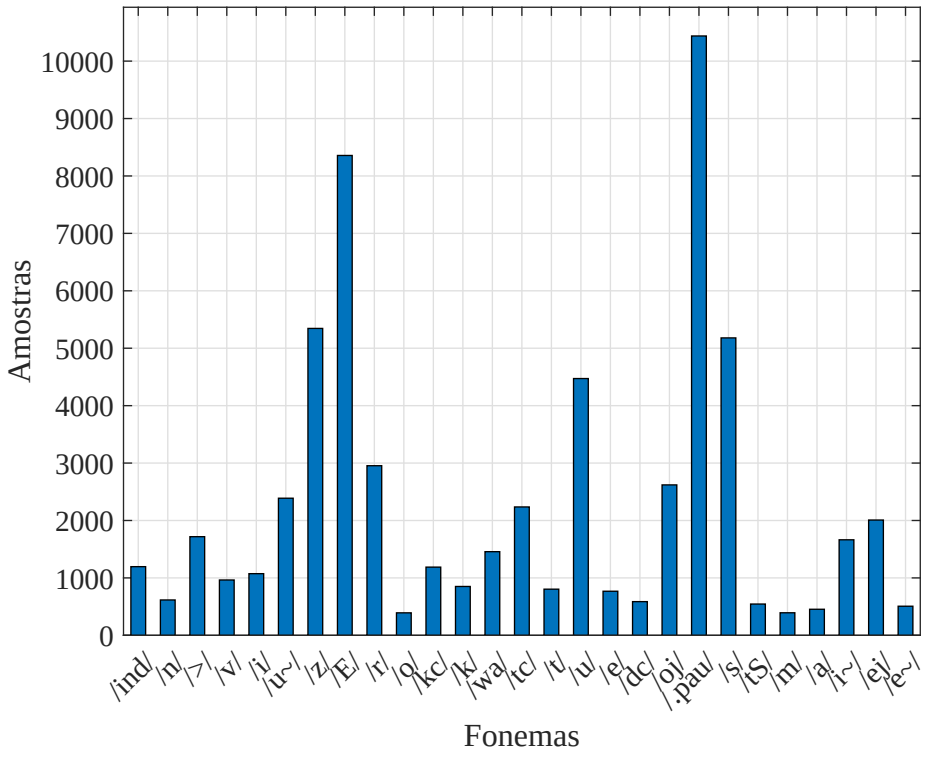


Figura 33: Quantidade de amostras por fonema para um grupo de treinamento do banco de dados de fala

A presença de desequilíbrio entre as classes de um conjunto de dados, geralmente ocasiona um aprendizado tendencioso e favorável para as classes majoritárias, de modo que o desempenho do processo de classificação para as classes minoritárias é prejudicado. Este comportamento é uma consequência do uso de algoritmos de aprendizado que assumem uma distribuição balanceada entre as classes (JAPKOWICZ; STEPHEN, 2002). Do mesmo

modo, a pouca presença de dados minoritários dentro do espaço de representação dificulta o processo de generalização do classificador ao se tratar das classes minoritárias (WEISS, 2004).

Existe um conjunto de técnicas que abordam o problema de desequilíbrio de classes. São as soluções a nível de dados, conhecidas como técnicas de re-amostragem, onde são agregadas e/ou eliminadas observações dentro do conjunto de dados para reduzir o desequilíbrio entre classes. As técnicas são simples e melhoram o desempenho global do classificador (BATISTA; PRATI; MONARD, 2004). Este tipo de método abrange duas abordagens: a sub-amostragem ou *undersampling*, que elimina as observações da classe majoritária e a sobre-amostragem ou *oversampling*, que incrementa o número de instâncias dentro das classes minoritárias.

Considerando o primeiro método, é possível reduzir a classe majoritária usando duas estratégias. A primeira estratégia é baseada no princípio de aleatoriedade, onde a sub-amostragem elimina um sub-conjunto E de amostras aleatórias da classe majoritária, tal que $\mathcal{S}_{maj} = \mathcal{F}_{maj} - E$, onde $\mathcal{S}_{maj}$ representa a novo conjunto de amostras majoritárias e $\mathcal{F}_{maj}$ representa o conjunto de observações presentes na classe majoritária. A segunda estratégia elimina as amostras redundantes e aquelas que são consideradas como ruído usando os métodos de distância (CHYI, 2003), semelhantes ao processo descrito na Seção 4.2.2. Comparando o desempenho dos dois métodos mencionados, é importante destacar que uma eliminação aleatória de amostras do conjunto de dados das classes majoritárias pode resultar em um descarte de informações importantes para a definição das fronteiras de decisão (HE; GARCIA, 2009).

Por outro lado, a sobre-amostragem é estruturada de forma semelhante, baseando novamente no princípio de aleatoriedade, onde o subconjunto E de amostras selecionadas aleatoriamente dentro da classe minoritária é replicado, de modo que $\mathcal{S}_{min} = \mathcal{F}_{min} + E$, onde $\mathcal{S}_{min}$ representa o novo conjunto de amostras e $\mathcal{F}_{min}$ é o conjunto de amostras pertencentes à classe minoritária. Contudo, o uso de réplicas idênticas da classe minoritária pode gerar problemas de sobre-ajuste do classificador (BATISTA; PRATI; MONARD, 2004).

Embora estas técnicas consigam nivelar os índices de representatividade de cada classe para conjuntos de treinamento de pequenas dimensões (HULSE; KHOSHGOFTAAR, 2009), em um problema do tipo de RAF, caracterizado por um extenso conjunto de classes e índices de desequilíbrio elevados, estas medidas são insuficientes. Portanto,

recorre-se a uma terceira alternativa que consiste em dividir o classificador em vários classificadores treinados individualmente e cujas decisões são combinadas posteriormente. As características, comportamento e configuração deste tipo de sistemas são detalhados na Seção 4.3.

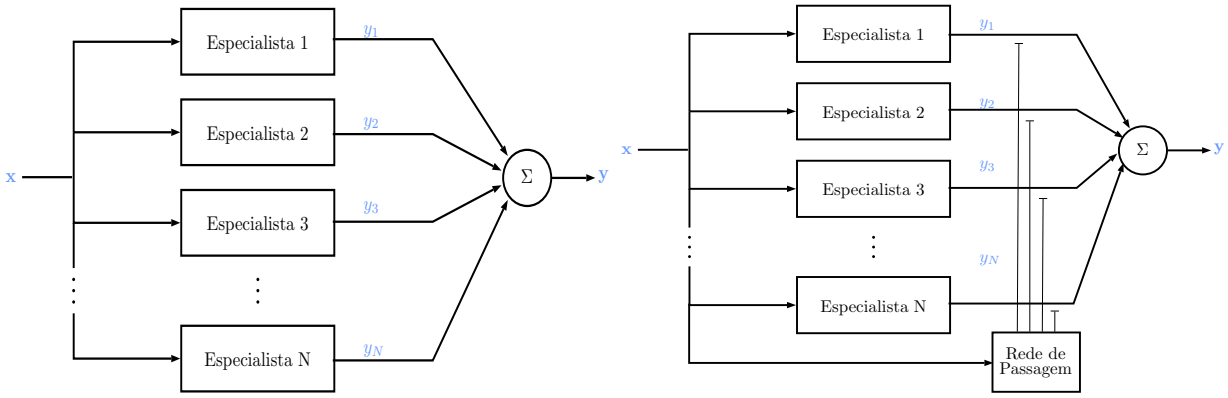
4.3 Comitê de máquinas

Os comitês de máquinas foram uma metodologia de classificação diferentes dos esquemas tradicionais para classificar padrões usando só uma estrutura de RNA. Este tipo de procedimento é baseado na premissa de que uma tarefa computacionalmente custosa pode ser simplificada, mediante o uso de diversas sub-tarefas computacionais simples e combinando suas sub-soluções para determinar uma solução da tarefa original (KITTLER et al., 1998). Um outro aspecto complicado no treinamento de um sistema de RAF, é que quantidade de dados usados é extenso, de modo que, um único componente pode não modelar o problema de forma eficiente (HAYKIN, 2001). Isso justifica o uso de um sistema composto de vários sub-sistemas onde cada um considera subconjuntos dos dados. Desse modo, além de mitigar o problema de memoria do dispositivo durante o processo de treinamento, é também solucionado até uma certa medida, o problema de desequilíbrio de classes apresentado na Seção 4.2.3.

Existem dois tipos de comitês de máquinas. Esta classificação depende de como são combinadas as saídas de cada rede especialista. As estruturas estáticas, onde as saídas de vários especialistas são combinadas por um mecanismo que não envolve a entrada, e as estruturas dinâmicas, onde a entrada está envolvida na atuação do mecanismo que integra as saídas dos especialistas individuais para gerar a saída global. A Figura 34 apresenta a arquitetura para os dois tipos de comitê de máquinas .

A solução proposta nesta dissertação está baseada na estrutura apresentada na Figura 34b, conhecida como *Composição de Especialistas*. Este tipo de arquitetura permite criar diferentes grupos de RNAs especializados no reconhecimento de uma área do mapa de decisão e ignorando as amostras externas, sendo diferente das estruturas estáticas, onde cada grupo especialista é focado em diminuir o erro global (MASOUDNIA; EBRAHIMPOUR, 2014).

O modelo de composição de especialistas é composto por N módulos de aprendizado supervisionado, conhecidos como redes especialistas, e uma unidade integradora



(a) Comitê de máquinas estáticas (b) Comitê de máquinas dinâmicas

Figura 34: Tipos de comitê de máquinas

denominada como rede de passagem. A divisão do espaço entre as especialistas pode ser linear ou não linear, podendo contemplar em alguns casos sobreposições entre regiões de decisão. A diferença dos sistemas estáticos, a rede de passagem é responsável pela ponderação das saídas dos especialistas, descompondo o espaço de entrada em diferentes sub-regiões. Com isto, a rede de passagem conhece e pode indicar automaticamente a região de cada especialista. A Figura 35 exemplifica a interação entre sistema de composição de especialistas e um espaço de decisão de entrada representado em dois vetores caraterísticos.

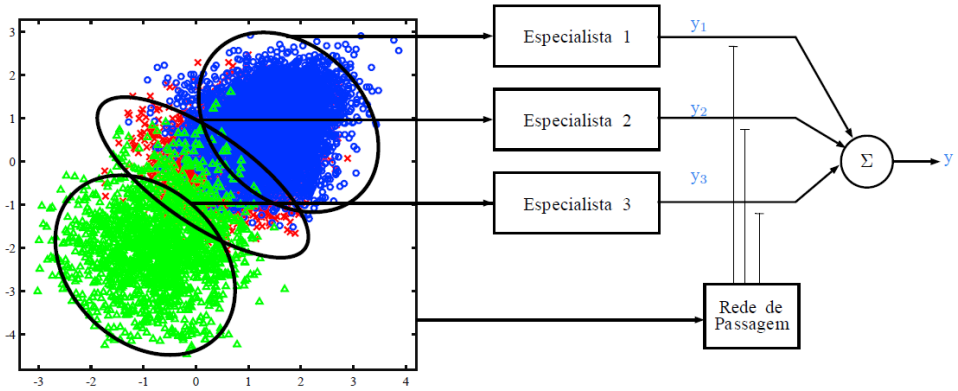


Figura 35: Decomposição do espaço de entrada dado um sistema de composição de especialistas

Considerando o funcionamento de cada rede especialista no respectivo campo de decisão, é importante notar a necessidade de um mecanismo que seja capaz de identificar para cada entrada de **x**, o especialista melhor condicionado para produzir a saída correta em termos probabilísticos. Esta identificação é realizada pela rede de passagem. A interpretação probabilística deste tipo de rede é de um sistema que calcula para cada espe-

cialista a probabilidade dele gerar a saída desejada baseando-se apenas no conhecimento do vetor de entrada. Estas probabilidades são expressas pelos coeficientes g_n, correspondentes às saídas da rede de passagem. Deste modo, este tipo de saídas são valores não negativos cuja soma é 1. A solução apresentada em (JACOBS; JORDAN; BARTO, 1991) contempla o uso da função *softmax*, definida pela Equação 31:

$$g_n = \frac{\exp(\sum_j w_{n,j} x_j)}{\sum_{n=1}^{N} \exp(w_{n,j} x_j)}, \tag{31}$$

onde x_j representa a componente j do vetor de entrada $\mathbf{x}$ e $w_{n,j}$ corresponde ao peso sináptico do neurônio n que pondera a entrada x_j.

A técnica de treinamento para um sistema de composição de especialistas é conhecida como *método desacoplado* (JORDAN; JACOBS, 1994), onde o treinamento para cada rede especialista e a de passagem é realizado independentemente. Assim, para cada rede especialista é usado o modelo de treinamento de propagação resiliente apresentado na Seção 4.1.5. Entretanto, para a rede de passagem é usado o mesmo treinamento, mas maximizando a função de custo descrita a seguir:

$$E = -\sum_{i=1}^{N} d_i \log y_i, \tag{32}$$

onde d_i é a saída desejada, correspondente a uma re-amostragem feita para o banco de dados de treinamento inicial e y_i representa a saída atual da rede de passagem obtida, compondo as N saídas correspondentes às redes de especialistas do modelo.

Por último, a função de combinação mede a saída ponderada para cada especialista, dando como ganhador aquela rede especialista que apresentar o índice de saída ponderado. Esta decisão finalmente é modelada por uma rede recorrente básica que tem como função acumular as saídas passadas para obter um novo "especialista"que condicione a decisão de classificação de acordo com o comportamento das saídas passadas. Este procedimento é detalhado na próxima seção.

4.4 Função dinâmica de pós-processamento

Um sinal de natureza dinâmica, como é um sinal de fala, deve ser tratado por sistemas de reconhecimento que considerem as mudanças de informação no tempo e que verifique a dependência da amostra atual com as amostras passadas já classificadas. A relação fonética entre as amostras de som e o comportamento dinâmico do sinal de fala correspondente é amplamente estudada no Capitulo 1.

A RNA que se adapta às características dinâmicas de um conjunto de dados dependente do tempo é conhecida como uma *rede neural recorrente*, introduzida na Seção 4.1.3. Este tipo de redes é caracterizado por ter pelo menos um ciclo de retroalimentação. Assim, os dados de saída dependem das entradas da rede, das saídas anteriores e dos estados da rede atual ou anterior. Estes ciclos envolvem o uso de tempo discreto e de bifurcações, compostas por elementos de uma unidade de atraso.

Devido à limitação temporal supracitada das redes de múltipla camada, surge a necessidade de implementar *redes neurais artificiais dinâmicas* que tenham uma recorrência das saídas adaptativas não lineares, permitindo a caracterização de um sinal que varia no tempo. A dinâmica que envolve uma rede recorrente é descrita por:

$$\mathbf{y}_{t+1} = \sigma(\mathbf{y}_t, \mathbf{y}_{t-1}, \cdots, \mathbf{y}_{t-q+1}; \mathbf{x}_t, \mathbf{x}_{t-1}, \cdots, \mathbf{x}_{t-q+1}), \tag{33}$$

onde σ é uma função não linear, $\mathbf{x}$ é o vetor de entradas e $\mathbf{y}$ é o vetor de saídas que realimenta a entrada usando uma memória de linha de retardo de q unidades.

4.4.1 Algoritmo de aprendizado

Um dos métodos usados para o treinamento das redes neurais recorrentes é o algoritmo de propagação reversa de erro no tempo. Este algoritmo é uma extensão do método de propagação reversa de erro clássico, descrito na Seção 4.4.1 onde as atualizações dos pesos ocorrem em certos momentos específicos do tempo (HAYKIN et al., 2009).

A primeira etapa envolve um processo de desdobramento da rede. Este consiste em replicar várias vezes a rede recorrente para obter uma rede equivalente apresentada na Figura 36. Cada réplica representa um instante de tempo, sendo $t = 0$, o instante de tempo inicial e $t = k$ o tempo final.

Para cada instante de tempo é calculada a função de erro E^t_{ind} usando o algoritmo de propagação reversa de erro para as redes estáticas. Finalmente, a função de custo total da rede dinâmica é aplicada, somando cada uma dos erros E^t_{ind} obtidas para cada instante de tempo $0 < t \leq k$:

$$E_{\text{total}} = \sum_{t=0}^{k} E^t_{ind}, \tag{34}$$

onde E_{total} é o erro da rede dinâmica.

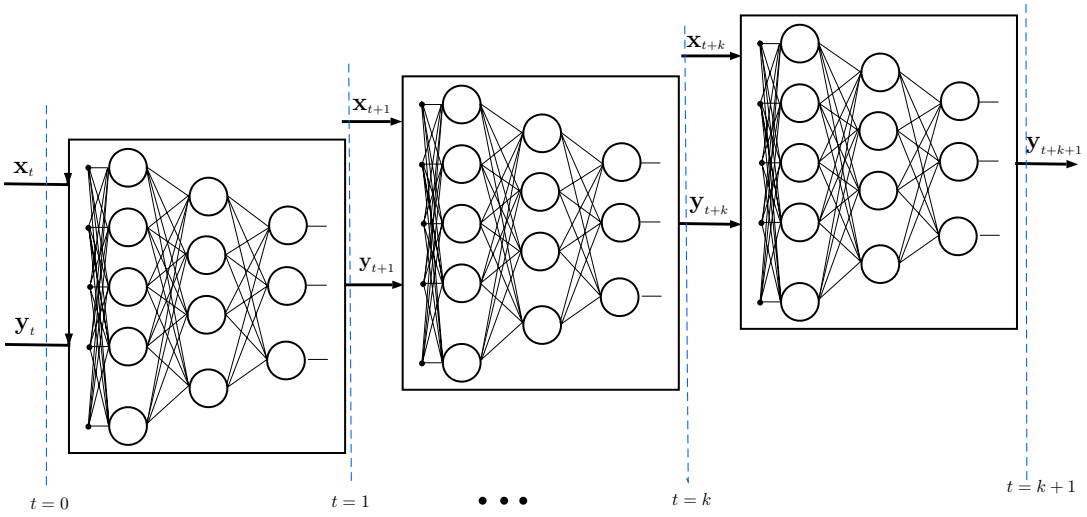


Figura 36: Representação do desdobramento da rede recorrente em uma rede com propagação reversa de erro

Em seguida, é aplicado o algoritmo de gradiente descendente para minimizar a função de custo da Equação 34. Com esta informação é possível obter a atualização dos pesos da rede dinâmica descrita na Equação 35:

$$\Delta w_{i,j} = -\eta \frac{\partial E_{\text{total}}}{\partial w_{i,j}} = -\eta \sum_{t=0}^{k} \frac{\partial E_{ind}^{t}}{\partial w_{i,j}}, \tag{35}$$

onde $w_{i,j}$ é o peso correspondente à conexão entre o neurônio i e o neurônio j.

Assim, as atualizações dos pesos dependem das entradas da rede e dos estados ocultos representados com atrasos de tempo. Com isto, o erro é propagado, tanto na arquitetura da rede como no tempo.

4.5 Considerações finais do capítulo

Neste capítulo foram apresentados os conceitos básicos e técnicos relacionados à implementação de uma arquitetura que permite o reconhecimento de fala baseada em redes neurais artificiais. Foram descritos os conceitos fundamentais para a compreensão de um sistema de aprendizagem RNA, como a topologia clássica e os algoritmos de aprendizagem usados durante o treinamento das RNAs propostas. Também foram descritas as estratégias usadas para a adequação do conjunto de dados de treinamento, de modo que sejam mitigados os efeitos nocivos gerados por problemas inerentes aos dados. Por último, foram apresentados dois métodos para o tratamento do problema de classificação, baseados no mapeamento e separação do espaço de representação, simplificando o processo de

treinamento de um sistema reconhecedor de fonemas pertencentes a fala Portuguesa e, por outro lado, a interpretação dinâmica do problema usando um sistema combinador de especialistas baseado em redes recorrentes artificiais.

O próximo capítulo apresenta os aspectos de implementação do sistema de pré-processado do sinal de fala e o reconhecedor fonético.

Capítulo 5

ASPECTOS DE IMPLEMENTAÇÃO

NESTE capítulo são apresentadas as características e considerações técnicas adotadas na implementação de um sistema de RAF. Desse modo, são descritas as características do banco de dados usado na etapa de treinamento. As considerações de implementação do estágio de pré-processamento também são descritas, de acordo com as limitações físicas da ferramenta de implementação e as observações técnicas descritas no Capítulo 3, bem como expostas as configurações dos modelos de composição usados nas experiências. Em seguida, é descrito o método combinador de especialistas usado e uma etapa de pós-processamento que considera o comportamento dinâmico de uma sequência de sons contidos na fala. Finalmente, são descritas as regras de medição de desempenho usadas nos resultados do sistema RAF.

5.1 Características do conjunto de treinamento

Um Corpora linguístico é um conjunto extenso e estruturado de exemplos reais de uso de uma linguagem determinada. As amostras que compõem os Corporas incluem textos e/ou gravações de fala transcritas que representam a dinâmica de uma língua específica. Um Corpora com conteúdo de fala engloba recopilações de entrevistas, conferências, diálogos, entre outros, vinculados a suas respetivas transcrições ortográficas e fonéticas (MAIR; HUNDT, 2000).

O Corpora com conteúdo de dados de fala usado na etapa de treinamento e teste do sistema de RAF, proposto nesta dissertação, é o CSLU: Spoltech Brazilian Portuguese Version 1.0 (SCHRAMM et al., 2006). Este banco de dados foi criado pela Universidade Federal do Rio Grande do Sul, em associação com a *Oregon Health & Science University*,

OHSU, e distribuída pelo *Linguistic Data Consortium*, LDC. O Corpora contém arquivos de som representados em forma de onda (.wav) com conteúdo de fala, arquivos em texto plano (.txt) que contêm a representação ortográfica do arquivo de som e arquivos de texto plano do tipo (.phn), onde estão contidas as representações fonéticas correspondentes.

As gravações de som contidas dentro do Corpora foram feitas em ambientes acústicos não controlados, a fim de incluir diferentes condições de ruido de fundo que permitam desenvolver aplicações generalizáveis para situações cotidianas. Com o mesmo objetivo, foram usados locutores com diferentes sotaques, correspondentes ao Português falado no território Brasileiro. Por outro lado, a fala gravada tem uma velocidade natural, cujas transcrições estão compostas por segmentos de texto extraídos de artigos publicados no jornal Folha de São Paulo e por respostas a perguntas concretas como por exemplo, o ano de nascimento, o número CEP da residência, a idade, entre outros.

Contudo, o banco de dados de fala *Spoltech* tem diversas dificuldades, sendo que alguns arquivos de fala não têm os registros correspondentes a sua descrição ortográfica e fonética, por outro lado, algumas etiquetas não correspondem a nenhum elemento conhecido dentro do alfabeto fonético padrão usado na LDC, *worldbet* (SCHRAMM et al., 2000), este é uma representação ASCII do sistema fonético IPA, apresentado no Capítulo 1, e descrito nas Tabelas 4 e 5 (HIERONYMUS, 1993). Ademais, durante a análise dos dados de entrada, foram achados diversos erros de etiquetado fonético.

Tabela 4: Fonemas presentes no Português Brasileiro usando a notação Worldbet - Vogais

	Anterior	Quase-Anterior	Central	Quase-Posterior	Posterior
Fechada	i				u
Quase-Fechada		I		U	
Meia Fechada	e				o
Meia			&		
Meia Aberta	E				>
Aberta	a				

Nasal	a~, e~, i~, o~, u~
Ditongos	aI, eI, EI, oI, >I, uI, aU, eU, EU, oU, iU, Ia, Ie, IU, Io, Ua, UI, Ue, Uo

Devido aos inconvenientes apresentados, foram estabelecidas um conjunto de ações sobre o Corpora para melhorar as condições do banco de dados. Desse modo, foi feita uma pré-seleção das gravações, onde foram rejeitados os arquivos de som que não apresentaram a descrição fonética correspondente. Para o conjunto de treinamento foram escolhidas as

Tabela 5: Consoantes presentes no Português Brasileiro e outros tipos de sons usando a notação Worldbet

	Bilabial	Labio-dental	Dental	Alveolar	Pós-alveolar	Retro-flexa	Paletal	Velar	Uvular	Faringal	Epiglotal	Glotal
Oclusiva	p,b		t,d					k,g				
Nasal	m			n			n~					
Fricativa		f,v		s, z	S, Z			x,G				h,hv
Africada					tS,tZ							
Vibrante				r								
Tepe				r(								
Aproximante				9								
Aprox. lateral				l			L					
Sons não falados	.pau(pausa), .laugh(riso), .bg(som de fundo), .br(respiração)											

gravações etiquetadas como "zipcode", as quais, contêm fala de pessoas enumerando o CEP de sua residência, sendo um conjunto de gravações com vocabulário restrito aos números em português e com uma duração aproximada de 15 minutos de fala contínua. Para o conjunto de teste foram escolhidas as gravações com as etiquetas *age* e *birthdate*, com um vocabulário semelhante ao conjunto de treinamento e com uma duração total de 10 minutos.

Adicionalmente, foram usadas as técnicas de aperfeiçoamento de dados do conjunto de treinamento apresentadas na Seção 4.2. Desse modo foram removidas as amostras com etiquetas desconhecidas e não correspondentes ao alfabeto *worldbet*, as amostras com baixa taxa de representação e por último, remover na medida do possível aquelas amostras mal etiquetadas.

5.2 Implementação do pré-processamento

Para o pré-processamento dos sinais escolhidos dentro do Corpora *Spoltech* foi desenvolvida e implementada uma solução *online*, visando generalizar o esquema de pré-processamento para sinais de entrada em tempo real, e, desse modo, gerar uma aplicação prática do sistema proposto. A implementação foi realizada em SIMULINK®, uma ferramenta de simulação e teste de sistemas embebidos, com capacidade para executar modelos em tempo real usando o *kernel Simulink Desktop Real-Time*™ (MATHWORKS, 2016)

Os primeiros parâmetros de configuração considerados são os pertinentes à aquisição de dados. Estes foram escolhidos conforme as características padrão oferecidas pela

maioria das placas de som contidas nos dispositivos inteligentes mais populares do mercado (KEHTARNAVAZ; PARRIS; SEHGAL, 2015). O esquema topológico apresentado na Figura 37 é uma representação típica de um digitalizador de sinal. Os detalhes relacionados a seu funcionamento foram apresentados na Seção 3.1.1.

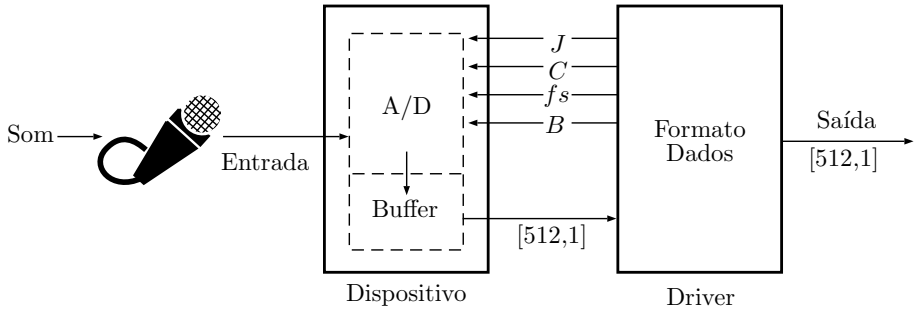


Figura 37: Diagrama de operação do sistema de captura de som de SIMULINK®

Uma das funções de SIMULINK® é disponibilizar um conjunto de blocos e bibliotecas personalizáveis. No caso de aquisição de dados de som, a ferramenta conta com um bloco cuja função é configurar o funcionamento do *driver* que controla as condições em que o sinal é digitalizado. Os parâmetros usados para este processo são:

- Tamanho do bloco de aquisição J: Este parâmetro indica o número de amostras apresentadas no bloco de saída. O tamanho deste bloco é de 512 amostras, as quais representam 32 ms do sinal de entrada, o tempo necessário para a análise e processamento de um fonema, de acordo com a Seção 1.1.2.

- Número de canais de entrada C: É configurado como sinal *mono*, dado que a fonte de entrada é originada por um microfone.

- Frequência de amostra fs: O dispositivo é configurado para amostrar o sinal usando uma taxa de 16 Khz, de acordo com as características espectrais do sinal de fala apresentadas na Seção 3.1.1 e as configurações padrão usadas comumente pelos dispositivos inteligentes, presentes na placa de som usada nos testes.

- Profundidade de bits B: Especifica a quantidade de bits usados na representação de uma amostra. A captura do sinal de entrada é feita usando uma resolução de 16 bits por amostra, correspondente aos padrões de captura da mídia digital (FARUQUE, 2015), evitando o uso excessivo de memória.

Assim, o sinal de entrada é digitalizado, de acordo as configurações apresentadas, e armazenado dentro de um *buffer* configurado automaticamente pelo controlador com capacidade de armazenamento de 16000 amostras, ou seja, 32 KB. A Figura 38 apresenta os métodos usados para o pré-processamento depois da digitalização do sinal de fala. Cada um dos passos contidos dentro deste processo está exposto no Capítulo 3.

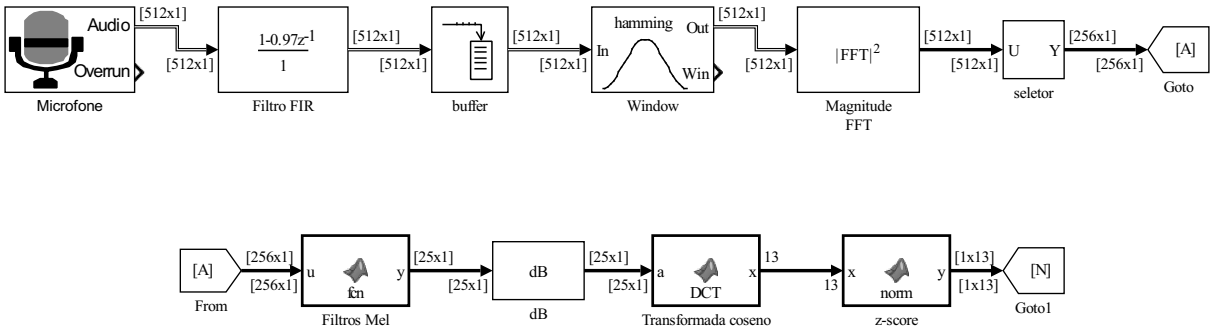


Figura 38: Estágios do pré-processamento implementados em SIMULINK.

A variável a apresentada na Seção 3.1.2 é configurada como $a = 0.97$ de acordo com as especificações dadas em (RABINER; JUANG, 1999). Seguidamente, o sinal filtrado é janelado de modo que possa ser analisado num tempo curto e sobreposto com os blocos vizinhos para manter uma correlação com amostras passadas. Assim, a sobreposição de amostras deve respeitar as condições teóricas, apresentadas na Seção 3.1.3, e deve atender as limitações da implementação. Neste caso, o sistema discreto multifrequência deve operar a passos de tamanho de múltiplos inteiros do tempo de amostragem do sistema, portanto $t_s = 16$ ms (MATLAB, 2016). Consequentemente, a subamostragem gerada pelo *buffer* deve corresponder como mínimo a um tempo de amostra de $t_{s1} = 32$ ms, ou seja, uma janela suavizada por um filtro Hamming, composta por 512 amostras e com uma sobreposição correspondente a 256 amostras.

Logo após, são calculados os vetores que caracterizarão o segmento atual do sinal de entrada. Inicialmente, é calculada a distribuição de energias dentro do espectro de frequências usando a densidade espectral de potência sobre o vetor janelado. Por praticidade, são consideradas só as primeiras 256 amostras, dado que a outra parte do vetor contém a mesma informação. Depois, são implementados 25 filtros de Mel, projetados e calculados *offline* de acordo com as equações e o procedimento apresentados na Seção 3.2. Cada filtro age sobre o vetor de entrada, gerando uma saída de 25 elementos representativos do sinal de entrada original. Ao final, com o intuito de reduzir a autocorrelação entre

os componentes do vetor de características para minimizar a complexidade do sistema de RAF proposto, é aplicada uma transformada discreta de cosseno de 13 características representativas, de acordo com a Equação 11 apresentada na Seção 3.

No último estágio, o vetor de características é normalizado usando a técnica *z-score*, apresentada na Seção 4.2.1, mediante as características estatísticas da média e o desvio padrão extraídos do conjunto de treinamento pré-processado. Cabe notar que este pré-processamento é aplicado tanto para sinais de entrada em tempo real, capturados pelo microfone, como para os sinais pertencentes ao conjunto de treinamento e teste. Nesse caso, cada bloco de comprimento igual a 512 amostras, pertencentes ao banco de dados, é apresentado usando a mesma frequência de amostragem configurada para o dispositivo de entrada, $f_s = 16$ Khz.

5.3 Configuração e treinamento da composição de especialistas

Nesta seção são estabelecidas as estruturas e configurações necessárias para a implementação de um sistema reconhecedor de fala, baseado no método de composição de especialistas. Esta implementação é dividida em cinco etapas: Inicialmente, é determinada a estrutura global do sistema de composição de especialistas, onde são estabelecidas as estratégias usadas na configuração dos sistemas de composição e apresentadas duas estruturas de distribuição de especialistas, com diferente nível de complexidade. Depois, são descritas as técnicas usadas para a definição topológica de cada especialista, isto é, o número de camadas, o número de neurônios em cada camada e o número de atrasos. Na etapa seguinte, são estipuladas as condições topológicas e funcionais das redes de passagem que ponderarão os resultados dos especialistas. Estes resultados são combinados de acordo com o peso de saída correspondente a cada especialista. Por último, é estabelecido um pós-processamento da saída, mediante o uso de uma rede dinâmica com retroalimentação.

5.3.1 Número de especialistas e profundidade hierárquica

A seleção da distribuição topológica de um modelo de composição de especialistas é uma das principais atividades dentro da configuração de um sistema de RAF baseado neste tipo de estruturas. O principal objetivo da construção do modelo de composição é determinar a quantidade de especialistas a serem usados e o grau de profundidade ou nível hierárquico.

Uma primeira solução foi proposta em (RASMUSSEN, 1996). Este método, baseado na especificação do número de neurônios dentro de uma camada oculta de uma MLP, seleciona o número de especialistas de acordo com o tamanho do vetor de características $\mathbf{x}$ e o comprimento do conjunto de treinamento. A formulação do número de especialistas I é descrito em (WATERHOUSE, 1998) como:

$$I = \frac{P}{(n+1)(k+1)}, \tag{36}$$

onde P é o comprimento do conjunto de treinamento, n é a dimensão do vetor $\mathbf{x}$ e k é a dimensão de y.

Por outro lado, para determinar o nível de hierarquia do modelo é usada uma estratégia baseada em uma arquitetura de árvore binária (WATERHOUSE, 1998). Para este caso, a profundidade da arvore b é determinada por:

$$b = \frac{\log(P + 2(n+1)) - \log(n-1) - \log(k+2)}{\log 2}, \tag{37}$$

onde são usadas as mesmas varáveis definidas na Equação 36.

A escolha do modelo de composições, baseado nas equações 36 e 37, pode ser superficial, uma vez que não leva em conta a complexidade da tarefa, podendo gerar, em alguns casos, grandes quantidades de parâmetros, resultando em funções potencialmente complexas, devido ao tamanho do conjunto de treinamento e a relação com a quantidade de entradas e saídas. Por outro lado, em outros casos há poucos parâmetros, resultando em um sub-ajuste dos dados. Desse modo, as formulações anteriormente mencionadas serão usadas como uma referência correspondente à máxima complexidade que o modelo pode atingir de acordo com suas características iniciais.

Para a definição topológica do comitê de máquinas foram definidas duas estratégias, que visam potenciar o reconhecimento de fonemas dentro de um espaço de decisão não linear. A primeira estratégia é baseada na proposta de reconhecimento fonético apresentada em (SCANLON; ELLIS; REILLY, 2007), que divide o espaço de decisão em grupos fonéticos e às experiências de classificação fonética apresentados em em (BONILLA; NEDJAH; MOURELLE, 2015). A segunda proposta é uma modificação hierarquizada que procura orientar a estratégia de reconhecimento de fonemas, considerando os níveis de desequilíbrio entre classes apresentados na Figura 33.

5.3.1.1 Topologia baseada em agrupamentos fonéticos MEF

Em primeira instância, foram consideradas as características espectrais dos diferentes grupos fonéticos expostos no Capítulo 1. Desse modo, foram criados sete sistemas especialistas, cada um especializado em certa região do espaço de decisão. A Tabela 6 apresenta a distribuição de cada classe dentro dos especialistas fonéticos. Também são expostos os índices de representatividade para cada classe fonética em relação ao total de amostras contidas dentro do conjunto de treinamento do sistema especialista.

Tabela 6: Distribuição das classes fonéticas dentro de cada sistema especialista

Especialistas	**Total amostras**	**Fonemas**	**Amostras classe**	**Índice (%)**
Vogais	16781	/o/	392	2,34
		/a/	454	2,63
		/e/	768	4,58
		/i/	1074	6,4
		/ɔ/	1718	10,24
		/u/	4472	26,65
		/ɛ/	8357	49,8
Fricativas	14988	/ʧ/	545	3,64
		/v/	964	6,43
		/r/	2954	19,71
		/s/	5180	34,56
		/z/	5345	35,66
Plosivas	2242	/d/	587	26.18
		/k/	804	35,86
		/t/	851	37,96
Nasais	1009	/m/	393	38,95
		/n/	616	61,05
Vogais Nasais	4559	/ẽ/	507	11,12
		/ĩ/	1664	36,50
		/ũ/	2388	52,28
Ditongos	6085	/u̯a/	1457	23,94
		/eɪ̯/	2008	33
		/oɪ̯/	2620	43,06
Pausa	10438	.pau	10438	100

Devido à persistência de desequilíbrio de classes entre os fonemas pertencentes aos grupos especialistas, principalmente para os grupos 'vogais' e 'fricativas', são executadas ações de sub-amostragem e sob-amostragem, apresentadas na Seção 4.2.3, no conjunto de treinamento de cada especialista para, desse modo, evitar um treinamento que gere classificadores tendenciosos. Finalmente, cada saída é ponderada pela rede de passagem, cuja função é determinar a probabilidade de ocorrência de cada especialista dentro do espaço de decisão. A Figura 39 apresenta o esquema de distribuição das máquinas especialistas e sua interação com a rede de passagem.

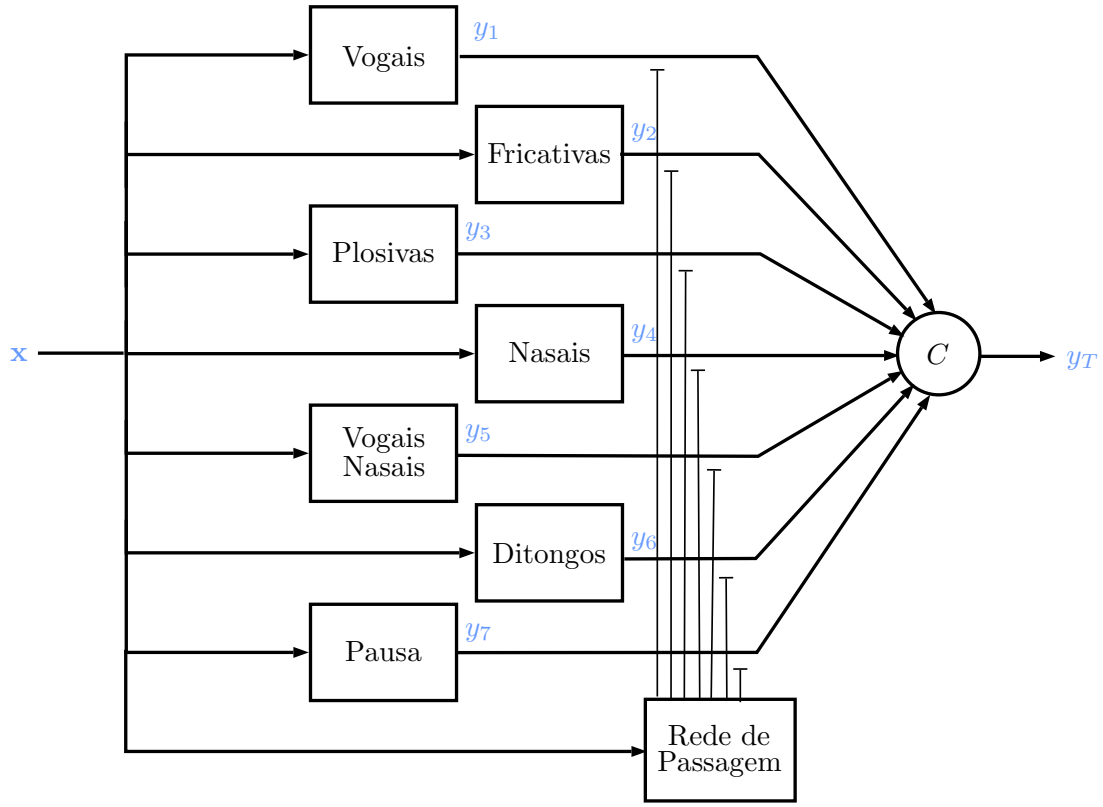


Figura 39: Componentes de especialistas distribuídos de acordo com os grupos fonéticos

5.3.1.2 Topologia baseada no desequilíbrio de classes MEHC

Esta abordagem considera a problemática de desequilíbrio de classes apresentada na Seção 4.2.3. Este modelo foi criado mediante o agrupamento de fonemas com características espectrais semelhantes e com pouco equilíbrio entre as amostras pertencentes ao conjunto de treinamento. O agrupamento de especialistas inicial foi baseado no dendrograma de agrupamento de distância média entre *clusters* apresentado em (LOPES; PERDIGÃO, 2012) e na ponderação da quantidade de amostras disponíveis. A Tabela 7 apresenta a distribuição de especialistas, o número de amostras por cada classe que compõe a máquina e o índice de representatividade de cada classe dentro do especialista. Com o intuito de nivelar os índices de representatividade, cada especialista com baixo peso é associado com outro agrupamento de características semelhantes, tanto em representatividade numérica como em componentes acústicos. Desse modo, mediante 3 níveis hierárquicos de profundidade, o modelo atinge um nível de equilíbrio razoável.

A Figura 40 expõe a distribuição de treze máquinas especialistas, com redes de passagem RP_k para $1 \leq k \leq K$, onde $K = 5$ é a quantidade de funções combinadoras, correspondentes a cada RP_k presente.

Tabela 7: Distribuição de amostras das classes fonéticas por cada especialista e em cada nível hierárquico

Especialistas			Hierarquia 1		Hierarquia 2		Hierarquia 3	
Grupos	**N Amostras**	**Índice (%)**	**N Amostras**	**Índice (%)**	**N Amostras**	**Índice (%)**	**N Amostras**	**Índice (%)**
/o/	392	40,01	846	39,89	2121	32,25	6577	11,64
/a/	454	53,66						
/ẽ/	507	39,76	1275	60,11				
/e/	768	60,24						
/i/	1074	39,23	2738	61,45	4456	67,75		
/ĩ/	1664	60,77						
/ɔ/	1718	100	1718	38,55				
/m/	393	38,95	1009	40,07	2518	32,64	7714	13,65
/n/	616	61,05						
/ɱ/	545	36,12	1509	59,93				
/v/	964	63,88						
/d/	587	26,18	2242	43,15	5196	67,36		
/k/	804	35,86						
/t/	851	37,96						
/r/	2954	100	2954	56,85				
/ũ/	2338	34,33	6810	100	6810	100	6810	12,65
/u/	4472	65,67						
/o̯a/	1457	23,94	6085	100	6085	100	6085	10,77
/oi̯/	2008	33						
/ei̯/	2620	43,06						
/s/	5180	49,22	10525	100	10525	100	10525	18,63
/z/	5345	50,72						
/ɛ/	8357	100	8357	100	8357	100	8357	14,79
Pausa	10438	100	10438	100	10438	100	10438	18,47

5.3.2 Definição topológica dos sistemas especialistas

Uma das configurações mais importantes e diretamente relacionada com o grau de desempenho de um modelo de composições é a configuração topológica de seus especialistas. Esta composição depende da complexidade do problema de classificação, da natureza própria do vetor de caraterísticas de cada classe e do comportamento em conjunto das amostras de entrada pertencentes ao bloco fonético.

Em primeira instância, é definida a profundidade da RNA especialista, ou seja, o número de camadas ocultas presentes. De acordo com a interpretação geométrica das fronteiras de decisão, apresentada em (LIPPMANN, 1987), uma rede com uma camada pode definir o espaço de decisão mediante hiperplanos. Com uma camada oculta e uma camada de saída, o modelo consegue limitar as classes mediante grandes regiões de classificação. Um modelo de rede com mais de duas camadas ocultas pode criar fronteiras de decisão mais complexas. Dada a natureza não linear do problema de classificação fonética, é estabelecida por conveniência uma profundidade de duas camadas ocultas. O uso de mais camadas pode derivar em um sobre-treinamento da rede devido a que o erro não consegue ser retro-propagado para todas as camadas. Portanto, os pesos das primeiras camadas não são adaptados corretamente (HAYKIN et al., 2009).

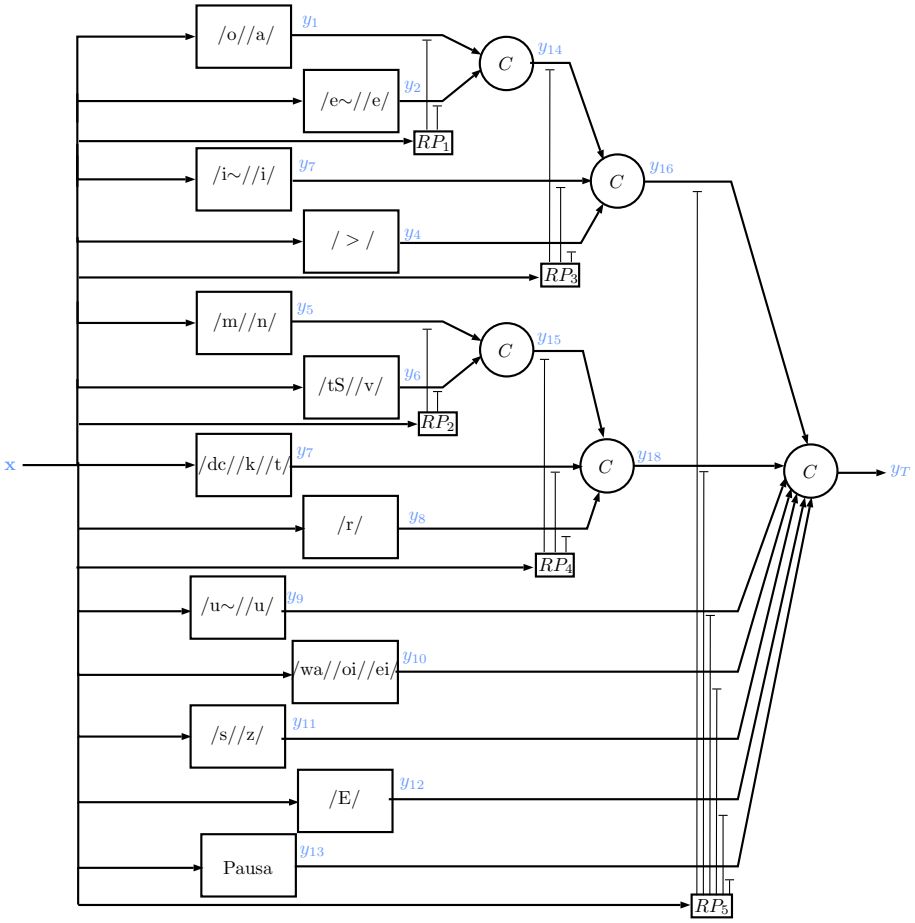


Figura 40: Componentes de especialistas distribuídos de acordo com a taxa de representatividade fonética e suas características espectrais

Definida a quantidade de camadas a serem usadas em cada rede especialista, o próximo parâmetro de configuração compete à definição do número de neurônios por camada. Uma estratégia inicial é usar a Equação 36 exposta na Seção 5.3.1, cuja função é dar uma aproximação da quantidade máxima de neurônios ocultos a ser usados sem afetar o desempenho da rede. Porém, para uma definição mais exata, foi definido um processo de seleção da topologia de rede entre diferentes hipóteses de configuração. A avaliação do desempenho de cada hipótese é dada mediante técnicas de validação cruzada, especificamente o método *hold out*, onde os dados são divididos em três conjuntos: Treinamento, validação e teste. O conjunto de treinamento é usado para o processo de aprendizado do especialista. Entretanto, o conjunto de validação é usado para gerar uma parada antecipada e desse modo evitar um sobre-treinamento da rede. Por último, o conjunto de teste mede o grau de generalização da rede para novas entradas.

Assim, cada hipótese é testada iterativamente 30 vezes de modo que o erro de validação possa ser analisado estatisticamente. Idealmente, aquele modelo que atinge o

menor erro de validação é selecionado. Não obstante, com o intuito de selecionar o modelo com menor erro e mais simples possível, é usado um teste de significância estatística, conhecido como teste T, onde são comparadas a hipótese com menor erro e outro modelo mais simples, mas com um erro de validação tolerável. De modo que, caso exista um nível de significância entre as hipóteses, é selecionada a hipótese com menor erro de validação, caso contrário, é selecionada a hipótese mais simples.

O último aspecto a considerar dentro da definição topológica dos especialistas é o comportamento dinâmico de cada fonema no tempo. Estes aspectos foram tratados no Capítulo 1. Deste modo, os grupos fonéticos que são constantes no eixo temporal são abordados mediante redes estáticas que avaliam individualmente cada vetor de características entrante. Por outro lado, aqueles grupos fonéticos que contêm uma variação espectral no tempo, como os ditongos, devem ser classificados de acordo com o conjunto de amostras passadas, de modo que possa ser avaliado como um bloco fonético. A Figura 41 apresenta o esquema que representa o modo em que os dados passados são tomados pela RNA especialista.

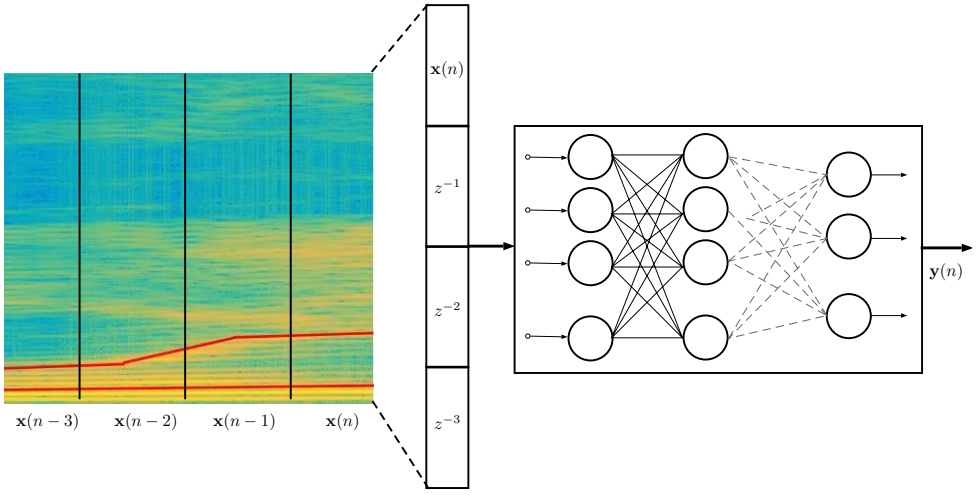


Figura 41: Deslocamento temporal da amostras dos ditongos para a rede especialista

Assim, a definição da quantidade de atrasos é determinada de acordo com o número amostras necessárias para representar o comportamento dinâmico do fonema. Este parâmetro na média é de 3 amostras, ou seja, 3 atrasos. Deve-se considerar que o número de atrasos na camada de entrada aumenta a complexidade do especialista.

5.3.3 Aspectos de implementação das redes de passagem

Uma rede de passagem tem a função de determinar a probabilidade de ocorrência de uma classe de acordo com as características do vetor de características. Como já foi descrito na Seção 4.3, a rede de passagem define o especialista com maior relevância mediante a ponderação de suas saídas. Devido a que as saídas de uma rede de passagem apresentam uma medida de probabilidade, o conjunto de etiquetado deve ser expressado em termos de zeros com uma única saída expressada como a unidade.

Para a implementação de uma rede de passagem dentro de um sistema de classificação de múltiplas classes, são necessárias tantas saídas como elementos a serem classificados. Uma camada oculta, cuja configuração determina de modo geral a distribuição das amostras de cada classe no espaço de decisão. Um conjunto de etiquetas referenciadas à saída desejada, expressadas entre 0 e 1. O número de neurônios da camada oculta é determinado do mesmo modo que a configuração dos especialistas, ou seja, um teste iterativo para um conjunto de hipóteses, onde é selecionada a topologia com menor entropia cruzada de validação. A Tabela 8 apresenta as configurações usadas nas experiências MEF e MEHC.

Tabela 8: Configuração topológica das redes de passagem

	# RP	Neurônios na camada1	Neurônios na camada de saída
MEF	1	20	7
MEHC	1	15	2
	2	10	2
	3	20	3
	4	15	3
	5	25	7

5.3.4 Função combinadora

Após a escolha do conjunto de redes especialistas a compor o comitê, o passo seguinte é a escolha dos métodos para a combinação de suas saídas. Este mecanismo deve reunir em totalidade o conjunto de soluções parciais atingidas por cada especialista para dar uma resposta ao problema original. Dado que cada especialista está centrado em uma área especifica do espaço de decisão pré-determinado, é estabelecido um sistema simples de votação, onde cada saída do módulo especialista, ponderada por uma rede de passagem,

indica sua probabilidade de ocorrência na saída final de acordo com a amostra processada. Desse modo, a saída global do sistema é associada ao módulo especialista que ofereça maior confiança.

Por outro lado, na topologia apresentada como MEHC, cada comparador usado dentro dos níveis hierárquicos, está configurado do mesmo modo, onde a saída ganhadora é aquela classe ponderada pela rede de passagem correspondente, com maior confiança, ou seja, maior peso de saída. Para identificar o especialista e a classe ganhadora no nível hierárquico final, é criado um vetor com 3 componentes gerado por cada especialista, conhecido como vetor identificador. O primeiro dado indica o índice do especialista, o segundo dado apresenta a subclasse selecionada, correspondente ao especialista relacionado no primeiro campo, e o terceiro espaço do vetor representa o valor de saída do neurônio relacionado à subclasse do especialista. Cabe notar que as ponderações feitas pela rede de passagem são calculadas usando a terceira posição do vetor identificador e este só é modificado momentaneamente para tomar a decisão de votação. Para o seguinte nível hierárquico o vetor do especialista ganhador mantém seus dados iniciais. A Figura 42 apresenta a dinâmica dos comparadores para um sistema hierárquico, onde é apresentado um exemplo do funcionamento do vetor de identificador associado a cada especialista.

Cada cor da Figura 42 representa um nível hierárquico diferente. Neste exemplo, a classe ganhadora é aquela correspondente ao fonema /ei/ pertencente ao especialista 10. Cabe notar que existe uma possibilidade de que o especialista não identifique o vetor de características de entrada dentro de seu campo de decisão. Nesse caso, o especialista etiqueta a saída como /-/.

5.3.5 Pós-processamento usando uma rede recorrente

Muitos algoritmos de treinamento das RNAs não são capazes de implementar mapeamentos dinâmicos, como por exemplo o algoritmo de retropropagação simples, que pode apenas aprender mapeamentos estáticos. O tempo pode ser representado pelo efeito que acarreta no processamento de sinais, que significa oferecer características dinâmicas ao mapeamento realizado pela rede, tornando-a sensível a sinais que variem com o tempo. Para uma RNA ser considerada dinâmica, é preciso que possua memória (ELMAN, 1990).

As redes recorrentes possuem realimentação entre os neurônios de camadas diferentes. Estas redes possuem capacidade de memória, reproduzindo seu comportamento

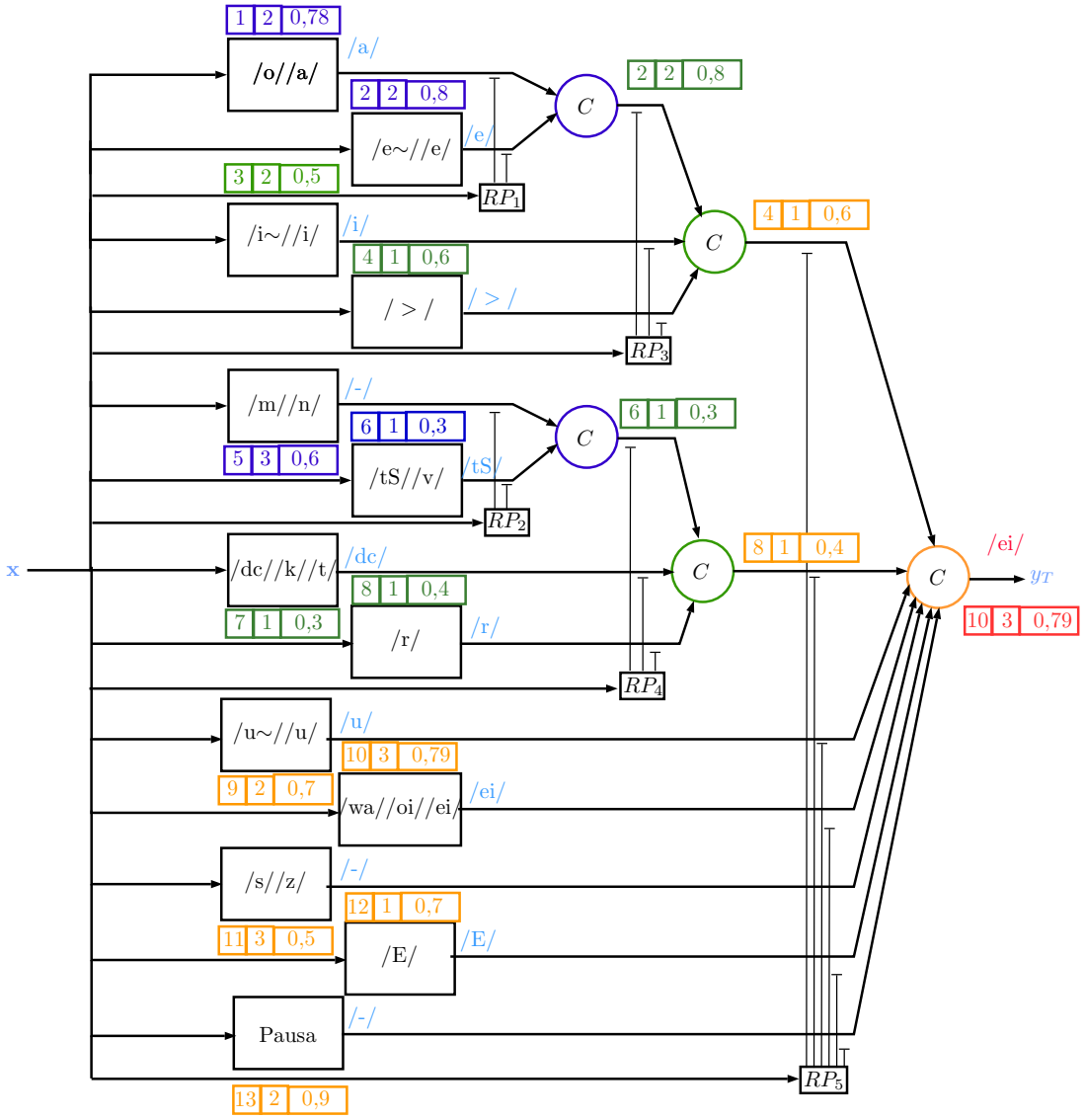


Figura 42: Estratégia de combinação de especialistas para níveis hierárquicos

com relação aos padrões anteriormente apresentados e, desta forma, possibilitando processar sequências de informações. O modelo adotado para este pós-processamento é uma rede Hopfield, apresentado no Capítulo 4, na Figura 27, de uma camada oculta e 4 neurônios, com uma componente de atraso para a entrada e saída de 3 tempos. O método de treinamento é o algoritmo de propagação de erro no tempo, usando o conjunto de dados usados para a configuração dos sistemas especialistas, mas apresentado de um modo sequencial e ordenado de modo que possa aprender o comportamento sequencial de um conjunto de sons falados.

5.4 Medidas de desempenho um sistema de RAF

Na avaliação de um sistema de RAF são importantes as medidas de desempenho que não só permitam uma análise cuidadosa e uma boa interpretação do sistema, mas que também estejam adequadas à tarefa a ser avaliada. Neste caso é importante determinar a taxa de acerto do fonema reconhecido e a correspondência da posição do fonema dentro da sequência de fala analisada. Por outro lado, também é necessário medir o grau de influência do arranjo de especialistas dentro de um sistema de RAF, mediante o cálculo dos desempenhos individuais de cada especialista e a comparação de diferentes hipóteses topológicas do comitê de máquinas, tomando em consideração o número de especialistas e a organização hierárquica destes dentro do comitê. As próximas 2 seções descrevem separadamente as medidas usadas.

5.4.1 Taxa de acurácia fonética

Nos sistemas de reconhecimento de fala, a medida de avaliação de desempenho usada é a taxa de erro fonético PER (*Phoneme Error Rate*) ou a taxa de acurácia correspondente, PAR (*Phoneme Accuracy Rate*). Esses tipos de métricas requerem uma sequência fonética reconhecida e um vetor de transcrições correspondentes às locuções de teste. A comparação entre a sequência de fonemas reconhecidos e a sequência de etiquetas de referência é feita usando um algoritmo de programação dinâmica, conhecido como distância de Levenshtein (KOHONEN; SOMERVUO, 1998). Este algoritmo considera a quantidade de vezes que um fonema é substituído por outro, s, as vezes em que um fonema é deletado, a, e quantos fonemas foram inseridos, i, a soma destes eventos é apresentada como DL. O Algoritmo 3 apresenta o procedimento para contabilizar as diferenças entre dois vetores de fonemas, gerando como saída a quantidade de diferenças entre os vetores.

Depois de obter o índice de diferença entre os vetores de fonemas, é calculada a taxa de acurácia fonética:

$$PAR = 1 - PER = \left(1 - \frac{a+s+i}{N}\right) * 100 = \left(1 - \frac{DL}{N}\right) * 100, \tag{38}$$

onde N é o comprimento total da sequência de fonemas analisada e DL é a distância de Levenshtein, que considera as mudanças dadas por substituição, s, os fonemas inseridos, i, e os fonemas deletados, a.

Algoritmo 3 Distância de Levenshtein

Entrada y: Caracteres de fonemas calculados. d: Caracteres de etiquetas
Saída DL: Inteiro, Distância Levenshtein

```
1: r ← y.comprimento; c ← d.comprimento;
2: M(1 : r + 1, 1 : c + 1) = 0; custo = 0;
3: para todo k = 0 → r faça
4:    M(k, 0) = k;
5: fim para;
6: para todo j = 0 → c faça
7:    M(0, j) = j;
8: fim para;
9: para todo k = 1 → r faça
10:   para todo j = 1 → c faça
11:      se y(k)==d(j) então
12:         custo=0;
13:      senão
14:         custo=1;
15:      fim se;
16:      f(1) = M(k − 1, j) + 1; //Deletar
17:      f(2) = M(k, j − 1) + 1; //Inserir
18:      f(3) = M(k − 1, j − 1) + custo; //Substituir
19:      M(k, j) =Min(f);
20:   fim para;
21: fim para;
22: DL = M(r + 1, c + 1);
```

5.4.2 Medidas de desempenho para classificação

A exatidão de uma classificação pode ser avaliada mediante o cálculo do número de amostras corretamente classificadas (verdadeiros positivos), o número de amostras bem classificadas, mas que não pertencem à classe avaliada (verdadeiros negativos), e amostras que ou foram atribuídas de forma incorreta à classe (falsos positivos) ou que não foram reconhecidas como exemplos da classe (falsos negativos).

Em uma classificação multiclasse, uma amostra de entrada é etiquetada só como uma única classe c_n, pertencente a um conjunto de classes não sobrepostas **c**, de N elementos. Para uma classe individual, c_n, a avaliação do desempenho é definida por: verdadeiros positivos, VP_n, verdadeiros negativos, VN_n, falsos positivos, FP_n, e falsos negativos FN_n. A Tabela 9 apresenta o mapa de confusão para um problema de múltiplas classes, onde são comparadas a saída calculada y e a saída desejada d. Como exemplo, é apresentada a distribuição das características de classificação dos dados em função da classe 2, c_2.

Tabela 9: Matriz de confusão para classificação com múltiplas classes

		Saída y			
		C_1	C_2	**...**	C_n
Saída d	C_1	VN_2	$\mathbf{FP}_2$	FN	FN
	C_2	$\mathbf{FN}_2$	$\mathbf{VP}_2$	$\mathbf{FN}_2$	$\mathbf{FN}_2$
	...	FN	$\mathbf{FP}_2$	VN_2	FN
	C_n	FN	$\mathbf{FP}_2$	FN	VN_2

A capacidade de um sistema para reconhecer as amostras pertencentes à classe c_n é conhecida como sensibilidade, S. Nesta medida, a quantidade de amostras detectadas como verdadeiras positivas são relacionadas com o número de saídas desejadas etiquetadas como c_n. A Equação 39 apresenta esta relação:

$$S_n = \frac{VP_n}{VP_n + FN_n}, \tag{39}$$

onde os índices VP_n e FN_n correspondem à contagem de eventos de classificação dados para a classe n. Esta taxa de desempenho pode ser afetada por problemas de etiquetado relacionados com a classe n, sendo que estas amostras ruidosas alteram o comportamento do classificador como foi descrito na Seção 4.2.2.

Por outro lado, o índice de reconhecimento de amostras positivas reais, dado um conjunto de amostras classificadas como positivas, é conhecido como precisão. Esta medida está determinada pela seguinte equação:

$$P_n = \frac{VP_n}{VP_n + FP_n}, \tag{40}$$

onde P_n é a precisão calculada para a classe n. Este tipo de medição é útil para evidenciar uma classe que tenha sido sobreajustada, por exemplo, o caso em que qualquer amostra de um conjunto de teste de múltiplas características seja etiquetada como c_n. Neste exemplo, o conjunto de falsos positivos é considerável e, portanto, o índice de precisão para a classe n é baixo.

A acurácia total do sistema representa a taxa de acerto global, proporção de predições corretas em relação ao tamanho do conjunto de dados. Esta medida é calculada usando a Equação 41:

$$Ac = \frac{\sum_{n=1}^{N} VP_n}{\sum_{n=1}^{N} VP_n + FP_n} = \frac{\text{Total amostras bem classificadas}}{\text{Número total de amostras de teste}}, \tag{41}$$

onde N é o número total de classes, VP_n são as amostras bem preditas pertencentes à classe n e FP_n são os falsos positivos de cada classe, os quais, ao serem somados em conjunto, geram a totalidade de amostras que não foram bem classificadas.

5.5 Considerações finais do capítulo

Este capítulo apresentou todas as considerações físicas e técnicas usadas na implementação de um sistema de RAF, mediante a definição topológica e técnica dos elementos que constituem o sistema de pré-processamento. Dois modelos de composição de especialistas foram propostos baseados nas características espectrais e na distribuição das classes dentro do conjunto de treinamento. Além disso, é exposto o pós-processamento dinâmico que inclui um conhecimento das saídas passadas para decidir o valor da saída atual. No capítulo seguinte são apresentados os resultados nas configurações topológicas para cada rede especialista e os resultados da implementação dos modelos propostos neste capítulo.

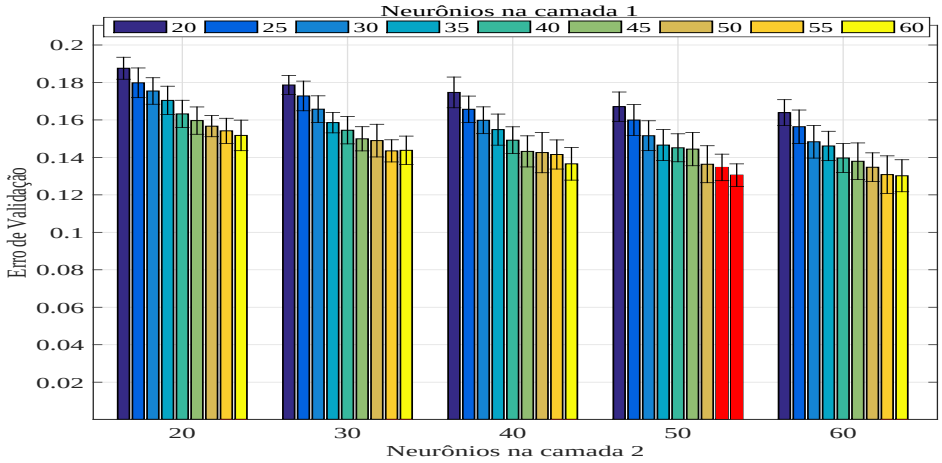
Capítulo 6

RESULTADOS

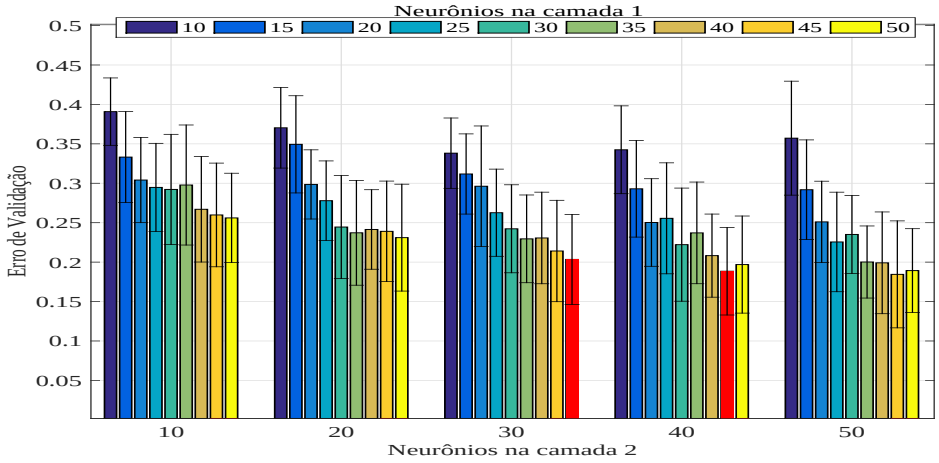
NSTE capítulo são apresentados e analisados os resultados alcançados com a implementação descrita no Capítulo 5. Na Seção 6.1 são expostos os resultados referentes aos testes das hipóteses de configuração topológica, feita para cada rede especialista contida dentro das duas propostas de modelos de mistura. Nas próximas seções, é feita a avaliação dos métodos propostos, mediante testes do modelo conduzidos em SIMULINK®, usando dois bancos de teste. Os resultados destes testes são analisados de acordo com as regras de medição descritas na Seção 5.4.2, do Capítulo 5, e são comparados os comportamentos gerados pelos modelos em cada experiência. Por último, os desempenhos das redes propostas são comparados com trabalhos relacionados no reconhecimento automático de fonemas.

6.1 Topologia das redes especialistas.

De acordo com as técnicas de configuração topológica, apresentadas na Seção 5.3.2 do Capítulo 5, usadas para o estabelecimento de cada rede especialista, são apresentadas 45 hipóteses relacionadas com a configuração topológica, constituídas por diferentes quantidades de neurônios distribuídos entre duas camadas ocultas. Estes testes visam obter a configuração topológica que gere menor erro de validação durante o treinamento da rede. Cada topologia é testada iterativamente 30 vezes, de modo que possa ser analisada estatisticamente. A Figura 43 apresenta um exemplo que expõe os erros de validação para cada hipótese testada. Neste caso, é selecionada a hipótese com menor erro de validação e outra hipótese correspondente a uma topologia mais simples que tenha um erro de validação aceitável.



(a) Erro de validação para o especialista em fonemas vogais.



(b) Erro de validação para o especialista em fonemas nasais.

Figura 43: Média do erro de validação para hipóteses apresentadas pelos especialistas.

Os detalhes dos testes estão contidos no Apêndice A, que apresenta todos os resultados por hipótese e as figuras correspondentes, onde são comparados os erros de validação para cada rede especialista. Em geral, o incremento do número de neurônios dentro das camadas ocultas ajuda a reduzir as taxas de erro de validação.

Para determinar o nível de significância da rede com menor erro de validação, em relação com à outra hipótese selecionada, é feito um teste de hipóteses conhecido como teste T. As Tabelas 10 e 11 apresentam uma consolidação dos resultados destes testes, feitos para cada especialista pertencente aos modelos de mistura propostos MEF e MEHC.

Estes testes foram implementados usando um nível de confiança $a = 0,025$ em um conjunto de amostras $n = 30$. Este comprimento corresponde ao número de vezes em que cada hipótese foi testada. O valor t, calculado de acordo com as condições anteriores é de

Tabela 10: Resultados do Teste T para as topologias candidatas dos especialistas pertencentes ao modelo MEF

Especialista	Topologia	$t_{a,n-1}$	H_0
Vogais	**(50-40)**	2,3588	Rejeitada
	(40-40)		
Fricativas	**(60-40)**	1,2355	Aceita
	(50-60)		
Plosivas	**(45-50)**	2,2043	Rejeitada
	(40-30)		
Nasais	**(50-30)**	1,3647	Aceita
	(45-40)		
Vog Nasais	**(50-30)**	0,8632	Aceita
	(50-40)		
Ditongos	**(55-50)**	1,2282	Aceita
	(60-60)		
Pausas	**(55-30)**	0,8543	Aceita
	(45-60)		

Tabela 11: Resultados do Teste T para as topologias candidatas dos especialistas pertencentes ao modelo MEHC

Especialista	Topologia	$t_{a,n-1}$	H_0
/o/ /a/	**(50-40)**	2,3588	Rejeitada
	(40-40)		
/ẽ/ /e/	**(30-50)**	1,1498	Aceita
	(45-40)		
/ĩ/ /i/	**(50-50)**	2,1323	Rejeitada
	(45-30)		
/ɔ/	**(25-30)**	1,9560	Aceita
	(35-30)		
/m/ /n/	**(50-30)**	1,3647	Aceita
	(45-40)		
/ʧ//v/	**(40-50)**	0,444	Aceita
	(45-50)		
/d/ /k/ /t/	**(50-30)**	0,8632	Aceita
	(50-40)		
/r/	**(45-30)**	0,8183	Aceita
	(50-40)		
/ũ//u/	**(40-50)**	0,9312	Aceita
	(50-50)		
/ʊ̯**a**//o**I**̯//e**I**̯/	**(55-50)**	1,2282	Aceita
	(60-60)		
/s/ /z/	**(30-40)**	1,9821	Aceita
	(50-50)		
/ɛ/	**(45-40)**	1,3432	Aceita
	(50-50)		
Pausas	**(55-30)**	0,8543	Aceita
	(45-60)		

$t_0 = 2,0452$. A hipótese nula $H_0 : \mu_1 = \mu_2$ é rejeitada no caso em que $t > t_0$. Dadas as condições anteriores, a maior parte dos testes nas redes especialistas indica que a topologia com menor erro de validação pode ser substituída por uma rede menos complexa, mas com um erro de validação aproximado.

Cabe destacar que os especialistas plosivas, nasais, ditongos e pausas, pertencentes ao modelo MEF, são equivalentes aos especialistas do modelo MEHC: /d//k//t/, /m//n/, /ʊ̯a//oɪ̯//eɪ̯/ e pausas, respetivamente. Portanto, são mantidas as configurações topológicas destes especialistas para ambos modelos.

6.2 Resultados do método de mistura do tipo MEF

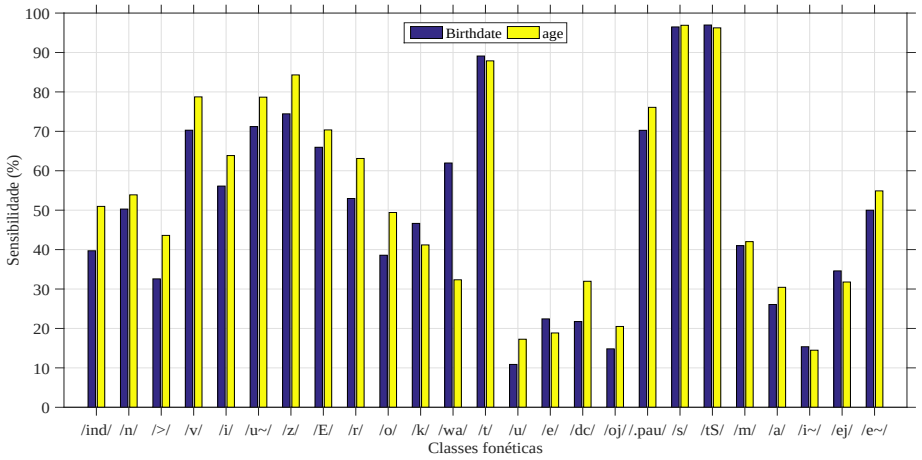
Os resultados apresentados nesta seção correspondem à primeira topologia proposta, apresentada na Seção 5.3.1.1. Este modelo foi treinado usando o banco de dados *zipcode* pertencente ao corpora *Spoltech* e testado com os bancos de dados *age* e *birthdate*, de modo que possa ser avaliado o nível de generalização dos modelos propostos.

Para determinar o influxo do pós-processamento dinâmico sobre o modelo de misturas, são realizadas duas experiências independentes usando só o MEF e o MEF mais o pós-processamento dinâmico no sinal de saída.

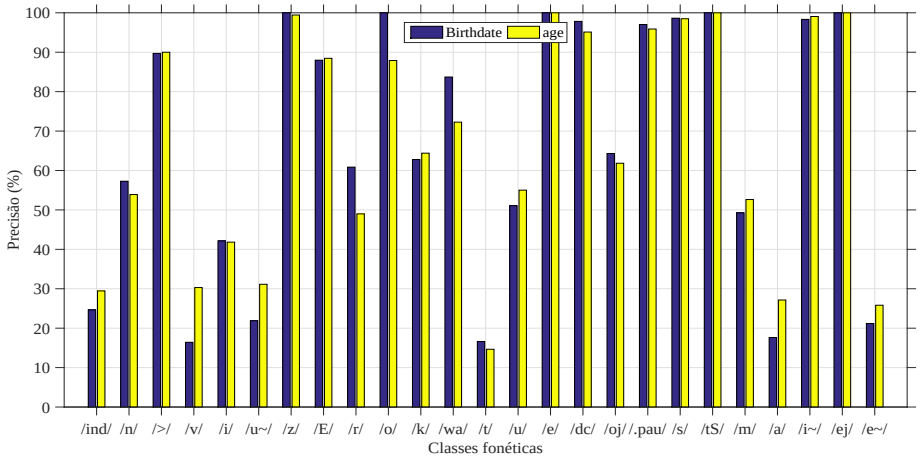
6.2.1 Resultados usando o modelo MEF sem pós-processamento

Nesta experiência, foi testado o modelo MEF de acordo com os dados obtidos na saída do comparador de especialistas, ou seja, é considerada a solução que identifica a classe fonética sem considerar as saídas passadas. A Figura 44 apresenta a taxa de sensibilidade e precisão obtidas para o teste deste modelo usando os bancos de dados *birthdate* e *age*.

De acordo com as Figuras 44(a) e 44(b), o modelo MEF apresenta um comportamento similar para ambos os testes, com uma pequena diferença de desempenho entre cada classificação fonética. Evidentemente aquelas classes fonéticas pertencentes a grupos com baixa taxa de representação obtiveram os desempenhos de sensibilidade mais baixos, sendo que as ponderações feitas pela rede de passagem dão prioridade àqueles fonemas pertencentes aos grupos fonéticos mais numerosos, ou seja, as vogais, os fonemas fricativos e as pausas.



(a) Sensibilidade para cada classe fonética usando o Modelo MEF



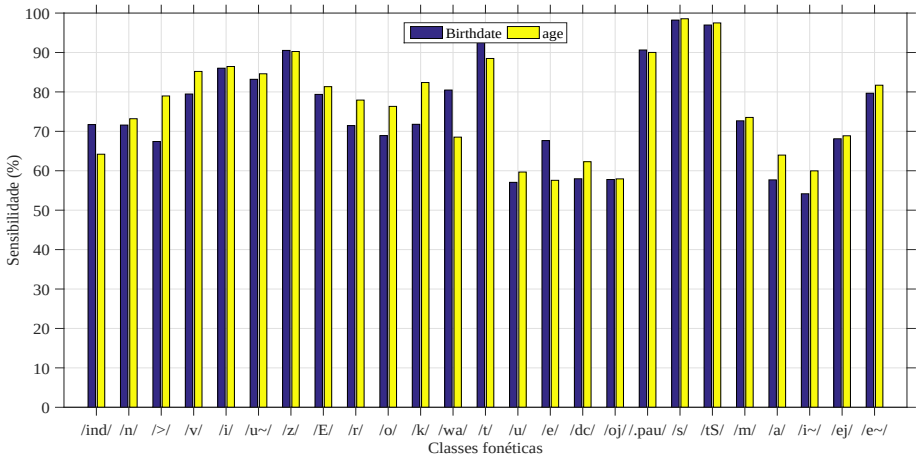
(b) Precisão para cada classe fonética usando o Modelo MEF

Figura 44: Resultados de classificação usando o modelo MEF

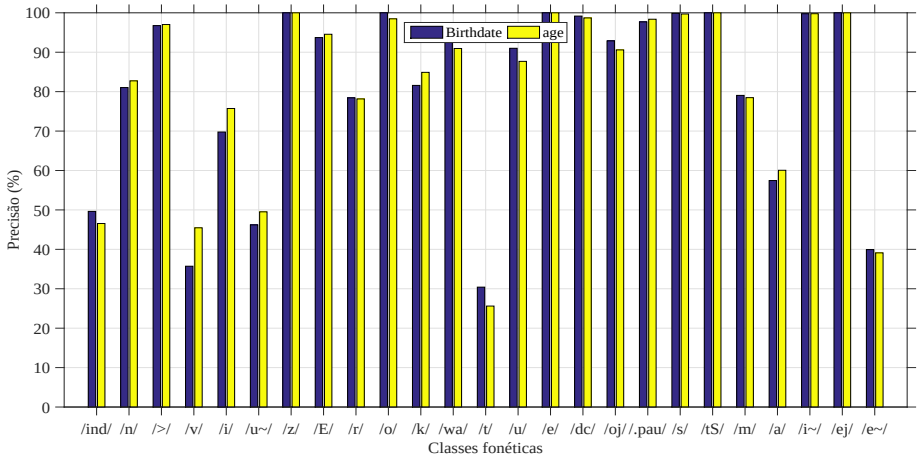
6.2.2 Resultados do modelo MEF com pós-processamento dinâmico

Neste caso, os resultados foram obtidos depois de um pós-processamento usando uma rede recorrente. Desse modo, são consideradas as saídas passadas como uma fonte adicional de informação que é somada às características frequenciais aportadas pelo modelo MEF. A Figura 45 apresenta a sensibilidade e precisão medidas usando os dois bancos de dados de teste da experiência passada.

As Figuras 45(a) e 45(b) evidenciam um incremento geral e considerável no desempenho do sistema de RAF, tanto na sensibilidade como na taxa de precisão na classificação de cada fonema. A Tabela 12 apresenta uma comparação entre os desempenhos dados nos testes das redes MEF e MEF com rede recorrente. Para uma comparação visual, os resultados destes modelos são apresentados na Figura 47. É claro que o modelo MEF



(a) Sensibilidade para cada classe fonética



(b) Precisão para cada classe fonética

Figura 45: Resultados de classificação usando o modelo MEF e a rede recorrente de pós-processamento

seguido pela rede recorrente supera o MEF sem pós-processamento dinâmico. A classe com maior melhoria na sensibilidade é a do fonema /u/ com 44,29% de melhoria. Quanto à precisão, a classe mais beneficiada com a implementação em pós-processamento é a do fonema /a/ com 36,36% de melhoria.

Em termos gerais, a acurácia do modelo com pós-processamento é melhorada, em grande parte, devido à diminuição de oscilação temporal entre classes como é apresentado na Figura 46, onde é exposto como o modelo de misturas reconhece individualmente cada amostra. Porém, devido à falta de relação com as amostras passadas, pode gerar uma oscilação, principalmente entre as classes foneticamente similares. O comportamento da rede com pós-processamento demostra a importância da interconexão da resposta atual com as amostras passadas. Consequentemente, os níveis de desempenho dados em sensibi-

Tabela 12: Comparação de desempenhos entre os modelos MEF e MEF+RNN

	MEF		MEF+RNN		Melhoria	
	Sensibilidade %	**Precisão %**	**Sensibilidade %**	**Precisão %**	**Sensibilidade %**	**Precisão %**
/ind/	45,332	27,044	67,952	48,093	22,620	21,049
/n/	52,087	55,586	72,395	81,883	20,308	26,297
/>/	38,091	89,853	73,198	96,884	35,107	7,031
/v/	74,515	23,363	82,333	40,592	7,818	17,230
/i/	59,984	41,978	86,232	72,748	26,248	30,770
/u~/	74,934	26,510	83,900	47,853	8,966	21,343
/z/	79,376	99,719	90,389	100,000	11,013	0,281
/E/	68,160	88,222	80,353	94,133	12,192	5,911
/r/	58,043	54,926	74,689	78,290	16,646	23,364
/o/	43,993	93,947	72,623	99,237	28,630	5,289
/k/	43,918	63,589	77,082	83,243	33,164	19,654
/wa/	47,156	77,991	74,503	92,687	27,347	14,696
/t/	88,491	15,645	90,489	28,014	1,998	12,369
/u/	14,071	53,045	58,369	89,346	44,298	36,301
/e/	20,641	100,000	62,611	100,000	41,971	0,000
/dc/	26,853	96,474	60,133	98,937	33,280	2,463
/oj/	17,667	63,078	57,858	91,732	40,191	28,654
/.pau/	73,174	96,463	90,326	98,029	17,152	1,566
/s/	96,696	98,565	98,391	99,788	1,695	1,224
/tS/	96,610	100,000	97,235	100,000	0,625	0,000
/m/	41,505	50,943	73,100	78,765	31,595	27,822
/a/	28,252	22,388	60,834	58,753	32,582	36,364
/i~/	14,923	98,699	57,070	99,766	42,147	1,067
/ej/	33,190	100,000	68,495	100,000	35,305	0,000
/e~/	52,447	23,507	80,688	39,517	28,241	16,010

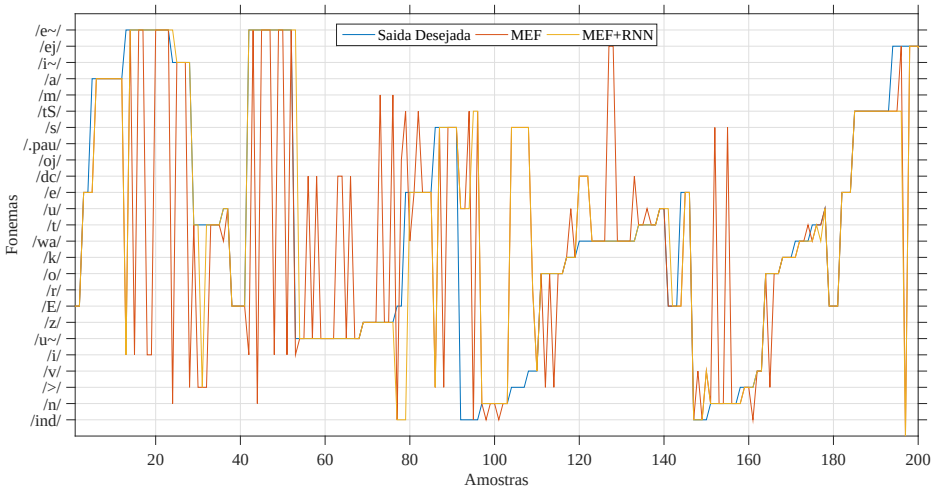
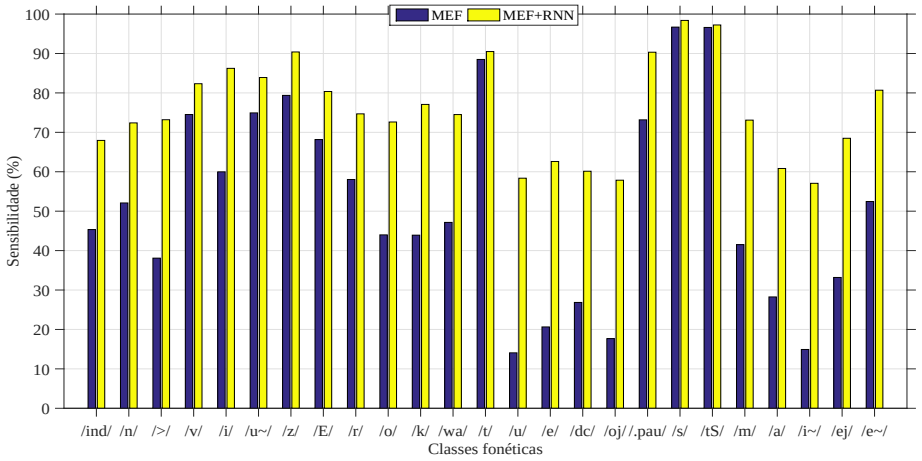
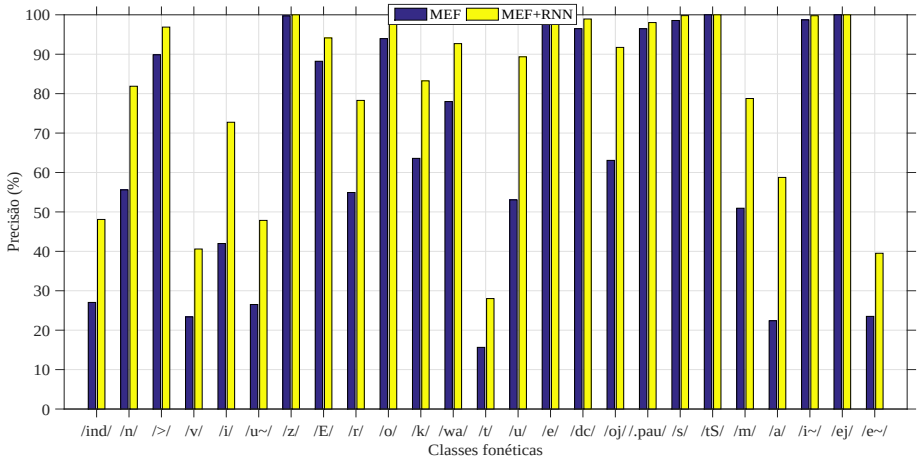


Figura 46: Exemplo do comportamento da saída do modelo MEF com e sem pós-processamento vs a saída desejada

lidade e precisão, apresentados na Tabela 12 são melhoradas substancialmente. Os níveis de acurácia global e PER também são melhorados, sendo que o modelo MEF apresenta uma acurácia de $55,090\%$ e um PER de $54,521\%$. Por outro lado, o modelo MEF+RNN apresenta melhores resultados de acurácia, $77,753\%$, e de PER, $21,362\%$.



(a) Comparação entre níveis de sensibilidade



(b) Comparação entre níveis de precisão

Figura 47: Comparação entre modelos MEF com e MEF+RNN

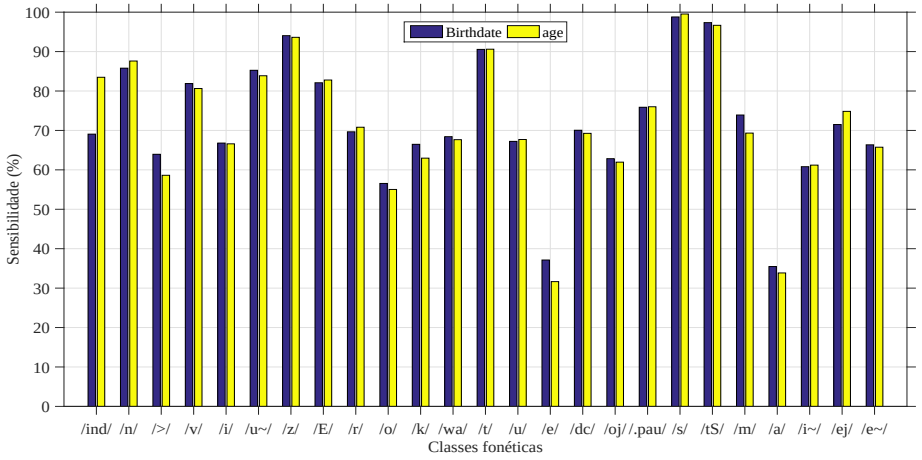
6.3 Resultados do método de mistura do tipo MEHC

Nesta seção são apresentados os resultados obtidos do modelo MEHC, apresentado na Seção 5.3.1.2, usando os bancos de dados de teste, *birthdate* e *age*. Do mesmo modo que com as experiências anteriores, foi testado o modelo MEHC independentemente, com o objetivo de visualizar o comportamento da saída para o MEHC, e para o mesmo modelo com sua saída pós-processada.

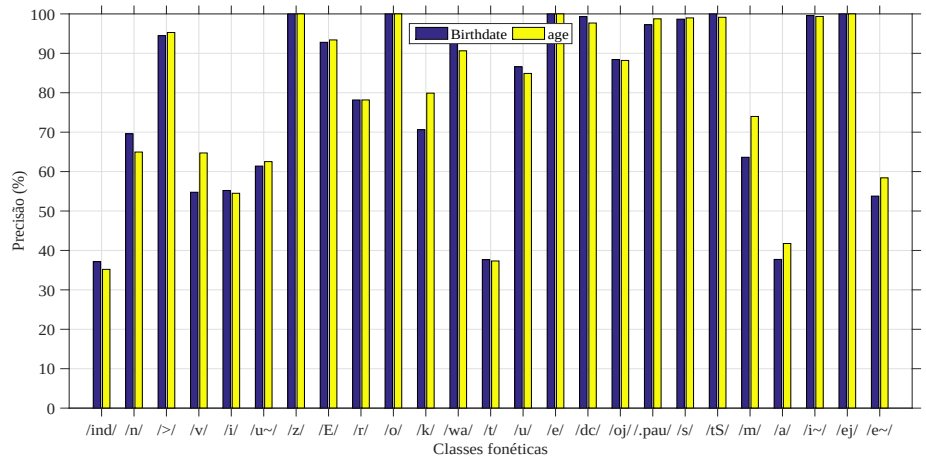
6.3.1 Resultados usando o modelo MEHC sem pós-processamento

Nesta experiência foi testado o modelo MEHC individualmente, usando os dois bancos de dados de testes, a fim de visualizar o comportamento da classificação estática, dado

um sistema de aprendizado treinado com um banco de dados desequilibrado. A Figura 48 apresenta a sensibilidade e precisão das experiências.



(a) Sensibilidade para cada classe fonética usando o Modelo MEHC



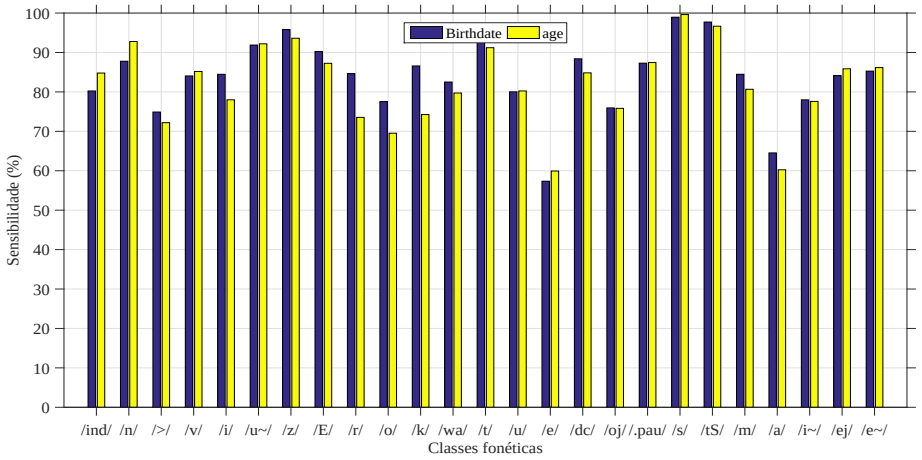
(b) Precisão para cada classe fonética usando o Modelo MEHC

Figura 48: Resultados do processo de classificação usando o modelo MEHC

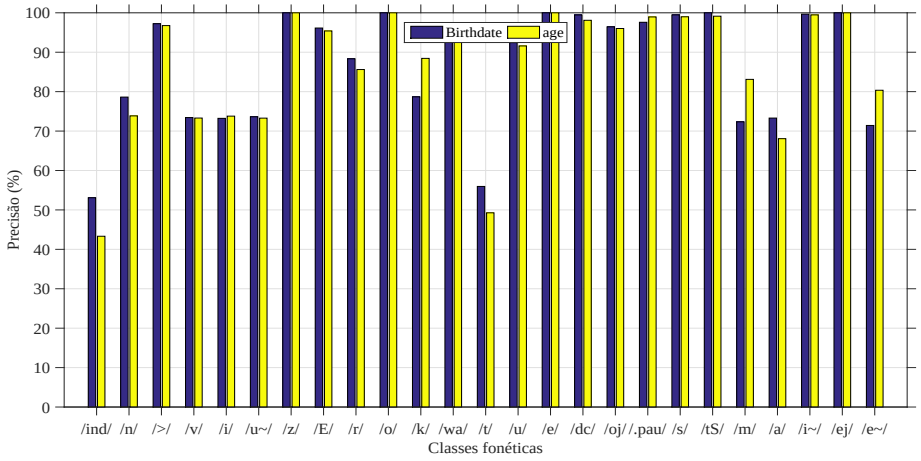
De acordo com as Figuras 48(a) e 48(b), o modelo MEHC expõe um comportamento estável para ambas as experiências. Neste caso, as classes fonéticas com pouca quantidade de amostras dentro do conjunto de treinamento apresentaram os níveis de sensibilidade mais baixos.

6.3.2 Resultados do modelo MEHC com pós-processamento dinâmico

Os resultados desta experiência foram obtidos depois do pós-processamento executado por uma rede recorrente. A Figura 49 apresenta os níveis de sensibilidade e precisão para cada classe fonética de acordo com o banco de teste usado.



(a) Sensibilidade para cada classe fonética usando o Modelo MEHC+RNN



(b) Precisão para cada classe fonética usando o Modelo MEHC+RNN

Figura 49: Resultados do processo de classificação usando o modelo MEHC com pós-processamento da saída

As Figuras 49(a) e 49(b) mostram um incremento nas taxas de sensibilidade e precisão para cada fonema, melhorando o desempenho do modelo MEHC. A Tabela 13, apresenta uma comparação entre os desempenhos dados para o modelo MEHC e o mesmo modelo com um componente de pós-processamento adicional. A comparação visual é exposta na Figura 46.

Estes resultados indicam um melhoramento geral produzido pela etapa de pós-processamento, diminuindo os níveis de oscilação apresentados do mesmo modo que os expostos na Figura 46. Assim, os desempenhos já melhorados pelo modelo MEHC são reforçados pelo pós-processamento. As classes fonéticas com maiores melhorias sobre sua sensibilidade são /a/ e /e/, ambas caracterizadas por sua pouca representatividade no conjunto de dados de treinamento. Cabe notar que outro grupo de classes fonéticas com

Tabela 13: Comparação de desempenhos entre os modelos MEHC e MEHC+RNN

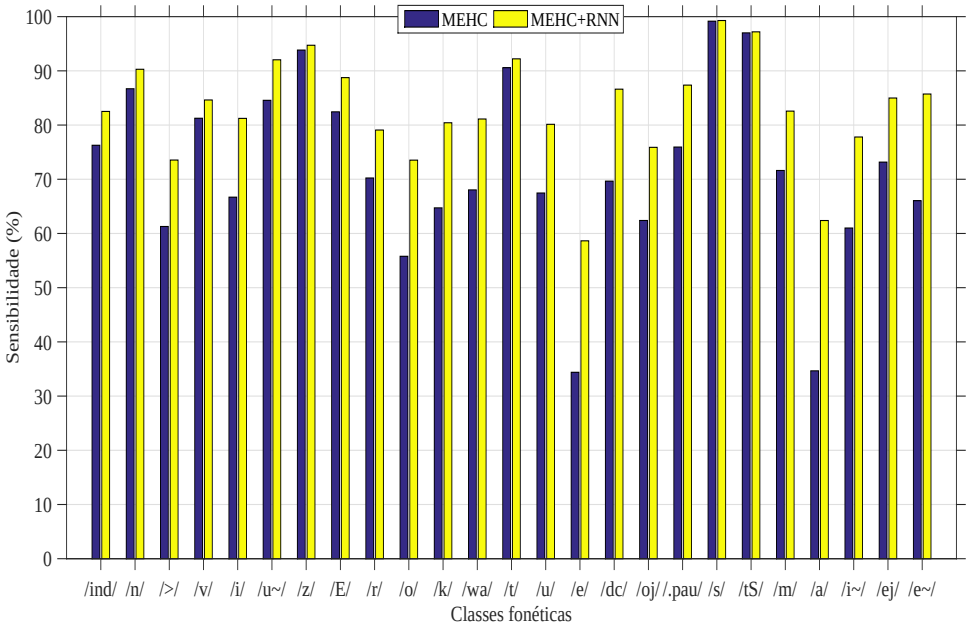
	MEHC		MEHC+RNN		Melhoria	
	Sensibilidade %	**Precisão %**	**Sensibilidade %**	**Precisão %**	**Sensibilidade %**	**Precisão %**
/ind/	76,274	36,176	82,513	48,204	6,240	12,028
/n/	86,702	67,271	90,290	76,240	3,588	8,968
/>/	61,288	94,876	73,545	96,990	12,257	2,114
/v/	81,257	59,747	84,627	73,362	3,370	13,615
/i/	66,699	54,859	81,231	73,512	14,532	18,653
/u∼/	84,564	61,964	92,041	73,455	7,477	11,491
/z/	93,833	100,000	94,722	100,000	0,889	0,000
/E/	82,442	93,098	88,749	95,769	6,306	2,670
/r/	70,234	78,174	79,082	86,972	8,849	8,798
/o/	55,792	100,000	73,527	100,000	17,735	0,000
/k/	64,725	75,261	80,425	83,562	15,699	8,300
/wa/	68,039	92,506	81,117	96,718	13,077	4,212
/t/	90,581	37,496	92,216	52,611	1,635	15,115
/u/	67,462	85,744	80,137	92,012	12,675	6,268
/e/	34,391	100,000	58,643	100,000	24,252	0,000
/dc/	69,655	98,501	86,621	98,780	16,966	0,279
/oj/	62,397	88,311	75,889	96,231	13,492	7,921
/.pau/	75,943	97,997	87,370	98,276	11,427	0,279
/s/	99,161	98,829	99,284	99,231	0,123	0,402
/tS/	97,008	99,573	97,197	99,573	0,189	0,000
/m/	71,620	68,814	82,572	77,729	10,952	8,915
/a/	34,661	39,745	62,389	70,686	27,729	30,941
/i∼/	61,005	99,454	77,795	99,571	16,790	0,117
/ej/	73,164	100,000	84,988	100,000	11,824	0,000
/e∼/	66,050	56,100	85,725	75,893	19,675	19,793

pouca representatividade, como /o/, /m/ e /tS/ teve um bom desempenho durante o processo de classificação. Isto pode indicar que o conjunto de vetores de características, correspondentes às classes /a/ e /e/, contém componentes ruidosos. De acordo com (BARANDELA et al., 2004), as classes pouco representativas de um sistema multi-classificatório têm uma maior tendência a apresentar padrões ruidosos e, portanto, diminuem o desempenho de um sistema de classificação de padrões.

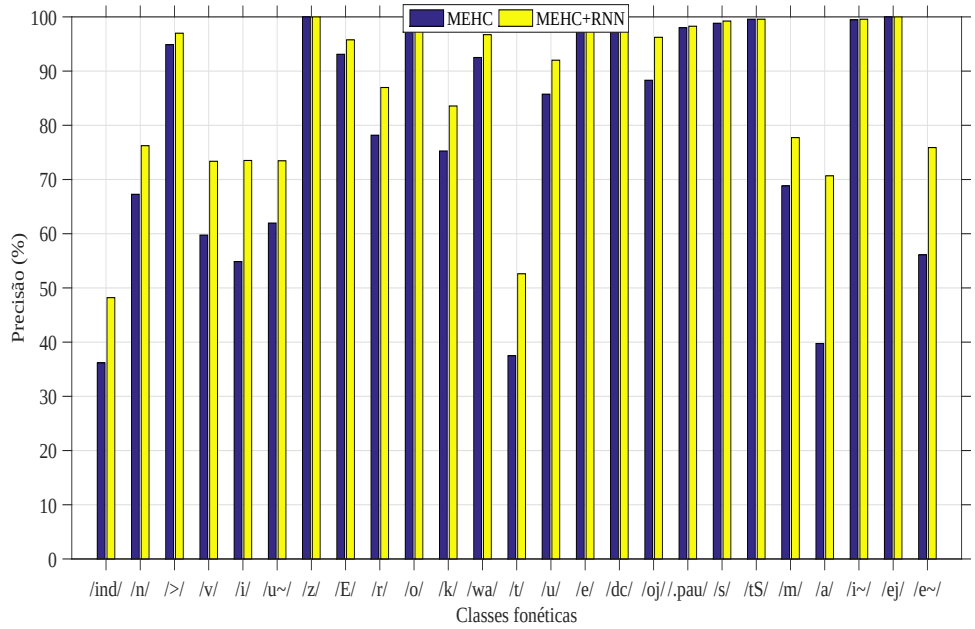
Os níveis de acurácia global e PER também são melhorados, sendo que, o modelo MEHC apresenta uma acurácia de $76,885\%$ e um PER de $33,418\%$. Por outro lado, o modelo MEHC+RNN apresenta melhores resultados de acurácia e de PER, sendo $86,525\%$ e $13,128\%$, respetivamente.

6.4 Comparações dos modelos de mistura propostos

A Figura 51 apresenta a comparação entre os desempenhos medidos nos modelos MEF e MEHC, evidenciando, que uma topologia de misturas hierárquicas, que considera o desequilíbrio de classes de um conjunto de treinamento, melhora os níveis de reconhecimento fonético, comparado com uma topologia baseada nas características acústicas dos grupos fonéticos.



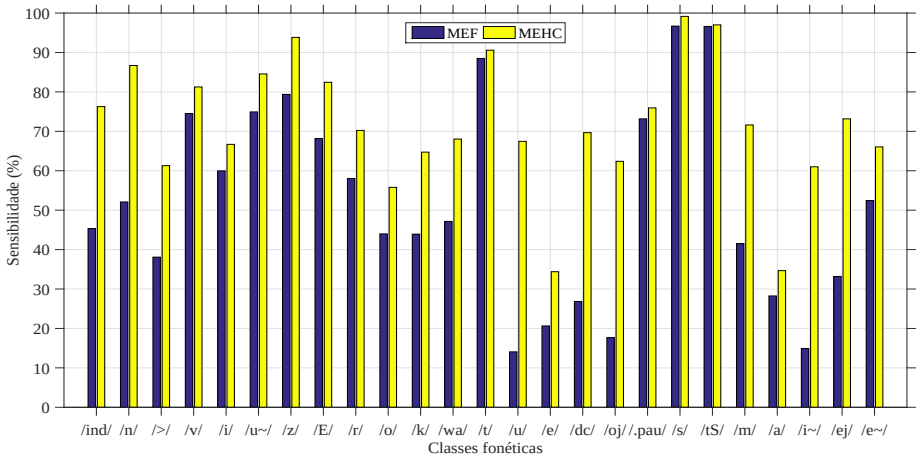
(a) Comparação entre níveis de sensibilidade



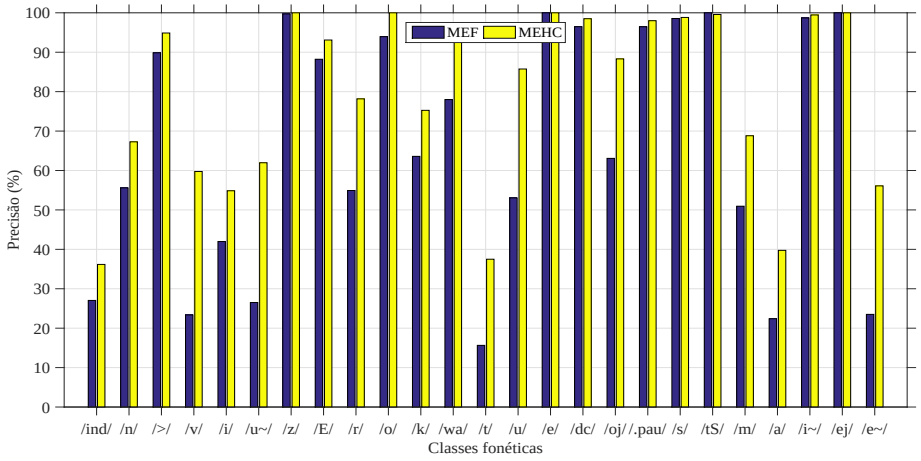
(b) Comparação entre níveis de precisão

Figura 50: Comparação entre modelos MEHC com e sem RNN no pós-processamento

Finalmente, são comparados os níveis de desempenho global, ou seja, a acurácia de cada modelo proposto e a taxa de erro fonético PER. Nesta comparação é claro que o melhor modelo para implementar um sistema de RAF, baseado em reconhecimento fonético, é aquele que une o modelo de misturas hierárquico baseado em desequilíbrio de classes, MEHC com um pós-processamento, usando uma rede neural recorrente RNN.



(a) Comparação entre níveis de precisão sensibilidade



(b) Comparação entre níveis de precisão

Figura 51: Comparação de desempenho entre modelos de mistura MEF e MEHC

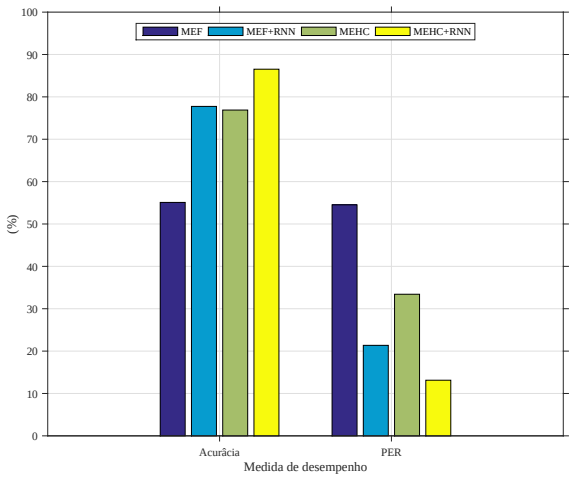


Figura 52: Desempenho global de acurácia e PER para os modelos propostos com e sem pós-processamento

6.5 Comparação dos resultados relacionados

A topologia proposta, junto com as técnicas de tratamento do conjunto de treinamento e o pós-processamento usados, geraram uma acurácia de 86,525% e uma taxa de erro fonético PER de 13,128%. A Tabela 14 apresenta alguns trabalhos relacionados com o reconhecimento de fonemas, onde são detalhadas as técnicas usadas no pré-processamento e os métodos de classificação. A informação apresentada indica um índice de acurácia entre 63,07% e 78,9%. Isto demostra que o sistema proposto nesta dissertação gera uma melhoria de 7,62%, apresentada na Figura 53 . Porém, deve-se considerar que os conjuntos de dados de treinamento usados nestes trabalhos estão contidos no corpora TIMIT, para reconhecimento de fala em Inglês, com 35 classes fonéticas, diferente do corpora usado nesta dissertação.

Tabela 14: Comparação de configuração e acurácia para trabalhos relacionados ao reconhecimento de fala

Ref	Ano	Técnica de pré-processamento	Técnica de Classificação	Acurácia (%)
(MESSAOUD; HAMIDA, 2010)	2010	MFCC	CDHMM	63,07
(TÓTH, 2011)	2011	MFCC	HMM-MLP	77,83
(UMARANI; RAVIRAM; WAHIDABANU, 2011)	2011	PLP	SMLP	78,9
(CLARKSON; MORENO, 1999)	1999	MFCC	SVM	77,6

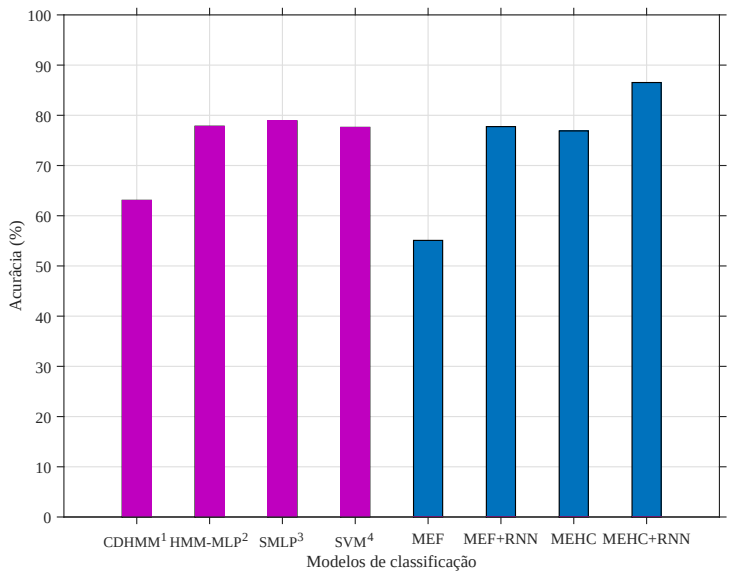


Figura 53: Comparação entre os diferentes modelos de classificação dos trabalhos relacionados e os modelos MEHC

6.6 Considerações finais do capítulo

Neste capítulo foram apresentados os resultados obtidos na avaliação de diferentes configurações topológicas, baseadas no número de neurônios nas camadas ocultas. Esses resultados indicam que, na média, um conjunto de 80 neurônios, distribuídos entre duas camadas, gera as melhores taxas de erro de validação, comparado com outras configurações.

Quatro modelos de RAF foram avaliados baseados em reconhecimento de fonemas: MEF, MEF-RNN, MEHC e MEHC-RNN. Os resultados obtidos indicam que o melhor desempenho é para o modelo MEHC-RNN, tanto na classificação individual de cada fonema, como na avaliação geral do modelo. Os resultados indicaram que uma distribuição de redes especialistas, baseada no desequilíbrio do conjunto de treinamento, aprimora o processo de aprendizagem, gerando um resultado superior, comparado com uma distribuição baseada em características fonéticas. Além disso, a implementação de um sistema de pós-processamento dinâmico, melhora as taxas de acurácia e PER, devido ao uso do conhecimento das saídas passadas, o que diminui a oscilação de classificação entre classes fonéticas semelhantes.

É importante destacar que o Apêndice A apresenta o conjunto de dados completos relacionados aos diferentes testes feitos para determinar a melhor configuração topológica para cada rede especialista, as redes de passagem e uma apresentação detalhada dos resultados dos testes usando matrizes de confusão para cada conjunto de dados testado. No capítulo seguinte serão expostas as conclusões pertinentes ao desenvolvimento e resultados desta dissertação, assim como os trabalhos futuros relacionados à proposta do sistema de RAF.

Capítulo 7

CONCLUSÕES

ESTA dissertação abordou o reconhecimento automático de fala mediante a classificação de unidades fonéticas. A realização deste trabalho é incentivada pela necessidade de usar métodos que permitem uma transcrição *online* de fala para texto. A possibilidade de trabalhar com o reconhecimento de grandes números de classes fonéticas acarreta a necessidade de um sistema de reconhecimento complexo que possa modelar comportamentos físicos não lineares e dependentes do tempo. De forma geral, o reconhecimento automático de fala mediante máquinas especialistas com pós-processamento dinâmico simplifica o processo de aprendizado e reforça o reconhecimento das classes fonéticas, facilitando assim a transcrição *online*.

7.1 Conclusões

As topologias e métodos propostos permitem classificar um sinal de fala, capturado em tempo real, em um extenso conjunto de classes fonéticas, gerando uma saída de classificação sequencial. Para o treinamento *offline* foi usado um conjunto de dados de fala que apresenta um desequilíbrio tendo em vista as classes consideradas. Para a implementação e o teste *online* do modelo proposto, foi usada a ferramenta SIMULINK®, com capacidade de execução de modelos em tempo real. Durante as experiências realizadas usando os modelos projetados, foram testados dois bancos de dados, *birthdate* e *age*, pertencentes ao Corpora *Spoltech*. Cada um desses bancos de dados apresenta aproximadamente 7 minutos de fala contínua, permitindo avaliar o desempenho do modelo mediante o uso de novos dados na generalização do sistema de RAF proposto.

O modelo proposto permite classificar um grande conjunto de fonemas mediante a divisão do espaço de decisão. Esta distribuição de espaço é feita usando um conjunto

de sistemas especialistas, de modo que cada sistema possa avaliar independentemente as características de entrada. Cada especialista gera um resultado que é ponderado por meio da observação probabilística do espaço de decisão pela rede de passagem. Desse modo, a classe fonética com maior probabilidade de saída é potencializada com relação às demais saídas. Quando o modelo de misturas escolhe o especialista ganhador, e portanto, a classe ganhadora, é feita uma avaliação das quatro saídas anteriores, utilizando uma rede neural recorrente que determina a classificação final de acordo com o comportamento temporal das saídas consideradas.

Foram projetados dois modelos de mistura de especialistas. O primeiro, conhecido como MEF, considera a distribuição de especialistas usando as características frequenciais da entrada. O segundo modelo, conhecido como MEHC, além de considerar as características físicas da entrada, também leva em conta o desequilíbrio de classes no conjunto de treinamento. Isto permite resolver alguns dos problemas devidos a os dados desbalanceados, tais como, o sub-treinamento e o sobre-treinamento de classes.

Para a configuração topológica de cada rede especialista, foram testadas 45 hipóteses diferentes de configuração de camadas e número de neurônios. Cada hipótese foi testada 30 vezes para obter um bom nível de variância. Desse modo, mediante observações estatísticas e comparações de hipóteses, foi possível escolher a topologia que proporciona o menor erro de validação.

Métricas de avaliação foram usadas para avaliar os dois modelos de mistura de especialistas propostos, assim como o aporte do pós-processamento dinâmico no desempenho do sistema de RAF. Cada modelo foi testado com e sem o pós-processamento dinâmico, usando dois conjuntos de teste diferentes. Os resultados obtidos foram analisados considerando a taxa de sensibilidade, a precisão na classificação de cada classe fonética, a acurácia geral do modelo e a taxa de erro fonético.

Os resultados obtidos demostraram um desempenho superior do modelo de misturas MEHC em comparação com MEF. Este resultado é justificado pela alta sensibilidade e a melhor precisão na classificação fonética. Nota-se que as classes que foram sobretreinadas pelo MEF, tais como /t/ e /u∼/, e aqueles fonemas com baixa representatividade e tendo um comportamento difuso, tais como /a/ e /o/, tiveram melhorias no seus níveis de precisão e sensibilidade.

Durante as experiências realizadas que incluíram o pós-processamento dinâmico, foi observada a mitigação do comportamento oscilatório gerado pela classificação estática dos modelos de mistura, onde a classe de saída variava entre etiquetas de fonemas espectralmente similares. Esta alteração na saída afetou diretamente a taxa de erro fonético e a sensibilidade na classificação de cada classe fonética, melhorando assim o desempenho geral do sistema de reconhecimento de fala contínua.

Os métodos propostos apresentam restrições relacionadas ao conjunto de treinamento. A disposição hierárquica dos especialistas depende principalmente da análise de distribuição das classes para o conjunto de treinamento considerado. A disposição dos especialistas e o sistema hierárquico devem ser planejados de novo, caso o conjunto de treinamento seja diferente do conjunto analisado nesta dissertação. Ademais, a transcrição de fala para texto feita pelo sistema de RAF proposto é restrita às representações fonéticas e não aos grafemas (representações alfabéticas).

Os resultados de acurácia do modelo MEHC com pós-processamento dinâmico conforme implementado, foram comparados com aqueles apresentados em (MESSAOUD; HAMIDA, 2010), (TÓTH, 2011), (UMARANI; RAVIRAM; WAHIDABANU, 2011) e (CLARKSON; MORENO, 1999). Naqueles trabalhos de reconhecimento fonético com independência do falante, foi usado o Corpora TIMIT, abrangendo 39 classes fonéticas correspondentes à língua Inglesa. Tanto no caso desta dissertação quanto no caso dos trabalhos comparados, foi usado o método MFCC. No entanto no caso do trabalho (UMARANI; RAVIRAM; WAHIDABANU, 2011), foi usado o método MLP. Nota-se que nestes trabalhos usam diferentes técnicas de classificação: desde métodos probabilísticos, tal como HMM até modelos de classificação baseados em aprendizagem de padrões. O sistema proposto proporciona uma melhoria de $7,62\%$ em termos de acurácia. Porém, esta diferença em termos de acurácia pode ter sido ocasionada pelos fatos que os idiomas são diferentes (Português *vs.* Inglês) e os números de classes fonéticas são diferentes (25 *vs.* 35) pela diferença em tamanho do conjunto de classes do MEHC desenhado para o reconhecimento de 25 classes e para uma linguagem diferente. Mesmo assim, vale enfatizar que a taxa de acurácia do sistema de RAF proposto supera o desempenho esperado para um problema de classificação complexo, tal como a classificação fonética.

7.2 Trabalhos Futuros

Nesta seção, são citadas algumas possíveis melhorias na metodologia proposta com intuito de melhorar ainda mais o desempenho do sistema de RAF.

É incentivada a realização de novos ensaios para avaliar diferentes aspectos do modelo proposto, tais como a adaptação de um modelo de linguagem que vincule o reconhecimento acústico com o vocabulário presente na língua Portuguesa assim como a embarcação deste modelo em um sistema de transcrição de palavras *online*.

Também é incentivada a implementação de uma topologia dinâmica na criação do especialista em fonemas plosivos que avalie o tempo de silêncio prévio à explosão de energia de fala. Nessa circunstância é possível que seja necessário incluir um novo conjunto de fonemas que diferenciem as plosivas localizadas no início das palavras daquelas que aparecem dentro da palavra. Esta implementação melhoraria o reconhecimento de fonemas plosivos que, no sistema propostos, são as vezes classificados como pausas.

As redes neurais profundas têm tido sucesso na classificação de padrões. Porém o treinamento deste tipo de redes implica no uso de grandes quantidades de memória RAM e requer velocidade de processamento, exigindo ferramentas computacionalmente superiores ou dedicadas que possam treinar as redes de modo eficiente. Entretanto, uma das vantagens dos modelos de mistura de especialistas e modelos hierárquicos propostos é a distribuição e simplificação das tarefas de classificação. Em conjunto com o uso de redes profundas, isto pode melhorar o tempo de treinamento e o uso de recursos computacionais em comparação com o treinamento de uma grande rede neural profunda. Assim, para melhorar a acurácia de cada sistema especialista é recomendado usar uma rede neural profunda com um pré-treinamento, que possa modelar de modo mais preciso a classificação da fala no espaço de decisão.

A exploração de um modelo de linguagem permitirá a consolidação de um conjunto de fonemas para gerar uma palavra, mediante a detecção de pausas que indiquem o início e final de palavras assim como a consideração dos diferentes sotaques que podem levar modelos fonéticos correspondentes a uma mesma palavra. Isto gerará uma aplicação mais consolidada de reconhecimento *online* de fala, onde não só impactará aplicações de transcrições fonéticas e detecção de problemas de dicção, mas também aplicações comerciais, tais como sistemas de geração de subtítulos automática.

REFERÊNCIAS

ABDEL-HAMID, O. et al. Convolutional neural networks for speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, IEEE, v. 22, n. 10, p. 1533–1545, 2014.

ASSOCIATION, I. P. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Trumington Street, Cambridge, United Kingdom: Cambridge University Press, 1999.

BABER, C. *Speech technology in control room systems: a human factors perspective*. [S.l.]: Ellis Horwood Limited, 1991. (Ellis Horwood series in information technology).

BARANDELA, R. et al. The imbalanced training sample problem: Under or over sampling? In: SPRINGER. *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Berlin, 2004. p. 806–814.

BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, ACM, v. 6, n. 1, p. 20–29, 2004.

BISOL, L. *Introdução a estudos de fonologia do português brasileiro*. Rio Grande do Sul, Brasil: Edipucrs, 2005.

BONILLA, D. A.; NEDJAH, N.; MOURELLE, L. d. M. Reconhecimento automático de fala em português usando redes neurais artificiais profundas. In: ABRICOM. *XII Congresso Brasileiro De Inteligência Computacional CBIC 2015*. Curitiba, Brasil, 2015.

BORGES, M. V. Fonética e fonologia do português: roteiro de estudos e guia de exercícios. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, SciELO Brasil, v. 16, n. 1, p. 183–188, 2000.

BRAGA, A. d. P.; CARVALHO, A.; LUDERMIR, T. B. *Redes neurais artificiais: teoria e aplicações*. Travessa do Ouvidor, 11 Rio de Janeiro, Brasil: Livros Técnicos e Científicos, 2000.

CASTELLVÍ, M. T. C. i. *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos*. Girona, España: Documenta Universitária, 2010.

CECCARELLI, M.; HOUNSOU, J. T. Sequence recognition with radial basis function networks: experiments with spoken digits. *Neurocomputing*, Elsevier, v. 11, n. 1, p. 75–88, 1996.

CHEN, W.-Y.; CHEN, S.-H.; LIN, C.-J. A speech recognition method based on the sequential multi-layer perceptrons. *Neural Networks*, Elsevier, v. 9, n. 4, p. 655–669, 1996.

CHURCHLAND, P.; SEJNOWSKI, T. *Computational neuroscience. The computational brain*. One Rogers Street Cambridge MA, USA: Cambridge, MA: The MIT Press, 1992.

CHYI, Y. Classification analysis techniques for skewed class distribution problems. *Department of Information Management, National Sun Yat-Sen University*, 2003.

CLARKSON, P.; MORENO, P. J. On the use of support vector machines for phonetic classification. In: IEEE. *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. Civic Plaza, Hyatt Regency - Phoenix, Arizona, USA, 1999. v. 2, p. 585–588.

COSI, P.; BENGIO, Y.; MORI, R. D. Phonetically-based multi-layered neural networks for vowel classification. *Speech Communication*, Elsevier, v. 9, n. 1, p. 15–29, 1990.

CUTAJAR, M. et al. Comparative study of automatic speech recognition techniques. *Signal Processing, IET*, IET, v. 7, n. 1, p. 25–46, 2013.

DAMPER, R. I. Speech technology?implications for biomedical engineering. *Journal of medical engineering & technology*, Taylor & Francis, v. 6, n. 4, p. 135–149, 1982.

DANIEL, J.; JAMES, H. Speech and language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition, 2nd Ed., Prentice Hall*, 2009.

DELCROIX, M.; NAKATANI, T.; WATANABE, S. Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing. *Audio, Speech, and Language Processing, IEEE Transactions on*, IEEE, v. 17, n. 2, p. 324–334, 2009.

ELMAN, J. L. Finding structure in time. *Cognitive science*, Elsevier, v. 14, n. 2, p. 179–211, 1990.

FARUQUE, S. Pulse code modulation (pcm). In: *Radio Frequency Source Coding Made Easy*. New York City,USA: Springer, 2015. p. 65–90.

FRATTALI, C.; ASSOCIATION, A. S.-L.-H. et al. *Functional assessment of communication skills for adults (ASHA FACS)*. 2200 Research Boulevard Rockville, USA: ASHA, 1995.

GIBSON, J. D. *Mobile communications handbook*. Boca Raton, Florida, USA: CRC press, 2012.

GRAMSS, T.; STRUBE, H. W. Recognition of isolated words based on psychoacoustics and neurobiology. *Speech Communication*, Elsevier, v. 9, n. 1, p. 35–40, 1990.

HAGAN, M. T. et al. *Neural network design*. Statler Office Bldg, 20 Park Plaza Boston, MA 02116 USA: PWS publishing company Boston, 2014.

HANES, M. D.; AHALT, S. C.; KRISHNAMURTHY, A. K. Acoustic-to-phonetic mapping using recurrent neural networks. *Neural Networks, IEEE Transactions on*, IEEE, v. 5, n. 4, p. 659–662, 1994.

HAYKIN, S. S. *Redes neurais*. Porto Alegre, Brasil: Bookman, 2001.

HAYKIN, S. S. et al. *Neural networks and learning machines*. 221 River St, Hoboken, NJ 07030, USA: Pearson Education Upper Saddle River, 2009.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 21, n. 9, p. 1263–1284, 2009.

HIERONYMUS, J. L. Ascii phonetic symbols for the world's languages: Worldbet. *Journal of the International Phonetic Association*, v. 23, n. 3, p. 12–54, 1993.

HUANG, X. et al. *Spoken language processing: A guide to theory, algorithm, and system development*. Upper Saddle River, N.J,USA: Prentice Hall PTR, 2001.

HULSE, J. V.; KHOSHGOFTAAR, T. Knowledge discovery from imbalanced and noisy data. *Data & Knowledge Engineering*, Elsevier, v. 68, n. 12, p. 1513–1542, 2009.

JACOBS, R. A.; JORDAN, M. I.; BARTO, A. G. Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, Elsevier, v. 15, n. 2, p. 219–250, 1991.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. *Intelligent data analysis*, IOS Press, v. 6, n. 5, p. 429–449, 2002.

JIMENEZ-MARQUEZ, S.; LACROIX, C.; THIBAULT, J. Statistical data validation methods for large cheese plant database. *Journal of dairy science*, Elsevier, v. 85, n. 9, p. 2081–2097, 2002.

JORDAN, M. I.; JACOBS, R. A. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, MIT Press, v. 6, n. 2, p. 181–214, 1994.

KANDEL, E. R. et al. *Principles of neural science*. New York, USA: McGraw-hill New York, 2000.

KEHTARNAVAZ, N.; PARRIS, S.; SEHGAL, A. *Smartphone-Based Real-Time Digital Signal Processing*. 1210 Fifth Ave. Ste 250, San Rafael, CA, USA: Morgan & Claypool Publishers, 2015.

KITTLER, J. et al. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 20, n. 3, p. 226–239, 1998.

KNAGENHJELM, P.; BRAUER, P. Classification of vowels in continuous speech using mlp and a hybrid net. *Speech Communication*, Elsevier, v. 9, n. 1, p. 31–34, 1990.

KOHONEN, T.; SOMERVUO, P. Self-organizing maps of symbol strings. *Neurocomputing*, Elsevier, v. 21, n. 1, p. 19–30, 1998.

KONG, S.-G.; KOSKO, B. Differential competitive learning for centroid estimation and phoneme recognition. *Neural Networks, IEEE Transactions on*, IEEE, v. 2, n. 1, p. 118–124, 1991.

LIPPMANN, R. An introduction to computing with neural nets. *IEEE Assp magazine*, IEEE, v. 4, n. 2, p. 4–22, 1987.

LOPES, C.; PERDIGÃO, F. Broad phonetic class definition driven by phone confusions. *EURASIP Journal on Advances in Signal Processing*, Springer, v. 2012, n. 1, p. 1–12, 2012.

MAIR, C.; HUNDT, M. *Corpus linguistics and linguistic theory*. Amsterdam, Netherlands: Rodopi, 2000.

MARIANI, J. *Language and Speech Processing*. 111 River Street Hoboken, NJ,USA: John Wiley & Sons, 2013.

MARTIN, R.; HEUTE, U.; ANTWEILER, C. *Advances in digital speech transmission*. 111 River Street Hoboken, NJ,USA: John Wiley & Sons, 2008.

MASOUDNIA, S.; EBRAHIMPOUR, R. Mixture of experts: a literature survey. *Artificial Intelligence Review*, Springer, v. 42, n. 2, p. 275–293, 2014.

MASTERS, T. *Practical neural network recipes in C++*. Burlington, Massachusetts, USA: Morgan Kaufmann, 1993.

MATHWORKS, I. *SIMULINK, Simulation and Model-Based Design*. 2016.

MATLAB, S. *Sample Times in Systems*. 2016. Acessado em: 2016-05-30. Disponível em: <http://www.mathworks.com/help/simulink/ug/managing-sample-times-in-systems.html>.

MENDEL, J. M.; FU, K. S. et al. *Adaptive, learning, and pattern recognition systems: theory and applications*. New York, USA: Academic Press New York, 1970.

MESSAOUD, Z. B.; HAMIDA, A. B. Cdhmm parameters selection for speaker-independent phone recognition in continuous speech system. In: IEEE. *Melecon 2010-2010 15th IEEE Mediterranean Electrotechnical Conference*. Grand Hotel Excelsior, Valletta, Malta, 2010. p. 253–258.

MPORAS, I. et al. Comparison of speech features on the speech recognition task. *Journal of Computer Science*, Citeseer, v. 3, n. 8, p. 608–616, 2007.

OPPENHEIM, A. V. et al. *Discrete-time signal processing*. Englewood Cliffs, NJ, USA: Prentice hall, 1989.

OSHAUGHNESSY, D. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, Elsevier, v. 41, n. 10, p. 2965–2979, 2008.

PETERSON, G. E.; BARNEY, H. L. Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, Acoustical Society of America, v. 24, n. 2, p. 175–184, 1952.

PRINCIPI, E. et al. An integrated system for voice command recognition and emergency detection based on audio signals. *Expert Systems with Applications*, Elsevier, v. 42, n. 13, p. 5668–5683, 2015.

RABINER, L.; JUANG, B.-H. Fundamentals of speech recognition. Prentice hall, 1993.

RABINER, L. R.; JUANG, B.-H. *Fundamentals of speech recognition*. ?Beijing, China: Tsinghua University Press, 1999.

RASMUSSEN, C. E. *Evaluation of Gaussian processes and other methods for non-linear regression*. Tese (Doutorado) — University of Toronto, 1996.

RENCHER, A. C. *Methods of multivariate analysis*. 111 River Street Hoboken, NJ,USA: John Wiley & Sons, 2003.

RIEDMILLER, M.; BRAUN, H. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In: IEEE. *Neural Networks, 1993., IEEE International Conference On*. San Francisco, CA, USA, 1993. p. 586–591.

SAHA, G.; CHAKROBORTY, S.; SENAPATI, S. A new silence removal and endpoint detection algorithm for speech and speaker recognition applications. In: IIT. *Proceedings of the 11th National Conference on Communications (NCC)*. New Delhi,India, 2005. p. 291–295.

SCANLON, P.; ELLIS, D. P.; REILLY, R. B. Using broad phonetic group experts for improved speech recognition. *IEEE transactions on audio, speech, and language processing*, IEEE, v. 15, n. 3, p. 803–812, 2007.

SCHRAMM, M. et al. *Spoltech Brazilian Portuguese Version 1.0 LDC2006S16. Web Download*. Linguistic Data Consortium, 2006. Acessado em: 15-06-2016. Disponível em: <https://catalog.ldc.upenn.edu/LDC2006S16>.

SCHRAMM, M. C. et al. A Brazilian Portuguese language corpus development. In: ISCA. *Sixth International Conference on Spoken Language Processing*. Beijing, China, 2000.

SHANNON, C. E. Communication in the presence of noise. *Proceedings of the IRE*, IEEE, v. 37, n. 1, p. 10–21, 1949.

SILVA, I. d.; SPATTI, D. H.; FLAUZINO, R. A. *Redes neurais artificiais para engenharia e ciências aplicadas*. Av. Diógenes Ribeiro de Lima, 3294 - 05083-010 São Paulo, SP - Brasil: Artliber, 2010.

SILVA, T. C.; YEHIA, H. C. Sonoridade em artes, saúde e tecnologia. *Revista Docência do Ensino Superior*, v. 1, n. 15, p. 62–74, 2015.

SINISCALCHI, S. M. et al. Exploiting deep neural networks for detection-based speech recognition. *Neurocomputing*, Elsevier, v. 106, n. 6, p. 148–157, 2013.

SOLERA-UREÑA, R. et al. Robust asr using support vector machines. *Speech Communication*, Elsevier, v. 49, n. 4, p. 253–267, 2007.

SOUSA, E. d. Para a caracterização fonético-acústica da nasalidade no português do brasil. *Master. Universidade Estadual de Campinas Campinas*, 1994.

TORRE, D. *Segmentación y Etiquetado Fonéticos Automáticos. Un enfoque Basado en Modelos Ocultos de Markov y Refinamiento Posterior de las Fronteras Fonéticas. Resumen de tesis doctoral. Universidad Politécnica de Madrid*. 2012.

TÓTH, L. A hierarchical, context-dependent neural network architecture for improved phone recognition. In: IEEE. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Prague Congress Centre Prague, Czech Republic, 2011. p. 5040–5043.

TWYCROSS, J. P. *Integrated innate and adaptive artificial immune systems applied to process anomaly detection*. Nottingham, UK: University of Nottingham, 2007.

UMARANI, S.; RAVIRAM, P.; WAHIDABANU, R. Implementation of hmm and radial basis function for speech recognition. In: IEEE. *Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009. International Conference on*. TBD Chennai, India, 2011. p. 1–4.

VARGA, I. et al. Asr in mobile phones-an industrial approach. *Speech and Audio Processing, IEEE Transactions on*, IEEE, v. 10, n. 8, p. 562–569, 2002.

VIRTANEN, T.; SINGH, R.; RAJ, B. *Techniques for noise robustness in automatic speech recognition*. 111 River Street Hoboken, NJ,USA: John Wiley & Sons, 2012.

WACHTER, M. D. et al. Template-based continuous speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 15, n. 4, p. 1377–1390, 2007.

WANG, D. et al. Embedded speech recognition system on 8-bit mcu core. In: IEEE. *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. Fairmont Queen Elizabeth Hotel in Montreal, Quebec, Canada, 2004. v. 5, p. V–301.

WANG, W.-J.; LIAO, Y.-F.; CHEN, S.-H. Rnn-based prosodic modeling for mandarin speech and its application to speech-to-text conversion. *Speech Communication*, Elsevier, v. 36, n. 3, p. 247–265, 2002.

WATERHOUSE, S. R. *Classification and regression using mixtures of experts*. Tese (Doutorado) — Departament of Engineering, University of Cambridge, 1998.

WEISS, G. M. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, ACM, v. 6, n. 1, p. 7–19, 2004.

WOODLAND, P. C.; SMYTH, S. An experimental comparison of connectionist and conventional classification systems on natural data. *Speech Communication*, Elsevier, v. 9, n. 1, p. 73–82, 1990.

ZAHORIAN, S. A.; NOSSAIR, Z. B. A partitioned neural network approach for vowel classification using smoothed time/frequency features. *Speech and Audio Processing, IEEE Transactions on*, IEEE, v. 7, n. 4, p. 414–425, 1999.

ZAMANI, B. et al. Optimized discriminative transformations for speech features based on minimum classification error. *Pattern Recognition Letters*, Elsevier, v. 32, n. 7, p. 948–955, 2011.

ZHANG, Y.; TOGNERI, R.; ALDER, M. Phoneme-based vector quantization in a discrete hmm speech recognizer. *IEEE Transactions on speech and audio processing*, IEEE, v. 5, n. 1, p. 26–32, 1997.

APÊNDICE A – Seleção dos modelos topológicos para cada especialista

Este apêndice apresenta os testes feitos para a seleção do modelo topológico de cada sistema especialista, incluindo os erros de validação gerados em cada iteração de treinamento, a média e o desvio padrão de cada teste e as comparações de significância entre hipóteses mediante o teste T. Estes dados e testes são relacionados com as técnicas de seleção topológica apresentadas na Seção 5.3.2 do Capitulo 5.

A.1 Especialistas do modelo MEF e MEHC

Nesta seção são apresentados os resultados do treinamento de um conjunto de hipóteses, caracterizadas pela distribuição de diferentes números de neurônios entre as camadas escondidas. Cada hipótese foi testada 30 vezes com os mesmos conjuntos de treinamento e validação. Esta operação é feita para decidir a configuração da quantidade de neurônios presentes para cada sistema especialista projetado no Capitulo 5, tanto para o modelo MEF como para o modelo MEHC.

Cada uma das Figuras apresentadas descreve o erro de validação médio e o desvio padrão para cada hipótese. Cada barra representa a quantidade de neurônios na primeira camada oculta e cada agrupação de barras está relacionada com a número de neurônios presentes na segunda camada. De acordo com o menor erro de validação apresentado nos testes, é selecionada a hipótese com menor erro de validação e com um desvio padrão aceitável.

Por último, é escolhida outra hipótese com configuração mais simples e com uma boa taxa de erro de validação. Desse modo, compara-se a semelhança das hipóteses mediante o uso de um teste T. No caso em que o nível de significância seja mínimo, a hipótese mais simples é aceitada. Caso contrário, é implementada a rede com menor erro de validação.

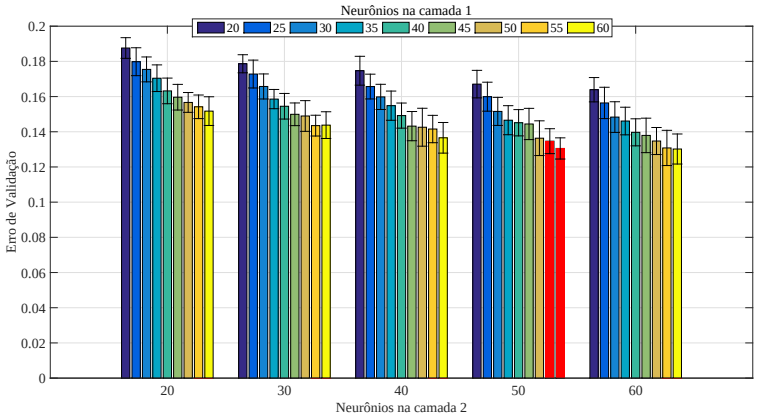


Figura 54: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em vogais.

Tabela 15: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em vogais

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	20	20	0,187598126560199	0,00586959553272975
2	25	20	0,179813301868354	0,00792870279234380
3	30	20	0,175445283867983	0,00707042241297064
4	35	20	0,170434434999417	0,00757108476516767
5	40	20	0,163219252754611	0,00728128118368426
6	45	20	0,159641930740510	0,00729030145427895
7	50	20	0,156667575526145	0,00563188190530507
8	55	20	0,154184021904824	0,00670146480199368
9	60	20	0,151732793820778	0,00815635803386493
10	20	30	0,178646719824464	0,00513456745323153
11	25	30	0,172804690436240	0,00791641073340926
12	30	30	0,165745416259402	0,00711813074758733
13	35	30	0,158573044478849	0,00547006674653105
14	40	30	0,154498671250158	0,00727865541894470
15	45	30	0,149926510250817	0,00644570984564476
16	50	30	0,148962613688536	0,00872388062573293
17	55	30	0,143487657703374	0,00590398766843664
18	60	30	0,143773940929457	0,00759456302416384
19	20	40	0,174688490827195	0,00819845993282337
20	25	40	0,165690610971020	0,00703997103241441
21	30	40	0,159819823792172	0,00715249969273991
22	35	40	0,154828840567573	0,00829977098782787
23	40	40	0,149189118074979	0,00713551396585505
24	45	40	0,143205883198569	0,00829988630983798
25	50	40	0,142580023572059	0,0107747943902211
26	55	40	0,141554431665126	0,00777285058352381
27	60	40	0,136568591596720	0,00870328778365973
28	20	50	0,167082307623703	0,00781865998670110
29	25	50	0,159933055789242	0,00823435261944008
30	30	50	0,151572320888396	0,00797029020031831
31	35	50	0,146572970189783	0,00825145298087227
32	40	50	0,145126763635046	0,00746572869038420
33	45	50	0,144428414673952	0,00887982412267029
34	50	50	0,136348905404997	0,00986705743652994
35	55	50	0,134650834248570	0,00708180573170037
36	60	50	0,130521058961092	0,00602092287392288
37	20	60	0,163898770442421	0,00690522838672247
38	25	60	0,156381024826426	0,00890807905605346
39	30	60	0,148349082649881	0,00870890592205313
40	35	60	0,146102025139114	0,00784530304594407
41	40	60	0,139683449900716	0,00768507725324002
42	45	60	0,137929081463738	0,00979108371111687
43	50	60	0,134742741023593	0,00768185248847104
44	55	60	0,130810800213740	0,0100091548413671
45	60	60	0,130198886798008	0,00853860370367860

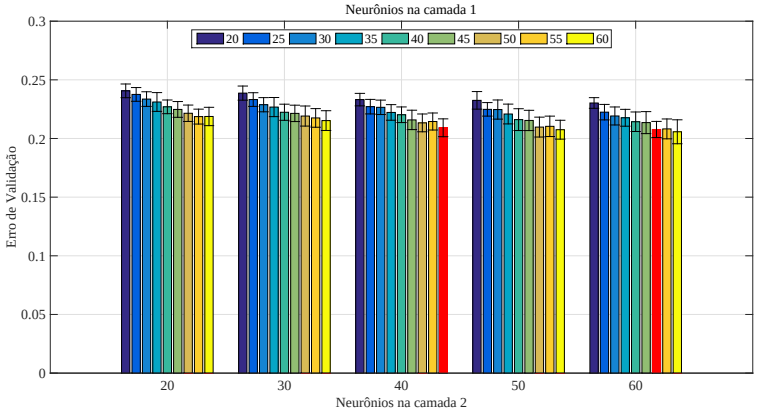


Figura 55: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em fricativas.

Tabela 16: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em fricativas

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	20	20	0,252620919	0,252620919
2	25	20	0,236824481	0,236824481
3	30	20	0,235940617	0,235940617
4	35	20	0,226351119	0,226351119
5	40	20	0,225412977	0,225412977
6	45	20	0,221355005	0,221355005
7	50	20	0,213410431	0,213410431
8	55	20	0,213374869	0,213374869
9	60	20	0,21508923	0,21508923
10	20	30	0,246130518	0,246130518
11	25	30	0,233451927	0,233451927
12	30	30	0,227397964	0,227397964
13	35	30	0,220339687	0,220339687
14	40	30	0,213606132	0,213606132
15	45	30	0,214911956	0,214911956
16	50	30	0,208951183	0,208951183
17	55	30	0,208234764	0,208234764
18	60	30	0,208774663	0,208774663
19	20	40	0,240187201	0,240187201
20	25	40	0,231431246	0,231431246
21	30	40	0,223543119	0,223543119
22	35	40	0,215105159	0,215105159
23	40	40	0,210021906	0,210021906
24	45	40	0,213868266	0,213868266
25	50	40	0,196242385	0,196242385
26	55	40	0,203656139	0,203656139
27	60	40	0,19773845	0,19773845
28	20	50	0,237705846	0,237705846
29	25	50	0,226025252	0,226025252
30	30	50	0,214916015	0,214916015
31	35	50	0,217409094	0,217409094
32	40	50	0,218473188	0,218473188
33	45	50	0,208405021	0,208405021
34	50	50	0,202042845	0,202042845
35	55	50	0,200532859	0,200532859
36	60	50	0,194392632	0,194392632
37	20	60	0,227204848	0,227204848
38	25	60	0,222867435	0,222867435
39	30	60	0,216075014	0,216075014
40	35	60	0,211681133	0,211681133
41	40	60	0,200502339	0,200502339
42	45	60	0,20029579	0,20029579
43	50	60	0,199050808	0,199050808
44	55	60	0,185440702	0,185440702
45	60	60	0,188451631	0,188451631

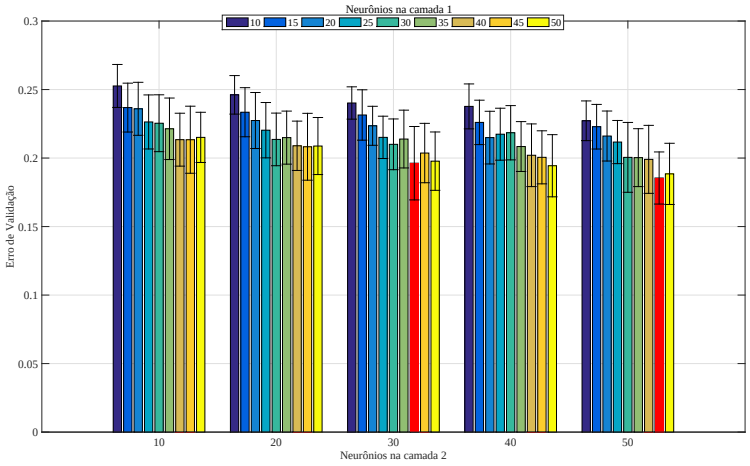


Figura 56: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em fonemas plosivos.

Tabela 17: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em plosivos

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,252620919	0,252620919
2	15	10	0,236824481	0,236824481
3	20	10	0,235940617	0,235940617
4	25	10	0,226351119	0,226351119
5	30	10	0,225412977	0,225412977
6	35	10	0,221355005	0,221355005
7	40	10	0,213410431	0,213410431
8	45	10	0,213374869	0,213374869
9	50	10	0,21508923	0,21508923
10	10	20	0,246130518	0,246130518
11	15	20	0,233451927	0,233451927
12	20	20	0,227397964	0,227397964
13	25	20	0,220339687	0,220339687
14	30	20	0,213606132	0,213606132
15	35	20	0,214911956	0,214911956
16	40	20	0,208951183	0,208951183
17	45	20	0,208234764	0,208234764
18	50	20	0,208774663	0,208774663
19	10	30	0,240187201	0,240187201
20	15	30	0,231431246	0,231431246
21	20	30	0,223543119	0,223543119
22	25	30	0,215105159	0,215105159
23	30	30	0,210021906	0,210021906
24	35	30	0,213868266	0,213868266
25	40	30	0,196242385	0,196242385
26	45	30	0,203656139	0,203656139
27	50	30	0,19773845	0,19773845
28	10	40	0,237705846	0,237705846
29	15	40	0,226025252	0,226025252
30	20	40	0,214916015	0,214916015
31	25	40	0,217409094	0,217409094
32	30	40	0,218473188	0,218473188
33	35	40	0,208405021	0,208405021
34	40	40	0,202042845	0,202042845
35	45	40	0,200532859	0,200532859
36	50	40	0,194392632	0,194392632
37	10	50	0,227204848	0,227204848
38	15	50	0,222867435	0,222867435
39	20	50	0,216075014	0,216075014
40	25	50	0,211681133	0,211681133
41	30	50	0,200502339	0,200502339
42	35	50	0,20029579	0,20029579
43	40	50	0,199050808	0,199050808
44	45	50	0,185440702	0,185440702
45	50	50	0,188451631	0,188451631

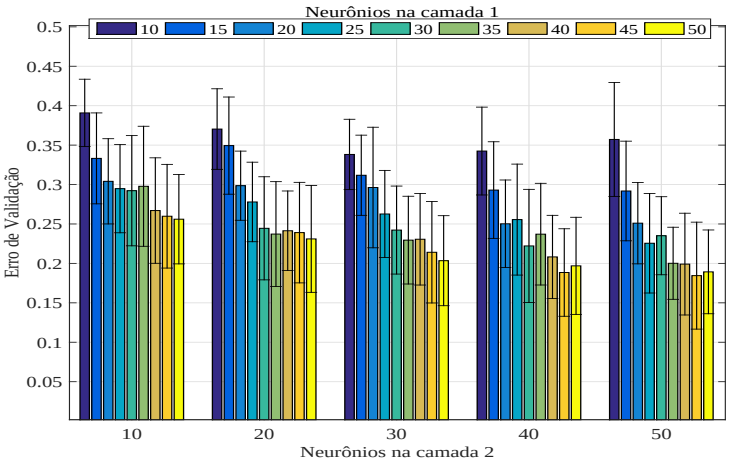


Figura 57: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista dos fonemas Nasais.

Tabela 18: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em fonemas nasais

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,391	0,002
2	15	10	0,333	0,003
3	20	10	0,304	0,003
4	25	10	0,295	0,003
5	30	10	0,292	0,005
6	35	10	0,298	0,006
7	40	10	0,267	0,004
8	45	10	0,260	0,004
9	50	10	0,256	0,003
10	10	20	0,370	0,003
11	15	20	0,349	0,004
12	20	20	0,299	0,002
13	25	20	0,278	0,003
14	30	20	0,245	0,004
15	35	20	0,237	0,004
16	40	20	0,241	0,003
17	45	20	0,239	0,004
18	50	20	0,231	0,005
19	10	30	0,338	0,002
20	15	30	0,312	0,003
21	20	30	0,296	0,006
22	25	30	0,263	0,003
23	30	30	0,242	0,003
24	35	30	0,230	0,003
25	40	30	0,231	0,003
26	45	30	0,214	0,004
27	50	30	0,203	0,003
28	10	40	0,342	0,003
29	15	40	0,293	0,004
30	20	40	0,250	0,003
31	25	40	0,256	0,005
32	30	40	0,222	0,005
33	35	40	0,237	0,004
34	40	40	0,208	0,003
35	45	40	0,188	0,003
36	50	40	0,197	0,004
37	10	50	0,357	0,005
38	15	50	0,292	0,004
39	20	50	0,251	0,003
40	25	50	0,226	0,004
41	30	50	0,235	0,002
42	35	50	0,200	0,002
43	40	50	0,199	0,004
44	45	50	0,184	0,005
45	50	50	0,189	0,003

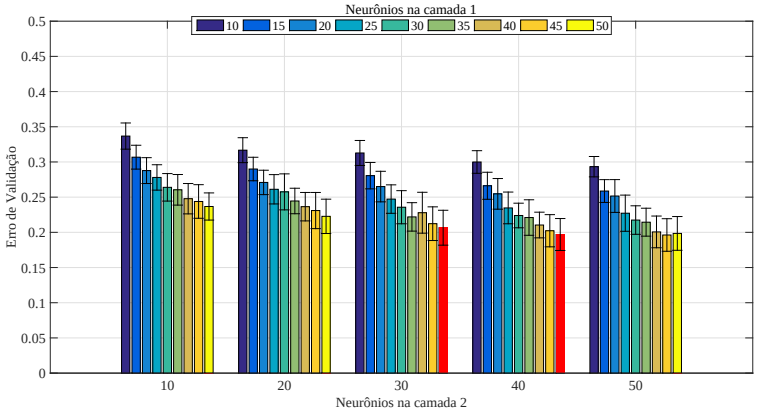


Figura 58: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista dos fonemas vogais nasais.

Tabela 19: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em vogais nasais

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,33678243	0,33678243
2	15	10	0,306792285	0,306792285
3	20	10	0,28772804	0,28772804
4	25	10	0,278005676	0,278005676
5	30	10	0,263951253	0,263951253
6	35	10	0,260429337	0,260429337
7	40	10	0,247793672	0,247793672
8	45	10	0,243855759	0,243855759
9	50	10	0,236720186	0,236720186
10	10	20	0,316744318	0,316744318
11	15	20	0,289976943	0,289976943
12	20	20	0,270920117	0,270920117
13	25	20	0,26118261	0,26118261
14	30	20	0,257508443	0,257508443
15	35	20	0,244544902	0,244544902
16	40	20	0,236457358	0,236457358
17	45	20	0,230967811	0,230967811
18	50	20	0,222744174	0,222744174
19	10	30	0,312771961	0,312771961
20	15	30	0,280540649	0,280540649
21	20	30	0,265036081	0,265036081
22	25	30	0,24726351	0,24726351
23	30	30	0,235700935	0,235700935
24	35	30	0,221949418	0,221949418
25	40	30	0,227870996	0,227870996
26	45	30	0,212299084	0,212299084
27	50	30	0,206564404	0,206564404
28	10	40	0,29995198	0,29995198
29	15	40	0,266167301	0,266167301
30	20	40	0,254677488	0,254677488
31	25	40	0,234698205	0,234698205
32	30	40	0,223942194	0,223942194
33	35	40	0,221037655	0,221037655
34	40	40	0,210404146	0,210404146
35	45	40	0,202280275	0,202280275
36	50	40	0,196924046	0,196924046
37	10	50	0,293291175	0,293291175
38	15	50	0,258671408	0,258671408
39	20	50	0,25162724	0,25162724
40	25	50	0,227198553	0,227198553
41	30	50	0,217460816	0,217460816
42	35	50	0,214427583	0,214427583
43	40	50	0,200624285	0,200624285
44	45	50	0,196241299	0,196241299
45	50	50	0,198471841	0,198471841

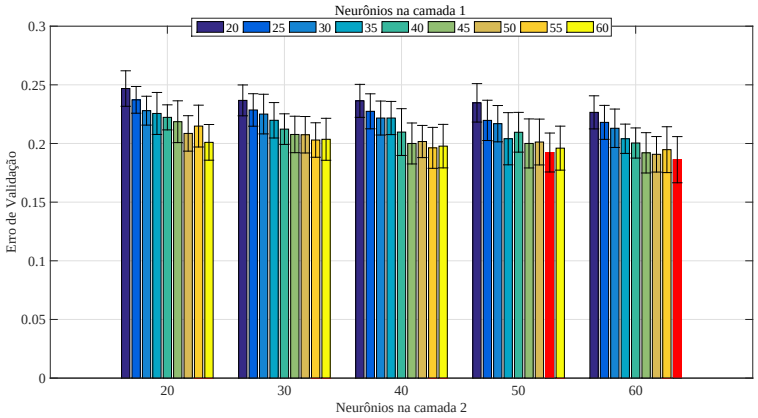


Figura 59: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em ditongos sem atraso.

Tabela 20: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em ditongos sem atraso

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	20	20	0,246845118	0,015130709
2	25	20	0,237212044	0,011392764
3	30	20	0,227977964	0,012302258
4	35	20	0,225617183	0,017888616
5	40	20	0,222278182	0,010677124
6	45	20	0,218564309	0,017817322
7	50	20	0,20858848	0,015109554
8	55	20	0,214843706	0,017829943
9	60	20	0,200975434	0,015215957
10	20	30	0,236758781	0,013155691
11	25	30	0,228584364	0,01384636
12	30	30	0,22508217	0,016892475
13	35	30	0,219790916	0,01509588
14	40	30	0,212256972	0,013042995
15	45	30	0,207704717	0,015538017
16	50	30	0,207456923	0,015539051
17	55	30	0,202966218	0,014696975
18	60	30	0,203585872	0,017893929
19	20	40	0,236347881	0,01406721
20	25	40	0,227434213	0,014915903
21	30	40	0,221702647	0,014475627
22	35	40	0,221721146	0,014147106
23	40	40	0,209732181	0,01994159
24	45	40	0,200023393	0,017457627
25	50	40	0,201685646	0,013711272
26	55	40	0,196290243	0,017440707
27	60	40	0,19775604	0,018542917
28	20	50	0,234656395	0,016383782
29	25	50	0,219707402	0,017165452
30	30	50	0,216896061	0,015475149
31	35	50	0,204065853	0,022237681
32	40	50	0,209577185	0,016948472
33	45	50	0,200048902	0,020895292
34	50	50	0,201257408	0,019535927
35	55	50	0,192315584	0,016594851
36	60	50	0,196037325	0,018748748
37	20	60	0,226555616	0,01409978
38	25	60	0,218009953	0,014479041
39	30	60	0,212988727	0,01636755
40	35	60	0,204035929	0,012487126
41	40	60	0,200417473	0,012852862
42	45	60	0,192048932	0,017183884
43	50	60	0,190783082	0,015139123
44	55	60	0,194738212	0,019544375
45	60	60	0,186199558	0,019675933

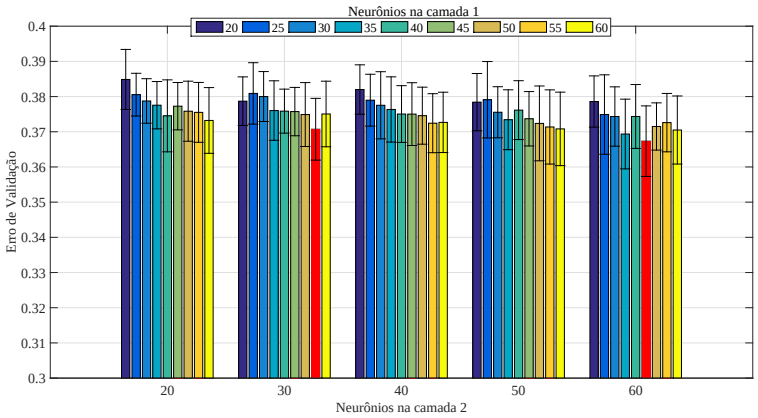


Figura 60: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista em pausas.

Tabela 21: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em pausas

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	20	20	0,384863161661806	0,00853357028263589
2	25	20	0,380545445822502	0,00606889973923540
3	30	20	0,378748795677883	0,00632864201097678
4	35	20	0,377552067638725	0,00672244815454489
5	40	20	0,374524047276326	0,0102254573107892
6	45	20	0,377271844097014	0,00672226329264599
7	50	20	0,375839461006818	0,00853429597138512
8	55	20	0,375510790410375	0,00850787003105729
9	60	20	0,373214636553135	0,00933705861898788
10	20	30	0,378691382273519	0,00690540190076041
11	25	30	0,380899305403448	0,00872207431880557
12	30	30	0,380009041973366	0,00709463414033629
13	35	30	0,376028270513692	0,00845390696630934
14	40	30	0,375860862009278	0,00625277689545508
15	45	30	0,375734946999437	0,00686909573073883
16	50	30	0,374891699687401	0,00908307589703320
17	55	30	0,370715752564622	0,00877327904750807
18	60	30	0,375054096145323	0,00931544003763668
19	20	40	0,382008351974169	0,00701251491254315
20	25	40	0,378972663765701	0,00736905703728069
21	30	40	0,377527108057955	0,00953617191403599
22	35	40	0,376338966745449	0,00926772500108846
23	40	40	0,375038610630309	0,00805987126081969
24	45	40	0,375016466376238	0,00892011516830973
25	50	40	0,374569370231072	0,00810215186155742
26	55	40	0,372436824873098	0,00838187492505210
27	60	40	0,372665767567841	0,00857067032420245
28	20	50	0,378418647189130	0,00813568704129243
29	25	50	0,379099625106677	0,0108538302027344
30	30	50	0,375546552945083	0,00725169682503553
31	35	50	0,373417279386138	0,00850921365137205
32	40	50	0,376145802616375	0,00836396359478797
33	45	50	0,373691398989492	0,00772614259201137
34	50	50	0,372381091739337	0,0106262042048180
35	55	50	0,371361832122860	0,0105436874033968
36	60	50	0,370816094282066	0,0104429815315358
37	20	60	0,378595856978769	0,00727478291089035
38	25	60	0,374893798519901	0,0112851598159271
39	30	60	0,374333760731990	0,00842864458506215
40	35	60	0,369359753040633	0,00990440810226411
41	40	60	0,374347766549674	0,00907099560658213
42	45	60	0,367329677738277	0,0100232616626086
43	50	60	0,371502325356150	0,00670923098226773
44	55	60	0,372595545914370	0,00829411705428834
45	60	60	0,370497100810724	0,00967308030174959

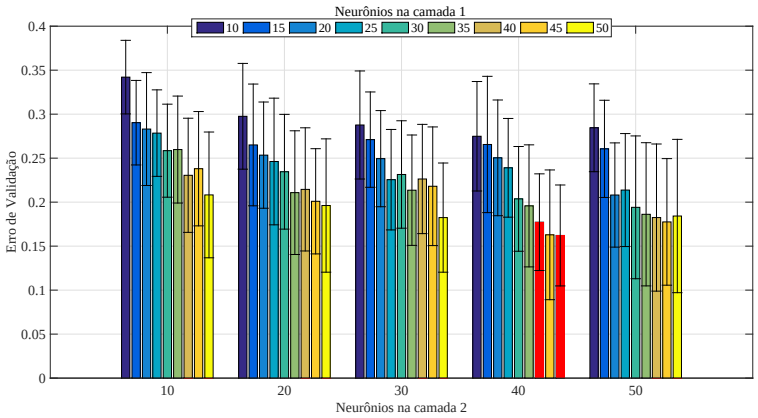


Figura 61: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /o//a/ do modelo MEHC.

Tabela 22: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /o//a/

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,342101720600986	0,0418192663583553
2	15	10	0,290318799783163	0,0480351982085338
3	20	10	0,283083637386568	0,0641559421231223
4	25	10	0,278473444680675	0,0491554105979258
5	30	10	0,258498034159941	0,0528688709660580
6	35	10	0,259781006689338	0,0607860531410984
7	40	10	0,230520357450695	0,0648701642879101
8	45	10	0,238010832248884	0,0649651988927746
9	50	10	0,208237299846726	0,0714694808358658
10	10	20	0,297585038761600	0,0601003463409371
11	15	20	0,265042394289217	0,0691806213498021
12	20	20	0,253443102176646	0,0603670915833524
13	25	20	0,246211470681488	0,0719811284269108
14	30	20	0,234555093475517	0,0652203553930046
15	35	20	0,210801909024355	0,0703022586119771
16	40	20	0,214509660198935	0,0699944037591064
17	45	20	0,200969171322142	0,0597944554333227
18	50	20	0,196162118299300	0,0757668047628796
19	10	30	0,287705254199343	0,0614047522075336
20	15	30	0,271064588047839	0,0541811908471528
21	20	30	0,249428153840253	0,0546357736416687
22	25	30	0,225539118224596	0,0570926319313684
23	30	30	0,231448214039592	0,0610535037099354
24	35	30	0,213589937281228	0,0627294809508230
25	40	30	0,226319694062147	0,0620860421834256
26	45	30	0,218061599556475	0,0674333593906955
27	50	30	0,182435231030971	0,0620460830891156
28	10	40	0,274891165096531	0,0622207192167783
29	15	40	0,265556175113309	0,0774568062647188
30	20	40	0,250350547626869	0,0657608516422805
31	25	40	0,239026625099050	0,0560959726362219
32	30	40	0,203713470273594	0,0595328331548305
33	35	40	0,195777906894679	0,0693543909811617
34	40	40	0,177219650700936	0,0549445267002734
35	45	40	0,162866139164122	0,0737339843445040
36	50	40	0,162147888436076	0,0573580865981845
37	10	50	0,284526831831604	0,0499073885132524
38	15	50	0,260566563432615	0,0552396136281614
39	20	50	0,208117941249673	0,0592359563831904
40	25	50	0,213748835600335	0,0642041924429032
41	30	50	0,194111706366429	0,0811581779818738
42	35	50	0,186196340001486	0,0813755187536274
43	40	50	0,182473552268462	0,0836197602289644
44	45	50	0,177544124574225	0,0719307091576891
45	50	50	0,184233183389810	0,0871292362907194

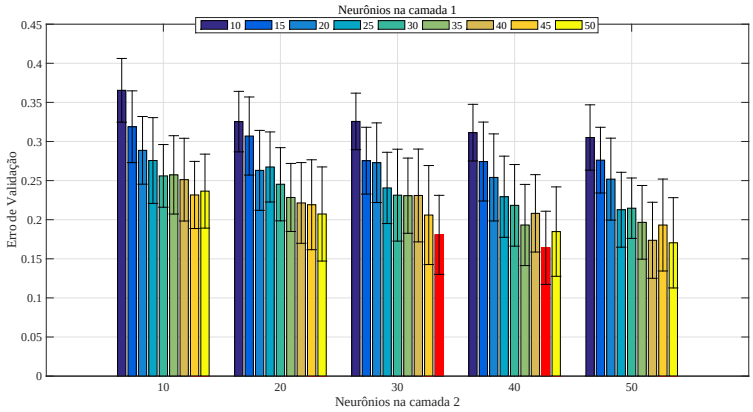


Figura 62: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ẽ/ /e/ do modelo MEHC.

Tabela 23: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ẽ/ /e/

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,365451	0,040729
2	15	10	0,31887	0,045896
3	20	10	0,288689	0,043266
4	25	10	0,275684	0,05484
5	30	10	0,256007	0,040088
6	35	10	0,257322	0,05009
7	40	10	0,251206	0,052896
8	45	10	0,231635	0,042927
9	50	10	0,23654	0,047313
10	10	20	0,32544	0,038675
11	15	20	0,30699	0,049922
12	20	20	0,263093	0,051085
13	25	20	0,267433	0,044783
14	30	20	0,245343	0,046774
15	35	20	0,228447	0,043505
16	40	20	0,221415	0,051613
17	45	20	0,219129	0,057485
18	50	20	0,207298	0,060175
19	10	30	0,325682	0,036122
20	15	30	0,27548	0,042709
21	20	30	0,27287	0,050925
22	25	30	0,240645	0,045564
23	30	30	0,231381	0,058763
24	35	30	0,230703	0,048083
25	40	30	0,23098	0,059328
26	45	30	0,205965	0,063301
27	50	30	0,180655	0,050518
28	10	40	0,311315	0,036227
29	15	40	0,274371	0,050465
30	20	40	0,254064	0,055655
31	25	40	0,229376	0,051885
32	30	40	0,218362	0,052187
33	35	40	0,193236	0,051894
34	40	40	0,208119	0,049478
35	45	40	0,164056	0,046783
36	50	40	0,184791	0,057162
37	10	50	0,305094	0,041794
38	15	50	0,276156	0,042012
39	20	50	0,251872	0,052396
40	25	50	0,212739	0,04788
41	30	50	0,214714	0,038602
42	35	50	0,196647	0,047107
43	40	50	0,173687	0,04861
44	45	50	0,193177	0,058786
45	50	50	0,170446	0,057721

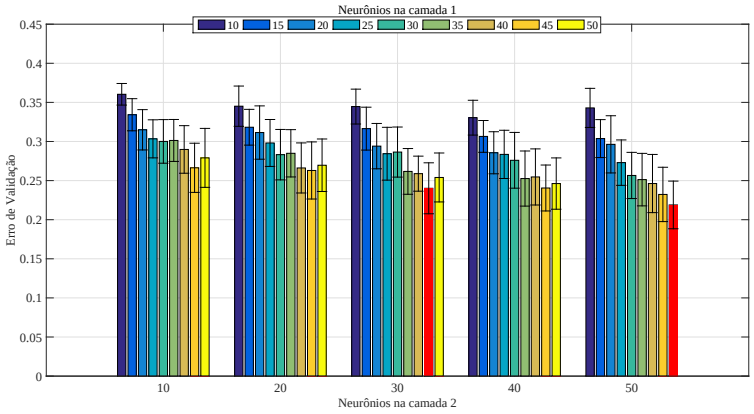


Figura 63: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ĩ/ /i/ do modelo MEHC.

Tabela 24: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ĩ/ /i/

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,360372	0,013809
2	15	10	0,334213	0,020537
3	20	10	0,314958	0,02577
4	25	10	0,3034	0,024325
5	30	10	0,300106	0,027877
6	35	10	0,301335	0,026824
7	40	10	0,289835	0,030418
8	45	10	0,266356	0,031434
9	50	10	0,279057	0,037689
10	10	20	0,345146	0,025752
11	15	20	0,318241	0,022867
12	20	20	0,311442	0,034093
13	25	20	0,298125	0,030003
14	30	20	0,283164	0,032239
15	35	20	0,284908	0,030237
16	40	20	0,266173	0,032009
17	45	20	0,262955	0,036509
18	50	20	0,269639	0,03357
19	10	30	0,344656	0,022296
20	15	30	0,316438	0,027419
21	20	30	0,294079	0,028982
22	25	30	0,284361	0,033773
23	30	30	0,28648	0,03202
24	35	30	0,261749	0,029239
25	40	30	0,258863	0,022409
26	45	30	0,240165	0,032584
27	50	30	0,254006	0,031291
28	10	40	0,330458	0,022257
29	15	40	0,306504	0,020339
30	20	40	0,285595	0,026886
31	25	40	0,283519	0,030836
32	30	40	0,275961	0,035644
33	35	40	0,252618	0,035235
34	40	40	0,254676	0,035874
35	45	40	0,240543	0,029363
36	50	40	0,2462	0,032809
37	10	50	0,342965	0,024999
38	15	50	0,303717	0,024253
39	20	50	0,296363	0,036564
40	25	50	0,272961	0,029035
41	30	50	0,256576	0,029529
42	35	50	0,251251	0,033557
43	40	50	0,246249	0,037198
44	45	50	0,23231	0,034744
45	50	50	0,218884	0,03047

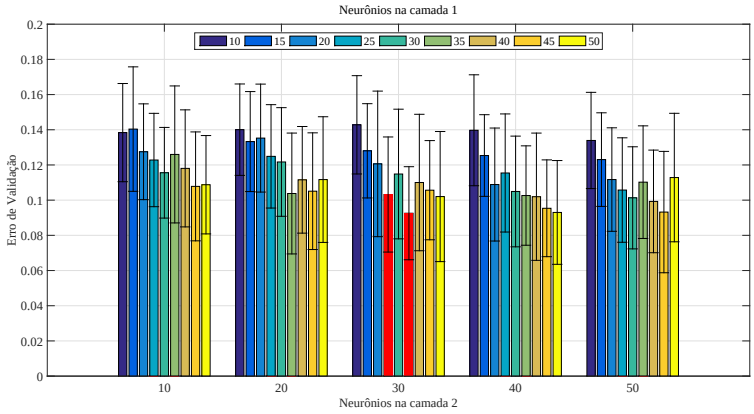


Figura 64: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ɔ/ do modelo MEHC.

Tabela 25: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ɔ/

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,138383	0,027902
2	15	10	0,140386	0,035366
3	20	10	0,127504	0,02722
4	25	10	0,122808	0,026526
5	30	10	0,115571	0,02576
6	35	10	0,125983	0,038926
7	40	10	0,118058	0,033258
8	45	10	0,107827	0,030949
9	50	10	0,108764	0,027944
10	10	20	0,14005	0,025974
11	15	20	0,133286	0,028414
12	20	20	0,135269	0,030678
13	25	20	0,124891	0,029375
14	30	20	0,121696	0,030867
15	35	20	0,10378	0,03434
16	40	20	0,111553	0,030301
17	45	20	0,105094	0,03316
18	50	20	0,111688	0,035709
19	10	30	0,142814	0,027932
20	15	30	0,128038	0,026786
21	20	30	0,12061	0,041363
22	25	30	0,1032	0,032707
23	30	30	0,11485	0,036838
24	35	30	0,092566	0,026442
25	40	30	0,110021	0,038771
26	45	30	0,10566	0,028184
27	50	30	0,102036	0,036967
28	10	40	0,13972	0,031518
29	15	40	0,125358	0,023205
30	20	40	0,108873	0,032106
31	25	40	0,115442	0,033559
32	30	40	0,104931	0,031456
33	35	40	0,102619	0,028242
34	40	40	0,10195	0,036164
35	45	40	0,095378	0,027502
36	50	40	0,093017	0,029489
37	10	50	0,133925	0,027343
38	15	50	0,123042	0,026597
39	20	50	0,111702	0,029408
40	25	50	0,105746	0,02968
41	30	50	0,101321	0,029008
42	35	50	0,110227	0,031998
43	40	50	0,099284	0,029159
44	45	50	0,093225	0,034486
45	50	50	0,112845	0,036509

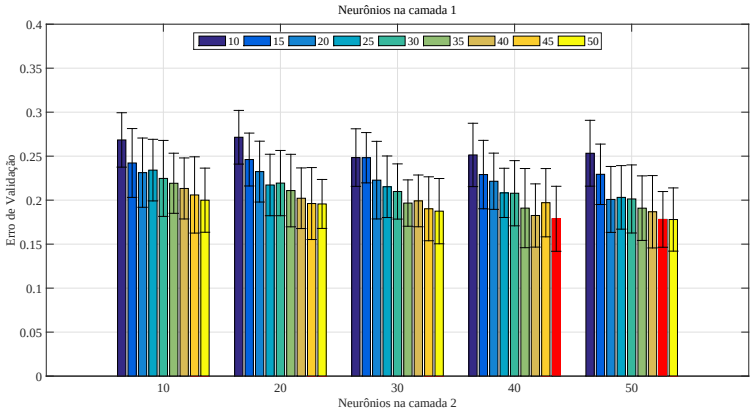


Figura 65: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ʧ//v/ do modelo MEHC.

Tabela 26: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ʧ//v/

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,268435	0,030919
2	15	10	0,242248	0,039121
3	20	10	0,231247	0,039357
4	25	10	0,234139	0,035006
5	30	10	0,224704	0,04318
6	35	10	0,219193	0,034158
7	40	10	0,213294	0,034677
8	45	10	0,205907	0,043334
9	50	10	0,199959	0,036443
10	10	20	0,271495	0,030509
11	15	20	0,246164	0,03005
12	20	20	0,232436	0,034582
13	25	20	0,217231	0,034934
14	30	20	0,219395	0,037078
15	35	20	0,2109	0,041204
16	40	20	0,202168	0,034455
17	45	20	0,196138	0,040899
18	50	20	0,195684	0,027827
19	10	30	0,248387	0,032712
20	15	30	0,248228	0,028584
21	20	30	0,222739	0,044101
22	25	30	0,215223	0,034977
23	30	30	0,209844	0,03138
24	35	30	0,196682	0,026343
25	40	30	0,199165	0,029412
26	45	30	0,19019	0,036341
27	50	30	0,187496	0,037102
28	10	40	0,251395	0,036004
29	15	40	0,229085	0,038873
30	20	40	0,221483	0,031955
31	25	40	0,208262	0,027975
32	30	40	0,207864	0,037047
33	35	40	0,190961	0,044997
34	40	40	0,182563	0,035931
35	45	40	0,19711	0,038795
36	50	40	0,178818	0,036943
37	10	50	0,253317	0,037428
38	15	50	0,229381	0,034336
39	20	50	0,200821	0,037471
40	25	50	0,203145	0,036103
41	30	50	0,201359	0,038686
42	35	50	0,190901	0,036682
43	40	50	0,186801	0,041123
44	45	50	0,178125	0,031661
45	50	50	0,177983	0,035924

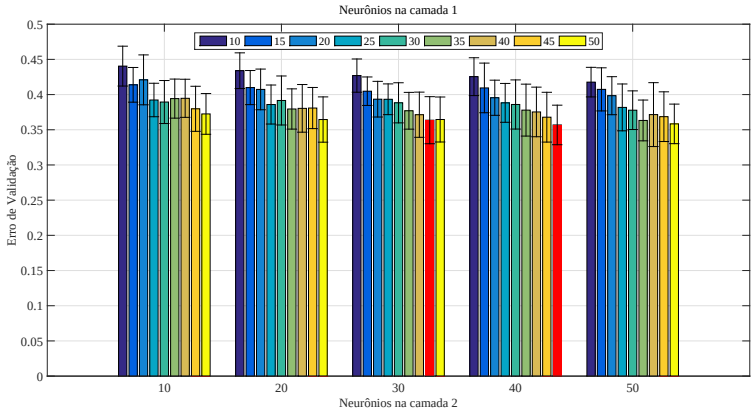


Figura 66: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /r/ do modelo MEHC.

Tabela 27: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /r/

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,440437	0,028314
2	15	10	0,413854	0,024646
3	20	10	0,42092	0,035445
4	25	10	0,392297	0,023775
5	30	10	0,389401	0,030548
6	35	10	0,394263	0,027746
7	40	10	0,394709	0,027068
8	45	10	0,379782	0,032006
9	50	10	0,372518	0,028923
10	10	20	0,43392	0,025366
11	15	20	0,409939	0,024222
12	20	20	0,407331	0,028891
13	25	20	0,385859	0,027717
14	30	20	0,391607	0,034807
15	35	20	0,37953	0,028591
16	40	20	0,380418	0,03386
17	45	20	0,380871	0,029226
18	50	20	0,364532	0,032172
19	10	30	0,426962	0,02361
20	15	30	0,404786	0,020235
21	20	30	0,393451	0,025334
22	25	30	0,393319	0,021837
23	30	30	0,388291	0,028493
24	35	30	0,376993	0,026124
25	40	30	0,371373	0,032096
26	45	30	0,36359	0,033434
27	50	30	0,364584	0,031962
28	10	40	0,42544	0,026878
29	15	40	0,409456	0,035259
30	20	40	0,395487	0,024969
31	25	40	0,388216	0,027599
32	30	40	0,385924	0,03491
33	35	40	0,377922	0,036888
34	40	40	0,375324	0,035235
35	45	40	0,367871	0,035371
36	50	40	0,356842	0,028092
37	10	50	0,417687	0,021012
38	15	50	0,4074	0,030623
39	20	50	0,398407	0,027005
40	25	50	0,381777	0,033273
41	30	50	0,377799	0,027457
42	35	50	0,36325	0,028967
43	40	50	0,371534	0,045384
44	45	50	0,368597	0,03531
45	50	50	0,358351	0,028102

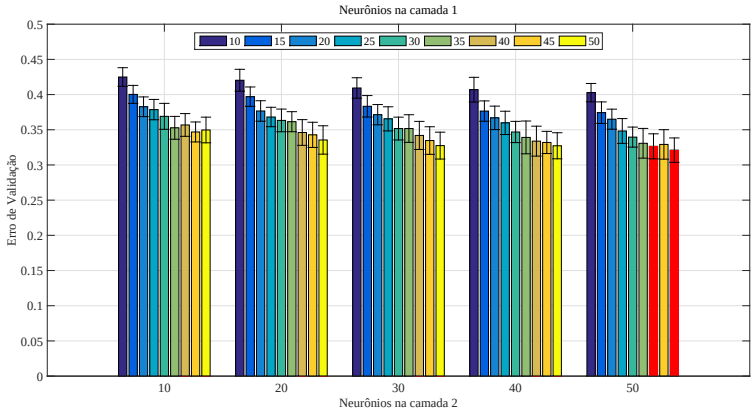


Figura 67: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ũ//u/ do modelo MEHC.

Tabela 28: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ũ//u/

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,424964	0,013218
2	15	10	0,400269	0,012772
3	20	10	0,382682	0,013986
4	25	10	0,378655	0,014447
5	30	10	0,369086	0,018421
6	35	10	0,352759	0,016237
7	40	10	0,356834	0,01616
8	45	10	0,346891	0,014176
9	50	10	0,349698	0,018231
10	10	20	0,420352	0,015541
11	15	20	0,397043	0,013754
12	20	20	0,376705	0,014537
13	25	20	0,368149	0,013784
14	30	20	0,363298	0,016088
15	35	20	0,361463	0,014236
16	40	20	0,34613	0,018252
17	45	20	0,342782	0,017881
18	50	20	0,335514	0,020149
19	10	30	0,409376	0,014527
20	15	30	0,383356	0,015214
21	20	30	0,37143	0,014396
22	25	30	0,365498	0,017144
23	30	30	0,351726	0,016197
24	35	30	0,351711	0,019631
25	40	30	0,341957	0,019953
26	45	30	0,33467	0,019603
27	50	30	0,327486	0,019088
28	10	40	0,406965	0,017547
29	15	40	0,376551	0,014452
30	20	40	0,36689	0,016681
31	25	40	0,359785	0,016544
32	30	40	0,346845	0,015083
33	35	40	0,339146	0,023274
34	40	40	0,333831	0,021252
35	45	40	0,331978	0,015753
36	50	40	0,327253	0,018526
37	10	50	0,402741	0,013027
38	15	50	0,374312	0,015267
39	20	50	0,365029	0,014233
40	25	50	0,348255	0,017616
41	30	50	0,339566	0,01423
42	35	50	0,330797	0,021114
43	40	50	0,326481	0,017771
44	45	50	0,329177	0,020936
45	50	50	0,321042	0,017347

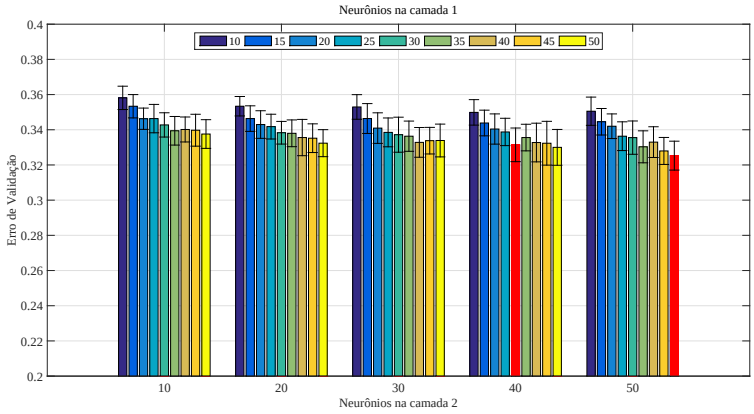


Figura 68: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista /ũ//u/ do modelo MEHC.

Tabela 29: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em /ũ//u/

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,358128	0,00663
2	15	10	0,353368	0,006623
3	20	10	0,346271	0,006049
4	25	10	0,346346	0,008063
5	30	10	0,34276	0,006898
6	35	10	0,339424	0,008126
7	40	10	0,340161	0,007094
8	45	10	0,339744	0,009064
9	50	10	0,337576	0,008153
10	10	20	0,35338	0,005521
11	15	20	0,346349	0,007261
12	20	20	0,34297	0,00784
13	25	20	0,341791	0,007037
14	30	20	0,338309	0,006437
15	35	20	0,33799	0,007606
16	40	20	0,335577	0,010308
17	45	20	0,335213	0,008153
18	50	20	0,332385	0,007674
19	10	30	0,352964	0,00699
20	15	30	0,346387	0,008485
21	20	30	0,340949	0,008685
22	25	30	0,338512	0,008188
23	30	30	0,337191	0,009955
24	35	30	0,336346	0,008589
25	40	30	0,332834	0,008454
26	45	30	0,333799	0,007554
27	50	30	0,3339	0,009314
28	10	40	0,349879	0,007238
29	15	40	0,343846	0,007292
30	20	40	0,340426	0,008589
31	25	40	0,338758	0,00777
32	30	40	0,331432	0,009561
33	35	40	0,335574	0,007557
34	40	40	0,332746	0,010988
35	45	40	0,332352	0,012443
36	50	40	0,330004	0,010179
37	10	50	0,350539	0,008024
38	15	50	0,344555	0,007531
39	20	50	0,342008	0,007002
40	25	50	0,336345	0,008171
41	30	50	0,335529	0,009462
42	35	50	0,330311	0,009083
43	40	50	0,332976	0,008734
44	45	50	0,327964	0,007665
45	50	50	0,325298	0,008214

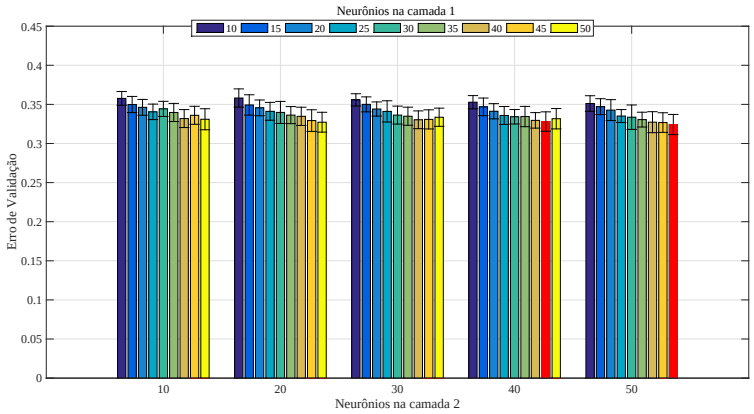


Figura 69: Média e desvio padrão dos erros de validação para diferentes distribuições de neurônios entre a camada 1 e 2 para a rede especialista εdo modelo MEHC.

Tabela 30: Média e desvio padrão dos erros de validação para diferentes configurações topológicas do especialista em ε

			Erro Validação	
# Hipótese	**# neurônios Camada 1**	**# neurônios Camada 2**	**Média**	**Desvio Padrão**
1	10	10	0,357713	0,008804
2	15	10	0,349855	0,010334
3	20	10	0,346264	0,010102
4	25	10	0,340469	0,009905
5	30	10	0,344252	0,009659
6	35	10	0,339668	0,011594
7	40	10	0,331879	0,011467
8	45	10	0,336058	0,011497
9	50	10	0,33101	0,013427
10	10	20	0,358158	0,011668
11	15	20	0,349306	0,012995
12	20	20	0,345538	0,010079
13	25	20	0,341159	0,011403
14	30	20	0,339755	0,014078
15	35	20	0,336289	0,010912
16	40	20	0,33475	0,01166
17	45	20	0,329303	0,013834
18	50	20	0,327228	0,012679
19	10	30	0,355787	0,007837
20	15	30	0,350042	0,009563
21	20	30	0,344096	0,009062
22	25	30	0,34106	0,01357
23	30	30	0,336321	0,011477
24	35	30	0,334863	0,011564
25	40	30	0,330325	0,011355
26	45	30	0,330825	0,012228
27	50	30	0,333575	0,011623
28	10	40	0,3528	0,008441
29	15	40	0,346832	0,011281
30	20	40	0,341241	0,009725
31	25	40	0,335852	0,011441
32	30	40	0,334188	0,009234
33	35	40	0,334459	0,012979
34	40	40	0,329571	0,009914
35	45	40	0,328071	0,012401
36	50	40	0,331688	0,013009
37	10	50	0,351145	0,009919
38	15	50	0,34709	0,010192
39	20	50	0,342705	0,013377
40	25	50	0,335143	0,008239
41	30	50	0,333647	0,015604
42	35	50	0,330537	0,009458
43	40	50	0,327253	0,013415
44	45	50	0,32676	0,012537
45	50	50	0,324245	0,012915

A.2 Matrizes de confusão

Tabela 31: Matriz de confusão correspondente ao modelo MEF usando o banco de dados *Birthdate*

	/ind/	/n/	/>/	/v/	/i/	/u~/	/z/	/E/	/r/	/o/	/k/	/wa/	/t/	/u/	/e/	/dc/	/oj/	/.pau/	/s/	/tS/	/m/	/a/	/i~/	/ej/	/e~/	soma
/ind/	856	3	0	186	13	177	0	21	33	0	36	2	719	35	0	0	0	46	0	0	5	2	0	0	22	2156
/n/	31	177	1	50	1	53	0	0	5	0	0	0	22	1	0	0	0	0	0	0	1	3	0	0	7	352
/>/	296	25	244	5	0	72	0	4	0	0	8	47	4	4	0	0	0	0	0	0	1	15	0	0	24	749
/v/	41	6	0	291	3	43	0	0	1	0	0	3	7	7	0	1	0	0	0	0	4	2	0	0	5	414
/i/	33	8	0	46	289	65	0	3	9	0	0	0	29	1	0	0	4	0	0	0	4	0	0	0	24	515
/u~/	39	6	0	74	6	623	0	2	4	0	3	0	83	27	0	0	2	0	0	0	1	1	0	0	4	875
/z/	51	7	0	378	2	39	1800	0	8	0	0	0	82	10	0	0	0	0	33	0	0	1	0	0	7	2418
/E/	371	0	0	9	20	50	0	2277	94	0	0	0	40	0	0	0	16	0	0	0	6	181	0	0	388	3452
/r/	263	8	0	100	0	84	0	60	768	0	0	0	107	7	0	0	3	0	0	0	11	15	0	0	24	1450
/o/	11	5	1	30	0	51	0	5	13	103	8	0	18	13	0	0	9	0	0	0	0	0	0	0	0	267
/k/	60	2	0	11	0	60	0	0	1	0	167	0	45	12	0	0	0	0	0	0	0	0	0	0	0	358
/wa/	69	9	25	3	0	28	0	11	0	0	3	365	0	7	0	0	0	0	0	0	0	45	0	0	24	589
/t/	21	0	0	7	0	6	0	0	4	0	1	0	368	3	0	0	0	0	0	0	0	0	0	0	3	413
/u/	359	19	0	169	0	677	0	5	107	0	17	0	373	214	0	0	0	1	0	0	13	3	0	0	11	1968
/e/	39	1	0	5	18	22	0	17	6	0	0	0	31	0	61	0	21	0	0	0	1	1	0	0	49	272
/dc/	39	3	0	91	0	9	0	0	0	0	0	0	20	0	0	45	0	0	0	0	0	0	0	0	0	207
/oj/	204	7	1	19	67	403	0	25	63	0	5	0	39	36	0	0	164	0	0	0	8	1	0	0	64	1106
/.pau/	442	0	0	54	0	16	0	0	1	0	14	0	103	15	0	0	0	1538	0	0	0	1	0	0	5	2189
/s/	13	0	0	54	4	5	0	2	4	0	3	0	0	0	0	0	0	0	2390	0	0	0	0	0	2	2477
/tS/	3	0	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	256	0	0	0	0	0	264
/m/	11	0	0	19	1	22	0	2	5	0	0	0	29	2	0	0	0	0	0	0	66	0	2	0	2	161
/a/	69	4	0	13	0	6	0	23	16	0	0	19	11	1	0	0	0	0	0	0	0	61	0	0	11	234
/i~/	92	16	0	98	71	300	0	4	3	0	1	0	16	16	0	0	0	0	0	0	10	0	118	0	23	768
/ej/	38	0	0	29	189	0	0	68	95	0	0	0	45	0	0	0	36	0	0	0	1	3	0	318	97	919
/e~/	24	3	0	27	2	35	0	59	22	0	0	0	21	8	0	0	0	0	0	0	2	11	0	0	214	428
Soma	3475	309	272	1772	686	2847	1800	2588	1262	103	266	436	2212	419	61	46	255	1585	2423	256	134	346	120	318	1010	**Acurácia**
Sensibilidade	39,70315399	50,28409091	32,57676903	70,28985507	56,11650485	71,2	74,44168734	65,9617613	52,96551724	38,57677903	46,64804469	61,96943973	89,10411622	10,87398374	22,42647059	21,73913043	14,82820976	70,26039287	96,48768672	96,96969697	40,90378882	26,06837607	15,36458333	34,60282916	50	55,08979641
Precisão	24,63309353	57,2815534	89,70588235	16,4221219	42,12827988	21,88268353	100	87,98299845	60,85578447	100	62,78195489	83,71559633	16,63652803	51,07398568	100	97,82608696	64,31372549	97,03470032	98,638052	100	49,25373134	17,6300578	98,33333333	100	21,18811881	

Tabela 32: Matriz de confusão correspondente ao modelo MEF usando o banco de dados *Age*

	/ind/	/n/	/>/	/v/	/i/	/u~/	/z/	/E/	/r/	/o/	/k/	/wa/	/t/	/u/	/e/	/dc/	/oj/	/.pau/	/s/	/tS/	/m/	/a/	/i~/	/ej/	/e~/	soma
/ind/	901	2	0	50	2	39	0	32	45	0	47	2	485	45	0	0	0	78	0	0	4	4	0	0	32	1768
/n/	12	187	8	50	2	21	0	0	6	0	0	0	45	2	0	0	0	0	0	0	1	5	0	0	8	347
/>/	100	54	450	5	0	300	0	7	0	23	3	47	3	3	0	0	0	0	0	0	1	12	0	0	24	1032
/v/	31	6	0	500	5	34	0	0	3	0	1	3	23	8	0	4	0	0	0	0	4	7	0	0	6	635
/i/	10	2	0	46	325	32	0	4	7	0	0	0	34	3	0	0	8	0	0	0	4	0	0	0	34	509
/u~/	20	6	0	27	7	756	0	9	5	0	3	0	79	37	0	0	6	0	0	0	1	2	0	0	3	961
/z/	15	8	0	200	8	45	2300	0	8	0	0	0	90	12	0	0	0	0	32	0	0	2	0	0	8	2728
/E/	200	0	0	9	23	23	0	2706	345	0	0	0	45	0	0	0	45	0	3	0	6	120	0	0	321	3846
/r/	187	9	5	100	0	4	0	45	878	0	0	0	91	8	0	0	7	0	0	0	21	12	0	0	24	1391
/o/	29	4	7	30	0	3	0	3	34	167	4	0	32	23	0	0	2	0	0	0	0	0	0	0	0	338
/k/	80	7	0	11	2	56	0	0	3	0	208	0	93	45	0	0	0	0	0	0	0	0	0	0	0	505
/wa/	178	15	30	3	0	34	0	23	0	0	4	185	0	3	0	0	0	0	0	0	0	76	0	0	21	572
/t/	12	0	0	7	0	4	0	0	4	0	2	0	290	4	0	0	3	0	0	0	0	0	0	0	4	330
/u/	400	23	0	239	0	344	0	5	243	0	23	0	312	340	0	0	0	3	0	0	12	4	0	0	21	1969
/e/	10	1	0	5	15	34	0	34	7	0	0	0	34	0	56	0	32	0	0	0	2	2	0	0	65	297
/dc/	39	3	0	91	0	4	0	0	0	0	0	0	29	0	0	78	0	0	0	0	0	0	0	0	0	244
/oj/	245	2	0	19	78	340	0	52	54	0	2	0	19	45	0	0	235	0	0	0	8	2	0	0	45	1146
/.pau/	376	0	0	53	0	24	0	0	2	0	21	0	96	17	0	0	0	1890	0	0	0	2	0	0	3	2484
/s/	14	0	0	21	4	4	13	1	5	0	2	0	0	0	0	0	0	0	2285	0	0	0	0	0	9	2358
/tS/	2	0	0	6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	231	0	0	0	0	0	240
/m/	23	0	0	23	1	33	0	3	7	0	0	0	34	5	0	0	0	0	0	0	100	0	1	0	8	238
/a/	100	7	0	15	0	3	0	24	26	0	0	19	23	1	0	0	0	0	0	0	0	98	0	0	6	322
/i~/	12	8	0	100	71	278	0	5	4	0	3	0	56	12	0	0	0	0	0	0	23	0	106	0	54	732
/ej/	23	0	0	30	232	0	0	54	85	0	0	0	12	0	0	0	42	0	0	0	1	2	0	245	45	771
/e~/	40	3	0	10	2	12	0	52	21	0	0	0	54	5	0	0	0	0	0	0	2	11	0	0	258	470
Soma	3059	347	500	1650	777	2428	2313	3059	1792	190	323	256	1979	618	56	82	380	1971	2320	231	190	361	107	245	999	**Acurácia**
Sensibilidade	50,9615385	53,8904899	43,6046512	78,7401575	63,8506876	78,6680541	84,3108504	70,3588144	63,1200575	49,408284	41,1881188	32,3426573	87,8787879	17,2676486	18,8552189	31,9672131	20,5061082	76,0869565	96,90415606	96,25	42,0168067	30,4347826	14,4808743	31,7769131	54,893617	60,13418214
Precisão	29,45407	53,8904899	90	30,3030303	41,8275418	31,1367381	99,4379594	88,4602811	48,9955357	87,8947368	64,3962848	72,265625	14,6538656	55,0161812	100	95,1219512	61,8421053	95,890411	98,49137931	100	52,6315789	27,1468144	99,0654206	100	25,8258258	

Tabela 33: Matriz de confusão correspondente ao modelo MEF com pós-processamento dinâmico usando o banco de dados *Birthdate*

	/ind/	/n/	/>/	/v/	/i/	/u~/	/z/	/E/	/r/	/o/	/k/	/wa/	/t/	/u/	/e/	/dc/	/oj/	/.pau/	/s/	/tS/	/m/	/a/	/i~/	/ej/	/e~/	soma
/ind/	1546	3	0	45	13	139	0	21	33	0	36	2	213	35	0	0	0	46	0	0	5	2	0	0	17	2156
/n/	31	252	1	43	1	14	0	0	5	0	0	0	4	0	0	0	0	0	0	0	1	0	0	0	0	352
/>/	187	4	505	2	0	3	0	4	0	0	8	13	3	3	0	0	0	0	0	0	1	1	0	0	15	749
/v/	37	6	0	329	3	23	0	0	1	0	0	0	9	0	0	1	0	0	0	0	0	2	0	0	3	414
/i/	3	3	0	23	443	3	0	3	4	0	0	0	21	1	0	0	4	0	0	0	4	0	0	0	3	515
/u~/	12	2	0	43	3	728	0	2	2	0	3	0	53	21	0	0	1	0	0	0	1	1	0	0	3	875
/z/	31	4	0	123	2	2	2189	0	3	0	3	0	53	1	0	0	0	0	3	0	0	1	0	0	3	2418
/E/	241	0	0	2	13	23	0	2740	34	0	0	0	23	0	0	0	3	0	0	0	3	46	0	0	324	3452
/r/	221	4	0	31	0	43	0	31	1036	0	0	0	53	3	0	0	1	0	0	0	3	5	0	0	19	1450
/o/	3	3	1	23	0	32	0	4	5	184	0	0	2	7	0	0	3	0	0	0	0	0	0	0	0	267
/k/	42	2	0	7	0	42	0	0	1	0	257	0	3	4	0	0	0	0	0	0	0	0	0	0	0	358
/wa/	41	3	14	3	0	12	0	3	0	0	1	474	0	2	0	0	0	0	0	0	0	32	0	0	4	589
/t/	21	0	0	0	0	3	0	0	3	0	0	0	382	1	0	0	0	0	0	0	0	0	0	0	3	413
/u/	219	3	0	65	0	214	0	5	87	0	0	0	241	1123	0	0	0	1	0	0	3	3	0	0	4	1968
/e/	13	0	0	2	1	2	0	13	6	0	0	0	21	0	184	0	14	0	0	0	1	0	0	0	15	272
/dc/	23	3	0	32	0	9	0	0	0	0	0	0	20	0	0	120	0	0	0	0	0	0	0	0	0	207
/oj/	160	7	1	5	31	130	0	21	27	0	2	0	32	14	0	0	639	0	0	0	4	1	0	0	32	1106
/.pau/	123	0	0	23	0	0	0	0	0	0	4	0	54	0	0	0	0	1984	0	0	0	1	0	0	0	2189
/s/	6	0	0	31	2	1	0	1	2	0	0	0	0	0	0	0	0	0	2433	0	0	0	0	0	1	2477
/tS/	3	0	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	256	0	0	0	0	0	264
/m/	10	0	0	0	0	12	0	1	3	0	0	0	15	2	0	0	0	0	0	0	117	0	1	0	0	161
/a/	41	4	0	5	0	0	0	16	9	0	0	13	5	1	0	0	0	0	0	0	0	135	0	0	5	234
/i~/	63	7	0	47	54	126	0	4	3	0	1	0	14	12	0	0	0	0	0	0	2	0	416	0	19	768
/ej/	23	0	0	21	67	0	0	35	45	0	0	0	32	0	0	0	23	0	0	0	1	3	0	626	43	919
/e~/	15	1	0	12	2	14	0	20	12	0	0	0	3	4	0	0	0	0	0	0	2	2	0	0	341	428
Soma	3115	311	522	921	635	1576	2189	2924	1321	184	315	502	1256	1234	184	121	688	2031	2436	256	148	235	417	626	854	**Acurácia**
Sensibilidade	71,7068646	71,5909091	67,423231	79,468599	86,0194175	83,2	90,5293631	79,3742758	71,4482759	68,9138577	71,7877095	80,475382	92,4939467	57,0630081	67,6470588	57,9710145	57,7757685	90,6349931	98,22365765	96,969697	72,6708075	57,6923077	54,1666667	68,117519	79,6728972	77,75288988
Precisão	49,6308186	81,0289389	96,743295	35,7220413	69,7637795	46,1928934	100	93,7072503	78,4254353	100	81,5873016	94,4223108	30,4140127	91,0048622	100	99,1735537	92,877907	97,685869	99,87684729	100	79,0540541	57,4468085	99,7601918	100	39,9297424	

Tabela 34: Matriz de confusão correspondente ao modelo MEF com pós-processamento dinâmico usando o banco de dados *Age*

	/ind/	/n/	/>/	/v/	/i/	/u~/	/z/	/E/	/r/	/o/	/k/	/wa/	/t/	/u/	/e/	/dc/	/oj/	/.pau/	/s/	/tS/	/m/	/a/	/i~/	/ej/	/e~/	soma
/ind/	1135	2	0	50	2	39	0	32	45	0	47	2	315	45	0	0	0	34	0	0	4	4	0	0	12	1768
/n/	12	254	3	36	2	14	0	0	3	0	0	0	7	2	0	0	0	0	0	0	1	5	0	0	8	347
/>/	100	6	815	5	0	45	0	4	0	4	3	15	3	3	0	0	0	0	0	0	1	4	0	0	24	1032
/v/	31	6	0	541	5	14	0	0	3	0	1	3	5	8	0	2	0	0	0	0	4	7	0	0	5	635
/i/	5	2	0	12	440	9	0	4	7	0	0	0	17	3	0	0	3	0	0	0	2	0	0	0	5	509
/u~/	15	1	0	27	1	813	0	9	8	0	3	0	39	37	0	0	2	0	0	0	1	2	0	0	3	961
/z/	15	8	0	173	1	14	2462	0	3	0	0	0	38	2	0	0	0	0	4	0	0	2	0	0	6	2728
/E/	200	0	0	9	4	23	0	3128	56	0	0	0	13	0	0	0	9	0	3	0	6	74	0	0	321	3846
/r/	187	1	3	21	0	4	0	8	1084	0	0	0	38	8	0	0	7	0	0	0	4	2	0	0	24	1391
/o/	5	4	0	30	0	3	0	3	27	258	0	0	4	2	0	0	2	0	0	0	0	0	0	0	0	338
/k/	23	1	0	6	2	36	0	0	3	0	416	0	15	3	0	0	0	0	0	0	0	0	0	0	0	505
/wa/	32	7	19	3	0	23	0	8	0	0	4	392	0	3	0	0	0	0	0	0	0	23	0	0	58	572
/t/	12	0	0	7	0	4	0	0	4	0	2	0	292	3	0	0	2	0	0	0	0	0	0	0	4	330
/u/	198	4	0	100	0	198	0	6	56	0	4	0	210	1175	0	0	0	3	0	0	6	4	0	0	5	1969
/e/	10	1	0	5	4	3	0	21	7	0	0	0	29	0	171	0	32	0	0	0	2	2	0	0	10	297
/dc/	21	2	0	41	3	4	0	0	0	0	0	0	21	0	0	152	0	0	0	0	0	0	0	0	0	244
/oj/	128	2	0	4	24	210	0	13	21	0	2	0	19	23	0	0	664	0	0	0	2	2	0	0	32	1146
/.pau/	187	0	0	12	0	3	0	0	2	0	5	0	24	12	0	0	0	2236	0	0	0	2	0	0	1	2484
/s/	5	0	0	21	2	1	0	1	2	0	0	0	0	0	0	0	0	0	2324	0	0	0	0	0	2	2358
/tS/	2	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	234	0	0	0	0	0	240
/m/	14	0	0	3	1	21	0	3	1	0	0	0	10	1	0	0	0	0	0	0	175	0	1	0	8	238
/a/	53	2	0	7	0	3	0	16	7	0	0	19	4	1	0	0	0	0	0	0	0	206	0	0	4	322
/i~/	12	3	0	34	34	145	0	0	0	0	3	0	23	4	0	0	0	0	0	0	12	0	439	0	23	732
/ej/	23	0	0	30	54	0	0	29	34	0	0	0	12	0	0	0	12	0	0	0	1	2	0	531	43	771
/e~/	13	1	0	10	2	12	0	23	14	0	0	0	2	5	0	0	0	0	0	0	2	2	0	0	384	470
Soma	2438	307	840	1190	581	1642	2462	3308	1387	262	490	431	1140	1340	171	154	733	2273	2331	234	223	343	440	531	982	**Acurácia**
Sensibilidade	64,1968326	73,1988473	78,9728682	85,1968504	86,4440079	84,5993757	90,2492669	81,3312533	77,9295471	76,3313609	82,3762376	68,5314685	88,4848485	59,6749619	57,5757576	62,295082	57,9406632	90,0161031	98,55810008	97,5	73,5294118	63,9751553	59,9726776	68,8715953	81,7021277	78,98829718
Precisão	46,5545529	82,7361564	97,0238095	45,4621849	75,7314974	49,5127893	100	94,5586457	78,1542898	98,4732824	84,8979592	90,9512761	25,6140351	87,6865672	100	98,7012987	90,5866303	98,3721953	99,6996997	100	78,4753363	60,058309	99,7727273	100	39,1038697	

Tabela 35: Matriz de confusão correspondente ao modelo MEHC usando o banco de dados *Birthdate*

	/ind/	/n/	/>/	/v/	/i/	/u~/	/z/	/E/	/r/	/o/	/k/	/wa/	/t/	/u/	/e/	/dc/	/oj/	/.pau/	/s/	/tS/	/m/	/a/	/i~/	/ej/	/e~/	soma
/ind/	1489	3	0	86	13	76	0	21	33	0	36	2	287	35	0	0	0	46	0	0	5	2	0	0	22	2156
/n/	31	302	1	0	1	0	0	0	5	0	0	0	0	1	0	0	0	0	0	0	1	3	0	0	7	352
/>/	196	25	479	5	0	12	0	4	0	0	8	0	0	4	0	0	0	0	0	0	1	15	0	0	0	749
/v/	41	6	0	339	3	0	0	0	1	0	0	3	7	7	0	1	0	0	0	0	4	2	0	0	0	414
/i/	33	8	0	25	344	50	0	3	9	0	0	0	10	1	0	0	4	0	0	0	4	0	0	0	24	515
/u~/	39	6	0	34	6	746	0	2	4	0	3	0	0	27	0	0	2	0	0	0	1	1	0	0	4	875
/z/	51	7	0	32	2	0	2274	0	8	0	0	0	0	10	0	0	0	0	33	0	0	1	0	0	0	2418
/E/	371	0	0	0	20	2	0	2834	94	0	0	0	40	0	0	0	16	0	0	0	6	33	0	0	36	3452
/r/	263	8	0	20	0	0	0	60	1010	0	0	0	53	7	0	0	3	0	0	0	11	15	0	0	0	1450
/o/	11	5	1	15	0	41	0	0	13	151	8	0	0	13	0	0	9	0	0	0	0	0	0	0	0	267
/k/	60	2	0	0	0	0	0	0	1	0	238	0	45	12	0	0	0	0	0	0	0	0	0	0	0	358
/wa/	69	9	25	0	0	28	0	0	0	0	3	403	0	7	0	0	0	0	0	0	0	45	0	0	0	589
/t/	21	0	0	7	0	0	0	0	4	0	1	0	374	3	0	0	0	0	0	0	0	0	0	0	3	413
/u/	359	19	0	0	0	129	0	5	50	0	17	0	49	1323	0	0	0	1	0	0	13	3	0	0	0	1968
/e/	39	1	0	5	0	0	0	17	6	0	0	0	31	0	101	0	21	0	0	0	1	1	0	0	49	272
/dc/	39	3	0	0	0	0	0	0	0	0	0	0	20	0	0	145	0	0	0	0	0	0	0	0	0	207
/oj/	204	7	1	2	67	21	0	0	20	0	5	0	39	36	0	0	695	0	0	0	8	1	0	0	0	1106
/.pau/	442	0	0	12	0	0	0	0	1	0	14	0	38	15	0	0	0	1661	0	0	0	1	0	0	5	2189
/s/	13	0	0	2	4	0	0	2	4	0	3	0	0	0	0	0	0	0	2447	0	0	0	0	0	2	2477
/tS/	3	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	257	0	0	0	0	0	264
/m/	11	0	0	0	1	22	0	2	0	0	0	0	0	2	0	0	0	0	0	0	119	0	2	0	2	161
/a/	69	4	0	13	0	6	0	23	16	0	0	19	0	1	0	0	0	0	0	0	0	83	0	0	0	234
/i~/	92	16	0	18	71	47	0	4	3	0	1	0	0	16	0	0	0	0	0	0	10	0	467	0	23	768
/ej/	38	0	0	0	89	0	0	18	10	0	0	0	0	0	0	0	0	0	0	0	1	3	0	657	67	919
/e~/	24	3	0	0	2	35	0	59	0	0	0	0	0	8	0	0	0	0	0	0	2	11	0	0	284	428
Soma	4008	434	507	619	623	1215	2274	3054	1292	151	337	427	993	1528	101	146	786	1708	2480	257	187	220	469	657	528	**Acurácia**
Sensibilidade	69,06307978	85,79545455	63,95193591	81,88405797	66,7961165	85,25714286	94,04466501	82,09733488	69,65517241	56,55430712	66,48044693	68,42105263	90,55690073	67,22560976	37,13235294	70,04830918	62,83905967	75,87939698	98,78885749	97,34848485	73,91304348	35,47008547	60,80729167	71,49075082	66,35514019	76,8849246
Precisão	37,1506986	69,58525346	94,47731755	54,76575121	55,21669342	61,39917695	100	92,79633268	78,17337461	100	70,6231454	94,3793911	37,66364552	86,58376963	100	99,31506849	88,42239186	97,24824356	98,66935484	100	63,63636364	37,72727273	99,57356077	100	53,78787879	

Tabela 36: Matriz de confusão correspondente ao modelo MEHC usando o banco de dados *Age*

	/ind/	/n/	/>/	/v/	/i/	/u~/	/z/	/E/	/r/	/o/	/k/	/wa/	/t/	/u/	/e/	/dc/	/oj/	/.pau/	/s/	/tS/	/m/	/a/	/i~/	/ej/	/e~/	soma
/ind/	1476	1	0	12	13	43	0	26	23	0	2	1	120	12	0	0	0	23	0	0	3	1	0	0	12	1768
/n/	25	304	1	0	1	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	1	4	0	0	8	347
/>/	285	45	605	23	0	24	0	8	0	0	3	0	0	5	0	0	0	0	0	0	4	30	0	0	0	1032
/v/	62	9	0	512	5	0	0	0	2	0	0	5	13	14	0	4	0	0	0	0	3	6	0	0	0	635
/i/	33	8	0	25	339	48	0	5	7	0	0	0	12	1	0	0	6	0	0	0	2	0	0	0	23	509
/u~/	48	7	0	38	9	806	0	4	4	0	3	0	0	30	0	0	3	0	0	0	1	2	0	0	6	961
/z/	65	8	0	42	5	0	2554	0	11	0	0	0	0	16	0	0	0	0	24	0	0	3	0	0	0	2728
/E/	391	0	0	0	28	2	0	3184	98	0	0	0	45	0	0	0	26	0	0	0	6	36	0	0	30	3846
/r/	255	4	0	29	0	0	0	56	985	0	0	0	41	3	0	0	1	0	0	0	6	11	0	0	0	1391
/o/	15	4	1	23	0	60	0	0	18	186	12	0	0	10	0	0	9	0	0	0	0	0	0	0	0	338
/k/	68	7	0	0	0	0	0	0	9	0	318	0	70	33	0	0	0	0	0	0	0	0	0	0	0	505
/wa/	72	9	27	0	0	35	0	0	0	0	4	387	0	6	0	0	0	0	0	0	0	32	0	0	0	572
/t/	17	0	0	4	0	0	0	0	1	0	1	0	299	3	0	0	0	0	0	2	0	0	0	0	3	330
/u/	367	25	0	0	0	123	0	5	34	0	15	0	53	1333	0	0	0	1	0	0	10	3	0	0	0	1969
/e/	42	3	0	8	0	0	0	20	8	0	0	0	35	0	94	0	27	0	0	0	3	3	0		54	297
/dc/	45	6	0	0	0	0	0	0	0	0	0	0	24	0	0	169	0	0	0	0	0	0	0	0	0	244
/oj/	212	7	1	2	67	21	0	0	21	0	5	0	43	45	0	0	710	0	0	0	9	3	0	0	0	1146
/.pau/	462	0	0	19	0	0	0	0	3	0	32	0	46	23	0	0	0	1888	0	0	0	4	0	0	7	2484
/s/	3	0	0	1	2	0	0	1	1	0	2	0	0	0	0	0	0	0	2347	0	0	0	0	0	1	2358
/tS/	5	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	232	0	0	0	0	0	240
/m/	19	0	0	0	4	33	0	4	0	0	0	0	0	6	0	0	0	0	0	0	165	0	3	0	4	238
/a/	83	5	0	32	0	9	0	21	23	0	0	34	0	6	0	0	0	0	0	0	0	109	0	0	0	322
/i~/	95	13	0	18	71	47	0	4	5	0	1	0	0	14	0	0	0	0	0	0	4	0	448	0	12	732
/ej/	23	0	0	0	75	0	0	7	5	0	0	0	0	0	0	0	23	0	0	0	0	1	0	577	60	771
/e~/	25	3	0	0	3	38	0	64	0	0	0	0	0	9	0	0	0	0	0	0	6	13	0	0	309	470
Soma	4193	468	635	791	622	1289	2554	3409	1260	186	398	427	801	1570	94	173	805	1912	2371	234	223	261	451	577	529	**Acurácia**
Sensibilidade	83,4841629	87,6080692	58,624031	80,6299213	66,6011788	83,8709677	93,6217009	82,7873115	70,8123652	55,0295858	62,970297	67,6573427	90,6060606	67,6993398	31,6498316	69,2622951	61,9546248	76,0064412	99,533503	96,6666667	69,3277311	33,8509317	61,2021858	74,8378729	65,7446809	77,5206801
Precisão	35,2015264	64,957265	95,2755906	64,7281922	54,5016077	62,5290923	100	93,399824	78,1746032	100	79,8994975	90,6323185	37,3283396	84,9044586	100	97,6878613	88,1987578	98,7447699	98,9877689	99,1452991	73,9910314	41,7624521	99,3348115	100	58,4120983	

Tabela 37: Matriz de confusão correspondente ao modelo MEHC com pós-processamento dinâmico usando o banco de dados *Birthdate*

	/ind/	/n/	/>/	/v/	/i/	/u~/	/z/	/E/	/r/	/o/	/k/	/wa/	/t/	/u/	/e/	/dc/	/oj/	/.pau/	/s/	/tS/	/m/	/a/	/i~/	/ej/	/e~/	soma
/ind/	1730	3	0	12	13	76	0	21	33	0	36	2	128	35	0	0	0	46	0	0	5	2	0	0	14	2156
/n/	31	309	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	7	352
/>/	150	12	561	5	0	4	0	4	0	0	8	0	0	4	0	0	0	0	0	0	1	0	0	0	0	749
/v/	41	6	0	348	3	0	0	0	1	0	0	3	2	3	0	1	0	0	0	0	4	2	0	0	0	414
/i/	12	8	0	5	435	21	0	3	4	0	0	0	1	1	0	0	3	0	0	0	4	0	0	0	18	515
/u~/	12	6	0	21	6	804	0	2	4	0	3	0	0	11	0	0	0	0	0	0	1	1	0	0	4	875
/z/	51	7	0	21	2	0	2317	0	3	0	0	0	0	3	0	0	0	0	13	0	0	1	0	0	0	2418
/E/	218	0	0	0	16	2	0	3115	34	0	0	0	32	0	0	0	13	0	0	0	6	14	0	0	2	3452
/r/	129	8	0	17	0	0	0	34	1227	0	0	0	12	7	0	0	3	0	0	0	7	6	0	0	0	1450
/o/	5	3	1	12	0	23	0	0	8	207	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	267
/k/	23	0	0	0	0	0	0	0	1	0	310	0	12	12	0	0	0	0	0	0	0	0	0	0	0	358
/wa/	31	9	14	0	0	16	0	0	0	0	3	486	0	7	0	0	0	0	0	0	0	23	0	0	0	589
/t/	10	0	0	7	0	0	0	0	4	0	1	0	385	3	0	0	0	0	0	0	0	0	0	0	3	413
/u/	216	12	0	0	0	79	0	5	34	0	9	0	28	1575	0	0	0	1	0	0	9	0	0	0	0	1968
/e/	39	1	0	5	0	0	0	5	6	0	0	0	34	0	156	0	12	0	0	0	1	1	0	0	12	272
/dc/	12	1	0	0	0	0	0	0	0	0	0	0	11	0	0	183	0	0	0	0	0	0	0	0	0	207
/oj/	137	1	0	2	43	17	0	0	12	0	5	0	17	26	0	0	840	0	0	0	5	1	0	0	0	1106
/.pau/	231	0	0	2	0	0	0	0	1	0	7	0	26	8	0	0	0	1911	0	0	0	0	0	0	3	2189
/s/	13	0	0	2	4	0	0	2	0	0	3	0	0	0	0	0	0	0	2451	0	0	0	0	0	2	2477
/tS/	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	258	0	0	0	0	0	264
/m/	6	0	0	0	1	10	0	2	0	0	0	0	0	2	0	0	0	0	0	0	136	0	2	0	2	161
/a/	45	0	0	0	0	5	0	14	7	0	0	12	0	0	0	0	0	0	0	0	0	151	0	0	0	234
/i~/	67	6	0	12	34	23	0	0	3	0	1	0	0	4	0	0	0	0	0	0	7	0	599	0	12	768
/ej/	26	0	0	0	34	0	0	12	6	0	0	0	0	0	0	0	0	0	0	0	0	1	0	773	67	919
/e~/	21	1	0	0	2	12	0	21	0	0	0	0	0	2	0	0	0	0	0	0	1	3	0	0	365	428
Soma	3259	393	577	474	594	1092	2317	3240	1389	207	394	503	688	1704	156	184	871	1958	2464	258	188	206	601	773	511	**Acurácia**
Sensibilidade	80,2411874	87,7840909	74,8998665	84,057971	84,4660194	91,8857143	95,8229942	90,2375435	84,6206897	77,5280899	86,5921788	82,5127334	93,220339	80,0304878	57,3529412	88,4057971	75,9493671	87,300137	98,9503432	97,7272727	84,4720497	64,5299145	77,9947917	84,1131665	85,2803738	86,52453902
Precisão	53,083768	78,6259542	97,2270364	73,4177215	73,2323232	73,6263736	100	96,1419753	88,336933	100	78,680203	96,6202783	55,9593023	92,4295775	100	99,4565217	96,4408726	97,5995914	99,4724026	100	72,3404255	73,3009709	99,6672213	100	71,4285714	

Tabela 38: Matriz de confusão correspondente ao modelo MEHC com pós-processamento dinâmico usando o banco de dados *Age*

	/ind/	/n/	/>/	/v/	/i/	/u~/	/z/	/E/	/r/	/o/	/k/	/wa/	/t/	/u/	/e/	/dc/	/oj/	/.pau/	/s/	/tS/	/m/	/a/	/i~/	/ej/	/e~/	soma
/ind/	1499	1	0	12	13	33	0	26	23	0	2	1	107	12	0	0	0	23	0	0	3	1	0	0	12	1768
/n/	12	322	1	0	1	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	1	4	0	0	3	347
/>/	185	34	745	21	0	19	0	8	0	0	3	0	0	0	0	0	0	0	0	0	4	13	0	0	0	1032
/v/	45	6	0	541	5	0	0	0	1	0	0	0	13	11	0	4	0	0	0	0	3	6	0	0	0	635
/i/	33	2	0	12	397	28	0	5	4	0	0	0	10	1	0	0	3	0	0	0	2	0	0	0	12	509
/u~/	12	0	0	27	3	886	0	4	2	0	3	0	0	15	0	0	3	0	0	0	1	2	0	0	3	961
/z/	65	8	0	42	5	0	2554	0	11	0	0	0	0	16	0	0	0	0	24	0	0	3	0	0	0	2728
/E/	391	0	0	0	14	2	0	3356	48	0	0	0	1	0	0	0	1	0	0	0	6	26	0	0	1	3846
/r/	255	4	0	14	0	0	0	56	1023	0	0	0	23	3	0	0	1	0	0	0	2	10	0	0	0	1391
/o/	15	4	1	19	0	43	0	0	8	235	5	0	0	5	0	0	3	0	0	0	0	0	0	0	0	338
/k/	54	3	0	0	0	0	0	0	6	0	375	0	43	24	0	0	0	0	0	0	0	0	0	0	0	505
/wa/	34	9	22	0	0	30	0	0	0	0	1	456	0	2	0	0	0	0	0	0	0	18	0	0	0	572
/t/	17	0	0	2	0	0	0	0	1	0	1	0	301	3	0	0	0	0	0	2	0	0	0	0	3	330
/u/	213	12	0	0	0	78	0	5	26	0	13	0	34	1580	0	0	0	0	0	0	5	3	0	0	0	1969
/e/	42	3	0	0	0	0	0	15	8	0	0	0	23	0	178	0	13	0	0	0	1	2	0	0	12	297
/dc/	34	2	0	0	0	0	0	0	0	0	0	0	1	0	0	207	0	0	0	0	0	0	0	0	0	244
/oj/	146	7	1	1	23	21	0	0	11	0	5	0	32	21	0	0	869	0	0	0	7	2	0	0	0	1146
/.pau/	234	0	0	13	0	0	0	0	3	0	13	0	23	21	0	0	0	2172	0	0	0	1	0	0	4	2484
/s/	3	0	0	1	2	0	0	0	0	0	2	0	0	0	0	0	0	0	2349	0	0	0	0	0	1	2358
/tS/	5	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	232	0	0	0	0	0	240
/m/	19	0	0	0	3	14	0	2	0	0	0	0	0	3	0	0	0	0	0	0	192	0	3	0	2	238
/a/	65	3	0	15	0	3	0	16	12	0	0	14	0	0	0	0	0	0	0	0	0	194	0	0	0	322
/i~/	43	13	0	15	32	38	0	0	2	0	1	0	0	7	0	0	0	0	0	0	2	0	568	0	11	732
/ej/	19	0	0	0	37	0	0	2	4	0	0	0	0	0	0	0	12	0	0	0	0	0	0	662	35	771
/e~/	20	3	0	0	3	14	0	23	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	405	470
Soma	3460	436	770	738	538	1209	2554	3518	1195	235	424	471	611	1725	178	211	905	2195	2373	234	231	285	571	662	504	**Acurácia**
Sensibilidade	84,7850679	92,795389	72,1899225	85,1968504	77,9960707	92,1956296	93,6217009	87,2594904	73,5442128	69,5266272	74,2574257	79,7202797	91,2121212	80,2437786	59,9326599	84,8360656	75,8289703	87,4396135	99,6183206	96,6666667	80,6722689	60,2484472	77,5956284	85,8625162	86,1702128	84,9998094
Precisão	43,3236994	73,853211	96,7532468	73,3062331	73,7918216	73,2837055	100	95,3951109	85,6066946	100	88,4433962	96,8152866	49,2635025	91,5942029	100	98,1042654	96,0220994	98,952164	98,988622	99,1452991	83,1168831	68,0701754	99,474606	100	80,3571429	