Universidade Federal da Bahia
Escola Politécnica
Departamento de Engenharia Elétrica
Programa de Pós-Graduação em Engenharia Elétrica

CONTRIBUIÇÃO AO DESENVOLVIMENTO DE UM MÓDULO AUDITIVO PARA ROBÓTICA ASSISTIVA

Autor: Elmo Alberto Teixeira Borges Junior
Orientadores: Jés J. F. Cerqueira
Eduardo F. Simas Filho

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica, para preencher os requisitos à obtenção do Título de Mestre em Engenharia Elétrica

Salvador - BA, Dezembro, 2020

Elmo Alberto Teixeira Borges Junior

CONTRIBUIÇÃO AO DESENVOLVIMENTO DE UM MÓDULO AUDITIVO PARA ROBOTICA ASSISTIVA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal da Bahia, como parte dos requisitos, para a obtenção do título de Mestre em Engrenharia.

Universidade Federal da Bahia

Departamento de Engenharia Elétrica

Programa de Pós-Graduação em Engenharia Elétrica

Orientador: Jés J. F. Cerqueira

Coorientador: Eduardo F. Simas Filho

Salvador

2020

B732 Borges Junior, Elmo Alberto Teixeira.
Contribuição ao desenvolvimento de um módulo auditivo para robótica assistiva / Elmo Alberto Teixeira Borges Junior. – Salvador, 2020.
83 f.: il. color.

Orientador: Prof. Dr. Jés J. F. Cerqueira.
Coorientador: Prof. Dr. Eduardo F. Simas Filho.

Dissertação (mestrado) – Universidade Federal da Bahia. Escola Politécnica, 2020.

1. Redes neurais artificiais. 2. Emoções - reconhecimento. 3. Descritores. 4. Robótica. I. Cerqueira, Jés J. F. II. Simas Filho, Eduardo F. III. Universidade Federal da Bahia. IV. Título.

CDD: 621.382

Elmo Alberto Teixeira Borges Junior

CONTRIBUIÇÃO AO DESENVOLVIMENTO DE UM MÓDULO AUDITIVO PARA ROBOTICA ASSISTIVA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal da Bahia, como parte dos requisitos, para a obtenção do título de Mestre em Engrenharia.

Trabalho aprovado. Salvador, 08 de Dezembro de 2020:

Prof. Dr. Jés de Jesus Fiais Cerqueira
Orientador - UFBA

Prof. Dr. Eduardo F. Simas Filho
Coorientador - UFBA

Prof. Dr. Tiago Trindade Ribeiro
Avaliador Interno - UFBA

Prof. Dr. Jugurta R. Montalvão Filho
Avaliador Externo - UFS

Salvador
2020

AGRADECIMENTOS

Agradeço a todas as pessoas que contribuíram para que chegasse ao fim desta jornada. Entre elas destaco: minha mãe, Vera Lúcia dos Santos Borges, pela paciência, amor, compreensão; meu pai, Elmo Alberto Teixeira Borges, pelo amor e ensinamentos sobre a vida; minha esposa, Lubiana de Oliveira Silva Borges, pela amizade, companhia, afeto e amor. Agradeço aos amigos e colegas que fiz em todas as instituições de ensino por onde passei durante minha vida acadêmica. Agradeço aos mestres que passaram em minha vida pelos ensinamentos. Agradeço ao meu orientador, Jés Cerqueira por aceitar compartilhar seu conhecimento, experiência e pela orientação em si, ao coorientador Eduardo Furtado de Simas Filho, pela orientação desde a época da graduação e pelo apoio no projeto. Por fim, agradeço a todos os outros que fizeram e fazem parte de minha vida, muito obrigado!

RESUMO

A voz é uma importante ferramenta de interação entre seres humanos e entre homens e máquinas. Neste contexto, este trabalho consiste na aplicação de técnicas de pós-processamento de descritores extraídos do sinais de voz, como análise de componentes principais e análise de componentes independentes, com o objetivo de redução do número de parâmetros e aumento na eficiência de um sistema automático para reconhecimento de locutores e emoções, por meio do sinal de voz, baseado em classificadores neurais artificiais, rasos e profundos, como contribuição ao desenvolvimento de um módulo auditivo para um robô de assistência. Inicialmente são extraídos descritores característicos de sinais de voz, como os coeficientes cepstrais na frequência mel e frequência fundamental (*pitch*), a partir dos arquivos de áudio utilizados. Após essa etapa, são aplicadas análise de componentes principais e análise de componentes independentes aos vetores de atributos apresentados a entrada da rede neural *shallow*. A fim de comparação entre classificadores, também, são utilizadas redes neurais recorrentes profundas com células de memória LSTM e BLSTM. Os resultados obtidos mostram que a utilização de técnicas de processamento estatístico de sinais auxiliam no aumento da eficiência do classificadores neurais artificiais podendo ser utilizadas para a tarefa de reconhecimento automático de emoções e locutores alcançando uma eficiência de discriminação máxima de 98%, para os locutores, e 92%, para as emoções.

Palavras-chave: Redes Neurais Artificiais, Extração de Descritores, MFCC, Reconhecimento de Emoções, Reconhecimento Automático de Emoções, Análise de Componentes Independentes.

ABSTRACT

The voice is an important tool for interaction between human beings and machines. In this context, this work consists of applying post-processing techniques for descriptors extracted from voice signals, such as principal component analysis and independent component analysis, with the objective of reducing the number of parameters and increasing the efficiency of an automatic system, for speaker and emotion recognition, through the voice signal, based on artificial neural classifiers, shallow and deep, as a contribution to the development of an auditory module for an assistance robot. Initially, characteristic descriptors of voice signals are extracted, such as mel cepstral coefficients and fundamental frequency (pitch), from the available audio files. After this step, the principal component analysis and the independent component analysis are applied to the attribute vectors presented at the input of the shallow neural network. In order to compare classifiers, also, deep recurrent neural networks with LSTM and BLSTM memory cells are also used. The results obtained show that the use of statistical signal processing techniques help to increase the efficiency of the artificial neural classifiers and can be used for an automatic task of recognition of emotions and speakers, reaching a maximum discrimination efficiency of 98%, for the speakers , and 92%, for emotions.

Keywords: Artificial Neural Networks, Descriptor Extraction, MFCC, Automatic Speaker Recognition, Automatic Emotion Recognition, Independent Component Analysis.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplos de robôs de assistência. Fonte: Autor. Baseado em RobotLAB (2020), TheGuardian (2020) e RIKEN (2020). 6
Figura 2 – Diagrama das informações contidas no sinal de voz. Fonte: Autor. Baseado em Oliveira (2018a). 7
Figura 3 – Diagrama das etapas iniciais para obtenção de descritores do sinal de voz. Fonte: Autor baseado em Togneri e Pullella (2011). 9
Figura 4 – Diagrama de obtenção dos MFCC. Fonte: Autor. 9
Figura 5 – Diagrama de blocos da PCA. Fonte: Autor. 12
Figura 6 – Modelo de um neurônio. Fonte: (HAYKIN et al., 2009). 17
Figura 7 – Modelo adaptado de um neurônio. Fonte: (HAYKIN et al., 2009). . . . 18
Figura 8 – Rede Neural *Feedforward* com uma única camada de neurônios. Fonte: (HAYKIN et al., 2009). 20
Figura 9 – Rede Neural *Feedforward* com uma camada de neurônios oculta. Fonte: (HAYKIN et al., 2009). 21
Figura 10 – Modelo adaptado de uma rede recorrente simples (SRN). Fonte: Autor (baseado em Haykin et al. (2009)). 22
Figura 11 – Modelo de uma rede neural multicamadas. Fonte: Autor. 22
Figura 12 – Exemplo de uma rede neural profunda. Fonte: Autor. 23
Figura 13 – Exemplo em grafos de uma RNN. Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016). 24
Figura 14 – Matriz de confusão para duas classes. Fonte: Autor. 26
Figura 15 – Diagrama de blocos do sistema de classificação de emoções e locutor. Fonte: Autor. 28
Figura 16 – Composição do *dataset Voxceleb1*. Fonte: Adaptado de Nagrani, Chung e Zisserman (2017) 29
Figura 17 – Diagrama representativo do sistema de validação cruzada com 10 folds. Fonte: Autor 31
Figura 18 – Diagrama de blocos simplificado do processo de ajuste de parâmetros do módulo de identificação do locutor. Fonte: Autor 32
Figura 19 – Diagrama de blocos simplificado do processo de ajuste de parâmetros do módulo de identificação de emoções. Fonte: Autor 35
Figura 20 – Variação no número de descritores para formação do vetor **vt-1**. Fonte: Autor 38
Figura 21 – Variação no número de nerônios na camada oculta utilizando o **vt-1**. Fonte: Autor 39

Figura 22 – Curva de carga da PCA para o vetor **vt-1** extraído dos sinais do banco (OLIVEIRA, 2018a). Fonte: Autor 41
Figura 23 – Matriz de correlação para os vetores de parâmetros do banco (OLIVEIRA, 2018a). Fonte: Autor 42
Figura 24 – Valores médios das componentes independentes para cada locutor. Fonte: Autor 42
Figura 25 – Curva de carga da PCA para sinais do banco Voxceleb1. Fonte: Autor . 43
Figura 26 – Matriz de correlação para os parâmetros antes da aplicação da ICA para sinais do Banco de Dados Voxceleb1. Fonte: Autor 43
Figura 27 – Curva de carga da PCA para sinais do *dataset* SAVEE. Fonte: Autor . 49
Figura 28 – Matriz de correlação antes da ICA para emoções do banco de dados SAVEE . Fonte: Autor. 49
Figura 29 – Valores médios das componentes independentes para cada Emoção no banco de dados SAVEE . Fonte: Autor. 49
Figura 30 – Curva de carga da PCA para sinais do banco Emo-DB. Fonte: Autor . 50
Figura 31 – Matriz de correlação para os parâmetros antes da ICA para emoções do banco de dados Emo-DB. Fonte: Autor 50
Figura 32 – Valores médios das componentes independentes para cada Emoção no banco de dados Emo-DB . Fonte: Autor. 51

LISTA DE TABELAS

Tabela 1 – Distribuição dos arquivos de áudio na base de dados SAVEE. 30
Tabela 2 – Distribuição dos arquivos de áudio na base de dados Emo-DB. 30
Tabela 3 – Descritores e parâmetros propostos para o classificador MLP. 31
Tabela 4 – Parâmetros de treinamento da rede neural MLP. 33
Tabela 5 – Parâmetros de treinamento da rede neural profunda. 34
Tabela 6 – Configurações das redes neurais LSTM e BLSTM testadas. 35
Tabela 7 – Tabela de correspondência entre siglas e emoções para os bancos de dados SAVEE e Emo-DB. 36
Tabela 8 – Produto das Eficiências (PEf), em %, da rede neural MLP para os vetores **vt-1** e **vt-2**. 38
Tabela 9 – Produto das Eficiências (PEf), em %, da rede neural MLP para os vetores de características 1 e 2. 39
Tabela 10 – Produto das Eficiências (PEf),em %, para a rede neural MLP em diferentes base de dados para o vetor **vt-1**. 40
Tabela 11 – Produto das Eficiências (PEf), em %, para a rede neural MLP em diferentes base de dados para o vetor **vt-2**. 40
Tabela 12 – Matriz de confusão (em %) para uma rede neural MLFN com 70 neurônios na camada oculta para o banco de dados Oliveira (2018a). . 40
Tabela 13 – Produto das Eficiências médias (PEf), em %, para uma rede neural MLP com e sem aplicação de pré-processamento estatístico ao vetor de entrada. 44
Tabela 14 – Configurações das redes LSTM e BLSTM. 44
Tabela 15 – Valores médios (em %) do PEf para os bancos de dados (OLIVEIRA, 2018a) e Voxceleb1 para a configuração t-RNN 1 e vetores de caracteristicas **vt-1** e **vt-2**. 45
Tabela 16 – Matriz de confusão (valores em %) para rede com banco de dados (OLIVEIRA, 2018a). 45
Tabela 17 – Comparação do PEF, em %, para os diversos classificadores. 46
Tabela 18 – Comparação dos tempos (em segundos) de processamento para os diversos classificadores. 47
Tabela 19 – Produto das Eficiências (em %) para a rede neural MLP em diferentes base de dados e vetores de características. 48
Tabela 20 – Matriz de confusão do banco de dados SAVEE para uma rede neural MLP com 70 neurônios na camada escondida e **vt-1**. (Valores em %) . 48
Tabela 21 – Produto das Eficiências máximas (PEf), em %, para rede MLP com e sem pré-processamento estatístico. 51

Tabela 22 – Matriz de confusão (valores em %) do banco de dados SAVEE e pré-processamento por P-ICA. 52
Tabela 23 – Matriz de confusão (valores em %) do banco de dados Emo-DB e pré-processamento por P-ICA. 52
Tabela 24 – Valores médios, em %, do PEf para a rede neural recorrente t-RNN 1 e os *dataset* SAVEE e Emo-DB. 53
Tabela 25 – Matriz de confusão (valores em %) dos resultados para o banco de dados Emo-DB para uma rede **t-RNN 1**. 53
Tabela 26 – Comparação dos valores médios do PEf, em %, para os *dataset* SAVEE e Emo-DB. 54
Tabela 27 – Comparação dos tempos de processamento (em segundos) para os diversos classificadores. 54

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Networks*
AC	Acurácia de um classificador
Adam	*Adaptive moment estimation*
ASR	*Automatic Speech Recognition*
CNN	*Convolutional Neural Network*
P-ICA	*Independent Component Analysis with compaction*
DL	*Deep Learning*
DNN	*Deep Neural Network*
DSP	*Digital Signal Processing*
DSCNN	*Deep Stride Convolutional Neural Network*
E	Emoção
Emo-DB	*Berlin Emotion Database*
Emo	Emoção
GMM	*Gaussian Mixture Models*
HMM	*Hidden Markov Models*
IA	Inteligência Artificial
ICA	*Independent Component Analysis*
IEMOCAP	*Interactive Emotional Dyadic Motion Capture*
RAVDESS	*Ryerson Audio-Visual Database of Emotional Speech and Song*
KNN	*k-Nearest Neighbor*
L	Locutor
Loc	Locutor
LPC	*Linear Predictor Coefficients*

LVQ	*Learning Vector Quantization*
LSTM	*Long Short Term Memory*
BILSTM	*Bidirectional Long Short Term Memory*
MEDC	*Mel-Energy spectrum Dynamic Coefficients*
MCAF	*Mel Cepstral Affinity Features*
MFCC	*Mel Frequency Cepstral Coefficients*
MFSC	*Mel Frequency Spectrum Coefficients*
MMResLSTM	*Mutimodal Residual LSTM*
ML	*Machine Learning*
MLFN	*Multi Layer Feedforward Network*
MLP	*Multilayer Perceptron*
MPSoC	*Multiprocessor System-on-Chip*
MSE	*Mean Squared Error*
NCF	*Normalized Correlation Function*
PCA	*Principal Component Analysis*
p-CNN	*pré-treined Convolutional Neural Network*
PEF	média geométrica das eficiências do classificador
RAL	Reconhecimento Automático do Locutor
RAS	Robótica Assistiva Social
REV	Reconhecimento de emoções através da voz
Rprop	*Resilient Backpropagation*
RNA	Redes Neurais Artificiais
ResNet	*Residual Neural Network*
RNN	*recurrent neural network*
SAVEE	*Surrey Audio-Visual Expressed Emotion*
SAR	*Socially Assistive Robotics*

SER	*Speech Emotion Recognition*
SLFN	*Single Layer Feedforward Network*
SOM	*Self-Organizing Map*
SRS	*Speaker Recognition System*
SVM	*Suport Vector Machine*
SGD	*Stochastic Gradient Descent*
t-RNN	*treined Recurrent Neural Network*
TDNN	*Time-Delay Neural Networks*
vt-1	Vetor de atributos do sinal de voz composto pelos MFCC, Delta, Delta Delta e *Pitch*
vt-2	Vetor de atributos do sinal de voz composto pelos MFCC, Delta e Delta Delta

LISTA DE SÍMBOLOS

α	Constante do filtro passa-alta
Δ	Coeficientes delta
$\Delta c_j(n)$	Coeficientes delta de ordem n
$\Delta\Delta$	Coeficientes delta delta
$\Delta\Delta c_j(n)$	Coeficientes delta delta de ordem n
Γ	Função de mapeamento
φ	Função de ativação
μ	Vetor com MFCC, delta e delta-delta
η	Fator da *Resilient Backpropagation*
θ	Probabilidade
ϑ	Coeficiente limiar de um neurônio
ξ	Potencial de um neurônio
ω	Coeficiente de peso de um neurônio
a_{ij}	Elemento de $\mathbf{A}$
b_k	Bias do neurônio k
C	*Mel Frequency Cepstral Coefficients*
F	Frequência em Hertz
f	Função (genérica)
$f(\xi)$	Função sigmoide
f_0	frequência fundamental
MSE	*Mean squared error*
t	Tempo em segundos
K	Número de filtros mel

L	Função de perda que mede o erro da saída
X	Sinal de áudio
o	Erro da saída da rede neural recorrente
y	Saída alvo da rede neural recorrente
x	Coeficiente de viés de um neurônio
$w(n)$	Janela de *Hamming*
Y	*Fast Fourier Transform*
$y(t)$	Sinal filtrado
$x(t)$	Sinal de entrada
x_j	Vetor sinal de entrada
n	Número de amostras de cada *frame*
$X(k, n)$	Espectro do sinal
K	Número de filtros do banco mel
s	Tempo em segundos.
N	Número de amostras em cada *frame* do sinal
f_{hz}	Frequência do sinal em Hertz
f_{mel}	Frequência do sinal na escala mel
$\hat{f}$	Função que mapeia x em θ
$c_j(n)$	Coeficientes mel cepstrais de ordem n
u_k	Combinador linear do neurônio k
y_k	Saída do neurônio k
w_{kj}	Pesos sinápticos do neurônio k
v_k	Potencial de ativação do neurônio k
EF_i	Eficiência da classificação obtida para a classe i
z^{-1}	Função de transferência que aplica um retardo unitário ao sinal de entrada.

$\mathbf{U}$	Matriz de pesos sinápticos de um modelo de rede neural
$\mathbf{W}$	Matriz de pesos sinápticos de uma rede neural recorrente
$\mathbf{V}$	Matriz de pesos sinápticos de um modelo de rede neural recorrente

SUMÁRIO

1 INTRODUÇÃO 1
1.1 Justificativa 2
1.2 Objetivo 3
1.2.1 Objetivos Gerais 3
1.2.2 Objetivos Específicos 3

2 REVISÃO BIBLIOGRÁFICA 4
2.1 Robótica de Assistência 5
2.2 Reconhecimento de Locutor Através da Voz 6
2.3 Reconhecimento de Emoções através da voz 8
2.4 Extração de descritores do sinal de voz 8
2.4.1 Coeficientes Cepstrais de Frequência na Escala Mel 9
2.4.2 Coeficientes *Delta* e *Delta Delta* 11
2.5 Frequência Fundamental (*Pitch*) 11
2.6 Processamento Estatístico 11
2.6.1 Análise de Componentes Principais 12
2.6.2 Análise de Componentes Independentes 13
2.7 Aprendizado de Máquina 15
2.7.1 Redes Neurais Artificiais 15
2.7.2 Arquitetura de redes 19
2.7.3 Redes Neurais Multicamadas 19
2.7.4 Algoritmos de treinamento 20
2.8 Aprendizado Profundo 23
2.8.1 Redes Neurais Recorrentes 24
2.9 Parâmetros utilizados para avaliação do sistema 25
2.9.1 Acurácia 25
2.9.2 Produto das Eficiências 25
2.9.3 Matriz de confusão 25

3 METODOLOGIA 27
3.1 Modelo Proposto 27
3.2 Base de dados 28
3.3 Vetor de características utilizado 30
3.4 Validação Cruzada 30
3.5 Identificação do locutor 32

3.5.1 Classificador Neural MLP . . . 33
3.5.2 Classificador Neural Profundo . . . 34
3.6 Identificação de Emoções . . . 34

4 RESULTADO . . . 37
4.1 Identificação do Locutor . . . 37
4.1.1 Classificador Neural Artificial MLP . . . 37
4.1.1.1 Variação do número de descritores . . . 37
4.1.1.2 Variação do número de neurônios na camada oculta . . . 39
4.1.1.3 Avaliação dos descritores em diferentes bases de dados . . . 39
4.1.1.4 Redes Neurais MLP com pré-processamento estatístico . . . 41
4.1.2 Classificador Neural Artificial Profundo . . . 44
4.1.2.1 Análise dos resultados . . . 45
4.2 Identificação de Emoções . . . 47
4.2.1 Classificador MLP . . . 47
4.2.1.1 Redes Neurais MLP com pré-processamento estatístico . . . 48
4.2.2 Classificador Neural Artificial Profundo . . . 52
4.2.3 Análise dos Resultados . . . 53

5 CONCLUSÕES . . . 55

Referências . . . 57

ANEXOS 62

ANEXO A – ARTIGO DO AUTOR . . . 63

1 INTRODUÇÃO

A robótica de assistência é uma área da robótica que estuda, projeta e cria dispositivos que auxiliam e dão suporte a serem humanos. Áreas de interesse em robótica assistiva incluem robôs de reabilitação, cadeira de rodas robótica e outras formas de mobilidade, robôs companheiros, braços manipuladores para deficientes físicos e robôs educacionais (FEIL-SEIFER; MATARIC, 2005). Por exemplo, no Brasil, segundo o IBGE (Instituto Brasileiro de Geografia e Estatística), em 2030, aproximadamente 24% da população possuirá algum grau de deficiência e 7% da população terá mais de 64 anos. Nesse contexto, uma área que apresenta bastante interesse é a robótica assistiva social *(Socially Assistive Robotics - SAR)*. A robótica assistiva social (RAS) objetiva a construção de dispositivos dotados de capacidades para fornecer assistência às pessoas por meio de interação social, por exemplo, criando meios para que crianças com transtornos do espectro autista possam interagir com outras crianças.

A interação através da voz é uma das principais formas de comunicação entre os seres humanos. Esta forma de comunicação é tão versátil que nem sempre é feita com o uso de vocábulos. Assim, quando um módulo de audição artificial é projetado ele deve ser capaz de realizar algumas tarefas que os seres humanos realizam naturalmente através da audição (OLIVEIRA, 2018a). Logo, um robô de assistência que seja capaz de interpretar informações transmitidas através da voz possui uma grande ferramenta de interação com seres humanos. Neste sentido, no processo de interação homem-máquina utilizando informações da fala, algumas tarefas são particularmente importantes, entre elas: identificar o locutor (SILVA et al., 2015) e reconhecer o estado afetivo da pessoa (DAHAKE; SHAW; MALATHI, 2016).

A partir dos avanços tecnológicos que possibilitaram a montagem de sistemas complexos em um único chip, a produção em larga escala e o barateamento dos dispositivos eletrônicos, atualmente é possível construir módulos para robôs de assistência com baixo custo. Esses robôs em geral, são compostos de vários módulos que podem ser construídos a partir de sistemas embarcados MPSoCs (*Multiprocessor System-on-Chip*) ou de computadores montados em uma única placa.

Diversos estudos tem sido realizados nessas áreas: Schueler, Silveira e Cataldo (2018) recorreram a características extraídas do sinal glotal para treinar o modelo de reconhecimento de locutor utilizando HMM *(Hidden Markov Models)*; Aouani e Ayed (2018) utilizaram os MFCCs *(Mel Frequency Cepstral Coefficients)* como descritores do sinais de áudio, *Support Vector Machine*, *Deep Support Vector Machine* e *auto-encoders* como classificadores para a tarefa de reconhecimento de emoções; Oliveira, Cerqueira e

Filho (2018b) propuseram um módulo auditivo artificial multidispositivo para a realização das tarefas de identificação de locutor, identificação da fonte sonora, e reconhecimento do estado afetivo; Silva et al. (2015) empregaram coeficientes cepstrais na escala mel, um preditor LPC (*Linear Predictor Coding*) para a predição dos MFCC, coeficientes DELTA e coeficientes DELTA DELTA como descritores da fala e modelagem do locutor.

No desenvolvimento de sistemas automáticos para utilização em plataformas modulares para robótica de assistência, parâmetros como a dimensão do vetor de características apresentado ao classificador e a eficiência de discriminação entre classes são muito relevantes. Este trabalho consiste na utilização de técnicas de pós-processamento de sinais, *Principal Component Analysis* (PCA) e *Independent Component Analysis* (ICA), e de *Deep Learning* como contribuições ao desenvolvimento de um módulo auditivo artificial para utilização em um robô socialmente assistivo tendo como base o estudo, a investigação, o levantamento bibliográfico sobre robótica assistiva, técnicas de extração de descritores de sinais de voz, Redes Neurais Artificiais Rasas e Redes Neurais Artificiais Profundas para fins de comparação. A pesquisa bibliográfica é utilizada para capturar o estado da arte sobre o reconhecimento Automático do Locutor Através da Voz (RAL), o reconhecimento de emoções através da voz (REV), a extração de descritores do sinal de voz, o aprendizado profundo, redes neurais artificiais e redes neurais artificiais profundas.

Este documento está organizado conforme descrito a seguir: O capítulo 2 trata do referencial teórico, trazendo conceitos sobre processamento digital de sinais, extração de características de sinais de voz, sistemas de classificação e aprendizado de máquina alicerçado no estudo bibliográfico sobre o assunto. O capítulo 3 traz a metodologia utilizada no desenvolvimento da dissertação. O capítulo 4 apresenta as resultados alcançados. O capítulo 5 mostra as conclusões acerca dos temas abordados no trabalho.

1.1 JUSTIFICATIVA

Os seres humanos possuem como principais ferramentas sensoriais a visão, a audição, o olfato, o tato e o paladar. Já nas relações sociais a comunicação verbal ou oral é essencial. Por meio da voz é possível determinar a localização do falante, com quem se fala, o que se fala e até mesmo o que o falante está sentindo. Robôs de assistência com a habilidade de processar os sinais de áudio oriundos do trato vocal de forma similar ao sistema auditivo humano podem exercer uma interação homem-rôbo de forma mais natural, auxiliando nas tarefas cotidianas de forma mais eficaz.

Máquinas socialmente assistivas que possuam um módulo que consiga realizar essa interação de uma forma natural para o usuário são importantes. Um robô de assistência capaz de reproduzir aspectos da audição humana, como: localização, distinção, sensibilidade e comunicação, podem ser utilizado em diversas áreas da robótica de assistência, como

no auxílio de idosos, pacientes em recuperação em hospitais ou em residências, crianças com transtorno do espectro autistas, aumentando o sucesso das terapias e diminuindo os custos de internações. Essa não é uma tarefa simples, pois o módulo auditivo deve possuir tamanho e peso reduzidos, a fim de facilitar o embarque em uma plataforma robótica, com grande capacidade de processamento de informações e possuir sensores para captura dessas informações no ambiente.

1.2 OBJETIVO

Este trabalho, quanto a finalidade, pode ser dividido em: objetivos gerais e objetivos específicos.

1.2.1 Objetivos Gerais

Este trabalho objetiva o estudo de técnicas de processamento de sinais e classificação de padrões para identificação de emoções e falantes por meio do sinal de voz. Dessa forma, as técnicas utilizadas poderão compor um módulo auditivo utilizado como parte de um robô socialmente assistivo que seja capaz de realizar a identificação de locutor e emoção através do sinal de voz.

1.2.2 Objetivos Específicos

Os objetivos específicos desse trabalho estão descritos abaixo:

1. Realizar estudos e extrair do sinal de voz os coeficientes MFCC *(Mel Frequency Cepstral Coefficient)*, coeficientes Delta e coeficientes Delta Delta;
2. Realizar estudos e identificar o locutor através do sinal de voz utilizando técnicas de processamento estatístico (PCA e ICA), redes neurais artificiais *feedfoward* com uma camada oculta e rede neurais artificiais profundas comparando os resultados obtidos;
3. Realizar estudos e identificar a emoção através do sinal de voz utilizando técnicas de processamento estatístico (PCA e ICA), redes neurais artificiais *feedfoward* com uma camada oculta e rede neurais artificiais profundas comparando os resultados obtidos;

2 REVISÃO BIBLIOGRÁFICA

Essa dissertação de mestrado aborda diversos assuntos, entre eles destacam-se: reconhecimento de emoções na fala *(Speech Emotion Recognition - SER)*, reconhecimento automático de locutor através da voz *(Automatic Speaker Recognition - ASR)*, Redes Neurais Artificiais *(Artificial Neural Networks - ANN)*, Aprendizagem Profunda *(Deep Learning - DL)*, Aprendizagem de Máquina *(Machine Learning - ML)*, Processamento Digital de Sinais de áudio *(Digital Signal Processing - DSP)*, Descritores de sinais de voz, como contribuição ao desenvolvimento de um módulo auditivo para robótica de assistência. Este capítulo tem como objetivo mostrar o estado da arte dos temas em epígrafe, justificando a abordagem utilizada nesta dissertação.

Neste contexto, diversos trabalhos têm tratado de reconhecimento automático de locutores e reconhecimento automático de emoções através da voz. Em Campos (2017) utilizou-se Coeficientes Cepstrais de Frequência Mel *(Mel-frequency Cepstral Coefficients - MFCCs)* e Coeficientes de Predição Linear Perceptual *(Linear Predictor Coefficients - LPCs)* como descritores do sinal de voz e K-Vizinhos mais próximos *(k-Nearest neighbor - KNN)* como técnica de classificação. Em Mafra (2002) foi proposto o reconhecimento automático de locutor independente do texto por meio de extração de descritores do sinal de voz e classificação de uma rede neural de arquitetura *Self-Organizing Map (SOM)*. Em NERI (2019) empregou-se um sistema de segmentação de locutores baseado em *Mel Cepstral Affinity Features (MCAF)*, como descritores da fala. Em Oliveira (2018a), também, foram usadas técnicas de extração de características da voz, como os Coeficientes Cepstrais de Frequência Mel, e redes neurais artificiais para reconhecimento automático do locutor e emoções. No trabalho Jagiasi et al. (2019) implementou-se um sistema de reconhecimento do locutor independente do texto baseado em *Frequency Cepstral Coefficients*, *Deep Neural Network* e *Convolucional Neural Network*. Em Toruk e Gokay (2019) abordou-se o problema de reconhecimento de locutor através de frases curtas por meio da utilização de *Time-Delay Neural Networks (TDNN)*. Em Bonilla, Nedjah e Mourelle (2015), para uma tarefa de reconhecimento automático de fala *(automatic speech recognition - ASR)*, utilizou MFCCs como atributos de entrada para redes neurais profundas *(Deep Neural Network - DNN)* e redes *MLP (Multilayer Perceptron)*, comparando os resultados obtidos. Já Abdel-Hamid et al. (2014) utilizou redes neurais convolucionais *(Convolutional Neural Network - CNN)*, incluindo redes pré-treinadas, e extração de atributos baseadas na energia logarítmica para a tarefa de ASR.

Em relação ao tema de reconhecimento automático de emoções através do sinal de voz podem-se citar diversos trabalhos como: Oliveira (2018a), valeu-se de técnicas de

extração de atributos do sinal de voz como os MFCC, *DELTA* e *DELTA DELTA*, redes neurais artificiais (RNA) para reconhecimento automático das emoções, utilizando as bases de dados de áudio *Berlin Database of Emotional Speech (Emo-DB)* e *Surrey Audio-Visual Expressed Emotion (SAVEE)*. Em Badshah et al. (2017) utilizaram-se espectrogramas gerados por meio dos sinais de fala e redes neurais artificiais profundas como método para reconhecimento de emoções a partir do sinal de voz. Nesse trabalho o *dataset* utilizado foi o *(Emo-DB)*. Em Likitha et al. (2017) foram empregados MFCC, suas médias e desvio padrão para reconhecimento de emoções de um falante através de sua voz. Em Kwon et al. (2020) aplicou-se uma nova abordagem especialmente desenvolvida para a discriminação de espectrogramas por meio de um arquitetura de redes neurais convolucionais chamada *Deep Stride Convolutional Neural Network (DSCNN)* valendo-se das base de dados *Interactive Emotional Dyadic Motion Capture (IEMOCAP)* e *Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. Em Ma et al. (2019) os autores utilizaram redes LSTM residuais mutimodais (MMResLSTM) para o reconhecimento de emoções obtendo resultados promissores.

Dentre os autores citados acima, somente em Oliveira (2018a) os temas de reconhecimento automático de locutor e emoções por meio do sinal de voz foram abordados conjuntamente. Nos trabalhos que tratam exclusivamente do reconhecimento de locutor, observa-se que a maioria utilizou os MFCCs como descritores do sinal de voz, mas diferentes técnicas de classificação. Já para o reconhecimento de emoções, os trabalhos mostrados utilizam MFCC ou espectrogramas como representações dos sinais de voz e principalmente os *dataset* (conjunto de dados) Emo-DB e SAVEE para projeto do sistema de classificação.

Como contribuições, essa dissertação utiliza descritores de sinais de voz baseados em análise tonal, como o *pitch*, para o reconhecimento de emoções e a utilização de técnicas de *Blind Source Separation*(BSS)(Separação Cega de Fontes) para redução de dimensão dos vetores apresentados aos sistemas de classificação e aumento de eficiência do classificador. Outros pontos importantes são: a abordagem dos temas de reconhecimento automático de locutor e emoções de forma conjunta, com a utilização de bancos de dados de vozes extraídas de conversas em ambiente não controlado, auferindo maior robustez ao classificador, de modo que, futuramente, possa ser implementados em um módulo único.

Os tópicos que seguem tem como objetivo apresentar conceitos importantes ao melhor entendimento sobre os assuntos abordados nesse trabalho, para tanto introduzir-se-á assuntos importantes ao tema.

2.1 ROBÓTICA DE ASSISTÊNCIA

Os sistemas robóticos são dispositivos criados para auxílio ou substituição dos seres humanos em tarefas repetitivas ou perigosas. Os robôs são bastantes comuns em nossa

sociedade, são encontrados em linhas de montagens, em atividades em ambientes inóspitos (altas temperaturas, altas pressões, presença de atmosfera perigosa a vida), em buscadores de informações, em centrais de atendimento entre outras aplicações.

A robótica de assistência é uma subárea da robótica que inclui robôs de reabilitação, robôs de cadeira de rodas, robôs companheiros e robôs educacionais. Segundo Feil-Seifer e Mataric (2005), uma importante motivação para Robótica de Assistência (RAS) é o menor risco na interação humano robô devido à ausência de contato, por isso, esses sistemas são mais facilmente testados e implantados. A Figura 1 mostra exemplos de robôs de assistência utilizados em diversas aplicações.

Esses robôs, em geral, são compostos de vários módulos que podem ser construídos a partir de sistemas baseados em microprocessadores *MPSoCs (Multiprocessor System-on-Chip)*. Os benefícios desses sistemas que o tornaram tão populares são: possibilidade de atualização do software (programa), também denominado de *firmware*, sem a necessidade de modificação do *hardware*, possibilidade de utilização do mesmo dispositivo (*hardware*) para aplicações com diferentes programas, melhoria de funcionalidades e correção de erros a partir da atualização do programa, local ou remotamente.

Figura 1: Exemplos de robôs de assistência. Fonte: Autor. Baseado em RobotLAB (2020), TheGuardian (2020) e RIKEN (2020).

2.2 RECONHECIMENTO DE LOCUTOR ATRAVÉS DA VOZ

O reconhecimento biométrico é uma importante forma de reconhecimento do indivíduo. A impressão digital é largamente utilizada para esta tarefa, mas outras formas de identificação mais naturais, como voz e face, são também amplamente empregadas. Segundo Cai, Cai e Li (2018), em geral, a fala contém, além de informações lexicais, atributos paralinguísticos como idioma, canal, emoção. Na Figura 2 é possível observar um diagrama representativo das informações contidas no sinal de voz.

Os sistemas de reconhecimento de locutores utilizam características da voz para identificar indivíduos. Esses sistemas são baseados na modelagem da voz dos locutores

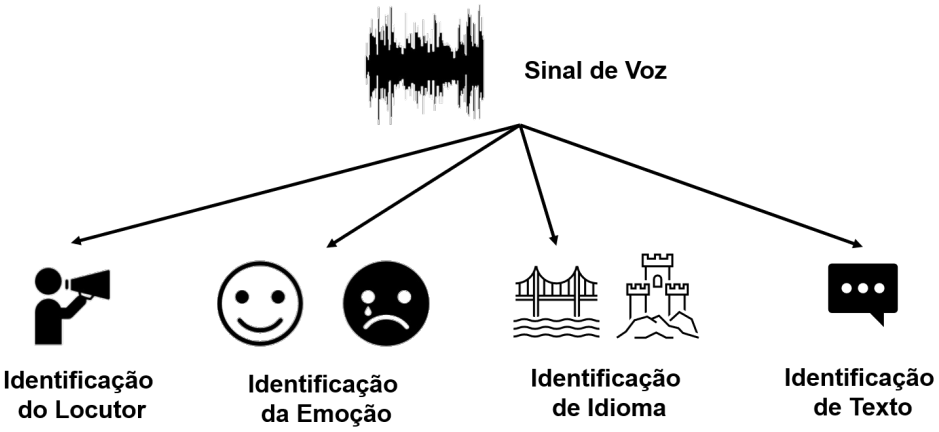

Figura 2: Diagrama das informações contidas no sinal de voz. Fonte: Autor. Baseado em Oliveira (2018a).

e são divididos em diversas tarefas, entre elas: Identificação do locutor,verificação do locutor e detecção de fala (TORUK; GOKAY, 2019). Na identificação do locutor, o sinal de voz de um indivíduo é usado para identificar quem é esse indivíduo através de características extraídas do sinal vocal. Em geral, o sistema de reconhecimento de locutor opera em duas fases: treino e teste. Na primeira fase, o discurso associado ao falante é apresentado ao sistema de reconhecimento para o treinamento, na segunda o sistema treinado e apresentado a sinais desconhecidos para identificação. A métrica de desempenho utilizada em tais sistemas é a taxa de identificação (porcentagem média de identificação correta dos falantes) (TOGNERI; PULLELLA, 2011). Na etapa de teste, uma amostra de discurso é apresentada ao sistema de classificação que define se a amostra pertence ou não a um determinado locutor (IRIYA, 2014).

Outra forma de classificar os sistemas de reconhecimento de voz é dividindo-os em: sistemas dependentes de texto e sistemas independentes de texto. Em sistemas dependentes de texto as frases de reconhecimento são fixas. Em sistemas independentes de texto, não há restrições nas palavras que os falantes podem usar (OLIVEIRA, 2018a). Para a utilização de sinais de voz em sistemas de reconhecimento do locutor é necessária a estimação de descritores capazes de carregar informações discriminantes de cada locutor, auxiliando a tomada de decisão do sistema de classificação. Nesse contexto, descritores que realizam a análise do espectro do sinal de voz são amplamente utilizados na literatura, dentre estes se destacam pelos resultados: Coeficientes de Predição Linear (LPC), Coeficientes Cepstrais com espaçamento de frequências Mel (MFCC) (CAMPOS, 2017) (OLIVEIRA, 2018a) (SCHUELER; SILVEIRA; CATALDO, 2018) e Coeficientes de Predição Linear Perceptual (*Perceptual Linear Prediction (PLP)*) (TOGNERI; PULLELLA, 2011).

2.3 RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA VOZ

De acordo Cai, Cai e Li (2018) a emoção também é uma informação contida no sinal de voz. Como no caso do reconhecimento de locutor, o reconhecimento de emoções por meio do sinal vocal é uma tarefa de processamento de sinal, extração de características discriminantes e classificação (IRIYA, 2014). Nesse contexto, técnicas de identificação do locutor podem ser utilizadas para reconhecimento de emoções. Segundo Cowie et al. (2001), as características de destaque no reconhecimento de emoções no discurso são: Nível de voz e *Pitch* de Voz. Alguns trabalhos utilizam os MFCC como componentes do vetor de características apresentadas ao sistema de classificação, pode-se destacar nesse sentido os trabalhos de Iriya (2014), Dahake, Shaw e Malathi (2016), Likitha et al. (2017), pelos resultados teóricos, e Oliveira (2018a) tanto pelos resultados teóricos alcançados quanto pela realização de testes práticos.

2.4 EXTRAÇÃO DE DESCRITORES DO SINAL DE VOZ

A fim de realizar o reconhecimento do locutor e das emoções contidas no sinal de voz é necessário extrair informações discriminantes de modo a facilitar classificação. Para tanto, serão estudadas características da voz amplamente utilizadas na literatura em sistemas automáticos de reconhecimento de vozes e emoções.

Segundo Togneri e Pullella (2011), o processo básico em todas as formas de reconhecimento de falantes e de discurso é a extração de vetores de características uniformemente espaçadas no tempo a partir de uma forma de onda amostrada no domínio do tempo. Independente dos atributos estimados a partir do sinal os passos iniciais indicados são:

1. Pré-ênfase: Ao sinal de entrada é aplicado um filtro passa altas de 1ª ordem, conforme descrito na equação 2.1, com o objetivo de atenuar as baixas frequências, compensando o processo de produção da fala humana, que tende atenuar as altas frequências (TOGNERI; PULLELLA, 2011).

$$y(n) = x(n) - 0,97x(n-1), \tag{2.1}$$

   onde $x(n)$ é o sinal de entrada, $y(n)$ é o sinal filtrado e n são amostras de tempo.

2. Enquadramento: O sinal é segmentado em quadros com duração fixa sobrepostos. Os valores típicos de duração dos quadros são entre 20 e 30 ms com sobreposição de 10 a 20 ms.

3. Janelamento: Cada quadro (*frame*) é multiplicado por uma função de janela. O janelamento é necessário para suavizar as bordas dos quadros da etapa anterior (TOGNERI; PULLELLA, 2011).

A Figura 3 mostra um diagrama que ilustra tratamento de um sinal para a extração de características. Nos próximos tópicos serão abordados temas referentes aos descritores dos sinais de fala utilizados neste trabalho.

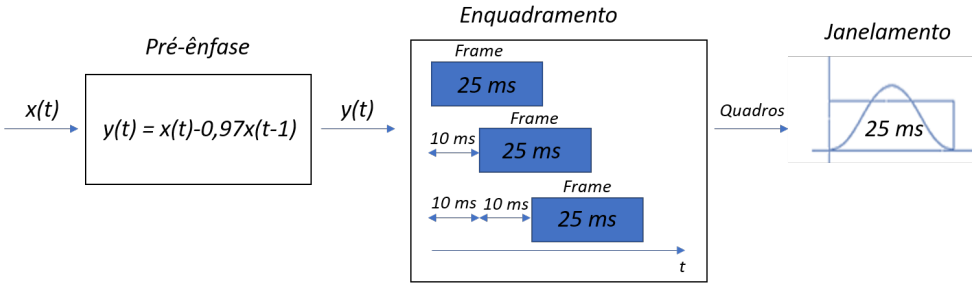

Figura 3: Diagrama das etapas iniciais para obtenção de descritores do sinal de voz. Fonte: Autor baseado em Togneri e Pullella (2011).

2.4.1 Coeficientes Cepstrais de Frequência na Escala Mel

Os principais parâmetros utilizados no reconhecimento de locutor são os Coeficientes Mel Ceptrais Espaçados na Frequência Mel (TOGNERI; PULLELLA, 2011) ou Coeficientes Cepstrais na Frequência Mel (LERCH, 2012). Segundo Lerch (2012), os MFCC são utilizados no processamento de sinais de fala desde a década de 80 do século passado. Conforme Togneri e Pullella (2011), os Coeficientes Cepstrais na Frequência Mel são baseados na percepção humana e no processo da audição. Na literatura, esses parâmetros também são bastante utilizados em sistemas para Recuperação de Informações Musicais (*MIR – Musical Information Retrival*) (LERCH, 2012). A Figura 4 mostra o diagrama de obtenção desses parâmetros.

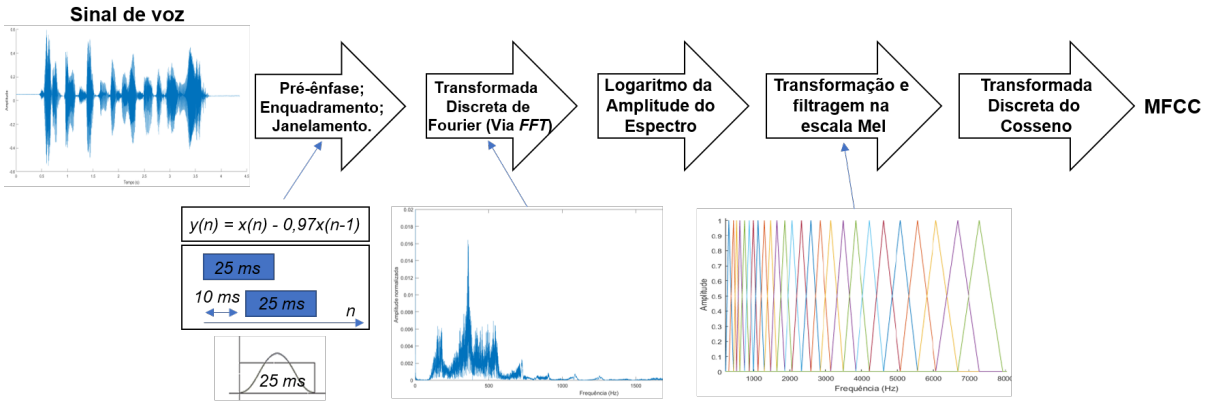

Figura 4: Diagrama de obtenção dos MFCC. Fonte: Autor.

De acordo com Togneri e Pullella (2011) e Oliveira (2018a) além das etapas descritas no tópico anterior as seguintes etapas são necessárias para o cálculo dos *MFCC*:

1. Utilização da janela de *Hamming* para a etapa de janelamento: Os *frames* obtidos da etapa anterior (enquadramento) são multiplicados por janelas de *Hamming* com o objetivo de atenuação das bordas dos quadros. Esta etapa garante transições mais suaves e menos distorcidas após a transformação para o domínio espectral (DFT) realizada adiante. A expressão que define a janela de *Hamming* é mostrada abaixo (CHAKRABORTY; TALELE; UPADHYA, 2014):

$$w(n) = 0,54 - 0,46\cos(\frac{2n\pi}{N-1}), \tag{2.2}$$

onde $n = [0, N-1]$ e N é a quantidade de amostras em cada *frame* sendo que as constantes $0,54$ e $0,46$ são próprias do tipo de janela (LATHI; GREEN, 2005).

2. Transformação para o domínio da frequência: Cada um dos *frames* que passaram através da janela de *Hamming* são convertidos para o domínio da frequência (espectro), usando a transformada rápida de *Fourier*. No processamento de fala, a informação de fase é ignorada e somente a magnitude da FFT é considerada (TOGNERI; PULLELLA, 2011).

3. Bancos de filtros espaçados na escala mel: Os coeficientes de magnitude da FFT são convertidos para a escala mel pela multiplicação por filtros triangulares espaçados de acordo com a Equação 2.3. Este processo origina os coeficientes espectrais de frequência de Mel (*Mel frequency spectrum coefficients ou MFSC*). A faixa de frequências dos filtros compreende o intervalo de 0 a $f_s/2$ (TOGNERI; PULLELLA, 2011) (LIU et al., 2013). A quantidade de filtros pode variar, por exemplo: em (MAFRA, 2002) foram utilizados 32 filtros; em (OLIVEIRA, 2018a) 26; 20 em (DAVIS; MERMELSTEIN, 1980) e (MARTINEZ et al., 2012). Segundo (TOGNERI; PULLELLA, 2011) podem ser utilizados 30 ou mais filtros. Neste trabalho foram utilizados 42 filtros em uma faixa de 300 a 8kHz (frequência audível). Por meio da equação 2.3 é realizada a conversão da frequência em Hertz para a frequência mel.

$$f_{mel} = 1125 \ln\left(1 + \frac{f_{Hz}}{700}\right), \tag{2.3}$$

onde f_{Hz} é a frequência em Hz e f_{mel} é a frequência na escala Mel (LERCH, 2012).

4. Análise cepstral: Converte os valores dos coeficientes espectrais de frequência de Mel em coeficientes cepstrais usando a Transformada Discreta do Cosseno, conforme equação abaixo:

$$c_j = \sum_{k=1}^{K} \log\left(X(k)\right) \cos\left[j\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \tag{2.4}$$

onde $X(k)$ é a magnitude espectral, K é o número de filtros do banco mel e j é a ordem do MFCC.

Segundo Togneri e Pullella (2011), em geral, são extraídos 12 MFCC por quadro, mas outras configurações também são utilizadas como 13, 14, 15, 19 e 20 coeficientes. Nesse trabalho serão testados todas as configurações anteriores sendo observado a relação entre eficiência de classificação e tempo de extração para definição do número de coeficientes.

2.4.2 Coeficientes *Delta* e *Delta Delta*

A fim de aumentar a eficiência do classificador, além dos coeficientes cepstrais na frequência mel "estáticos", estimam-se aproximações da velocidade e a aceleração dos MFCCs por meio dos parâmetros Delta e Delta Delta. Esses descritores derivados dos MFCC são adicionados , pois fornecem informações sobre variação da mudança da fala, Delta, e sobre a variação da velocidade da fala, Delta Delta (OLIVEIRA, 2018a). Os coeficientes dinâmicos podem ser obtidos através das Equações 2.5 e 2.6 (KIM; MOREAU; SIKORA, 2006).

$$\Delta c_j(n) = \frac{\sum_{p=1}^{P} c(n+p) - c(n-p)}{2\sum_{p=1}^{P} p^2} \tag{2.5}$$

$$\Delta\Delta c_j(n) = \frac{\sum_{p=1}^{P} \Delta c(n+p) - \Delta c(n-p)}{2\sum_{p=1}^{P} p^2} \tag{2.6}$$

onde $c_j(n)$ são os coeficientes mel cepstrais de ordem n e $P = 2$.

2.5 FREQUÊNCIA FUNDAMENTAL (*PITCH*)

O *pitch* ou frequência fundamental (f_0) é um atributo muito importante no reconhecimento de emoção a partir do sinal de voz. O *pitch* possui informações sobre a emoção, pois depende da tensão das pregas vocais e da pressão do ar necessária para abertura das cordas vocais (ar subglótico), sendo diferente nos estados emocionais básicos (Alegria, Tristeza, Medo, Nojo, Surpresa, Raiva, Desprezo) (PAN; SHEN; SHEN, 2012).

Na literatura, existem diferentes métodos para a estimação da frequência fundamental que variam em termos de robustez ao ruído, precisão e custo computacional. Para estimar o *pitch* foi utilizado o método da Função de Correlação Normalizada (*Normalized Correlation Function (NCF)*) e uma banda de frequências de 60-320 Hz (VERVERIDIS; KOTROPOULOS; PITAS, 2004).

2.6 PROCESSAMENTO ESTATÍSTICO

A forma como as informações são apresentadas ao sistemas de classificação influenciam em sua capacidade de discriminação entre classes. Na análise de sinais multidimensionais é necessário encontrar a melhor forma de disposição dos dados de modo a torná-los

mais acessíveis. As Análise de Componentes Principais (PCA) e Análise de Componentes Independentes (ICA) são técnicas muito utilizadas para a compressão de dados e extração de características.

Em diversos trabalhos é possível observar o uso de técnicas de processamento estatístico para melhorar os resultados de classificação. Por exemplo, em Oliveira et al. (2020) utilizou-se a técnica de PCA para compactação e redução da correlação entre os sinais ultrassônicos. Em Moura et al. (2009) utilizou-se Análise de componentes independentes para processamento e remoção de interferência em sinais de sonar passivo. Em ambos os casos a rede neural treinada usando características extraídas através do uso de ICA e PCA aumentou a eficiência da discriminação se comparado à mesma rede neural treinada diretamente com os sinais.

Neste trabalho foram utilizadas as análises de componentes principais e componentes independentes com o objetivo de remover a redundância entre os atributos de entrada para a rede neural. A seguir estas técnicas são brevemente descritas.

2.6.1 Análise de Componentes Principais

A análise de componentes principais *(Principal Component Analysis - PCA)* é uma técnica estatística de processamento de sinais diretamente ligada à transformação de *Karhunen-Loève* (JOLLIFE, 2002). A PCA realiza uma transformação linear tal que os dados projetados sejam não correlacionados e grande parcela da energia (variância) esteja concentrada num pequeno número de componentes.

A análise de componentes principais é bastante utilizada para compactação de informação. Como a PCA projeta os sinais em componentes ordenados por energia, uma medida geralmente utilizada para reduzir a dimensão dos dados selecionando apenas os componentes de maior energia, de modo que o sinal recuperado a partir da informação compactada tenha pequeno erro médio quadrático se comparado ao sinal original.

A Figura 5 mostra um diagrama de obtenção da *PCA*. Os sinais $\mathbf{X}$ são projetados pela matriz $\mathbf{B}$, originando as componentes $\mathbf{Z}$ não correlacionados e ordenados pela energia.

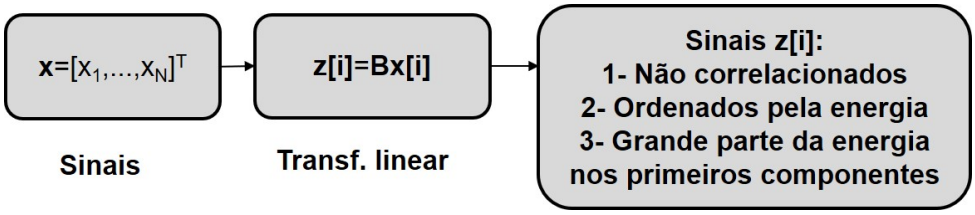

Figura 5: Diagrama de blocos da PCA. Fonte: Autor.

2.6.2 Análise de Componentes Independentes

A análise de componentes independentes, ou *Independent Component Analisis (ICA)*, é um método que busca encontrar uma representação linear, de modo que os novos componentes sejam estatisticamente independentes e não Gaussianos. Essa representação parece capturar a estrutura essencial dos dados em muitas aplicações, incluindo extração de atributos e separação de sinais (HYVÄRINEN; OJA, 2000).

A ICA está intimamente relacionada ao método chamado separação cega da fonte (*Blind Source Separation - BSS*) ou separação cega do sinal. É um dos métodos, talvez o mais utilizado, para realizar a decomposição de um sinal multidimensional em suas componentes originais mutuamente independentes (KAMATH; RAVINDRAN; ANDERSON, 2004). Por exemplo, se em um local fechado há duas fontes sonoras e realiza-se a captura desses sinais a partir de dois pontos distintos, os sinais capturados por esses dois sensores são uma soma ponderada dos sinais emitidos por essas pessoas. Essa situação pode ser exemplificada matematicamente pelas equações:

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \tag{2.7}$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \tag{2.8}$$

onde $x_1(t)$ **e** $x_2(t)$ correspondem aos sinais capturados pelos sensores utilizados, $s_1(t)$ e $s_2(t)$ são os sinais oriundos das fontes sonoras a_{11}, a_{12}, a_{21} e a_{22} são parâmetros que dependem das distâncias dos sensores e das fontes sonoras.

Nesse contexto, a técnica de Análise de Componentes Independentes pode ser utilizada para estimar os parâmetros a_{ij} com base na informação de independência estatística das fontes, o que permite separar os dois sinais originais $s_1(t)$ e $s_2(t)$ de suas misturas $x_1(t)$ e $x_2(t)$ (HYVÄRINEN; OJA, 2000).

Segundo Hyvärinen e Oja (2000), para uma definição matemática rigorosa devemos utilizar um modelo estatístico de "variáveis latentes". Supondo que observa-se n misturas lineares $x_1, ..., x_n$ de n componentes independentes. Os sinais dos sensores são definidos como uma amostra de uma variável aleatória. Usando-se uma notação matricial para definir o problema, indicando por $\mathbf{x}$ o vetor aleatório cujos elementos são as misturas $x_1, ..., x_n$ e da mesma forma por $\mathbf{s}$ o vetor aleatório com elementos $s_1, ..., s_n$ e escrevendo os elementos a_{ij} como uma matriz $\mathbf{A}$. O modelo de mistura acima é escrito como:

$$\mathbf{x} = \mathbf{As} \tag{2.9}$$

O modelo ICA é um modelo generativo, o que significa que ele descreve como os dados observados são gerados por um processo de mistura dos componentes s_i (HYVARINEN, 1999). Como variáveis latentes as componentes independentes não podem ser

observadas diretamente, assim como a matriz de mistura **A** também é desconhecida. Tudo o que observamos é o vetor aleatório **x**, e devemos estimar ambos **A** e **s**. O ponto de partida para a ICA é a suposição de que os componentes s_i são estatisticamente independentes. Então, depois de estimar a matriz **A**, podemos calcular sua inversa, e obter o componente independente simplesmente por (HYVARINEN, 1999).

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{2.10}$$

Os algoritmos para a obtenção dos componentes da ICA, em sua maioria, computam a PCA numa etapa anterior, facilitando o problema de obtenção de estimativas de ordem superior. Considerando sua rápida execução, mesmo com um número grande de sinais, e sua eficiência na estimação dos componentes independentes, neste trabalho será utilizando o algoritmo FastICA (HYVARINEN, 1999). O FastICA é um algoritmo de aprendizagem que busca encontrar um vetor **a** de forma que $a^T x$ maximize a não-gaussianidade. Para apenas uma componente o algoritmo pode ser descrito da seguinte forma:

1. Estima-se o vetor a de forma aleatória;
2. Faz-se $a^+ = E\left\{xG(a^T x)\right\} - E\left\{G'(a^T x)\right\} a$;
3. Sendo $a = a^+ / \left\|a^+\right\|$;
4. Retornar ao passo 2 até que o algoritmo convirja.

onde **a** é o vetor dos parâmetros dos sensores e das fontes. Comparando com o modelo de um neurônio, o vetor **a** são os pesos sinápticos. A função G e sua derivada (G') são definidas da seguinte forma:

$$G(a^T x) = \frac{1}{c_1} \times \log\cosh(c_1(a^T x)) \tag{2.11}$$

$$G'(a^T x) = tanh(c_1(a^T x)) \tag{2.12}$$

onde c_1 é uma constante com valor igual a 1 e G é uma função não quadrática utilizada no algoritmo. Para múltiplas componente o algoritmo pode ser descrito da seguinte forma:

1. $a^+ = a/\sqrt{\|aa^T\|}$;
2. $a = \frac{3}{2}a - \frac{1}{2}aa^T a$;
3. Retornar ao passo 2 até que o algoritmo convirja.

2.7 APRENDIZADO DE MÁQUINA

Aprendizado de máquina ou *Machine Learning (ML)*, do inglês, é uma grande área de estudos que compreende, entre outros, ferramentas matemáticas capazes de dotar sistemas computacionais de meios de tomada de decisões baseadas em informações coletadas anteriormente. Segundo Mitchell (1997b), desde a invenção dos computadores imaginava-se, por exemplo, a sua utilização para descoberta de padrões a partir de registros médicos para identificação doenças graves antes da sua manifestação. Por outro lado, uma compreensão bem sucedida de como fazer os computadores aprenderem abririam muitos novos usos para essas máquinas. Além disso, um entendimento detalhado dos algoritmos de processamento de informações para aprendizado de máquina pode levar a uma melhor compreensão das habilidades e dificuldades de aprendizagem humana (MITCHELL, 1997a).

As técnicas de aprendizado de máquina podem ser classificadas de diferentes formas, dependendo da maneira como o conhecimento é organizado e generalizado. Nos próximos capítulos serão abordadas técnicas de *Machine Learning* utilizadas neste trabalho.

2.7.1 Redes Neurais Artificiais

As redes neuras artificiais (*Artificial neural networks – ANN*) são modelos inspirados no funcionamento do sistema nervoso de animais. Os menores elementos das redes neurais artificiais são os neurônios que podem ser organizados de diferentes formas, criando redes diferentes. Diferente de um computador digital convencional, o cérebro é um computador altamente complexo, não linear e paralelo (HAYKIN et al., 2009). A generalização de uma rede neural se deve ao fato de que ela pode produzir saídas adequadas a entradas que não estavam presentes durante o processo de aprendizagem. Apesar de características úteis na resolução de problemas complexos, as redes neurais artificiais ainda estão longe de conseguir igualar o cérebro humano (SVOZIL; KVASNICKA; POSPICHAL, 1997). Segundo Haykin et al. (2009), o uso de redes neurais artificiais oferece as seguintes propriedades de interesse:

1. **Não-linearidade**: um neurônio artificial pode ser linear ou não-linear, constituindo, assim, redes lineares ou não-lineares. A não-linearidade pode ser distribuída por toda a rede, sendo importante para acessar informações da estatística de ordem superior (HAYKIN et al., 2009).

2. **Mapeamento entrada-saída**: aprendizado supervisionado consiste na modificação dos pesos sinápticos de uma rede neural pela aplicação de um conjunto de amostras de treino rotuladas. Apresenta-se para a rede um conjunto de entrada e os respectivos sinais de saída desejados. Os pesos sinápticos são alterados a fim de conseguir o menor erro entre a saída real e a desejada, usando um procedimento estatístico

apropriado. O treinamento da rede é repetido várias vezes para o conjunto de entrada até que não haja mudanças significativas.

3. **Adaptabilidade**: as redes neurais têm a capacidade de modificar seus pesos sinápticos em resposta a modificações do meio ambiente. Uma rede neural treinada para operar num ambiente especifico pode ser retreinada para atender a modificações do ambiente.

4. **Resposta a Evidências**: uma rede neural pode fornecer informações não somente sobre qual padrão particular selecionar, mas também sobre a confiança na decisão tomada.

5. **Informação Contextual**: cada neurônio é afetado pela atividade de todos os outros neurônios na rede.

6. **Tolerância a Falhas**: uma rede neural, implementada em um *hardware*, é capaz de ser tolerante a falhas ou de realizar computação robusta (HAYKIN et al., 2009). Uma rede neural operando em condições adversas, perda de um neurônio ou algumas conexões, tem seu desempenho afetado, mas devido ao armazenamento distribuído das informações o resultado global não é totalmente prejudicado (HAYKIN et al., 2009) .

Existem duas formas principais de treinamento: treinamento supervisionado e não supervisionado. No treinamento supervisionado são utilizados conjuntos de dados rotulados, ou seja, a cada sinal de entrada é associado uma saída alvo desejada e o ajuste dos coeficientes de peso é feito de tal forma que as saídas calculadas e desejadas sejam as mais próximas possíveis. No treinamento não supervisionado a saída desejada não é conhecida, ou seja, não há exemplos rotulados da função a ser aprendida pela rede (HAYKIN et al., 2009).

O modelo de um neurônio pode ser visto na Figura 6. As saídas y_k variam de acordo com a função de ativação utilizada, por exemplo, entre $[0, 1]$ ou $[-1, 1]$.

A saída do combinador linear u_k pode ser definida matematicamente pela equação abaixo:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{2.13}$$

os valores w_{kj} são denominados pesos sinápticos e os x_j são os sinais de entrada. Nesse modelo de neurônio, os pesos sinápticos são ajustado para que se chegue aos valores de saída desejados. Por outro lado, a saída y_k pode ser definida conforme a equação abaixo:

$$y_k = \varphi(u_k + b_k) \tag{2.14}$$

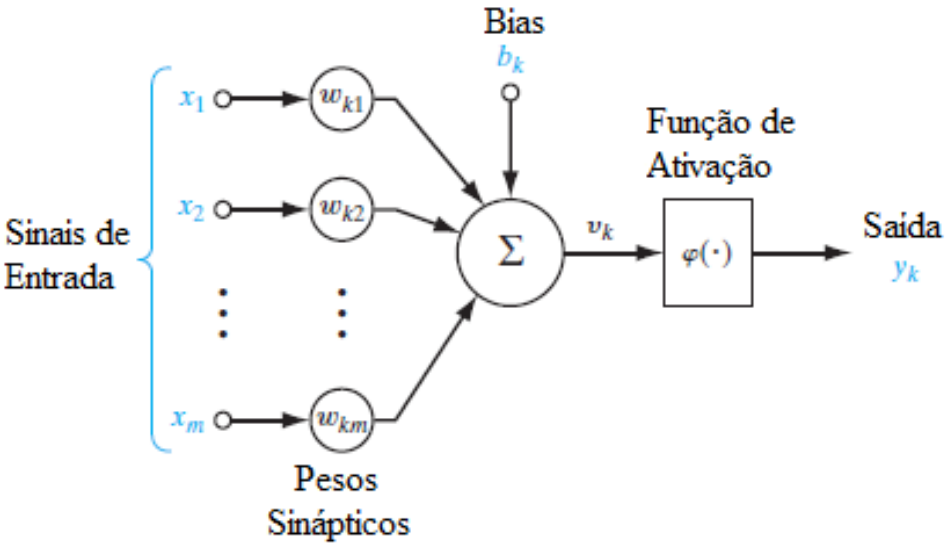

Figura 6: Modelo de um neurônio. Fonte: (HAYKIN et al., 2009).

a variável b_k é o *bias* (viés); $\varphi()$ é a função de ativação. O uso do *bias* tem o efeito de aplicar uma transformação afim à saída u_k do combinador linear no modelo da Equação 2.14, como mostrado por Haykin et al. (2009).

$$v_k = u_k + b_k \tag{2.15}$$

Na Equação 2.15, verifica-se que dependendo se a polarização b_k é positiva ou negativa a relação entre o campo local induzido, ou o potencial de ativação, v_k do neurônio k e a saída do combinador linear u_k é modificada. Logo, pode-se reformular o modelo do neurônio k como mostrado na Figura 7. O efeito do bias é explicado de forma resumida como:

1. Adicionando um novo sinal de entrada fixado em +1, e
2. Adicionando um novo peso sináptico igual ao bias b_k

Embora os modelos das Figuras 6 e 7 possuam aparências distintas, são matematicamente equivalentes.

a função de ativação, denotada por (φ), define a saída de um neurônio em termos do campo local induzido. A seguir, identificamos alguns tipos básicos de funções de ativação:

- Função limiar:

$$\varphi(v) = \begin{cases} 1, & v \geq 0, \\ 0, & v < 0 \end{cases} \tag{2.16}$$

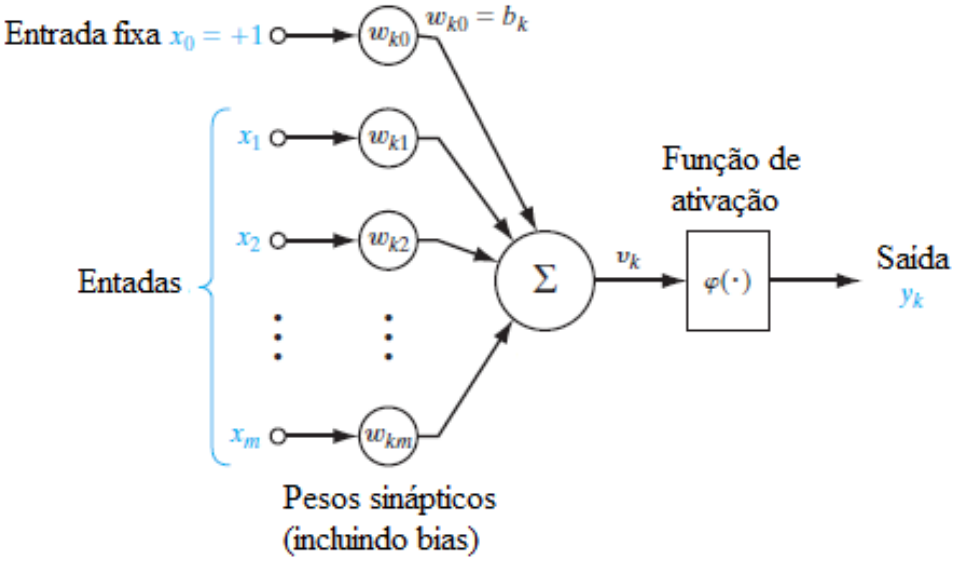

Figura 7: Modelo adaptado de um neurônio. Fonte: (HAYKIN et al., 2009).

- Função sinal:

$$\varphi(v) = \begin{cases} -1, & v < 0, \\ 0, & v = 0, \\ 1, & v > 0 \end{cases} \tag{2.17}$$

- Função sigmóide. A função sigmóide é uma das formas mais comuns de funções de ativação usada na construção de redes neurais, sendo definida como uma função estritamente crescente. Exemplo de função sigmóide são a função tangente hiperbólica e a função logística, definida por (HAYKIN et al., 2009):

$$\varphi(v) = \frac{1}{1 + exp(-av)} \tag{2.18}$$

$$\varphi(v) = \tanh(v) \tag{2.19}$$

- Função Softmax. A função de ativação Softmax é normalmente utilizada na camada de saída (classificação) pois pode gerar uma estimativa de massa de probabilidades dos valores das saída com valores não nulos e soma igual a 1.

$$\varphi_i(v) = \frac{\exp(v_i)}{\sum_{j=1}^{k} \exp(v_j)} \tag{2.20}$$

- Função Retificadora Linear. A função *rectified linear function* (ReLu) é uma função de ativação bastante utilizada em *Deep Learning*. Para qualquer valor de entrada menor que zero, a saída é sempre zero, sendo definida da seguinte forma:

$$\varphi(v) = \max\{0, v\} \tag{2.21}$$

- Leaky ReLu. A função *Leaky ReLU* é uma modificação da função ReLU que busca solucionar o problema da morte do gradiente adicionando um fator de escala na função que passa a apresentar valores diferente de zero para valores negativos de v.

$$\varphi(v, \alpha) = \max\{\alpha v, v\} \tag{2.22}$$

Nessa dissertação foram utilizadas as funções de ativação tangente hiperbólica, *softmax* e retificadora linear. No Capítulo 3 será abordado os motivos para cada escolha. Já na seção subseção 2.7.2 serão discutidas formas de organização dos neurônios.

2.7.2 Arquitetura de redes

Dependendo da forma como os neurônios são organizados as redes neurais podem exibir as seguintes arquiteturas (HAYKIN et al., 2009):

1. Redes Alimentadas Adiante com Camada Única (*Single-Layer Feedforward Networks*): forma mais simples de uma rede em forma de camadas. Possui apenas uma camada de entrada e uma camada de saída.

2. Redes Alimentadas Adiante com Múltiplas Camadas (*Multilayer Feedforward Networks*): nesse tipo de rede além da camada de entrada e da camada de saída há presença de uma ou mais camadas ocultas, cujos nós são chamados de neurônios ocultos ou unidades ocultas.

3. Redes Recorrentes (*Recurrent Networks*): é um tipo de arquitetura de rede que possui pelo menos um laço de realimentação. Esses laços de realimentação, também chamados de Loops de feedback, possuem elementos de retardo de tempo unitário (*unit-time delay*), que resultam em um comportamento dinâmico não linear, desde que a rede neural contenha unidades não lineares. A Figura 10 mostra um diagrama simplificado de uma rede neural recorrente, onde z^{-1} é a função de transferência que aplica um retardo unitário ao sinal de entrada.

2.7.3 Redes Neurais Multicamadas

Redes Neurais *Feedforward* ou *Multilayer Perceptrons (MLP)* tem como objetivo a aproximação de funções. Dado uma função $z = f(x)$, de x em z, uma rede *feedforward* mapeia $z = \hat{f}(x, \theta)$ armazenando conhecimento sobre os valores de θ que melhor aproximam a função (GOODFELLOW; BENGIO; COURVILLE, 2016). A rede do tipo perceptron é

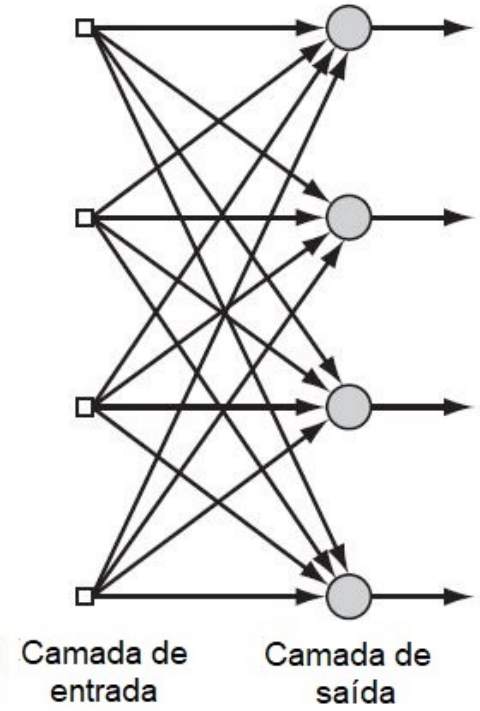

Figura 8: Rede Neural *Feedforward* com uma única camada de neurônios. Fonte: (HAYKIN et al., 2009).

o caso mais simples das redes *feedforward* por não possuírem camada escondida e por só serem capazes de resolver problemas linearmente separáveis.

As *Multilayer Perceptrons (MLP)* são redes *Feedforward* com uma ou mais camadas escondidas dando a elas um maior poder computacional. Os três pontos a seguir destacam os recursos básicos dos perceptrons multicamadas (HAYKIN et al., 2009):

1. O modelo de cada neurônio na rede inclui uma função de ativação não linear que é diferenciável.
2. A rede contém uma ou mais camadas ocultas além dos nós de entrada e saída.
3. A rede exibe um alto grau de conectividade, cuja extensão é determinada pelos pesos sinápticos da rede.

2.7.4 Algoritmos de treinamento

O número de iterações e a escolha do algoritmo de treinamento possuem muita importância no resultado da rede. Neste tópico discute-se sobre dois tipos de algoritmos de treinamento utilizados com redes de múltiplas camadas. São eles o *Backpropagation* e o *Resilient Propagation*.

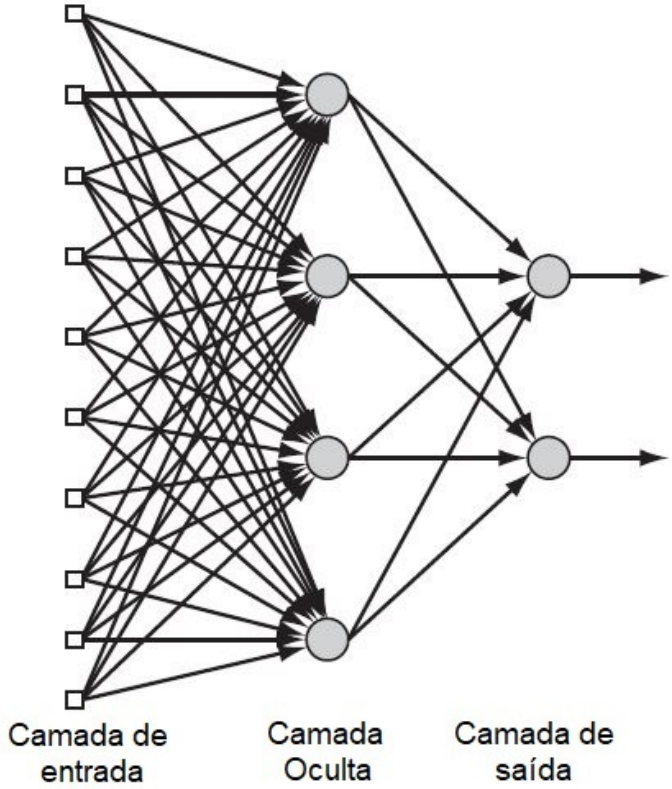

Figura 9: Rede Neural *Feedforward* com uma camada de neurônios oculta. Fonte: (HAYKIN et al., 2009).

Segundo Haykin et al. (2009), o desenvolvimento do algoritmo de retropropagação representa um marco nas redes neurais, pois fornece um método computacional eficiente para o treinamento de perceptrons de múltiplas camadas, aplicando uma correção aos pesos sinápticos através do mapeamento dos pares entrada-saída. Na aplicação desse algoritmo observa-se dois passos: o passo para frente ou propagação e o passo para trás ou retropropagação (HAYKIN et al., 2009).

Ao iniciar o treinamento, no passo para frente, é realizada a inicialização dos pesos da rede neural com valores aleatórios. Um vetor de entrada é apresentado e os valores de saída são comparados aos valores desejados. Os pesos sinápticos se mantêm inalterados e os sinais de saída são calculados individualmente neurônio por neurônio.

O passo para trás começa na saída da rede, onde é calculado o erro (diferença entre a saída real e a saída alvo). Os sinais de erro são propagados na direção da entrada da rede, camada por camada, adaptando-se os pesos através de ajustes recursivos. O processo é realizado até os pesos sinápticos e os bias se estabilizarem, e o erro médio quadrático do conjunto de treinamento convergir para um valor mínimo. Uma época pode ser definida como a passagem de todos os dados do conjunto de treinamento pela rede neural.

Várias modificações no algoritmo *Bachpropagation* foram propostas. Dentre elas

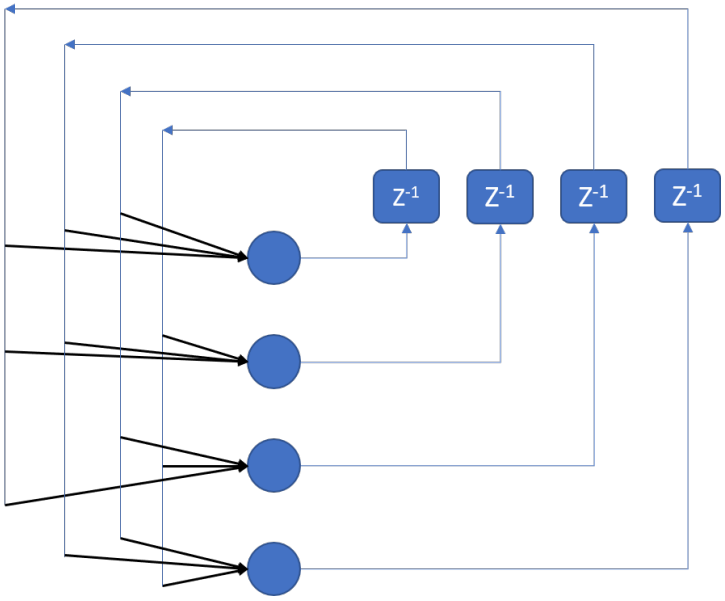

Figura 10: Modelo adaptado de uma rede recorrente simples (SRN). Fonte: Autor (baseado em Haykin et al. (2009)).

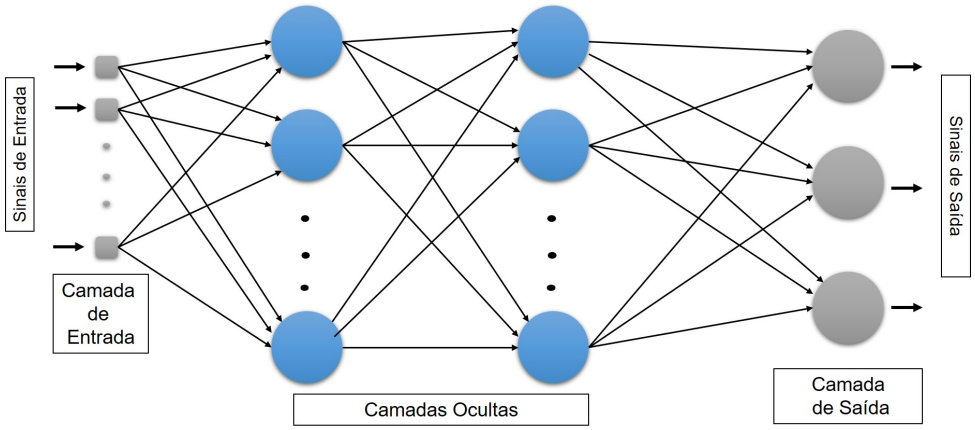

Figura 11: Modelo de uma rede neural multicamadas. Fonte: Autor.

citam-se: o *Levenberg-Marquardt Backpropagation*, *Bayesian Regularization Backpropagation*, o *Resilient Backpropagation* e o *Gradient Descent Backpropagation*. A propagação resiliente é um eficiente esquema de aprendizagem, que realiza uma adaptação direta do peso com base em informações do gradiente local. A diferença fundamental entre os métodos é a redução da influência do gradiente no esforço de adaptação. Isso se deve à atualização do valor de cada peso individualmente, determinando somente o tamanho da atualização do peso que evolui adaptativamente durante o processo de aprendizagem baseada na função de erro.

O algoritmo de propagação resiliente *Resilient Propagation* elimina os efeitos da amplitude das derivadas parciais, levando em consideração somente o sinal destas derivadas para decisão do sentido de atualização dos pesos sinápticos.

2.8 APRENDIZADO PROFUNDO

O Aprendizado Profundo, ou *Deep learnig*, é uma poderosa ferramenta para o aprendizado supervisionado e não supervisionado, por sua estrutura complexa e encadeada pode representar funções com auto grau de complexidade (GOODFELLOW; BENGIO; COURVILLE, 2016). A aprendizagem profunda é uma área particular da aprendizagem de máquina em que técnicas de extração de características não são necessariamente aplicadas aos dados brutos antes da apresentação dessas informações aos algorítimos de classificação. Esses algorítimos são dotados de ferramentas que buscam extrair elementos que auxiliam na resolução do problema e sua profundidade está relacionada ao número de camadas da estrutura.

As técnicas de aprendizado profundo são utilizadas com resultados satisfatórios em diversas abordagem. Dentre elas: Reconhecimento automático de Voz (ASR), Reconhecimento de emoções na voz (SER), Processamento de linguagem natural, Robótica, Visão Computacional, Bioinformática, Máquinas de busca, Vídeo games (GOODFELLOW; BENGIO; COURVILLE, 2016), Séries Temporais e Detecção de Movimento (Cai; Cai; Li, 2018).

Uma rede neural artificial para aprendizado profundo, em geral, é constituída de camadas conectadas sequencialmente, ou seja, a saída de uma camada torna-se entrada para a camada seguinte. A aprendizagem se dá pelo ajuste dos pesos das camadas. A Figura 12 mostra uma representação de uma rede neural profunda.

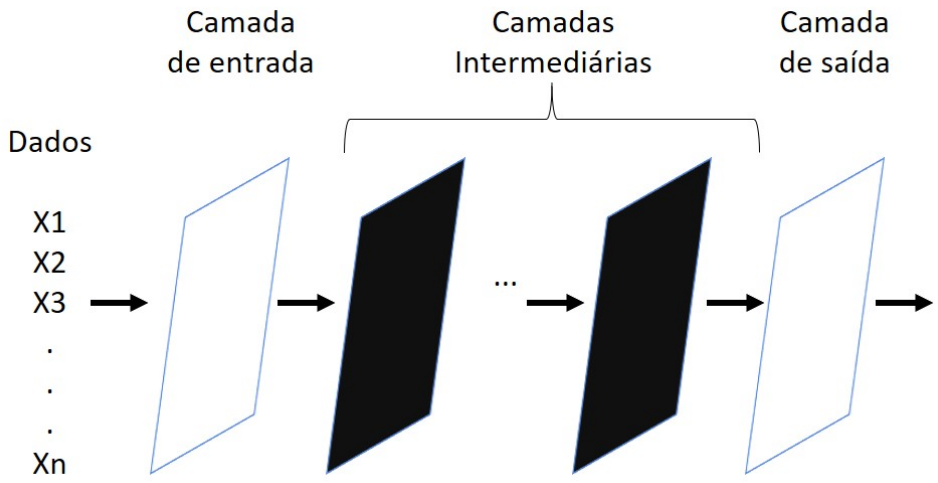

Figura 12: Exemplo de uma rede neural profunda. Fonte: Autor.

Nos próximos tópicos serão abordados tipos específicos de redes neurais artificiais profundas bastante utilizados em tarefas de *SER* e *ASR*.

2.8.1 Redes Neurais Recorrentes

Redes Neurais Recorrentes, ou *Recurrent neural networks (RNN)*, são uma família de redes neurais para processamento de dados sequenciais (GOODFELLOW; BENGIO; COURVILLE, 2016). Redes Neurais Recorrentes (RNR) são redes especializadas no processamento de entradas sequenciais, como fala, linguagem, séries temporais de valores e etc. As RNRs podem processar longas e variáveis sequências com muito mais eficiência do que redes neurais não especializadas em dados sequenciais (GOODFELLOW; BENGIO; COURVILLE, 2016). Como características especiais dessas redes podemos destacar o fato delas possuírem conexões recorrentes, ou seja, ligações entre os neurônios podem formar ciclos (recorrências).

Segundo Santana (2017), havia um problema no aprendizado de redes recorrentes devido à dificuldade de aprender dependências de longo alcance, causando os problemas de desaparecimento e explosão do gradiente durante a retropropagação do erro em redes com múltiplas camadas. Para a solução deste problema, Wöllmer et al. (2013) propôs uma nova arquitetura RNN denominada *Long Short-Term Memory*, que possui a capacidade de armazenar informações por um período mais longo de tempo e pode aprender grande quantidade de informações relevantes. A Figura13 mostra representações das redes neurais recorrentes. Na esquerda, observa-se um representação simplificada com uma retroalimentação na camada h. Já na direita, verifica-se um detalhamento maior da rede. Cada instância de tempo depende do seu estado anterior $(t-1)$, ou seja, a camada oculta do passo anterior alimenta o passo atual.

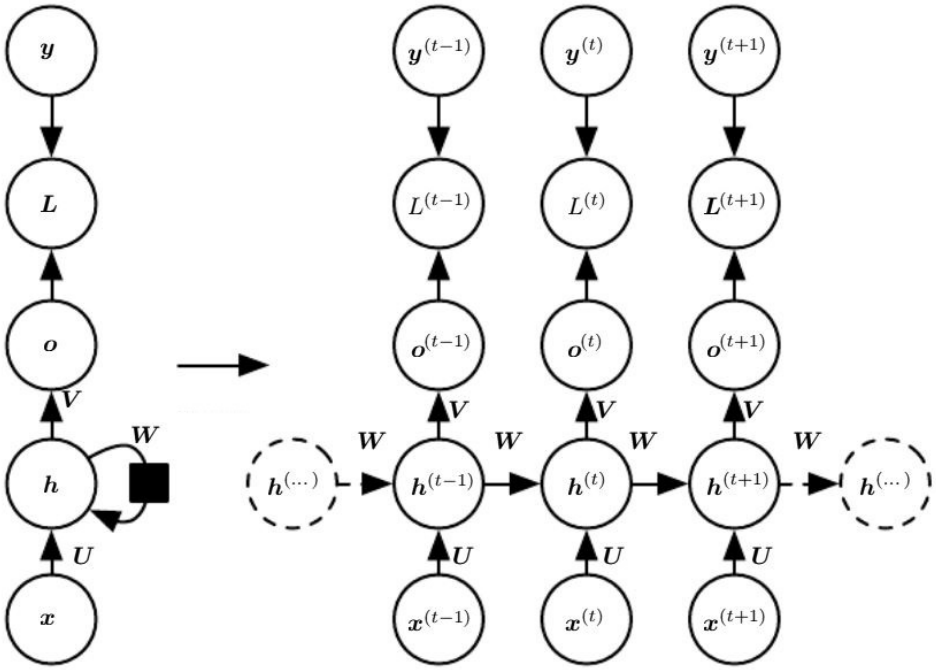

Figura 13: Exemplo em grafos de uma RNN. Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016).

onde x é a entrada, h é a camada oculta, **U**, **W** e **V** são matrizes de pesos, L é a função de perda que mede o erro da saída o em relação ao alvo y.

Redes neurais recorrentes com *Long Short-Term Memory* e *bidirectional Long Short-Term Memory* tem mostrado boa performance em tarefas de reconhecimento de discursos, falante e emoções (HUANG et al., 2016). No Capítulo 3 serão mostrados detalhes sobre as redes recorrentes utilizadas.

2.9 PARÂMETROS UTILIZADOS PARA AVALIAÇÃO DO SISTEMA

Para avaliar o desempenho dos classificadores utilizados, a matriz de confusão, a média geométrica das eficiências (PEF) e a acurácia foram calculados.

2.9.1 Acurácia

A acurácia produz uma avaliação geral dos acertos. É calculada somando o número de acertos de classificação para cada classe e dividindo pelo número total de dados disponíveis no *dataset*.

$$AC = \left(\frac{\sum_{i=1}^{k} EF_i}{NT} \right) 100 \tag{2.23}$$

onde EF_i é a eficiência da classificação, ou seja, número de acertos do classificador obtido para a classe i e NT é o número total de amostras utilizada.

2.9.2 Produto das Eficiências

A média geométrica das eficiências (PEF) fornece uma medida do desempenho geral do classificador:

$$PEF = \left(\sqrt[k]{\prod_{i=1}^{k} EF_i} \right) 100 \tag{2.24}$$

onde EF_i é a eficiência da classificação obtida para a classe i. A média geométrica é a preferida neste caso, uma vez que é mais sensível a baixa eficiência de uma única classe.

2.9.3 Matriz de confusão

A matriz de confusão é uma ferramenta que permite a visualização das eficiências de discriminação (na diagonal principal) e erros de classificação (em posições fora da diagonal), para cada classe, em um problema de classificação. Neste sentido, a classe

predita está localizada na vertical e a classe verdadeira na horizontal. A Figura 14 mostra a uma representação da matriz de confusão utilizada.

	Classe A	**Classe B**
Classe A	**Acerto da Classe A (em %)**	Erro da Classe A (em %)
Classe B	Erro da Classe B (em %)	**Acerto da Classe B (em %)**

Figura 14: Matriz de confusão para duas classes. Fonte: Autor.

3 METODOLOGIA

Neste Capítulo o sistema proposto é apresentado, as ferramentas utilizadas, seus ajustes, configurações e os motivos de suas escolhas são justificadas. Uma parte das ferramentas e ajustes inicias propostos são baseados no trabalho de Oliveira (2018a).

3.1 MODELO PROPOSTO

O modelo proposto baseia-se na extração de características do sinal de voz, extração de características derivadas (Delta e Delta Delta), no pós processamento dos descritores para a obtenção de redução de dimensionalidade e de informação do espaço de características (ICA), a utilização de classificadores neurais rasos (MLP) e profundos (RNN). A Figura 15 ilustra o sistema de classificação de emoções e locutor proposto nesse trabalho.

As características estimadas a partir do sinal de voz foram os Coeficientes Mel Cepstrais e suas derivadas, o Delta (velocidade) e Delta Delta (aceleração), e o *pitch* de voz. A escolha desses descritores é devida a sua grande utilização em diversos estudos sobre reconhecimento de locutor e emoções, como por exemplo: Oliveira (2018a), Aouani e Ayed (2018), Oliveira, Cerqueira e Filho (2018b), Silva et al. (2015) e Likitha et al. (2017).

Com o objetivo de obter um controle sobre as etapas do processo de obtenção dos descritores dos sinais de voz buscou-se escrever os próprios códigos. Esses códigos foram implementados (programados) no MATLAB 2020b devido a familiaridade do autor com o ambiente de programação e a possibilidade de utilização da grande biblioteca de funções nativas. Todos os códigos desenvolvidos para o projeto estão disponíveis em: <https://https://github.com/EATBJ/codigosestrado>.

Diversas ferramentas são utilizadas na literatura para o reconhecimento automático de locutores e emoções através da voz, dentre elas: Em Campos (2017) K-Vizinhos mais próximos *(k-Nearest neighbor - KNN)*. Em Mafra (2002) uma rede neural *Self-Organizing Map (SOM)*. Em Oliveira (2018a), redes neurais artificiais MLP. No trabalho Jagiasi et al. (2019) são utilizadas *Deep Neural Network* e *Convolucional Neural Network*. Em Toruk e Gokay (2019) utilizam-se *Time-Delay Neural Networks (TDNN)*. Já Abdel-Hamid et al. (2014) utilizou redes neurais convolucionais *(Convolutional Neural Network - CNN)*, incluindo redes pré-treinadas. Em Kwon et al. (2020) redes neurais convolucionais *Deep Stride Convolutional Neural Network (DSCNN)*. Em Ma et al. (2019) os autores utilizaram redes LSTM residuais mutimodais (MMResLSTM).

Dentre as técnicas listadas anteriormente, optou-se pela utilização de redes neurais perceptron de múltiplas camadas e redes neurais recorrentes profundas como ferramentas

para classificação de padrões. As redes MLP foram escolhidas pela sua rápida implementação, por apresentarem resultados satisfatórios em problemas com poucas classes e por serem afetadas pelos parâmetros extraídos e pelo pré-processamento utilizado. Essas características serão exploradas mais profundamente neste trabalho. As redes neurais profundas são o estado da arte em sistemas de classificação, por isso utilizaremos as redes mais populares para fins de comparação.

Para a extração dos MFCC, delta e delta-delta foram utilizados: janelas de 30ms, 50% de superposição entre janelas adjacentes, 40 filtros para o banco de filtros Mel abrangendo aproximadamente de 300 a $8kHz$ (frequência audível). Já para a obtenção da frequência fundamental foram utilizados: janelas de 30ms e 50% de superposição entre janelas adjacentes.

Os estudos foram divididos em etapas de identificação do locutor e identificação da emoção, de modo que, cada sistema possa ser desenvolvido e implementado separadamente.

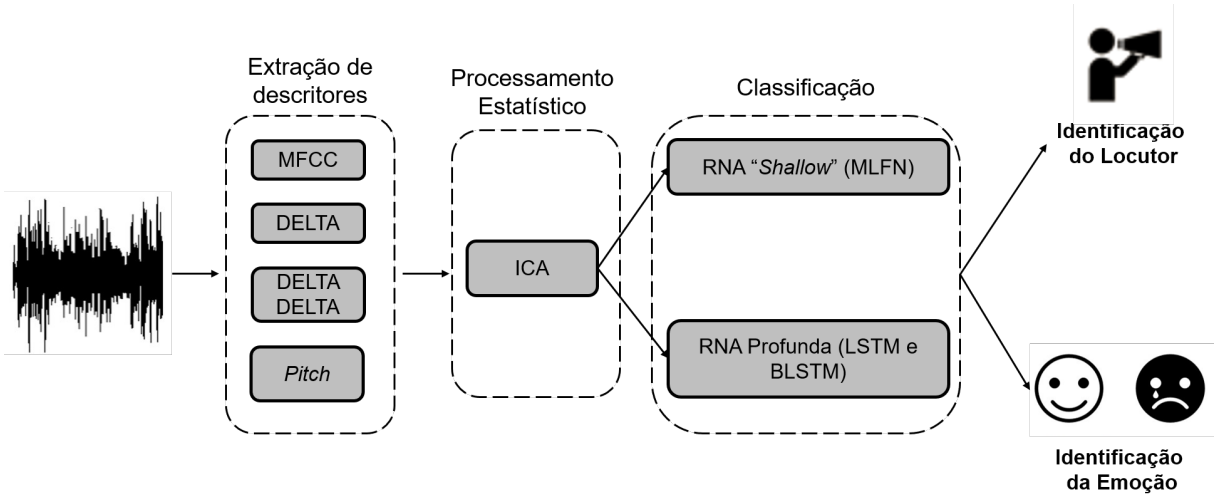

Figura 15: Diagrama de blocos do sistema de classificação de emoções e locutor. Fonte: Autor.

3.2 BASE DE DADOS

Banco de dados são conjuntos de informações organizadas que possuem algum significado. Uma base de dados de discursos (*speech database*), ou banco de falas, é um banco de dados de arquivos de áudios de discurso humano, contendo frases completas ou simplesmente palavras (OLIVEIRA, 2018a).

O acesso a um banco de vozes (falas) provê sinais escolhidos, processados e muitas vezes já testados em outras implementações, sendo uma fonte confiável de informações para treinamento e teste de sistemas de classificação.

Diversos bancos de dados de vozes estão disponíveis para as tarefas de reconhecimento de locutor e emoções. Cita-se, entre eles, o Banco de falas *Voxceleb1*, disponível em Nagrani, Chung e Zisserman (2017), composto de mais de 100.000 sinais de áudio extraído de 1.251 locutores, distribuídos conforme Figura 16. Em outra base de dados (OLIVEIRA, 2018a), as gravações foram feitas por 5 diferentes locutores em um estúdio com equipamentos profissionais. Cada um dos locutores repetiu as mesmas 100 palavras, criando um banco de voz com 500 arquivos de áudio. A gravação foi feita com uma frequência de amostragem de 44100 Hz e 16 bits. Esses bancos de dados foram utilizados neste trabalho devido a disponibilidade de amostras de sinais de voz e pela utilização em outros trabalhos relacionados ao tema abordado.

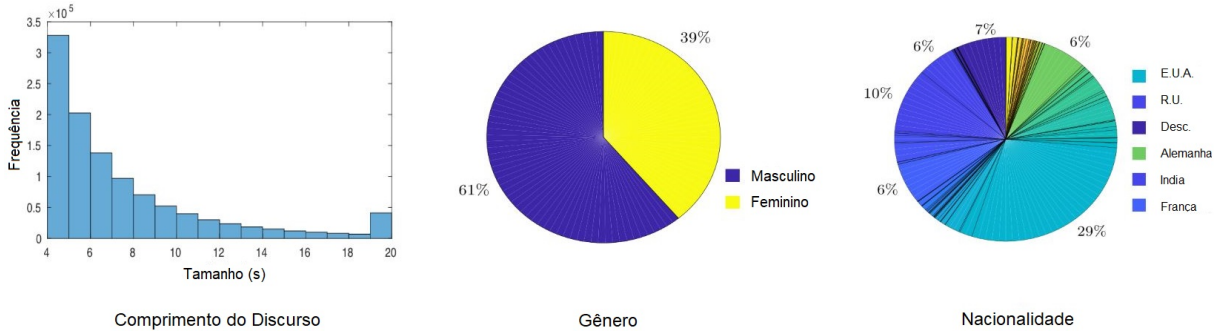

Figura 16: Composição do *dataset Voxceleb1*. Fonte: Adaptado de Nagrani, Chung e Zisserman (2017)

Já para a tarefa de reconhecimento de emoções, o banco de falas também pode ser chamado de banco de emoções (*emotion database*). Como exemplos de bancos de emoções públicos utilizados na literatura pode-se citar: o *Surrey Audio-Visual Expressed Emotion(SAVEE)* (JACKSON; HAQ, 2014) e o *Berlin Database of Emotional Speech (Emo-DB)* (BURKHARDT et al., 2005a), ambos disponíveis online. O banco SAVEE disponibiliza 4 locutores interpretando 7 diferentes emoções falando inglês britânico: raiva, nojo, medo, alegria, neutralidade, tristeza e surpresa, os sinais foram gravados com uma frequência de amostragem de 44100 Hz, conforme Tabela 1. Já o EMO-DB possui 10 locutores, 5 homens e 5 mulheres, falando em alemão e interpretando 7 emoções: raiva, nojo, medo, alegria, calma, tristeza e tédio, gravado com uma taxa de amostragem de 16000 Hz (BURKHARDT et al., 2005a).As Tabelas 1 e 2 mostram a quantidade e distribuição dos arquivos nas bases de dados. Os *dataset* SAVEE e EMO-DB foram utilizados devido a sua utilização em outros trabalhos relacionados ao tema abordado, facilitando, assim, a comparação dos resultados.

Tabela 1: Distribuição dos arquivos de áudio na base de dados SAVEE.

Emoção	**Qtd. de arquivos**
Raiva	60
Nojo	60
Medo	60
Felicidade	60
Neutro	120
Tristeza	60
Surpresa	60
Total	480

Tabela 2: Distribuição dos arquivos de áudio na base de dados Emo-DB.

Emoção	**Qtd. de arquivos**
Raiva (*Anger*)	127
Ansiedade/Medo(*Anxiety/Fear*)	69
Tédio (*Boredom*)	81
Nojo (*Disgust*)	46
Felicidade (*Happiness*)	71
Neutro (*Neutral*)	79
Tristeza (*Sadness*)	62
Total	535

3.3 VETOR DE CARACTERÍSTICAS UTILIZADO

Nos experimentos realizados foram utilizados duas configurações de vetores de atributos do sinal de voz. Na Tabela 3 é possível verificar os atributos e as quantidades utilizadas como entrada para os classificadores neurais. A configuração **vt-1** possui 46 descritores extraídos de cada *frame*, já a configuração **vt-2** possui 45 características. A diferença entre os vetores testados está na utilização, ou não, do atributo frequência fundamental (f_0) (*Picth*).

3.4 VALIDAÇÃO CRUZADA

Para medir a eficiência dos sistemas de classificação são utilizadas diversas ferramentas. A divisão do banco de dados em conjuntos de treino e teste é uma abordagem bastante utilizada. Em geral, os dados são subdivididos em proporções de 10 a 30% para o conjunto de teste e de 90 a 70% para o conjunto de treino.

A validação cruzada é usualmente utilizada no processo de aprendizagem. Nessa técnica, o conjunto de dados disponível é particionado aleatoriamente em uma amostra de treinamento e uma amostra de teste. O *dataset* pode, ainda, ser subdividido em um conjunto de validação, usado para testar ou validar o modelo, nesse caso, para evitar

Tabela 3: Descritores e parâmetros propostos para o classificador MLP.

vt-1

Descritores	**Parâmetros**	**Qtd. por *frame***
MFCC	Coeficientes	15
Δ	Coeficientes	15
$\Delta\Delta$	Coeficientes	15
Pitch	Amplitude	1
Total		46

vt-2

Descritores	**Parâmetros**	**Qtd. por *frame***
MFCC	Coeficientes	15
Δ	Coeficientes	15
$\Delta\Delta$	Coeficientes	15
Total		45

a possibilidade de sobreajuste, o desempenho do classificador selecionado é medido no conjunto de teste, que é diferente do subconjunto de validação (HAYKIN et al., 2009).

A Figura 17 mostra o diagrama representativo da distribuição das realizações por validação cruzada utilizando a técnica de *k-fold*, sendo k igual a 10, utilizada neste trabalho. A partir de uma das base de dados disponíveis, a cada iteração têm-se uma organização aleatória do novo conjunto de treinamento e teste. Deste modo, a cada iteração são aferidas medidas de avaliação de desempenho, Acurácia e Produto das eficiências e, por fim, a média desses parâmetros retorna uma avaliação mais precisa do desempenho do classificador.

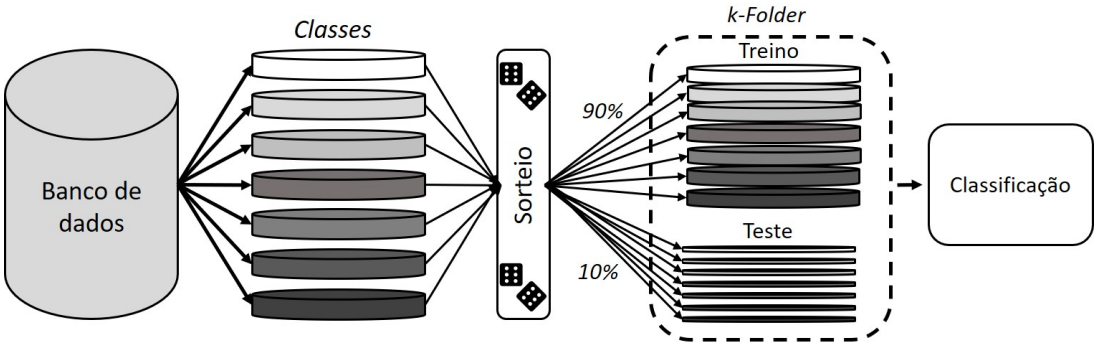

Figura 17: Diagrama representativo do sistema de validação cruzada com 10 folds. Fonte: Autor

3.5 IDENTIFICAÇÃO DO LOCUTOR

Essa parte do módulo tem como objetivo identificar o falante, para tanto, nessa fase do trabalho, realizou-se a implementação de um sistema onde extraem-se características, treinam-se redes neurais MLP, Redes Neurais Recorrentes LSTM e Redes Neurais Recorrentes BILSTM .

Para extração de descritores do sinal de voz e treinamento dos classificadores foram utilizados bancos de dados de sinais de fala humana, sendo eles:

1. Banco de falas utilizado em Oliveira (2018a) composto de 5 locutores e 500 sinais, 100 sinais para cada falante, com frequência de amostragem de 44,1 kHz; e

2. Banco de falas Voxceleb1, disponível em Nagrani, Chung e Zisserman (2017), composto de mais de 100.000 sinais e 1.251 locutores, distribuídos conforme Figura 16 nessa etapa utilizaremos 1000 sinais para fins de comparação entre os bancos de dados. A fim de comparação entre as bases de dados Oliveira (2018a) e Voxceleb1 foram realizados experimentos com 5, 8, 11 e 14 locutores para a *database* Voxceleb1.

Os sinais que compõem os banco de dados foram divididos em dois conjuntos, treino e teste, sendo 90% dos arquivos para treinamento e 10% dos arquivos para teste. A fim de melhorar os resultados, utilizou-se a técnica de validação cruzada com 10-folds. Um diagrama simplificado das etapas do processo de programação do módulo de identificação do locutor pode ser verificada na Figura 18.

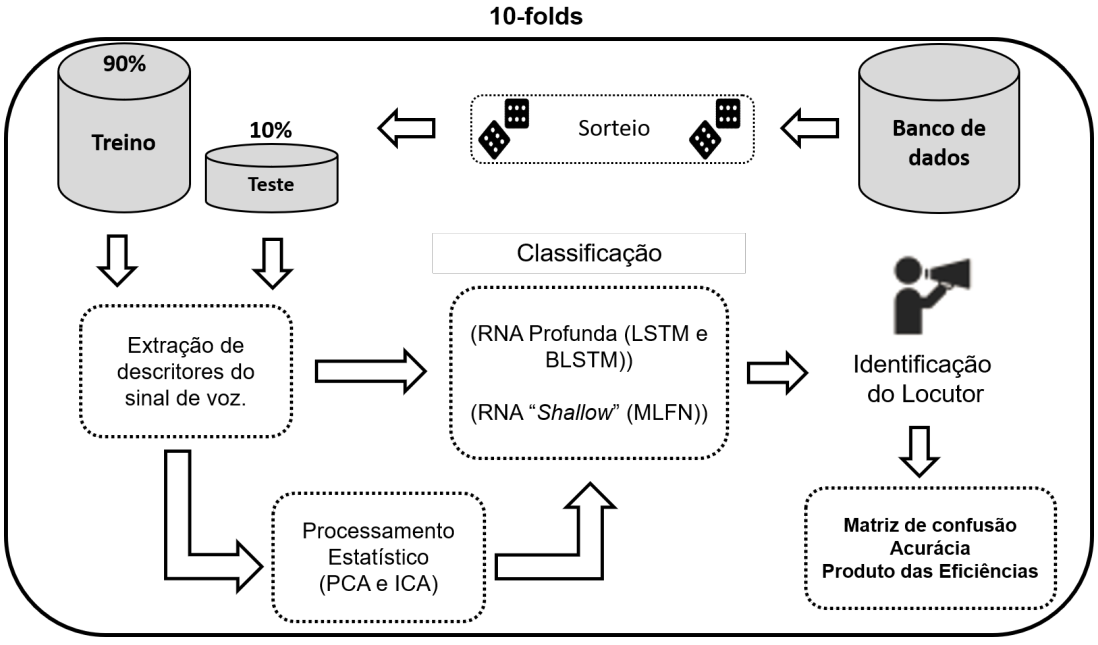

Figura 18: Diagrama de blocos simplificado do processo de ajuste de parâmetros do módulo de identificação do locutor. Fonte: Autor

A fim de obter os melhores resultados foram realizados vários testes com diferentes configurações, variando-se desde o número de descritores aos tipos de classificadores e entre os classificadores, número de neurônios e de camadas ocultas. Estes experimentos foram baseados em trabalhos como (OLIVEIRA; CERQUEIRA; FILHO, 2018b), (HERSHEY et al., 2017), (NERI, 2019) e (PENHA, 2018).

3.5.1 Classificador Neural MLP

A rede neural utilizada foi uma rede do tipo *feedforward* com uma camada de entrada, uma camada intermediária (oculta) e uma camada de saída, o algorítimo de treinamento utilizado foi o *Resilient Backpropagation* pela sua velocidade e baixa utilização de memória em relação a outros algorítimos baseados na descida do gradiente. A função de ativação utilizada foi a tangente hiperbólica com inicialização normalizada pois, para o treinamento de redes neurais "*shallow*", alcança melhor performance que a função sigmoide devido ao fato de ser centrada no zero. Outra motivação para a utilização das ferramentas citadas acima são os resultados alcançados no trabalho de Oliveira (2018a). O número de iterações e a taxa de aprendizagem foram escolhidos, inicialmente, de acordo com Oliveira (2018a). Na Tabela 4 é possível verificar as configurações dos parâmetros da rede neural MLP.

Tabela 4: Parâmetros de treinamento da rede neural MLP.

Parâmetros	**Configuração**
Função de ativação	tangente hiperbólica(VOGL et al., 1988)
Algoritmo de treinamento	*Resilient Backpropagation*
Número de neurônios	70
Número máximo de iterações	500
Taxa de aprendizagem	0,01
Cálculo do erro	Erro médio quadrático (MSE)

Diversas consultas foram realizadas na literatura para obter um valor de partida para o número de coeficientes mel cepstrais. No trabalho NERI (2019) foram utilizados 12 e 19 coeficientes MFCC, em Oliveira (2018a) foram utilizados 15 MFCC como descritores dos sinais de fala. Para os testes de performance dos descritores iniciou-se com 12 (doze) coeficientes MFCC. Depois, variou-se o número de coeficientes MFCC de 12 a 17, sempre concatenados com os coeficientes dinâmicos (DELTA e DELTA DELTA) e o *Pitch*.

Após a definição das quantidades de descritores e parâmetros, investigou-se o número de neurônios na camada oculta das redes neurais MLP de forma empírica. O número de neurônios na camada oculta teve como base o valor de 40 neurônios, conforme melhor resultado apresentado por Oliveira (2018a).

Afim de melhorar a eficiência de discriminação do classificador *Multi Layer Perceptron* foram utilizados métodos de pré-processamento estatístico, ICA e PCA, para redução da informação redundante e diminuição do número de parâmetros de entrada da rede neural.

3.5.2 Classificador Neural Profundo

Com o objetivo de obtenção dos melhores resultados, buscou-se comparar classificadores neurais artificiais. Além das redes neurais *Perceptron* de múltiplas camadas utilizou-se redes neurais recorrentes LSTM e BILSTM nos estudos. As redes neurais com aprendizado profundo são bastante utilizadas em aplicações de reconhecimento de imagens, visão computacional e análise de textos.

Na Tabela 5 é possível verificar os parâmetros de treinamento da rede neural profunda. Esses parâmetros foram ajustado de acordo com (HE et al., 2016), (MATHWORKS, 2020) e testes empíricos. Nos experimentos realizados, variou-se o número de iterações e a taxa de aprendizagem da rede. A tabela 6 mostra os tipos de camadas da rede LSTM e BLSTM utilizadas nos experimentos.

Tabela 5: Parâmetros de treinamento da rede neural profunda.

Parâmetros	**Configuração**
Função de ativação	ReLu e Sofmax
Algoritmo de otimização	*Adam (Adaptive moment estimation) e SGD (Stochastic Gradient Descent)*
Número máximo de iterações	30
Taxa de aprendizagem	0,0001
Função de custo	Entropia cruzada

3.6 IDENTIFICAÇÃO DE EMOÇÕES

Este submódulo tem como função identificar a emoção estereotipada no sinal de voz, para tanto, nessa fase do trabalho, realizou-se a implementação de um sistema, onde, extrai-se características, treinam-se redes neurais MLP, Redes Neurais Recorrentes LSTM e BILSTM.

A Figura 19 mostra um diagrama de blocos simplificado do módulo de identificação de emoções. Observa-se que a partir dos sinais disponíveis nos bancos de dados são formados os conjuntos de treino e testes. Estes conjuntos são rearranjados, de modo que, a cada *fold*, os arquivos de treino e teste sejam diferentes, mantendo-se a proporção de 90% dos arquivos para treinamento e 10% dos arquivos para teste. Os descritores utilizados foram os MFCC, Delta, Delta Delta e *Pitch*.

Tabela 6: Configurações das redes neurais LSTM e BLSTM testadas.

t-RNN 1

QTD	Descrição
1	Camada de entrada
1	Camada com célula LSTM com 500 unidades e 50% de *dropout*
1	Camada com célula LSTM com 200 unidades e 50% de *dropout*
7	Camada totalmente conectada com função de ativação softmax
1	Camada de Saída

t-RNN 2

QTD	Descrição
1	Camada de entrada
3	Camada com célula LSTM com 250 unidades e 50% de *dropout*
7	Camada totalmente conectada com função de ativação softmax
1	Camada de Saída

t-RNN 3

QTD	Descrição
1	Camada de entrada
3	Camada com célula BLSTM com 250 unidades e 50% de *dropout*
7	Camada totalmente conectada com função de ativação softmax
1	Camada de Saída

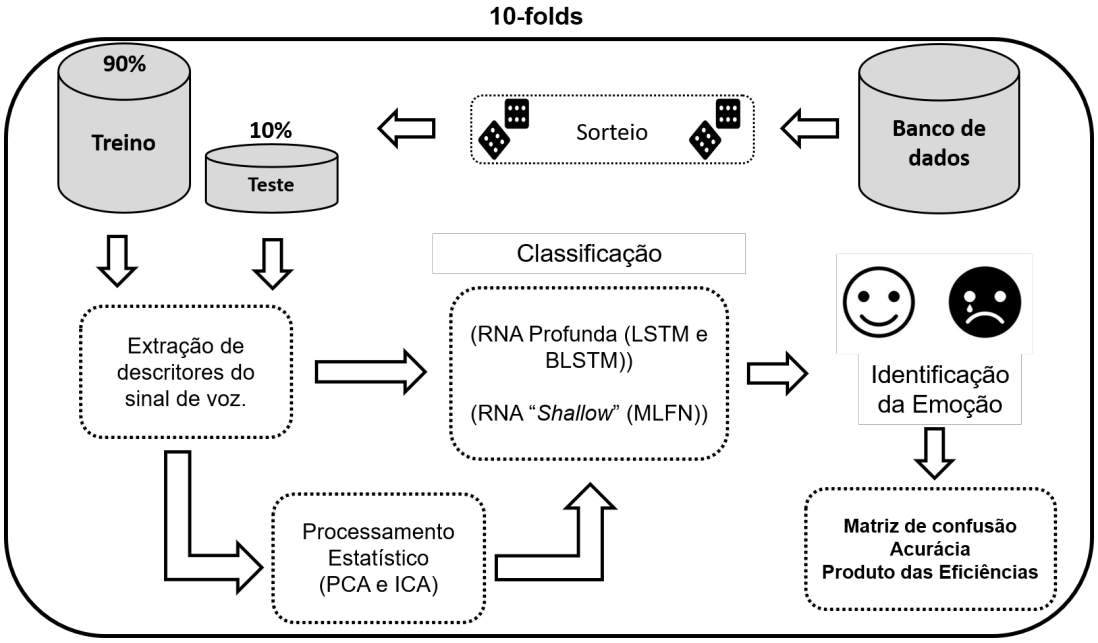

Figura 19: Diagrama de blocos simplificado do processo de ajuste de parâmetros do módulo de identificação de emoções. Fonte: Autor

Para esta etapa são necessários bancos de dados de emoções, por isso foram utilizados as seguintes bases de dados de sinais:

1. *Surrey Audio-Visual Expressed Emotion(SAVEE)*: com 4 locutores, idioma inglês

britânico, interpretando 7 emoções diferentes (raiva, nojo, medo, alegria, neutralidade, tristeza e surpresa), gravadas com uma frequência de amostragem de 44100 Hz (JACKSON; HAQ, 2014).

2. *Berlin Database of Emotional Speech (Emo-DB)*: com 10 locutores, 5 homens e 5 mulheres, falando em alemão e interpretando 7 emoções (raiva, nojo, medo, alegria, calma, tristeza e tédio), gravado com uma taxa de amostragem de 16000 Hz (BURKHARDT et al., 2005b).

Para facilitar a identificação das emoções contidas em cada banco serão utilizadas siglas para identificar as emoções. A tabela 7 mostra a relação entre as emoções contidas nos sinais em cada base de dados e as siglas utilizadas no trabalho.

Tabela 7: Tabela de correspondência entre siglas e emoções para os bancos de dados SAVEE e Emo-DB.

Sigla	**SAVEE**	**EMO-DB**
E01	*Raiva (A)*	*Ansiedade (An)/Medo (M)*
E02	*Nojo (D)*	*Nojo (D)*
E03	*Medo (M)*	*Felicidade (H)*
E04	*Felicidade (H)*	*Tédio (B)*
E05	*Neutra (N)*	*Neutra (N)*
E06	*Tristeza (S)*	*Tristeza (S)*
E07	*Surpresa (S)*	*Raiva (A)*

A fim de obter os melhores resultados foram realizados vários testes com diferentes configurações, variando-se desde o número de descritores aos tipos de classificadores e entre os classificadores, número de neurônios e de camadas ocultas. Nos próximos capítulos descreveremos os resultados obtidos nas diversas configurações utilizadas comparando-os com resultados de outros trabalhos.

4 RESULTADOS

Neste capítulo serão apresentados os resultados obtidos nas tarefas de reconhecimento automático de locutor e emoções através do sinal de voz utilizando a metodologia descrita no Capítulo 3.

Para a realização dos testes foi utilizado um computador com Sistema Operacional *Windows 10 Home Single Language*, processador Intel Core i7-8565U CPU @1.80GHz com 16 GB de memória DDR4.

4.1 IDENTIFICAÇÃO DO LOCUTOR

Neste tópico serão apresentados os resultados alcançados na tarefa de identificação automática do locutor.

4.1.1 Classificador Neural Artificial MLP

Nesta seção são descritos os testes realizados e os resultados alcançados na tarefa de reconhecimento automático do locutor com a utilização de um classificador neural MLP. Como orientação para as análises realizadas, além de pesquisa bibliográfica, foram utilizados os ajustes propostos nos melhores resultados alcançados em Oliveira (2018a). Para o caso do classificador neural MLP, tem-se: 40 neurônios na camada oculta, *Resilient Backpropagation* como Algoritmo de treinamento e Erro médio quadrático para o cálculo do erro. A partir dessas informações, os experimentos realizados buscaram definir o número de descritores do sinal de voz e a quantidade de neurônios na camada oculta, além de verificar os efeitos da aplicação de análise de componentes principais e análise de componentes independentes como pós-processamento ao vetor de características apresentado a entrada do classificador neural.

4.1.1.1 Variação do número de descritores

Com o objetivo de estimar o número de descritores do sinal de voz com o qual seria possível obter o melhor resultado de classificação, foram realizadas consultas na literatura e testes de performance. Para formação do vetor de características foram utilizados os coeficientes MFCC, Delta, Delta Delta e o *Pitch*, extraídos do sinal de voz, conforme definido no Capítulo 2 e Capítulo 3. Em NERI (2019) foram utilizados 12 coeficientes MFCC, Jagiasi et al. (2019) utilizaram 13 MFCC e Oliveira (2018a) utilizou 15 MFCC, 15 Coeficientes Delta e 15 Coeficientes DELTA DELTA como descritores de sinais da fala. A esses coeficientes foram concatenados os valores do *pitch*, realizando-se testes empíricos

iniciando com 37 parâmetros, seguindo-se de testes com 40, 43, 46, 49 e 52 parâmetros do sinal de voz.

Para a definição do número de parâmetros do vetor de características foram utilizados 40 neurônios na cama oculta, pois observou-se que o melhor resultado encontrado em Oliveira (2018a) foi alcançado utilizando essa quantidade de neurônios. A base de dados utilizada foi a disponibilizada em Oliveira (2018a).

Conforme a Figura 20, o melhor resultado para o classificador neural MLP ocorreu quando utilizou-se 46 parâmetros (15 primeiros MFCC, 15 coeficientes DELTA, 15 coeficientes DELTA DELTA e o *Pitch*). Nos testes realizados, todas as configurações foram processadas 10 vezes obtendo um produto das eficiências máxima de 95% e média de 92%. Observa-se que a configuração escolhida apresenta a maior média e a menor amplitude no resultado.

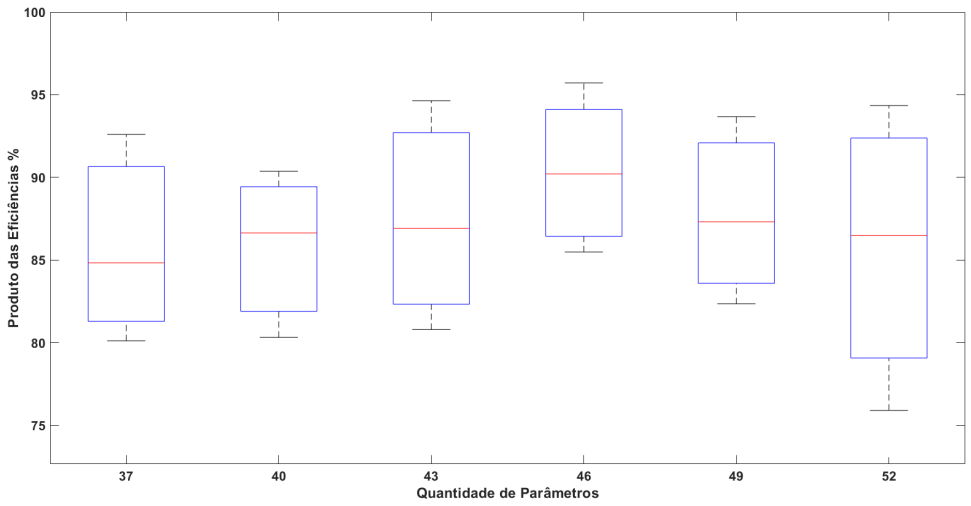

Figura 20: Variação no número de descritores para formação do vetor **vt-1**. Fonte: Autor

Nesta etapa também foram testados os efeitos da utilização do *pitch* de voz na identificação automática do locutor. Conforme Capítulo 3, os vetores **vt-1** e **vt-2** se diferenciam pela concatenação, ou não, do *pitch* aos coeficientes MFCC e suas derivadas. Na Tabela 8 comparam-se os resultados alcançados entre os vetores de características **vt-1** e **vt-2**.

Tabela 8: Produto das Eficiências (PEf), em %, da rede neural MLP para os vetores **vt-1** e **vt-2**.

Vetor de caract.	Parâmetros	PEf max.	PEf med.
vt-1	46	95	91 ± 3
vt-2	45	94	91 ± 5

4.1.1.2 Variação do número de neurônios na camada oculta

Definido os descritores e suas quantidades, partiu-se para a investigação do número de neurônios na camada oculta das redes neurais MLP. Foram realizados diversos testes variando-se o número de neurônios na camada oculta. A Figura 21 mostra os resultados obtidos.

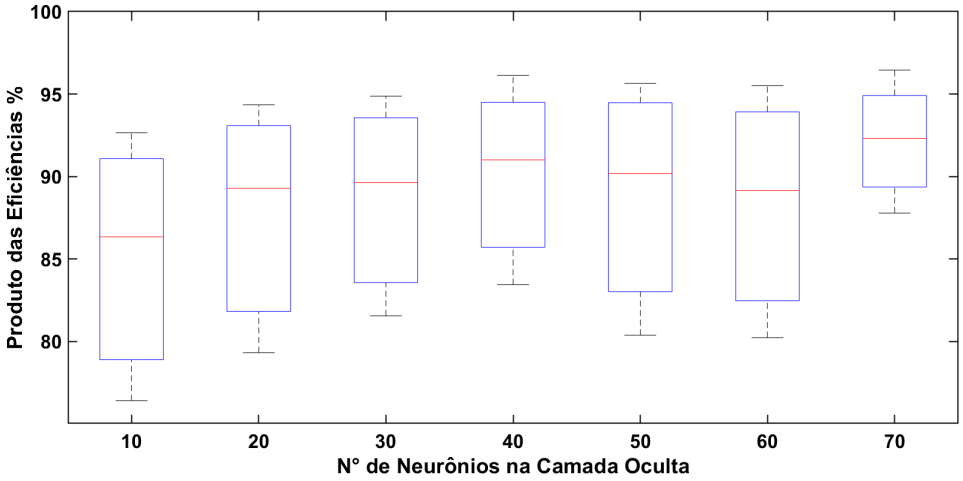

Figura 21: Variação no número de nerônios na camada oculta utilizando o **vt-1**. Fonte: Autor

Observa-se, na Figura 21, que o melhor resultado ocorreu quando foram utilizados 70 neurônios na camada oculta. Nos testes realizados, todas as configurações foram processadas 10 vezes obtendo um Produto das Eficiência máximo de 98%, médio de 94% e mínimo de 88% na classificação . Por meio da Tabela 9 é possível observar os valores obtidos nos testes realizados com os vetores de parâmetros **vt-1** e **vt-2**. Verifica-se, neste caso, aumento das eficiências máximas e médias em comparação quando foram utilizados 40 neurônios na camada oculta.

Tabela 9: Produto das Eficiências (PEf), em %, da rede neural MLP para os vetores de características 1 e 2.

Vetor de caract.	**Parâmetros**	**PEf max.**	**PEf med.**
vt-1	46	98	94 ± 3
vt-2	45	96	92 ± 3

4.1.1.3 Avaliação dos descritores em diferentes bases de dados

A partir dos resultados obtidos na Tabela 9, realizou-se estudos com os vetores **vt-1** e **vt-2** em bases de dados diferentes. Observa-se na Tabela 10 que o aumento no número de classes foi acompanhado pela diminuição na eficiência do classificador neural. Por exemplo,

para o caso com 14 locutores diferentes houve uma redução 12 pontos percentuais em relação ao experimento com 5 classes.

Tabela 10: Produto das Eficiências (PEf),em %, para a rede neural MLP em diferentes base de dados para o vetor **vt-1**.

Base de dados	**Locutores**	**PEf max.**	**PEf med.**
Oliveira (2018a)	5	98	94 ± 3
Voxceleb1	5	96	95 ± 1
Voxceleb1	8	92	88 ± 2
Voxceleb1	11	90	86 ± 2
Voxceleb1	14	86	84 ± 2

Com o objetivo de verificar a relevância do *pitch* de voz para a identificação automática do locutor realizou-se experimentos com o vetor **vt-2**. Este vetor é composto pelo coeficientes MFCC e suas derivadas, conforme pode ser observado no Capítulo 3. A Tabela 11 traz os resultados obtidos. Verifica-se a diminuição dos valores médios e máximos dos produtos das eficiências em relação aos resultados alcançados quando o vetor **vt-1** foi empregado.

Tabela 11: Produto das Eficiências (PEf), em %, para a rede neural MLP em diferentes base de dados para o vetor **vt-2**.

Base de dados	**Locutores**	**PEf max.**	**PEf med.**
(OLIVEIRA, 2018a)	5	97	89 ± 5
Voxceleb1	5	97	89 ± 5
Voxceleb1	8	93	87 ± 4
Voxceleb1	11	87	79 ± 5
Voxceleb1	14	81	73 ± 5

Por meio da matriz de confusão abaixo, Tabela 12, verifica-se que houve uma confusão de 10% entre os locutores Loc2 e Loc3, indicando um resultado de discriminação das classes para o classificador neural MLP alinhado ao trabalho de Oliveira (2018a) que obteve uma acurácia máxima de 97% para o mesmo problema.

Tabela 12: Matriz de confusão (em %) para uma rede neural MLFN com 70 neurônios na camada oculta para o banco de dados Oliveira (2018a).

	Loc1	**Loc2**	**Loc3**	**Loc4**	**Loc5**
Loc1	**100**	0	0	0	0
Loc2	0	**90**	10	0	0
Loc3	0	0	**100**	0	0
Loc4	0	0	0	**100**	0
Loc5	0	0	0	0	**100**

Na próxima seção será analisado o emprego de processamento estatístico, por meio da análise de componentes principais e análise de componentes independentes, com o objetivo de diminuição da informação redundante entre os dados a fim de melhorar a eficiência do classificador.

4.1.1.4 Redes Neurais MLP com pré-processamento estatístico

Afim de melhorar a eficiência de discriminação do classificador foram utilizados métodos de pré-processamento estatístico, como análise de componentes principais (PCA) e análise de componentes independentes (ICA), para redução da informação mútua e diminuição do número de parâmetros de entrada da rede neural. A Figura 22 mostra a curva de carga dos componentes da PCA para os descritores extraídos dos sinais do conjunto de treinamento do banco de dados Oliveira (2018a). Verifica-se que 99,9% da energia (variância) está armazenada nos 15 primeiros componentes.

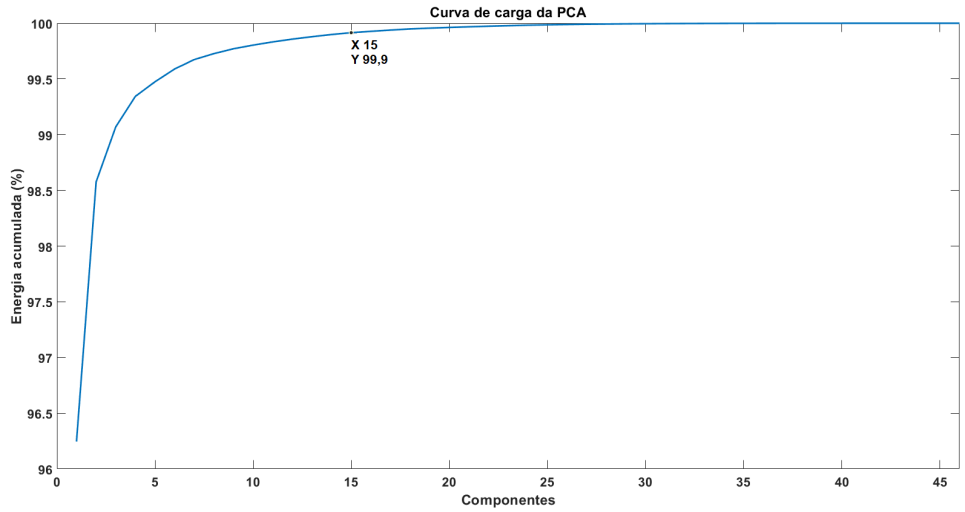

Figura 22: Curva de carga da PCA para o vetor **vt-1** extraído dos sinais do banco (OLIVEIRA, 2018a). Fonte: Autor

Na Figura 23 é possível observar o coeficiente de correlação entre os vetores extraídos dos sinais de áudio da base de dados (OLIVEIRA, 2018a). Verifica-se uma alta correlação entre parâmetros diferentes. Neste caso, quanto mais próximo de 1 ou -1, maior é a correlação entre os parâmetros. O coeficiente de correlação é um valor que fornece uma medida do grau de dependência entre variáveis. A ocorrência de sobreposição de informações dificulta a separação entre as classes.

Analisando os valores médios das componentes independentes para cada locutor, Figura 24, verifica-se que as componentes independentes possuem informações relevantes para distinguir sinais de diferentes locutores. Por exemplo, é possível discriminar os locutores 1, 2 e 3 entre si e entre os demais locutores. Por outro lado, observa-se semelhanças entre os locutores 4 e 5.

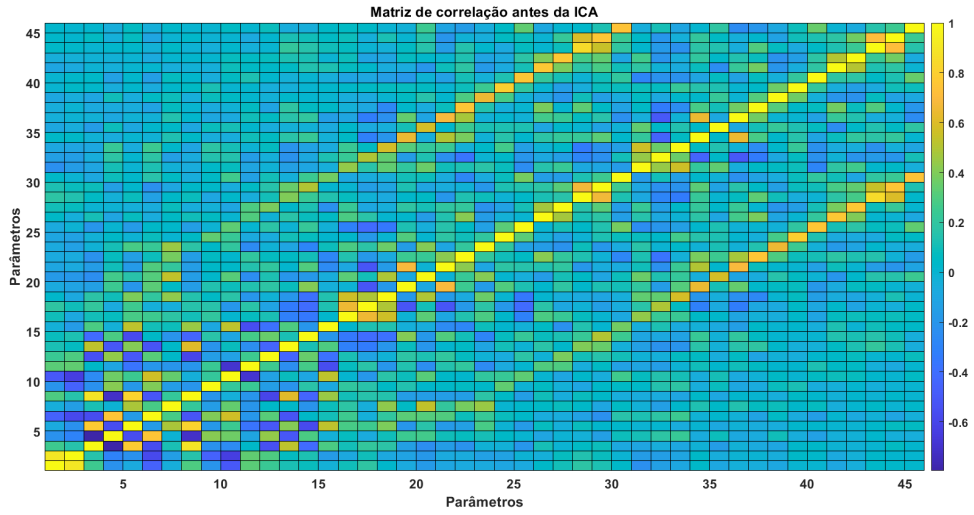

Figura 23: Matriz de correlação para os vetores de parâmetros do banco (OLIVEIRA, 2018a). Fonte: Autor

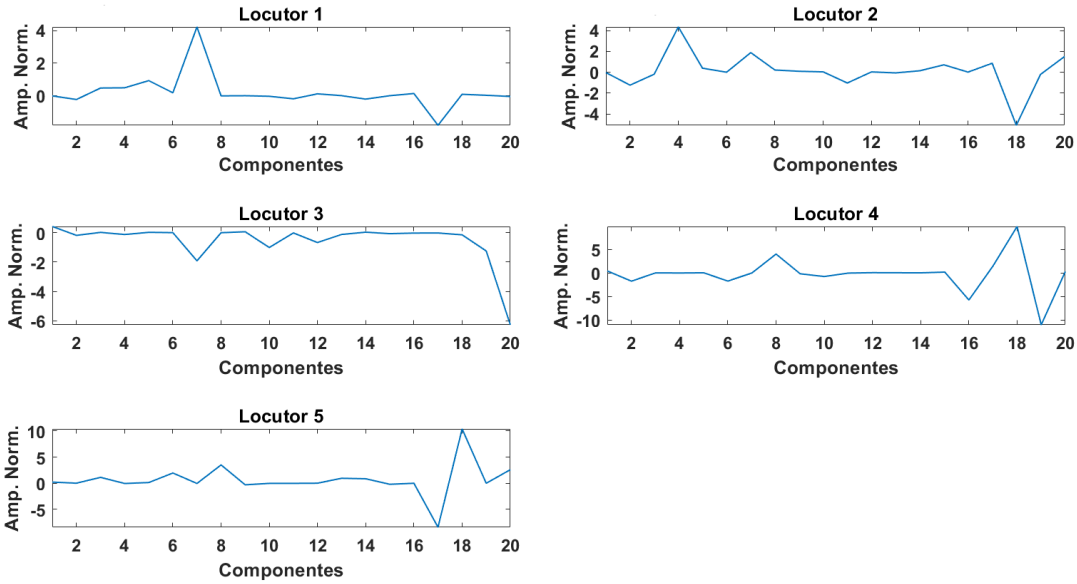

Figura 24: Valores médios das componentes independentes para cada locutor. Fonte: Autor

Para sinais do *dataset* Voxceleb1, Figura 25, é possível verificar que, neste caso, 99,9% da energia (variância) está armazenada em aproximadamente 43% dos componentes principais. Do mesmo modo que no banco de dados (OLIVEIRA, 2018a) verificou-se a possibilidade de redução no vetor de entrada do classificador com a aplicação desse tipo de transformação.

Na Figura 26 observar-se os coeficientes de correlação entre os parâmetros extraídos dos sinais de áudio do banco de dados Voxceleb1. Como no caso dos sinais do banco de dados (OLIVEIRA, 2018a), verifica-se que há elevada correlação entre parâmetros diferentes.

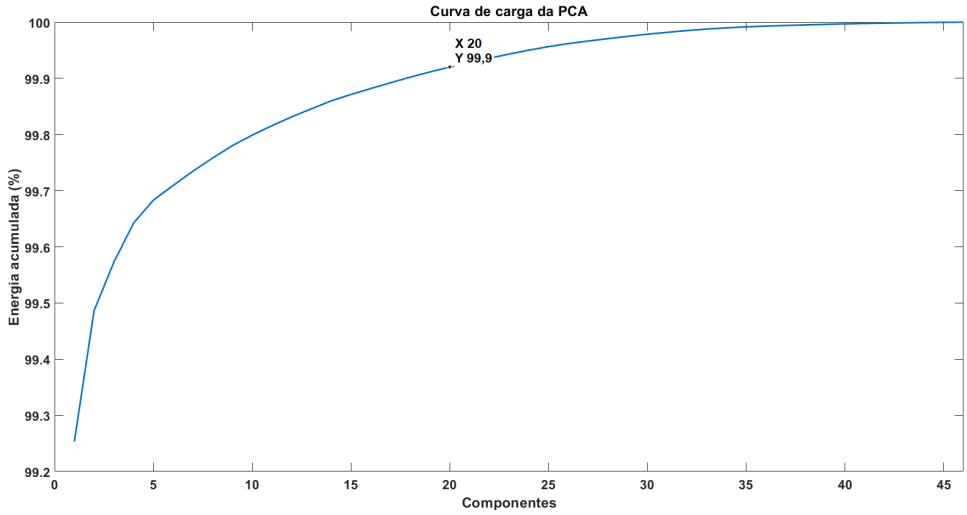

Figura 25: Curva de carga da PCA para sinais do banco Voxceleb1. Fonte: Autor

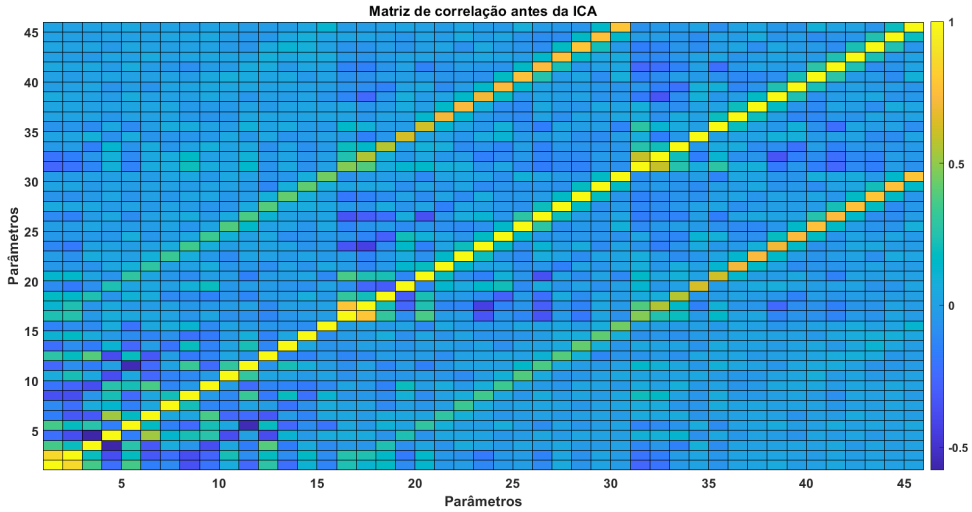

Figura 26: Matriz de correlação para os parâmetros antes da aplicação da ICA para sinais do Banco de Dados Voxceleb1. Fonte: Autor

Uma comparação entre os valores médios do produto das eficiências para os vetores de características com e sem aplicação de pré-processamento estatístico pode ser verificado nas Tabela 13. Os vetores **vt-1** e **vt-2** são compostos, respectivamente, dos MFCC e suas derivadas e dos MFCCs, suas derivadas e do *pitch* de voz. Os resultados são mostrados através das médias e desvios padrão do produto das eficiências do classificador MLFN. Observa-se que a utilização de técnicas de redução da informação redundante e diminuição do número de parâmetros de entrada aumentou a eficiência de discriminação do classificador neural MLFN em até 4 pontos percentuais. Por outro lado, é possível verificar que o aumento do número de classes foi acompanhado pelo decréscimo na eficiência do classificador mesmo quando utilizadas técnicas de processamento estatístico.

Na próxima seção serão mostrados os resultados dos testes para o classificador neural profundo.

Tabela 13: Produto das Eficiências médias (PEf), em %, para uma rede neural MLP com e sem aplicação de pré-processamento estatístico ao vetor de entrada.

vt-1		Proces. Estatístico		
Base de dados	**Locutores**	**Sem**	**ICA**	**P-ICA**
Oliveira (2018a)	5	94 ± 3	96 ± 2	96 ± 2
Voxceleb1	5	93 ± 2	96 ± 2	96 ± 2
Voxceleb1	8	91 ± 4	90 ± 4	92 ± 3
Voxceleb1	11	86 ± 1	88 ± 2	90 ± 3
Voxceleb1	14	81 ± 1	84 ± 1	85 ± 4

vt-2		Proces. Estatístico		
Base de dados	**Locutores**	**Sem**	**ICA**	**P-ICA**
Oliveira (2018a)	5	92 ± 6	94 ± 4	95 ± 3
Voxceleb1	5	90 ± 6	94 ± 4	95 ± 2
Voxceleb1	8	89 ± 4	90 ± 2	91 ± 2
Voxceleb1	11	83 ± 6	89 ± 3	90 ± 2
Voxceleb1	14	81 ± 5	83 ± 2	84 ± 3

4.1.2 Classificador Neural Artificial Profundo

Neste tópico foram utilizadas técnicas de classificação baseadas em *Deep Learning (DL)* para comparação dos resultados obtidos quando foram utilizados classificadores neurais *feedforward* multicamadas com somente um única camada oculta.

A partir dos parâmetros utilizados na tarefa de identificação do locutor por meio das redes neurais artificiais do tipo MLP, buscou-se avaliar a melhor configuração para classificadores neurais profundos *LSTM (Long Short Term Memory)* e *BILSTM (Bidirectional Long Short Term Memory)*. Diversos testes foram realizados, variando-se o número de camadas, o número de neurônios e a composição do vetor de entradas. Neste caso, os dados processados são entendidos como uma sequência do número de parâmetros do sinal de voz pela a quantidade de sinais disponíveis. A Tabela 14 mostra as configurações utilizadas para as redes LSTM e BLSTM.

Tabela 14: Configurações das redes LSTM e BLSTM.

Parâmetros da rede	**Configuração**
Número de camadas	11, 12 e 14
Dimensão Camada de entrada	45 e 46
Número de Células de memória	500, 250 e 200

O melhor resultado foi alcançado com a configuração t-RNN 1. Todas as configuração utilizadas estão disponíveis no Capítulo 3. Observa-se, na Tabela 15, os valores médios do produto das eficiências dos testes realizados com rede neural recorrente com célula de memória LSTM. Os resultados variaram conforme o número de classes, locutores,

apresentados a rede e a aplicação, ou não, de pré-processamento. Por exemplo, nos casos dos vetores **vt-1** e **vt-2** houveram variações de 7 pontos entre os melhores resultados com pré-processamento e sem pré-processamento.

Tabela 15: Valores médios (em %) do PEf para os bancos de dados (OLIVEIRA, 2018a) e Voxceleb1 para a configuração t-RNN 1 e vetores de caracteristicas **vt-1** e **vt-2**.

vt-1		Proces. Estatístico		
Base de dados	**N° falantes**	**Sem**	**ICA**	**P-ICA**
Oliveira (2018a)	5	82 ± 2	87 ± 3	89 ± 3
Voxceleb1	5	80 ± 3	86 ± 2	88 ± 3
Voxceleb1	8	76 ± 4	78 ± 5	80 ± 5
Voxceleb1	11	74 ± 2	76 ± 2	76 ± 3
Voxceleb1	14	68 ± 3	71 ± 5	73 ± 4

vt-2		Proces. Estatístico		
Base de dados	**Locutores**	**Sem**	**ICA**	**P-ICA**
Oliveira (2018a)	5	81 ± 2	86 ± 3	88 ± 5
Voxceleb1	5	72 ± 4	74 ± 2	75 ± 2
Voxceleb1	8	68 ± 3	71 ± 6	72 ± 7
Voxceleb1	11	68 ± 4	70 ± 4	74 ± 5
Voxceleb1	14	67 ± 2	69 ± 2	70 ± 4

Na Tabela 16 observa-se a matriz de confusão para o banco de dados Voxceleb1 com 5 locutores. Verifica-se que as menores eficiências foram alcançadas pelos locutores 4 e 5, com 80%. Nesse caso, verificou-se confusões entre os locutores 4 e 5 e entre os locutores 2 e 5 (10%).

Tabela 16: Matriz de confusão (valores em %) para rede com banco de dados (OLIVEIRA, 2018a).

	Loc.1	**Loc.2**	**Loc.3**	**Loc.4**	**Loc.5**
Loc.1	**100**	0	0	0	0
Loc.2	0	**90**	0	0	10
Loc.3	0	0	**100**	0	0
Loc.4	0	0	0	**90**	10
Loc.5	0	0	0	20	**80**

4.1.2.1 Análise dos resultados

Para a tarefa de identificação automática do locutor independente do texto foram realizados diversos experimentos a fim de se obter o melhor sistema de classificação. Neste sentido, comparou-se os resultados das eficiências médias obtidas por meio da Tabela 17. Observa-se que os Classificadores MLP+ e t-RNN 1+ obtiveram melhor resultado

de discriminação que os classificadores MLP e t-RNN 1. O símbolo + após a sigla do classificador significa que houve processamento dos vetores de características apresentados a rede neural MLP e a rede neural recorrente (t-RNN 1). Foi possível verificar que técnicas de processamento estatístico de sinais como o ICA possuem um impacto positivo na eficiência dos classificadores. Esse resultado pode ser atribuído ao fato das componentes independentes serem não-correlacionadas. O uso desse tipo de técnica aumentou o desempenho médio do classificador MLP+ em até 4 pontos percentuais em comparação com o classificador neural MLP sem a utilização de pré-processamento estatístico. Já o desempenho do classificador t-RNN 1+ foi elevado em até 8 pontos percentuais em comparação com o classificador neural t-RNN 1 sem a utilização de pré-processamento estatístico.

Em outra análise, verificou-se o impacto do uso do *pitch* como descritor do sinal de voz. Observou-se que, na maioria dos testes, o vetor de características **vt-1** obteve maiores médias e menores desvios padrão do que o vetor **vt-2**. Porém o aumento não foi expressivo, variando entre 1 ou 2 pontos percentuais.

Tabela 17: Comparação do PEF, em %, para os diversos classificadores.

vt-1		Classificadores			
Base de dados	Locutores	MLP	MLP+	t-RNN 1	t-RNN 1+
Oliveira (2018a)	5	94 ± 3	96 ± 2	82 ± 4	89 ± 3
Voxceleb1	5	93 ± 2	96 ± 2	80 ± 3	88 ± 3
Voxceleb1	8	91 ± 4	95 ± 3	77 ± 3	80 ± 5
Voxceleb1	11	86 ± 1	90 ± 3	74 ± 2	76 ± 3
Voxceleb1	14	81 ± 3	85 ± 4	68 ± 3	73 ± 4
vt-2		Classificadores			
Base de dados	Locutores	MLP	MLP+	t-RNN 1	t-RNN 1+
Oliveira (2018a)	5	94 ± 3	96 ± 3	81 ± 2	88 ± 5
Voxceleb1	5	94 ± 1	95 ± 3	72 ± 4	75 ± 2
Voxceleb1	8	90 ± 4	92 ± 3	68 ± 3	74 ± 5
Voxceleb1	11	85 ± 1	90 ± 3	68 ± 4	72 ± 7
Voxceleb1	14	80 ± 2	85 ± 4	67 ± 2	70 ± 4

A Tabela 18 mostra os tempos médios de execução das rotinas de extração de descritores e reconhecimento do locutor para os *dataset* Oliveira (2018a) e Voxceleb1 utilizando rede neural *Feedfoward Multi Layer Perceptron* e rede neural recorrente. É possível verificar que o maior impacto foi decorrente da utilização do classificador neural recorrente com célula de memória LSTM. A aplicação de pós-processamento não influenciou significativamente o tempo de execução do algorítimo de reconhecimento do locutor.

Comparando-se com outros trabalhos, os resultados foram superiores aos obtidos por (OLIVEIRA; CERQUEIRA; FILHO, 2018b) que obteve uma acurácia máxima de 94% e um produto das eficiências máxima de 93% na identificação dos locutores do bancos de dados Oliveira (2018a).

Tabela 18: Comparação dos tempos (em segundos) de processamento para os diversos classificadores.

		Classificadores			
Base de dados	Locutores	MLP	MLP+	t-RNN 1	t-RNN 1+
(OLIVEIRA, 2018a)	5	$1,21 \pm 0,03$	$1,22 \pm 0,02$	$1,97 \pm 0,03$	$1,98 \pm 0,04$
Voxceleb1	5	$1,23 \pm 0,10$	$1,23 \pm 0,10$	$1,96 \pm 0,02$	$1,92 \pm 0,15$
Voxceleb1	8	$1,32 \pm 0,03$	$1,34 \pm 0,10$	$2,42 \pm 0,34$	$2,40 \pm 0,30$
Voxceleb1	11	$1,41 \pm 0,05$	$1,42 \pm 0,04$	$2,94 \pm 0,36$	$2,89 \pm 0,43$
Voxceleb1	14	$1,84 \pm 0,21$	$1,85 \pm 0,33$	$3,01 \pm 0,09$	$3,03 \pm 0,12$

4.2 IDENTIFICAÇÃO DE EMOÇÕES

Essa parte do módulo tem como objetivo identificar a emoção estereotipada na fala, para tanto, assim como na etapa de identificação do locutor, realizou-se a implementação de um sistema, onde, são extraídas características, treinam-se redes neurais MLP e redes neurais recorrentes LSTM e avaliam-se os resultados.

4.2.1 Classificador MLP

Como na tarefa de identificação do locutor foram realizados testes para o alcance do melhor resultado do classificar neural *Single Hidden Layer Feedforward Neural Network (SLFN)*. Os resultados obtidos são mostrados na Tabela 19.

Nos experimento de identificação automática das emoções foram utilizados o mesmo número e tipo de descritores usados na etapa de identificação do locutor, são eles: 45 descritores (MFCC, DELTA e DELTA DELTA) e 46 descritores (MFCC, DELTA, DELTA DELTA e *Pitch*). As configurações das redes neurais foram as mesmas utilizadas na etapa de identificação do locutor.

Na Tabela 19 verifica-se os melhores resultados, nos testes realizados, todas as configurações foram processadas 10 vezes. Observa-se que houve um melhor desempenho de classificação para o vetor **vt-1** utilizando sinais do banco EMO-DB obtendo um PEf médio de 78% e máximo de 87%.

Na matriz de confusão abaixo, Tabela 20, verifica-se que houve confusão entre diversas emoções do banco de dados SAVEE. Nesse sentindo, destacam-se as confusões entre as emoções raiva (E01) e surpresa (E07), emoções negativas, com 33,7% e entre as emoções raiva, medo (E03) surpresa com 33,7% e 16,7%, respectivamente. A emoção surpresa obteve uma menor eficiência de classificação comparando-se com as demais classes. Esse resultado pode ser atribuído a similaridade entre essas emoções dificultando a tarefa do classificador em reconhecer os padrões.

Tabela 19: Produto das Eficiências (em %) para a rede neural MLP em diferentes base de dados e vetores de características.

vt-1

Base de dados	PEf max.	PEf med.
SAVEE	77	71 ± 4
EMO-DB	87	78 ± 5

vt-2

Base de dados	PEf max.	PEf med.
SAVEE	73	71 ± 1
EMO-DB	84	77 ± 6

Tabela 20: Matriz de confusão do banco de dados SAVEE para uma rede neural MLP com 70 neurônios na camada escondida e **vt-1**. (Valores em %)

	E01	**E02**	**E03**	**E04**	**E05**	**E06**	**E07**
E01	**66,7**	0,0	0,0	0,0	0,0	0,0	33,3
E02	0,0	**83,3**	0,0	0,0	16,7	0,0	0,0
E03	0,0	0,0	**83,3**	0,0	0,0	16,7	0,0
E04	16,7	0,0	0,0	**83,3**	0,0	0,0	0,0
E05	0,0	0,0	0,0	0,0	**100,0**	0,0	0,0
E06	0,0	0,0	16,7	0,0	0,0	**83,3**	0,0
E07	33,7	0,0	16,7	0,0	0,0	0,0	**50,0**

4.2.1.1 Redes Neurais MLP com pré-processamento estatístico

Afim de melhorar a eficiência de discriminação do classificador foram utilizados métodos de pré-processamento estatístico, ICA e PCA, para redução da informação redundante e diminuição do número de parâmetros de entrada da rede neural.

Em uma primeira análise, foi avaliado, por meio da Figura 27, a curva de carga dos componentes da PCA para sinais do banco de dados SAVEE. É possível verificar que 99,9% da energia (variância) está armazenada nos 20 primeiros componentes.

Já em uma outra análise, verificou-se, através da Figura 28, a correlação entre os atributos extraídos do conjunto de treino do banco de dados SAVEE. Observou-se alta correlação entre parâmetros diferentes. A partir dessas informações verifica-se que a utilização de ICA pode auxiliar o classificador, pois nesta técnica as componentes independentes são não-correlacionadas.

Na Figura 29 verifica-se que as componentes independentes possuem informações relevantes para distinguir sinais de diferentes emoções. Quando seus valores médios são mostrados por meio de um gráfico é possível observar, através dos perfis obtidos, que há semelhanças entre as emoções medo (E03) e surpresa (E07) e diferença entre a emoção tristeza (E06) e as demais emoções.

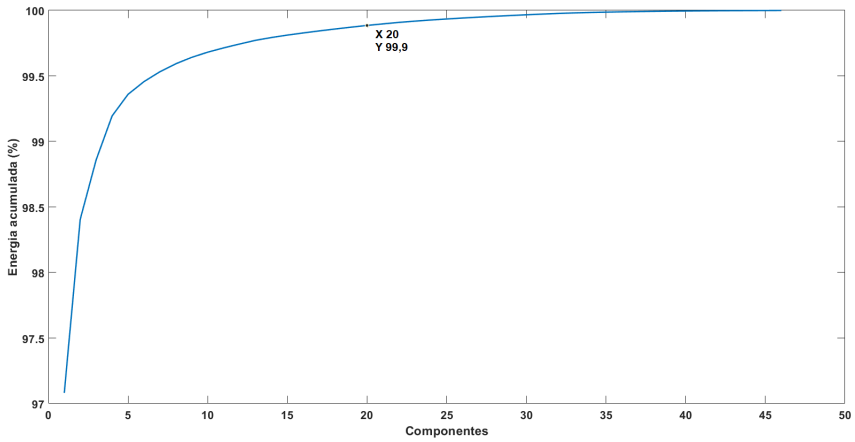

Figura 27: Curva de carga da PCA para sinais do *dataset* SAVEE. Fonte: Autor

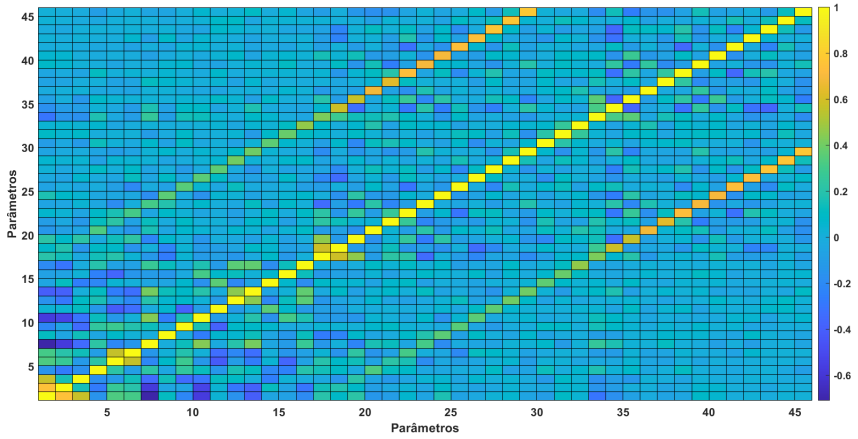

Figura 28: Matriz de correlação antes da ICA para emoções do banco de dados SAVEE . Fonte: Autor.

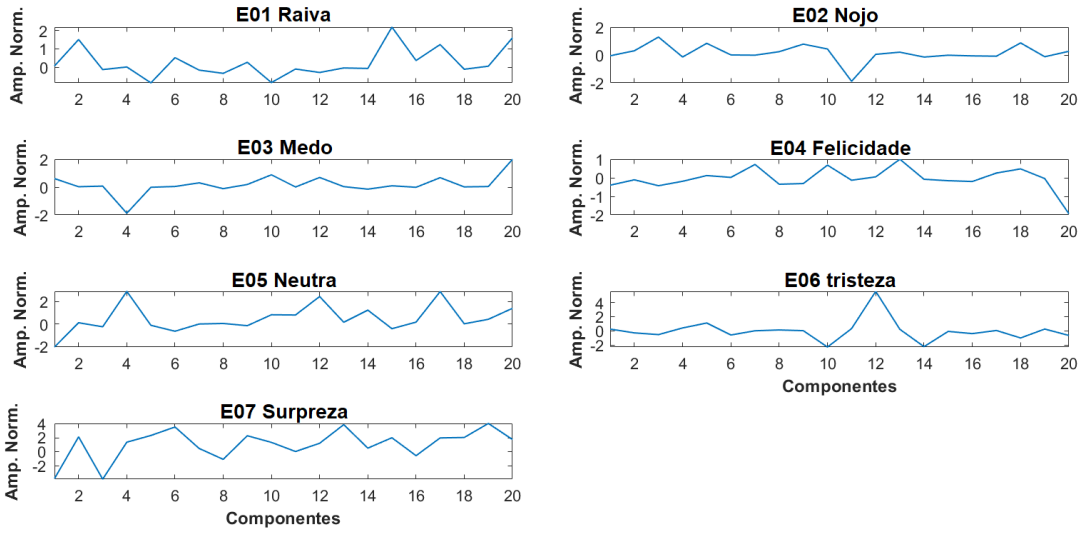

Figura 29: Valores médios das componentes independentes para cada Emoção no banco de dados SAVEE . Fonte: Autor.

Já para sinais do banco de dados Emo-DB , Figura 30, é possível verificar que, neste caso, 99,9% da energia (variância) está armazenada nos 15 primeiros componentes.

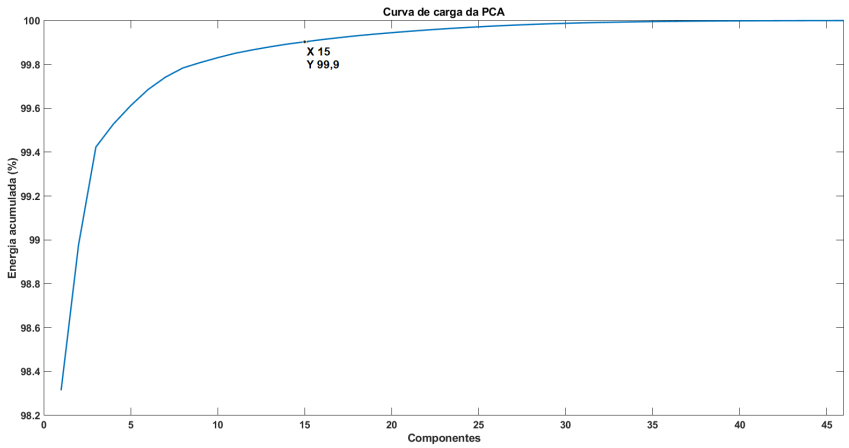

Figura 30: Curva de carga da PCA para sinais do banco Emo-DB. Fonte: Autor

Na Figura 31 observar-se os coeficientes de correlação entre os parâmetros extraídos dos sinais de áudio do banco de dados Emo-DB. De forma análoga aos atributos extraídos do banco de dados SAVEE, verifica-se que a utilização de ICA pode auxiliar o classificador, pois nesta técnica as componentes independentes são não-correlacionadas.

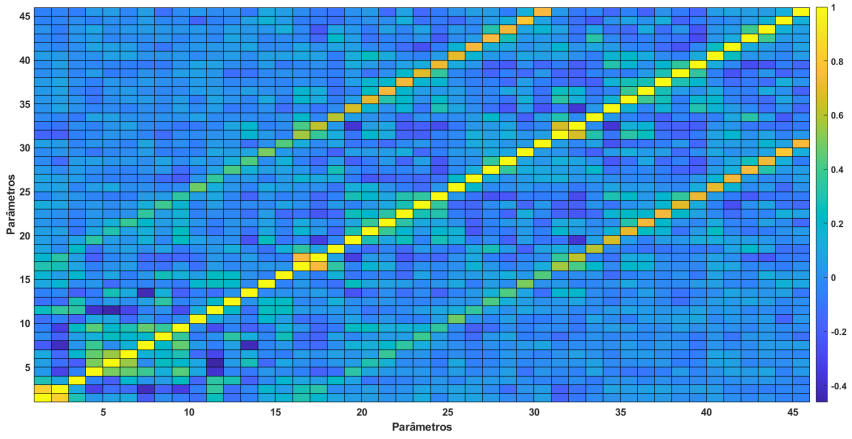

Figura 31: Matriz de correlação para os parâmetros antes da ICA para emoções do banco de dados Emo-DB. Fonte: Autor

Por meio da Figura 32 é possível observar os gráficos dos valores médios das componente independente para cada emoção do banco de dados Emo-DB. Nota-se que as emoções nojo (E02) e tristeza (E06) possuem um perfil diferente das outras emoções. Por outro lado há semelhanças entre as emoções felicidade (E03), raiva (E07) e tristeza (E06) e entre as emoções neutra (E05) e raiva (E06).

Uma comparação entre os resultados para vetores de características com e sem processamento estatístico pode ser verificado na Tabela 21. Observa-se que a utilização de técnicas de redução da informação redundante e diminuição do número de parâmetros

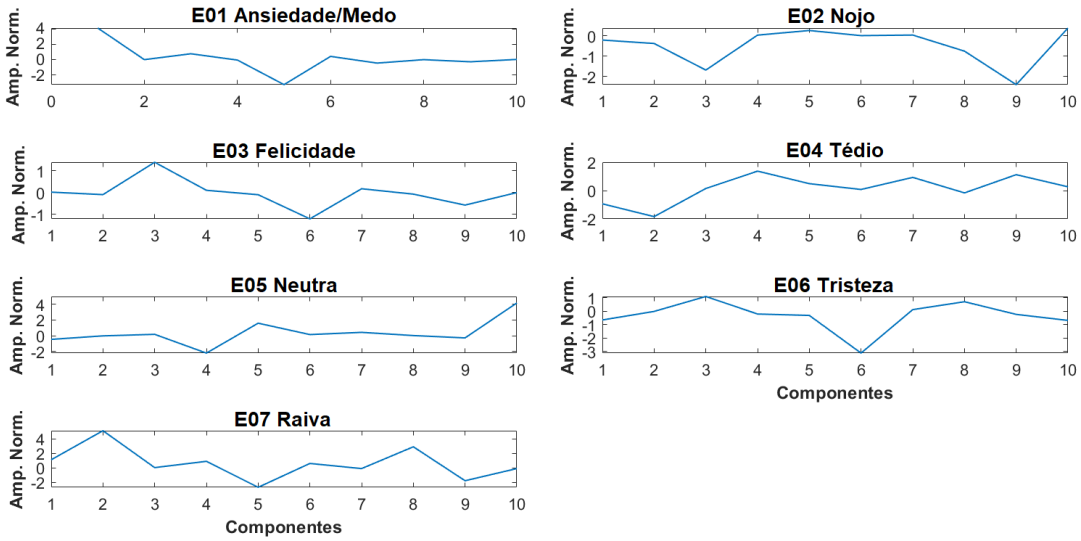

Figura 32: Valores médios das componentes independentes para cada Emoção no banco de dados Emo-DB . Fonte: Autor.

de entrada aumentou a eficiência de discriminação do classificador MLFN. Por exemplo, para o banco Emo-DB houve um aumento de eficiência na discriminação em até 8 pontos percentuais. Já no caso do *database* SAVEE, houve um aumento de eficiência de até 16 pontos percentuais quando utilizou-se pré-processamento estatístico. Os melhores resultados foram alcançados para o vetor de características **vt-1**.

Tabela 21: Produto das Eficiências máximas (PEf), em %, para rede MLP com e sem pré-processamento estatístico.

vt-1	**Proces. Estatístico**		
Base de dados	**Sem Proc.**	**ICA**	**P-ICA**
SAVEE	70 ± 3	83 ± 5	86 ± 5
EMO-DB	79 ± 4	86 ± 4	87 ± 5

vt-2	**Proces. Estatístico**		
Base de dados	**Sem Proc.**	**ICA**	**P-ICA**
SAVEE	71 ± 2	81 ± 6	80 ± 7
EMO-DB	77 ± 6	84 ± 5	84 ± 7

Na Tabela 22 observa-se a matriz de confusão para sinais com pré-processamento para o banco de emoções SAVEE. É possível verificar que a emoção raiva não foi confundida com nenhuma outra emoção, já a emoção felicidade foi confundida com a emoção raiva e neutra.

Já para o banco de emoções Emo-DB, matriz de confusão é mostrada na Tabela 23. Observa-se que as emoções desgosto, tédio, tristeza e raiva também não foram confundidas

Tabela 22: Matriz de confusão (valores em %) do banco de dados SAVEE e pré-processamento por P-ICA.

	E01	**E02**	**E03**	**E04**	**E05**	**E06**	**E07**
E01	**100,0**	0,0	0,0	0,0	0,0	0,0	0,0
E02	0,0	**100,0**	0,0	0,0	0,0	0,0	0,0
E03	0,0	0,0	**100,0**	0,0	0,0	0,0	0,0
E04	33,3	0,0	0,0	**50,0**	16,7	0,0	0,0
E05	0,0	0,0	0,0	0,0	**100,0**	0,0	0,0
E06	0,0	0,0	0,0	0,0	0,0	**100,0**	0,0
E07	0,0	0,0	0,0	0,0	0,0	0,0	**100,0**

com nenhuma outra emoção, já a emoção medo foi confundida com a emoção tédio e neutra.

Tabela 23: Matriz de confusão (valores em %) do banco de dados Emo-DB e pré-processamento por P-ICA.

	E01	**E02**	**E03**	**E04**	**E05**	**E06**	**E07**
E01	**83,3**	0,0	0,0	0,0	16,7	0,0	0,0
E02	0,0	**100,0**	0,0	0,0	0,0	0,0	0,0
E03	0,0	16,7	**83,3**	0,0	0,0	0,0	0,0
E04	0,	0,0	0,0	**100**	0,0	0,0	0,0
E05	16,7	0,0	0,0	0,0	**83,3**	0,0	0,0
E06	0,0	0,0	0,0	0,0	0,0	**100,0**	0,0
E07	0,0	0,0	0,0	0,0	0,0	0,0	**100,0**

4.2.2 Classificador Neural Artificial Profundo

Neste tópico foram utilizadas técnicas de classificação baseadas em *Deep Learning (DL)* para comparação dos resultados com classificadores neurais *shallow* com somente um única camada oculta.

A partir dos parâmetros utilizados na tarefa de identificação de emoções utilizando redes neurais artificiais do tipo MLP, buscou-se avaliar a melhor configuração para classificadores neurais profundos do tipo *Long Short Term Memory (LSTM)* e *Bidirecional Long Short Term Memory (BLSTM)*.

Através da Tabela 24 é possível observar resultados dos testes dos classificadores neurais recorrentes com diferentes bases de dados. Para a rede neural recorrente o melhor resultado foi alcançado com a rede t-RNN 1. A configuração da rede t-RNN 1 pode ser verificada no Capítulo 3.

Na Tabela 25 verifica-se a matriz de confusão para resultados dos testes dos classificadores neurais recorrentes com base de dados Emo-DB. Neste caso houveram

Tabela 24: Valores médios, em %, do PEf para a rede neural recorrente t-RNN 1 e os *dataset* SAVEE e Emo-DB.

vt-1

Base de dados	**Sem Proc.**	**ICA**	**P-ICA**
SAVEE	65 ± 3	78 ± 3	79 ± 3
EmoDB	69 ± 2	78 ± 3	80 ± 4

vt-2

Base de dados	**Sem Proc.**	**ICA**	**P-ICA**
SAVEE	62 ± 1	73 ± 5	75 ± 2
EmoDB	63 ± 3	75 ± 2	78 ± 3

confusões entre diversas emoções, por exemplo: emoções medo e neutra; emoções tédio e tristeza; emoções neutra e raiva; emoções tristeza e neutra; emoções raiva, medo e nojo. Contudo, o classificador confundiu a emoção neutra e tristeza com outras três emoções.

Tabela 25: Matriz de confusão (valores em %) dos resultados para o banco de dados Emo-DB para uma rede **t-RNN 1**.

	E01	**E02**	**E03**	**E04**	**E05**	**E06**	**E07**
E01	**100,0**	0,0	0,0	0,0	0,0	0,0	0,0
E02	0,0	**83,3**	0,0	0,0	16,7	0,0	0,0
E03	0,0	0,0	**100,0**	0,0	0,0	0,0	0,0
E04	0,0	0,0	0,0	**83,3**	0,0	16,7	0,0
E05	0,0	0,0	0,0	0,0	**83,3**	0,0	16,7
E06	0,0	0,0	0,0	0,0	16,7	**83,3**	0,0
E07	16,7	16,7	0,0	0,0	0,0	00,0	**66,7**

4.2.3 Análise dos Resultados

Para a tarefa de identificação de emoções através do sinal de voz foram realizados diversos experimentos a fim de obter o melhor sistema de classificação, maior discriminação (maior e PEf) e menor tempo de processamento. Neste sentido, comparou-se os resultados do produto das eficiências médias obtidas por meio da Tabela 26. Observa-se que ambos os classificadores neurais, MLP e RNN, obtiveram melhor desempenho quando foi utilizado o vetor de características **vt-1**. O melhor resultado foi alcançado para o classificador neural MLP+, pois obteve uma eficiência média maior que o classificador t-RNN 1+ e houve diminuição da confusão entre emoções negativas e positivas. Sendo assim, o *pitch* de voz pode ser considerado um descritor importante na identificação de emoções por meio do sinal de voz. Em outra análise, foi possível verificar que a aplicações de técnicas como o ICA possuem um impacto positivo na eficiência dos classificares utilizados. O uso desse tipo de análise aumentou o produto das eficiências em até 15 pontos percentuais. Durante

os testes verificou-se que 43% das componentes carregam cerda de 99% da informação do sinal original.

Tabela 26: Comparação dos valores médios do PEf, em %, para os *dataset* SAVEE e Emo-DB.

vt-1	Classificador			
Base de dados	**MLP**	**MLP+**	**t-RNN 1**	**t-RNN 1+**
SAVEE	70 ± 3	86 ± 5	65 ± 3	79 ± 3
EmoDB	79 ± 4	87 ± 5	69 ± 2	80 ± 4

vt-2	Classificador			
Base de dados	**MLP**	**MLP+**	**t-RNN 1**	**t-RNN 1+**
SAVEE	71 ± 2	80 ± 7	62 ± 1	75 ± 2
EmoDB	77 ± 6	84 ± 7	63 ± 3	78 ± 3

Tabela 27: Comparação dos tempos de processamento (em segundos) para os diversos classificadores.

Base de dados	**MLP**	**MLP+**	**t-RNN 1**	**t-RNN 1+**
SAVE	$1,83 \pm 0,18$	$1,84 \pm 0,04$	$2,97 \pm 0,05$	$2,94 \pm 0,024$
Emo-DB	$1,92 \pm 0,75$	$1,95 \pm 0,05$	$2,95 \pm 0,03$	$2,99 \pm 0,12$

A partir dos resultados obtidos é possível verificar que a utilização de técnicas de pré-processamento estatísticos aumentou a eficiência dos classificadores neurais sem elevar significativamente o tempo de processamento. Comparando-se com outros trabalhos os resultados foram superiores aos obtidos por Oliveira, Cerqueira e Filho (2018b) que alcançaram uma acurácia de 63% e um produto das eficiências de 55% na identificação da emoção estereotipada no sinal de voz para o banco de dados SAVEE.

5 CONCLUSÕES

Diversos estudos estão sendo desenvolvidos no campo do reconhecimento automático de locutores e emoções por meio dos sinais de voz. Várias áreas tem se beneficiado desses estudos, por exemplo: robótica assistiva, reconhecimento biométrico, automação de sistemas de atendimento e processamento de linguagem natural. Na ultima década técnicas de *Machine Learning* estão sendo largamente empregadas nessas tarefas destacando-se redes neurais MLP, redes neurais convolucionais e redes neurais recorrentes. Apesar de avanços, identificar emoções a partir de sinais de áudio não é uma tarefa simples principalmente para a utilização em sistemas embarcados. Neste sentido, ferramentas capazes de unir eficiência de discriminação e baixo custo computacional são necessárias para o desenvolvimento dessa área de estudos. Este trabalho se propôs a aplicar técnicas de processamento de sinais e identificar classificadores neurais que possam ser utilizados em um módulo auditivo embarcado para um robô de assistência.

Através dos resultados apresentados verifica-se que os objetivos desse trabalho foram alcançados. Primeiramente, realizou-se a extração de características (MFCC, Delta, Delta Delta e *Pitch*) dos sinais de áudio contidos nos bancos de dados. Depois, desenvolveu-se um algorítimo computacional capaz de realizar a identificação do locutor e emoções através do sinal de voz utilizando técnicas de processamento estatístico, redes neurais artificiais MLP e rede neurais artificiais profundas comparando os resultados obtidos. Para essas tarefa foram utilizados bancos de dados de áudio disponíveis publicamente, como por exemplo, SAVEE e Emo-DB para identificação de emoções e Voxceleb1 para a identificação do locutor.

Quanto aos resultados alcançados na tarefa de identificação do locutor, verificou-se uma média geométrica das eficiências máxima de 98% para o sistema com classificador MLP proposto quando foi utilizado análise de componentes independentes e 5 classes. Já para o classificador neural recorrente observou-se uma eficiência máxima de 92%. Comparando-se com outros trabalhos afins, observa-se que os resultados alcançados são relevantes, tanto pela eficiência alcançada quanto pelas técnicas utilizadas para o alcance desses valores. Neste sentido, vale destacar a utilização de análise de componentes independentes como pré processamento para o vetor de parâmetros apresentado ao classificador MLP e redes neurais recorrentes para a identificação de padrões nos vetores de características extraídos dos sinais de voz.

Para a tarefa de identificação da emoção verificou-se uma eficiência de 93% para o sistema com classificador MLP quando foram utilizados análise de componentes independentes e banco de dados Emo-DB. Para a base de dados SAVEE a eficiência máxima foi

de 91%. O uso de ICA também diminuiu a influência da composição dos bancos de dados. Observou-se que quando foi utilizado somente o classificador MLP houve uma diferença acentuada na entre os bancos de dados (77% e 87%). Por outro lado, quando foi utilizado o pré processamento essa diferença foi minimizada (91% e 93%). Já para o classificador recorrente observou-se uma eficiência máxima de 85%. Em outra análise, a utilização do *Pitch* de voz como atributo do vetor de características apresentado ao classificador neural aumentou as eficiências médias e máximas para em até 6 pontos percentuais.

Em ambas as tarefas observou-se que o uso de técnicas como PCA e ICA, para diminuição da dependência estatística e dimensionalidade do vetor de entrada dos classificadores MLFN obteve êxito no aumento de eficiência do classificador neural com diminuição do vetor de característica. Neste sentido, destaca-se que o tempo de processamento em ambos os casos, com e sem processamento, foram muito próximos. No entanto, o tempo de execução mostrou-se um parâmetro pouco robusto, pois depende do hardware e do software utilizado. Os resultados para as redes neurais recorrentes devem ser melhor explorados já que a eficiência do classificador foi menor e o tempo de processamento maior que o classificador MLP.

Em trabalhos futuros pretende-se utilizar as ferramentas desenvolvidas em um sistema capaz de ser utilizado em um hardware para robótica assistiva, por isso, outros testes devem ser efetuados para a implementação de um módulo embarcado. Propõe-se, também, em trabalhos futuros a utilização da linguagem de programação interpretada *Python*. Métricas de desempenho de sistemas de classificação não são, por si só, suficientes para a tomada de decisão sobre este tipo de aplicação, sendo necessário, por exemplo, medida da quantidade de operações em ponto flutuante e simplificação de funções.

REFERÊNCIAS

ABDEL-HAMID, O. et al. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, IEEE, v. 22, n. 10, p. 1533–1545, 2014. Citado 2 vezes nas páginas 4 e 27.

AOUANI, H.; AYED, Y. B. Emotion recognition in speech using mfcc with svm, dsvm and auto-encoder. In: IEEE. *4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. [S.l.], 2018. p. 1–5. Citado 2 vezes nas páginas 1 e 27.

BADSHAH, A. M. et al. Speech emotion recognition from spectrograms with deep convolutional neural network. In: IEEE. *international conference on platform technology and service (PlatCon)*. [S.l.], 2017. p. 1–5. Citado na página 5.

BONILLA, D. A.; NEDJAH, N.; MOURELLE, L. de M. Reconhecimento automático de fala em português usando redes neurais artificiais profundas. In: Bastos Filho, C. J. A.; POZO, A. R.; LOPES, H. S. (Ed.). *Anais do 12 Congresso Brasileiro de Inteligência Computacional*. Curitiba, PR: ABRICOM, 2015. p. 1–6. Citado na página 4.

BURKHARDT, F. et al. A database of german emotional speech. In: *Ninth European Conference on Speech Communication and Technology*. [S.l.: s.n.], 2005. Citado na página 29.

BURKHARDT, F. et al. A database of german emotional speech. In: *Ninth European Conference on Speech Communication and Technology*. [S.l.: s.n.], 2005. Citado na página 36.

Cai, D.; Cai, Z.; Li, M. Deep speaker embeddings with convolutional neural network on supervector for text-independent speaker recognition. In: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. [S.l.: s.n.], 2018. p. 1478–1482. Citado 3 vezes nas páginas 6, 8 e 23.

CAMPOS, V. d. A. *Arcabouço para reconhecimento de locutor baseado em aprendizado não supervisionado*. 85 p. Dissertação (Mestrado) — Universidade Estadual Paulista (UNESP), 2017. Citado 3 vezes nas páginas 4, 7 e 27.

CHAKRABORTY, K.; TALELE, A.; UPADHYA, S. Voice recognition using mfcc algorithm. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, v. 1, n. 10, p. 158–161, 2014. Citado na página 10.

COWIE, R. et al. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, Citeseer, v. 18, n. 1, p. 32–80, 2001. Citado na página 8.

DAHAKE, P. P.; SHAW, K.; MALATHI, P. Speaker dependent speech emotion recognition using mfcc and support vector machine. In: IEEE. *International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*. [S.l.], 2016. p. 1080–1084. Citado 2 vezes nas páginas 1 e 8.

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, IEEE, v. 28, n. 4, p. 357–366, 1980. Citado na página 10.

FEIL-SEIFER, D.; MATARIC, M. J. Defining socially assistive robotics. In: IEEE. *9th International Conference on Rehabilitation Robotics (ICORR)*. [S.l.], 2005. p. 465–468. Citado 2 vezes nas páginas 1 e 6.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>. Citado 4 vezes nas páginas iv, 19, 23 e 24.

HAYKIN, S. S. et al. *Neural networks and learning machines/Simon Haykin*. [S.l.]: New York: Prentice Hall, 2009. Citado 10 vezes nas páginas iv, 15, 16, 17, 18, 19, 20, 21, 22 e 31.

HE, K. et al. Deep residual learning for image recognition. In: IEEE. *IEEE conference on computer vision and pattern recognition*. [S.l.], 2016. p. 770–778. Citado na página 34.

HERSHEY, S. et al. Cnn architectures for large-scale audio classification. In: IEEE. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2017. p. 131–135. Citado na página 33.

HUANG, Z. et al. Speaker adaptation of rnn-blstm for speech recognition based on speaker code. In: IEEE. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2016. p. 5305–5309. Citado na página 25.

HYVARINEN, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, IEEE, v. 10, n. 3, p. 626–634, 1999. Citado 2 vezes nas páginas 13 e 14.

HYVÄRINEN, A.; OJA, E. Independent component analysis: algorithms and applications. *Neural networks*, Elsevier, v. 13, n. 4-5, p. 411–430, 2000. Citado na página 13.

IRIYA, R. *Análise de sinais de voz para reconhecimento de emoções*. Tese (Doutorado) — Universidade de São Paulo, 2014. Citado 2 vezes nas páginas 7 e 8.

JACKSON, P.; HAQ, S. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014. Citado 2 vezes nas páginas 29 e 36.

JAGIASI, R. et al. Cnn based speaker recognition in language and text-independent small scale system. In: IEEE. *Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. [S.l.], 2019. p. 176–179. Citado 3 vezes nas páginas 4, 27 e 37.

JOLLIFE, I. T. *Principal Component Analysis*. [S.l.]: Springer Science Business Media, LLC, 2002. Citado na página 12.

KAMATH, S.; RAVINDRAN, S.; ANDERSON, D. V. Independent component analysis for audio classification. In: IEEE. *3rd IEEE Signal Processing Education Workshop. 2004 IEEE 11th Digital Signal Processing Workshop, 2004.* [S.l.], 2004. p. 352–355. Citado na página 13.

KIM, H.-G.; MOREAU, N.; SIKORA, T. *MPEG-7 audio and beyond: Audio content indexing and retrieval*. [S.l.]: John Wiley Sons, 2006. Citado na página 11.

KWON, S. et al. A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 20, n. 1, p. 183, 2020. Citado 2 vezes nas páginas 5 e 27.

LATHI, B. P.; GREEN, R. A. *Linear systems and signals*. [S.l.]: Oxford University Press New York, 2005. Citado na página 10.

LERCH, A. *An introduction to audio content analysis: Applications in signal processing and music informatics*. [S.l.]: Wiley-IEEE Press, 2012. Citado 2 vezes nas páginas 9 e 10.

LIKITHA, M. et al. Speech based human emotion recognition using mfcc. In: IEEE. *international conference on wireless communications, signal processing and networking (WiSPNET)*. [S.l.], 2017. p. 2257–2260. Citado 3 vezes nas páginas 5, 8 e 27.

LIU, Q. et al. Research on different feature parameters in speaker recognition. *Journal of Signal and Information Processing*, Scientific Research Publishing, v. 4, n. 2, p. 106–110, 2013. Citado na página 10.

MA, J. et al. Emotion recognition using multimodal residual lstm network. In: IEEE. *27th ACM International Conference on Multimedia*. [S.l.], 2019. p. 176–183. Citado 2 vezes nas páginas 5 e 27.

MAFRA, A. T. *Reconhecimento automático de locutor em modo independente de texto por Self-Organizing Maps*. 80 p. Dissertação (Mestrado) — Universidade de São Paulo, 2002. Citado 3 vezes nas páginas 4, 10 e 27.

MARTINEZ, J. et al. Speaker recognition using mel frequency cepstral coefficients (mfcc) and vector quantization (vq) techniques. In: IEEE. *22nd International Conference on Electrical Communications and Computers (CONIELECOMP)*. [S.l.], 2012. p. 248–251. Citado na página 10.

MATHWORKS. *Train Deep Learning Network to Classify New Images*. 2020. Disponível em: <https://www.mathworks.com/help/deeplearning/ug/train-deep-learning-network-to-classify-new-images.html>. Acesso em: 12 nov. 2020. Citado na página 34.

MITCHELL, T. M. Does machine learning really work? *AI Magazine*, Public Knowledge Project, v. 18, n. 3, p. 11–20, 1997. Citado na página 15.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, 1997. Citado na página 15.

MOURA, N. N. de et al. Independent component analysis for passive sonar signal processing. In: *Advances in Sonar Technology*. [S.l.]: InTech, 2009. p. 91–108. Citado na página 12.

NAGRANI, A.; CHUNG, J. S.; ZISSERMAN, A. Voxceleb: a large-scale speaker identification dataset. In: IEEE. *INTERSPEECH*. [S.l.], 2017. p. 2616–2620. Citado 3 vezes nas páginas iv, 29 e 32.

NERI, L. V. *Extração de características para segmentação de locutores*. 114 p. Tese (Doutorado) — Universidade Federal de Pernambuco, 2019. Citado 3 vezes nas páginas 4, 33 e 37.

OLIVEIRA, M. A. et al. Ultrasound-based identification of damage in wind turbine blades using novelty detection. *Ultrasonics*, Elsevier, p. 106–166, 2020. Citado na página 12.

OLIVEIRA, M. L. L.; CERQUEIRA, J. J. F.; FILHO, E. F. S. Simulation of an artificial hearing module for an assistive robot. In: SPRINGER. *Proceedings of SAI Intelligent Systems Conference*. [S.l.], 2018b. p. 852–865. Citado 5 vezes nas páginas 2, 27, 33, 46 e 54.

OLIVEIRA, M. L. L. d. *Proposta de um módulo auditivo artificial para um robô assistivo*. 127 p. Dissertação (Mestrado) — Universidade Federal da Bahia, 2018a. Citado 25 vezes nas páginas iv, v, vi, 1, 4, 5, 7, 8, 9, 10, 11, 27, 28, 29, 32, 33, 37, 38, 40, 41, 42, 44, 45, 46 e 47.

PAN, Y.; SHEN, P.; SHEN, L. Speech emotion recognition using support vector machine. *International Journal of Smart Home*, Citeseer, v. 6, n. 2, p. 101–108, 2012. Citado na página 11.

PENHA, D. d. P. *Rede neural convolucional aplicada à identificação de equipamentos residenciais para sistemas de monitoramento não-intrusivo de carga*. 71 p. Dissertação (Mestrado) — Universidade Federal do Pará, 2018. Citado na página 33.

RIKEN. *New robot to reduce burden on care facilities*. 2020. Disponível em: <https://www.riken.jp/medialibrary/riken/import/jp/info/release/press/2009/090827/image/01.jpg>. Acesso em: 30 nov. 2020. Citado 2 vezes nas páginas iv e 6.

ROBOTLAB. *NAO power V6 educator pack*. 2020. Disponível em: <https://www.robotlab.com/hubfs/Nao%20Power%20V6-1.png>. Acesso em: 30 nov. 2020. Citado 2 vezes nas páginas iv e 6.

SANTANA, L. M. Q. d. *Aplicação de redes neurais recorrentes no reconhecimento automático da fala em ambientes com ruídos*. 69 p. Dissertação (Mestrado) — Universidade Federal de Sergipe, 2017. Citado na página 24.

SCHUELER, C. F.; SILVEIRA, F.; CATALDO, E. Verificação de locutor utilizando parâmetros extraídos do sinal glotal em conjunto com a técnica mfcc. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, v. 6, n. 1, 2018. Citado 2 vezes nas páginas 1 e 7.

SILVA, M. F. d. et al. *Aplicação do método de fusão para verificação de locutor independente de texto*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2015. Citado 3 vezes nas páginas 1, 2 e 27.

SVOZIL, D.; KVASNICKA, V.; POSPICHAL, J. Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, Elsevier, v. 39, n. 1, p. 43–62, 1997. Citado na página 15.

THEGUARDIAN. *Robot looks after residents at Italian care home*. 2020. Disponível em: <https://www.theguardian.com/technology/gallery/2015/dec/21/robot-looks-after-residents-at-italian-care-home-in-pictures#img-4>. Acesso em: 30 nov. 2020. Citado 2 vezes nas páginas iv e 6.

TOGNERI, R.; PULLELLA, D. An overview of speaker identification: Accuracy and robustness issues. *IEEE circuits and systems magazine*, IEEE, v. 11, n. 2, p. 23–61, 2011. Citado 6 vezes nas páginas iv, 7, 8, 9, 10 e 11.

TORUK, M. M.; GOKAY, R. Short utterance speaker recognition using time-delay neural network. In: IEEE. *16th International Multi-Conference on Systems, Signals & Devices (SSD)*. [S.l.], 2019. p. 383–386. Citado 3 vezes nas páginas 4, 7 e 27.

VERVERIDIS, D.; KOTROPOULOS, C.; PITAS, I. Automatic emotional speech classification. In: IEEE. *IEEE international conference on acoustics, speech, and signal processing*. [S.l.], 2004. p. 593–596. Citado na página 11.

VOGL, T. P. et al. Accelerating the convergence of the back-propagation method. *Biological cybernetics*, Springer, v. 59, n. 4-5, p. 257–263, 1988. Citado na página 33.

WÖLLMER, M. et al. Feature enhancement by bidirectional lstm networks for conversational speech recognition in highly non-stationary noise. In: IEEE. *IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.], 2013. p. 6822–6826. Citado na página 24.

Anexos

ANEXO A – ARTIGO DO AUTOR

Borges J., E.; Cerqueira, J.; Simas F., E.; Fernandes J., A.; Oliveira, A. Reconhecimento de Emoções em Sinais de Fala utilizando Redes Neurais e Análise de Componentes Independentes. Anais do 14 Congresso Brasileiro de Inteligência Computacional, p. 1, 2019.

RESUMO

Com o uso cada vez maior de máquinas é necessário o desenvolvimento e aprimoramento de técnicas para reconhecimento de padrões de características físicas, objetivando uma maior interação entre usuários finais e as máquinas requerendo, dessas ultimas, a percepção das ações e reações humanas e respondendo a elas de maneira apropriada. O objetivo desse trabalho foi desenvolver um sistema de classificação de emoções a partir de sinais de voz, comparar diversas combinações de descritores de sinais de música e voz, realizar processamento estatístico por ICA dos vetores de parâmetros extraídos dos sinais de voz e classificação por meio de um classificador neural supervisionado com arquitetura MLP (Multilayer Perceptron).

Uma contribuição importante do trabalho foi a utilização de estatística de alta ordem (ICA) em uma etapa de processamento estatístico com objetivo de redução da redundância entre os parâmetros utilizados. A técnica de Análise de Componentes Independentes foi aplicada ao vetor de característica antes da etapa de classificação, contribuindo para a redução da dependência estatística e a compactação dos dados, já que, o algoritmo da ICA calcula a análise de componentes principais (PCA) em uma etapa preliminar.

Os resultados encontrados indicam que os coeficientes MFCC “estáticos” e “dinâmicos”, são características que possuem maior capacidade de discriminação das emoções contidas no sinal de fala do que quando combinada às características propostas. Contudo, observa-se que alguns componentes, dentre os 15 extraídos para cada descritor, totalizando 45 coeficientes, possuem maior capacidade de diferenciação entre as emoções do que outros. O algoritmo da ICA utilizado reduziu a dimensionalidade do vetor de características, pois verificou-se que 99,9% da energia (variância) estava armazenada nos 20 primeiros componentes da PCA, obtendo-se um aumento de eficiência do classificador neural com diminuição do vetor de característica. Houve aumento da eficiência (acurácia) do classificador após aplicação da ICA em 2,8 pontos percentuais. Para 50% das emoções houve aumento da eficiência e para 33,3% das emoções o resultado permaneceu constante. Em uma comparação com outros trabalhos foi possível observar que, em alguns casos, obtiveram maior acerto do sistema de classificação, mas reduzindo-se o número de emoções,

o que significa um problema de separação entre classes muito mais simples, ou o número de parâmetros que compõem o vetor apresentado a entrada do classificador. Em futuros trabalhos pretende-se abordar diferentes tipos e arquiteturas de classificadores neurais, como redes profundas, testar novos descritores, coeficientes MFCC, MFCC Delta e Delta Delta específicos e utilizar outros bancos de dados de falas.