CENTRO UNIVERSITÁRIO FEI

THIAGO SPILBORGHS BUENO MEYER

# RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA FALA APLICADO A ROBÔS DE ASSISTÊNCIA DOMÉSTICA

São Bernardo do Campo

2022

THIAGO SPILBORGHS BUENO MEYER

# RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA FALA APLICADO A ROBÔS DE ASSISTÊNCIA DOMÉSTICA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação do Centro Universitário da FEI como requisito à obtenção do título de Mestre em Engenharia Elétrica. Orientado pelo Professor Doutor Plinio Thomaz Aquino Junior.

São Bernardo do Campo

2022

Spilborghs Bueno Meyer, Thiago.
Reconhecimento de emoções através da fala aplicado a robôs de assistência doméstica / Thiago Spilborghs Bueno Meyer. São Bernardo do Campo, 2022.
75 f. : il.

Dissertação - Centro Universitário FEI.
Orientador: Prof. Dr. Plinio Thomaz Aquino Junior.

1. Reconhecimento de Emoções. 2. Fala. 3. Aprendizado Profundo. 4. Interação Humano-Robô. 5. Redes Neurais. I. Thomaz Aquino Junior, Plinio, orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

centro universitário FEI

# APRESENTAÇÃO DE DISSERTAÇÃO
# ATA DA BANCA EXAMINADORA

**Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica**

**Mestrado**

**PGE-10**

**Aluno:** Thiago Spilborghs Bueno Meyer **Matrícula:** 119114-7

**Título do Trabalho:** Reconhecimento de emoções através da fala aplicado a robôs de assistência doméstica.

**Área de Concentração:** Inteligência Artificial Aplicada à Automação e Robótica

**Orientador:** Prof. Dr. Plínio Thomaz Aquino Junior

**Data da realização da defesa:** 23/02/2022

**ORIGINAL ASSINADA**

**Avaliação da Banca Examinadora:**

A banca foi realizada no dia 23 de fevereiro de 2022 com início às 14h00. O aluno realizou a apresentação com sucesso, seguida pela arguição pelos membros da banca. O aluno respondeu as questões de forma adequada, demonstrando conhecimento do tema e ajudando a entender os pontos de atenção do trabalho. Foram sugeridas melhorias em relação ao texto para a versão final, documentada nas versões do texto dos membros da banca e nas atas individuais. O aluno foi aprovado.

São Bernardo do Campo, / / .

| **MEMBROS DA BANCA EXAMINADORA** | |
|---|---|
| Prof. Dr. Plínio Thomaz Aquino Junior | Ass.: _____ |
| Prof. Dr. Guilherme Alberto Wachs Lopes | Ass.: _____ |
| Prof. Dr. Mario Minami | Ass.: _____ |

A Banca Julgadora acima-assinada atribuiu ao aluno o seguinte resultado:

APROVADO ☒ REPROVADO ☐

**VERSÃO FINAL DA DISSERTAÇÃO**

**APROVO A VERSÃO FINAL DA DISSERTAÇÃO EM QUE FORAM INCLUÍDAS AS RECOMENDAÇÕES DA BANCA EXAMINADORA**

_____

Aprovação do Coordenador do Programa de Pós-graduação

_____

Prof. Dr. Carlos Eduardo Thomaz

A meu avô, Pedro, que tanto me incentivou.

## AGRADECIMENTOS

Agradeço primeiramente à minha família pelo incansável apoio, aos professores Ivandro e Carlos que muitas vezes me elucidaram em diversas questões, à equipe RoboFEI@Home pelas pontuais informações e ao meu orientador, o professor Plinio Thomaz Aquino Junior pelas informações de grande relevância ao desenvolvimento de meu trabalho e, finalmente, a todos que mesmo indiretamente acrescentaram algo aos meus estudos.

## RESUMO

Por meio da fala, que privilegia a natureza funcional e interativa do texto, é possível averiguar as circunstâncias espaço-temporais, as condições de produção e recepção do discurso, os propósitos explícitos como informar, explicar, convencer etc. Condições essas que permitem aproximar a interação entre humanos à interação entre humanos e robôs tornando-a natural e sensível às informações. No entanto, não basta compreender o que é falado, faz-se necessário o reconhecimento de emoções para a interação desejada. Verificou-se a validez do uso de redes neurais para seleção de características e para o reconhecimento de emoções. Para isso propõe-se o uso de Redes Neurais e comparação de modelos, como redes neurais recorrentes e redes neurais profundas, com intuito de realizar a classificação das emoções através dos sinais de fala para verificar a qualidade do reconhecimento. Espera-se possibilitar a implementação em robôs de um ambiente doméstico, como o robô HERA da equipe RoboFEI@Home, que tem como foco robôs de serviço autônomos para o ambiente doméstico. Foram realizados testes utilizando-se apenas os Coeficientes Cepstrais da Frequência-Mel, bem como testes com diversas características do Delta-MFCC, contraste espectral e o espectrograma-Mel. Para realizar o treinamento, validação e testes das redes neurais, usufruiu-se a base de dados eNTERFACE'05, que possui 42 locutores de 14 nacionalidades diferentes falando o idioma inglês. Os dados da base escolhida são vídeos que, para o uso nas redes neurais, foram convertidos em áudios. Constatou-se como resultado uma classificação de 52% de acertos quando empregada a rede neural profunda, quando verificado o uso da rede neural recorrente, sendo a classificação com acurácia igual 44%. Os resultados apresentam maior acurácia quando apenas os Coeficientes Cepstrais da Frequência-Mel são usados para a classificação, utilizando o classificador com a Rede Neural Profunda e em apenas um caso é possível observar um maior acerto por parte da Rede Neural Recorrente, que se dá no uso de diversas características e na configuração de 73 para o tamanho do *Batch* e 100 épocas de treinamento.


Palavras-chave: Reconhecimento de Emoções. Fala. Aprendizado Profundo. Interação Humano-Robô.Redes Neurais.

## ABSTRACT

Through speech, which privileges the functional and interactive nature of the text, it is possible to ascertain the spatio-temporal circumstances, the conditions of production and reception of the discourse, the explicit purposes such as informing, explaining, convincing etc. These conditions allow bringing the interaction between humans closer to the Human-Robot interaction, making it natural and sensitive to information. However, it is not enough to understand what is said, it is necessary to recognize emotions for the desired interaction. The validity of the use of neural networks for feature selection and emotion recognition was verified. For this purpose, it is proposed the use of Neural Networks and comparison of models, such as recurrent neural networks and deep neural networks, in order to carry out the classification of emotions through speech signals to verify the quality of recognition. It is expected to enable the implementation in robots in a domestic environment, such as the HERA robot from the RoboFEI@Home team, which focuses on autonomous service robots for the domestic environment. Tests were performed using only the Mel-Frequency Cepstral Coefficients, as well as tests with several characteristics of Delta-MFCC, spectral contrast and the melspectrogram. To carry out the training, validation and testing of the neural networks, the eNTERFACE'05 database was used, which has 42 speakers from 14 different nationalities speaking the English language. The data from the chosen database are videos that, for use in neural networks, were converted into audios. It was found as a result a classification of 52% of correct answers when using the deep neural network, when the use of the recurrent neural network was verified, with the classification with accuracy equal to 44%. The results are more accurate when only the Mel-Frequency Cepstral Coefficients are used for the classification, using the classifier with the Deep Neural Network and in only one case it is possible to observe a greater accuracy by the Recurrent Neural Network, which occurs in the use of various features and setting 73 for *Batch* size and 100 training epochs.


Keywords: Emotion Recognition. Speech. Deep Learning. Human-Robot Interaction. Neural Networks

## LISTA DE ILUSTRAÇÕES

Figura 1 – Robô de Assistência Doméstica HERA . . . . . 15
Figura 2 – Diagrama do Funcionamento do Sistema Proposto . . . . . 45
Figura 3 – Representação da Rede Neural Recorrente utilizando apenas MFCCs . . . . 46
Figura 4 – Representação da Rede Neural Recorrente . . . . . 47
Figura 5 – Representação da Rede Neural Profunda utilizando apenas MFCCs . . . . . 47
Figura 6 – Representação da Rede Neural Profunda . . . . . 48
Figura 7 – Gráfico de Comparação da Acurácia da DNN e RNN utilizando apenas MFCCs 53

## LISTA DE TABELAS

Tabela 1 – Tabela de Comparação das Bases de Dados . . . . . 38
Tabela 3 – Dados de reposta da Rede Neural Profunda utilizando apenas MFCCs . . . . 51
Tabela 4 – Dados de reposta da Rede Neural Recorrente utilizando apenas MFCCs . . . 52
Tabela 5 – Matriz de Confusão Para Rede Neural Profunda utilizando apenas MFCCs . 53
Tabela 6 – Matriz de Confusão Para Rede Neural Recorrente utilizando apenas MFCCs 54
Tabela 7 – Matriz de Confusão Para Rede Neural Profunda . . . . . 54
Tabela 8 – Matriz de Confusão Para Rede Neural Recorrente . . . . . 55
Tabela 9 – Tabela de Comparação entre resultados positivos das configurações e as emoções . . . . . 55

## LISTA DE ABREVIATURAS

| | |
|---|---|
| ANN | Redes Neurais Artificiais (*Artificial Neural Networks*) |
| BayesNet | Rede de Bayes (*Bayes Network*) |
| BSF | Características Bi-Espectrais (*Bi-Spectral Features*) |
| CNN | Rede Neural de Convolução (*Convolutional Neural Networks*) |
| DCT | Transformada Discreta de Cosseno (*Discrete Cosine Transform*) |
| DFT | Transformada Discreta de Fourier (*Discrete Fourier Transform*) |
| DNN | Rede Neural Profunda (*Deep Neural Network*) |
| DNN-HMM | Modelos escondidos de Markov em conjunto com Redes Neurais profundas (*Deep Neural Networks - Hidden Markov Model*) |
| EKIsomap | Mapeamento Isométrico com Núcleo Aprimorado (*Enhanced Kernel Isometric Mapping*) |
| FFT | Transformada Rápida de Fourier (*Fast Fourier Transform*) |
| GMM | Modelo de Mistura Gaussiana (*Gaussian Mixture Model*) |
| HERA | Robô Assistente de Ambiente Doméstico (*Home Environment Robot Assistant*) |
| HMM | Modelo Escondido de Markov (*Hidden Markov Model*) |
| HOS | Espectro de Ordem Maior (*Higher Order Spectrum*) |
| K-NN | k Vizinhos Próximos (*k-Nearest Neighbours*) |
| LFLB | Bloco de Aprendizado de Características Locais (*Local Feature Learning Block*) |
| LFPC | Coeficiente de Log de Frequência de Potência (*Log-Frequency Power Coefficient*) |
| LogMMSE | Amplitude Log-Espectral de Erro Quadrático Médio Mínimo (*Log-Spectrum Minimal Medium Square Error*) |
| LPCC | Coeficientes Preditivos de percepção linear (*Linear Prediction Cepstrum Coefficient*) |
| LSTM | Memória de Curto-Longo Prazo (*Long-Short Term Memory*) |
| MFCCs | Coeficientes Cepstrais da Frequência Mel (*Mel Frequency Cepstral Coefficients*) |
| MLP | Perceptron de Multicamadas (*MultiLayered Perceptron*) |
| MMSE | Erro Quadrático Médio Mínimo (*Minimal Medium Square Error*) |

| | |
|---|---|
| MSF | Características de Modulação Espectral (*Modulation Spectral Features*) |
| PCA | Análise de Componente Principal (*Principal Component analysis*) |
| RBM | máquinas restritas de Boltzmann (*Restrict Boltzman Machines*) |
| RF | Florestas Aleatórias (*Random Forests*) |
| RMS | Raiz Quadrada Média (*Root mean Square*) |
| RNN | Rede Neural Recorrente (*Recurrent Neural Networks*) |
| SER | Reconhecimento de emoções através da fala (*Speech Emotion Recognition*) |
| SERS | Sistema Reconhecimento de emoções através da fala (*Speech Emotion Recognition System*) |
| SOFMNN | Rede Neural de Mapeamento de Características Auto-Organizadora (*Self Organizing Feature Mapping Neural Network*) |
| ST | Espectro Temporais (*Spectrum Temporal*) |
| SVM | Máquinas de Vetor de Suporte (*Suport Vector Machine*) |
| TEO | Operador de Energia Teager (*Teager Energy Operator*) |
| ZCR | Taxa de Cruzamento em Zero (*Zero Crossing Rate*) |

## SUMÁRIO


**1 Introdução** . . . . . 13

**2 Revisão Bibliográfica** . . . . . 19

2.1 Sistema de Reconhecimento de Emoções Através da Fala . . . . . 19

2.2 Seleção e Extração de Características . . . . . 20

2.3 Trabalhos Relacionados . . . . . 30

**3 Conceitos** . . . . . 37

3.1 Bases de Dados . . . . . 37

3.2 Classificadores . . . . . 39

3.3 Redes Neurais Profundas . . . . . 41

3.4 Características da Fala . . . . . 42

**4 Metodologia** . . . . . 44

**5 Experimentos** . . . . . 50

**6 Resultados** . . . . . 51

**7 Conclusões** . . . . . 57

**REFERÊNCIAS** . . . . . 59

**APÊNDICE A – Tabela de Trabalhos Relacionados** . . . . . 63

# 1 INTRODUÇÃO

A interação entre humanos e robôs tem se desenvolvido, cada vez mais, por meio da fala, pois ela possibilita, assim como nas interações humanas, a averiguação de circunstâncias que permitem a aproximação necessária para o reconhecimento das emoções. Esse reconhecimento possibilita a compreensão sobre o que o usuário deseja e a interpretação que permitem a emissão de uma resposta através da fala, o que torna possível a realização conjunta de tarefas entre humanos e robôs. As emoções transmitidas por meio da voz agregam ao que foi falado informações acerca do estado de espírito de um usuário. Sendo assim, para o aprimoramento da interação entre humanos e robôs utilizou-se o método chamado Reconhecimento de emoções através da fala (SER, *Speech Emotion Recognition*).

Esse método é muito empregado em aplicações nas quais a compreensão do sentimento do usuário é necessária, como em tutoriais para aplicações que desfrutam da sensação do usuário para definir passos futuros ou adaptações. Por exemplo, quando o usuário se sente inseguro, o sistema reconhece essa insegurança através da fala e, ao invés de continuar as explicações do resto do sistema, recapitula o conteúdo passado; ou seja, baseia-se na compreensão do sentimento para se adaptar ao usuário. Os sistemas que usufruem de tais dados podem ser inseridos em contextos como o computador de bordo de um carro, para que o carro inicie processos de segurança ao usuário, assim como podem ser inseridos em robôs, para que haja uma adaptação no comportamento e resposta baseados no que o usuário está sentindo durante a interação.

Como visto em El Ayadi, Kamel e Karray (2011), a tarefa de reconhecimento de emoção através da fala é muito complexa devido às diversas características que definem a emoção expressa através de uma fala. Mesmo com essa complexidade, ainda não fica claro quais aspectos têm maior influência na determinação da emoção. Outra problemática no reconhecimento de emoções através da voz é a variação acústica entre as vozes diversas de usuários, visto que há diferenças entre tonicidade de voz, cadência da fala e até a forma com que a pessoa se comunica. Tais fatores são parte das características que podem ser aplicadas para a definição de emoção (EL AYADI; KAMEL; KARRAY, 2011).

Assim como é demonstrado em Haytham M. Fayek, Lech e Cavedon (2017), utilizou-se Redes Neurais para a resolução do problema, devido às diversas características que podem ser observadas quando realiza-se o método SER. É possível empregar as Redes Neurais para tratamento de dados diversos e classificação de valores, como reconhecimento de fala, de objetos ou mesmo de pessoas. Portanto, elas podem ser empregadas para agrupar e classificar as características presentes na fala de uma pessoa. É necessário separar em classes as emoções a

serem definidas. Para isso será estabelecido um modelo de definição de emoções, assim como é feito em Lijiang Chen et al. (2012), que utilizam um conjunto das propostas de Ekman (1992) e Fox (1991). Esse conjunto se baseia na proposta de que existem emoções básicas, assim como as cores, por isso é chamada de Teoria da paleta, na qual emoções são combinações entre as emoções básicas.

A relevância nesse desenvolvimento se destaca, sobretudo, a partir da época de isolamento social de 2020, em que se pensa a respeito de formas de aplicações de automações em diversas áreas, visto que, além da discussão sobre o uso de robôs em diversos ambientes de interação, há um aprofundamento cada vez maior nesses estudos. Totens são requisitados para inspeção de saúde até robôs de serviço para o cuidado de humanos em locais onde há pessoas do grupo de risco, a fim de evitar a exposição a potenciais contaminações. Para estas máquinas, a interação com o ser humano é o núcleo do uso, no qual o ser humano deve interagir e obter uma resposta, seja ela visual, auditiva ou física e nessa interação o desejo é de que o humano se sinta confortável, consiga interagir de maneira natural e tenha respostas direcionadas para o seu sentimento ou estado emocional atual. Para que isso ocorra é necessário ter um sistema de reconhecimento de emoções.

Para este trabalho o Robô Assistente de Ambiente Doméstico (HERA, *Home Environment Robot Assistant*), desenvolvido pela equipe RoboFEI@Home para Interação Humano-Robô e realização de tarefas cooperativas em ambientes domésticos, cuja comunicação se realiza pelo reconhecimento da fala, Aquino Jr et al. (2019), foi o que alavancou a curiosidade e o estímulo para essa pesquisa, apresentando a possibilidade de uma futura alteração na interação conforme o estado emocional do usuário perante o robô.

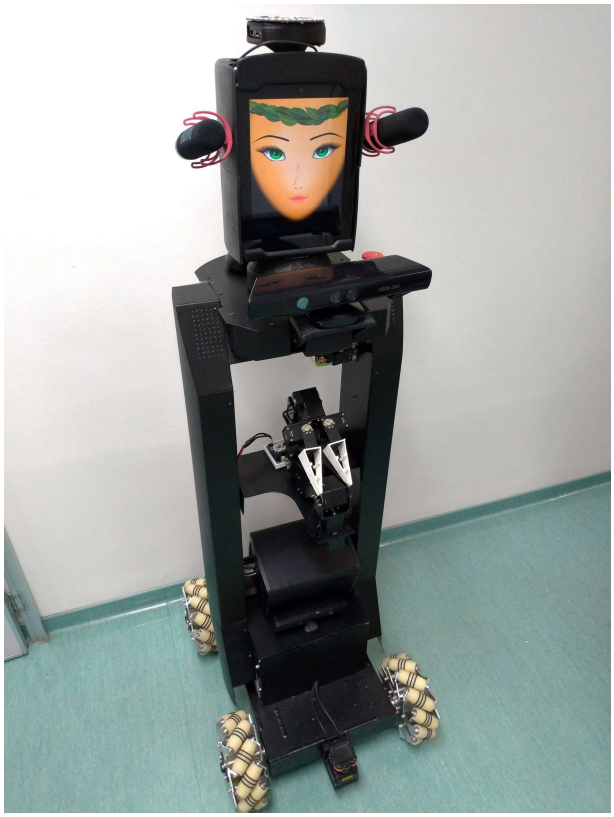
HERA possui uma base omnidirecional, que permite uma locomoção sem restrições de direção, além de diversos sensores para interação com o ambiente e com os usuários. Entre estes sensores estão lasers, para reconhecimento de profundidade e mapeamento de ambientes; câmeras, para reconhecimento de faces, objetos e pessoas; microfones, para reconhecimento de fala e localização de fontes sonoras. Além desses sensores, o robô também possui um manipulador, para realizar a interação com objetos e realizar gestos e um *tablet* no qual é representado o rosto do robô HERA para interação com usuários (AQUINO JR et al., 2019).

O robô HERA foi uma das motivações para iniciar a pesquisa de reconhecimento de emoções através da fala. HERA, ilustrado na figura 1, foi desenvolvido para o trabalho de assistência pessoais em ambientes domésticos, de maneira a auxiliar e realizar, em qualquer

tarefa, atividades de maneira autônoma. Uma vez que é projetado para a realização da Interação Humano-Robô e tarefas de cooperação (AQUINO JR et al., 2019).

A equipe RoboFEI@Home começou a projetar e trabalhar com robôs de assistência doméstica em meados de 2015 e atualmente trabalha com o robô HERA participando de competições como a RoboCup@Home da (ROBOCUP-FEDERATION, 2020). Um exemplo de pesquisa é o sistema de localização de fontes sonoras em Meyer e Junior (2019) e sua análise em Meyer e Aquino-Junior (2020), no qual o usuário realiza a interação com o robô e este identifica a direção na qual o usuário está localizado para realizar uma interação usufruindo de sua parte frontal para comunicação, apresentado seu rosto e podendo identificar o usuário através das câmeras, realizando um reconhecimento de face com o usuário ativo.

Figura 1 – Robô de Assistência Doméstica HERA



Fonte: Autor

O robô HERA possui diversos sistemas para interação com o espaço e usuários, como a visão que se utiliza de câmeras e sensores para poder realizar o reconhecimento de faces e detecção e reconhecimento de objetos; e para a interação com tais objetos reconhecidos é empregado um manipulador robótico projetado pela equipe, de maneira a controlar o objeto e encaminhá-lo para pessoas ou locais. O robô possui a base omnidirecional que permite a navegação em qualquer direção a partir do cálculo de somatória de forças exercido pelas rodas

da base e, para a identificação de locais e possíveis barreiras, é empregado um laser que faz a varredura do ambiente no qual o robô se encontra. Além disso, para a comunicação com os usuários, o robô possui um sistema de reconhecimento e síntese de fala para a interação (AQUINO JR et al., 2019).

HERA dispõe de microfones da marca Rode para capturar a voz e realizar o reconhecimento de fala, identificando quais palavras foram proferidas, afim de compreendê-las, interpretá-las e realizar a ação requisitada por esse comando de voz. Além dos microfones direcionais da Rode, HERA também possui um conjunto de microfones presentes na placa multissensorial *Matrix Creator*, que são aplicados para realizar a localização de uma fonte sonora através do áudio coletado (MEYER; AQUINO-JUNIOR, 2020).

Um robô de assistência doméstica tem como possível cenário, em diversos momentos, a interação com todos os usuários presentes naquele ambiente, assim como ocorre durante a competição de robótica de serviço focada em assistentes domésticos como a competição RoboCup-Federation (2020). É verificado também que durante a interação Humano-Robô há uma variação muito grande na complexidade e duração de uma fala, como por exemplo quando pedimos um copo com refrigerante utilizando a frase "Robô, busque um copo de Coca-Cola"ou mesmo quando pedimos algo mais complexo, como "Robô, verifique se há alguém no quarto e traga uma maçã que está na geladeira para João que está na sala". Esse cenário de interação é um dos casos apresentados durante a competição, no qual o robô deve executar as tarefas requisitadas, assim como seria dentro de uma casa onde o robô estaria situado no dia-a-dia. Nestes cenários de uso com robôs de serviço, é necessário refletir sobre o uso do SER visando a variação de usuários, portanto um sistema que não leva em consideração a possibilidade de diferentes usuários interagindo com o sistema, não cumpre com os requisitos para uso no projeto do robô utilizado como base neste trabalho.

Há os robôs de uso nos lares que tiveram grande aumento de consumo na última década, principalmente os com função de aspiradores de pó, como o iRobot (2020), e os de interação social como os robôs Sony-Corporation (2020) e o SoftBank-Robotics (2020). Para estes de interação social é comum aplicar o reconhecimento da fala para realizar a comunicação humano-robô.

Ressalta-se que robôs de serviço são estruturas mecânicas, elétricas e computacionais com sensores que realizam tarefas em ambientes diversos, interagindo com objetos, cenário e até mesmo com usuários, e no caso deste trabalho o foco dos robôs é no ambiente doméstico.

Para a interação do robô Hera com humanos, é utilizado um sistema de reconhecimento de fala, que realiza o reconhecimento do pedido e interação do usuário utilizando o idioma inglês. Em pesquisas anteriores no robô Hera, foram utilizadas redes Neurais Recorrentes, utilizando LSTM, e Redes Neurais Profundas para realizar esse reconhecimento de fala. Devido ao direcionamento da Competição na qual a equipe RoboFEI@Home participa, a RoboCup, é utilizado como idioma de interação, o inglês e por conta desse direcionamento, o domínio utilizado é baseado também nesse idioma para este trabalho.

Propõe-se aqui o uso de Redes Neurais para a classificação de emoções através da fala em uma conversa entre usuário e um sistema. Seja ele um robô físico ou virtual, como em um atendimento online, defina o sentimento da pessoa que está realizando a comunicação. Visa-se ainda contribuir com trabalhos futuros que busquem aprimorar a Interação Humano-Robô baseando-se de alterações no comportamento, baseados em fatores como a aproximação do robô ao usuário e a emoção do usuário perante o robô. Neste trabalho, portanto, busca-se, sobretudo, realizar a classificação das emoções presentes na fala, utilizando-se das características dos sinais de fala de um usuário, para que se cumpra com assertividade o objetivo de uso. Busca-se também, realizar a comparação na acurácia dos sistemas utilizando classificadores Rede Neural Profunda (DNN, *Deep Neural Network*) e Memória de Curto-Longo Prazo (LSTM, *Long-Short Term Memory*), e estudando diferentes configurações para o treinamento desses classificadores.

O uso do sistema SER apresenta desafios de classificação e seleção de características que se discutirá ao longo do capítulo 3, contudo faz-se necessário pensar nos desafios presentes ao se utilizar e aplicar este sistema a um robô de serviço móvel. Alguns dos desafios são o ambiente que é diferente dos ambientes de estúdios, com a acústica projetada, na qual algumas bases de dados são empregadas; os ruídos presentes em um ambiente doméstico que variam, tanto por influência interna, quanto por sons externos. No uso de um robô doméstico, também é previsto que as emoções podem ser alteradas devido à aproximação do robô, caso o usuário não se sinta confortável, por exemplo.

Dividiu-se o estudo em capítulos, sendo o capítulo dois uma introdução aos trabalhos já realizados na área, a fim de apresentar os que tiveram propostas relacionadas a este. O capítulo três apresenta os conceitos utilizados para essa elaboração, contendo desde problemáticas e técnicas até métodos usados ao longa da história da área e seu estado da arte. No quatro se descreve a metodologia empregada, explicando como o sistema fora projetado, qual base de dados se emprega e se apresenta a proposta do trabalho com relação aos sistemas que foram usados. No capítulo cinco é apresentado o formato e quais experimentos foram realizados. No

capítulo seis, assim como no capítulo anterior, apresenta-se os resultados dos testes. Por fim, no capítulo sete, apresenta-se as conclusões a que se chegou.

# 2 REVISÃO BIBLIOGRÁFICA

Neste capítulo, serão apresentados trabalhos revisados relacionados à proposta em desenvolvimento, a fim de destacá-los, bem como o detalhamento de Reconhecimento do Sistema SER, a importância da extração e seleção de características dentre outras pesquisas relacionadas à proposta deste.

## 2.1 SISTEMA DE RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA FALA

Em Chaudhary et al. (2015) o Sistema Reconhecimento de emoções através da fala (SERS, *Speech Emotion Recognition System*) é definido como um sistema de reconhecimento de padrões, o que significa que as passagens e partes existentes em um sistema de reconhecimento de padrões estará presente em um SERS. Definir quais serão as emoções significantes a serem classificadas é a maior preocupação quando se trabalha com SERS, visto que bases típicas de emoções possuem 300 estados emocionais, porém para evitar entrar em um problema de classificação com tantas classes é comum seguir a Teoria da Paleta, no qual as emoções são definidas por composições de emoções primárias, assim como cores. Essa teoria resume as emoções primárias como sendo: raiva, aversão, medo, alegria, tristeza e surpresa.

No trabalho de Akçay e Oğuz (2020), o SER é definido como uma coleção de metodologias que processa e classifica um sinal de uma fala, a fim de detectar emoções presentes. Um sistema SER precisa de um classificador e um sistema de aprendizado para sua aquisição de reconhecimento das emoções humanas através da voz. Deve-se definir quais emoções serão classificadas e as modelar à classificação, o que pode ser um problema, já que é um dos desafios que se encontra em aberto até na área da psicologia. A partir das definições realizadas de bases e emoções que serão discretizadas e classificadas pelo sistema de SER é necessário ainda realizar um pré-processamento.

Para El Ayadi, Kamel e Karray (2011), o SERS é muito útil para aplicações na qual a Interação Humano-Máquina deve ser natural e nas aplicações nas quais é desejada a adaptação da resposta do sistema baseado na frustração ou irritação na voz do usuário falante. Com as tarefas possíveis definidas, é necessário reconhecer então que o SER é muito desafiador de se realizar, devido a razões como: quais características devem ser utilizadas para classificar as emoções, variações acústicas presentes nas frases e estilo de fala, pois podem atrapalhar na classificação; bem como a maneira com a qual uma emoção é expressa que pode variar dependendo da cultura e ambiente do usuário.

Em Bhavan et al. (2019), é visto que a tarefa de reconhecimento de emoções na fala é um assunto de grande interesse há tempos e esse desafio se mostrou muito complexo e multidimensional, devido aos problemas que foram observados em outros trabalhos já revisados, incluindo o problema da gama de emoções transmitidas para a forma verbal que podem ter alterações e ainda quais características devem ser empregadas para a análise das emoções presentes na fala.

Para Haytham M. Fayek, Lech e Cavedon (2017), assim como para outros autores, o SER apresenta um problema de classificação. Contudo, definiu-se que este problema pode ser de natureza estática ou dinâmica, o que motivou duas formulações populares, a da formulação de processamento baseada em turnos e a da formulação de processamento baseada em quadros.

O SER pode ser empregado como uma aplicação à parte ou como parte de um sistema integrado. O processamento com base em quadro é mais robusto, visto que não depende da segmentação da fala de entrada em enunciados e pode modelar emoções dinamicamente. Independente da base do processamento, o maior esforço atrelado às pesquisas recentes é voltado à seleção de características ótimas (FAYEK, Haytham M.; LECH; CAVEDON, 2017)).

Quando aplicado à Interação Humano-Robô, o SER possui o foco de tornar esta interação mais natural, pois possui uma entrada a mais que a tradicional, que é o conteúdo verbal, que realiza alterações sutis no contexto da fala que afetam a reação de qualquer humano nesta interação. A área de interações orientadas por emoções possui algumas vertentes como as emoções faciais como também as emoções da fala (RAMAKRISHNAN; EL EMARY, 2013).

Segundo Ramakrishnan e El Emary (2013), o SER deixa de ser um tópico secundário, na última década, e torna-se um tópico muito importante na área de Interação Humano-Computador e na área de processamento da fala.

## 2.2 SELEÇÃO E EXTRAÇÃO DE CARACTERÍSTICAS

Em Akçay e Oğuz (2020) é dito que as características podem ser separadas em quatro categorias distintas, sendo elas as Características prosódicas, as Características Espectrais, as Características de qualidade de voz e as Características baseadas em Operador de Energia Teager (TEO, *Teager Energy Operator*).

As características prosódicas são aquelas percebidas pelos humanos, como a entonação e o ritmo. As mais utilizadas são baseadas na frequência fundamental, energia e duração. Verifica-se que existe uma correlação entre as características prosódicas e os estados emocio-

nais, visto que elas apresentam as mudanças ocorridas durante toda a fala (AKÇAY; OĞUZ, 2020).

Características espectrais são as obtidas através da transformação do sinal do domínio do tempo em um sinal do domínio da frequência fazendo uso da transformada de Fourier. Essas características são extraídas de partes da fala que variam de 20 a 30 milissegundos (AKÇAY; OĞUZ, 2020).

Qualidade de voz apresenta características que são determinadas pelas propriedades físicas do trato vocal. Mudanças involuntárias produzem sinais de fala que podem ser diferenciados através do uso de distintas propriedades com *jitter*, brilho e a proporção de harmônicos para ruído (AKÇAY; OĞUZ, 2020).

Akçay e Oğuz (2020) também apresentam outras medidas utilizadas para verificar as características de qualidade de voz como o Quociente de Amplitude Normalizado, o Quociente *Quasi Open*, a características da diferença entre os dois primeiros harmônicos do diferencial do espectro de fonte glotal, o Quociente de dispersão máxima, a inclinação espectral, o parâmetro parabólico espectral e o parâmetro da forma do modelo Liljencrants-Fant da dinâmica de pulso glotal.

As características que dependem do TEO são empregadas para detectar o estresse em uma fala, uma vez que uma situação estressante afeta a tensão muscular do locutor, o que resulta em uma alternação na dispersão de ar durante a produção de sons (AKÇAY; OĞUZ, 2020).

Em Wu, Falk e Chan (2011) se propõe extrair características modulares espectrais para a realização do SER, no qual dois tipos de Características de Modulação Espectral (MSF, *Modulation Spectral Features*) são calculadas para um método de curto prazo, fazendo uso da média das medidas espectrais e parâmetros de predição linear. Para cada quadro, é empregada uma escala de unidade de energia e seis medidas são utilizadas para cada base de quadro.

A primeira medida define a distribuição de energia da fala conforme a frequência modular. A segunda medida espectral é o achatamento espectral definido pela razão da média geométrica da energia espectral com a média aritmética, no qual um valor próximo a 1 (um) representa um espectro achatado e um valor próximo a 0 (zero) sugere um espectro com uma grande variabilidade nas amplitudes observadas. A terceira medida é o centroide espectral que reflete o centro de massa do espectro para cada canal modular. Os canais modulares são definidos pelas frequências presentes na fala. Além das seis medidas, também se emprega a análise de predição linear aplicada na seleção de canais modulares para a extração do segundo conjunto

de características de modulação espectral. Após a utilização de ambos os métodos é possível obter quarenta e uma características calculadas para cada quadro (WU; FALK; CHAN, 2011).

Somadas ao conjunto de MSF, também são consideradas as características de curto prazo no qual temos os Coeficientes Cepstrais da Frequência Mel (MFCCs, *Mel Frequency Cepstral Coefficients*), que foi aplicado com sucesso na área de reconhecimento de fala automático e popular no uso para reconhecimento de emoções. As características mais comuns a serem aplicadas quando se trata do uso de MFCCs são a média e desvio padrão das primeiras 13 (treze) MFCCs e seus deltas. Como adição aos MFCCs, Coeficientes Preditivos de percepção linear (LPCC, *Linear Prediction Cepstrum Coefficient*) são extraídos da fala, tornando-se uma alternativa para a comparação de características de curto prazo. Além dessas características, podem ser extraídas também as características prosódicas, tais como: frequência fundamental, intensidade e ritmo de fala (WU; FALK; CHAN, 2011).

Para El Ayadi, Kamel e Karray (2011) há quatro questões principais que devem ser consideradas na extração de características: a região de análise, melhores tipos de características, efeito do pré-processamento do áudio e a suficiência das características acústicas com relação à linguística, informação do discurso e características faciais.

Como os sinais das falas não são estacionários, é normal aplicar a segmentação de quadros para estes sinais e para esses quadros o sinal é considerado praticamente estacionário. Algumas características como tom e energia são extraídas de cada quadro e são chamadas de características locais. Em contrapartida a essas características extraídas de cada quadro estão as características globais, que são extraídas de cada enunciado a ser classificado. A maioria dos pesquisadores chegaram a um acordo de que as características globais são superiores às locais em termo de acurácia e tempo de classificação, além do número de características ser muito menor. Porém alguns pesquisadores já demonstraram que as globais são eficazes apenas quando se quer distinguir emoções com alta excitação de emoções com baixa excitação (EL AYADI; KAMEL; KARRAY, 2011).

Segundo El Ayadi, Kamel e Karray (2011), as características de fala podem ser divididas em quatro categorias, sendo elas: contínuas, qualitativas, espectrais e baseadas em energia. Para o reconhecimento de emoções é comum se empregar de uma combinação de características provenientes de diferentes categorias para representar o sinal da fala. Quando se trata de características contínuas, pode-se agrupar estas em cinco subcategorias como relacionada a tom, formante (máximo espectral), baseado em energia, tempo e articulação.

Para as características qualitativas para a voz, é entendido que o conteúdo emocional presente em um enunciado é altamente relacionado com essas características e são então divididas em quatro subcategorias, sendo estas o nível da voz; o a frequência fundamental da voz; as estruturas temporais e frases, fonemas, palavras; e limites de características. Contudo, existe um problema quanto ao papel dessas características no reconhecimento de emoções por dois motivos, um deles é que os rótulos para descrever a tensão, quão ofegante é a voz e quão dura é a voz são rótulos que dependem muito do compreendimento por parte do pesquisador; o segundo motivo é a dificuldade de classificar automaticamente esses termos diretamente do sinal da fala (EL AYADI; KAMEL; KARRAY, 2011).

As características espectrais podem ser extraídas de diversas maneiras, como no uso de coeficientes de predição linear, coeficientes de predição linear autocorrelatos unilateralmente, método de coerência de curto tempo e mínimos quadrados modificados por Yule-Walker. As características baseadas em TEO são muito importantes quando há a necessidade de qualificar o estresse presente em uma fala. Conclui-se que a escolha das características para o reconhecimento de emoções depende muito da tarefa de classificação que é considerada, principalmente quanto a quais serão as emoções a serem definidas para o sistema (EL AYADI; KAMEL; KARRAY, 2011).

Em Lijiang Chen et al. (2012) definem-se características tradicionais e características de domínio de frequência. Para as tradicionais são exemplificadas a energia, taxa de cruzamento zero, tom e formantes. Para as características de domínio de frequência são exemplificados o centroide espectral e a frequência de corte espectral, que refletem a distribuição de frequência em sinais de fala. Além dessas características, também se realizou a correlação de densidade e dimensão fractal, na qual a correlação reflete a distribuição espectral a curto tempo do sinal analisado e a dimensão fractal reflete a não-linearidade do sinal. E por último, a banda de energia baseada nas frequências-Mel são empregados devido à sua consistência com a percepção humana quanto às características de frequência de um som.

Para Ramakrishnan e El Emary (2013), um modelo de referência bom para o reconhecimento de emoções é o sistema auditivo humano. Em seu trabalho, o objetivo foi simular a percepção humana de emoções e identificar possíveis características que podem ser identificadas na fala independentemente do idioma, da pessoa ou mesmo do contexto no qual a fala está inserida. A frequência fundamental da fala pura e o contorno de energia podem ser utilizados como são e tem o nome de características de curto tempo, ou mesmo estas são usadas para criar

as outras características pelo uso de funções, muitas vezes estatísticas, sobre a sequência de valores de um segmento, conhecidas como características estatísticas globais.

A escolha do tipo de características que serão manipuladas influencia no tipo de classificador a ser empregado, por exemplo. Caso sejam extraídas características estatísticas globais, o indicado é aplicar um classificador estático, como as Máquinas de Vetor de Suporte (SVM, *Suport Vector Machine*), já no uso de características de curto tempo é indicado a implementação de classificadores dinâmicos como o Modelo Escondido de Markov (HMM, *Hidden Markov Model*) (RAMAKRISHNAN; EL EMARY, 2013).

De acordo com Ramakrishnan e El Emary (2013), há alguns grupos de características principais. Descritores de baixo nível consistem em características espectrais, sinal de tempo bruto, formantes, frequência fundamental, energia, MFCCs e qualidade da voz. Esses descritores são definidos através do enunciado como um todo e é reconhecido que estados emocionais diferentes possuem padrões prosódicos diferentes. Outro tipo de característica é a de relação com a duração de pausa, na qual a duração e distribuição dos segmentos com e sem voz são relacionados às características retiradas. É comum a utilização de características como número de regiões com e sem voz, número de quadros com e sem voz, razão entre quadros com voz e o total de quadros do enunciado, razão entre regiões com voz e o total de regiões e tremor na voz. As características Zipf são aplicadas para o dimensionamento melhor do ritmo e métrica da fala, definidas empiricamente por G. K. Zipf. Além destas características que são baseadas nos padrões acústicos verificados, é possível obter características linguísticas, quando o texto falado é analisado, de maneira a compreender as informações presentes na fala, como o uso de certas palavras ou alterações gramaticais.

No trabalho de Chaudhary et al. (2015), as emoções presentes na fala de um locutor são representadas por grande número de parâmetros, por isso a extração das características em um sistema de reconhecimento de emoções é uma parte bastante importante na elaboração do mesmo. As características, assim como para textciteramakrishnan 2013 speech, podem ser divididas em duas categorias: curto prazo e longo prazo. As características prosódicas são conhecidas como os indicadores primários do estado emocional passado pelo locutor e estudo de emoções presentes na fala concluem que a frequência fundamental, energia, intervalo, formante, MFCCs e o LPCC são características importantes para o reconhecimento das emoções presentes.

Uma das principais características da fala que indica uma emoção é a energia; seu estudo depende do curto prazo e a amplitude média verificada a curto prazo. A frequência fundamental,

também referida como forma de onda glotal, é outro indicativo para a emoção passada pela fala, podendo ser verificada na vibração produzida pela corda vocal quando se está reproduzindo um som ou fala e é caracterizada pela frequência do tom e pela velocidade do ar glotal. Os números de harmônicos presentes no espectro são diretamente associados à frequência fundamental. O LPCC apresenta detalhes sobre as características presentes no trato vocal de uma pessoa e este é alterado conforme a emoção a ser identificada. A vantagem ao utilizar o LPCC está no baixo uso de poder computacional, visto que seu algoritmo é mais eficiente e descreve as vogais de maneira superior. O MFCCs é muito usado em sistemas de reconhecimento de fala e de emoções através da fala, pois para frequências baixas a resolução e robustez a ruídos pode ser atingida. Se empregado como entrada todo o conjunto de características extraídas pode não ser garantido o desempenho do sistema para a classificação, por isso é necessária uma seleção sistemática para reduzir o número de características usadas e as características restantes podem ser aplicadas para aumentar a acurácia da classificação (CHAUDHARY et al., 2015).

Em Perez-Gaspar, Caballero-Morales e Trujillo-Romero (2016), a combinação entre SER e visão foi empregada de maneira que seu sistema de reconhecimento de emoções fosse multimodal. É verificado que para o SER as propriedades espectrais das vogais no enunciado são muito efetivas e possuem propriedades muito distintas entre os estados emocionais. Por isso a modelagem de vogais baseadas em emoções específicas foi considerada, tendo uma transcrição fonética da base de dados disponível.

Porém, o trabalho de Perez-Gaspar, Caballero-Morales e Trujillo-Romero (2016) foca no uso do idioma espanhol mexicano, tanto na base quanto no reconhecimento, tendo então um modelo específico de representações de fonemas e seus correspondentes emocionais baseados em mexicanos falando espanhol mexicano. Baseando-se nesse trabalho, seria necessário utilizar uma base de dados de brasileiros falando português brasileiro ou brasileiros falando inglês americano para o uso em conjunto com reconhecimento de voz presente no robô HERA (AQUINO JR et al., 2019).

Em Mao et al. (2017), é decorrido que a aplicação das características verificadas em bases de dados não necessariamente se faz presentes no mundo real, principalmente porque os sinais de fala não são muito similares em termos de locutores, tipos de emoções, situação de gravação e grau de espontaneidade. Porém, a adaptação de domínio é bem eficiente para esse tipo de problema. Para a extração de características foram empregadas redes neurais profundas, porém a divergência não foi explicitamente reduzida e para isso foi utilizada a divergência Kullback-Leibler.

Na aplicação de características para classificação foram usadas as acústicas baseadas na fala como um sinal bruto, tendo 12 funcionais no qual são utilizados Descritores de Baixo Nível. No final, é possível verificar 384 atributos (MAO et al., 2017).

Para a identificação de diferentes estados emocionais presentes nos sinais de fala é necessário realizar a extração das características marcantes presentes. Para isso os sinais foram normalizados e segmentados em quadros e as partes que não continham voz foram removidas para uma melhor classificação, e as características bi-espectrais foram extraídas tanto da onda da fala quanto da onda glotal (CK et al., 2017).

As Características Bi-Espectrais (BSF, *Bi-Spectral Features*) são extraídas do sinal, no qual estas são definidas como as frequências de domínio de terceira ordem do sinal que está sendo representado. A partir dessas características são definidas 14 características derivadas, como a magnitude média, a entropia de fase, o peso central entre outras. As BSF são funções de duas frequências e durante seu cálculo é obtido como resultado uma Transformada Discreta de Fourier (DFT, *Discrete Fourier Transform*) e seu resultado é uma matriz da frequência 1 pela frequência 2. BSF também são utilizadas no uso de Espectro de Ordem Maior (HOS, *Higher Order Spectrum*) que possui características vantajosas no uso com processamento de sinais de fala. Uma característica importante é a resistência a ruídos, além disso o HOS é muito empregado no processo de reconstrução de fase do sinal e essa informação é usada para a distinção de dois sinais diferentes. Por isso é possível concluir que as BSF são propícias para a diferenciação de estados emocionais diferentes e estresses, reduzindo a confusão entre os estados (CK et al., 2017).

No trabalho de Torres-Boza et al. (2018), a codificação de um estado emocional é definida como altamente complexa e apenas parcialmente compreendida. É possível definir que há duas estratégias para a extração de características perceptualmente motivadas. A primeira sendo a análise do áudio pelos modelos auditivos genéricos para obter as características específicas do conteúdo do áudio e a segunda extrai as características baseadas na correlação de dimensionamento de alto nível da fala.

Para as tarefas de reconhecimento é comum se utilizar as características como MFCCs, pois se verifica que o Coeficiente de Log de Frequência de Potência (LFPC, *Log-Frequency Power Coefficient*) na comparação com os MFCCs e o LFPC para o reconhecimento das emoções básicas concede resultados melhores de desempenho. Outras características também podem ser utilizadas para se averiguar as emoções presentes em uma fala. Além dos MFCCs, as características das dimensões perceptuais podem ser utilizadas para o reconhecimento de

emoções como a frequência fundamental, volume, identificação fonética e qualidade de voz. Em geral, observa-se que os valores da simetria glotal são bem efetivos para a classificação de emoções (TORRES-BOZA et al., 2018).

Para o trabalho de Torres-Boza et al. (2018), as características escolhidas para realizar o reconhecimento de emoções foram o conjunto F200 e o conjunto FPH. O F200 trata-se de um conjunto reduzido de características prosódicas, no qual são utilizadas as características mais comuns para medição de emoções e este é muito utilizado para o reconhecimento de emoções em segmentos de fala mais longos, como frases com maior duração e seu resultado é um vetor de 200 dimensões, enquanto o conjunto FPH é proposto como um complemento para o F200, a partir da qualidade de voz, medidas espectrais, medidas derivadas do modelo auditório e outras medidas baseadas em entonação.

O sinal de fala contém um número muito grande de parâmetros que refletem as características emocionais e um dos desafios é saber quais características devem ser utilizadas para o reconhecimento emocional. Para o trabalho de Kerkeni et al. (2018), as características modulares espectrais e as características MFCCs foram utilizadas para extrair os aspectos emocionais presentes. Para os MFCCs, são extraídos os primeiros 12 coeficientes da Transformada Discreta de Cosseno (DCT, *Discrete Cosine Transform*) para o processo de classificação. Como resultado obtém-se um vetor de 60 dimensões.

Já as características espectrais modulares são extraídas de uma representação espectro temporal de longo termo inspirada na parte auditiva, no qual são emulados os processamentos Espectro Temporais (ST, *Spectrum Temporal*) que são realizados pelo sistema auditivo humano e considera as frequências acústicas regulares. A representação das ST é formada pela medida de energia dos sinais já decompostos, como uma função de frequência acústica regular e uma frequência de modulação. No total são calculadas 95 características MSF dessa representação (KERKENI et al., 2018).

Um sistema de reconhecimento de emoções para múltiplas línguas, ou poliglota, tem um aspecto essencial no seu design que é a metodologia de modelagem de relações entre as emoções e as características da fala entre os idiomas utilizados. Para isso é necessário que o sistema tenha uma percepção emocional inspirada em um modelo de três camadas incorporando as características acústicas, primitivas semânticas e as dimensões emocionais. Para isso se estudou quatro corpus emocionais utilizando-se bases em diferentes idiomas como japonês, alemão, mandarim e inglês (LI; AKAGI, 2019).

As características acústicas podem ser divididas em prosódicas e espectrais. Para as características prosódicas, é possível agrupá-las em cinco categorias como frequências fundamentais, espectro de potência, envelope de potência, tempo e qualidade de voz. Para as espectrais utiliza-se o conjunto MSF e dois domínios são utilizados para cálculo, como o domínio de frequência de modulação e o domínio de frequência acústica. Com as características acústicas é possível obter 215 atributos para o sinal de fala sendo examinado (LI; AKAGI, 2019).

Para a execução do modelo de três camadas proposto por Li e Akagi (2019), é necessária a avaliação de primitivas semânticas e dimensões emocionais. Para a primeira foram utilizadas pessoas avaliando as frases reproduzidas e analisadas, além de terem sido selecionadas aleatoriamente, de maneira a ter apenas o modo com que foi falada a frase reproduzida; enquanto que para a avaliação de dimensionamento de emoções, a valência e empolgação foram empregadas, de modo a ter pessoas classificando as frases escutadas e a forma como são faladas, criando-se a correlação com o dimensionamento das emoções.

Em Hacine-Gharbi e Ravier (2019), é discutido que as características mais utilizadas para o reconhecimento de emoções são as características espectrais de curto prazo como o LPCC, o MFCCs e as características prosódicas. Porém um passo importante que talvez seja necessário para um sistema que faça esse reconhecimento seja a seleção das características relevantes após o passo de extração dos atributos, isso para que não haja redundância nas classes emocionais quando coletadas informações sobre as características que definem as mesmas. Esse processo diminui o tempo de computação e a capacidade de memória necessária para realizar o reconhecimento, além aumentar a acurácia do sistema (HACINE-GHARBI; RAVIER, 2019).

Zhao, Mao e Chen (2019) não fazem uso, por outro lado, da extração de características por métodos manuais, mas por características aprendidas. Isto é, as características são extraídas por diferentes redes profundas como máquinas restritas de Boltzmann (RBM, *Restrict Boltzman Machines*) baseadas em DNN e Rede Neural de Convolução (CNN, *Convolutional Neural Networks*), cujo uso deste tipo de característica tem se tornado cada vez mais popular.

Para o aprendizado profundo de características foi combinado o uso de um Bloco de Aprendizado de Características Locais (LFLB, *Local Feature Learning Block*) e a LSTM, tendo como foco a aprendizagem de características locais e globais dos áudios puros e espectrogramas de log-mel. Para o aprendizado das características locais foi utilizado o LFLB, um substituto das CNN, para a extração de características emocionais. Para as características globais utiliza-se a LSTM, uma arquitetura de Rede Neural Recorrente (RNN, *Recurrent Neural Networks*), para aprender as dependências contextuais de longo-prazo, visto que a LSTM é explicitamente

desenvolvida para o aprendizado de dependências de longo-prazo de sequências (ZHAO; MAO; CHEN, 2019).

Um dos aspectos mais importantes para o SER é a obtenção de características da fala que não sejam dependentes do locutor ou mesmo do conteúdo da fala e que possam eficientemente caracterizar o conteúdo emocional presente no enunciado a ser reconhecido emocionalmente. Para o trabalho de Özseven (2019) utiliza-se um método de seleção de características baseado nas emoções e suas mudanças de propriedades acústicas.

Em Bhavan et al. (2019), constata-se que o conjunto correto de características extraídas é a chave para o sucesso para a análise de emoções em uma fala. Para se analisar dados de fala usando técnicas de aprendizado de máquina, foram extraídas características espectrais das bases de dados e um vetor de características pra cada fala. As características extraídas foram os MFCCs, o Delta e o Delta-Delta dos MFCCs, além dos centroides espectrais. Como também os áudios que são então divididos em quadros de 25 ms e para cada quadro calcula-se: DFT, a potência espectral baseada em periodograma estimada, o banco de filtros de Espaço-Mel e o Logaritmo dos 26 valores obtidos do banco de filtros junto com a DCT (BHAVAN et al., 2019).

Luefeng Chen et al. (2020) adota uma não personalização das características emocionais da fala baseada na derivativa para sustentar as características tradicionais emocionais personalizadas. Isso resulta em 16 aspectos básicos e 12 características estatísticas. Características derivativas são menos afetadas de pessoa para pessoa, o que é visto como características não personalizadas. Para pessoas específicas, o uso de características personalizadas têm um bom resultado, porém caso seja utilizado com pessoas desconhecidas não pertencentes à base de dados, o reconhecimento tem uma queda na acurácia (CHEN, Luefeng et al., 2020).

O pré-processamento é o primeiro passo após a coleta de dados e algumas técnicas desse passo são focadas na extração de características, porém outras são focadas na normalização das características a fim de reduzir o efeito de diferentes locutores no processo de reconhecimento de emoções. Um método é o *framing*, que consiste em separar o dado da fala em quadros de 20 a 30 ms, possibilitando a extração de características locais, e esses quadros com tamanho fixo possibilitam o uso de classificadores como as Redes Neurais Artificiais (ANN, *Artificial Neural Networks*), além de reter a informação emocional na fala (AKÇAY; OĞUZ, 2020).

Após o uso do *framing*, geralmente se utiliza uma função de janela nos quadros criados a partir dos dados base. Essa função tem o papel de reduzir os efeitos de perda de dados durante a Transformada Rápida de Fourier (FFT, *Fast Fourier Transform*) por conta da descontinuidade das pontas dos sinais. Além da função de janela, é necessário compreender que a locução ou

fala possui três partes: com voz, sem voz e silêncio. A parte com voz é gerada pela vibração da corda vocal criando uma excitação periódica durante a pronúncia de fonemas. A parte sem voz é o resultado do ar que passa nas cordas vocais e produz sons turbulentos e transitórios e a parte do silêncio é a ausência de sons significantes para a identificação. A identificação das partes com voz entre os momentos de silêncio e sem voz é o que pode ser chamado de detecção de fala ou detecção de atividade vocal (AKÇAY; OĞUZ, 2020).

Akçay e Oğuz (2020) descreve também o processo utilizado no pré-processamento que é a normalização do dado, no caso característica, momento muito importante utilizado para a redução na variabilidade de locutor e gravação sem que seja perdido o poder classificatório das características. Esse processo aumenta a habilidade de generalizar as características a serem retiradas da fala. Outro aspecto importante é a redução de ruído, visto que o ruído pode afetar a taxa de reconhecimento. Para isso é muito comum se utilizar técnicas como Erro Quadrático Médio Mínimo (MMSE, *Minimal Medium Square Error*) e Amplitude Log-Espectral de Erro Quadrático Médio Mínimo (LogMMSE, *Log-Spectrum Minimal Medium Square Error*), cujo objetivo desses métodos é minimizar a distorção entre o sinal puro e o estimado.

Por fim, Akçay e Oğuz (2020) aponta que a seleção de características e a redução de dimensão é um passo importante para o reconhecimento de emoções devido à quantidade de características possíveis a serem verificadas e a incerteza do melhor conjunto de características para modelar as emoções a serem reconhecidas e classificadas. Essa seleção de características é o processo de escolha dos subconjuntos mais úteis e relevantes do conjunto de características escolhido nos quais os atributos desnecessários, redundantes e irrelevantes são identificados e removidos.

## 2.3 TRABALHOS RELACIONADOS

Para Jang e Kwon (2006) os sinais de fala contém emoções além das informações tradicionais retiradas através de uma fala que são os dados linguísticos presentes. A emoção presente em uma fala torna a comunicação natural, enfatiza a intenção do locutor e demonstra seu estado emocional. Um sistema de SER pode ser utilizado em robôs inteligentes que respondam a comandos do usuário de acordo com seu estado emocional ou até mesmo em um reprodutor de música para sugerir músicas e listas de músicas para o usuário.

A emoção através da fala pode ser reconhecida através de informações linguísticas ou mesmo por informações acústicas calculadas através do sinal do enunciado utilizado pelo usuá-

rio ao se comunicar com o sistema. Para o trabalho de Jang e Kwon (2006) foram utilizadas informações como tom, energia, formato, tempo, duração, tremor, brilho, MFCCs, coeficiente de predição linear codificado e energia Teager. Para realizar a classificação foi utilizada uma SVM e o desempenho do sistema foi comparada com o desempenho de ouvintes humanos. Observou-se uma acurácia de 58,6% e a classificação dos humanos para com o teste obteve uma acurácia de 60,4% (JANG; KWON, 2006).

No trabalho de Kwon et al. (2007) é introduzido um sistema de interação emocional para um robô de serviço. O trabalho viabilizou um robô para compreensão de estados emocionais para poder interagir emocionalmente com o usuário de maneira a expressar as emoções e compreender as emoções passadas a ele durante sua interação. Essas emoções normalmente são expressas por meio de expressões faciais, voz, linguagem, gestos e sinais fisiológicos. O sistema proposto possui três módulos, sendo eles um sistema de SER, um sistema de reconhecimento de emoções linguísticas e o reconhecimento de emoções por meio de informações de toque.

Robôs de serviço são os que operam de maneira autônoma para realizar tarefas para os usuários em um ambiente diário de sua vida. Para coexistir com os humanos e oferecer-lhes serviços, o robô deve compreender o estado emocional destes. Tais informações dos estados emocionais podem ser obtidas de diversas maneiras como pelas expressões faciais, gestos e através da fala. Com isso verifica-se que o SER é necessário para a Interação Humano-Robô em conjunto com o reconhecimento de fala (PARK; KIM; OH, 2009).

Park, Kim e Oh (2009) propõe o uso de uma classificação utilizando-se vetores de características em conjunto com o Modelo de Mistura Gaussiana (GMM, *Gaussian Mixture Model*) nos quais os testes são propostos para a identificação de duas classes para as emoções em uma fala, sendo as negativas e não-negativas. Os experimentos empregaram apenas exemplos de falas curtas sem conteúdo emocional intrínseco.

O trabalho de Huahu, Jue e Jian (2010) tem como foco o uso do SER para categorizar cinco emoções de fala, sendo aplicado para uma plataforma de um robô doméstico inteligente, de modo que este possa reconhecer as informações emocionais contidas nos sinais de fala durante a Interação Humano-Robô. Nesse trabalho dispõe-se um modelo híbrido de HMM com uma Rede Neural de Mapeamento de Características Auto-Organizadora (SOFMNN, *Self Organizing Feature Mapping Neural Network*). Robôs domésticos inteligentes podem se comunicar com o proprietário do lar de maneira natural e de modo amigável, podem realizar tarefas para toda a família e ser amigo de seres humanos.

Os testes foram conduzidos para quatro pessoas que realizaram as gravações de frases para a análise emocional. Depois das gravações, duas outras pessoas, que não passaram pelo processo de gravação, fizeram parte do processo de diferenciação das informações de fala gravadas e aquelas falas sem características com significância emocional foram removidas. Através da integração do HMM e SOFMNN foi possível obter um resultado mais satisfatório quando comparado com o uso do HMM sozinho (HUAHU; JUE; JIAN, 2010).

No trabalho de Zhang, Zhao e Lei (2013) propõe-se a redução de características da fala para o SER empregando-se do método Mapeamento Isométrico com Núcleo Aprimorado (EKI-somap, *Enhanced Kernel Isometric Mapping*) para a Interação Humano-Robô. Nesse trabalho realiza-se a extração de dados de treino e de teste, no qual ambos passam pela fase de extração de características acústicas e redução de dimensão usufruindo do método proposto. Para a classificação é empregada uma SVM em conjunto com a redução do vetor de características dos dados de treino.

As características utilizadas foram as prosódicas e de qualidade de voz retiradas da base de dados de Berlim. Conclui-se que a dimensão é importante no SER e a redução dela é uma estratégia importante para realizar o processamento de dados nesta tarefa. Os resultados da pesquisa demonstraram que o uso do método proposto é promissora e há uma possibilidade alta de aplicação do sistema de SER para a Interação Humano-Robô (ZHANG; ZHAO; LEI, 2013).

O uso de Modelos escondidos de Markov em conjunto com Redes Neurais profundas (DNN-HMM, *Deep Neural Networks - Hidden Markov Model*) tem apresentado resultados promissores na aplicação em tarefas de reconhecimento de fala quando comparado com o uso de GMM em conjunto com os HMM. No trabalho de Li et al. (2013) é verificado o uso das DNN-HMM com as RBM para o SER.

Em comparação com os modelos que usam HMM é possível se verificar que os modelos que os empregam não são eficientes, estatisticamente falando, e é essa melhora que o modelo apresentado por Li et al. (2013) traz para a área de SER. Quando o número de camadas ocultas é definido corretamente, o modelo DNN-HMM possui resultados satisfatórios. É aplicado um modelo misto híbrido DNN-HMM utilizando como características os MFCCs e foi possível obter um valor de 53,89% de acerto (LI et al., 2013).

A análise de características é muito menos estudada na área de reconhecimento de emoções quando comparada ao campo de reconhecimento de fala. Em diversos trabalhos estuda-se empiricamente a seleção de características acústicas para a classificação (HAN; YU; TASHEV, 2014). Nesse trabalho emprega-se uma DNN para receber as características acústicas tradi-

cionais e criar distribuições probabilísticas para os estados emocionais a nível de segmentos para os enunciados de fala. Para a classificação é implementada então uma rede neural, com apenas uma camada escondida, chamada de aprendizado de máquina extremo que possui um desempenho superior quando comparada ao desempenho das SVM.

No trabalho de Haytham M Fayek, Lech e Cavedon (2015) é proposto um sistema de reconhecimento de emoções através da voz para reconhecimento em tempo real aplicando aprendizado profundo. Propõe-se o uso de quadros de 1 segundo de duração para a classificação e estudo das emoções retirados de espectrogramas, brutos ou sem tratamento, da fala. Usufruiu-se da base de dados eNTERFACE'05 e, através do uso de uma DNN, foi possível obter 60,53% de acerto na classificação de emoções através da fala utilizando-se uma técnica de aprendizado profundo de ponta-a-ponta por meio dos áudios para o aprendizado da rede.

Tradicionalmente as máquinas não possuem a associação de racionalidade ligada aos seus sistemas de decisão, porém, recentemente com pesquisas de neurociência, uma área emergiu e é conhecida como computação afetiva, na qual as máquinas devem reconhecer, expressar, modelar, comunicar e responder aos indicadores de emoções dos usuários. O principal foco dessa área são robôs e sua conexão com humanos que envolve a interação e um *framework* com resposta emocional baseada nas emoções humanas verificadas na fala (RÁZURI et al., 2015).

Em Rázuri et al. (2015) propõe-se a comparação de diferentes classificadores mais conhecidos com as características mais comumente aplicadas em diversas pesquisas para capturar as nuances emocionais de falas. Para a redução de tamanho das características da fala e melhorar os resultados dos classificadores utilizou-se a saída de uma árvore de decisão para seleção de características. Por fim, foi verificado que o Perceptron de Multicamadas (MLP, *MultiLayered Perceptron*), as SVM e as Rede de Bayes (BayesNet, *Bayes Network*) obtiveram os melhores resultados e quando é proposto para o uso em um robô, tanto as SVM quanto a BayesNet são as melhores opções devido à sua facilidade de implementação e baixa complexidade computacional.

Em Perez-Gaspar, Caballero-Morales e Trujillo-Romero (2016), um sistema multimodal para reconhecimento de emoções é proposto, tendo como principal foco o uso em interações humano-robô. Alguns robôs são analisados, como o robô Kismet, que foi desenvolvido para a pesquisa de como a percepção e interação com humanos é alterada com base nas emoções expressadas pelo sistema robótico e Jibo que realizava tarefas e era vendido como um robô de companhia. Para esse trabalho reconhece-se emoções multimodais, utilizando-se o sistema de visão e audição para tal reconhecimento de emoções na interação com o robô, no caso um robô

humanoide. Após os testes, verificou-se um reconhecimento com 97% de precisão, empregando uma base de dados usuário provenientes do México falando espanhol, e esta base foi utilizada na aplicação testando pessoas mexicanas falando o idioma espanhol.

No trabalho de Trigeorgis et al. (2016) se faz o estudo do uso de CNN com redes LSTM para aprender automaticamente a melhor representação diretamente do sinal de voz em comparação com os métodos tradicionais de seleção e extração de características. A base RECOLA é usada para realizar experimentos e foi possível obter resultados melhores quando comparado ao design de características de métodos tradicionais, demonstrando uma eficácia melhor no aprendizado de características necessárias para a tarefa. Ao contrário de outros trabalhos, o sinal utilizado para realizar o reconhecimento é a onda de maneira pura, tendo apenas sido feito o pré-processo para variações de altura do sinal dos leitores.

Emoções possuem um papel importante na comunicação entre humanos e estas podem ser expressas de maneira diferente, dependendo da cultura do locutor, contexto e condições do ambiente em que circulam. Por esses motivos, modelos aprendidos através de dados não estruturados são muito efetivos. Reconhecimento de padrões complexos tiveram um grande avanço nos últimos anos, como é visto no avanço do reconhecimento de fala. Um grande problema, porém, para aplicações em robótica é a diferença entre os dados de treino e de teste, como a maneira com o qual tipos de microfones se comportam conforme o ambiente e suas condições (LAKOMKIN et al., 2018).

O trabalho de Lakomkin et al. (2018) tem como objetivo avaliar o desempenho do estado da arte do modelo neural de SER em uma configuração de emulação de cenários do mundo real, como a interação em um ambiente ruidoso. São propostas também a avaliação de técnicas de acréscimo de dados de fala e o análise dos efeitos deles no desempenho em modelos neurais em condições ideais. Conclui-se que a desempenho sem o acréscimo de dados e com o uso de dados limpos é significantemente menor e a robustez do sistema tem um aumento quando as técnicas propostas são aplicadas. Por isso é crucial incluir esse acréscimo nos dados de treinamento quando se pretende empregar o sistema em um robô (LAKOMKIN et al., 2018).

O reconhecimento de afeto é um componente importante para uma melhor interação entre humano e máquinas, sendo vital considerar emoções durante a interação. As Redes Neurais profundas emergiram nos últimos anos e apresentaram diversas melhoras na área de aprendizado de máquina. Estudos recentes utilizam DNN para realizar o SER, porém muitos empregam características produzidas manualmente. Aqui se propõe uma arquitetura de Redes Neurais convolucionais recorrentes de ponta-a-ponta para o reconhecimento contínuo de emoções. O

modelo proposto atingiu os resultados verificados pelo estado da arte, tendo como comparação estudos que tiveram como base os testes da base de dados RECOLA (TZIRAKIS; ZHANG; SCHULLER, 2018).

A área de robótica tem se espalhado e se desenvolvido bastante, recentemente e cada vez mais é esperado que a Interação Humano-Robô se torne mais humana e natural e entenda a intenção do usuário. Além disso é esperado que os robôs possam expressar e até mesmo reagir à emoções, uma vez que essas são laço importante entre os humanos e robôs durante a interação que pode ser verificado através dos sinais da fala, expressões faciais e até mesmo sinais psicológicos (CHEN, Luefeng et al., 2020). Neste se apresentou o uso de Florestas Aleatórias (RF, *Random Forests*) para o reconhecimento de emoções através da fala e características personalizadas e não-personalizadas são usadas para o SER. Realiza-se a identificação de informações humanas para a soma das características identificadas para a realização do reconhecimento de emoções (CHEN, Luefeng et al., 2020).

Para a extração do conjunto de características é utilizada a ferramenta openSMILE e características básicas como energia Raiz Quadrada Média (RMS, *Root mean Square*), Taxa de Cruzamento em Zero (ZCR, *Zero Crossing Rate*), proporção de harmônicos da fala e MFCCs são obtidas. A estrutura de aplicação dos testes é um robô móvel, uma estação de trabalho computacional, um computador pessoal, um roteador e um equipamento de transmissão de dados. Os dados são capturados pelo robô, transmitidos para a estação de trabalho que insere as informações recebidas no sistema de SER para reconhecer a emoção. A resposta é então devolvida para o robô compreender a emoção humana e ter alguma reação. O computador pessoal é empregado para controle de problemas no sistema principal, na estação de trabalho, quando não há supervisão humana presente na mesma. Ao final é verificado que o sistema conseguiu identificar as emoções básicas e caso seja desejada uma maior acurácia, o uso de um sistema multimodal é necessária (CHEN, Luefeng et al., 2020).

Os trabalhos estudados para a Revisão Bibliográfica e para os Conceitos foram selecionados e agrupados em uma tabela, apresentada em Apêndice A, visando observar os objetivos dos trabalhos, a data de publicação, quais foram as características utilizadas, quais os métodos de seleção de características, quais classificadores foram empregados e quais bases de dados foram usados pelos pesquisadores da área. Devido à sua extensão, a tabela foi repartida de maneira a apresentar as características observadas para cada trabalho, tendo como a primeira coluna o nome dos autores do trabalho analisado e as consequentes sendo: Título do Trabalho, Ano, Objetivo do Trabalho, Características Utilizadas, Seleção de características, Classificado-

res e Base de Dados. Foi possível observar que características com base na frequência Mel são populares entre os autores que buscam trabalhar com SER. É possível observar também que os classificadores baseados em SVM e Redes Neurais são bastante empregados.

# 3 CONCEITOS

Nesta seção, são discutidos os conceitos utilizados para o trabalho, passando por seções como Bases de Dados, Classificadores e Redes Neurais Profundas.

## 3.1 BASES DE DADOS

Bases de dados são muito importantes para o processo de classificação das emoções do SER. As bases podem ser classificadas em três tipos: atuação, indução e natural. As bases de dados de atuação são feitas por atores, profissionais ou semiprofissionais, em salas acústicas de estúdios. Essas bases são as mais fáceis de serem criadas, porém constatou-se que as emoções atuadas não são tão próximas das emoções reais e isso diminui a acurácia do reconhecimento de emoções reais (AKÇAY; OĞUZ, 2020).

Enquanto as bases de dados de indução são as que o locutor é colocado em uma situação de simulação emocional na qual há o estímulo de diversas emoções. Essas emoções, por não serem completamente atuadas, são mais próximas das reais. Por último, as bases de dados naturais que são obtidas de *talk shows*, gravações de *call-centers*, entrevistas em rádios e outras fontes similares em que há interação natural. Por questões éticas e legais, é mais difícil obter tais dados (AKÇAY; OĞUZ, 2020).

Um exemplo de base de dados é a Livingstone e Russo (2018), que conta com 24 atores representando o sotaque de inglês presente nos Estados Unidos da América, balanceado entre gêneros e com 7 emoções classificadas. Em Cao et al. (2014), seis emoções foram classificadas utilizando-se de noventa e um atores com diversidade étnica. Os dados dessa base de dados são usados para classificação humana dos sentimentos, tendo como informação exemplos de áudio, imagem e áudio com imagem. Também é possível se empregar a base de dados MSP-Improv corpus, no qual foram usados doze atores para gravar diversos exemplos audiovisuais de emoções, sendo estas classificadas em cinco classes.

Para o trabalho corrente, o idioma principal utilizado foi o inglês, devido a sua aplicação para o reconhecimento de fala, que ocorre durante a interação por meio da voz com o robô HERA. Algumas bases que se encaixam com a premissa do trabalho são: Survey Audio-Visual Expressed Emotion (SAVEE), Toronto Emotional Speech Database (TESS), Electromagnetic Articulography Database (EMA), eNTERFACE'05 Audio-Visual Emotion Database e AFEW Database.

Tabela 1 – Tabela de Comparação das Bases de Dados

| **Base de Dados** | **Idioma** | **Tamanho** | **Custo** | **Emoções** | **Modalidade** | **Formato** |
|---|---|---|---|---|---|---|
| *Surrey Audio-Visual Expressed Emotion (SAVEE)* | Inglês | 14 Homens e 120 frases (total de 1680 frases) | Grátis | Comum, Desgosto, Felicidade, Ira, Medo, Neutro, Surpresa, Tristeza | Atuação | Áudio / Visual |
| *Toronto Emotional Speech Database (TESS)* | Inglês | 2 Mulheres e 2800 frases (total de 5600 frases) | Grátis | Ira, Desgosto, Neutro, Medo, Felicidade, Tristeza, Prazer, Supresa | Atuação | Áudio |
| *Electromagnetic Articulography Database (EMA)* | Inglês | 2 Mulheres e 1 Homem, 14 frases para o Homem e 10 frases para a Mulher (total de 34 frases) | Grátis Para Pesquisas | Ira, Felicidade, Tristeza e Neutro | Atuação | Áudio / Dado de Movimento Articulatório |
| eNTERFACE'05 *Audio-Visual Emotion Database* | Inglês | 34 Homens e 8 Mulheres (Total de 1116 vídeos) | Grátis | Ira, Desgosto, Medo, Felicidade, Tristeza e Surpresa | Indução | Áudio / Visual |
| AFEW Database | Inglês | 330 Locutores com 1426 Frases | Grátis | Ira, Desgosto, Surpresa, Medo, Felicidade, Neutro e Tristeza | Natural | Áudio / Visual |

Fonte: Autor

Para uma melhor visualização das informações das bases que encaixam na premissa do trabalho corrente, foi gerada a tabela 1, que apresenta o nome da base de dados, o idioma utilizado nas frases gravadas, o tamanho da base em questão de frases ou vídeos, o valor para uso da base, as emoções classificadas, a modalidade da base de dados e por último o formato dos dados presentes na base de dados.

A base de dados escolhida para realizar o treinamento e os testes foi a eNTERFACE'05 Audio-Visual Emotion Database que conta com 42 locutores, 34 homens e 8 mulheres, de 14 nacionalidades diferentes. A base possui 1116 vídeos com falas que variam de 2 a 5 segundos de duração (MARTIN et al., 2006). Conta ainda com dados áudio visuais tendo seis emoções representadas, sendo estas: ira, desgosto, medo, felicidade, tristeza e surpresa.

## 3.2 CLASSIFICADORES

Classificadores são algoritmos e técnicas que são empregados para realizar a classificação e, ou, reconhecimento de classes, sejam estas múltiplas ou binárias. Para a classificação de emoções através da fala, podem ser utilizados diversos classificadores, como foi verificado durante a revisão dos trabalhos relacionados. Alguns desses possíveis classificadores serão citados nos próximos parágrafos para apresentação de outras possibilidades usadas por pesquisadores que realizaram o trabalho de reconhecimento de emoções através da fala.

Em estudos recentes, Akçay e Oğuz (2020) descreve que os sistemas de reconhecimento de emoções classificam as emoções dadas uma fala ou oração. Além dos classificadores tradicionais e dos algoritmos de aprendizado profundo, alguns algoritmos de aprendizado de máquina são utilizados para essa tarefa, porém não há um algoritmo de aprendizado de máquina que seja aceito comumente e os estudos quanto a esses algoritmos são empíricos.

Para os classificadores tradicionais é comum empregar os HMM, GMM, SVM e ANN. Porém ainda é possível ver métodos de classificação utilizando Árvores de Decisão, k Vizinhos Próximos (K-NN, *k-Nearest Neighbours*), k-means e Naive Bayes (AKÇAY; OĞUZ, 2020).

Em Wu, Falk e Chan (2011) estuda-se o uso de SVM para a classificação das emoções em conjunto ao uso das MSF classificando sete emoções humanas diferentes. No uso das MSF em conjunto com as características prosódicas obteve-se uma acurácia de 91.6%.

Já em Lijiang Chen et al. (2012) se faz uma comparação no uso do discriminador Fisher e a Análise de Componente Principal (PCA, *Principal Component analysis*) para testar uma ANN e uma SVM na tarefa de SER. Nesse estudo foi possível verificar que ao aplicar Fisher

ao invés do PCA, os resultados foram melhores enquanto as SVM são melhores ao tratar de um reconhecimento que independe do locutor no reconhecimento de emoções quando comparado a uma ANN.

Assim como em Lijiang Chen et al. (2012), no trabalho de Kerkeni et al. (2018) é feito uma comparação, porém nesse foram utilizados três classificadores para a comparação de duas bases de dados, sendo estes: Regressão linear de multivariáveis, SVM e RNN. Como resultado é visto que ao empregar os MFCCs e MS, o melhor resultado vem da utilização do classificador aplicando regressão linear, atingindo um total de 90,05% de acerto no reconhecimento.

Em Özseven (2019) foi proposto um método de seleção de características e os classificadores usados para a classificação de emoções a partir da proposta realizada foram a SVM, o K-NN e o MLP. Nessa comparação verificou-se que o K-NN exigiu uma carga de trabalho menor, o MLP teve maior taxa de acerto e o SVM teve menor carga de cálculo para realizar a tarefa de SER .

Bhavan et al. (2019) propõe a utilização de uma SVM em conjunto com um núcleo Gaussiano para o SER. Ao usar os MFCCs e centroides espectrais para a classificação em conjunto com um seletor de características para formar o conjunto de melhores características, o conjunto com apenas características acústicas foi criado e empregado para os testes. Após testes, verificou-se, em média, 84.1% de acurácia utilizando de Centroides Espectrais e MFCCs como características para classificação.

Em pesquisas recentes, o desempenho dos algoritmos de aprendizado profundo tem superado os algoritmos tradicionais de aprendizado de máquina e a grande vantagem ao se empregar esses algoritmos é a não necessidade de extrair características e realizar a seleção de características, já que estas são selecionadas automaticamente pelos algoritmos de aprendizado profundo. Os tipos de algoritmos mais utilizados quando se fala na área de SER são as CNN e as RNN (AKÇAY; OĞUZ, 2020).

Trigeorgis et al. (2016) propõe o uso da combinação de CNN em conjunto com as Redes LSTM para o aprendizado automático das melhores características para a representação das emoções através da fala, tendo um desempenho superior às técnicas tradicionais baseadas em técnicas de processamento de sinais. O trabalho de Trigeorgis et al. (2016) baseou-se na RECOLA, tendo seus testes focados nessa mesma base de dados.

Algumas pesquisas como a de Torres-Boza et al. (2018) e a de Luefeng Chen et al. (2020) estudam o uso de *Machine Learning* no SER.

Em Torres-Boza et al. (2018) propõe-se uma arquitetura HSC usufruindo-se duas bases de dados conhecidas, a VAM-Audio Database e a AVEC2012 Challenge Database. De início, características perceptuais de alto-nível foram propostas para uma melhor representação da percepção do conteúdo emocional. No uso das FPHs já houve uma melhora na predição quando comparado com o uso de características prosódicas sozinhas, e no uso combinado de ambas a representação emocional foi melhorada. Quando foi empregado o sistema de aprendizado não supervisionado, verificou-se que foi possível obter características úteis para a distinção das emoções.

Neste trabalho utilizou-se classificadores baseados em Redes Neurais Profundas e Redes Neurais Recorrentes, devido a recorrência no uso de redes neurais profundas para a classificação de emoções através da fala e os resultados positivos apresentados, como é visto em Akçay e Oğuz (2020). Ambos serão explicados na Seção de Redes Neurais Profundas.

## 3.3 REDES NEURAIS PROFUNDAS

Técnicas de aprendizado de máquina têm sido aplicadas em diversas áreas como reconhecimento de padrões, processamento de linguagem natural e aprendizado computacional. Na última década, o aprendizado de máquina tem influenciado a vida diária com os mecanismos de busca, visão computacional e reconhecimento de caracteres. O conceito de aprendizado profundo é originário do estudo de ANN (LIU et al., 2017).

ANN são criadas ao empregar *perceptrons*, que são neurônios artificiais, para produzir ativações baseadas em valores para realizar cálculos e obter-se um valor esperado. A criação das técnicas de aprendizado profundo ocorre em 2006 com a ideia de Hinton ao criar o aprendizado guloso de camadas para um pré-treino da rede antes de realizar o treino de camada em camada (LIU et al., 2017).

A arquitetura das redes profundas é comumente aplicadas no reconhecimento de fala e modelagem acústica para classificação de áudio, assim como são muito utilizadas na área de processamento de imagens (LIU et al., 2017).

Para Witten et al. (2017), o aprendizado profundo teve um grande impacto na área de reconhecimento de fala e visão computacional e para outras áreas como processamento de linguagem natural se começa a observar seus benefícios. O aprendizado profundo formula um problema dentro de uma arquitetura de rede no qual a camada de saída define uma função de perda necessária para o aprendizado. Métodos de aprendizado profundo são baseados em redes

nas quais o algoritmo de *backpropagation* tem o papel de computar os gradientes e atualizar os parâmetros do modelo a partir de pequenos subconjuntos do conjunto de treinamento.

Para este trabalho, definiu-se pela aplicação de uma rede neural profunda e uma rede neural profunda com uma camada LSTM, tornando-a uma rede neural recorrente profunda.

A LSTM é uma arquitetura usada em redes neurais recorrentes que se utiliza de uma memória para armazenar valores durante intervalos definidos. Uma vantagem da LSTM é a indiferença apresentada ao tratar de intervalos durante treinamento e classificação. Isso se deve à arquitetura da célula, que possui uma estrutura em cadeia e que apresenta células para manipulação da informação. A manipulação da informação a partir da memória é feita pelos portões dessa estrutura, sendo estes: Portão de Esquecimento, Portão de Entrada e Portão de Saída (ACADEMY, 2021).

Detalhes da configuração das redes neurais empregadas estão descritas no capítulo Metodologia.

## 3.4 CARACTERÍSTICAS DA FALA

Para realizar a extração do MFCCs, primeiro se realiza a separação do sinal em quadros de 20 a 40 milissegundos. Para cada quadro é, então, extraído o conjunto de MFCCs. Para cada quadro é realizado a DFT, utilizando-se a equação abaixo para realizar o cálculo.

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N} \tag{1}$$

Nesta equação s(n) é o sinal do domínio do tempo, h(n) é a janela de análise de tamanho N e K é o tamanho do DFT. A estimativa de potência espectral baseada no periodograma para cada quadro da fala é definida pela equação (2). Após esse cálculo é computado o banco de filtros do espaço Mel e para cada energia é realizado o logaritmo, resultando no Log das energias dos bancos de filtros. Por fim, é aplicado a Transformada Discreta de Cosseno desses valores para a obtenção dos Coeficientes Cepstrais (LYONS, 2012).

$$P_i(k) = \frac{1}{N} |S_i(k)|^2 \tag{2}$$

Outra característica utilizada foi o Delta-MFCC, no qual é realizado o cálculo dos coeficientes delta que pode ser verificado através da equação (3) (LYONS, 2012).

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{3}$$

Para o Espectrograma-Mel, primeiro é computado o espectrograma de magnitude a partir do sinal de fala usado, para isso é utilizado na equação 1 para os segmentos da fala que será analisada, gerando então um espectrograma. Após o cálculo do espectrograma, o mesmo é gerado em Hz, porém como é desejado trabalhar com a frequência Mel, é utilizada a equação abaixo para converter de Hertz para a Escala Mel.

$$mel = 2595log_{10}(1 + \frac{hz}{700}) \tag{4}$$

Para o contraste espectral, o espectrograma é dividido em quadros e estes são divididos em sub-bandas. Para cada sub-banda é estimado o contraste de energia, comparando a média de energia do pico e do vale de energia, onde são verificadas as k sub-bandas. Para realizar a extração do pico espectral e vale espectral, são utilizadas as equações abaixo, representando o cálculo do pico e vale respectivamente.

$$Pico_k = log\left\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,i}\right\} \tag{5}$$

$$Vale_k = log\left\{\frac{1}{\alpha N}\sum_{ß=1}^{\alpha N} x'_{k,N-i+1}\right\} \tag{6}$$

Após o cálculo do Pico e Vale, é realizada a diferença entre os mesmos.

Para o cálculo do Espectrograma-Mel e do Contraste espectral foi utilizado, assim como para o cálculo dos MFCCs e do Delta-MFCCs, a biblioteca Librosa para linguagem de programação Python. A biblioteca Librosa permite realizar o cálculo de características a partir de um sinal de fala (MCFEE et al., 2015). Em total, foram utilizados 40 valores obtidos nos MFCCs, 40 valores do delta do MFCCs, 128 valores para o espectrograma-Mel e 7 valores para o contraste espectral, totalizando em um conjunto de 215 valores para o conjunto de características utilizadas.

# 4 METODOLOGIA

O foco deste trabalho é realizar a análise de sentimentos e emoções aplicando a análise em um sistema de SER. Para que se desenvolvesse este trabalho, diversos sistemas estudados por outros autores com pesquisa na área do SER foram analisados e observou-se que pesquisas atuais buscaram utilizar técnicas de aprendizado profundo para a realização da seleção de características. Esses sistemas propostos nos trabalhos de outros autores provaram que houve melhora no desempenho de aprendizado e classificação das emoções. Portanto, com base nas referências de sistemas de classificação de emoções através da fala foi empregada uma DNN e uma RNN para realizar a coleta das melhores características e classificação das emoções através da fala. Escolheu-se a utilização de uma abordagem quantitativa, avaliando a acurácia dos classificadores com base na classificação realizada pelo sistema escolhido, com o intuito de se verificar a qualidade do sistema proposto quando se pensa no uso das bases atualmente disponíveis e as falas no idioma inglês.

Considerou-se falas de diferentes durações, assim como previsto na tarefa *General Purpose Service Robot* (GPSR) na RoboCup@Home (ROBOCUP-FEDERATION, 2020), na qual são geradas frases a serem requeridas por um usuário através da fala, com a possibilidade de ser frases complexas como "Vá para a mesa da sala de jantar, ache a garrafa de água e leve para a pessoa localizada no sofá", ou frases simples como "Busque um copo para mim". A diferença de tempo nas interações é considerada devido às características como duração, espectro temporal estarem presentes nas características manuseadas em outras pesquisas como em Ramakrishnan e El Emary (2013) e Kerkeni et al. (2018).

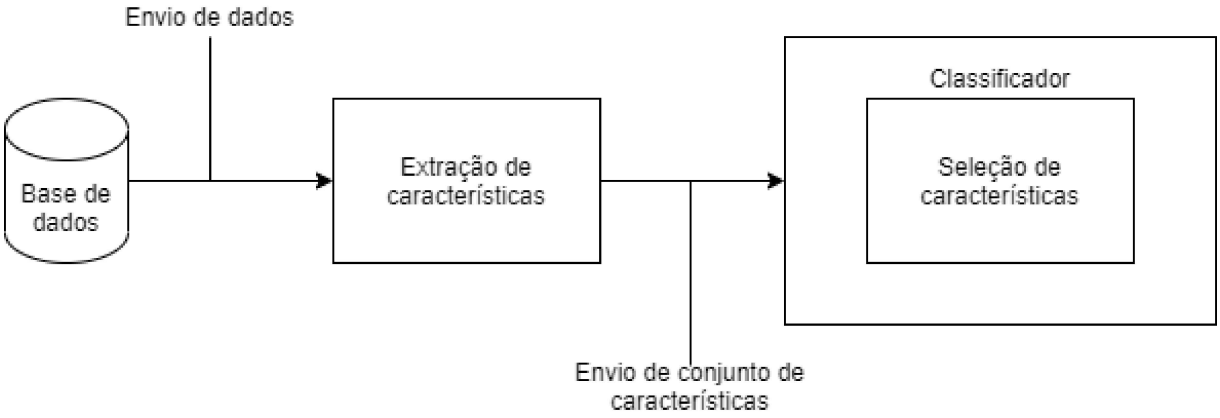
Após os estudos de trabalhos da área de Reconhecimento de Emoções, verificou-se que um dos maiores desafios fora a escolha de características para representar as emoções passadas pela voz em uma fala. Nos primeiros testes, utilizaram-se algumas características como MFCCs, Delta-MFCC, o Espectrograma na escala Mel, Contraste Espectral de Energia. Para os testes, realizou-se a comparação entre o uso de uma rede neural LSTM e uma Rede Neural Profunda. O uso de redes neurais providencia resultados com acurácia alta, mesmo que a carga de cálculo seja maior que quando se emprega SVM, como foi verificado em Özseven (2019).

Para a implementação elencou-se a linguagem de programação Python, visando a utilização futura em conjunto com outros pacotes feitos para uso no robô HERA, além do uso em outras áreas de pesquisa com interação através da fala. Para a implementação das redes neurais neste trabalho empregou-se o Keras, uma API feita para o uso do Tensorflow e criação de

redes neurais, tendo como facilitar a programação das redes, além de ser muito empregado por diversas empresas como a NASA (KERAS, 2021).

O diagrama ilustrado na figura 2 representa o treinamento do sistema proposto. Os dados são recolhidos da base de dados e enviados para o módulo de extração de características que realiza a extração de características definidas e as envia como um conjunto para a aplicação como entrada das redes neurais durante a fase de treino. A seleção das características é realizada pela rede neural, que decide qual a melhor representação para cada classe, ou emoção no caso, com base nas características apresentadas.

Figura 2 – Diagrama do Funcionamento do Sistema Proposto




Fonte: Autor

Para realizar a validação da rede neural, 10% dos dados da base de dados escolhida foram separados para realizar os testes de validação para ambas redes neurais, a LSTM e a DNN.

Durante a realização de testes o procedimento realizado é semelhante ao empregado durante o treinamento, no qual os dados de teste são colhidos e enviados para a extração de características. Após isso, o conjunto de características é enviado para a rede neural treinada que realizará a classificação conforme os pesos aprendidos durante o treinamento. A resposta do sistema é apresentada pela rede neural, classificando os áudios utilizados para teste conforme a possível emoção apresentada.

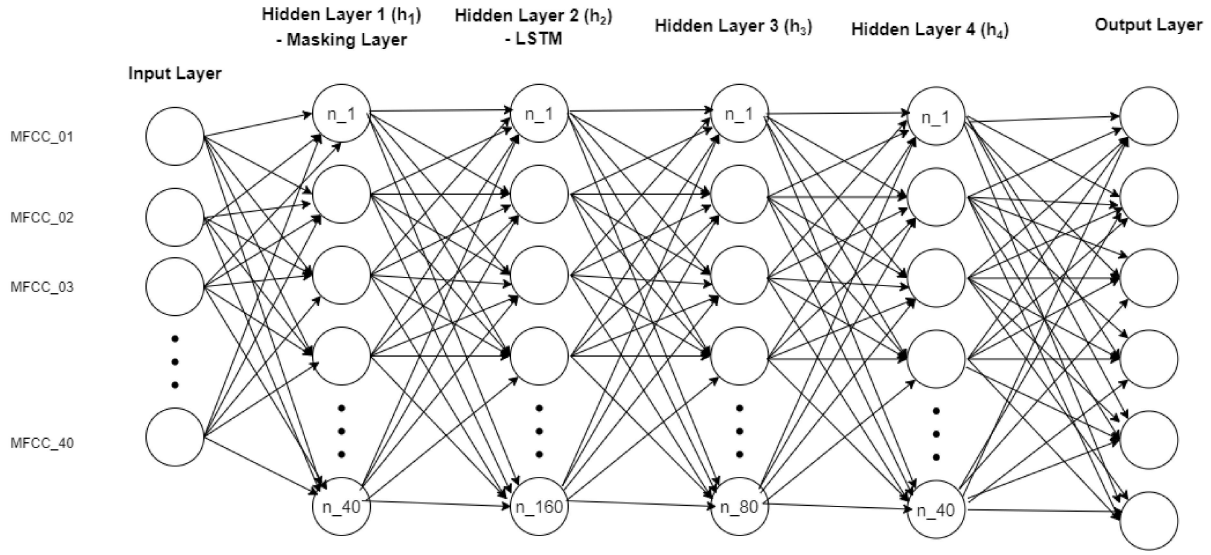
Implementou-se duas redes neurais para a realização dos testes, uma rede neural LSTM profunda e uma rede neural profunda.

A rede neural LSTM possui uma entrada que alimenta uma primeira camada da rede, a máscara. A máscara identifica os valores que deverão ser ignorados durante o treinamento e classificação. Tornou-se necessário o uso de uma camada de máscara devido à normalização da

entrada a ser analisada. Os áudios possuem tamanhos diferentes e as matrizes de características retiradas também apresentavam tamanhos distintos. Foi então realizada uma normalização no tamanho da matriz de característica, aplicando um tamanho único para todas as matrizes, porém isso resulta em adicionar valores que poderiam prejudicar a classificação, para isso a camada de máscara da rede foi inserida, visando observar e remover os valores que foram incluídos durante a normalização para a classificação.

A segunda camada da rede neural recorrente é uma camada LSTM, que recebe os dados de saída da camada de máscara e trabalha com espaços de tempo para classificação dos dados recebidos. No caso dos testes usando apenas as características dos MFCCs, essa camada possui 160 unidades para classificação como é visto na figura 3. Como é visto na figura 4, as unidades empregadas nas camadas dependem do número de características utilizadas como entrada para a rede neural.

Figura 3 – Representação da Rede Neural Recorrente utilizando apenas MFCCs

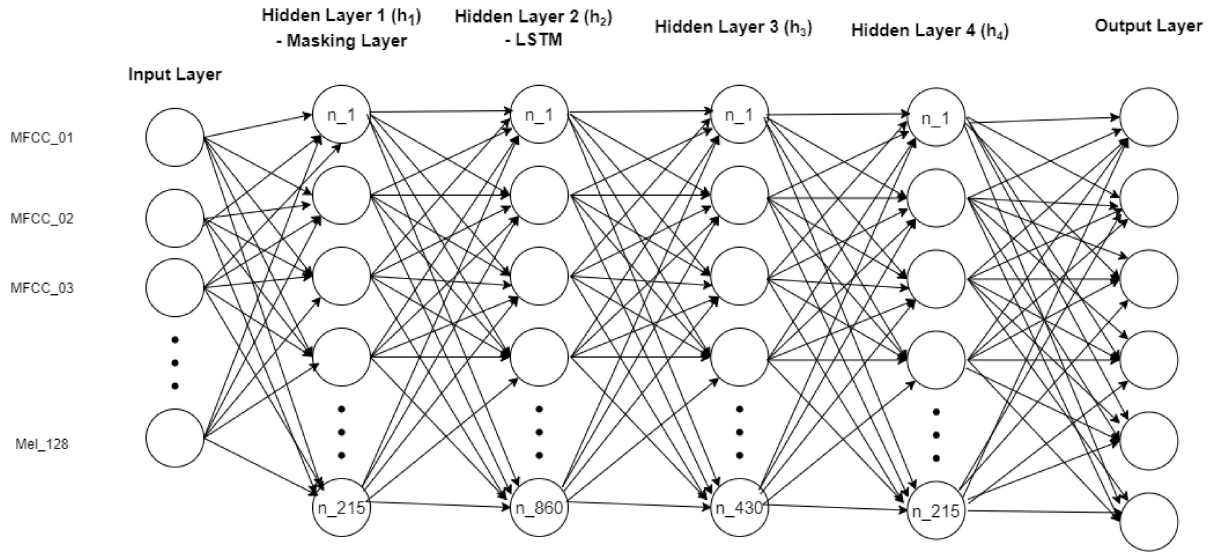



Fonte: Autor

São também usadas duas camadas escondidas após a camada LSTM, que tem como saída final a camada de saída com o resultado da classificação de emoções através da fala. Assim como a segunda camada, essas duas escondidas se baseiam na quantidade de características para gerar o número de unidades escondidas utilizadas.

Enquanto que na rede neural profunda realizou-se a implementação de três camadas densas escondidas que realizam a classificação das emoções através da fala. Como é visto nas figuras 5 e na figura 6, as unidades escondidas utilizadas nas camadas dependem também do

Figura 4 – Representação da Rede Neural Recorrente

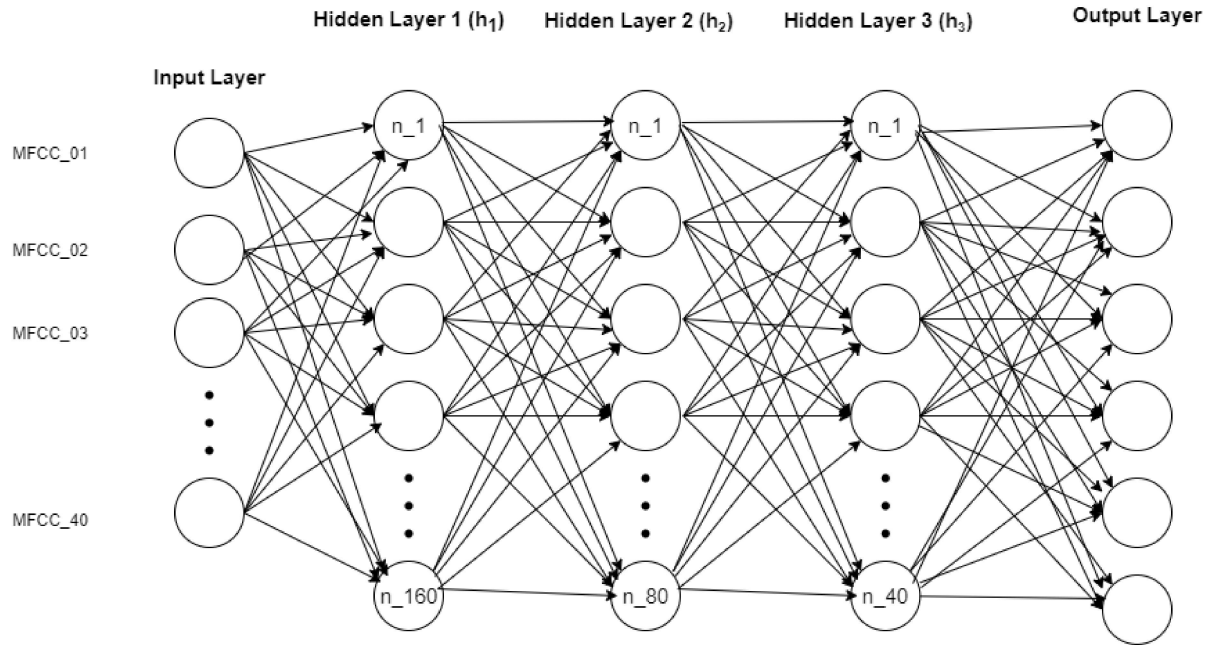



Fonte: Autor

número de características usadas com entrada para a rede neural. Para as camadas escondidas empregou-se a equação de ativação ReLU, cuja fórmula é $f(x) = max(0, x)$.

Para a camada de saída foi aplicada a função Softmax a fim de realizar a classificação das seis emoções presentes na base de dados escolhida.

Figura 5 – Representação da Rede Neural Profunda utilizando apenas MFCCs

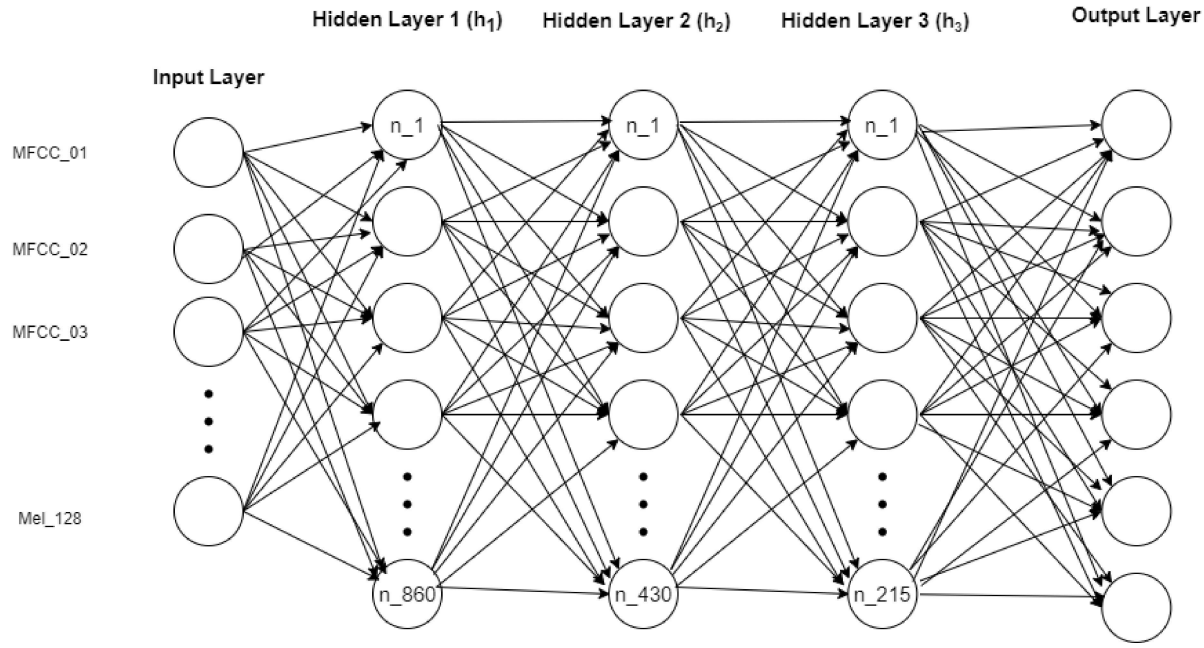



Fonte: Autor

Para realizar os experimentos, a base foi convertida para o formato WAV e separada em três grupos, sendo estes: Treinamento, Teste e Validação. Para os dados de Treinamento e Validação, é definido um valor para representar qual classe o áudio pertence, assim podendo ser utilizado no treinamento e na validação da rede neural.

Para realizar os experimentos, os áudios foram separados e para cada um foram extraídos os MFCCs, Delta-MFCC, Espectrogramas-Mel e Contraste Espectral. Os valores obtidos por estas características foram agrupados de maneira a serem usados para a classificação com as redes neurais. Para a análise dos resultados dos experimentos, realizou-se o cálculo da acurácia da classificação, e o tempo de cada etapa colheu-se durante o treinamento, extração de características e teste. Para os testes, empregou-se diversas configurações, com o intuito de verificar a melhor para a classificação utilizando as redes neurais.

Figura 6 – Representação da Rede Neural Profunda




Fonte: Autor

Após os testes preliminares com a rede neural realizando a classificação das emoções através da fala utilizando a implementação na linguagem de programação Python, foi possível obter resultados prévios de classificação do sistema. Para estes testes foram usadas duas opções, o uso do banco de filtros da frequência Mel dos áudios da base de dados e a transformada desses filtros em MFCCs. Durante os testes preliminares foram utilizados os MFCCs e foi obtida uma acurácia de 49%, os testes preliminares usaram uma entrada de tamanho igual a 40 para cada áudio de treinamento, validação e teste da rede neural. Os testes preliminares foram realizados

empregando uma arquitetura com 2 camadas escondidas com 40 e 20 unidades escondidas, respectivamente.

Para realizar os experimentos, foi definido a utilização da variação do *Batch Size* e a variação das épocas de treinamento das redes neurais. A variação escolhida para o tamanho do *Batch* e número de épocas foi realizada de maneira empírica, gerando o conjunto de valores $BS = [13,26,73,146]$. Para as épocas de treinamento foi empregado o conjunto $EP = [100, 500, 1000]$. Foram analisados outros valores para o tamanho do *Batch* e quantidade de épocas e apenas os valores com maiores destaques na classificação preliminar foram selecionados, de maneira a obter ambos conjuntos apresentados.

A base de dados originalmente possui vídeos com extensão AVI, que foram convertidos para áudios em 16 kHz para realizar a extração de características.

A base de dados não possui divisão previamente descrita para treinamento, validação e teste dos dados, consequentemente fez-se necessário separar os dados em 3 conjuntos com 80% dos dados para treinamento, 10% dos dados para validação e 10% dos dados para os testes.

Para compreender a qualidade da classificação das emoções, foram utilizadas algumas métricas de avaliação, sendo estas: Acurácia, Precisão (Precision), Sensibilidade (Recall) e F-Score. A acurácia é utilziada com o intuito de indicar a performance do modelo, apresentando quanto do modelo foi classificado de maneira correta, porém é possível que seja enganosa devido ao balanceamento da quantidade de frases utilizadas para cada classe ou emoção. A precisão indica nas classificações quais foram as classificadas corretamente, apontando a precisão do sistema em relação a falsos positivos. O recall apresenta a proporção daquelas classificações que estão corretas dentre as classes, utilizando o resultado positivo. O F-Score realiza a média harmônica entre a precisão e recall.

Nos capítulos seguintes, serão descritos os experimentos realizados, seus resultados e conclusões baseadas nos testes e resultados apresentados.

# 5 EXPERIMENTOS

Para realizar os experimentos, utilizou-se a base de dados eNTERFACE'05 Audio-Visual Emotion que conta com 42 locutores de 14 nacionalidades falando o idioma inglês, com acesso gratuito e que pertence à modalidade de indução.

Com esta base de dados realizou-se o treinamento da rede neural para seleção de características, visando aplicar os filtros da frequência-Mel, devido à sua relevância para a área de SER, como é verificado na tabela do Apêndice A.

Realizou-se testes com a base de dados, empregando-a para o treinamento, avaliação e teste. Para realizar este teste, a base foi convertida do formato original, vídeos com extensão AVI, para áudios em 16 kHz que são usados como padrão para a extração de características.

A base de dados não possui uma divisão prévia para treinamento, validação e teste dos dados, portanto fez-se necessário realizar a separação dos dados que foi realizada com 80% dos dados para treinamento, 10% dos dados para validação e 10% dos dados para os testes. Os áudios foram separados em seis classes: ira, desgosto, medo, felicidade, tristeza e surpresa; classes baseadas nas emoções apresentadas na base de dados utilizada. Ambas as arquiteturas apresentadas foram implementadas para realizar a classificação das emoções através da fala nos áudios da base de dados, tendo como característica empregada os MFCCs presentes nos áudios. Usou-se as variações das redes neurais para a verificação da acurácia do sistema, como o número de épocas empregadas para o treinamento e o tamanho do *Batch* utilizado com base nos conjuntos $EP = [100, 500, 1000]$ para o número de épocas e $BS = [13{,}26{,}73{,}146]$ para o tamanho do *Batch*.

Os experimentos, incluindo o treinamento, foram realizados em um ambiente virtual, usufruindo um Laptop que possui um processador Intel i5 e 8Gb de memória RAM. Para o treinamento foi utilizado, apenas, o poder de processamento do processador.

Foram realizados também experimentos que empregam mais características como o Delta-MFCC, o espectrograma-Mel e o contraste espectral, apresentados previamente na seção de conceitos. Os testes foram realizados utilizando-se as duas arquiteturas previamente explicadas: a rede neural recorrente LSTM profunda e a rede neural profunda.

Por fim, nos testes com a base de dados, verificou-se a acurácia dos sistemas finais de reconhecimento de emoções através da fala como também a velocidade de execução e resposta do sistema para um futuro estudo de uso de reconhecimento de emoções através da fala dentro do contexto de interação Humano-Robô adaptativa.

# 6 RESULTADOS

Após os testes preliminares, foram realizados os experimentos descritos no capítulo anterior fazendo uso das características dos MFCCs como entrada para as redes neurais propostas.

Utilizando a arquitetura e configuração proposta na figura 5, foi realizado o teste de classificação de emoções e a partir desses testes a tabela 3 foi criada. Com base nessa tabela é possível verificar que as maiores acurácias desse sistema ocorreram ao se empregar o *Batch* com tamanho igual a 13 e 100 ou 500 épocas, com 52% de acurácia e tempo de treinamento igual a 9,74 segundos e 41,64 segundos respectivamente, seguido pela configuração de 73 de *Batch* com 500 épocas, com 50% de acurácia e tempo de treino de 15 segundos. É possível observar as maiores acurácias na classificação das emoções observando o gráfico presente na figura 7, no qual são apresentados as acurácias conforme a configuração empregada, baseado na Tabela 3.

Tabela 3 – Dados de reposta da Rede Neural Profunda utilizando apenas MFCCs

| Épocas | *Batch* | Tempo de Extração de características (s) | Tempo de Treino (s) | Tempo de Teste (s) | Acurácia |
|---|---|---|---|---|---|
| **100** | **13** | **0,04767** | **9,74004** | **0,00064** | **52%** |
| **500** | **13** | **0,04748** | **41,64035** | **0,00068** | **52%** |
| 1000 | 13 | 0,04935 | 80,98221 | 0,00064 | 49% |
| 100 | 26 | 0,04833 | 6,25143 | 0,00056 | 46% |
| 500 | 26 | 0,04732 | 25,86067 | 0,00067 | 40% |
| 1000 | 26 | 0,04928 | 50,47201 | 0,00056 | 42% |
| 100 | 73 | 0,05068 | 4,57003 | 0,00065 | 45% |
| **500** | **73** | **0,04916** | **15,77564** | **0,00068** | **50%** |
| 1000 | 73 | 0,04841 | 29,86641 | 0,00063 | 48% |
| 100 | 146 | 0,05069 | 3,82995 | 0,00057 | 47% |
| 500 | 146 | 0,04927 | 12,62355 | 0,00077 | 47% |
| 1000 | 146 | 0,04752 | 23,69477 | 0,00055 | 50% |

Fonte: Autor

Testou-se também, utilizando-se a arquitetura e configuração proposta na figura 3, no qual é apresentada a configuração de rede neural recorrente que usa uma camada de LSTM. Foi realizado o teste de classificação de emoções e assim como foi feito para a rede neural profunda, a tabela 4 foi criada. Com essa tabela é possível verificar que as maiores acurácia do sistema empregando a RNN foi ao se utilizar o *Batch* com tamanho igual a 26 e 1000, com 37,01% de acurácia e tempo de treinamento igual a 12064 segundos. Observou-se que a configuração de

13 de *Batch* com 1000 épocas também alcançou a mesma acurácia, com 23571,15 segundos de tempo de treinamento.

Tabela 4 – Dados de reposta da Rede Neural Recorrente utilizando apenas MFCCs

| Épocas | *Batch* | Tempo de Extração de características (s) | Tempo de Treino (s) | Tempo de Teste (s) | Acurácia |
|---|---|---|---|---|---|
| 100 | 13 | 0,04794 | 2269,43299 | 0,00372 | 30% |
| 500 | 13 | 0,04769 | 11597,76966 | 0,00422 | 28% |
| **1000** | **13** | **0,05851** | **23571,15133** | **0,00389** | **37%** |
| 100 | 26 | 0,04856 | 1237,90399 | 0,00347 | 22% |
| 500 | 26 | 0,04846 | 6165,07455 | 0,00351 | 30% |
| **1000** | **26** | **0,04884** | **12063,99766** | **0,00383** | **37%** |
| 100 | 73 | 0,04793 | 571,23469 | 0,00371 | 22% |
| 500 | 73 | 0,04860 | 2869,44939 | 0,00347 | 34% |
| 1000 | 73 | 0,04905 | 5722,90928 | 0,00370 | 35% |
| 100 | 146 | 0,04922 | 410,64077 | 0,00368 | 20% |
| 500 | 146 | 0,04875 | 2042,41335 | 0,00349 | 28% |
| 1000 | 146 | 0,04817 | 4101,04633 | 0,00371 | 34% |

Fonte: Autor

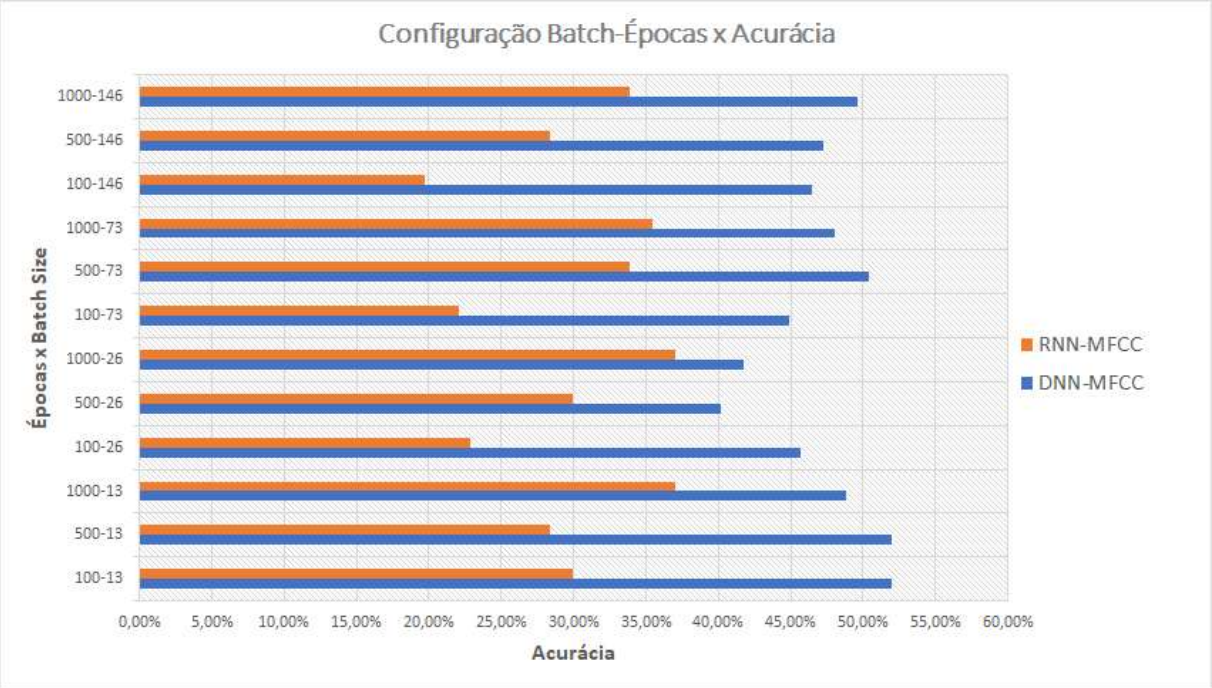
Para uma melhor visualização da comparação entre a acurácia da DNN e a LSTM, utilizando a mesma configuração de épocas e tamanho de *Batch*, foi gerado o gráfico presente na figura 7, que apresenta a acurácia de cada modelo conforme a configuração.

Para compreender como o sistema está se portando com relação à classificação de emoções através da fala e suas classes, quando empregada apenas uma característica, foi gerado uma matriz de confusão para os dois modelos apresentados, uma para a DNN e uma para a RNN. Para cada configuração da DNN e da RNN, foi gerada uma matriz de confusão e a partir das melhores acurácias apresentadas anteriormente, foram selecionadas duas configurações de cada modelo para fazer a comparação do F-Score, ou F-Measure, apresentando então a maior acurácia de cada modelo.

Após a criação da matriz de confusão para as configurações apresentadas da rede neural profunda, obteve-se que a matriz que possuía os maiores F-Score, no caso a matriz que representa a configuração com *Batch* igual a 13 e 500 épocas para treinamento, que está representada na tabela 5.

A partir da matriz de confusão 5, foi realizado o cálculo do F-Score, utilizando alguns métodos como média. Foram obtidos os seguintes valores ao realizar o cálculo do F-Score:

Figura 7 – Gráfico de Comparação da Acurácia da DNN e RNN utilizando apenas MFCCs




Fonte: Autor

Tabela 5 – Matriz de Confusão Para Rede Neural Profunda utilizando apenas MFCCs

| Esperado x Resposta | Ira | Desgosto | Medo | Felicidade | Tristeza | Surpresa |
|---|---|---|---|---|---|---|
| Ira | 12 | 2 | 0 | 2 | 0 | 2 |
| Desgosto | 3 | 9 | 1 | 4 | 1 | 3 |
| Medo | 4 | 3 | 7 | 3 | 4 | 0 |
| Felicidade | 4 | 2 | 2 | 5 | 1 | 3 |
| Tristeza | 0 | 4 | 0 | 0 | 16 | 6 |
| Surpresa | 0 | 0 | 2 | 2 | 3 | 17 |

Fonte: Autor

0,523; 0,520; 0,514. Esses valores são, respectivamente, referentes ao cálculo da média com peso, micro e macro. Além dos valores para o F-Score, foram calculados o Precision e o Recall da matriz apresentada, obtendo o resultado de 0,521 para Precision e 0,520 para Recall. Além destes valores, realizou-se o cálculo de cada classe, obtendo-se um vetor com os valores 0,588, 0,465, 0,512, 0,343, 0,605, 0,571 para as classes Ira, Desgosto, Medo, Felicidade, Tristeza e Surpresa, respectivamente.

Para a RNN verificou-se que entre as duas configurações apresentadas, a que possuía um melhor F-Score era a configuração com 26 de *Batch* e 1000 épocas de treinamento. Para realizar o cálculo do F-Score, gerou-se a matriz de confusão apresentada na tabela 6.

Com base na matriz de confusão gerada para a RNN utilizando apenas os MFCCs, realizou-se o cálculo do F-Score e com isso foram obtidos os valores do cálculo empregando-se a média baseada em pesos, Micro e Macro. Para a média com peso obteve-se 0,358, para a Micro 0,370 e para a média Macro 0.352. Quando analisada a acurácia das classes individualmente, verificou-se que a Tristeza obteve a maior acurácia, com 0,554.

Tabela 6 – Matriz de Confusão Para Rede Neural Recorrente utilizando apenas MFCCs

| Esperado x Resposta | Ira | Desgosto | Medo | Felicidade | Tristeza | Surpresa |
|---|---|---|---|---|---|---|
| Ira | 7 | 1 | 3 | 4 | 1 | 2 |
| Desgosto | 0 | 4 | 6 | 1 | 4 | 6 |
| Medo | 3 | 1 | 4 | 0 | 9 | 4 |
| Felicidade | 1 | 1 | 5 | 5 | 1 | 4 |
| Tristeza | 2 | 0 | 2 | 0 | 18 | 4 |
| Surpresa | 3 | 2 | 3 | 1 | 6 | 9 |

Fonte: Autor

Tabela 7 – Matriz de Confusão Para Rede Neural Profunda

| Esperado x Resposta | Ira | Desgosto | Medo | Felicidade | Tristeza | Surpresa |
|---|---|---|---|---|---|---|
| Ira | 16 | 0 | 0 | 1 | 0 | 1 |
| Desgosto | 4 | 6 | 3 | 3 | 1 | 4 |
| Medo | 2 | 3 | 7 | 4 | 2 | 3 |
| Felicidade | 2 | 3 | 0 | 10 | 0 | 2 |
| Tristeza | 0 | 2 | 7 | 2 | 12 | 3 |
| Surpresa | 1 | 1 | 2 | 3 | 3 | 14 |

Fonte: Autor

Quando realizado o cálculo de Precision e Recall, obteve-se, respectivamente, os valores 0,378 e 0,370.

Ao realizar os testes adicionando outras características, verificou-se que a maior acurácia da RNN superou a configuração equivalente da DNN, em questão de acerto. Porém, em todos os outros casos, obteve-se uma acurácia maior quando utilizada a rede neural profunda para a classificação.

Nos testes com outras características obteve-se uma acurácia de 51% para a configuração de *Batch* igual a 2 e 500 épocas de treinamento para a DNN. O caso em que a RNN, obteve uma acurácia maior que a DNN foi com o tamanho do *Batch* igual a 73 e usando 100 épocas para treinamento da rede, no qual a RNN obteve uma acurácia de 44%, enquanto a DNN obteve um resultado de 43% de acerto empregando a mesma configuração.

Utilizando-se a configuração com maior acurácia da DNN, foi possível gerar a matriz de confusão ilustrada na tabela 7. A partir da matriz de confusão gerada, calculou-se o F-Score, que apresentou 0,501 para o peso, 0,512 para o Micro e 0,504 para o cálculo Macro. A partir da matriz de confusão foi possível calcular o valor de Precision e Recall, observando-se os valores 0,510 e 0,512. Quando verificado a acurácia individual das classes observou-se uma acurácia de 0,744 para Ira, 0,333 para Desgosto, 0,35 para Medo, 0,5 para Felicidade, 0,545 para Tristeza e 0,549 para Surpresa.

Tabela 8 – Matriz de Confusão Para Rede Neural Recorrente

| Esperado x Resposta | Ira | Desgosto | Medo | Felicidade | Tristeza | Surpresa |
|---|---|---|---|---|---|---|
| Ira | 7 | 0 | 4 | 4 | 1 | 2 |
| Desgosto | 1 | 7 | 4 | 4 | 0 | 5 |
| Medo | 1 | 1 | 14 | 0 | 1 | 4 |
| Felicidade | 2 | 2 | 2 | 6 | 1 | 4 |
| Tristeza | 1 | 3 | 5 | 1 | 11 | 5 |
| Surpresa | 2 | 2 | 4 | 4 | 1 | 11 |

Fonte: Autor

Para a configuração da RNN, gerou-se a matriz de confusão apresentada na tabela 8. A partir dessa matriz de confusão possibilitou-se a realização do cálculo do F-Score, obtendo-se como resultado 0,442 para o cálculo com peso, 0,441 para o Micro e 0,436 para o Macro. Quando verificada a acurácia de cada classe foi possível observar 0,438 para a Ira, 0,389 para Desgosto, 0,519 para Medo, 0,333 para Felicidade, 0,537 para Tristeza e 0,4 para Surpresa. Além do cálculo do F-Score, realizou-se também o cálculo de Precision que apresentou o resultado de 0,478 e Recall que apresentou um valor igual a 0,441.

Como é possível observar pela tabela 9, a emoção Tristeza e Ira, foram as que obtiveram uma melhor classificação e as melhores configurações do sistema foram as configurações da DNN utilizando apenas os valores dos MFCCs e usufruindo das outras caraterísticas indicadas

Tabela 9 – Tabela de Comparação entre resultados positivos das configurações e as emoções

| Emoção x Configuração | DNN-MFCC | RNN-MFCC | DNN | RNN |
|---|---|---|---|---|
| Ira | 0,585 | 0,412 | 0,744 | 0,438 |
| Desgosto | 0,439 | 0,267 | 0,333 | 0,389 |
| Medo | 0,424 | 0,182 | 0,350 | 0,519 |
| Felicidade | 0,303 | 0,357 | 0,500 | 0,333 |
| Tristeza | 0,627 | 0,554 | 0,545 | 0,537 |
| Surpresa | 0,618 | 0,340 | 0,549 | 0,400 |

Fonte: Autor

anteriormente. Tendo em vista o objetivo de comparar as classificações da DNN e da LSTM usando apenas os valores dos MFCCs e o conjunto dos valores dos MFCCs somado às características definidas previamente, foi possível observar que a DNN em ambos casos obteve um resultado melhor. é possível observar a diferença na acurácia quando analisado o gráfico presente na figura 7, no qual em todos os cenários a DNN empregando apenas os MFCCs obteve uma taxa de acerto maior que a LSTM ao usufruir de apenas os MFCCs.

É possível visualizar esse resultado nas médias do F-Score de cada emoção classificada, obtendo 0,5 para a DNN utilizando apenas os valores dos MFCCs e 0,504 quando usado o conjunto com mais características. Com base nos resultados, é inferido que o uso de mais características pode influenciar em uma melhora na classificação das emoções, como é visto no uso da LSTM, tendo um aumento de 8,4% na taxa de acerto médio. Porém no uso da DNN, foi observado um aumento de 0,412% na taxa de acerto quando utilizadas mais características somadas aos valores dos MFCCs.

# 7 CONCLUSÕES

Neste trabalho objetivou-se a implementação e análise de um sistema de reconhecimento de emoções comparando-a com a utilização de uma DNN e uma rede neural recorrente LSTM, que são classificadores empregados na área de reconhecimento e classificação de emoções por meio da fala. Para possibilitar a realização deste trabalho, pesquisou-se o que é um sistema de reconhecimento de emoções e seus componentes. Escolheu-se a aplicação de técnicas tradicionais de extração de características e deixar a cargo da rede neural a escolha destas, para a melhor representação do sentimento através da fala. Para as características, foram escolhidos 40 valores dos MFCCs, 40 valores dos Delta-MFCCs, os Espectrogramas-Mel e o Contraste Espectral.

Escolheu-se realizar a contribuição na pesquisa de comparação na qualidade de reconhecimento de emoções quando são empregados dois classificadores baseados em redes neurais, sendo estes uma RNN e uma DNN. Além da contribuição verificada no uso dos dois classificadores que utilizaram como base de aprendizado a base de dados eNTERFACE'05 Audio-Visual Emotion. Estudou-se também a variação na qualidade de reconhecimento de emoções perante a variação de características, realizando-se a comparação do uso de diversas características com o uso de apenas 40 valores dos MFCCs.

Com base nos testes realizados, verificou-se que a classificação, usando apenas as características da frequência-Mel, obteve 52% de acerto. Após os testes verificou-se ainda que durante o treinamento das redes neurais utilizando apenas os MFCCs, a DNN obteve a maior acurácia em todas as configurações testadas, obtendo 0,523 para o F-Score, 0,521 em Precision e 0,520 em Recall. Porém, ao se empregar outras características como o Delta-MFCC, o espectrograma-Mel e o contraste espectral somado aos MFCCs, observou-se que em um caso a RNN obteve uma acurácia maior quando comparado à mesma configuração da DNN, alcançando um resultado de 44% de acerto. Para esta configuração, observou-se um resultado de 0,442 para o cálculo do F-Score, 0,478 para Precision e 0,441 para Recall.

Durante a pesquisa dos trabalhos relacionados, foi observado que Trigeorgis et al. (2016) utilizou CNN com LSTM como classificador para realizar o reconhecimento de emoções através da fala, utilizando o sinal da fala como características para reconhecimento. Ao classificar Valência e Excitação, foi possível verificar uma acurácia de 68,6% para excitação, e utilizando o conjunto de validação foi obtido 74,1% para a base utilizada, a base de dados RECOLA.

Já no trabalho de Haytham M Fayek, Lech e Cavedon (2015), foi proposto o uso de DNN para a classificação em tempo real da fala utilizando como base de dados a eNTERFACE'05.

Nesse trabalho foi obtido uma classificação com 60,53% de acerto utilizando quadros de 1 segundos de duração.

Os resultados dos testes possibilitaram a observação de um reconhecimento com maior acurácia por parte da DNN quando comparada à acurácia da LSTM, com exceção a um caso no qual a precisão da LSTM fora superior à precisão da DNN. Tais resultados indicam que, nesta configuração, para o caso da RNN e da DNN empregadas, o crescimento do número de características não aumenta a qualidade da classificação de emoções quando se utiliza a base de dados eNTERFACE'05 Audio-Visual Emotion, visto que a maior acurácia fora apresentada no uso de, apenas, 40 valores dos MFCCs.

Todos os códigos-fonte programados para realizar este trabalho estão disponíveis em um repositório de livre acesso (MEYER, 2021). Neste repositório estão disponíveis os códigos-fonte, as divisões usadas e também o link para acesso à base utilizada. A correção de dados realizada neste trabalho, alterando o nome incorreto de arquivos de vídeo de um usuário presente na base também estão documentadas.

A área de pesquisa que alavancou o interesse para realizar este trabalho voltado ao reconhecimento de emoções através da fala foi a Interação Humano-Robô voltada a robôs de serviço doméstico, uma vez que robôs de serviço podem usufruir da informação emocional do usuário para realizar diferentes cenários de interação, como analisar se o usuário está feliz e questionar como foi o dia, ou mesmo verificar se o usuário está triste e perguntar se há algo que o robô pode fazer, realizando essa adaptação na interação com o usuário, tornando-a mais próxima de uma interação com humanos. Para trabalhos futuros pretende-se realizar o estudo de características que devem ser aplicadas a um sistema robótico, visando a classificação de emoções, por meio da fala e visão. Além do estudo de outras características é desejado realizar testes em cenários reais de interação com o robô físico e estudar outras possibilidades de modelos para implantação em um robô doméstico e serviços que fazem o uso da fala para interação, como computadores de bordo e canais de atendimento que utilizam da fala para interação entre usuário e sistema. É desejado para trabalhos futuros realizar testes com bases de dados como a *Survey Audio-Visual Expressed Emotion* (SAVEE), *Toronto Emotional Speech Database* (TESS), *Electromagnetic Articulography Database* (EMA), eNTERFACE'05 *Audio-Visual Emotion Database* e *AFEW Database* além de outras não citadas.

# REFERÊNCIAS

ACADEMY, Data Science. **Deep Learning Book**. 2021. Disponível em: <https://www.deeplearningbook.com.br/>.

AKÇAY, Mehmet Berkehan; OĞUZ, Kaya. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. **Speech Communication**, Elsevier, v. 116, 2020.

AQUINO JR, Plinio Thomaz et al. HERA: Home Environment Robot Assistant, 2019. Disponível em: <https://www.researchgate.net/publication/333931333_HERA_Home_Environment_Robot_Assistant>.

BHAVAN, Anjali et al. Bagged support vector machines for emotion recognition from speech. **Knowledge-Based Systems**, p. 104886, ago. 2019. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0950705119303533>.

CAO, Houwei et al. CREMA-D: Crowd-sourced emotional multimodal actors dataset. **IEEE transactions on affective computing**, IEEE, v. 5, n. 4, p. 377–390, 2014.

CHAUDHARY, ANKUSH et al. Speech emotion recognition. **J Emerg Technol Innov Res**, v. 2, n. 4, p. 1169–1171, 2015.

CHEN, Lijiang et al. Speech emotion recognition: Features and classification models. **Digital Signal Processing: A Review Journal**, 2012.

CHEN, Luefeng et al. Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. **Information Sciences**, Elsevier, v. 509, p. 150–163, 2020.

CK, Yogesh et al. Bispectral features and mean shift clustering for stress and emotion recognition from natural speech. **Computers and Electrical Engineering**, 2017.

EKMAN, Paul. Are There Basic Emotions? **Psychological Review**, v. 99, n. 3, p. 550–553, 1992.

EL AYADI, Moataz; KAMEL, Mohamed S; KARRAY, Fakhri. Survey on speech emotion recognition: Features, classification schemes, and databases. **Pattern Recognition**, Elsevier, v. 44, n. 3, p. 572–587, 2011.

FAYEK, Haytham M.; LECH, Margaret; CAVEDON, Lawrence. Evaluating deep learning architectures for Speech Emotion Recognition. **Neural Networks**, 2017.

_____. Towards real-time speech emotion recognition using deep neural networks. In: IEEE. 2015 9th international conference on signal processing and communication systems (ICSPCS). [S.l.: s.n.], 2015. p. 1–5.

FOX, Nathan A. If It's Not Left, It's Right: Electroencephalograph Asymmetry and the Development of Emotion. **American Psychologist**, 1991.

HACINE-GHARBI, Abdenour; RAVIER, Philippe. On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion

recognition. **Journal of King Saud University - Computer and Information Sciences**, jul. 2019. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1319157819304537>.

HAN, Kun; YU, Dong; TASHEV, Ivan. Speech emotion recognition using deep neural network and extreme learning machine. In: FIFTEENTH annual conference of the international speech communication association. [S.l.: s.n.], 2014.

HUAHU, X.; JUE, G.; JIAN, Y. Application of Speech Emotion Recognition in Intelligent Household Robot. In: 2010 International Conference on Artificial Intelligence and Computational Intelligence. [S.l.: s.n.], 2010. v. 1, p. 537–541.

IROBOT. **Roomba série s**. [S.l.: s.n.]. Disponível em: <https://www.irobot.com.br/roomba/s-series>. Acesso em: 19 set. 2020.

JANG, Kwang-Dong; KWON, Oh-Wook. Speech emotion recognition for affective human-robot interaction. **SPECOM'2006**, p. 419–422, 2006.

KERAS. **Keras: the python deep learning API**. 2021. Disponível em: <https://keras.io/>.

KERKENI, Leila et al. Speech Emotion Recognition: Methods and Cases Study. In: p. 175–182.

KWON, Dong-Soo et al. Emotion interaction system for a service robot. In: IEEE. RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication. [S.l.: s.n.], 2007. p. 351–356.

LAKOMKIN, Egor et al. On the robustness of speech emotion recognition for human-robot interaction with deep neural networks. In: IEEE. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). [S.l.: s.n.], 2018. p. 854–860.

LI, Longfei et al. Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. In: IEEE. 2013 Humaine association conference on affective computing and intelligent interaction. [S.l.: s.n.], 2013. p. 312–317.

LI, Xingfeng; AKAGI, Masato. Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model. **Speech Communication**, 2019.

LIU, Weibo et al. A survey of deep neural network architectures and their applications. **Neurocomputing**, Elsevier, v. 234, p. 11–26, 2017.

LIVINGSTONE, Steven; RUSSO, Frank. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. **PLOS ONE**, v. 13, e0196391, mai. 2018.

LYONS, James. **Mel Frequency Cepstral Coefficient (MFCC) tutorial**. 2012. Disponível em: <http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/>.

MAO, Qirong et al. Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. **Speech Communication**, 2017.

MARTIN, O. et al. The eNTERFACE' 05 Audio-Visual Emotion Database. In: 22ND International Conference on Data Engineering Workshops (ICDEW'06). [S.l.: s.n.], 2006. p. 8–8.

MCFEE, Brian et al. librosa: Audio and music signal analysis in python. In: PROCEEDINGS of the 14th python in science conference. [S.l.: s.n.], 2015. v. 8.

MEYER, Thiago SB; JUNIOR, Plinio Thomaz Aquino. Sound Source Localization and Tracking for the@ Home Service Robot. In: II BRAHUR and III Brazilian Workshop on Service Robotics. [S.l.: s.n.], 2019. p. 59–64.

MEYER, Thiago Spilborghs Bueno. **SER - python3**. Mai. 2021. Disponível em: <https://gitlab.com/Thiago_Spilborghs/ser-python3>.

MEYER, Thiago Spilborghs Bueno; AQUINO-JUNIOR, Plinio Thomaz. Analysis of Sound Source Localization and Tracking for the@ Home Service Robot in Multiple Distances, 2020.

ÖZSEVEN, Turgut. A novel feature selection method for speech emotion recognition. **Applied Acoustics**, 2019.

PARK, Jeong-Sik; KIM, Ji-Hwan; OH, Yung-Hwan. Feature vector classification based speech emotion recognition for service robots. **IEEE Transactions on Consumer Electronics**, IEEE, v. 55, n. 3, p. 1590–1596, 2009.

PEREZ-GASPAR, Luis-Alberto; CABALLERO-MORALES, Santiago-Omar; TRUJILLO-ROMERO, Felipe. Multimodal emotion recognition with evolutionary computation for human-robot interaction. **Expert Systems with Applications**, Elsevier, v. 66, p. 42–61, 2016.

RAMAKRISHNAN, Srinivasan; EL EMARY, Ibrahiem MM. Speech emotion recognition approaches in human computer interaction. **Telecommunication Systems**, Springer, v. 52, n. 3, p. 1467–1478, 2013.

RÁZURI, Javier G et al. Speech emotion recognition in emotional feedbackfor human-robot interaction. **International Journal of Advanced Research in Artificial Intelligence (IJARAI)**, v. 4, n. 2, p. 20–27, 2015.

ROBOCUP-FEDERATION. **RoboCup@Home**. 2020. Disponível em: <https://www.robocup.org/domains/3>.

SOFTBANK-ROBOTICS. **Pepper the humanoid and programmable robot**. [S.l.: s.n.]. Disponível em: <https://www.softbankrobotics.com/emea/en/pepper>. Acesso em: 19 set. 2020.

SONY-CORPORATION. **aibo**. [S.l.: s.n.]. Disponível em: <https://us.aibo.com/>. Acesso em: 19 set. 2020.

TORRES-BOZA, Diana et al. Hierarchical sparse coding framework for speech emotion recognition. **Speech Communication**, 2018.

TRIGEORGIS, G. et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2016. p. 5200–5204.

TZIRAKIS, Panagiotis; ZHANG, Jiehao; SCHULLER, Bjorn W. End-to-end speech emotion recognition using deep neural networks. In: IEEE. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2018. p. 5089–5093.

WITTEN, Ian H et al. Chapter 10—Deep learning. **Data mining (fourth edition). London: Morgan Kaufmann**, p. 417–66, 2017.

WU, Siqing; FALK, Tiago H.; CHAN, Wai Yip. Automatic speech emotion recognition using modulation spectral features. **Speech Communication**, 2011.

ZHANG, Shiqing; ZHAO, Xiaoming; LEI, Bicheng. Speech emotion recognition using an enhanced kernel isomap for human-robot interaction. **International Journal of Advanced Robotic Systems**, SAGE Publications Sage UK: London, England, v. 10, n. 2, p. 114, 2013.

ZHAO, Jianfeng; MAO, Xia; CHEN, Lijiang. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. **Biomedical Signal Processing and Control**, 2019.

**APÊNDICE A –** TABELA DE TRABALHOS RELACIONADOS

| **Autores** | **Título do Trabalho** | **Ano** |
|---|---|---|
| Jang e Kwon | Speech emotion recognition for affective human-robot interaction | 2006 |
| Kwon et al. | Emotion interaction system for a service robot | 2007 |
| Park, Kim e Oh | Feature vector classification based speech emotion recognition for service robots | 2009 |
| Huahu, Jue e Jian | Application of Speech Emotion Recognition in Intelligent Household Robot | 2010 |
| El ayadi, Kamel e Karray | Survey on speech emotion recognition: Features, classification schemes, and databases | 2011 |
| Wu, Falk e Chan | Automatic speech emotion recognition using modulation spectral features | 2011 |
| Chen et al. | Speech emotion recognition: Features and classification models | 2012 |
| Li et al. | Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition | 2013 |
| Zhang, Zhao e Lei | Speech emotion recognition using an enhanced kernel isomap for human-robot interaction | 2013 |
| Han, Yu e Tashev | Speech emotion recognition using deep neural network and extreme learning machine | 2014 |
| Fayek, Lech e Cavedon | Towards real-time speech emotion recognition using deep neural networks | 2015 |
| Rázuri et al. | Speech emotion recognition in emotional feedbackfor human-robot interaction | 2015 |

| | | |
|---|---|---|
| Perez-Gaspar, Caballero-Morales e Trujillo-Romero | Multimodal emotion recognition with evolutionary computation for human-robot interaction | 2016 |
| Trigeorgis et al. | Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network | 2016 |
| CK et al. | Bispectral features and mean shift clustering for stress and emotion recognition from natural speech | 2017 |
| Fayek, Lech e Cavedon | Evaluating deep learning architectures for Speech Emotion Recognition | 2017 |
| Mao et al. | Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition | 2017 |
| Lakomkin et al. | On the robustness of speech emotion recognition for human-robot interaction with deep neural networks | 2018 |
| Torres-Boza et al. | Hierarchical sparse coding framework for speech emotion recognition | 2018 |
| Tzirakis, Zhang e Schuller | End-to-end speech emotion recognition using deep neural networks | 2018 |
| Bhavan et al. | Bagged support vector machines for emotion recognition from speech | 2019 |
| Hacine-Gharbi e Ravier | On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion recognition | 2019 |
| Li e Akagi | Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model | 2019 |
| Ozseven | A novel feature selection method for speech emotion recognition | 2019 |

| | | |
|---|---|---|
| Zhao, Mao e Chen | Speech emotion recognition using deep 1D {&} 2D CNN LSTM networks | 2019 |
| Chen et al. | Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction | 2020 |

| **Autores** | **...** | **Objetivo do Trabalho** |
|---|---|---|
| Jang e Kwon | ... | Avaliar o desempenho do método de reconhecimento de emoções para interação humano-robô afetiva |
| Kwon et al. | ... | Criar um framework para interação emocional |
| Park, Kim e Oh | ... | Criar uma classificação eficiente para o vetor de características na áera de SER em robôs de serviço |
| Huahu, Jue e Jian | ... | Realizar o reconhecimento de 5 emoções através da fala |
| El ayadi, Kamel e Karray | ... | Estudar e analisar o estado da arte da área de SER |
| Wu, Falk e Chan | ... | Analisar o uso de Características modulares espectrais (MSFs) para o reconhecimento automático de informações afetivas através da fala |
| Chen et al. | ... | Criar um modelo de reconhecimento de emoções atrvés da fala de três níveis |
| Li et al. | ... | Estudar e analisar o uso de DNN-HMMs para o reconhecimento de emoções atrvés da fala com RBM como base de pré-treinamento não-supervisionado e com pré-trinamento discriminativo |

| | | |
|---|---|---|
| Zhang, Zhao e Lei | ... | Utilizar e testar o método EKIsomap para a redução de dimensão das características utilizadas visando o reconhecimento de emoções através da fala aplicado na interação humano-robô |
| Han, Yu e Tashev | ... | Utilizar DNNs para a extração de características de alto nível de dados puros para veririfcar a efetividade duranta o reconhecimento de emoções através da fala |
| Fayek, Lech e Cavedon | ... | Criar um sistema de SER em tempo real de Aprendizado profundo ponta-a-ponta |
| Rázuri et al. | ... | Avaliar 6 classificadores para reconhecer emoções através de características não verbais |
| Perez-Gaspar, Caballero-Morales e Trujillo-Romero | ... | Explorar as implicações de utilizar uma based de dados padrão para a avaliação de técnicas de reconheciemento de emoções, realizar a extensão na otimização evolucionária de ANNs e HMMs para o desenvolvimento de um sistema multimodal para reconhecimento de emoções, definir uma linha para o desenvolvimento de uma base de dados emocional de falas e expressões faciais, definir regras para transcrição fonética para fala mexicana e avaliar a adequação de um sistema mutimodal para o contexto de diálogo entre um robô humanoide e um usuário humano. |

| | | |
|---|---|---|
| Trigeorgis et al. | ... | Solucionar o problema de extração de características emocionais relevantes com consciência de contexto, por meio da combinação de CNN e LSTM para aprendizado automático da melhor representação do sinal de fala diretamente da representação pura de tempo. |
| CK et al. | ... | Aplicar mean shift clustering nas BSFs para aumentar a habilidade discriminatória das características extraídas. |
| Fayek, Lech e Cavedon | ... | Explorar Arquiteturas de redes neurais recorrentes e feed-forward e suas variações empiricamente em sistemas de SER |
| Mao et al. | ... | Aplicar um método de Adaptação de domínio chamado de EDFLM para o SER. |
| Lakomkin et al. | ... | Avaliar a robustez do estado-da-arte de modelos neurais de reconhecimento de emoções acústicas no cenário de Interação Humano-Robô |
| Torres-Boza et al. | ... | Propor um esquema de HSC para representação de características do áudio para reconhecimento de emoções através da fala |
| Tzirakis, Zhang e Schuller | ... | Apresentar um novo modelo para o reconhecimento de emoções através da fala contínuo |
| Bhavan et al. | ... | Realizar o reconhecimento de emoções em 3 bases dde dados |

| | | |
|---|---|---|
| Hacine-Gharbi e Ravier | ... | Utilizar um critério baseado em informação mútua para estimar o número mínimo de características que exemplificam a variedade de indíces de classes a serem reconhecidas em um reconhecimento de emoções através da fala |
| Li e Akagi | ... | Apresentar um esquema para reconhecimento de emoções através da fala multilinguístico |
| Ozseven | ... | Propor um método estatístico de seleção de caraterísticas baseado na alteração de emoções em características acústicas |
| Zhao, Mao e Chen | ... | Utilizar características profundas de emoção para aprendizado e reconhcimento de emoções |
| Chen et al. | ... | Utilizar o algoritmo de Floresta Aleatória múltipla com duas camadas fuzzy para reconhecimento de emoções através da fala |

| **Autores** | ... | **Características Utilizadas** |
|---|---|---|
| Jang e Kwon | ... | Tom, Energia, Formato, Tempo, Duração, Tremor, Brilho, MFCC, LPC, Energia Teager |
| Kwon et al. | ... | RSS, SFM |
| Park, Kim e Oh | ... | Tom. Log de energia, Razão de cruzamento em zero e MFCC de 12 dimensões |
| Huahu, Jue e Jian | ... | - |
| El ayadi, Kamel e Karray | ... | - |

| | | |
|---|---|---|
| Wu, Falk e Chan | ... | MSF (Média de energia, Planicidade espectral, Centroide Espectral, Centroide de Modulação Espectral, Coeficiente de regressão linear e Raiz quadrada do erro médio), LPCC, MFCC, PLP, Prosodic Features. |
| Chen et al. | ... | Energia, Razão de cruzamento em zero, Energia x ZCR, Tom, Primeiro 3 formantes, centroides espectrais, frequência de corte espectral, densidade de correlação, dimensão fractal e Energia de 5 bandas de Frequencia-Mel. Primeira e segunda derivada das características, Máximo, mínimo, média, desvio padrão, distorção e curtose. |
| Li et al. | ... | MFCC |
| Zhang, Zhao e Lei | ... | Tom, Intensidade, duração, três primeiros formantes, distribuição de energia espectral, razação de harmônicos e ruído, irregularidade de tom e irregularidade de amplitude. |
| Han, Yu e Tashev | ... | MFCC |
| Fayek, Lech e Cavedon | ... | - |
| Rázuri et al. | ... | Fluxo Espectral, Centroide Espectral, Ponto de Saída Espectral, RMS, SCV, ZCR, Compactação, MFCC, Método de momentos, LPC, 2DMM, Frequência mais forte baseado em Cruzamentos em Zero, 2DMM de MFCCs, Fração de quadros de baixa energia, Frequência mais forte por FFT máximo, Frequência mais forte por Centroide espectral. |

| | | |
|---|---|---|
| Perez-Gaspar, Caballero-Morales e Trujillo-Romero | ... | Classificado por fonemas |
| Trigeorgis et al. | ... | - |
| CK et al. | ... | BSFs |
| Fayek, Lech e Cavedon | ... | - |
| Mao et al. | ... | LLDs correspondentes a 384 atributos retirados pelo openEAR |
| Lakomkin et al. | ... | 32 características de LLDs do eGEMAPS a partir do OpenSMILE |
| Torres-Boza et al. | ... | F200 e FPH |
| Tzirakis, Zhang e Schuller | ... | MFCC |
| Bhavan et al. | ... | MFCCs, Delta e Delta-Delta MFCCs e Centroides espectrais |
| Hacine-Gharbi e Ravier | ... | MFCC, Delta e Delta-Delta MFCCs |
| Li e Akagi | ... | Prosodicas IS16, Espectrais MSF |
| Ozseven | ... | 1582 openSMILE |
| Zhao, Mao e Chen | ... | LFLB e LSTM |
| Chen et al. | ... | F0, ZCR, energia RMS e MFCC (1-12) |

| **Autores** | **...** | **Seleção de Características** | **Classificadores** | **Base de Dados** |
|---|---|---|---|---|
| Jang e Kwon | ... | - | SVM | Criada |
| Kwon et al. | ... | - | - | - |
| Park, Kim e Oh | ... | - | GMM | Emotional Prosody Speech and Transcripts of LDC |
| Huahu, Jue e Jian | ... | - | HMM/SOFMNN | Criada |
| El ayadi, Kamel e Karray | ... | - | - | - |

| | | | | |
|---|---|---|---|---|
| Wu, Falk e Chan | ... | Fisher Disccriminant Ratio, Sequential Forward Feature Selection, Linear Discriminant Analysis | SVM | Berlin Emotional Speech database, VAM database |
| Chen et al. | ... | Fisher e Principal Component Analysis | SVM e ANN | Beihang University Database of Emotional Speech |
| Li et al. | ... | - | DNN-HMM (RBM) e DNN-HMM (DPT) | eNTERFACE'05 e Berlin emotion database |
| Zhang, Zhao e Lei | ... | EKIsomap | SVM | Berlin Emotional Speech database |
| Han, Yu e Tashev | ... | DNN | ELM | IEMOCAP |
| Fayek, Lech e Cavedon | ... | - | DNN | eNTERFACE'05 e SAVEE |
| Rázuri et al. | ... | - | SVM, Árvore de Decisão, ANN, BayesNet, K-NN, Naive Bayes. | eNTERFACE05 |
| Perez-Gaspar, Caballero-Morales e Trujillo-Romero | ... | - | HMM, GA+HMM | Criada |
| Trigeorgis et al. | ... | - | CNN+LSTM | RECOLA |
| CK et al. | ... | - | ELM, KNN, PNN e GRNN, | BES, SAVEE e SUSAS |
| Fayek, Lech e Cavedon | ... | - | FFN, LSTM-RNNs e ConvNets | IEMOCAP |
| Mao et al. | ... | Back Propagation Network / EDFLM | SVM | FAU AEC, ABC, Emo-DB |

| | | | | |
|---|---|---|---|---|
| Lakomkin et al. | ... | - | CNN e RNN | IEMOCAP |
| Torres-Boza et al. | ... | SC | SVR | VAM-Audio e AVEC2012 |
| Tzirakis, Zhang e Schuller | ... | CNN | CRNN | RECOLA |
| Bhavan et al. | ... | wrapper-based | SVM (GK) | Berlin EmoDB, RAVDESS e IITKGP-SEHSC |
| Hacine-Gharbi e Ravier | ... | MI (MMI, CMI, JMI e TMI) | GMM/HMM | Berlin EmoDB |
| Li e Akagi | ... | FDR, SFFS | ANFIS (Adaptative neuro fuzzy inference system) | Fujitsu DB, Emo-DB, CASIA, SAVEE |
| Ozseven | ... | Proposta por mudança de características | SVM, MLP e K-NN | Emo-DB, eNTERFACE05, EMOVO e SAVEE |
| Zhao, Mao e Chen | ... | - | CNN LSTM | Berlin Emo-DB e IEMOCAP |
| Chen et al. | ... | FCM | TLFMRF | CASIA e Berlin Emo-DB |