UNIVERSIDADE FUMEC
FACULDADE DE CIÊNCIAS EMPRESARIAIS
MESTRADO PROFISSIONAL EM SISTEMAS DE INFORMAÇÃO
E GESTÃO DO CONHECIMENTO

RONITON REZENDE OLIVEIRA

RECONHECIMENTO DE EMOÇÕES NA FALA A PARTIR DE GRAVAÇÕES TELEFÔNICAS: COMPARAÇÃO ENTRE ALGORITMOS DE EXTRAÇÃO DE CARACTERÍSTICAS ACÚSTICAS E MÉTODOS DE CLASSIFICAÇÃO

Belo Horizonte – MG
2021

RONITON REZENDE OLIVEIRA

RECONHECIMENTO DE EMOÇÕES NA FALA A PARTIR DE GRAVAÇÕES TELEFÔNICAS: COMPARAÇÃO ENTRE ALGORITMOS DE EXTRAÇÃO DE CARACTERÍSTICAS ACÚSTICAS E MÉTODOS DE CLASSIFICAÇÃO

Dissertação apresentada à Banca Avaliadora do Curso de Mestrado em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC como requisito parcial para obtenção do grau de Mestre

Área de concentração: Tecnologia e Sistema de Informação

Linha de pesquisa: Cognição e Aprendizado de Máquina

Orientador: Prof. Dr. Luiz Cláudio Gomes Maia

Belo Horizonte – MG
2021

Dados Internacionais de Catalogação na Publicação (CIP)

O48r Oliveira, Roniton Rezende, 1986-

Reconhecimento de emoções na fala a partir de gravações telefônicas: comparação entre algoritmos de extração de características acústicas e métodos de classificação / Roniton Rezende Oliveira. - Belo Horizonte, 2021.

110 f. : il.

Orientador: Luiz Cláudio Gomes Maia

Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento), Universidade FUMEC, Faculdade de Ciências Empresariais, Belo Horizonte, 2021.

1. Emoções. 2. Comunicação. 3. Classificação. I. Título. II. Maia, Luiz Cláudio Gomes. III. Universidade FUMEC, Faculdade de Ciências Empresariais.

CDU: 62:007

Ficha catalográfica elaborada pela Biblioteca Universitária-FUMEC

Dissertação intitulada "**RECONHECIMENTO DE EMOÇÕES NA FALA A PARTIR DE GRAVAÇÕES TELEFÔNICAS: COMPARAÇÃO ENTRE ALGORITMOS DE EXTRAÇÃO DE CARACTERÍSTICAS ACÚSTICAS E MÉTODOS DE CLASSIFICAÇÃO**" de autoria de Roniton Rezende Oliveira, aprovada pela banca examinadora constituída pelos seguintes professores**:**

_____

Prof. Dr. Luiz Cláudio Gomes Maia – Universidade FUMEC
(Orientador)

_____

Prof. Dr. Fernando Silva Parreiras – Universidade FUMEC
(Examinador Interno)

_____

Prof. Dr. Wladmir Cardoso Brandão – PUC MINAS
(Examinador Externo)

_____

Prof. Dr. Fernando Silva Parreiras
Coordenador do Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC

Belo Horizonte, 24 de fevereiro de 2021.

Luiz Maia.

Fernando Silva Parreiras

Wladmir Cardoso Brandão

ABARIS

REQUESTED

TITLE	**Assinatura de ata e contra-capas Universidade**
FILE NAME	**1cacc4fc-4863-45f2-9c91-37508387116a.pdf**
RESQUEST ID	**signature_request_d694912d-d95d-4570-a850-c6e0**
RESQUESTED BY	**Júlio César Teixeira e Silva**
STATUS	**Completed**

Professor (luiz.maia@fumec.br)

SENDED
02/03/2021
17:53:54UTC±0

SIGNED
09/03/2021
19:48:59UTC±0
191.185.140.62

Professor (fernando.parreiras@fumec.br)

SENDED
09/03/2021
19:49:00UTC±0

SIGNED
09/03/2021
20:04:54UTC±0
187.111.30.10

Professor (wladbrandao@gmail.com)

SENDED
09/03/2021
20:04:54UTC±0

SIGNED
09/03/2021
22:39:41UTC±0
177.40.198.210

COMPLETED
09/03/2021
22:39:41 UTC±0
The document has been completed.

AGRADECIMENTOS

Ao Pai Celestial, por me conceder a VIDA e a coragem para enfrentar os desafios.

Aos meus pais, José Reis Donizetti de Oliveira e Zêomar Costa Rezende Oliveira, pelo apoio, carinho, amor, compreensão e exemplo de vida.

Às minhas irmãs, Roniele Rezende Oliveira e Ronilene Rezende Oliveira. Vocês são companheiras que sempre me entenderam.

Ao meu irmão, Ronielton Rezende Oliveira, pelas inúmeras discussões e sugestões, sem dúvidas, motivador e incentivador, um influenciador da minha trajetória de vida pessoal e profissional.

Ao meu orientador Prof. Dr. Luiz Claudio Maia, que dedicou horas de seu tempo para que este trabalho fosse realizado com qualidade, pelos comentários valiosos e pelas inúmeras considerações que resultaram nesta dissertação.

Ao Prof. Dr. Fernando Parreiras, pela instrução e apontamento do referencial teórico para a elaboração do instrumento de pesquisa, valiosos comentários e inúmeras considerações realizadas na disciplina de projeto de dissertação e no decorrer de todo o curso.

Aos professores do Curso de Mestrado em Sistema de Informação e Gestão do Conhecimento, por provocarem a busca pelos conhecimentos necessários à realização deste trabalho.

À equipe do Mestrado e Doutorado da Universidade FUMEC, por nos bastidores, prover a infraestrutura e a logística que favoreceu a absorção do aprendizado.

A todos os participantes anônimos da pesquisa, por colaborarem, direta ou indiretamente, para o êxito deste trabalho e viabilizarem o sucesso deste projeto.

“Já que se há de escrever... que ao menos não se esmaguem com palavras as entrelinhas.”

Clarice Lispector

RESUMO

Emoções humanas são uma condição complexa e momentânea que surge em experiências de caráter afetivo e provocam alterações em várias áreas do funcionamento psicológico e fisiológico, preparando o indivíduo para a ação. O reconhecimento de emoções na fala depende da eficiência e eficácia dos métodos utilizados. Essa dissertação propõe comparar os métodos de extração de característica acústicas *Mel-Frequency Cepstral Coefficients* (MFCC) e *Perceptual Linear Predictive* (PLP), para identificação dos atributos relevantes da fala e, dos métodos de classificação *Gaussian Mixture Model* (GMM) e *Support Vector Machine* (SVM), para o agrupamento de enunciados em categorias de emoção. Para execução, uma base de dados de gravações telefônicas de *call center* será rotulada em categorias de emoções (alegria, calma, raiva, surpresa e tristeza) e os algoritmos serão aplicados, em busca de informações que representem estados emocionais dos áudios.

Palavras-Chave: Reconhecimento de Emoção na Fala, Coeficientes Cepstrais de Frequência Mel, Predição Linear Perceptual, Modelo de Mistura Gaussiana, Máquinas de Vetores de Suporte.

ABSTRACT

Human emotions are a complex and instantaneous condition that arises in experiences of an affective character and can be the cause of several psychological and physiological changes in functioning, and leaves the individual ready for action. The recognition of emotions in speech depends on the efficiency and effectiveness of the methods used. This dissertation proposes to compare the methods of extraction of acoustic characteristics Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Predictive (PLP), to identify the relevant speech attributes and the methods of classification Gaussian Mixture Model (GMM) and Support Vector Machine (SVM), for grouping statements into categories of emotion. For this realization, a database of call center telephone recordings will be labeled in categories of emotions (joy, calm, anger, surprise and sadness) and the algorithms will be applied looking for information who can represent the human emotions states in audios.

Keywords: Speech Emotion Recognition, Mel-Frequency Cepstral Coefficients, Perceptual Linear Predictive, Gaussian Mixture Model, Support Vector Machine.

LISTA DE SIGLAS

ANN	*Artificial Neural Networks*
AP	*Acoustic-Prosodic*
ASR	*Automatic Speech Recognition*
DB	*Decibels*
DC	*Direct Current*
DCT	*Discrete Cosine Transform*
DFT	*Discrete Fourier Transform*
DS	*Design Science*
DT	*Decision Trees*
DSR	*Design Science Research*
EM	*Expectation-Maximization*
ERFTrees	*Ensemble Random Forest to Trees*
FDA	*Functional Data Analysis*
FFT	*Fast Fourier Transform*
GMM	*Gaussian Mixture Model*
GMVAR	*Gaussian Mixture Vector Autoregressive*
HFCC	*Human Factor Cepstral Coefficients*
HMM	*Hidden Markov Models*
k-NN	*k-Nearest Neighbors*
LBG	*Linde-Buzo-Gray*
LFCC	*Linear Frequency Cepstral Coefficients*
LFPC	*Log Frequency Power Coefficients*
LLD	*Low-Level Descriptors*
LPC	*Linear Predictive Coding*
LPCC	*Linear Predictive Cepstral Coefficients*
LS	*Semantic Labels*
LSF	*Line Spectral Frequencies*
LV-SVR	*Least Squares-Support Vector Regression*
MDS	*Multi-Dimensional Scaling*
MDT	*Meta Decision Tree*
MFCC	*Mel-Frequency Cepstral Coefficients*
MSF	*Modulation Spectral Features*
NN	*Neural Network*

PABX	*Private Automatic Branch Exchange*
PAD	*Position Arousal Dominance*
PCA	*Principle Component Analysis*
PLP	*Perceptual Linear Predictive*
PLPC	*Perceptual Linear Prediction Coefficients*
PNCC	*Power Normalized Cepstral Coefficients*
RASTA-PLP	*Relative Spectral Transform – Perceptual Linear Prediction*
RBF	*Radial Basis Function*
SVM	*Support Vector Machine*
VAD	*Voice Activity Detection*
VC	*Vapnik Chervonenkis*
WD-KNN	*Weighted Distance K-Nearest Neighbor*

LISTAS DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

SUMÁRIO

1 INTRODUÇÃO

O estado emocional do ser humano é um fator importante para a comunicação. Ele exerce influência significativa nas interações entre as pessoas, sejam elas por expressões faciais, características da voz ou conteúdo linguístico da comunicação verbal. As emoções influenciam tanto as características da voz quanto o conteúdo linguístico a ser transmitido. Modelar uma interface homem-máquina natural, requer reconhecer, interpretar e responder às emoções expressas na fala (Mirsamadi, Barsoum, & Zhang, 2017).

A fala é um sinal complexo que contém informações sobre a mensagem e o locutor. A linguagem é um dos principais meios para expressar as emoções. A maioria dos sistemas de reconhecimento de voz processa a fala neutra, gravada em estúdio, de forma eficaz. No entanto, seu desempenho é ruim, quando se trata de fala emocional. Isso se deve à dificuldade de modelagem e caracterização das emoções presentes na fala. Em uma conversa natural, a comunicação não verbal e a presença de emoções na voz são as responsáveis pela transmissão da intenção do discurso. Uma mensagem pode ser transmitida com semânticas diferentes e incorporar emoções apropriadas, dependendo de como foi dito (Schuller, 2018). Por exemplo, a expressão "*OK*" em inglês, dependendo do contexto, pode ser usada para expressar admiração, descrença, consentimento, desinteresse ou afirmação. Portanto, compreender o texto por si só não é suficiente para interpretar a semântica do enunciado falado.

É importante que os sistemas de reconhecimento de fala sejam capazes de processar as informações não linguísticas que são transmitidas junto às mensagens. Os seres humanos entendem a mensagem pretendida ao perceber as emoções subjacentes, além da informação fonética, usando pistas multimodais, ou seja, o tom da voz, a velocidade da fala, entre outras. As informações não linguísticas podem ser observadas por meio de: 1) expressões faciais no caso do vídeo; 2) expressões emocionais no caso da fala; e 3) pontuação no caso do texto escrito. Os objetivos básicos do processamento emocional da fala são: 1) compreender as emoções presentes na fala; e 2) sintetizar as emoções desejadas na fala de acordo com a mensagem pretendida (Schuller, Rigoll, & Lang, 2004). Da perspectiva da máquina, a compreensão das emoções da fala pode ser vista como uma classificação ou discriminação de emoções. A síntese de emoções pode ser vista como a incorporação de conhecimentos específicos de emoção durante a síntese da fala.

O reconhecimento de emoções na fala tem várias aplicações no dia a dia, sendo particularmente útil, para aplicações que requerem interação homem-máquina (Koolagudi, Maity, Kumar, Chakrabarti, & Rao, 2009). Por exemplo, sistemas de direção de veículos,

com objetivo de capturar as informações sobre o estado emocional do motorista e mantê-lo em alerta durante a direção, o que ajudaria a evitar acidentes, causados pelo nível de estresse do condutor (Schuller, Rigoll, & Lang, 2004); em sistemas médicos, com objetivo de usar o conteúdo emocional da fala de um paciente como uma ferramenta de diagnóstico para distúrbios (France, Shiavi, Silverman, Silverman, & Wilkes, 2000; Long, et al., 2017); entre outros. Especificamente em sistemas de atendimentos telefônicos, o uso das técnicas de reconhecimento de emoções na fala, tem o objetivo de identificar informações que represente a percepção emocional do cliente, quanto ao atendimento recebido (Lee & Narayanan, 2005).

O reconhecimento de emoções na fala depende da eficiência e eficácia do método e dos recursos empregados na extração de características dos sinais acústicos (El Ayadi, Kamel, & Karray, 2011). Técnicas de reconhecimento de padrão raramente são independentes do domínio do problema e a seleção adequada do método e dos recursos afetam significativamente o desempenho da classificação. Na literatura nota-se a predominância no uso dos métodos de extração de características acústicas *Mel-Frequency Cepstral Coefficients* (MFCC), *Linear Predictive Coding* (LPC) e *Perceptual Linear Predictive* (PLP) – que é derivada da técnica LPC – para identificação dos atributos relevantes da fala, e dos métodos de classificação *Gaussian Mixture Model* (GMM) e *Support Vector Machine* (SVM) para agrupamento dos enunciados em categorias de emoção (El Ayadi, Kamel, & Karray, 2011; Singh, Jain, & Tripath, 2014). Assim, foi proposta nessa pesquisa a realização de uma comparação entre os métodos de extração de características acústicas MFCC e PLP e dos métodos de classificação GMM e SVM para aferir qual método de extração de características acústicas, combinado ao método de classificação apresentou melhor desempenho no reconhecimento de emoções da fala. Para tal, foram analisadas gravações telefônicas de um *call center* em busca de informações que representem as emoções do cliente. A questão que direcionou essa pesquisa foi: *Qual o desempenho das técnicas de extração de características acústicas MFCC e PLP e dos métodos de classificação GMM e SVM no reconhecimento de emoções da fala em gravações telefônicas?*

1.1 Objetivos

1.1.1 Objetivo Geral

Comparar o desempenho das técnicas de extração de características acústicas MFCC e PLP e dos métodos de classificação GMM e SVM no reconhecimento de emoções da fala em gravações telefônicas.

1.1.2 Objetivos Específicos

a) Segmentar as gravações telefônicas em canais de áudio distintos e aplicar filtros para detecção de voz, remoção de ruídos e trechos de silêncio.
b) Criar base de dados de treinamento de áudio emocional baseados nas categorias básicas de emoção[1], utilizando o canal correspondente ao cliente.
c) Aplicar os algoritmos MFCC e PLP para extrair as características acústicas e aplicar os métodos de classificação GMM e SVM para agrupar os enunciados em categorias de emoção.
d) Comparar o resultado dos métodos de extração de características acústicas e classificação no reconhecimento de emoções da fala.

1.2 Lacuna

O reconhecimento de emoções na fala é uma tarefa desafiadora. Um problema comum aos métodos utilizados é como selecionar as características que representam os fatores emocionais do diálogo (Banse & Scherer, 1996). Pesquisas sobre o reconhecimento de emoções na fala têm se concentrado na busca por características que indiquem diferentes tipos emoções (Schuller, et al., 2007; Tahon & Devillers, 2016). A abordagem mais adotada tem sido extrair os recursos estatísticos no nível de enunciado, aplicar técnicas de redução de dimensão para obter uma representação compacta e realizar a classificação com algoritmo de aprendizado de máquina (Schuller, Arsić, Wallhoff, & Rigoll, 2006; Busso, Bulut, & Narayanan, 2013).

A extração de recursos consiste em duas etapas. Primeiro, uma série de características acústicas que se acredita serem influenciadas por emoções são extraídas em quadros curtos de 20 a 50ms, estes frequentemente são chamados de *Low-Level Descriptors* (LLD). Segundo, são aplicadas diferentes funções de agregação estatísticas aos descritores LLD e os resultados são concatenados em um vetor de característica longa no nível do enunciado.

[1] Nessa pesquisa a definição das classes de emoção a serem usadas no experimento ocorrerá após a análise dos arquivos de áudio. Ocorre que em bases de dados de áudio natural não é possível saber a priori quais são os tipos emoção existentes. A seção 2.2.3 detalhará os tipos de bases de dados de áudio existentes e suas características.

O papel das funções estáticas é descrever aproximadamente as variações e contornos dos descritores LLD durante o enunciado. A suposição é que o conteúdo emocional reside nas variações temporais, em vez de valores estáticos dos LLD de curto prazo. No entanto, um enunciado proveniente de chamadas telefônicas, geralmente, é composto por várias fontes de dados, como músicas, ruídos, palmas ou vários locutores, os quais podem afetar diretamente a classificação. Especificamente no caso de múltiplos locutores, ainda há um problema adicional, que é a predominância emocional do locutor que possui maior interação no áudio.

Nesta pesquisa foram propostas: (i) a segmentação das gravações telefônicas em arquivos de áudio de canais distintos, (ii) a aplicação de algoritmos de *Voice Activity Detection* (VAD) para remoção de trechos irrelevantes do enunciado, (iii) a aplicação dos algoritmos MFCC e PLP para extração das caraterísticas acústicas e (iv) a aplicação dos algoritmos GMM e SVM para agrupamento dos enunciados em categorias de emoção.

1.3 Justificativa

O reconhecimento de voz é realidade no cotidiano e diversas aplicações têm sido desenvolvidas para facilitar o dia a dia. Por exemplo, Siri da Apple; Alexa da Amazon; Watson da IBM, entre outras. A aplicação da técnica para classificação de emoções visa a obter o estado emocional das pessoas perante as situações as quais foram envolvidas. Quando empregada no reconhecimento de emoções em gravações telefônicas empresariais, se torna uma ferramenta para auxiliar à tomada de decisão, pois permite analisar a percepção emocional dos clientes com o objetivo de formular estratégias que visem melhorar os serviços oferecidos (Takeuchi, Subramaniam, Nasukawa, & Roy, 2007; Subramaniam, Faruquie, Ikbal, Godbole, & Mohania, 2009). A partir dessa perspectiva, surgiu como oportunidade de pesquisa, a aplicação de técnicas de extração de características acústicas e algoritmos de classificação, em bases de dados de chamadas telefônicas, para obtenção de trechos de áudio que representem a percepção emocional de clientes quanto ao atendimento recebido.

1.4 Aderência

O objetivo é a comparação de técnicas de extração de características acústicas e algoritmos de classificação aplicados em bases de dados de chamadas telefônicas. Essas técnicas integram a recuperação da informação e o aprendizado de máquina, especificamente, na área de linguística, e se enquadra na linha de pesquisa Tecnologia e

Sistemas de Informação e na trilha T2 – Cognição, Aprendizado de Máquina e Recuperação da Informação, do programa de Mestrado em Sistema de Informação e Gestão do Conhecimento da Universidade FUMEC.

1.5 Estrutura do trabalho

Esta pesquisa dividiu-se em seis capítulos, incluindo esta Introdução, na qual se apresentam a pergunta de pesquisa e o objeto em estudo. A primeira seção, Objetivos, apontou o geral e os específicos. A segunda seção, Lacuna, discutiu a escolha do tema e apresenta o problema a ser resolvido. A terceira seção, Justificativa, apresentou a relevância da pesquisa. A quarta seção, Aderência, apresentou o enquadramento da pesquisa no programa de Mestrado em Sistema de Informação e Gestão do Conhecimento da Universidade FUMEC. A quinta seção, Estrutura do Trabalho, apresentou a forma de organização do documento.

No segundo capítulo, desenvolveu-se a Fundamentação Teórica em três seções. A primeira, Emoções Humanas, apresentou os conceitos básicos de emoções humanas, suas principais teorias e definições, os tipos de emoções identificados na literatura e o problema de reconhecimento de emoções na fala. A segunda, Fundamentos para o Reconhecimento de Voz, apresentou os fundamentos para o reconhecimento de voz, o funcionamento dos sistemas de reconhecimento de voz, suas principais teorias e definições e os tipos de armazenamento de arquivos de voz. A terceira, Síntese do Capítulo, apresentou uma síntese dos principais tópicos abordados no capítulo.

No terceiro capítulo, desenvolveu-se a Metodologia em seis seções. A primeira, Classificação, indicou as características e demais aspectos metodológicos assumidos para a consecução da pesquisa. A segunda, *Design Science Research*, apresentou a metodologia *Design Science* (DS) e o método *Design Science Research* (DSR), pelos quais a pesquisa é sustentada e conduzida. A terceira, Base de Dados, apresentou as características dos dados que serão utilizados na pesquisa. A quarta, Procedimentos para Construção do Artefato, apresentou as etapas que serão executadas para condução do experimento. A quinta, Métricas de Desempenho, apresentou as métricas de desempenho para validação do experimento. A sexta, Síntese do Capítulo, apresentou uma síntese dos principais tópicos abordados no capítulo.

No quarto capítulo, desenvolveu-se a Revisão de Literatura em quatro seções. A primeira, Protocolo da Revisão de Literatura, apresentou o protocolo e as bases de dados científicas utilizadas para condução da revisão de literatura. A segunda, Resultado da

Pesquisa, apresentou os resultados em nível macro da revisão de literatura. A terceira, Relatório da Revisão de Literatura, apresentou o relatório dos trabalhos selecionados para análise, contendo as principais bases de dados de áudio utilizadas nos experimentos, os principais métodos de extração de características de acústicas, as principais abordagens de classificação e por fim, uma síntese dos trabalhos analisados. A quarta, Síntese do Capítulo, apresentou uma síntese dos principais tópicos abordados no capítulo.

No quito capítulo, desenvolveu-se a Implementação de Discussão dos Resultados em três seções. A primeira, Implementação, apresentou processo realizado para construção da base de dados de áudio e desenvolvimento do artefato. A segunda, Discussão dos Resultados, apresentou os resultados das comparações entre os métodos de extração de características acústicas MFCC e PLP e classificação GMM e SVM. A terceira, Síntese do Capítulo, apresentou uma síntese dos tópicos abordados no capítulo e dos resultados.

No sexto capítulo, desenvolveram-se as Considerações Finais em três seções. Neste apresentou-se a resposta à pergunta de pesquisa ao demostrar que o objetivo geral e os objetivos específicos foram alcançados. A primeira, Limitações de Pesquisa, apresentou as limitações encontradas no desenvolvimento da pesquisa. A segunda, Contribuições da Pesquisa, apresentou as contribuições da pesquisa para a academia e para o mercado. A terceira, Recomendações para Trabalhos Futuros, apresentou as sugestões de trabalhos futuros.

As referências indicaram as sessenta e cinco fontes consultadas. O apêndice A trouxe o código do artefato desenvolvido.

2 FUNDAMENTAÇÃO TEÓRICA

Foram abordados as emoções humanas e os fundamentos para o reconhecimento de voz.

2.1 Emoções humanas

Definir emoção pode parecer simples, visto que o termo é usado com frequência no cotidiano. Frases como “o dia está lindo”, “estou feliz por você”, “como você me faz feliz”, ilustram a situação. Porém, não há consenso na ciência psicológica, sobre a definição do termo emoção. Diversas teorias foram propostas, como: Jamesiana (James, 1884), Cannon-Bard (Cannon, 1927), Psicoevolucionistas (Darwin, 1872), cognitivistas (Schachter & Singer, 1962) e sociais (Gergen, 1985), que levaram ao entendimento na literatura que as emoções não são compreendidas como reação única, mas sim como processo de múltiplas variáveis. Com isso em mente, uma possível definição para o termo emoção, poderia ser: uma condição complexa e momentânea que surge em experiências de caráter afetivo, provocando alterações em várias áreas do funcionamento psicológico e fisiológico, preparando o indivíduo para a ação (Miguel, 2015).

Essa definição baseia-se nos princípios da teoria cognitiva (Schachter & Singer, 1962), que destaca a avaliação da situação como a principal característica da emoção. A avaliação da situação é uma atividade cognitiva da qual o indivíduo pode ter consciência ou não e tem efeito determinante na emoção (Prinz, 2007). Por exemplo, se o indivíduo recebe a notícia que fora demitido, pode entender a situação como uma consequência da competitividade, para a qual não se vê preparado e se sentir triste. Por outro lado, se interpretar que era um funcionário dedicado e competente e, mesmo assim, foi demitido, pode se sentir injustiçado e com raiva.

No contexto emocional, a avaliação da situação ocorre em decorrência de um evento e as reações ao evento são o reflexo das experiências individuais e sociais do indivíduo. Assim, se o evento possuir caráter afetivo, despertará no indivíduo expressões emocionais, que podem ser classificadas como: alegria, medo, surpresa, tristeza, nojo, raiva ou neutra.

2.1.1 Expressão emocional

Os aspectos do comportamento humano podem ser observados, pela: expressão facial, expressão gesticular e expressão vocal. Porém, como o foco dessa pesquisa é a análise de emoção na fala, apenas o aspecto vocal será considerado. A expressão vocal é uma das formas mais nítidas de manifestação de emoção, visto que diferentes estados emocionais,

implicam em diversas alterações na voz, por exemplo, as alterações na frequência, no volume e no ritmo da voz (Miguel, 2015). Ao expressar estados emocionais na voz, o orador busca persuadir o outro indivíduo, proporcionando assim eficácia e poder ao discurso.

Na literatura é comum se encontrar a nomenclatura "emoções básicas" para distinguir classes desse fenômeno. Porém, não existe consenso em relação a quantas e quais são as emoções básicas. Contudo, são identificados os seguintes estados:

- Alegria: Ocorre diante do ganho de um objeto ou evento de valor (Plutchik, 2002). Na voz é expressa tipicamente por tons altos e variados, com ritmo rápido e poucas pausas entre as palavras (Juslin & Laukka, 2003).
- Medo: Ocorre diante de um evento ameaçador, gerando sensação de incerteza ou falta de controle. Tipicamente resultará em resposta de fuga, na tentativa de restabelecer a segurança (Plutchik, 2002). Na voz é expressa por tons altos e variados, com ritmo mais rápido e volume alto (Juslin & Laukka, 2003).
- Surpresa: Ocorre diante de um evento inesperado ou da interrupção súbita de um estímulo (Plutchik, 2002). É uma emoção breve e pode durar apenas alguns segundos. Após compreensão do ocorrido, pode se combinar a outra emoção, positiva ou negativa, ou não ser seguida por nenhuma outra. Na voz é expressa por tom alto e variado, e com ritmo rápido (Juslin & Laukka, 2003).
- Tristeza: Ocorre na perda de objeto, evento ou ente querido. Gera a sensação de abandono e o sentimento de substituição, no caso de objetos ou eventos, e de resignação no caso de pessoas ou animais (Plutchik, 2002). Na voz é expressa por poucas variações de tom, tipicamente baixos e com discurso pausado e lento (Juslin & Laukka, 2003).
- Nojo: Ocorre por objetos ou situações consideradas como repulsivas e indesejáveis, com a tendência a expulsão ou remoção do objeto ou situação (Plutchik, 2002). Na voz é expressa por baixa variação do tom e o ritmo lento (Juslin & Laukka, 2003).
- Raiva: Surge ao se deparar com uma situação avaliada como hostil. Se a percepção for de interferência intencional, ao invés de acidental, o nível da raiva será maior. Pode ocorrer por frustação com pessoas ou objetos inanimados. Gera uma tendência de ataque, com a intenção de remover o impedimento, mudando a situação atual, de modo que destrua ou prejudique o alvo (Plutchik, 2002). Na voz, é expressa pela elevação do volume e tom, alta variabilidade do nível som e contorno de afinação e o ritmo das palavras tende a ser maior e com menos pausas (Juslin & Laukka, 2003).

- Neutra: São aquelas que quando produzidas não conduzem a reações agradáveis ou desagradáveis, mas facilitarão o aparecimento de estados emocionais subsequentes. Na voz, é expressa pela manutenção tom de voz, sem variação do nível ou intensidade do som e o ritmo das palavras tende a ser estável (Juslin & Laukka, 2003).

Modelos de emoções propõem o agrupamento de emoções básicas para a formação de emoções complexas. E se dá, devido ao fato da maior parte dos estados emocionais do ser humano ser formado por mais de uma emoção (Plutchik, 2002). Decepção pode ser uma mistura de surpresa e tristeza; remorso uma mistura de tristeza e nojo; e saudade uma mistura de alegria e tristeza. Porém, nem sempre a mistura das emoções básicas resultará na mesma emoção complexa, pois dependerá da intensidade e da avaliação subjetiva da pessoa.

Um atleta que conquistou uma vitória difícil pode “chorar de felicidade”, o que seria uma mistura de alegria e tristeza, mas não se pode dizer que o atleta está sentindo saudade. Por isso, é importante ressaltar que, quando se diz que uma emoção é básica, não está se referindo a um fenômeno único e isolado, cujas características são exatamente as mesmas e, portanto, qualquer diferença implicaria em ser outra emoção básica. Refere-se sim, a grupos que compartilham afetos, cognições e comportamentos suficientemente semelhantes.

Nessa pesquisa a definição das classes de emoção a serem usadas no experimento ocorrerá na etapa de treinamento, após a análise dos arquivos de áudios. Por se tratar de base de dados de áudio natural não é possível saber, a priori, quais são os tipos de emoção existentes, por tanto é necessária a análise dos arquivos áudio para identificação dos tipos de emoção existentes. A seção 2.2.3 apresentará os tipos de bases de dados de áudio existentes e suas características.

2.1.2 Reconhecimento de emoções na fala

O reconhecimento de emoções na fala consiste em processar um sinal voz, identificar as alterações fisiológicas da voz e classificá-la como pertencente a uma classe de emoção (Koolagudi & Rao, 2012). Esse problema tem sido tratado por meio de modelagens estatísticas ou de reconhecimento de padrões, análogo aos problemas de reconhecimento de voz e de locutor (Jain, Duin, & Mao, 2000; Gold, Morgan, & Ellis, 2011). Nesses, amostras de voz cujas emoções são conhecidas, são extraídas de um sinal acústico com parâmetros pré-determinados, para geração de modelos de emoções. Posteriormente, os modelos são comparados as amostras de emoções desconhecidas para classificação por similaridade. A unidade de trabalho pode ser pequenos segmentos do sinal de voz, denominados de quadros, ou frases completas que contenham manifestações claras de emoções e representem o

cenário com o qual o cérebro humano usa para discernir as emoções (El Ayadi, Kamel, & Karray, 2011). Os parâmetros de voz a serem utilizados são selecionados de acordo com as ideias das teorias de emoções e de resultados de estudos anteriores de reconhecimento de padrões de voz.

2.1.3 Identificação *versus* verificação

Problemas de reconhecimento, de qualquer natureza, podem ser representados de duas formas: identificação e verificação (Jain, Duin, & Mao, 2000; Petry, 2002). Por identificações de emoções, entende-se o problema de classificar a amostra como pertencente a uma classe de emoção dentre um conjunto de emoções, conforme Figura 1. Por classes de emoção, entende-se como o conjunto de amostras com características comuns que são previamente treinadas para criação de modelos de emoção. Já por verificação de emoções, entende-se como um processo de decisão, no qual determina se a amostra pertence ou não a determinada classe, ou seja, uma validação binária, contendo o modelo de emoção em foco e um modelo complementar, formado por amostras pertencentes de outras classes de emoção da amostra para treinamento, conforme Figura 2.

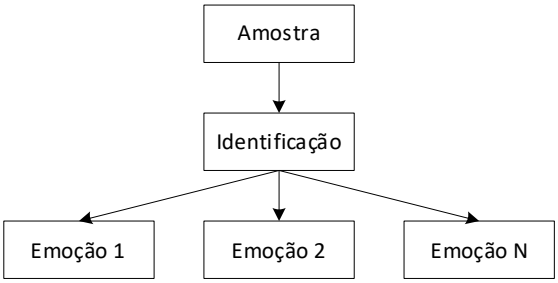

Figura 1. Modelo para identificação de emoções.
Fonte: Elaborado pelo autor.

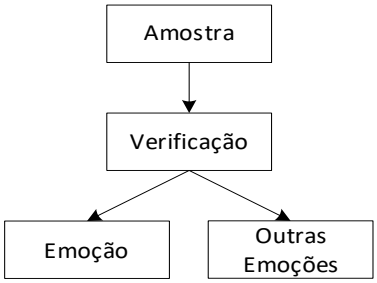

Figura 2. Modelo para verificação de emoções.
Fonte: Elaborado pelo autor.

Ambos os métodos são influenciados pelo tamanho da amostra, no caso dessa ser grande, haverá aumento exponencial na complexidade do reconhecimento, e no caso de a amostra ser pequena, tornará o reconhecimento do sistema bastante específico (Jain, Duin, & Mao, 2000). A solução tem sido adotar um conjunto de emoções formado por emoções básicas, que atende a uma variedade aplicações. A decisão pela implementação de um sistema de identificação ou de verificação, depende da aplicação. Por exemplo, o sistema de reconhecimento emocional da satisfação do cliente estaria interessado em detectar o maior número de emoções possíveis, assim a implementação seria de um modelo de identificação de emoções, já um sistema de detecção de mentiras, está interessado em reconhecer emoções

como ansiedade ou medo e o custo de falsos positivos é menor do que o de falsos negativos, assim utilizaria um modelo de verificação de emoções.

Nessa pesquisa foi aplicada a técnica identificação de emoções, pois o objetivo foi classificar uma amostra de categoria desconhecida como pertencente a uma classe de emoção.

2.2 Fundamentos para o reconhecimento de voz

O processo de reconhecimento de voz requer, como em todas as atividades de investigação, fundamentação teórica e aplicação prática. Esta seção apresentou os fundamentos e principais etapas do processo de reconhecimento de voz.

2.2.1 Sistemas para reconhecimento de voz

Os primeiros trabalhos desta área datam do início da década de 1950, quando, nos laboratórios Balashek, Bell, Biddulph e Davis foi desenvolvido um sistema de reconhecimento de dígitos isolados para um único locutor. Nas décadas seguintes outros trabalhos foram realizados. Na década de 1960 houve o desenvolvimento do sistema de reconhecimento de vogais de Suzuki e Nakata no Radio Research Lab. em Tóquio, o sistema de reconhecimento de fonemas de Sakai e Doshita na Universidade de Kyoto e o sistema de reconhecimento de dígitos nos laboratórios NEC. Na década de 1970 teve a formulação por Atal e Itakura dos conceitos fundamentais da técnica de *Linear Predictive Coding* (LPC) e o desenvolvimento do primeiro produto *Automatic Speech Recognition* (ASR) real chamado *VIP-100 System*, por Tom Martin na Threshold Technology. Na década de 1980 ocorreu o desenvolvimento dos métodos estatísticos para reconhecimento de voz, com ênfase para método *Hidden Markov Models* (HMM), que propiciou certo grau de convergência nos projetos subsequentes de sistemas de reconhecimento de voz (Juang & Rabiner, 2005).

Os sistemas para reconhecimento de voz possuem muitas aplicações, tais como: comandos e controles por voz; ditados; transcrição de discursos e diálogos interativos; e sendo classificados de diversas formas. Dentre eles podem ser destacados a habilidade em lidar com locutores específicos e não específicos (dependência de locutor e independência de locutor), com a aceitação de apenas locuções isoladas ou fala fluente (palavra isolada ou palavras conectadas), com o tamanho do vocabulário, com a perplexidade, com o nível de ruído e com a qualidade do transdutor (Martins J. A., 1997). A função de um sistema para reconhecimento de voz é receber um sinal acústico e interpretar ou produzir uma sequência de palavras ou frases que correspondam ao sinal aplicado (Yu & Deng, 2014).

A métrica mais importante para avaliação do desempenho dos sistemas de reconhecimento de voz é a taxa de erro por palavra, pois permite comparar diferentes sistemas, bem como avaliar as melhorias realizadas no próprio sistema. A tecnologia em reconhecimento de voz pode alcançar precisão em reconhecimento de dígitos isolados, independente de locutor, de até 3% de erro por palavra e sistema de ambientes de fala continua, independente de locutor, com vocabulário de palavras extenso e certas restrições gramaticais, são capazes de atingir até 80% de precisão no reconhecimento das palavras. Resultados que afirmam a utilidade potencial de um sistema para reconhecimento de voz (Martins R. M., 2014).

2.2.2 Funcionamento do sistema para reconhecimento de voz

Os sistemas modernos para reconhecimento de voz baseiam-se em técnicas modelagens estatísticas ou reconhecimento de padrões (Gold, Morgan, & Ellis, 2011). Nesses, os sinais acústicos são transformados em sequências de símbolos, analisados e estruturados em unidades computacionais. Um modelo genérico de sistema para reconhecimento de voz pode ser dividido em cinco partes: 1) aquisição do sinal sonoro; 2) pré-processamento; 3) extração de características; 4) classificação; e 5) decisão (Reynolds, 2002), conforme apresentado na Figura 3.

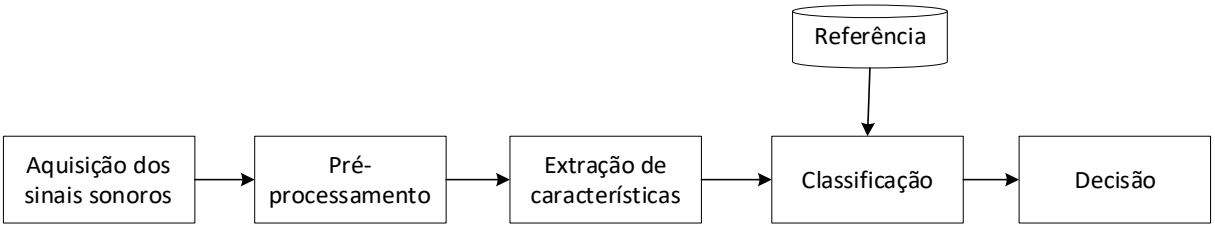

Figura 3. Modelo genérico de sistema para reconhecimento de voz.
Fonte: Adaptado de (Reynolds, 2002).

2.2.2.1 Aquisição do sinal de sonoro

O primeiro componente do sistema para reconhecimento de voz é a unidade de aquisição dos sinais sonoros. A função deste componente é captar as ondas sonoras e convertê-las para um sinal digital. As ondas sonoras são capturadas por transdutores – microfones e amplificadores – e convertidas em sinais elétricos. Em seguida são aplicados filtros de passa-baixa para suprimir os componentes de frequência superiores a 20 KHz, com objetivo de preservar o sinal da voz. Por fim, um amostrador realiza a conversão do sinal analógico em sinal digital (Petry, 2002; Reynolds, 2002), conforme a Figura 4.

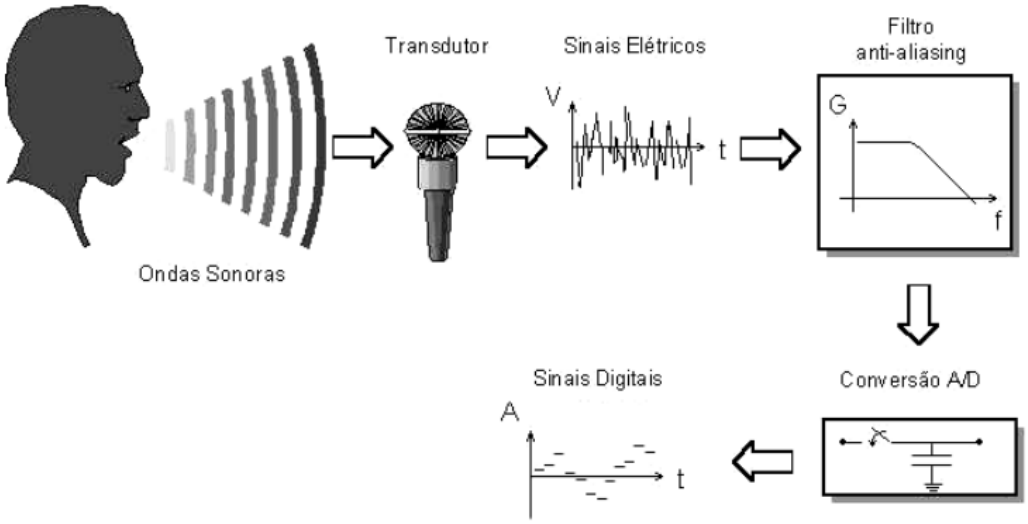

Figura 4. Processo de aquisição do sinal de voz.
Fonte: Petry (2002, p. 32).

2.2.2.2 Pré-processamento

Após a codificação do sinal, ocorre o pré-processamento, que visa a obtenção de um sinal acústico próximo ao da voz, por meio da remoção de ruídos e eliminação de componentes indesejáveis (Reynolds, 2002). A Figura 5 exibe o modelo genérico utilizado para o pré-processamento de voz, que será analisado a seguir.

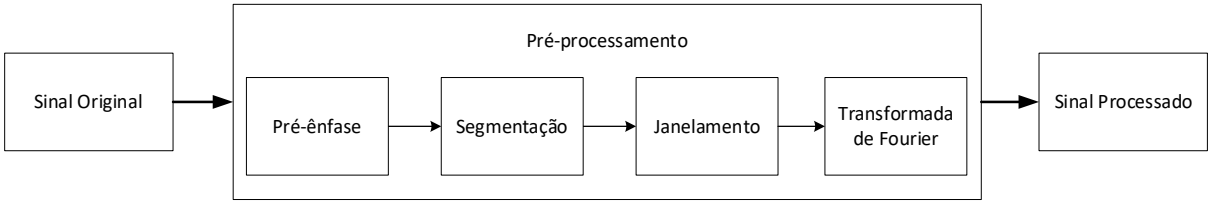

Figura 5. Modelo de pré-processamento de voz.
Fonte: Elaborado pelo autor.

2.2.2.2.1 Pré-ênfase

A pré-ênfase visa a aplicação de um filtro digital passa-alto de primeira ordem, para ressaltar as frequências mais altas e a remoção do componente *Direct Current* (DC) do sinal. Esse procedimento evitará a perda de dados durante o processo de segmentação, visto que a maior parte das informações está contida em frequências baixas (Gordillo, 2013). A Equação 1 representa a função de transferência aplicada no sinal, onde α determina a frequência de corte, com valores tipicamente variando ente 0,95 e 0,98.

$$H_{(z)} = 1 - \alpha z^{-1}, \qquad 0 \leq \alpha \leq 1$$

Equação 1. Representação da função de transferência aplicada na pré-ênfase.

2.2.2.2.2 Segmentação

A segmentação visa determinar com precisão o início e o fim de cada palavra, ou seja, distinguir no sinal acústico as partes que possuem ou não a presença de voz. Para isso, o sinal é segmentado em quadros – geralmente de 10 a 30ms, com deslocamento típico de 10ms, para evitar a perda de representação do segmento – que assumem características de estacionariedade. (Gordillo, 2013). Para calcular o número N de amostras que compõem cada quadro, multiplica-se a duração do segmento L_t, pela frequência de amostragem F_s, representado na Equação 2.

$$N = F_s(amostras/segundo) * L_t(segundos)$$

Equação 2. Representação da função aplicada para segmentação de quadros.

2.2.2.2.3 Janelamento

A segmentação do sinal produz o problema descontinuidade dos quadros, ou seja, o final do quadro passa a não corresponder ao o início do próximo (Gordillo, 2013). Para resolver este problema, aplica-se uma função de janelamento, conforme Figura 6.

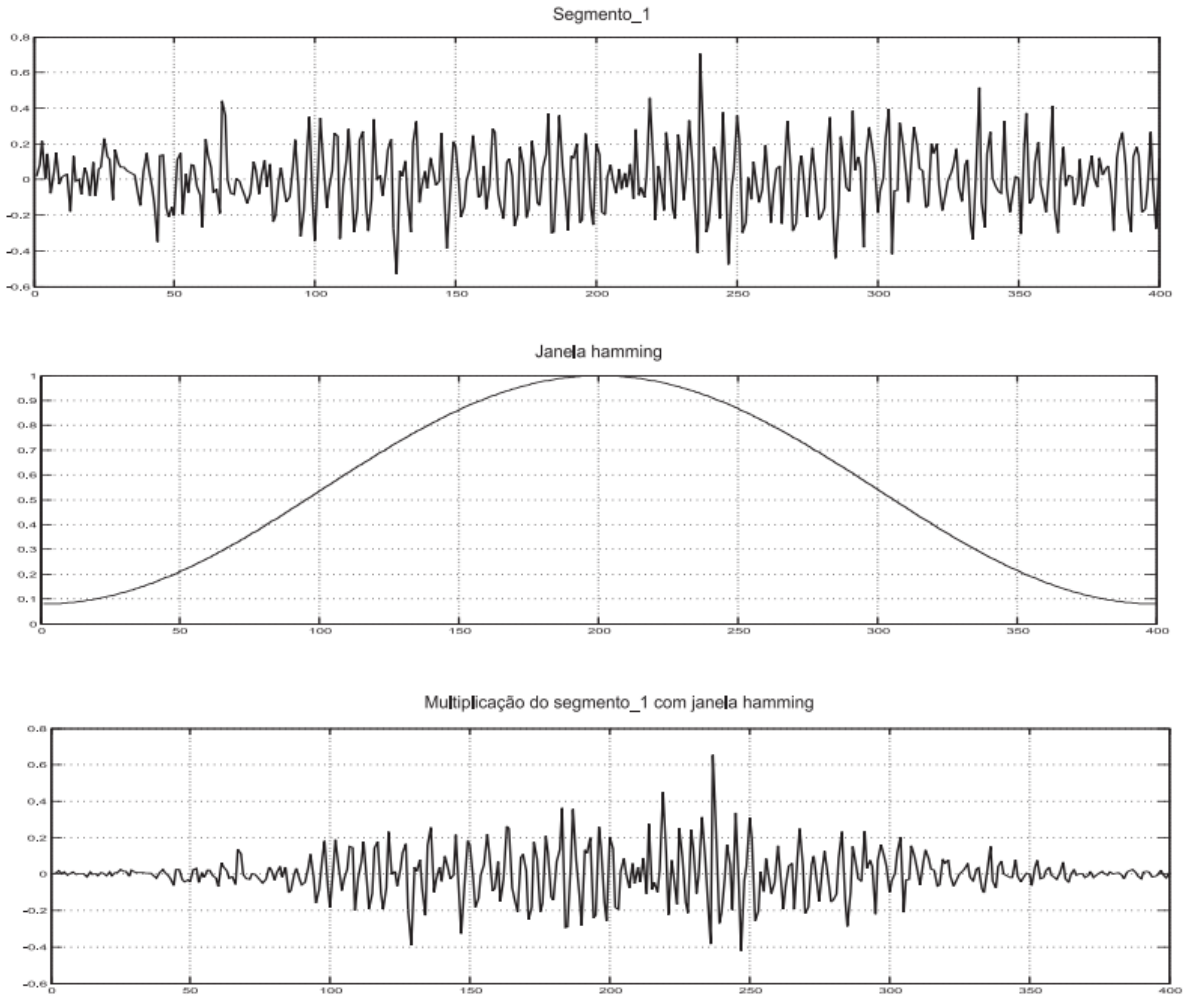

Figura 6. Segmento após aplicação da janela de *Hamming*.
Fonte: Adaptado de (Gordillo, 2013).

No reconhecimento de voz, existem diversos tipos de janelas, porém a mais utilizada é a de *Hamming*, que permite a detecção de erro de até dois bits e a correção de um bit (Gordillo, 2013). A aplicação da janela de *Hamming* é representada pela Equação 3, onde $a_0 = 0{,}53836$ é o valor ideal.

$$\omega[n] = a_0 - \underbrace{(1 - a_0)}_{a_1} . \cos\left(\frac{2\pi n}{N}\right), 0 \leq n \leq N$$

Equação 3. Representação matemática do método de *Hamming.*

2.2.2.2.4 Transformada de Fourier

A última etapa é aplicação do método *Discrete Fourier Transform* (DFT). Essa ação visa a transformação do sinal de voz em seus componentes de frequência, permitindo a diferenciação de vozes dos locutores e determinação das palavras que foram ditas, ou seja, a extração dos recursos essenciais (Gordillo, 2013). Uma questão associada à DFT é a

quantidade de operações aritméticas envolvidas no cálculo para longas amostras de sinal. Para resolver esse problema utilizam-se os algoritmos da *Fast Fourier Transform* (FFT), representado na Equação 4, que permite obter os mesmos resultados da DFT, mas em menor tempo e complexidade.

$$X(k) = \sum_{k=0}^{\frac{N}{2}-1} x(2n) W \frac{nk}{\frac{n}{2}} + W \frac{k}{n} \sum_{n=0}^{\frac{N}{2}-1} x(2n+1) W \frac{nk}{\frac{n}{2}}$$

Equação 4. Representação da equação transformada rápida de Fourier.

Na FFT o $x(n)$ é o sinal da voz, o N é o número de amostras na potência de dois e o k é a frequência. Essa equação apresenta os elementos dos índices pares e ímpares do sinal. Seus resultados são informações sobre a quantidade de energia de cada banda de frequência.

2.2.2.3 Extração de características

A etapa de extração e seleção das características acústicas dos sinais sonoros é fundamental no projeto de um sistema de reconhecimento de voz, uma vez que a avaliação direta do sinal, em virtude da grande quantidade de dados, não traz resultados significativos. O objetivo desta etapa é a redução do sinal digital a segmentos de dados, que caracterizem o sinal da voz (Petry, 2002). Diversas técnicas podem ser empregas para análise acústica, porém essa pesquisa abordará as técnicas de análise espectral.

Diversas técnicas de análise espectral podem ser utilizadas para extração de características do sinal de voz, como por exemplo: *Linear Predictive Coding* (LPC); Cepstral Analysis of Speech; *Mel-Frequency Cepstral Coefficients* (MFCC); *Linear Frequency Cepstral Coefficients* (LFCC); *Linear Predictive Cepstral Coefficients* (LPCC); *Human Factor Cepstral Coefficients* (HFCC); *Line Spectral Frequencies* (LSF); *Perceptual Linear Predictive* (PLP); *Relative Spectral Transform – Perceptual Linear Prediction* (RASTA-PLP); *Power Normalized Cepstral Coefficients* (PNCC); e outras, porém, notam-se que as técnicas predominantes na literatura e que são consideradas como mais precisas para sistemas de reconhecimento de voz são: MFCC, LPC e PLP – que é derivada da técnica LPC (Singh, Jain, & Tripath, 2014). Assim, para o desenvolvimento dessa pesquisa, optou-se pela utilização das técnicas MFCC e PLP.

2.2.2.3.1 *Mel Frequency Cepstral Coefficients* (MFCC)

O MFCC é uma das técnicas de extração de características mais usadas em sistemas de reconhecimento da voz, com base no domínio da frequência e é considerada mais precisa que técnicas que concatenam dados no domínio do tempo (Tiwari, 2010). Essa técnica utiliza uma escala chamada *Mel-Cepstral*, que visa transcrever as características perceptíveis pelo ouvido humano, ou seja, as de baixo nível, que são foneticamente mais importantes para a percepção humana do que as de alto nível (Tiwari, 2010). A Figura 7, apresenta as etapas da extração de características do MFCC.

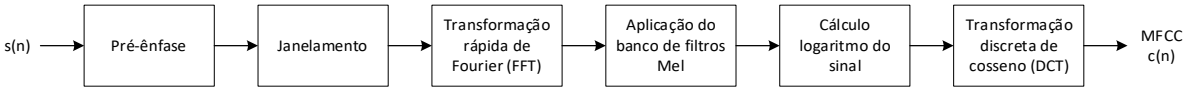

Figura 7. Etapas da extração de características do MFCC.
Fonte: Elaborado pelo autor.

O início do processo se dá com a realização das etapas de aquisição do sinal de voz e pré-processamento, que é dividido em: pré-ênfase; segmentação; janelamento; e transformada de Fourier. Com os resultados obtidos na FFT aplica-se um banco de filtros à potência espectral, que é formado por filtros triangulares, espaçados de acordo com a escala de frequência Mel, representada pela Equação 5, no qual *f* é a frequência de corte (Gordillo, 2013).

$$mel(f) = 2595 * \log10\left(1 + \frac{f}{700}\right)$$

Equação 5. Representação da escala de frequência Mel.

Após o cálculo, é realizado o *log* dos valores do espectro Mel, uma vez que a resposta humana ao nível do sinal é logarítmica. O uso do *log* faz com que o recurso seja menos sensível a variações na entrada, ou seja, evita as variações de energia que podem ocorrer devido à proximidade do microfone em relação a boca do locutor (Gordillo, 2013).

Na etapa final, os coeficientes do espectro do *log* Mel são convertidos novamente no domínio do tempo usando a *Discrete Cosine Transform* (DCT), representada na Equação 6, onde *N* é o comprimento do sinal *x(n)*. O uso da DCT se dá, pois, a maior parte da energia fica concentrada em poucos coeficientes, e a energia é uma propriedade importante dos sinais de voz (Gordillo, 2013).

$$C(k) = \alpha(k) \sum_{n=0}^{N-1} x(n) cos \frac{\pi \left(n + \frac{1}{2}\right) k}{N}, \text{para } 0 \leq k \leq N-1$$

Equação 6. Representação da equação transformação discreta de cosseno.

Para calcular o valor de *α(k),* utiliza-se a Equação 7.

$$\alpha(k) = \begin{Bmatrix} \sqrt{\frac{1}{N}} & \text{para k} = 0 \\ \sqrt{\frac{2}{N}} & \text{para } 1 \leq \text{k} \leq \text{N} - 1 \end{Bmatrix}$$

Equação 7. Valor de alfa na equação transformação discreta de cosseno.

Um vetor acústico de MFCC é computado para cada quadro e formado geralmente por 39 elementos, estáticos – 12 coeficientes cepstrais extraídos do MFCC – e dinâmicos – 1 coeficiente de energia, 13 coeficientes de velocidade (delta) e 13 coeficientes de aceleração (delta-delta). (Gordillo, 2013).

Os coeficientes dinâmicos são utilizados para captar as mudanças temporais bruscas presentes no espectro. A energia de um quadro, representada pela Equação 8, é calculada por meio da soma ao longo do tempo da capacidade das amostras no quadro (Gordillo, 2013).

$$Energia = \sum_{t=t_1}^{t_2} x^2[t]$$

Equação 8. Representação da equação energia do quadro.

Para o cálculo do coeficiente delta (Δ) é utilizado regressão linear sobre um quadro, ou seja, calculando a diferença entre os quadros anteriores e posteriores. O valor delta *d(t)* para um determinado valor cepstral *c(t)* no tempo *t* pode ser estimado pela Equação 9 (Gordillo, 2013).

$$d(t) = \frac{c(t+1) - c(t-1)}{2}$$

Equação 9. Cálculo do coeficiente delta.

Os parâmetros de segunda ordem, chamados de *delta-delta* (Δ^2), são obtidos replicando a derivada sobre os resultados obtidos na primeira derivação, calculada com a Equação 9 (Gordillo, 2013).

2.2.2.3.2 *Perceptual Linear Prediction* (PLP)

O PLP foi proposto por Hermansky e visa melhorar a estimativa do modelo de predição linear (LPC) considerando as características psicoacústicas do sistema auditivo humano (Motlícek, 2003; Borges, 2011). O sistema auditivo humano é constituído basicamente de três partes: o ouvido externo, ouvido médio e ouvido interno. O ouvido externo tem como função principal coletar os sons e conduzi-los, através do conduto auditivo, até o tímpano, que é a parte integrante do ouvido médio. Outra função do ouvido externo é fazer uma espécie de filtragem do som com o intuito de auxiliar a localização da origem das fontes sonoras. Também é função do ouvido externo servir de amplificador dos sinais da faixa de frequência da fala humana (entre 2k Hz e 5k Hz) (Motlícek, 2003). O ouvido médio tem a função de transformar as variações de pressão causadas pelas ondas sonoras recebidas pelo ouvido externo em energia mecânica que é então transmitida ao ouvido interno. Também é função do ouvido médio regular a diferença de pressão entre os ouvidos externo e interno. Por fim, o ouvido interno tem a função de transformar as ondas mecânicas do som recebido pelo ouvido médio em impulsos neurais codificados que são enviados ao cérebro para serem interpretados.

No modelo proposto por Hermansky, o espectro do sinal de voz é modificado de acordo com características psicoacústicas. A ideia é semelhante a utilizada no cálculo dos coeficientes *Mel-Cepestros*, entretanto, utiliza-se filtros assimétricos e com banda maior que a dos filtros triangulares para simular as bandas críticas e a escala Bark para espaçamento desses filtros. Além disso, incorporou também pré-ênfases e compressões com o objetivo de simular determinadas áreas do ouvido humano (Motlícek, 2003). A Figura 8 ilustra o modelo proposto.

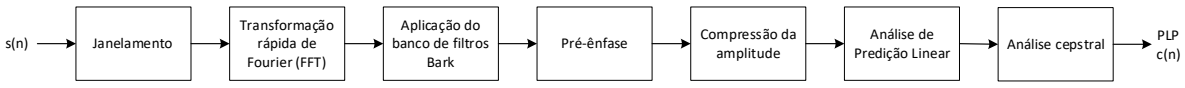

Figura 8. Etapas da extração de características do PLP
Fonte: Elaborado pelo autor.

O início do processo se dá com a realização das etapas de aquisição do sinal de voz, segmentação, janelamento e transformada de Fourier. Com os resultados obtidos na FFT aplica-se um banco de filtros à potência espectral, que é formado por filtros assimétricos,

espaçados de acordo com a escala de frequência Bark. Uma das primeiras constatações feitas em relação à percepção de frequência do sistema auditivo humano é que ele não percebe as componentes de frequência de forma linear e por este motivo se faz necessário a conversão do espectro do sinal da escala Hz para a escala Bark. A relação entre a escala Hz e a escala Bark é definida na Equação 10.

$$v = 6arcsin\left(\frac{f_{Hz}}{600}\right)$$

Equação 10. Relação entre a escala Hz e a escala Bark.

Na escala Bark, as componentes de baixa frequência têm uma relação de escala de frequência linear e as componentes de alta frequência têm uma relação de escala de frequência logarítmica (Borges, 2011). Além das componentes de frequência não serem percebidas de forma linear, também não são percebidas de forma isolada. Os sensores que captam as diferentes frequências funcionam como uma espécie de filtro de faixas de frequência. Cada sensor capta uma faixa de frequência diferente em torno de uma frequência central que é chamada de banda crítica. Para dada frequência central, as componentes vizinhas são interpretadas como pertencentes à sua banda crítica. Um filtro de banda crítica é definido através da Equação 11, onde *v* são as frequências na escala Bark e v_c a frequência central.

$$H_c(v) = f(x) = \begin{cases} 0 & para - 2{,}5 \leq v - v_c \\ 10^{(v - v_c + 0{,}5)} & para - 2{,}5 < v - v_c < -0{,}5 \\ 1 & para - 0{,}5 \leq v - v_c \leq 0{,}5 \\ 10^{-2{,}5(v - v_c - 0{,}5)} & para\ 0{,}5 < v - v_c < 1{,}3 \\ 0 & para\ v - v_c \geq 1{,}3 \end{cases}$$

Equação 11. Representação de filtro de banda crítica.

A resposta em frequência do espectro de potência do sinal sonoro após a análise em bandas críticas é representada pela Equação 12 (Borges, 2011).

$$I_k(v) = \sum_{v\,\epsilon\{v_c\}} H(v) X_k(v)$$

Equação 12. Resposta em frequência do espectro.

Em função da anatomia do sistema auditivo humano, o nível mínimo de percepção de uma componente de frequência varia. Este comportamento do nível mínimo de percepção é estimado pela curva descrita na Equação 13.

$$E(v) = \frac{(v^2 + 56{,}8\ 10^{+6})v^4}{(v^2 + 6{,}3\ 10^{+2})^2(v^2 + 0{,}38\ 10^{+9})}$$

Equação 13. Nível mínimo de percepção de uma componente de frequência.

Para frequências mais baixas o limiar de percepção em decibéis (db) de nível de pressão de sonoro está em torno de 9 db. Para as frequências médias o limiar está em torno de 0 db e para as frequências altas o limiar é de acima de 20 db. O espectro de potência do sinal após a pré-ênfase é dado pela Equação 14 (Borges, 2011).

$$EI_k(v) = I_k(v)E(v)$$

Equação 14. Representação do espectro de potência do sinal após a pré-ênfase.

Para finalizar o processamento psicoacústico ocorre a etapa de conversão da intensidade-sonoridade, realizada a partir da extração da raiz cúbica do espectro do sinal, representado pela Equação 15 (Borges, 2011).

$$L(v) = \sqrt[3]{EI_k(v)}$$

Equação 15. Representação da etapa de conversão da intensidade-sonoridade.

O término da análise de predição linear perceptual é realizado aplicando a transformação do espectro do sinal processado para o domínio temporal, por meio da transformada discreta de Fourier inversa e por fim, os coeficientes são então estimados pela análise de predição linear. Na utilização do PLP no reconhecimento de voz é comum a realização de transformações adicionais sendo que a mais comum é o PLP-Cepestro e suas primeiras e segundas derivadas.

2.2.2.4 Classificação

O reconhecimento de emoções na fala tem sido resolvido por meio de técnicas de análise estatísticas ou de reconhecimento de padrões, cujo modelo é gerado por meio de segmentos de áudio de emoções conhecidas e posteriormente confrontados a segmentos de áudio de emoções desconhecida (Iriya, 2014). Nos modelos estáticos, geralmente, a emoção

é modelada por uma densidade de probabilidade e a classificação se dá pela máxima verossimilhança. Na abordagem determinística, a classificação se dá otimizando alguma medida objetiva que compara os parâmetros de voz da instância com os parâmetros de voz do modelo. O tipo de parâmetro de voz, global ou dinâmico, dá origem a duas vertentes de classificação: estática e dinâmica.

2.2.2.4.1 Classificação estática

A diferença entre os métodos de classificação estáticos e dinâmicos está no tipo de parâmetro de voz usado, que são globais para os estáticos e dinâmicos (o contorno dos parâmetros de curto prazo) para os dinâmicos. Os conceitos de classificação estática e dinâmica podem ser quase confundidos com os de classificação estatística ou determinística, pois na maioria dos casos, métodos determinísticos utilizam parâmetros globais enquanto os métodos estatísticos utilizam parâmetros de curto prazo. Os métodos estáticos utilizam parâmetros globais, que em geral são cálculos de estatística descritiva calculados a partir do contorno de parâmetros de curto prazo, tais como: média, desvio padrão, máximo, mínimo, mediana, percentis, além de aspectos de duração e estatísticas em picos e vales. A necessidade de parâmetros globais advém dos parâmetros de curto prazo que não apresentarem individualmente uma informação útil para o modelo, além do contorno que pode ter comprimento variável. Ter um número variável de parâmetros não é aceitável para muitos métodos, que podem gerar matrizes ou vetores de tamanho fixo, de acordo com o número de parâmetros. Uma desvantagem da classificação estática é que como são geradas estatísticas globais do contorno, a informação temporal é perdida (Iriya, 2014). Exemplos de métodos de classificação estática são o *k-Nearest Neighbours* (k-NN), *Support Vector Machines* (SVM) e *Artifical Neural Networks* (ANN).

2.2.2.4.2 Classificação dinâmica

Métodos de classificação dinâmica utilizam todo o contorno dos parâmetros de curto prazo tanto para treinar os modelos quanto para a classificação em si. A classificação dinâmica é mais comum para métodos estatísticos, uma vez que toda a sequência de parâmetros de curto prazo pode ser utilizada para alimentar o algoritmo, de forma a estimar a densidade de probabilidade para a ocorrência da sequência de observações, ou calcular a verossimilhança desta sequência utilizando um modelo específico. Quanto mais pontos existirem, melhor é a estimativa da densidade de probabilidades e mais correto é o cálculo da verossimilhança. Exemplos de métodos de classificação dinâmica são os classificadores bayesianos simples, *Gaussian Mixture Models* (GMM) e *Hidden Markov Models* (HMM).

A vantagem de classificadores dinâmicos é que eles mantêm a informação temporal nas sequências de observações, quando combinados com as derivadas do contorno. Entretanto, nos modelos estatísticos é necessário assumir a forma da distribuição probabilística da observação, que é na verdade imprevisível, e a forma escolhida pode não ser um modelo realístico (Iriya, 2014).

2.2.2.4.3 Métodos de classificação

A seguir foram apresentados os métodos de classificação selecionados para o desenvolvimento dessa pesquisa.

2.2.2.4.3.1 *Gaussian Mixture Model* (GMM)

Os GMMs são um tipo particular de modelos de misturas, cuja importância é indiscutível para a área de processamento de voz e principalmente para os temas de reconhecimento de voz e de locutor. Modelos de misturas são modelos estatísticos que assumem que uma população de dados pertence a uma distribuição em forma de mistura, ou seja, uma distribuição formada por combinações lineares de diversas funções densidade de probabilidade. Os modelos GMMs representam uma distribuição como a soma ponderada de diversas gaussianas de diferentes médias e matrizes de covariância e pode ser representado por meio da Equação 16, onde *L* é a dimensão de cada observação (Iriya, 2014).

$$prob[x|\lambda] = \sum_{i=1}^{M} c_i \, \mathrm{g}(x, \mu_i, \Sigma_i)$$

$$\mathrm{g}(x, \mu_i, \Sigma_i) = \frac{1}{\sqrt{2\pi^L |\Sigma_i|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \sum\nolimits_i^{-1} (x-\mu_{i)}\right)$$

Equação 16. Representação modelo misturas gaussianas.

O modelo de mistura de gaussianas $\lambda = [c_i, \mu_i, \Sigma_i], 1 \leq i \leq M$ é definido pelos seguintes parâmetros: 1) c_i: O peso ou probabilidade a priori de cada gaussiana *m*; 2) μ_i: A média de cada gaussiana *m*; e 3) Σ_i: A matriz de covariância de cada gaussiana m. O GMM é largamente utilizado para estimar funções densidade de probabilidade desconhecidas, cuja tarefa de estimação torna-se a de estimar os pesos de cada gaussiana e suas respectivas médias e matrizes de covariância a partir dos dados observados. Também são utilizados como mecanismo de classificação bayesiana, no qual é escolhido o modelo que apresenta maior verossimilhança em relação à sequência de observações. No GMM podem ser utilizados tanto parâmetros globais quanto parâmetros de curto prazo, sendo o segundo

preferível na maioria dos casos, já que o modelo pode ser treinado por uma quantidade maior de dados (Iriya, 2014).

Muitas vezes, estes modelos podem ser considerados como um caso especial de HMM com apenas um estado e o algoritmo de *Baum-Welch* pode então ser utilizado para treinar as gaussianas. Entretanto, é mais comum a utilização do algoritmo *Expectation-Maximization* (EM) (Iriya, 2014). Embora o algoritmo EM seja sempre convergente, não é garantido que ele convirja para o máximo global. Normalmente ele converge para um máximo local, que pode não ser significativo. Portanto, a inicialização dos parâmetros iniciais tem impacto no resultado final do modelo e a escolha do algoritmo de agrupamento como o *K-Means* ou o *Linde-Buzo-Gray* (LBG) é importante para o processo. Outra solução para esse problema é a utilização de algoritmos que pós-processam o resultado final do EM tradicional ou permitem que o algoritmo seja utilizado recursivamente, de modo a tornar o algoritmo mais robusto.

2.2.2.4.3.2 *Support Vector Machines* (SVM)

O SVM é uma técnica utilizada em reconhecimento de padrões para decisões binárias, que tem sido largamente utilizada para o reconhecimento de emoções por meio da voz. A origem das SVM está fortemente ligada à teoria de minimização de riscos e à dimensão de *Vapnik Chervonenkis* (VC), que mede a capacidade da máquina de aprendizagem, isto é, sua habilidade de lidar com amostras que não foram usadas para o treinamento. As SVM fornecem uma solução ótima para o problema de separação de duas classes linearmente separáveis. Como observado na Figura 9 é possível pensar em diversas retas que separam as amostras para o caso de até duas dimensões, ou hiperplanos para dimensões maiores.

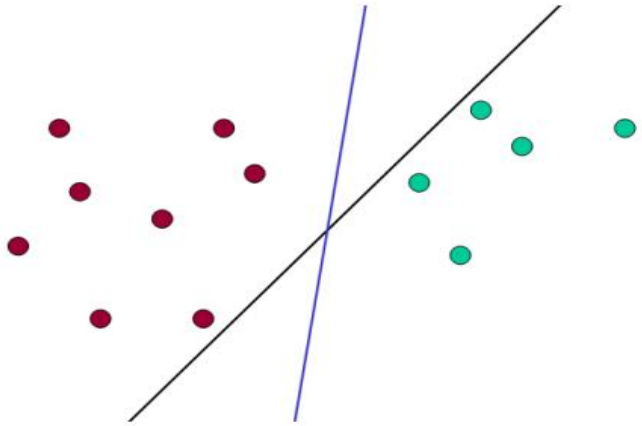

Figura 9. Separação linear de amostras de duas classes.
Fonte: Adaptada de (Iriya, 2014).

A escolha do hiperplano ótimo depende das amostras, na qual alguns pontos podem ser mais importantes que outros. Neste contexto, são definidos os vetores de suporte que nada mais são do que as amostras mais próximas à fronteira de decisão que alterariam o hiperplano ótimo caso fossem removidas. O algoritmo de treinamento das SVM procura pelo hiperplano que separa as classes com a maior margem, definida pela distância entre as amostras de cada classe mais próximas ao hiperplano, conforme Figura 10.

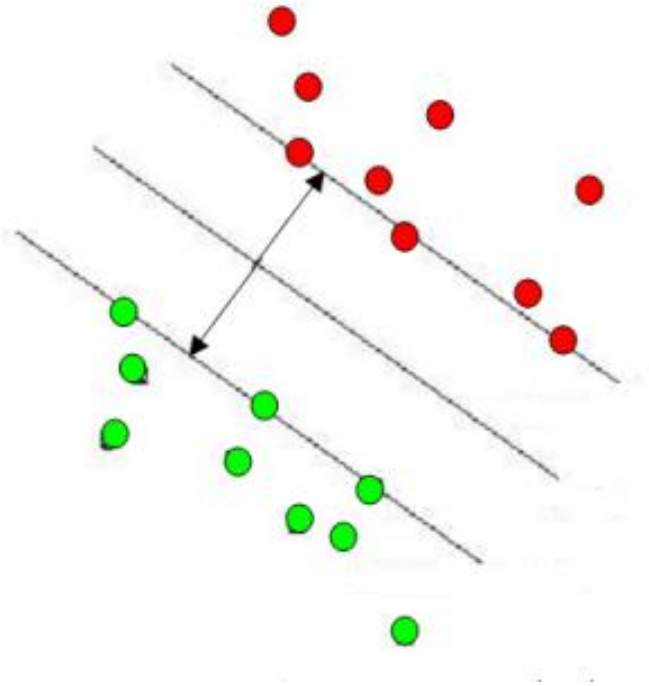

Figura 10. Separação por hiperplanos com máxima margem.
Fonte: Adaptada de (Iriya, 2014).

2.2.2.5 Decisão

As distâncias ou probabilidades obtidas na comparação com os padrões de referência são usadas para escolher o padrão que melhor corresponde ao padrão desconhecido. Para auxiliar na escolha do melhor candidato, podem-se usar restrições sintáticas e semânticas, por exemplo, o uso de uma gramática. Dessa forma, candidatos não razoáveis serão eliminados.

2.2.3 Armazenamento de arquivos de voz

Para caracterizar emoções, seja para síntese ou para reconhecimento, um banco de dados de fala emocional adequado é um pré-requisito obrigatório. Uma questão importante a ser considerada na avaliação dos sistemas de fala emocional é a qualidade do banco de

dados utilizado para desenvolvimento e teste da aplicação (Ververidis & Kotropoulos, 2006). Os objetivos e métodos de coleta de dados dos sistemas de fala emocional, variam de acordo com a motivação do sistema e as classificações dos tipos de bancos de dados são: 1) banco de dados de fala emocional simulado (baseado em ator); 2) banco de dados de fala emocional induzido; e 3) banco de dados de fala emocional natural (Koolagudi & Rao, 2012). A Figura 11 apresenta as características dos bancos de dados de fala emocional.

Tipo de banco de dados	Vantagens	Desvantagens
Simulado	• Padronizado. • Os resultados podem ser comparados facilmente. • Uma gama completa de emoções está disponível. • Grande variedade de bancos de dados, em diversos idiomas.	• As Emoções são artificiais. • Natureza episódica, sem contextualização no mundo real. • O discurso frequentemente é lido e não pronunciado de forma natural.
Induzido	• Próximo dos bancos de dados naturais. • A informação contextual está presente, mas é artificial.	• Nem todos os tipos de emoções estarão disponíveis. • Caso o orador saiba que está sendo gravado, o resultado poderá ser artificial.
Natural	• Dados capturados em aplicações reais. • Útil para modelagem de emoções do mundo real.	• Nem todos os tipos de emoções estarão disponíveis. • Questões de direitos autorais e privacidade. • Sobreposição de expressões. • Presença de ruído de fundo. • Contém emoções múltiplas e simultâneas. • Natureza difusa. • Difícil de modelar.

Figura 11. Principais características dos bancos de dados de fala emocional.
Fonte: Adaptador de (Koolagudi & Rao, 2012).

Os bancos de dados de fala simulada são criados com gravações de artistas. Nesses bancos de dados, eles são convidados a expressar frases linguisticamente neutras em diferentes tipos de emoção. A gravação é feita em diferentes sessões, com objetivo de considerar as variações no grau de expressividade e no mecanismo físico de produção da fala do ser humano. É um dos métodos mais fácies e confiáveis para coleta dados de fala emocional e contará com uma ampla gama de emoções. Este método representa mais de 60% dos bancos de dados utilizados nas pesquisas de fala emocional. As emoções coletadas com esse método são tipicamente intensas e incorporam a maioria dos aspectos relevantes para expressão de emoções (Schröder, 2001). Geralmente, as emoções simuladas são mais expressivas do que as reais.

Os bancos de dados de fala induzida são criados por simulação de situação emocional, sem conhecimento do sujeito. O sujeito é induzido a se envolver em uma conversa emocional com um moderador, no qual diferentes tipos de situações serão criados,

a fim de elicitar situações emocionais. Essa base de dados tende a ser mais natural do que a simulada, porém os assuntos tratados podem não ser propriamente expressivos, caso as partes saibam que estão sendo gravadas. Às vezes, essas bases de dados são criadas pedindo ao sujeito que se envolva em uma interação verbal com um computador, cujas respostas de fala, são por sua vez, controladas por um humano, sem o conhecimento do sujeito (Schröder, 2001).

Ao contrário dos bancos de dados anteriores, no banco de dados de fala natural, as emoções são expressas suavemente e, às vezes, de difícil caracterização e categorização. A captura desses dados pode ser realizada em diferentes ambientes, por exemplo, *call center*, *cockpit* de aviões durante situações anormais, consultório médico durante o diálogo do médico e o paciente, em conversas emocionais em locais públicos e assim por diante. O processo de anotação dessas bases é altamente subjetivo e a categorização é sempre discutível, além das próprias questões legais, como privacidade e direitos autorais, para utilização desses bancos de dados (Schröder, 2001).

O projeto e coleta dos dados de fala emocional dependem principalmente dos objetivos da pesquisa. Por exemplo, um banco de dados de fala emocional de um único locutor é suficiente para o propósito de síntese de fala emocional, ao passo que, para reconhecer emoções, é necessário um banco de dados com múltiplos locutores e diferentes estilos de expressão da emoção.

2.3 Síntese do capítulo

Foram abordados os principais tópicos teóricos e definições que corroboram para o desenvolvimento da pesquisa. Os principais tópicos foram:

- Emoções humanas: uma condição complexa e momentânea que surge em experiências de caráter afetivo e provocam alterações em várias áreas do funcionamento psicológico e fisiológico, preparando o indivíduo para a ação (Miguel, 2015). Em outras palavras, as emoções ocorrem como consequência da avaliação de uma situação e são reflexos das experiências individuais e sociais do indivíduo.
- Reconhecimento de emoções na fala: é o processo de processar um sinal voz, identificar as alterações fisiológicas da voz e classificá-las como pertencente a uma classe de emoção (Koolagudi & Rao, 2012).

- Sistemas de reconhecimento de voz: é um sistema que recebe um sinal acústico, interpreta e o classifica em classes pré-determinadas ou produz uma sequência de palavras ou frases que correspondam ao sinal aplicado (Yu & Deng, 2014).
- Funcionamento dos sistemas de reconhecimento de voz: nestes os sinais acústicos são transformados em sequências de símbolos, analisados e estruturados em unidades computacionais. Em um modelo genérico o sistema pode ser dividido em cinco partes: 1) aquisição do sinal sonoro; 2) pré-processamento; 3) extração de características; 4) classificação; e 5) decisão (Reynolds, 2002).
- Banco de dados de fala emocional: são um conjunto de dados utilizados para o processamento de sinais sonoros. Os objetivos e métodos de coleta de dados variam de acordo com a motivação do sistema. As classificações dos tipos de bancos de dados de fala emocional são: 1) banco de dados de fala emocional simulado (baseado em ator); 2) banco de dados de fala emocional induzido; e 3) banco de dados de fala emocional natural (Koolagudi & Rao, 2012).

3 METODOLOGIA

A seguir, foram apresentados os aspectos processuais seguidos para realização da pesquisa.

3.1 Classificação

A síntese do desenho da pesquisa é exibida na Figura 12.

Desenho da Pesquisa	Dissertação
Tipo	Exploratória
Abordagem	Quantitativa
Natureza	Aplicada
Procedimento	Experimental
Método	Design Science Research
Tamanho da Amostra	100 horas de gravações de chamadas telefônicas
Objeto de análise	Classificação do estado emocional em gravações telefônicas
Técnica	Aplicação de algoritmos de extração de características acústicas e algoritmos de classificação para obtenção do estado emocional do enunciado

Figura 12. Matriz Metodológica.
Fonte: elaborada pelo autor.

3.2 Design Science Research

Esta pesquisa é sustentada pela metodologia *Design Science* (DS) e conduzida pelo método *Design Science Research* (DSR). Os paradigmas DS e DSR foram discutidos nos últimos anos e são considerados, respectivamente, um quadro teórico e uma estratégia de pesquisa, capazes de orientar tanto a construção do conhecimento quanto aprimorar as práticas em sistemas de informação e de várias disciplinas, seja no campo gerencial ou tecnológico da ciência da informação (Wieringa, 2009; Bax, 2015).

No DSR a compreensão do problema e a sua solução são alcançados com a construção e aplicação de um artefato. Os artefatos são as representações simbólicas materializáveis, ou seja, construtos (entidades e relações), modelos (abstrações e representações), métodos (algoritmos e práticas) e instanciações (implementação de sistemas e protótipos) (March & Smith, 1995; Hevner, March, Park, & Ram, 2004). Por assumir abordagem pragmática o DSR não anseia alcançar verdades absolutas, teorias ou leis gerais, mas sim identificar e compreender os problemas do mundo real e propor soluções

apropriadas e úteis, capazes de avançar o conhecimento teórico na área (Hevner, March, Park, & Ram, 2004; Wieringa, 2009).

Em relação ao rigor e a relevância na DS observa-se que sete diretrizes são utilizadas para avaliação de uma pesquisa conduzida pelo método DSR (Hevner, March, Park, & Ram, 2004): projeto de artefato; relevância do problema; avaliação do projeto; contribuições da pesquisa; rigor da pesquisa; projeto como processo de pesquisa; e comunicação da pesquisa. A Figura 13 apresenta a abordagem de pesquisa de DSR de Hevner adaptada a essa pesquisa, a qual será explicada nos tópicos a seguir.

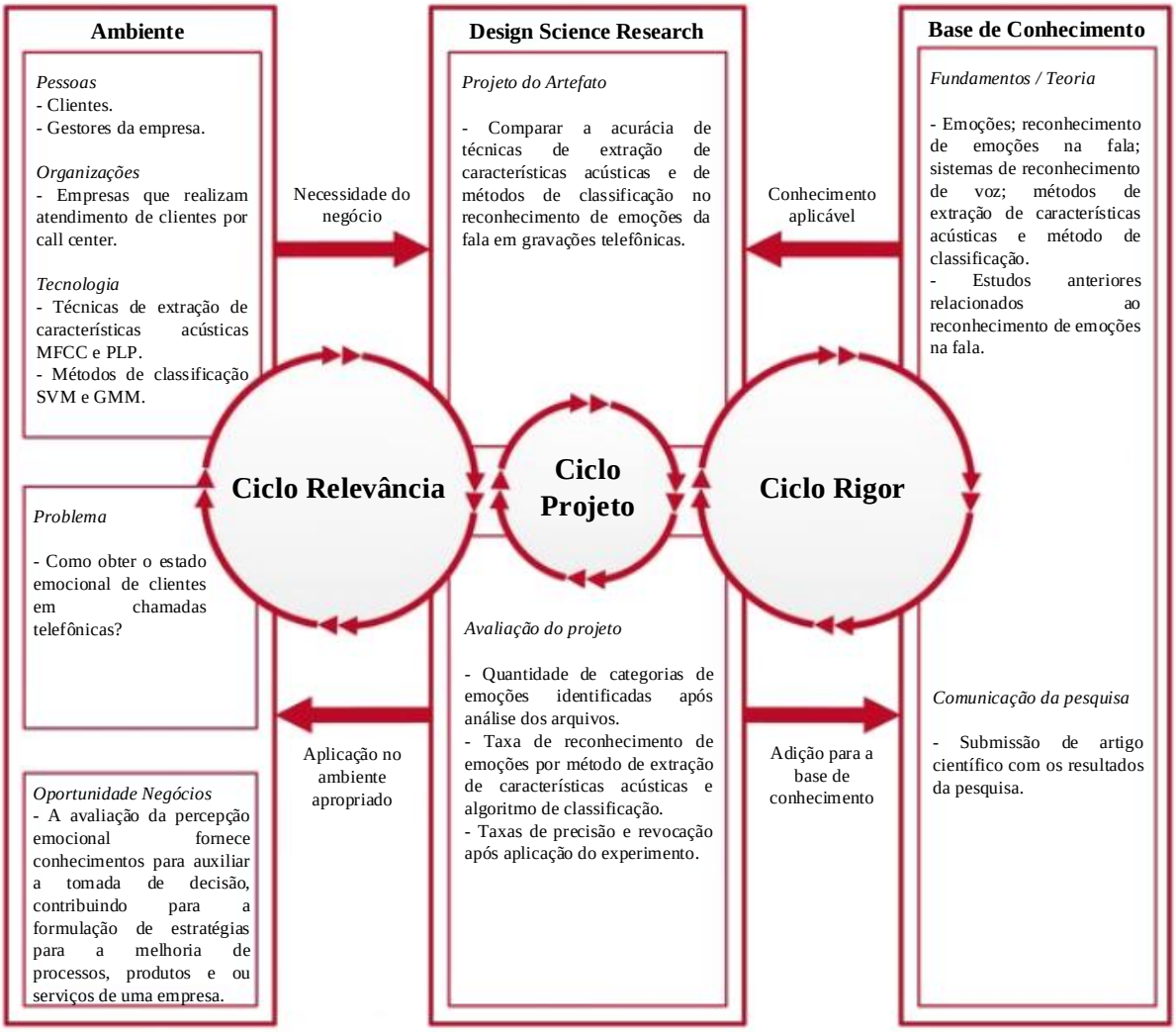

Figura 13. Abordagem de pesquisa científica baseada em Design Science Research.
Fonte: Adaptada de (Hevner, March, Park, & Ram, 2004).

3.2.1 Projeto de artefato

O projeto de artefato diz respeito à compreensão do problema que remete a análise a partir de uma perspectiva ampla, cujo pensamento sistêmico pode trazer contribuições relevantes. O principal resultado dessa etapa é a definição e a formalização do problema a ser solucionado, suas fronteiras (ambiente externo) e as soluções satisfatórias necessárias

(Hevner, March, Park, & Ram, 2004). Nesta etapa devem ser desenvolvidas uma, ou mais, alternativas de artefato para a solução dos problemas (Manson, 2006). Por consequência, o resultado é um conjunto de possíveis artefatos e a escolha de um, ou mais, para serem desenvolvidos.

O problema que motivou a realização desta pesquisa foi: *Como reconhecer emoções em gravações telefônicas?* E as perspectivas para solução eram: 1) *A transcrição de arquivos de áudio em texto e a aplicação de algoritmos de classificação para obtenção de trechos que representariam as emoções*; e 2) *Aplicação de algoritmos de extração de características acústicas em arquivos de áudio e o uso técnicas de classificação para obtenção de estados emocionais dos diálogos*. Dentre as perspectivas, optou-se pela segunda, pois a transcrição do áudio provoca a perda de informações relevantes, como, o tom da voz, a velocidade da fala, e outras, as quais são imprescindíveis para a classificação de emoções. Como produto final dessa pesquisa propõe-se a construção de um protótipo de sistema de reconhecimento de emoções da fala aplicáveis em gravações telefônicas.

3.2.2 Relevância do problema

A relevância do problema é a capacidade do conhecimento acadêmico, baseado em tecnologia, em proporcionar impacto no âmbito prático, ou seja, os resultados devem ser aplicáveis na solução de problemas relevantes (Hevner, March, Park, & Ram, 2004).

O problema de negócio que motivou esta pesquisa foi "*Como obter o estado emocional de clientes em chamadas telefônicas?*". A intenção é que o resultado corrobore na construção de um sistema de reconhecimento de emoções da fala, que possa ser aplicado em *call center*.

3.2.3 Avaliação do projeto

A avaliação do projeto é o processo rigoroso para verificação do artefato no ambiente para o qual foi concebido em relação às soluções que se propõem alcançar. O critério para validação do artefato se fundamenta na filosofia pragmática (Worren, Moore, & Elliott, 2002): proposições explícitas e causais; definição de regras para validação do artefato; e declarações explícitas de como os resultados são criados.

As métricas definidas para avaliação do artefato (projeto e experimento) nesta pesquisa foram: em aspectos qualitativos para o projeto: 1) *Definição da abrangência e profundidade da revisão de literatura*; 2) *Abrangência e profundidade da fundamentação teórica quanto aos tópicos "emoções, reconhecimento de emoções na fala, reconhecimento*

de voz e sistema de reconhecimento de voz". Nos aspectos quantitativos para o experimento: 1) *Quantidade de categorias de emoções identificadas após análise dos arquivos*; 2) *Taxa de reconhecimento de emoções por método de extração de características acústicas e algoritmo de classificação;* e 3) *Taxas de precisão e revocação após aplicação do experimento*.

3.2.4 Contribuições da pesquisa

A pesquisa deve fornecer contribuições claras e verificáveis para a área, além de demonstrar rigor metodológico para validação do artefato proposto (Hevner, March, Park, & Ram, 2004). A questão final, sustentada nos pilares de novidade, generalidade e importância, para avaliação do artefato é "*Quais são as novidades e contribuições do artefato para solução de problemas nas empresas?"*. Espera-se que o artefato proposto corrobore para a construção de um sistema de reconhecimento de emoções da fala a ser aplicado em *call center*.

Quanto à novidade, entende-se que as gravações telefônicas são fontes de informações pouco exploradas e que seu uso pode propiciar ganhos significativos de conhecimento para os negócios. Quanto à generalidade, entende-se que o artefato é aplicável a várias áreas de negócios. Quanto à importância, entende-se que a avaliação da percepção emocional das pessoas fornece conhecimentos para auxiliar a tomada de decisão, o que, apoiado na tecnologia, permite analisar grandes volumes de dados e pode contribuir para a formulação de estratégias que buscam: a melhoria dos processos, e a evolução de produtos e ou serviços de uma empresa.

3.2.5 Rigor da pesquisa

A pesquisa é baseada na aplicação de métodos rigorosos para a construção e avaliação dos artefatos. Na construção, o rigor é avaliado pela aplicabilidade e generalidade do artefato, sendo obtido por meio do uso eficaz de bases de conhecimento (fundamentos teóricos e metodologias de pesquisa). Na avaliação, o rigor é alcançado por métodos similares aos de avaliação das teorias comportamentais, no entanto, o objetivo principal é determinar como um artefato funciona e não teorizar ou provar algo sobre ele (Hevner, March, Park, & Ram, 2004).

Nesta pesquisa o rigor é aplicado com a revisão da literatura, que buscou elencar os principais métodos e suas características para construção do artefato e na validação com a definição de critérios e métricas de desempenho.

3.2.6 Projeto como processo de pesquisa

O projeto é um processo iterativo para encontrar a solução eficaz para um problema. A sua resolução pode ser vista como a utilização dos meios disponíveis para atingir os fins desejados, ao mesmo tempo em que deve satisfazer as leis existentes. Esses fatores são dependentes do problema e de seu ambiente e, invariavelmente, envolvem criatividade e inovação. Os meios são o conjunto de ações e recursos disponíveis para construir a solução. Os fins representam os objetivos e as restrições da solução. As leis são forças incontroláveis do ambiente. O projeto eficaz requer conhecimento do domínio da aplicação e domínio da solução (Manson, 2006). A tentativa é encontrar a melhor solução, ou constatar, que a solução ideal é irrealista, portanto, estratégias heurísticas devem ser usadas para encontrar soluções viáveis (Wieringa, 2009).

A pesquisa, muitas vezes, simplifica o problema, representando explicitamente apenas um subconjunto relevante de meios, fins e leis. O progresso do projeto é realizado expandindo iterativamente o escopo da pesquisa, ou seja, por meio do realismo, da exatidão dos meios, dos fins e das leis do ambiente. O conjunto de soluções possíveis para o problema é especificado como todos os meios que satisfazem as condições finais para as situações levantadas no projeto. Quando esses podem ser formulados apropriadamente e apresentados matematicamente, técnicas de pesquisas podem ser usadas para determinar a solução ótima para as condições especificadas. Porém, dada a natureza dos projetos de sistemas de informação, às vezes pode não ser possível representar claramente, os meios, os fins ou as leis relevantes para o projeto.

Nesta pesquisa os meios são: (i) a revisão de literatura; (ii) a identificação dos métodos de extração de características acústicas e de classificação; e (iii) as ferramentas para construção do artefato.

Os fins são: (i) a comparação entre técnicas de extração de características acústicas e (ii) métodos de classificação para o reconhecimento de emoções na fala.

As leis são as forças do ambiente, como a autorização para condução do experimento com conjunto de áudios privados.

3.2.7 Comunicação da pesquisa

A pesquisa deve contribuir para construção de conhecimento (Hevner, March, Park, & Ram, 2004). Assim, os resultados devem ser apresentados aos públicos orientados à tecnologia e à gestão. A divulgação proporciona o desenvolvimento de novos trabalhos sobre

o tema para que os profissionais técnicos implementem a solução em seu ambiente trabalho e para que os gestores decidam se a solução é apropriada ou não para suas empresas e negócios.

Nesta pesquisa a divulgação se deu por meio da submissão de artigo científico, que descreve os métodos utilizados e os resultados alcançados.

3.3 Base de dados

A aplicação de algoritmos de aprendizagem de máquina exige um conjunto de dados significativamente representativos para o problema proposto. Essa pesquisa utilizou um conjunto de gravações telefônicas obtidas de um provedor de internet, que se enquadra na categoria de Prestadoras de Pequeno Porte (PPP) da Anatel e atende Belo Horizonte e região metropolitana.

As gravações telefônicas foram extraídas do sistema de *Private Automatic Branch Exchange* (PABX) da Empresa[2], no período de 01/09/2019 a 30/12/2019. Foram compostas por 100 horas de chamadas telefônicas, com duração média de dois minutos e trinta segundos de áudio. O idioma dos áudios é português e o público foram clientes e funcionários da empresa, dos gêneros masculino e feminino e de várias idades. Os arquivos de áudio foram gravados no formato estéreo, no codec G.711 Alaw, a uma taxa de 16 bits e amostragem de 8.000 Hz, e possuem no mínimo dois locutores. São características dos arquivos, a alta incidência de ruído ambiental, baixa qualidade na gravação e a presença de voz artificial (máquina anunciadora e música em espera).

3.4 Procedimentos para construção do artefato

A Figura 14 apresenta os procedimentos para condução do experimento de reconhecimento de emoções na fala. Nesta, as tarefas em comum são compartilhadas entre as etapas de treinamento e classificação, as quais serão detalhadas a seguir.

[2] Este pesquisador possuí autorização da empresa, a qual solicitou a não divulgação de seu nome, para uso das gravações telefônicas em sua pesquisa, porém os arquivos de áudio não podem ser divulgados.

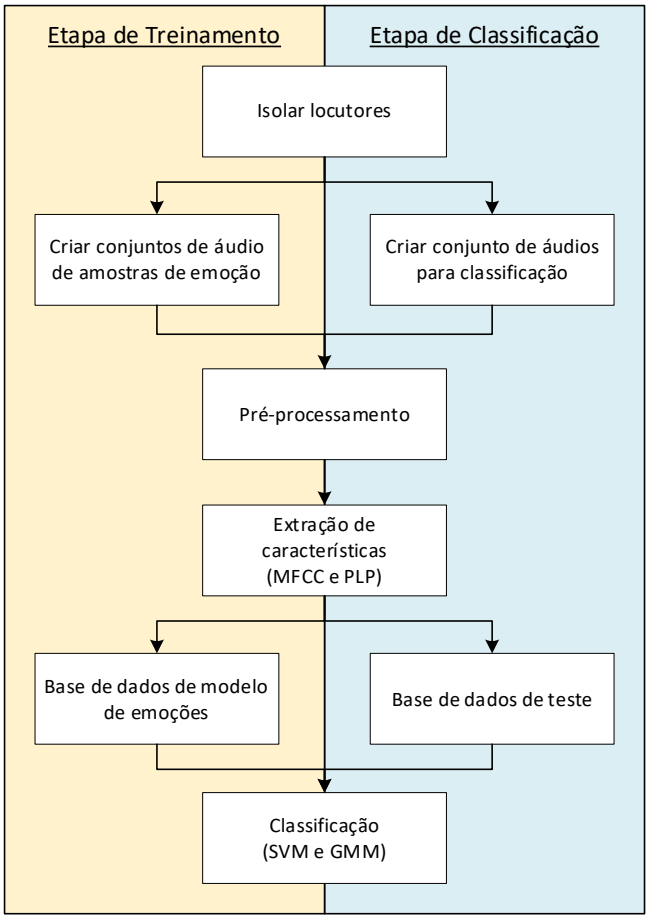

Figura 14. Processo de reconhecimento de emoções na fala.
Fonte: elaborado pelo autor.

3.4.1 Isolar locutores

O isolamento de locutores é realizado para treinamento e classificação. Essa etapa consistiu na divisão dos arquivos de áudio em dois canais, um representando o atendente e outro o cliente. O motivo da aplicação desta técnica é que um dos objetivos da pesquisa é a classificação dos estados emocionais com foco no cliente. Assim, foi removido o canal que representa o atendente e consequentemente, as vozes artificiais e ruídos oriundos do *call center*. Vozes artificiais oriundas do cliente (caixa postal) foram removidas na etapa de criação dos conjuntos de áudio. Para implementação das rotinas desta etapa será utilizada a linguagem de programação Python e a biblioteca: *Pydub*, que possibilita a manipulação de arquivos de áudio.

3.4.2 Conjunto de amostras de emoções e áudios para classificação

A criação dos conjuntos de áudio é especifica para cada etapa. Para criação do conjunto de treinamento os arquivos foram ouvidos integramente e realizado a extração de trechos de áudio que representem os tipos básicos de emoções. Por se tratar de áudio natural

não foi possível determinar antecipadamente quais tipos de emoção seriam encontradas, porém, por conhecimento prévio da base, partiu-se do princípio que existam as seguintes categorias de emoções: **alegria, calam, surpresa, tristeza e raiva**. A extração do trecho de áudio que representa a emoção foi realizada com o *software* Audacity, que permitiu a manipulação de um arquivo de áudio de forma rápida e intuitiva. Após extração, os trechos identificados foram armazenados em diretórios que representem as categorias encontradas. Para criação do conjunto de classificação foram excluídos os arquivos de áudio que possuíam apenas vozes artificiais (chamada atendida pela caixa postal) ou que continham apenas silêncio (chamada que houve o atendimento, mas não houve áudio) e os demais foram armazenados em um único diretório.

3.4.3 Pré-processamento

O pré-processamento foi realizado para os arquivos de treinamento e classificação. Essa etapa visou a obtenção de um sinal acústico próximo ao da voz e consistiu na remoção de ruídos, trechos de silêncio, janelamento e normalização dos arquivos de áudio. As rotinas de pré-processamento foram implementadas na linguagem de programação Python e utilizou as bibliotecas: *Pydub* – para leitura dos arquivos de áudio, *Numpy* – para transformação dos arquivos em vetores numéricos e *Scipy* – para aplicação das técnicas de janelamento, por meio do módulo *Signal Processing (scipy.signal)* e método *get_window*, e transformada de Fourier, por meio do módulo *Fast Fourier Transforms* (*scipy.fft*) e método *fft*.

3.4.4 Extração de características (MFCC e PLP)

A extração de características foi aplicada para os arquivos de treinamento e classificação. Essa etapa antecedeu a criação dos modelos e visou a obtenção das caraterísticas espectrais dos arquivos de áudio. Foram aplicados os algoritmos de MFCC e PLP para criação dos conjuntos de emoções. As rotinas desta etapa foram implementadas na linguagem Python e utilizaram a biblioteca *Simplified Python Audio-Features Extraction*, que permitiu a aplicação dos algoritmos de extração de caraterísticas acústicas: MFCC e PLP e forneceu módulos para aplicação dos bancos de filtros Mel e Bark.

3.4.5 Base de dados de emoções e teste

Após a extração das características foi utilizada a biblioteca a *Scikit-learn* em Python, para criação dos modelos de treinamento e teste. Para cada conjunto de emoção identificada

foi gerado um modelo de treinamento e por fim, gerado o modelo genérico de teste, com todos os arquivos de áudio.

3.4.6 Classificação (GMM e SVM)

A classificação foi realizada utilizando a biblioteca *Scikit-learn* e os módulo de aprendizagem de máquina GMM e SVM. A classificação foi realizada no modo multiclasse, no qual os modelos de treinamento de emoção foram comparados ao modelo genérico de teste. Esse teste visou comparar o desempenho dos classificadores com múltiplos tipos de emoção.

3.5 Métricas de desempenho

Trabalhos que envolvam comparações de qualquer natureza requerem a definição de qual é o sujeito de avaliação, bem como sob quais óticas este será avaliado. No contexto de reconhecimento de emoções é desejável que o sistema diferencie as locuções que contenham emoções similares, caso contrário poderá provocar erros na interpretação do conteúdo analisado. Quando o sistema falha neste tipo de classificação, diz-se que foi obtido um resultado falso positivo. Também se deseja que o sistema identifique e classifique a maior quantidade de locuções possíveis. Quando o sistema falha neste tipo de classificação, diz-se que foi obtido um resultado falso negativo.

A partir dessas definições foram propostas as métricas de precisão e revocação para avaliação dos modelos de aprendizagem de máquina, que correspondem às ausências de resultados falso-positivos e falso-negativos. Nessa pesquisa a precisão e a revocação foram calculadas para o cenário multiclasse. Adicionalmente foi aplicado a medida “F”, ou média harmônica, que visou calcular se um grupo contém todos os elementos de uma classe e apenas os elementos dessa classe, por intermédio da média harmônica ponderada dos valores de revocação e precisão.

3.5.1 Precisão

Pode ser vista como uma medida de exatidão, pois denota a ausência de falsos positivos, retornando a porcentagem de amostras rotuladas pelo classificador como c_i em relação a todas as amostras de teste que foram rotuladas pelo classificador como c_i.

$$Precisão\ c_i = \frac{VP_{ci}}{VP_{ci} + FP_{ci}}$$

Onde:

- VP_{ci} (Verdadeiros positivos) representa o número de amostras de teste que foram corretamente atribuídos à classe c_i;
- FP_{ci} (Falsos positivos) representa o número de amostras de teste de uma classe c_j ($c_j \neq c_i$), mas atribuídos à classe c_i.

3.5.2 Revocação

Pode ser vista como uma medida de completude, pois denota a ausência de falsos negativos retornando a porcentagem de amostras rotuladas pelo classificador como c_i em relação a todos as amostras de teste que pertencem à classe c_i.

$$Revocação\ c_i = \frac{VP_{ci}}{VP_{ci} + FN_{ci}}$$

Onde:

- VP_{ci} (Verdadeiros positivos) representa o número de amostras de teste que foram corretamente atribuídas à classe c_i;
- FN_{ci} (Falsos negativos) representa o número de amostras de teste da classe c_i mas rotuladas com uma classe c_j ($c_j \neq c_i$).

3.5.3 Medida F ou Média Harmônica

Na análise estatística da classificação, o escore "F" é uma medida que considera tanto a precisão como a revocação de um teste para calcular sua pontuação. É dada pela média harmônica entre a precisão e a revocação e atinge seu melhor e pior valor em 1 e 0, respectivamente.

$$F = 2 * \left(\frac{Precisão * Revoceção}{Precisão + Revocação}\right)$$

3.6 Síntese do capítulo

Foi abordada a metodologia que fundamenta a execução da pesquisa, os procedimentos para construção do artefato e as métricas para avaliação de desempenho do experimento. Os principais tópicos foram:

- Classificação da pesquisa: A pesquisa foi caracterizada como sendo do tipo exploratória, de abordagem quantitativa, natureza aplicada e procedimento

experimental. Para condução da pesquisa foi utilizada a metodologia *Design Science* e método *Design Science Research.* A amostra foi composta 100 horas de gravações de chamadas telefônicas e foi aplicada as técnicas de extração das características acústicas e classificação para obtenção do estado emocional dos enunciados.

- *Design Science Research*: é uma estratégia de pesquisa que a compreensão e solução de um problema é alcançada por meio da construção e aplicação de um artefato. Os artefatos são as representações simbólicas materializáveis, ou seja, construtos (entidades e relações), modelos (abstrações e representações), métodos (algoritmos e práticas) e instanciações (implementação de sistemas e protótipos) (March & Smith, 1995; Hevner, March, Park, & Ram, 2004). Pesquisas conduzidas pela metodologia DSR podem ser avaliadas pelo binômio rigor e a relevância. Na pesquisa a relevância e o rigor serão alcançados por meio das seguintes diretrizes: projeto de artefato; relevância do problema; avaliação do projeto; contribuições da pesquisa; rigor da pesquisa; projeto como processo de pesquisa; e comunicação da pesquisa (Hevner, March, Park, & Ram, 2004).
- Procedimento para construção do artefato: para construção do artefato foram executados os seguintes procedimentos: a) isolar os locutores: tem como objetivo a segmentação do arquivo de áudio em dois canais, que possibilitará a avaliação da parte que representa o cliente; b) criação dos conjuntos de treinamento e teste: tem como objetivo a identificação dos trechos de áudio que representem emoções e a rotulação em categorias emoções; c) pré-processamento: tem como objetivo a remoção de trechos de ruídos e silêncio para redução do custo computacional na análise dos dados; d) extração de características: visa a obtenção das caraterísticas espectrais dos arquivos de áudio; e) base de dados de emoções e teste: visa a criação dos modelos de treinamento e teste para a classificação; f) classificação: visa a aplicação de algoritmos de aprendizado de máquina para segmentação dos dados em classes de emoção.
- Métricas de desempenho: são indicadores que visam verificar se o resultado esperado foi atingido e identificar pontos de melhoria na aplicação. Para avaliação deste artefato as seguintes métricas foram propostas: a) precisão; b) revocação e c) medida “F” ou média harmônica.

4 REVISÃO DE LITERATURA

Para execução dessa pesquisa a literatura foi revisitada por meio da revisão sistemática.

4.1 Protocolo da revisão de literatura

4.1.1 Objetivo

Identificar na literatura a existência de estudos que proponham métodos para reconhecimento de emoções na fala e critérios para avaliação de desempenho dos métodos.

4.1.2 Questão de pesquisa

Quais métodos e técnicas foram propostas para reconhecimento de emoções na fala?

4.1.3 Critérios para seleção de fontes

Para seleção de fontes, considerou-se que deveriam estar disponíveis na internet, em bases de dados cientificas, no período de 01/2000 até 12/2019. Caso os trabalhos estivessem disponíveis em outros meios, mas que atendesse a critérios científicos, poderiam ser incluídos.

4.1.4 Métodos de busca de fontes

A pesquisa foi realizada pelo descritor (*Speech Emotion Recognition*) nos campos título e palavras-chaves determinadas pelos autores, em bases de dados de artigos científicos ou anais de eventos. Trabalhos de conclusão de curso de pós-graduação (mestrado e doutorado) também foram consultados.

4.1.5 Palavras-chaves

Foi usada a palavra-chave *Speech Emotion Recognition* (Reconhecimento de Emoções na Fala). O motivo do uso do termo em inglês é a maior existência de publicações em bases de dados internacionais.

4.1.6 Listagem de fontes

Foram realizadas pesquisas nas bases de dados: IEEE[3] e ScienceDirect[4].

4.1.7 Critérios de Inclusão dos trabalhos

Para inclusão dos trabalhos, deveriam ser completos, publicados em bases de dados cientificas e tratar do tema de reconhecimento de emoções na fala.

4.1.8 Critérios de Exclusão dos trabalhos

Foram excluídos os trabalhos que não possuíam experimentos práticos para testar suas hipóteses.

4.1.9 Estratégia para seleção da informação

Foi preenchido um formulário contendo as informações básicas de cada texto válido. As informações básicas eram dados bibliográficos, data de publicação, resumo e considerações do pesquisador que conduziu a revisão de literatura.

4.1.10 Intenção da revisão de literatura

Essa revisão buscou responder as seguintes questões: 1) Qual o estado da arte em reconhecimento de emoções na fala? 2) Quais são as técnicas comumente usadas para extração de características acústicas e classificação de arquivos de áudio? 3) Quais os tipos de bancos de dados acústicos são comumente usados para reconhecimento de emoções na fala?

4.2 Resultado da revisão de literatura

O resultado da revisão de literatura em bases científicas é exibido na Tabela 1. Na base de dados IEEE foram encontrados 365 registros e na base de dados ScienceDirect foram encontrados 110 registros.

Tabela 1. Quantidade de artigos encontrados

Palavra-chave	Base de pesquisa	Resultado
Speech Emotion Recognition	IEEE	365
	ScienceDirect	110

Fonte: elaborado pelo autor.

[3] https://ieeexplore.ieee.org/Xplore/guesthome.jsp
[4] https://www.sciencedirect.com/

Na análise dos artigos foi identificado que os métodos de extração de caraterísticas mais utilizados foram MFCC e o PLP, conforme Tabela 2.

Tabela 2. Métodos de extração de características mais utilizados

Método de extração de características	Resultados (%)
MFCC	25,68
PLP	15,42
LPC	13,59
LPCC	5,78
PNCC	5,55
RPLP	4,05
Outros	29,93

Fonte: elaborado pelo autor.

Quanto aos métodos de classificação, os mais utilizados foram SVM, HMM e GMM, conforme Tabela 3, porém, por decisão do pesquisador, optou-se pela comparação dos métodos SVM e GMM.

Tabela 3. Métodos de classificação mais utilizados

Classificador	Resultados (%)
SVM	15,78
HMM	12,20
GMM	10,80
ANN	9,20
MPL	8,62
k-NN	6,30
Outros	37,10

Fonte: elaborado pelo autor.

Também foi observado um número significativo de trabalhos que utilizavam técnicas de reconhecimento de padrões com redes neurais, porém, por decisão de pesquisa, optou-se na utilização de classificadores com abordagem estatística.

4.3 Relatório da revisão de literatura

Esta seção apresentou a síntese dos trabalhos relacionados a reconhecimento de emoções na fala, com foco nos métodos extração de categorias acústicas e classificadores utilizados.

4.3.1 Base de dados

Diferentes bancos de dados foram usados na validação dos experimentos, entre os quais se destacam os conjuntos: Berlin (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005); e FAU Aibo (Schuller, Steidl, & Batliner, 2009). O banco de dados acústico Berlin é um banco do tipo simulado, no idioma alemão e foi gravado no departamento de acústica técnica da Universidade de Berlim. É composto por vozes de 10 atores, sendo cinco masculinas e cinco femininas. Nele foram registradas cerca de 500 declarações emocionais que expressam os sentimentos de alegria, raiva, medo, indiferença, nojo, tristeza e uma versão neutra. O banco de dados acústico FAU Aibo é do tipo natural, no idioma inglês e consiste de nove horas de fala, de 51 crianças com idades entre 10 a 13 anos. Foi gravado a partir da interação das crianças com o robô de estimação Sony Aibo. Nesse banco as gravações foram segmentadas manualmente em partes sintaticamente significativas usando critérios sintático-prosódicos. Os dados foram anotados em 11 categorias de emoções, por cinco rotuladores humanos. As categorias do banco de dados incluem: felicidade, medo, nojo, raiva, tristeza e uma versão neutra.

4.3.2 Extração de características

Para extração de extração de características, têm sido empregados métodos, como: *Mel-Frequency Cepstral Coefficients*, *Prosodic Features*; *Linear Prediction Cepstral Coefficients*; *Perceptual Linear Predictive;* e outros, com objetivo de extrair o maior número de características para o reconhecimento de emoções humanas. Porém, observou-se que os mais utilizados são MFCC, LPC e PLP.

4.3.3 Abordagens de classificação

Para classificação das emoções, têm sido empregados métodos, como: *Support Vector Machine* (SVM); *Hidden Markov Model* (HMM); *Neural Network* (NN); *k-Nearest Neighbors* (k-NN); *Gaussian Mixture Model* (GMM); e outros. Porém, observou-se que para modelagem estática, os mais utilizados são SVM, HMM e GMM.

4.3.4 Trabalhos analisados

Nas últimas décadas pesquisas foram desenvolvidas com o intuito de aplicar métodos estatísticos para o reconhecimento de emoções na fala.

O uso do método *Log Frequency Power Coefficients* (LFPC) para representar os sinais de voz e o método *Hidden Markov Model* (HMM) como classificador (Nwe, Foo, &

De Silva, 2003). Neste o conjunto de dados de voz era privado e foi dividido em seis categorias de emoção: raiva, nojo, medo, alegria, tristeza e surpresa. A avaliação da proposta se deu comparando os resultados do método de extração características LFPC aos métodos *Linear Prediction Cepstral Coefficients* (LPCC) e *Mel-Frequency Cepstral Coefficients* (MFCC). O resultado demostrou que o método obteve precisão média de 78% na classificação de emoções.

A abordagem para o uso do método *Gaussian Mixture Vector Autoregressive* (GMVAR), que é uma mistura método *Gaussian Mixture Model* (GMM) com um vetor autorregressivo para resolução de problemas de classificação de reconhecimento de emoção da fala (El Ayadi, Kamel, & Karray, 2007). A principal motivação por trás do uso desse modelo é sua capacidade de modelar a dependência entre os vetores de recursos de fala extraídos, bem como a multimodalidade em sua distribuição. O conjunto de dados emocionais de Berlim foi usado para avaliação do GMVAR. A técnica proposta resultou em uma precisão de classificação de 76% contra 71% para o modelo *Hidden Markov Model* (HMM), 67% para o método *k-Nearest Neighbors* (k-NN), 55% para redes neurais do tipo *FeedForward*.

O modelo multidimensional utilizando emoções primitivas para reconhecimento de emoções de fala (Grimm, Kroschel, Mower, & Narayanan, 2007). As três dimensões foram construídas pela composição de três primitivas, denominadas de: valência, ativação e dominância. Assumiu-se que os valores dessas dimensões estariam na faixa de [-1, +1]. Um método livre de texto e baseado em imagens foi introduzido para avaliar as emoções primitivas e alcançar a melhor concordância entre os avaliadores. Para extrair recursos acústicos como altura, energia, velocidade de fala e características espectrais, foram empregadas regras de estimativa e lógica *fuzzy*. As abordagens foram validadas com dois conjuntos de dados, o EMA e o VAM, que foram gravados em um *talk-show* na TV alemã. Para mapear as primitivas de emoção para certa categoria de emoção, usou-se o classificador k-NN, que atingiu uma taxa de reconhecimento até 83,5%.

O método de florestas de decisão aleatória *Ensemble Random Forest To Trees* (ERFTrees) com um grande número de recursos para reconhecimento de emoção na fala sem referir-se a qualquer idioma ou informação linguística (Rong, Li, & Chen, 2009). Esse método é aplicado em pequenas bases de dados com grande número de recursos. Para avaliar o método proposto, aplicou o experimento em um conjunto de dados no idioma mandarim e os resultados apresentaram uma boa precisão no reconhecimento de emoção. O ERFTrees apresentou um desempenho superior a métodos populares de redução de dimensão, como *Principle Component Analysis* (PCA), *Multi-Dimensional Scaling* (MDS) e o ISOMap. O

método proposto alcançou uma taxa de reconhecimento para um conjunto de dados com vozes femininas de 82,54%.

O novo conjunto de recursos de harmonia para reconhecimento de emoções de fala (Yang & Lugger, 2010). Esses recursos dependem da percepção psicoacústica da teoria musical. A partir do contorno da altura estimada de um sinal de fala calculou-se a autocorrelação circular do histograma de altura, o qual mede a incidência de diferentes intervalos de dois tons, que causam uma impressão harmônica ou inarmônica. Na etapa de classificação, o classificador bayesiano usou uma regra de verossimilhança condicional de classe gaussiana. O resultado experimental no banco de dados de emoções de Berlim indicou uma melhoria no desempenho de reconhecimento.

Características espectrais de sinais emocionais para determinar emoções e caracterizar grupos (Albornoz, Milone, & Rufiner, 2011). Nessa pesquisa emoções foram agrupadas com base em características acústicas. Diferentes classificadores, como HMM, GMM e MLP, foram avaliados com configurações e recursos de entrada distintos para projetar novas técnicas hierárquicas para classificação de emoções. O resultado experimental no conjunto de dados de Berlim demostrou que a abordagem hierárquica atingiu o melhor desempenho em comparação aos classificadores padrões. Por exemplo, o desempenho do método HMM padrão atingiu 68,57%, enquanto no modelo hierárquico atingiu 71,75%.

A estrutura computacional hierárquica para reconhecer emoções (Lee, Mower, Busso, Lee, & Narayanan, 2011). A estrutura proposta mapeia um sinal de voz de entrada em uma das várias classes de emoção por meio de camadas subsequentes de classificações binárias. O principal conceito é resolver a tarefa de classificação da maneira mais fácil para diminuir a propagação do erro. Os bancos de dados AIBO e USC IEMOCAP foram empregados para avaliar o método de classificação. Sobre o modelo de linha de base SVM, o resultado absoluto apresentou uma melhora na precisão de 72,44% e 89,58%, em respectivos bancos de dados. O experimento demostrou que o método hierárquico relatado é eficiente para classificar o discurso emocional nos bancos de dados AIBO e USC IEMOCAP.

O método baseado em segmentos para reconhecimento de emoção na fala no idioma mandarim (Yeh, Pao, Lin, Tsai, & Chen, 2011). Nessa abordagem, a unidade de reconhecimento não é uma frase ou sentença, mas uma expressão emocional no diálogo. Essa foi realizada por meio dos seguintes passos. Primeiro, avaliou o desempenho de classificadores de reconhecimento de emoções de fala em frases curtas. Os resultados dos experimentos mostram que o classificador *Weighted Distance K-nearest neighbor* (WD-KNN) atingiu a melhor precisão para o reconhecimento de emoções em 5 classes do que as

cinco técnicas de classificação selecionadas. Em seguida, implementaram um sistema contínuo de reconhecimento de emoções de fala no idioma mandarim com um gráfico de radar de emoções baseado em WD-KNN; este sistema representou a intensidade de cada componente da emoção na fala. Esta abordagem mostrou como as emoções podem ser reconhecidas por sinais de fala e, por sua vez, como os estados emocionais podem ser visualizados.

O método para o reconhecimento de emoções da fala baseada em múltiplos classificadores utilizando a informação *Acoustic-Prosodic* (AP) e *Semantic Labels* (SL) (Wu & Liang, 2011). Nesse método de fusão, primeiro os recursos de AP foram extraídos e, em seguida, três tipos diferentes de classificadores – GMM, SVM e MLP – foram adotados como classificadores de nível básico. A *Meta Decision Tree* (MDT) foi então empregada para a fusão do classificador para obter a confiança de reconhecimento de emoção baseada em AP. Para reconhecimento baseado em SL, rótulos semânticos derivados de uma base de conhecimento em mandarim chamada HowNet foram usados para extrair automaticamente regras de associação de emoções das sequências de palavras reconhecidas. O modelo de entropia máxima foi usado para caracterizar a relação entre os estados emocionais e as regras de associação de emoções. Um método de fusão ponderada de produto foi usado para integrar os resultados de reconhecimento baseados em AP e SL para a decisão emocional final. Os resultados experimentais independentes do locutor revelam que o desempenho de reconhecimento de emoção com base na MDT atingiu 80,00% de reconhecimento. E por fim, que a combinação de informações das técnicas de AP e SL atingiu 83,55% de reconhecimento de emoções.

O método de reconhecimento de emoções da fala humana com o uso de *Modulation Spectral Features* (MSF) (Wu, Falk, & Chan, 2011). As características foram extraídas de um espectro-temporal de longo prazo inspirado na audição, utilizando um banco de filtros de modulação e um banco de filtros auditivos para decomposição da fala. Esse método obteve componentes de frequência acústica e frequência de modulação temporal para transmitir dados importantes que faltavam nas características espectrais de curto prazo tradicionais. Para o processo de classificação, foram adotados SVM com *Radial Basis Function* (RBF). Os bancos de dados Berlin e VAM foram empregados para avaliar o MSF. O resultado do experimento, demonstrou que o MSF possuí um desempenho promissor em comparação com MFCC e *Perceptual Linear Prediction Coefficients* (PLPC). Quando o MSF utilizou recursos prosódicos aumentados, houve uma melhoria considerável no desempenho do reconhecimento. Além disso, a taxa de reconhecimento geral alcançada para a classificação foi de 91,6%.

Melhorar o reconhecimento de emoções de fala de locutores independentes com uso de um método de reconhecimento de emoções em três níveis (Chen, Mao, Xue, & Cheng, 2012). Esse método classificou as emoções de rude a cortês e selecionou o recurso apropriado usando o método de Fisher. A razão Fisher foi utilizada como parâmetro de entrada para o classificador SVM. Empregou as técnicas: *Principal Component Analysis* (PCA) e a *Artificial Neural Network* (ANN) para reduzir a dimensionalidade. Foram realizados experimentos usando o conjunto de dados de fala emocional da universidade Beihang BHUDES. Quatro experimentos comparativos foram realizados, que incluíram: Fisher + SVM, PCA + SVM, Fisher + ANN e PCA + ANN. Como resultado obteve que na redução de dimensão, o método de Fisher foi superior ao método PCA e para classificação o SVM foi mais expansível que a ANN. As taxas de reconhecimento para os três níveis foram 86,5%, 68,5% e 50,2%.

O uso de modelos de referência neutros para detectar proeminências emocionais locais na frequência fundamental (Arias, Busso, & Yoma, 2014). Foi apresentada uma nova abordagem baseada na *Functional Data Analysis* (FDA), que visa capturar a variabilidade intrínseca dos contornos F0. Para um determinado contorno F0, *The Principal Component Analysis* (PCA) são calculados para usar como recursos para reconhecimento de emoção de fala. O resultado indica que a acurácia da abordagem proposta atingiu 75,8% na classificação binária. Isso significa 6,2% mais alto do que o sistema de *benchmark* treinado com estática F0 geral. A abordagem é avaliada pelo conjunto de dados SEMAINE. Os resultados indicam que o método pode ser eficaz se implementado em uma aplicação real.

O método de classificação SVM que sintetizava as informações sobre o reconhecimento de emoções para resolver o problema de classificação binária (Cao, Verma, & Nenkova, 2015). O método proposto instrui o algoritmo SVM para emoções específicas, tratando os dados de cada locutor como uma consulta distinta e, em seguida, mescla as previsões dos classificadores para aplicar a previsão multiclasse. Essa classificação possuí duas vantagens: (i) a primeira, para as etapas de treinamento e teste, pois obtém dados específicos de cada locutor; (ii) e a segunda, que considera que cada locutor pode expressar uma mistura de emoções para reconhecer a emoção dominante. Essa abordagem de classificação alcançou ganhos substanciais em termos de precisão em comparação ao SVM convencional. Foram realizados experimentos utilizando os conjuntos de dados públicos Berlin e LDC. Em ambos, tanto nos dados representados quanto nos dados espontâneos, que compreendem declarações emocionais neutras e intensas, a SVM baseada em classificação obteve maior precisão de reconhecimento de declarações emocionais do que os métodos de SVM convencionais. A precisão de equilíbrio alcançou 44,4%.

A abordagem computacional para o reconhecimento da emoção e análise das especificações da emoção em mídias sociais expressas como Wechat (Dai, Han, Dai, & Xu, 2015). Essa abordagem aproxima a emoção mista e as flutuações dinâmicas na *Position Arousal Dominance* (PAD), extraindo 25 características acústicas dos sinais de fala e empregando o modelo de treinamento *Least Squares-Support Vector Regression* (LV-SVR). Os resultados experimentais demonstram que a taxa média de reconhecimento de emoções foi de 82,43%.

Reconhecimento de emoção humana baseado na fala usando MFCC (Likitha, Gupta, Hasitha, & Raju, 2017). Essa abordagem utiliza o método MFCC e comparação estáticas para reconhecimento de emoções na fala. O banco de dados continha vozes de 60 pessoas com diferentes tipos de emoções. O início do processo se deu com a rotulação das amostras em categorias de emoção. Posteriormente, os sinais de voz foram analisados através da função de "*wavread*" do *software* Matlab, onde foram aplicadas as técnicas de janelamento de "Hamming", transformada Rápida de Fourier e a conversão da amostra para à escala de Mel, em que se obtiveram os coeficientes cepstrais de frequência Mel (MFCC). Em seguida foram obtidos o valor médio do MFCC e o desvio padrão do valor médio. Esses parâmetros foram usados para comparação binária entre os tipos de emoção encontrados. O método apresentou uma eficiência de 80% no reconhecimento de emoções.

Detectando depressão na fala: Comparação e combinação entre diferentes tipos de fala (Long, et al., 2017). Essa abordagem utiliza múltiplos métodos de extração de características acústicas (ZCR, MFCC, LPC e PLP) e o classificador SVM para detectar desordem depressiva. O banco de dados do experimento era composto por 74 amostras, sendo 37 de pacientes deprimidos e 37 de pacientes saudáveis, no idioma mandarim. No experimento foi examinado o poder discriminativo da fala em diferentes tipos de situações, como: leitura de textos, descrição de imagens e entrevista. Na leitura os participantes eram incentivados a lerem trechos de três artigos, em mandarim, que representavam trechos de emoções. Na descrição de imagens os participantes deveriam descrever as imagens apresentadas, classificando como positiva, negativa ou neutra. Nas entrevistas os participantes eram incentivados a responderem questões previamente formuladas. O experimento consistiu na aplicação dos métodos de extração de características nos áudios e a classificação pelo método SVM. Os resultados demostraram que os métodos de extração características testados foram robustos e conferiram alta precisão nos diferentes tipos de situações. Por fim, a abordagem apresentou acurácia de 78,02% no reconhecimento de emoções que representavam a depressão.

A Figura 15 apresenta a síntese dos trabalhos analisados. Quanto aos métodos de extração de características, observa-se o uso de técnicas de análise de curto e longo prazo e técnicas de análise espectral. Dentre as técnicas de análise espectral destacam-se os métodos MFCC, utilizado em sete trabalhos e PLP e seus derivados, utilizados em três trabalhos. Quanto aos métodos de classificação, observa-se o uso de técnicas de abordagem supervisionada e não supervisionada, com destaque para os métodos SVM, utilizado em seis trabalhos, GMM e seus derivados, utilizados em quatro trabalhos e HMM e k-NN, utilizados em três trabalhos.

Autor	**Método de extração de caraterística**	**Método de classificação**
(Nwe, Foo, & De Silva, 2003)	LFPC e MFCC	HMM
(El Ayadi, Kamel, & Karray, 2007)	MFCC	GMVAR, HMM, k-NN e FeedForward
(Grimm, Kroschel, Mower, & Narayanan, 2007)	Regras de estimativa e lógica fuzzy	k-NN
(Rong, Li, & Chen, 2009)	Características linguísticas, espectrais e contornos vocais	Arvore de decisão e floresta randômica
(Yang & Lugger, 2010)	Energia, estatísticas de *pitch*, estatísticas de duração, formante e ZCR	Classificador bayesiano
(Albornoz, Milone, & Rufiner, 2011)	MLS, MFCC e características prosódicas	HMM, GMM, MPL e modelo hierárquico
(Lee, Mower, Busso, Lee, & Narayanan, 2011).	Energia quadrada média, tom, relação harmônico-ruído e MFF	SVM
(Yeh, Pao, Lin, Tsai, & Chen, 2011).	Jitter, LPC, LPCC, MFCC, LFPC, PLP e Rasta-PLP	k-NN
(Wu & Liang, 2011)	AP e SL	GMM, SVM, MLP e MDT
(Wu, Falk, & Chan, 2011)	MFCC, PLPC e MSF	SVM e RBF
(Chen, Mao, Xue, & Cheng, 2012)	Energia, *pitch*, frequência de corte do espectro, densidade de correlação e MFF	SVM e ANN
(Arias, Busso, & Yoma, 2014)	FDA e PCA	Abordagem estatística
(Cao, Verma, & Nenkova, 2015)	Características prosódicas e espectrais	SVM
(Dai, Han, Dai, & Xu, 2015)	Energia de curto prazo, *pitch e* ZRC	LV-SVR
(Likitha, Gupta, Hasitha, & Raju, 2017)	MFCC	Abordagem estatística
(Long, et al., 2017)	ZCR, MFCC, LPC e PLP	SVM

Figura 15. Síntese dos trabalhos analisados
Fonte: elaborado pelo autor

4.4 Síntese do capítulo

Foram apresentados os procedimentos para condução da revisão de literatura, a intensão da revisão de literatura, os resultados da pesquisa em bases científicas e o relatório da revisão de literatura.

As questões apresentadas como intenção da revisão de literatura: 1) Qual o estado da arte em reconhecimento de emoções na fala? 2) Quais são as técnicas comumente usadas para extração de características acústicas e classificação de arquivos de áudio? – foram respondidas no tópico 4.2, que apresentou a quantidade de trabalhos encontrados em bases de dados científicas e os métodos mais utilizados para extração de características acústicas e classificação. Quanto à terceira pergunta, 3) Quais os tipos de bancos de dados acústicos são comumente usados para reconhecimento de emoções na fala? – foi respondida no tópico 4.3, onde são apresentadas as bases de dados comumente utilizadas nos trabalhos de reconhecimento de emoções da fala, as abordagens de extração de características e classificação, e a síntese dos trabalhos analisados, onde destacam-se o uso dos métodos MFCC e PLP e dos classificadores SVM e GMM.

5 IMPLEMENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Nos capítulos anteriores foram apresentados os conceitos associados ao estudo de reconhecimento de emoções na fala, às convenções e aos estudos encontrados na literatura. Neste capítulo foi apresentado os aspectos da implementação do artefato e dos resultados das simulações.

5.1 Implementação

A seguir, foi apresentado de forma macro as principais etapas da implementação. O código completo foi disponibilizado no Apêndice A.

5.1.1 Conjunto de treinamento e teste

Para criação dos conjuntos de treinamento e teste, 3 mil arquivos de áudio foram analisados e rotulados em categorias emoções. Na análise, os trechos que representavam emoções foram recortados e separados em diretórios correspondentes a cada categoria encontrada. Durante a análise, foram identificadas as seguintes categorias de emoções: **alegria**, **calma**, **raiva**, **surpresa** e **tristeza**. A validação do modelo se deu pela aplicação da técnica de *cross-validation*, cujos 80% dos arquivos foram utilizados para treinamento e 20% reservados para teste. A Tabela 4 apresentou a quantidade de amostras coletadas para criação do conjunto de treinamento e a

Tabela 5 apresentou a quantidade de amostras coletadas para a criação do conjunto de teste.

Tabela 4. Amostras do conjunto de treinamento

Categoria	**Quantidade de Amostras**
Alegria	470
Calma	500
Raiva	480
Surpresa	475
Tristeza	475
Total de Amostras	**2.400**

Fonte: Elaborado pelo autor.

Tabela 5. Amostras do conjunto de teste

Categoria	Quantidade de Amostras
Alegria	120
Calma	120
Raiva	120
Surpresa	120
Tristeza	120
Total de Amostras	**600**

Fonte: Elaborado pelo autor.

5.1.2 Extração de características

Nesta etapa, foram utilizados os métodos de extração de características MFCC e PLP para ressaltar os aspectos de informação que contribuíam para identificação dos diferentes tipos de emoção.

Para comparação das técnicas de extração de características foram selecionados um conjunto de 25 coeficientes de curto prazo. A seleção de 25 coeficientes se deu em razão do custo computacional para execução do experimento e o desempenho na classificação. Nas simulações, observou-se que um baixo número de coeficientes não apresentaram bons resultados na classificação. e o um alto número de coeficientes não resultaram em ganhos significativos na classificação, porém, aumentavam o custo computacional para execução do experimento.

Dado aos resultados dos coeficientes, foi aplicada a técnica *Principal Component Analysis* (PCA). O PCA foi uma técnica de análise multivariada que visou analisar a inter-relação entre um grande número de amostras e reduzir sua dimensão, preservando suas características principais. Como o experimento continha arquivos de áudio de durações variadas, foi necessário aplicar a técnica para redução de dimensão e preservação das características das amostras.

A implementação dos procedimentos foi realizada com as bibliotecas *Spafe* (*Simplified Python Audio Features Extraction*), para extração das características acústicas MFCC e PLP. A seguir, foi descrito brevemente os procedimentos realizados para extração das características acústicas.

5.1.2.1 Mel-Frequency Cepstral Coefficients (MFCC)

A técnica de extração de atributos *Mel-Frequency Cepstral Coefficients* (MFCC) (Milner & Shao, 2006) faz uma análise de características espectrais de curto prazo, baseando-se no uso do espectro da voz convertido para uma escala de frequências denominada MEL que é uma escala que visa imitar as características perceptíveis do ouvido humano. Estes coeficientes são uma representação definida como cepestro de um sinal

janelado no tempo, que tem sido derivado da aplicação da DFT, em escalas de frequência não lineares.

Para extração dos vetores de características MFCC, foi aplicada a técnica descrita na seção 2.2.2.3.1 e explicada brevemente a seguir:

- O sinal de voz é passado através do filtro de pré-ênfase com $\alpha = 0.97$. Essa etapa é utilizada para compensar a atenuação das componentes de alta frequência causadas pelo mecanismo da produção de voz.
- Depois do sinal ser filtrado, é necessário atenuar as descontinuidades causadas no início e no final do sinal de cada segmento, o qual é realizado aplicando a janela *Hamming* de 20 ms de comprimento, com deslocamento entre janelas de 10ms, obtendo-se assim vetores MFCC a cada 10ms.
- Após o janelamento do sinal, aplica-se a DFT para obtenção do espectro e da potência espectral.
- A etapa a seguir é a aplicação do banco filtros à potência espectral. O banco de filtros é formado por filtros triangulares, espaçados de acordo com a escala de frequência MEL. Cada filtro calcula a média do espectro em torno da frequência central e têm diferentes larguras de banda. Quanto maior é a frequência maior será a largura de banda.
- Após a aplicação do banco de filtros, calcula-se o log-energia da saída de cada um dos filtros MEL.
- Finalmente, obtém-se os coeficientes MFCC aplicando a transformada inversa do cosseno ao logaritmo dos coeficientes de energia obtidos no item anterior.

Na API, esse procedimento é realizado utilizando a chamada de função ilustrada na Figura 16.

```
_mfcc = mfcc.mfcc(signal, framerate, num_ceps=25, pre_emph=1, pre_emph_coeff=0.97,
             win_len=0.025, win_hop=0.01, win_type='hamming', nfilts=25, nfft=514,
            low_freq=None, high_freq=None, scale='constant', dct_type=2,
            use_energy=False, lifter=25, normalize=1)
```

Figura 16. Aplicação da técnica MFCC utilizando a biblioteca *Spafe*.

5.1.2.2 *Perceptual Linear Prediction* (PLP)

O PLP foi proposto por Hermansky e visa melhorar a estimativa do modelo de predição linear considerando as características psicoacústicas do sistema auditivo humano (Borges, 2011). No modelo proposto por Hermansky, o espectro do sinal de voz é modificado de acordo com características psicoacústicas. A ideia é semelhante a utilizada

no cálculo dos coeficientes Mel-Cepestros, entretanto, utiliza-se filtros assimétricos e com banda maior que a dos filtros triangulares para simular as bandas críticas e a escala Bark para espaçamento desses filtros.

Para extração dos vetores de características PLP, foi aplicada a técnica descrita na seção 2.2.2.3.2 e descrita brevemente a seguir:

- O início do processo se dá com a realização das etapas de aquisição do sinal de voz, segmentação, janelamento e transformada de Fourier.
- Com os resultados obtidos na FFT aplica-se um banco de filtros à potência espectral, que é formado por filtros assimétricos, espaçados de acordo com a escala de frequência Bark. Uma das primeiras constatações feitas em relação à percepção de frequência do sistema auditivo humano é que ele não percebe as componentes de frequência de forma linear e por este motivo se faz necessário a conversão do espectro do sinal da escala Hz para a escala Bark.
- Em seguida o sinal é processado em um filtro de pré-ênfase com $\alpha = 0.97$. Essa etapa é utilizada para compensar a atenuação das componentes de alta frequência causadas pelo mecanismo da produção de voz.
- O término da análise de predição linear perceptual é realizado aplicando a transformação do espectro do sinal processado para o domínio temporal, por meio da transformada discreta de Fourier inversa e por fim, os coeficientes são então estimados pela análise de predição linear.

Na API, esse procedimento é realizado utilizando a chamada de função ilustrada na Figura 17.

```
_plp = rplp.plp(signal, framerate, num_ceps=25, pre_emph=1, pre_emph_coeff=0.97, win_len=0.025,
            win_hop=0.01, modelorder=25, normalize=1)
```

Figura 17. Aplicação da técnica PLP utilizando a biblioteca *Spafe*.

5.1.3 Criação dos modelos de treinamento e teste

Nesta etapa, os arquivos de áudio em formato bruto foram submetidos às funções de extração de características acústicas e seus resultados foram armazenados em vetores bidimensionais, contendo, as características extraídas e os rótulos das emoções analisadas. Posteriormente, os vetores criados foram persistidos em arquivos físicos. A Figura 18 e a Figura 19 apresentam as funções de treinamento e teste.

```
def create_dataset_train(feature):
    x,y=[],[]

    print("Create training dataset - feature %s" % (feature))

    for emotion in observed_emotions:
        for file in glob.glob("../data/treinamento/%s/*.wav" % emotion):
            filename = os.path.basename(file)

            if emotions.get(filename.split("_")[0]) is None:
                continue

            if feature=='mfcc':
                mfcc = extract_features.MFCC()
                dataset = mfcc.mfcc(file)
            else:
                plp = extract_features.PLP()
                dataset = plp.plp(file)

            x.append(dataset)
            y.append(emotion)

            print('.', end='', flush=True)

    print('')

    return (np.array(x), y)
```

Figura 18. Função criação de base de treinamento

```
def create_dataset_test(feature):
    x,y=[],[]

    print("Create testing dataset - feature %s" % (feature))

    for file in glob.glob("../data/teste/*.wav"):
        filename = os.path.basename(file)

        if emotions.get(filename.split("_")[0]) is None:
            continue

        emotion = emotions[filename.split("_")[0]]

        if feature=='mfcc':
            mfcc = extract_features.MFCC()
            dataset = mfcc.mfcc(file)
        else:
            plp = extract_features.PLP()
            dataset = plp.plp(file)

        x.append(dataset)
        y.append(emotion)

        print('.', end='', flush=True)

    return (np.array(x), y)
```

Figura 19. Função de criação de base de teste.

5.1.4 Classificação

Após geração dos modelos, foram utilizados os métodos de classificação GMM e SVM para agrupamento dos dados em categorias de emoção. Para implementação dos procedimentos foi utilizado a biblioteca *Scikit-learn*.

5.1.4.1 Gaussian Mixture Model (GMM)

Modelos de misturas gaussianas são métodos estatísticos que assumem que uma população de dados pertence a uma distribuição em forma de mistura, ou seja, uma distribuição formada por combinações lineares de diversas funções densidade de probabilidade. Os modelos GMMs representam uma distribuição como a soma ponderada de diversas gaussianas de diferentes médias e matrizes de covariância.

O algoritmo de *Expectation Maximization* (EM) foi utilizado para estimar os parâmetros das gaussianas. Para a classificação, não foi necessário nenhum algoritmo em especial, bastando apenas calcular a probabilidade da sequência de observações segundo cada modelo, e escolher o modelo com maior verossimilhança.

No experimento foi utilizado o módulo *GaussianMixture* da biblioteca *Scikit-learn*, para realização das simulações. Para a inicialização das gaussianas foram testados os algoritmos *K-Means* e modo Aleatório, sendo escolhido o *K-Means* por ser mais estável. O número de componentes da mistura foi definido com o total de classes de emoções e a covariância foi testada com os tipos: a) *full* “cada componente tem sua própria matriz de covariância geral”, b) *tied* “todos os componentes compartilham a mesma matriz de covariância geral”, c) *diag* “cada componente tem sua própria matriz de covariância diagonal”, e d) *spherical* “cada componente tem sua própria variação única”, sendo que a “*full*” apresentou melhores resultados na classificação. A Figura 20 apresentou a função implementada.

```
def classifier_gmm(feature):
    _y_pred = []
    dataset_file = ("../data/models/dataset_%s.dump" % feature)

    if os.path.exists(dataset_file) is False:
        print("File %s not exists." % dataset_file)
        exit()

    with open(dataset_file, "rb") as f:
        data = pickle.load(f)
        x_train, y_train, x_test, y_test = data

        n_classes = len(np.unique(y_train))

        # Try GMMs using different types of covariances.
        model = mixture.GaussianMixture(n_components=n_classes, covariance_type='full', max_iter=500, n_init=3)

        # Train the other parameters using the EM algorithm.
        model.fit(x_train)

        y_pred = model.predict(x_test)

        for value in y_pred:
            _y_pred.append(observed_emotions[value])

        print("F1 Score: %f" % (f1_score(y_test, _y_pred, labels=np.unique(_y_pred), average='micro')))

        print("Confusion Matrix:")
        print(confusion_matrix(y_test, _y_pred))

        print("Classification Report:")
        print(classification_report(y_test, _y_pred))

        #Calculate the accuracy of our model
        accuracy=accuracy_score(y_true=y_test, y_pred=_y_pred)

        #Print the accuracy
        print("Accuracy GMM: {:.2f}%".format(accuracy*100))
```

Figura 20. Função de implementação do GMM.

5.1.4.2 Support Vector Machine (SVM)

O SVM também conhecido como Máquina de Suporte Vetorial, foi elaborado com o estudo proposto por Boser, Guyon e Vapnik em 1992. Ele é um algoritmo de aprendizado supervisionado, cujo objetivo é classificar determinado conjunto de amostras de dados que são mapeados para um espaço de características multidimensional usando uma função *kernel*. Nela, o limite de decisão no espaço de entrada é representado por um hiperplano em dimensão superior no espaço.

No experimento foi utilizado o módulo SVM da biblioteca *Scikit-learn*, para realização das simulações. Os testes foram realizados utilizando a técnica multiclasse, cujos arquivos de áudio foram treinados em um único modelo, formando um conjunto composto de emoções. Para validação foram usados os kernel (*linear*, *poly*, *rbf* e *sigmoid*), sendo que o *kernel* "*rbf*" apresentou melhor desempenho na classificação. A Figura 21 apresentou a função de SVM implementada.

```
def classifier_svm(feature):
    _y_train = []
    dataset_file = ("../data/models/dataset_%s.dump" % feature)

    if os.path.exists(dataset_file) is False:
        print("File %s not exists." % dataset_file)
        exit()

    print("Classifier SVM - %s" % (feature.upper()))

    with open(dataset_file, "rb") as f:
        data = pickle.load(f)

        x_train, y_train, x_test, y_test = data

        #Create a svm Classifier
        model = svm.SVC(kernel="rbf", C=1, gamma=0.5, decision_function_shape='ovo')

        #Train the model using the training sets
        model.fit(x_train, y_train)

        y_pred = model.predict(x_test)

        print("F1 Score: %f" % (f1_score(y_test, y_pred, labels=np.unique(y_pred), average='micro')))

        print("Confusion Matrix:")
        print(confusion_matrix(y_test, y_pred))

        print("Classification Report:")
        print(classification_report(y_test, y_pred))

        #Calculate the accuracy of our model
        accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)

        #Print the accuracy
        print("Accuracy SVM: %.2f%s" % (accuracy*100, '%'))
```

Figura 21. Função de implementação do SVM.

5.2 Discussão dos resultados

A simulação realizada com classificador supervisionado SVM em conjunto com o método de extração de características MFCC apresentou o melhor desempenho na classificação, com 67,33% de eficiência na classificação de emoções. O classificador não supervisionado GMM em conjunto com o método de extração de características PLP apresentou eficiência de 19,67% na classificação de emoções. O classificador supervisionado SVM em conjunto com o método de extração de características PLP apresentou eficiência de 20,00% na classificação de emoções. O classificador não supervisionado GMM em conjunto com o método de extração de características PLP apresentou eficiência de 19,83% na classificação de emoções. A Tabela 6 apresentou os resultados consolidados.

Tabela 6. Resultados consolidados

Método	Eficiência na Classificação
SVM e MFCC	67,33%
GMM e MFCC	19,67%
SVM e PLP	20,00%
GMM e PLP	19,83%

Fonte: Elaborado pelo autor.

5.2.1 Simulação SVM e MFCC

A simulação realizada com o classificador SVM e método de extração de características MFCC obteve uma taxa de acerto de 67,33% no reconhecimento de emoções. A Tabela 7 apresentou a matriz de confusão da simulação. A diagonal principal representa a quantidade amostras que foram classificadas com exatidão (Verdadeiro positivo). Considerando a seleção de uma classe alvo "Ex: Alegria", os demais valores da diagonal principal representaram a classificação correta de amostras não pertencentes a essa classe (Verdadeiro negativo). Os valores apresentados no eixo vertical, representaram a quantidade de amostras que pertenciam a outras classes e foram classificadas como pertencentes a classe alvo (Falso positivo) e os valores apresentados no eixo horizontal, representaram a quantidade de amostras da classe alvo, que foram classificadas como outro tipo de emoção (Falso negativo). De modo resumido, a matriz de confusão pode ser interpretada, pela análise da diagonal principal, que representou o total de emoções que foram identificadas para cada classe e pela somatória do eixo horizontal, que representou o erro na identificação de determinada classe.

Nesta simulação, a quantidade de acerto por emoções foram: a) alegria: 57 amostras; b) calma: 56 amostras; c) raiva: 113 amostras; d) surpresa: 119 amostras; e e) tristeza: 59 amostras, de 600 amostras analisadas.

Tabela 7. Matriz de confusão SVM e MFCC

	Alegria	**Calma**	**Raiva**	**Surpresa**	**Tristeza**
Alegria	**57**	2	59	0	2
Calma	0	**56**	60	0	4
Raiva	7	0	**113**	0	0
Surpresa	1	0	0	**119**	0
Tristeza	0	0	61	0	**59**

Fonte: Elaborado pelo autor.

As métricas de precisão, revocação e média F_1 foram calculadas por classe e foram apresentadas na Tabela 8.

Tabela 8. Relatório de classificação SVM e MFCC

	Precisão	Revocação	Média F	Amostras
Alegria	0,88	0,47	0,62	120
Calma	0,97	0,47	0,63	120
Raiva	0,39	0,94	0,55	120
Surpresa	1,00	0,99	1,00	120
Tristeza	0,91	0,49	0,64	120

Fonte: Elaborado pelo autor

A precisão é uma medida da exatidão do classificador. Representa para todas as ocorrências classificadas como positivas, qual porcentagem estava correta. A classe **Surpresa** apresentou a maior precisão na classificação, com 100% de acertos e a precisão global, obtido pela média ponderada, foi de 83%. A Figura 22 apresenta o percentual de precisão para todas as classes.

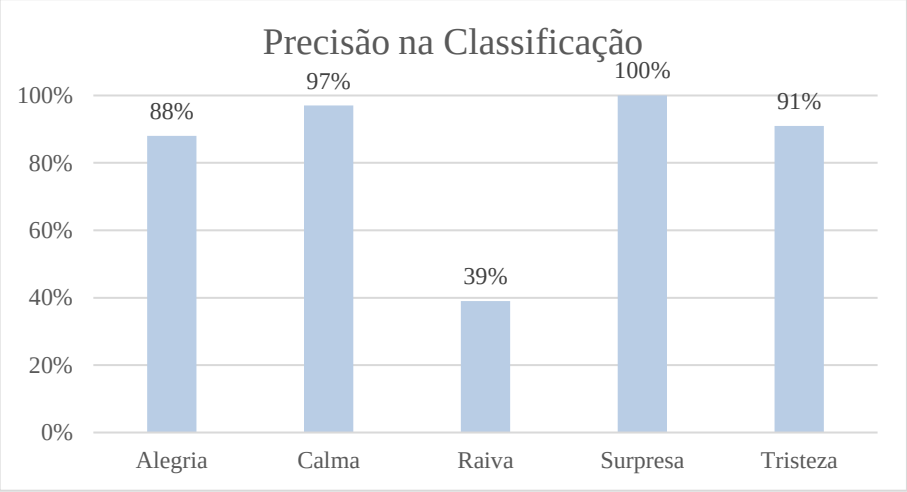

Figura 22. Precisão na Classificação SVM x MFCC

A revocação é uma medida da integridade do classificador. Para todas as instâncias que foram realmente classificadas como positivas, qual porcentagem foi classificada corretamente. A classe **Surpresa** apresentou a maior revocação na classificação, com 99% de acertos e a revocação global, obtida pela média ponderada, foi de 67%. A Figura 23 apresentou o percentual de revocação para todas as classes.

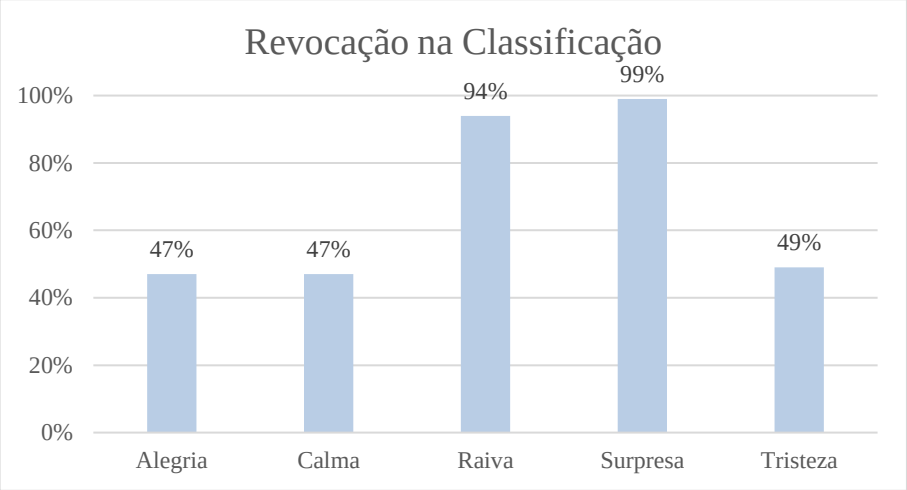

Figura 23. Revocação na Classificação SVM x MFCC

A média F_1 é uma média harmônica ponderada de precisão e recall, de modo que a melhor pontuação é 1,0 e a pior é 0,0. De modo geral, as pontuações F_1 são mais baixas do que as medidas de precisão, pois incorporam precisão e recall em seus cálculos. Como regra geral, a média ponderada de F_1 deve ser usada para comparar os modelos do classificador, não a precisão global. A classe **Surpresa** apresentou a maior média F_1 na classificação, com pontuação de 1,00 e a pontuação global, obtida pela média ponderada, foi de 0,69 pontos. A Figura 24 apresenta a pontuação da média F_1 para todas as classes.

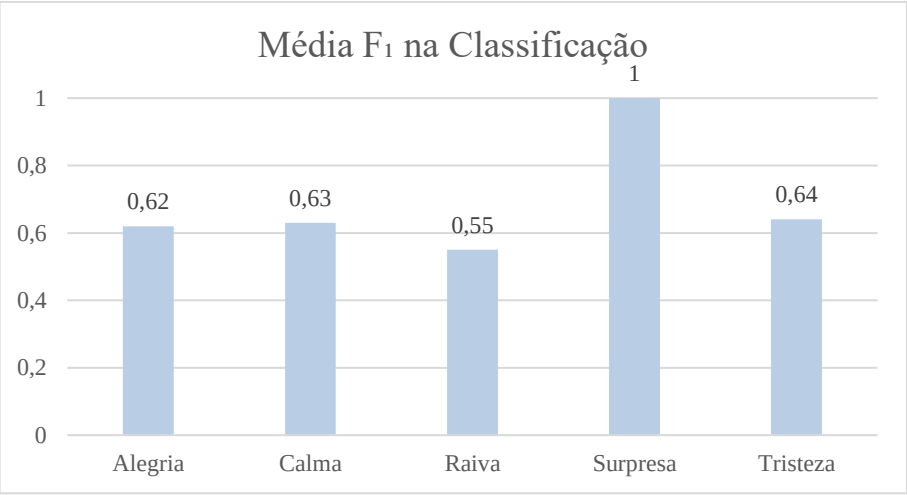

Figura 24. Média F_1 na Classificação SVM x MFCC

5.2.2 Simulação GMM e MFCC

A simulação realizada com os métodos GMM e MFCC obteve uma taxa de acerto de 19,67%. A Tabela 9 apresenta a matriz de confusão da simulação. Nesta, a quantidade de acerto por emoções foram: a) alegria: 7 amostras; b) calma: 24 amostras; c) raiva: 21 amostras; d) surpresa: 46 amostras; e e) tristeza: 20 amostras, de 600 amostras analisadas.

Tabela 9. Matriz de confusão GMM e MFCC

	Alegria	Calma	Raiva	Surpresa	Tristeza
Alegria	**7**	14	19	63	17
Calma	6	**24**	15	59	16
Raiva	8	16	**21**	47	28
Surpresa	8	26	21	**46**	19
Tristeza	6	15	20	59	**20**

Fonte: Elaborado pelo autor.

As métricas de precisão, revocação e média F_1 foram calculadas por classe e foram apresentadas na Tabela 10.

Tabela 10. Relatório de classificação GMM e MFCC

	Precisão	Revocação	Média F	Amostras
Alegria	0,20	0,06	0,09	120
Calma	0,25	0,20	0,22	120
Raiva	0,22	0,17	0,19	120
Surpresa	0,17	0,38	0,23	120
Tristeza	0,20	0,17	0,18	120

Fonte: Elaborado pelo autor.

A classe **Calma** apresentou a maior precisão na classificação, com 25% de acertos e a média global, obtida pela média ponderada, foi de 21%. A Figura 25 apresenta o percentual de precisão das classes.

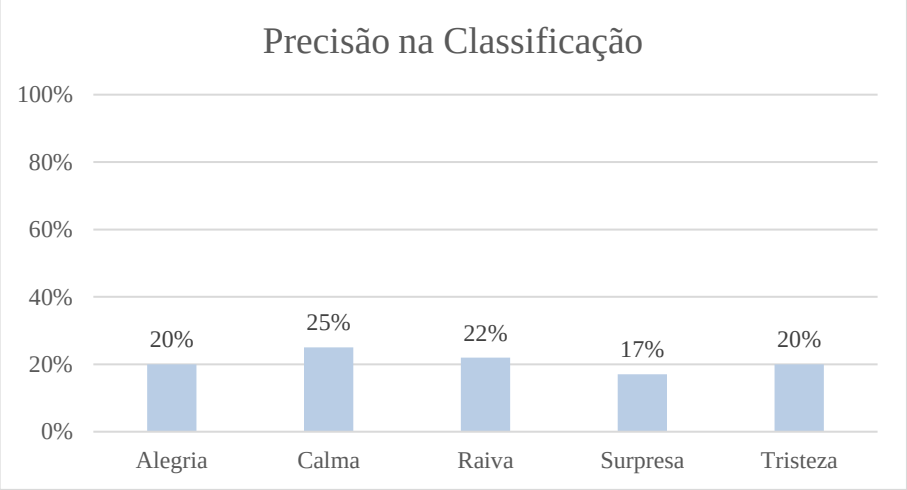

Figura 25. Precisão na Classificação GMM x MFCC

A classe **Surpresa** apresentou a maior revocação na classificação, com 38% de acertos e a média global, obtida pela média ponderada, foi de 20%. A Figura 26 apresentou o percentual de revocação das classes.

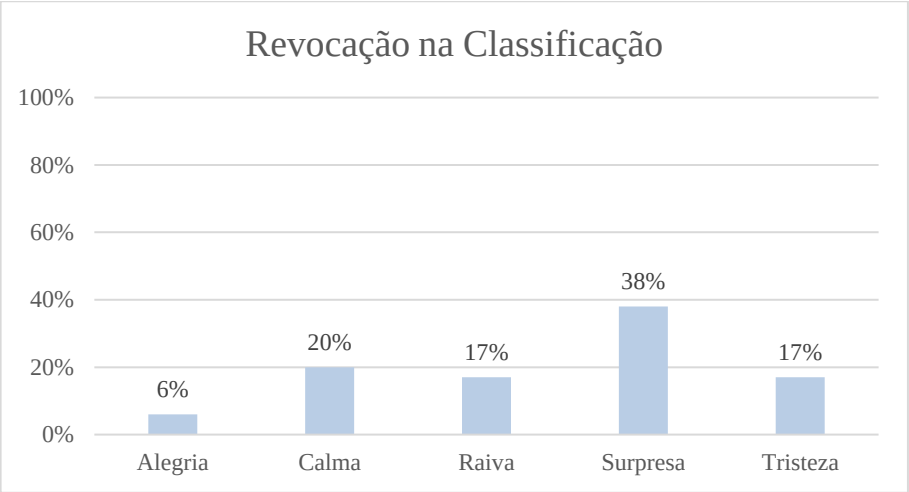

Figura 26. Revocação na Classificação GMM x MFCC

A classe **Surpresa** apresentou a maior média F_1 na classificação, com pontuação de 0,23 e a pontuação global, obtida pela média ponderada, foi de 0,18 pontos. A Figura 27 apresentou a pontuação da média F_1 para todas as classes.

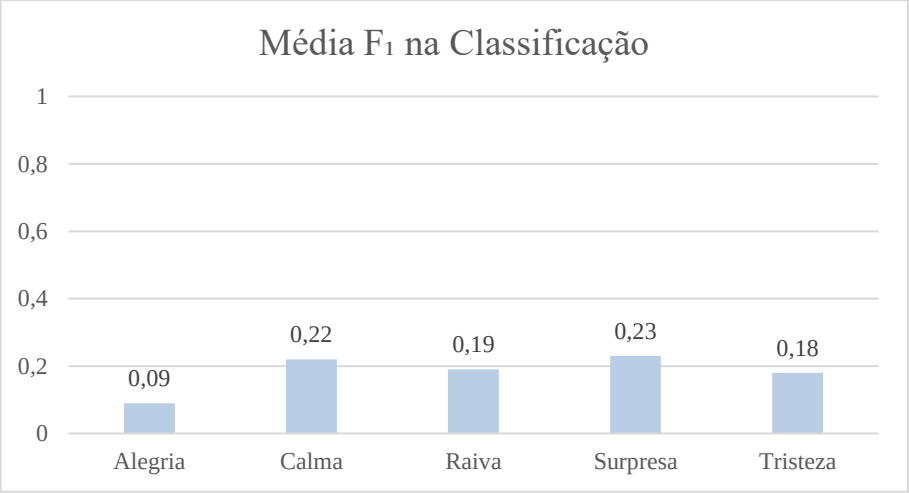

Figura 27. Média F_1 na Classificação GMM x MFCC

5.2.3 Simulação SVM e PLP

A simulação realizada com os métodos SVM e PLP obteve uma taxa de acerto de 20,00%. A Tabela 11 apresenta a matriz de confusão da simulação. Nela foi possível observar que o classificador não conseguiu distinguir os tipos das amostras e os classificou como sendo do tipo **Raiva**. Assim concluiu-se que os métodos utilizados não foram capazes de identificar emoções em base de áudio natural.

Tabela 11. Matriz de confusão SVM e PLP

	Alegria	Calma	Raiva	Surpresa	Tristeza
Alegria	**0**	0	120	0	0
Calma	0	**0**	120	0	0
Raiva	0	0	**120**	0	0
Surpresa	0	0	120	**0**	0
Tristeza	0	0	120	0	**0**

Fonte: Elaborado pelo autor.

As métricas de precisão, revocação e média F_1 foram calculadas por classe e foram apresentadas na Tabela 12.

Tabela 12. Relatório de classificação SVM e PLP

	Precisão	Revocação	Média F	Amostras
Alegria	0,00	0,00	0,00	120
Calma	0,00	0,00	0,00	120
Raiva	0,20	1,00	0,33	120
Surpresa	0,00	0,00	0,00	120
Tristeza	0,00	0,00	0,00	120

Fonte: Elaborado pelo autor

A classe **Raiva** apresentou precisão de 20% e a precisão global, obtida pela média ponderada, foi de 4%. A Figura 28 apresentou o percentual de precisão na simulação para todas as classes.

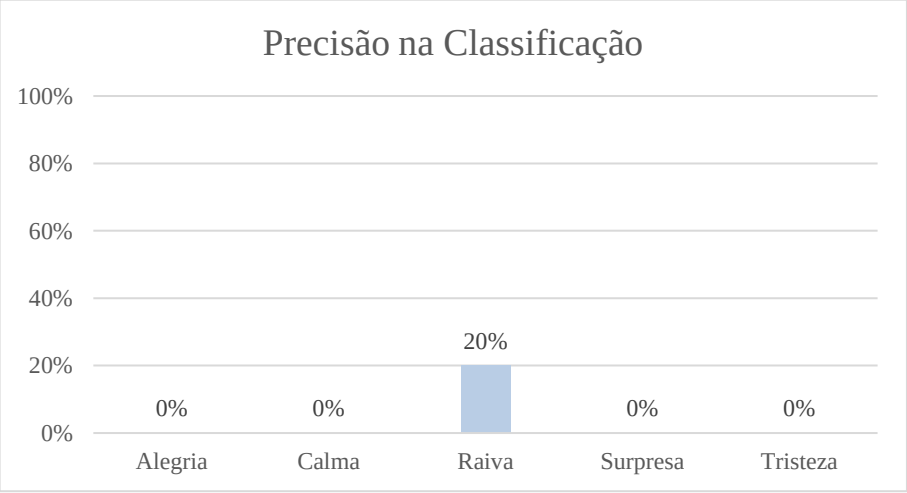

Figura 28. Precisão na Classificação SVM x PLP

A classe **Raiva** apresentou revocação de 100% e a revocação global, obtida pela média ponderada, foi de 20%. A Figura 29 apresenta o percentual de revocação na simulação para todas as classes.

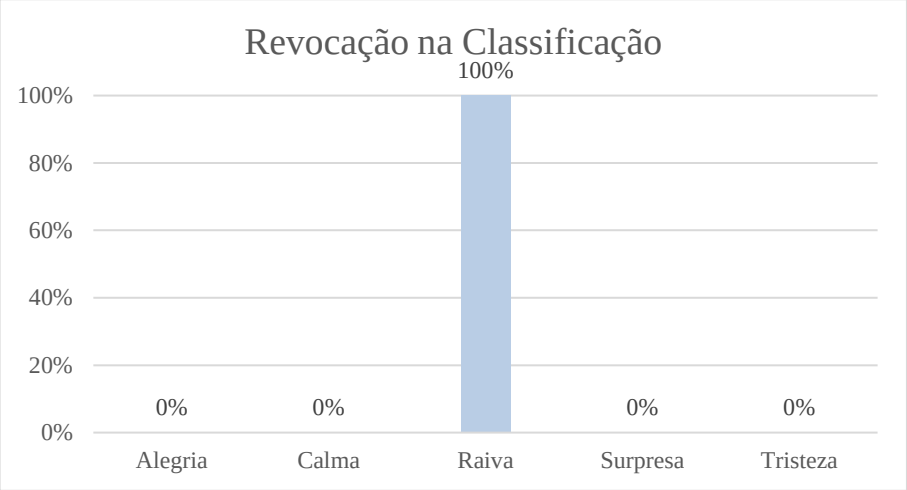

Figura 29. Revocação na Classificação SVM x PLP

A classe **Raiva** apresentou a pontuação de 0,33 de média F_1 e a pontuação global, obtida pela média ponderada, foi de 0,07 pontos. A Figura 30 apresenta a pontuação da média F_1 para todas as classes.

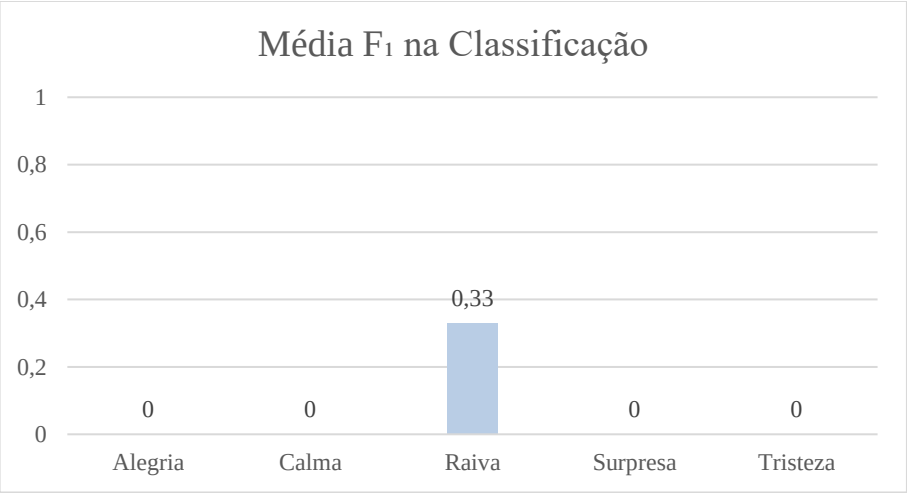

Figura 30. Média F_1 na Classificação SVM x PLP

5.2.4 Simulação GMM e PLP

A simulação realizada com os métodos SVM e PLP obteve uma taxa de acerto de 19,83%. A Tabela 13 apresentou a matriz de confusão da simulação. Nela foi possível observar que o classificador não conseguiu distinguir os tipos de emoções, classificando a maior parte como do tipo **Alegria**. Assim concluiu-se que os métodos utilizados não foram capazes de identificar emoções em base de áudio natural.

Tabela 13. Matriz de confusão GMM e PLP

	Alegria	**Calma**	**Raiva**	**Surpresa**	**Tristeza**
Alegria	**119**	1	0	0	0
Calma	120	**0**	0	0	0
Raiva	106	5	**0**	8	1
Surpresa	120	0	0	**0**	0
Tristeza	120	0	0	0	**0**

Fonte: Elaborado pelo autor.

As métricas de precisão, revocação e média F_1 foram calculadas por classe e foram apresentadas na Tabela 14.

Tabela 14. Relatório de classificação GMM e PLP

	Precisão	**Revocação**	**Média F**	**Amostras**
Alegria	0,20	0,99	0,34	120
Calma	0,00	0,00	0,00	120
Raiva	0,00	0,00	0,00	120
Surpresa	0,00	0,00	0,00	120
Tristeza	0,00	0,00	0,00	120

Fonte: Elaborado pelo autor.

A classe **Alegria** apresentou precisão de 20% e precisão global, obtida pela média ponderada, foi de 4%. A Figura 31 apresentou o percentual de precisão na simulação.

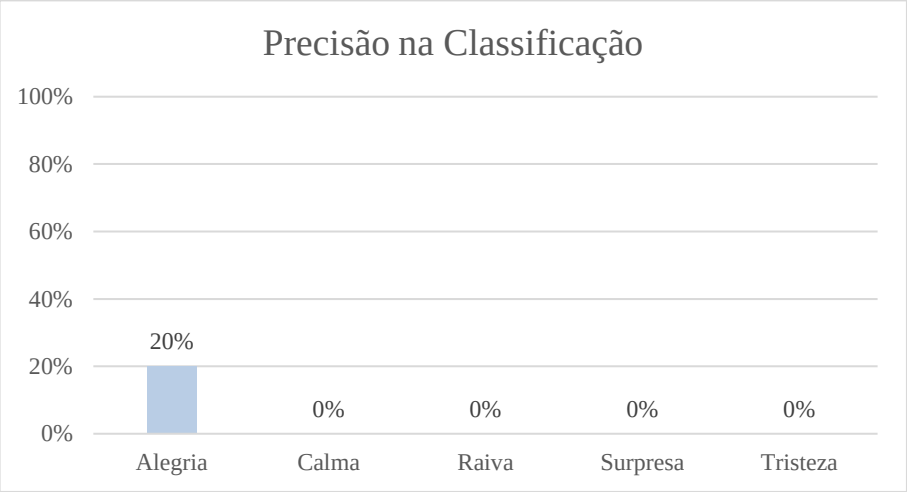

Figura 31. Precisão na Classificação GMM x PLP

A classe **Alegria** apresentou revocação de 99% e a revocação global, obtida pela média ponderada, foi de 20%. A Figura 32 apresentou o percentual de precisão na simulação.

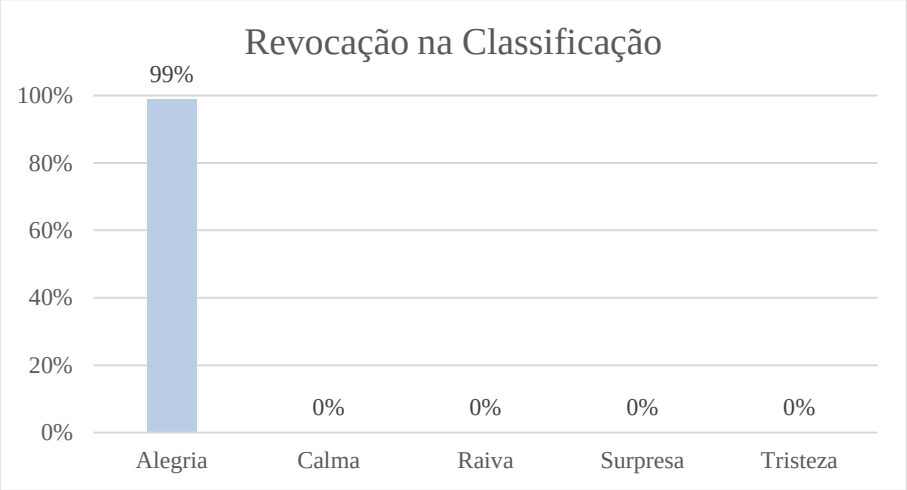

Figura 32. Revocação na Classificação GMM x PLP

A classe **Alegria** apresentou a pontuação de 0,34 de média F_1 e a pontuação global, obtida pela média ponderada, foi de 0,07 pontos. A Figura 33 apresentou a pontuação da média F_1 para todas as classes.

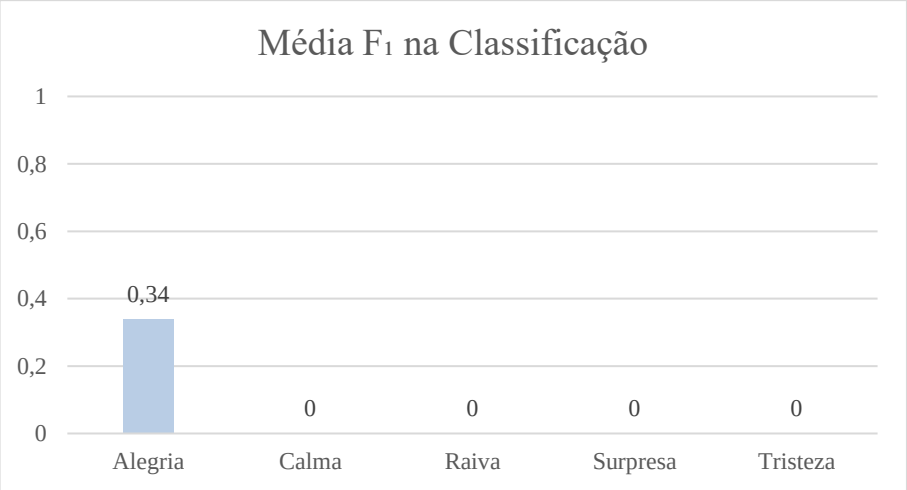

Figura 33. Média F_1 na Classificação GMM x PLP

5.3 Validação dos resultados

Como os resultados do experimento, em base de áudio natural, apresentaram divergência dos observados na literatura, definiu-se pela realização de validação cruzada, em base de áudio artificial. Para tal, foi selecionada a base de áudios simulada Toronto Emocional Speech Set (TESS) (Pichora-Fuller & Dupuis, 2020). A base é composta por um conjunto de 200 palavras-alvo e totalizam 2.800 expressões de emoções. As expressões foram pronunciadas por duas atrizes canadenses da região de Toronto, com idades entre 26 e 64 anos, formação universitária e musical e idioma inglês. Para construção base as atrizes foram instruídas a pronunciar a frase-base "Diga a palavra _____ ' e em seguida pronunciar

a palavra alvo. O conjunto de dados retratou sete categorias de emoções, que são: raiva, nojo, medo, felicidade, surpresa, tristeza e neutro.

Para comparação com os resultados obtidos no uso da base de áudio natural, foram selecionados os conjuntos de emoções: **raiva**, **felicidade**, **surpresa**, **tristeza** e **neutro** – que retratam o mesmo conjunto de emoções usadas nos testes com a base de áudios natural e totalizam 2.000 arquivos de expressões emocionais – e aplicados os métodos de extração de características acústicas MFCC e PLP e os métodos de classificação GMM e SVM. A validação do modelo se deu pela aplicação da técnica de *cross-validation*, cujos 80% dos arquivos foram utilizados para treinamento e 20% reservados para teste.

A simulação realizada com classificador supervisionado SVM em conjunto com o método de extração de características MFCC apresentou o melhor desempenho na classificação, com 78,00% de eficiência na classificação de emoções. O classificador não supervisionado GMM em conjunto com o método de extração de características PLP apresentou eficiência de 30,50% na classificação de emoções. O classificador supervisionado SVM em conjunto com o método de extração de características PLP apresentou eficiência de 91,50% na classificação de emoções. O classificador não supervisionado GMM em conjunto com o método de extração de características PLP apresentou eficiência de 22,00% na classificação de emoções. A Tabela 15 apresenta os resultados consolidados.

Tabela 15. Resultados consolidados base artificial

Método	**Eficiência na Classificação**
SVM e MFCC	78,00%
GMM e MFCC	30,50%
SVM e PLP	91,50%
GMM e PLP	22,00%

Fonte: Elaborado pelo autor.

5.3.1 Simulação SVM e MFCC

A simulação realizada com os métodos SVM e MFCC obteve uma taxa de acerto, em média ponderada, de 78,00%. A Tabela 16 apresentou a matriz de confusão da simulação. Nesta, a quantidade de acerto por emoções foram: a) alegria: 69; b) calma: 45; c) raiva: 82; d) surpresa: 52; e e) tristeza: 64, de 400 amostras analisadas.

Tabela 16. Matriz de confusão SVM e MFCC em base artificial

	Alegria	**Calma**	**Raiva**	**Surpresa**	**Tristeza**
Alegria	**69**	4	1	6	3
Calma	6	**45**	3	11	10
Raiva	0	3	**82**	1	1
Surpresa	7	14	4	**52**	3
Tristeza	1	6	1	3	**64**

Fonte: Elaborado pelo autor

As métricas de precisão, revocação e média F_1 foram calculadas por classe e foram apresentadas na Tabela 17**Erro! Fonte de referência não encontrada.**.

Tabela 17. Relatório de classificação SVM e MFCC em base artificial

	Precisão	**Revocação**	**Média F**	**Amostras**
Alegria	0,83	0,83	0,83	83
Calma	0,62	0,60	0,61	75
Raiva	0,90	0,94	0,92	87
Surpresa	0,71	0,65	0,68	80
Tristeza	0,79	0,85	0,82	75

Fonte: Elaborado pelo autor.

A classe **Raiva** apresentou a maior precisão na classificação, com 90% de acertos e a precisão global, obtida pela da média ponderada, foi de 78% de acertos. A Figura 34 apresentou o percentual de precisão das classes.

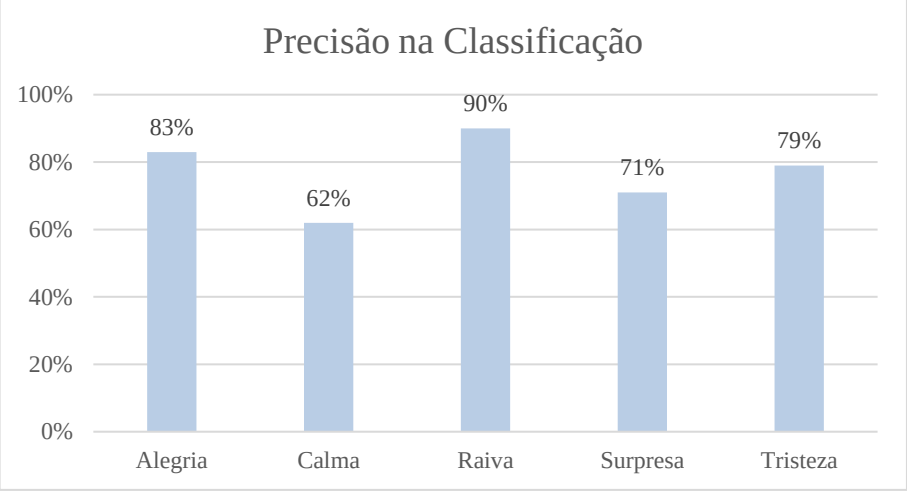

Figura 34. Precisão na classificação SVM e MFCC em base artificial

A classe **Raiva** apresentou a maior revocação na classificação, com 94% de identificação dos verdadeiros positivos e a revocação global, obtida pela média ponderada, foi de 78%. A Figura 35 apresentou o percentual de revocação das classes.

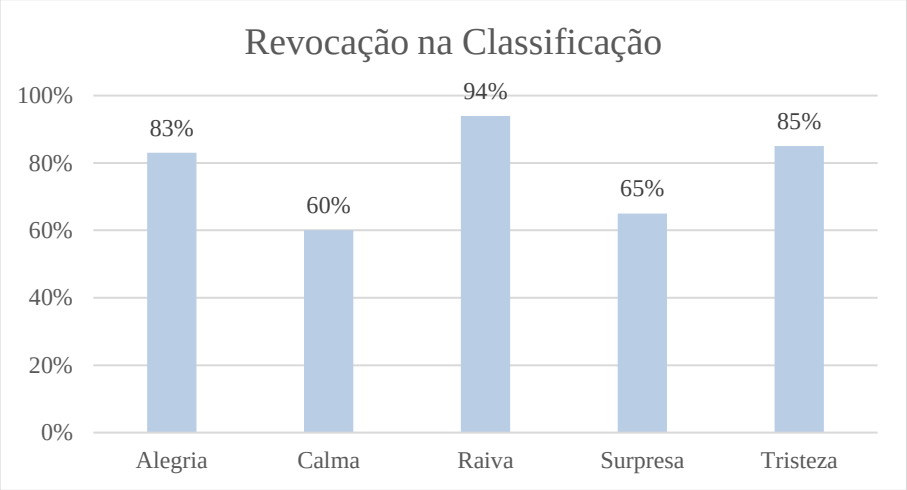

Figura 35. Revocação na classificação SVM e MFCC em base artificial

A classe **Raiva** apresentou a maior média F_1 na classificação, com pontuação de 0,92 e a pontuação global, obtida pela média ponderada, foi de 0,78 pontos. A Figura 36 apresentou a pontuação da média F_1 para todas as classes.

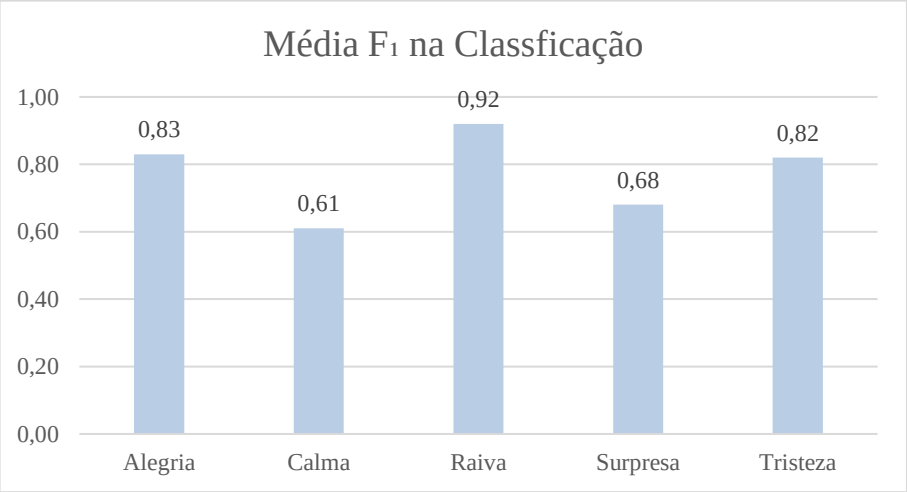

Figura 36. Média F_1 na classificação SVM e MFCC em base artificial

5.3.2 Simulação GMM e MFCC

A simulação realizada com os métodos GMM e MFCC obteve uma taxa de acerto, em média ponderada, de 30,50%. A Tabela 18 apresentou a matriz de confusão da simulação. Nela, a quantidade de acerto por emoções foram: a) alegria: 47 amostras; b) calma: 12 amostras; c) raiva: 5 amostras; d) surpresa: 24 amostras; e e) tristeza: 34 amostras, do total de 400 amostras analisadas.

Tabela 18. Matriz de confusão GMM e MFCC em base artificial

	Alegria	**Calma**	**Raiva**	**Surpresa**	**Tristeza**
Alegria	**47**	1	0	29	6
Calma	13	**12**	0	28	22
Raiva	44	8	**5**	9	21
Surpresa	34	7	0	**24**	15
Tristeza	8	29	0	4	**34**

Fonte: Elaborado pelo autor

As métricas de precisão, revocação e média F_1 foram calculadas por classe e foram apresentadas na Tabela 19.

Tabela 19. Relatório de classificação GMM e MFCC em base artificial

	Precisão	**Revocação**	**Média F**	**Amostras**
Alegria	0,32	0,57	0,41	83
Calma	0,21	0,16	0,18	75
Raiva	1,00	0,06	0,11	87
Surpresa	0,26	0,30	0,28	80
Tristeza	0,35	0,45	0,39	75

Fonte: Elaborado pelo autor.

A classe **Raiva** apresentou a maior taxa de precisão, com 100% de acertos, porém o resultado foi obtido com a identificação de cinco amostras como verdadeiras positivas e zero amostras como falso positivas. Além disso, a revocação dessa classe foi baixa, com 0,06% de acertos, o que indica que a maior parte de suas amostras foram identificadas como pertencentes a outro tipo de emoção.

A precisão global do experimento, obtido pela média ponderada, foi de 44% de acertos. A Figura 37 apresentou o percentual de precisão das classes.

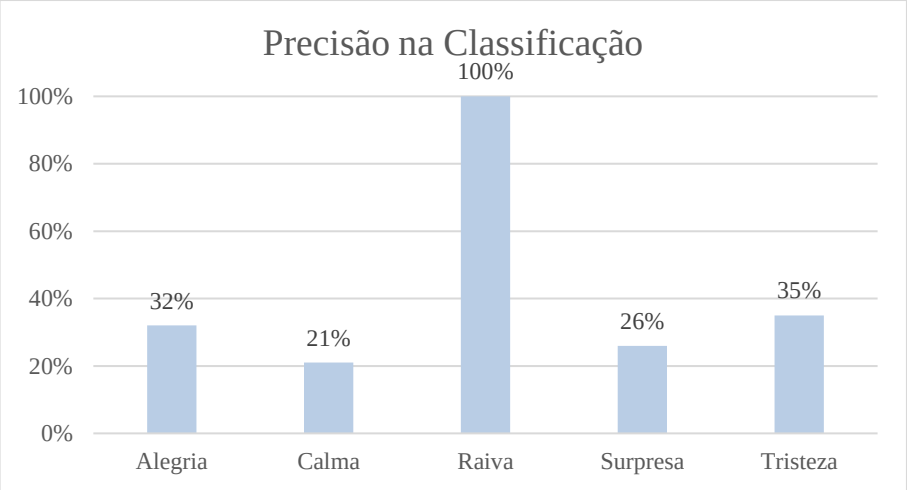

Figura 37. Precisão na classificação GMM e MFCC em base artificial

A classe **Alegria** apresentou a maior revocação na classificação, com 57% de identificação dos verdadeiros positivos e a revocação global, obtida pela média ponderada, foi de 30%. A Figura 38 apresentou o percentual de revocação das classes.

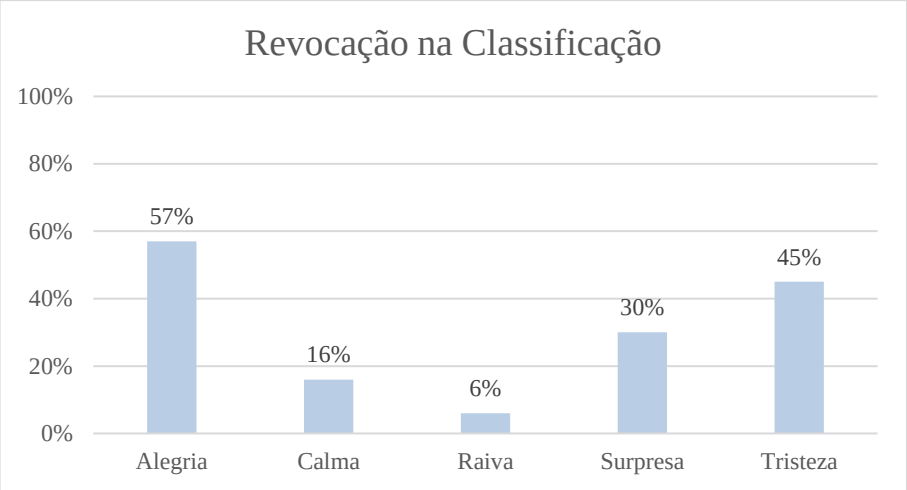

Figura 38. Revocação na classificação GMM e MFCC em base artificial

A classe **Alegria** apresentou a maior média F_1 na classificação, com pontuação de 0,41 e a pontuação global, obtida pela média ponderada, foi de 0,27 pontos. A Figura 39 apresentou a pontuação da média F_1 para todas as classes.

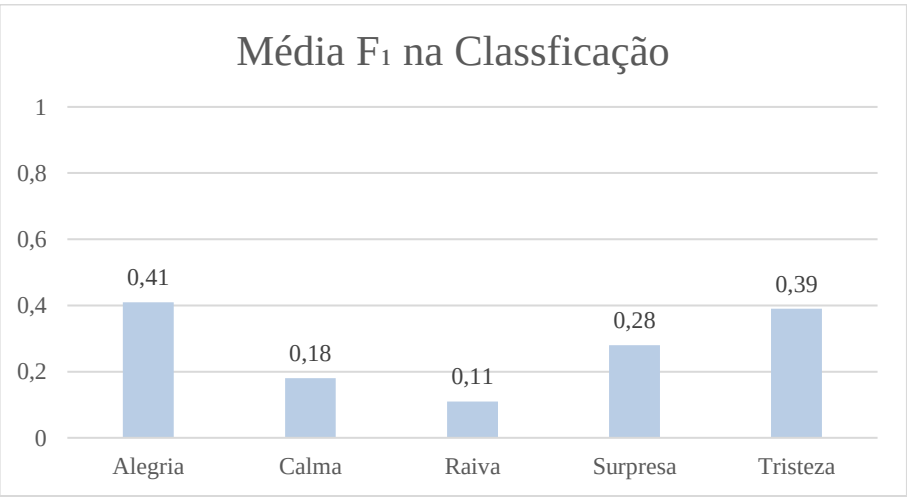

Figura 39. Média F_1 na classificação GMM e MFCC em base artificial

5.3.3 Simulação SVM e PLP

A simulação realizada com os métodos SVM e PLP obteve uma taxa de acerto, em média ponderada, de 91,50%. A Tabela 20 apresenta a matriz de confusão da simulação. Nesta, a quantidade de acerto por emoções foram: a) alegria: 79; b) calma: 72; c) raiva: 79; d) surpresa: 68; e e) tristeza: 68, de 400 amostras analisadas.

Tabela 20. Matriz de confusão SVM e PLP em base artificial

	Alegria	**Calma**	**Raiva**	**Surpresa**	**Tristeza**
Alegria	**79**	0	3	1	0
Calma	0	**72**	1	0	2
Raiva	2	4	**79**	2	0
Surpresa	4	4	3	**68**	1
Tristeza	0	6	0	1	**68**

Fonte: Elaborado pelo autor

As métricas de precisão, revocação e média F_1 foram calculadas por classe e foram apresentadas na Tabela 21.

Tabela 21. Relatório de classificação SVM e PLP em base artificial

	Precisão	**Revocação**	**Média F**	**Amostras**
Alegria	0,93	0,95	0,94	83
Calma	0,84	0,96	0,89	75
Raiva	0,92	0,91	0,91	87
Surpresa	0,94	0,85	0,89	80
Tristeza	0,96	0,91	0,93	75

Fonte: Elaborado pelo autor.

A classe **Tristeza** apresentou a maior precisão na classificação, com 96% de acertos e a precisão global, obtida pela da média ponderada, foi de 92% de acertos. A Figura 40 **Erro! Fonte de referência não encontrada.Erro! Fonte de referência não en contrada.Erro! Fonte de referência não encontrada.**apresentou o percentual de precisão das classes.

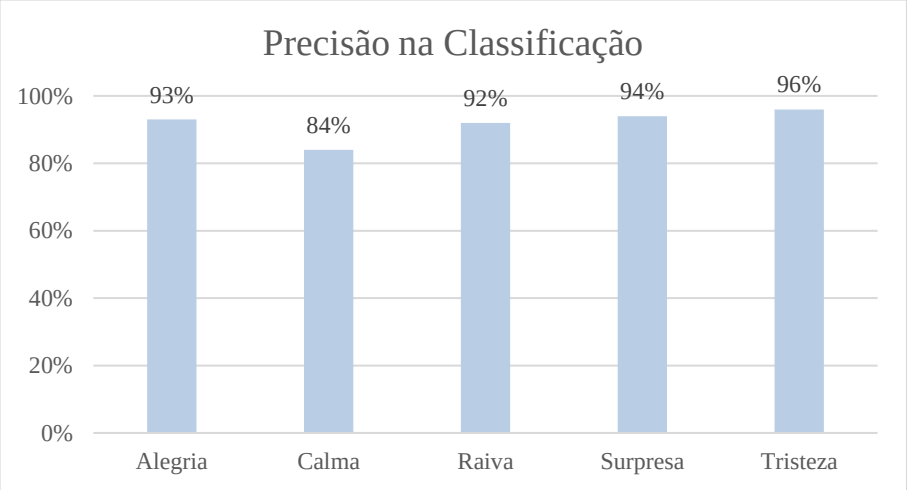

Figura 40. Precisão na classificação SVM e PLP em base artificial

A classe **Calma** apresentou a maior revocação na classificação, com 96% de identificação dos verdadeiros positivos e a revocação global, obtida pela média ponderada, foi de 92%. A Figura 41 apresentou o percentual de revocação das classes.

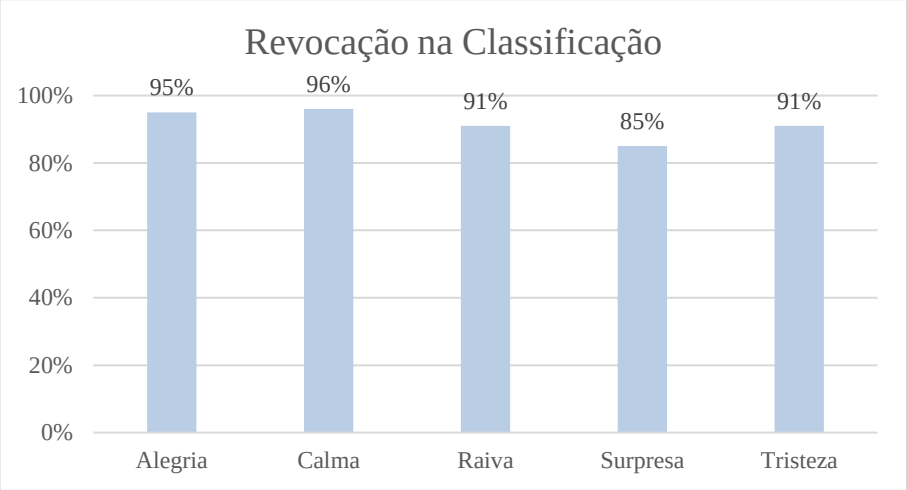

Figura 41. Revocação na classificação SVM e PLP em base artificial

A classe **Alegria** apresentou a maior média F_1 na classificação, com pontuação de 0,94 e a pontuação global, obtida pela média ponderada, foi de 0,92 pontos. A Figura 42 apresentou a pontuação da média F_1 para todas as classes.

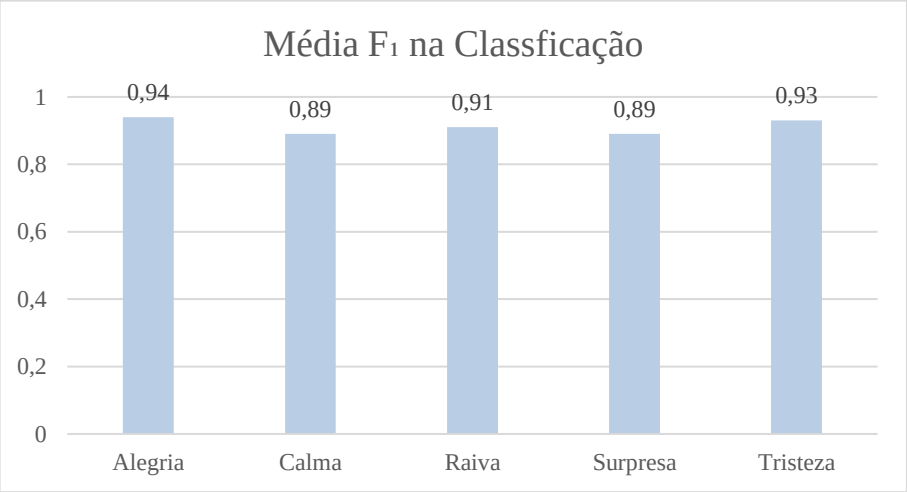

Figura 42. Média F_1 na classificação SVM e PLP em base artificial

5.3.4 Simulação GMM e PLP

A simulação realizada com os métodos GMM e PLP obteve uma taxa de acerto, em média ponderada, de 22,00%. A Tabela 22 apresentou a matriz de confusão da simulação. Nesta, a quantidade de acerto por emoções foram: a) alegria: 3; b) calma: 26; c) raiva: 3; d) surpresa: 28; e e) tristeza: 28, de 400 amostras analisadas.

Tabela 22. Matriz de confusão GMM e PLP em base artificial

	Alegria	**Calma**	**Raiva**	**Surpresa**	**Tristeza**
Alegria	**3**	4	18	20	38
Calma	40	**26**	0	0	9
Raiva	39	2	**3**	32	11
Surpresa	14	15	23	**28**	0
Tristeza	3	43	1	0	**28**

Fonte: Elaborado pelo autor

As métricas de precisão, revocação e média F_1 foram calculadas por classe e foram apresentadas na Tabela 23.

Tabela 23. Relatório de classificação GMM e PLP em base artificial

	Precisão	**Revocação**	**Média F**	**Amostras**
Alegria	0,03	0,04	0,03	83
Calma	0,29	0,35	0,32	75
Raiva	0,07	0,03	0,05	87
Surpresa	0,35	0,35	0,35	80
Tristeza	0,33	0,37	0,35	75

Fonte: Elaborado pelo autor.

A classe **Surpresa** apresentou a maior precisão na classificação, com 35% de acertos e a precisão global, obtida pela da média ponderada, foi de 21% de acertos. A Figura 43 apresentou o percentual de precisão das classes.

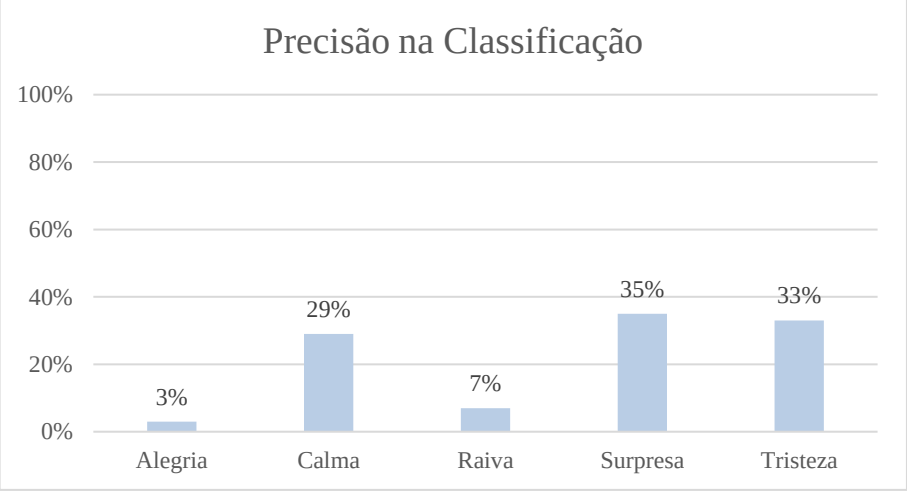

Figura 43. Precisão na classificação GMM e PLP em base artificial

A classe **Tristeza** apresentou a maior revocação na classificação, com 37% de identificação dos verdadeiros positivos e a revocação global, obtida pela média ponderada, foi de 22%. A Figura 44 apresentou o percentual de revocação das classes.

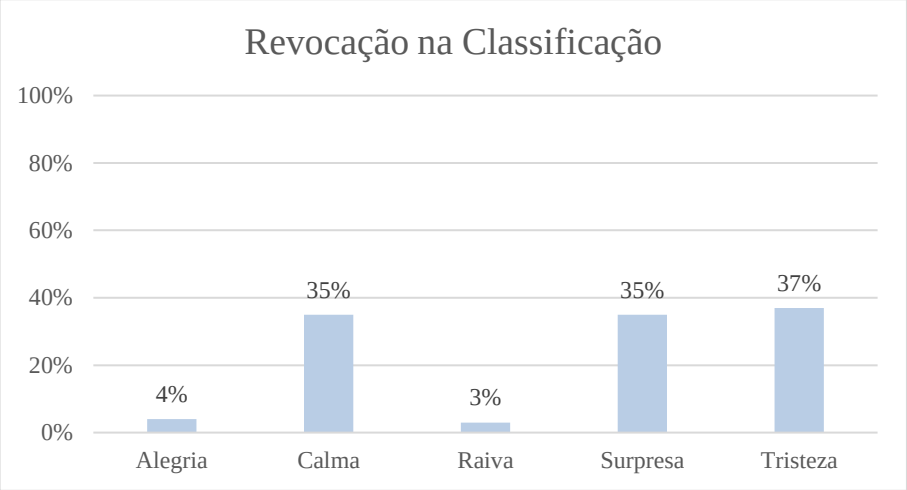

Figura 44. Revocação na classificação GMM e PLP em base artificial

As classes **Surpresa** e **Tristeza** apresentaram a maior média F_1 na classificação, com pontuação de 0,35 e a pontuação global, obtida pela média ponderada, foi de 0,21 pontos. A Figura 45 apresentou a pontuação da média F_1 para todas as classes.

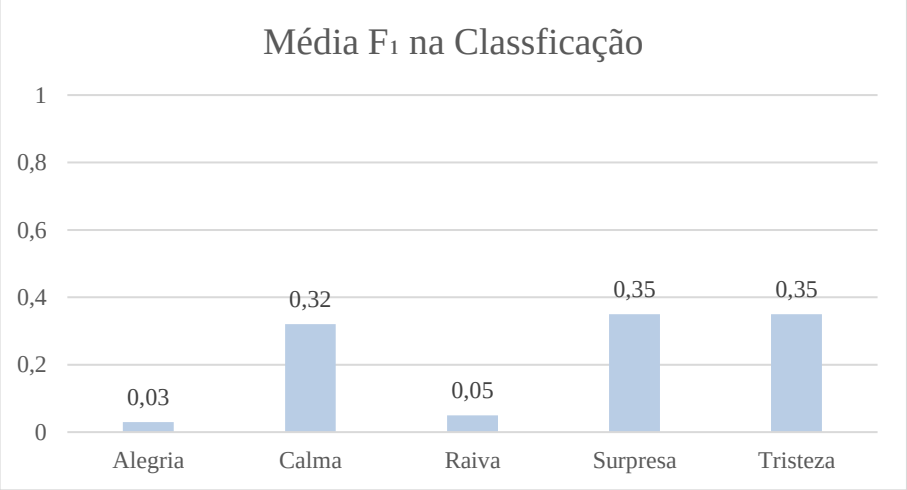

Figura 45. Média F_1 na classificação GMM e PLP em base artificial

5.4 Comparação dos resultados

A Tabela 24 apresentou os resultados das comparações sumarizados.

Tabela 24. Comparação entre Base Natural e Base Artificial

	SVM e MFCC	**GMM e MFCC**	**SVM e PLP**	**GMM e PLP**
Base Natural	67,33%	19,67%	20,00%	19,83%
Base Artificial	78,00%	30,50%	91,50%	22,00%

Fonte: Elaborado pelo autor.

A simulação realizada com classificador supervisionado SVM, combinado ao método de extração de características MFCC, apresentou desempenho de 78,00% no reconhecimento de emoções e comparado ao resultado obtido com a base de áudio natural,

observou-se a diferença de 10,67% de eficiência na classificação. Quando o classificador SVM foi combinado ao método de extração de características PLP, apresentou desempenho de 91,50% no reconhecimento de emoções e comparado a base de áudio natural, observou-se a diferença de 71,50% de eficiência na classificação. Outra diferença é que na base natural, a totalidade das amostras foram classificadas como pertencentes a classe **Raiva**, conforme Tabela 11 e na base artificial, o classificador foi capaz de identificar os tipos de emoções e separá-las em classes, conforme Tabela 20, o que indica que as características dos arquivos de áudio influenciam na classificação de emoções.

A simulação realizada com classificador não supervisionado GMM, combinado ao método de extração de características MFCC, apresentou desempenho de 30,50% no reconhecimento de emoções e comparado ao resultado obtido com a base de áudio natural, observou-se a diferença de 10,83% de eficiência na classificação. Quando o classificador GMM foi combinado ao método de extração de caraterísticas PLP, apresentou desempenho de 22% no reconhecimento de emoções e comparado a base de áudio natural, observou-se a diferença de desempenho de 2,17% de eficiência na classificação. Outra diferença foi que na base natural, a maior parte das amostras foram classificadas como pertencentes à classe **Alegria**, conforme Tabela 13 e na base artificial, o classificador foi capaz de identificar as emoções e separá-las em classes, conforme Tabela 22**Erro! Fonte de referência não e ncontrada.**, o que indica que as características dos arquivos de áudio influenciam na classificação de emoções.

5.5 Síntese do capítulo

Foram apresentados os procedimentos para implementação do artefato e os resultados obtidos. O conjunto de treinamento e testes foi composto por 3.000 mil arquivos de áudio. Destes, 80% foram utilizados para criação dos conjuntos de treinamento e 20% para criação do conjunto de teste. Para comparação das técnicas de extração de características MFCC e PLP, foram utilizados 25 coeficientes de curto prazo, essa escolha se deu pelo resultado obtido na classificação e o custo computacional para execução do experimento. Com o resultado das técnicas de extração de características foram criados os modelos de treinamento e teste e aplicado os algoritmos de classificação.

Em base de áudio natural, o melhor desempenho foi obtido com o uso do método de extração de caraterísticas MFCC e do classificador supervisionado SVM, com 67,33% de acertos. Em relação a simulação realizada com o método de extração de características MFCC e classificador não supervisionado GMM, os resultados demostraram que este

conjunto não obteve bom desempenho na classificação. Em comparação a simulação realizada com classificador supervisionado SVM, apresentou uma taxa de acertos de 19,67%, que é 47,66% inferior ao melhor resultado, tornando-se inviável para classificação emoções.

Nas simulações realizadas com o método de extração de características PLP, em base de áudio natural, os resultados demostraram que o método não foi eficiente para identificação de emoções. A simulação realizada com o classificador supervisionado SVM, obteve uma taxa de acerto 20%, porém, ao analisar a matriz de confusão da simulação, constatou-se que todas as amostras foram classificadas como pertencentes a um único tipo de emoção. A simulação realizada com o classificador não supervisionado GMM, obteve uma taxa de acerto de 19,83%, porém, ao analisar a matriz de confusão da simulação, constatou-se um comportamento similar ao do classificador SVM.

Como os resultados apresentaram divergência dos observados na literatura, optou-se pela realização de validação cruzada, em base de áudio artificial.

Em base de áudio artificial, a simulação realizada com o método de extração de características MFCC, combinada ao classificador supervisionado SVM, apresentou desempenho de 78,00% no reconhecimento de emoções. Quando comparado ao resultado obtido com a base de áudio natural, observou-se a diferença de 10,67% de eficiência na classificação, confirmando a eficiência dos métodos nos dois tipos de base de áudios. A simulação realizada com método de extração de características MFCC, combinada ao classificador GMM, apresentou um desempenho de 30,50% no reconhecimento de emoções. Quando comparado ao resultado obtido em base de áudio natural, observou-se a diferença de 10,83% na eficiência a classificação, confirmando que a combinação dos métodos não foi eficiente para identificação de emoções. A simulação realizada com o método de extração de caraterísticas PLP, combinada ao classificador supervisionado SVM, apresentou desempenho de 91,50% de acertos, demonstrando a eficiência do método para o reconhecimento de emoções em base de áudio artificial, porém, quando comparada ao resultado obtido em base de áudio natural, observou-se a diferença de 71,50% de eficiência na classificação. Outra diferença foi que na base de áudio natural, a totalidade das amostras foram classificadas como pertencentes a classe **Raiva**, o que demonstrou a ineficiência dos métodos para este tipo de amostra. A simulação realizada com o método de extração de características PLP, combinada ao classificador não supervisionado GMM, apresentou desempenho de 22,00%, indicando que a combinação de métodos não foi eficiente para classificação de emoções. Quando comparado ao resultado obtido em base de áudio natural, observou-se a diferença de 2,17% de eficiência, porém, na base de áudio natural, a maior

parte das amostras foram classificadas como pertencentes a classe **Alegria**, o que também indica que essa combinação de métodos não é adequada para este tipo de amostra.

Por fim, os resultados obtidos na comparação das bases áudio, confirmam que a diferença dos tipos de dados interfere no desempenho dos métodos e no reconhecimento de emoções.

6 CONSIDERAÇÕES FINAIS

O reconhecimento de emoções a partir de gravações de áudio fornece uma poderosa ferramenta para auxilio a tomada de decisão. Essa técnica permite que as empresas analisem as emoções dos clientes em gravações telefônicas e formulem estratégias para melhorar os serviços oferecidos (Subramaniam, Faruquie, Ikbal, Godbole, & Mohania, Business Intelligence from Voice of Customer, 2009).

A eficiência e eficácia dessas técnicas dependem dos métodos empregados para extração de características e classificação dos sinais acústicos (El Ayadi, Kamel, & Karray, 2011). Na literatura, observou-se a predominância do uso dos métodos de extração de características acústicas *Mel-Frequency Cepstral Coefficients* (MFCC) e *Perceptual Linear Predictive* (PLP) para identificação dos atributos relevantes da fala, e dos métodos de classificação *Gaussian Mixture Model* (GMM) e *Support Vector Machine* (SVM) para agrupamento dos enunciados em categorias de emoção (El Ayadi, Kamel, & Karray, 2011; Singh, Jain, & Tripath, 2014).

Diante dessa perspectiva a questão que direcionou essa pesquisa foi: “Qual o desempenho das técnicas de extração de características acústicas MFCC e PLP e dos métodos de classificação GMM e SVM no reconhecimento de emoções da fala em gravações telefônicas?”, a qual desencadeou uma pesquisa quantitativa, de abordagem exploratória e direcionada ao objetivo de: “Comparar o desempenho das técnicas de extração de características acústicas MFCC e PLP e dos métodos de classificação GMM e SVM no reconhecimento de emoções da fala em gravações telefônicas”.

Para desenvolvimento da pesquisa foram analisadas gravações telefônicas obtidas em um *call center* de um provedor de internet, que se enquadra na categoria de Prestadoras de Pequeno Porte (PPP) da Anatel e atende Belo Horizonte e região metropolitana.

Os resultados demonstraram que a combinação dos métodos: classificador supervisionado SVM e extração de características MFCC, em base de áudio natural, apresentaram o melhor desempenho na classificação de emoções, com 67,33% de acertos, que foi corroborado após aplicação dos métodos em base de áudio artificial e obtenção de 78% de acertos.

Quanto foi utilizado o classificador não supervisionado GMM, empregado ao método de extração de característica MFCC, em base de áudios natural, a taxa de acertos foi de 19,67% de acertos, demostrando a ineficiência dos métodos para classificação de emoções, que foi corroborado após a aplicação dos métodos em base de áudio artificial e obtenção de 30,50% de acertos.

Quando foi utilizado o classificador supervisionado SVM, empregado ao método de extração de caraterísticas PLP, em base de áudios natural, a taxa de acertos foi de 20%, porém, ao se analisar a matriz de confusão da simulação, constatou-se que a totalidade das amostras foram classificadas como pertencentes a uma classe de emoção. Quando os métodos foram aplicados em base de áudio artificial, obteve um desempenho de 91,50% de acertos, concluindo-se então que a combinação dos métodos foi eficaz para classificação de emoções, mas foi altamente dependente das características dos dados em que é aplicada.

Quando foi utilizado o classificador não supervisionado GMM, empregado ao método de extração de caraterísticas PLP, em base de áudios natural, a taxa de acertos foi de 19,83%, porém, ao analisar a matriz de confusão da simulação, constatou-se que a maior parte das amostras foram classificadas como pertencentes a uma classe de emoção. Quando os métodos foram aplicados em base de áudio artificial, obteve um desempenho de 22,00% e ao analisar a matriz de confusão da simulação, foi possível observar que o classificador, conseguiu separar as emoções em classes, porém obteve uma alta taxa de erro. Concluindo-se então que a combinação de métodos não foi eficaz para classificação de emoções.

6.1 Limitações da pesquisa

A limitação da pesquisa esteve na base de áudio utilizada, que devido as suas características, como: sobreposição de expressões; presença de ruído de fundo e emoções múltiplas e simultâneas, influenciaram no desempenho dos métodos analisados.

Outra limitação foi a quantidade de classificadores utilizados. A comparação de outros métodos poderia tornar experimento mais robusto, demostrando a eficácia da proposta.

Apesar das limitações acima citadas, os resultados foram satisfatórios, por se trata de uma base de dados de áudio natural.

6.2 Contribuições da pesquisa

Para a teoria, a principal contribuição apresentada por esta dissertação foi a proposta de um modelo com conceitos bem definidos provenientes da revisão da literatura. O tema utilizado Reconhecimento de Emoções na Fala aborda conceitos da ciência psicológica, fundamentos de reconhecimento de voz e técnicas para o processamento de linguagem natural, nesta pesquisa, representadas por gravações telefônicas.

A principal contribuição deste trabalho para a prática foi um modelo conceitual para reconhecimento de emoções na fala e um artefato de *software* que pode ser utilizado como exemplo para a construção de *softwares* ou novos experimentos.

6.3 Recomendações para trabalhos futuros

Como recomendações de estudos futuros, sugerem-se: a) utilização de outros métodos de extração das características acústicas e classificadores para classificação de emoções; b) transcrição do áudio e classificação para identificação dos problemas recorrentes de provedor de internet; e c) transcrição do áudio e aplicação de ferramentas de processamento de linguagem natural para criação de ferramentas de autoatendimento.

REFERÊNCIAS

Albornoz, E. M., Milone, D. H., & Rufiner, H. L. (2011). Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language, 25*(3), 556-570. doi:10.1016/j.csl.2010.10.001

Arias, J. P., Busso, C., & Yoma, N. B. (2014). Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech & Language, 1*(28), 278-294. doi:10.1016/j.csl.2013.07.002

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Pers. Soc. Psychol, 70*(3), 572-587.

Bax, M. P. (2015). Design science: filosofia da pesquisa em ciência da informação e tecnologia. *Ciência Da Informação, 2*(42), 298-312.

Borges, L. d. (2011). Extração de parâmetros característicos para detecção acústica de vazamento de água. *Tese de Doutorado, Escola Politécnica, Universidade de São Paulo*. São Paulo, SP, Brasil. doi:10.11606/T.3.2011.tde-19072011-110149

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *INTERSPEECH - 2005 - Eurospeech, 9th European Conference on Speech*, 1517-1520.

Busso, C., Bulut, M., & Narayanan, S. (2013). Toward effective automatic recognition systems of emotion in speech. *Social Emotions in Nature and Artifact*, 110-127.

Cannon, W. B. (1927). The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American Journal of Psychology, 39*(1/4), 106-124. doi:10.2307/1415404

Cao, H., Verma, R., & Nenkova, A. (2015). Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Computer Speech & Language, 29*(1), 186-202. doi:10.1016/j.csl.2014.01.003

Chen, L., Mao, X., Xue, Y., & Cheng, L. L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing, 22*(6), 1154-1160. doi:10.1016/j.dsp.2012.05.007

Dai, W., Han, D., Dai, Y., & Xu, D. (2015). Emotion recognition and affective computing on vocal social media. *Information & Management, 7*(52), 777-788. doi:10.1016/j.im.2015.02.003

Darwin, C. (1872). *The expression of the emotions in man and animals.* doi:10.1037/10001-000

El Ayadi, M. M., Kamel, M. S., & Karray, F. (2007). Speech emotion recognition using gaussian mixture vector autoregressive models. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, IV*, 957–960. doi:10.1109/icassp.2007.367230

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn, 44*(3), 572-587. doi:10.1016/j.patcog.2010.09.020

France, D. J., Shiavi, R., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering, 47*(7), 829-837. doi:10.1109/10.846676

Gergen, K. J. (1985). The social constructionist movement in modern psychology. *American Psychologist, 40*, 266-275. doi:10.1037/0003-066x.40.3.266

Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and audio signal processing: Processing and perception of speech and music* (2ª ed.). New Jersey, Estados Unidos: Wiley-Interscience.

Gordillo, C. D. (2013). Reconhecimento de voz contínua combinando os atributos MFCC e PNCC com métodos de robustez SS, WD, MAP e FRN. *Dissertação de mestrado, Pontifícia Universidade Católica do Rio de Janeiro*. Rio de Janeiro, RJ, Brasil. doi:10.17771/PUCRio.acad.23090

Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication, 49*(10-11), 787-800. doi:10.1016/j.specom.2007.01.010

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 1*(28), 75-105.

Iriya, R. (2014). Análise de sinais de voz para reconhecimento de emoções. *Dissertação de mestrado, Escola Politécnica da Universidade de São Paulo*. São Paulo, São Paulo, Brasil.

Jain, A. K., Duin, R. P., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 4-37. doi:10.1109/34.824819

James, W. (1884). What is an emotion? *Mind, os-IX*(34), 188–205. doi:10.1093/mind/os-IX.34.188

Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition – A brief history of the technology.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *D Appleton & Company*, 210-238. doi:10.1037/12203-005

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology, 15*(2), 99–117. doi:10.1007/s10772-011-9125-1

Koolagudi, S. G., Maity, S., Kumar, V. A., Chakrabarti, S., & Rao, K. S. (2009). IITKGP-SESC: Speech database for emotion analysis. In *Communications in Computer and Information Science* (pp. 485-492). Springer Berlin Heidelberg. doi:10.1007/978-3-642-03547-0_46

Lee, C. M., & Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing, 13*(2), 293-303. doi:10.1109/tsa.2004.838534

Lee, C.-C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication, 53*(9-10), 1162-1171. doi:10.1016/j.specom.2011.06.004

Likitha, M. S., Gupta, S. R., Hasitha, K., & Raju, A. U. (2017). Speech based human emotion recognition using MFCC. *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 2257-2260). Chennai, Índia: IEEE. doi:10.1109/WiSPNET.2017.8300161

Long, H., Guo, Z., Wu, X., Hu, B., Liu, Z., & Cai, H. (2017). Detecting depression in speech: Comparison and combination between different speech types. *2017 IEEE*

International Conference on Bioinformatics and Biomedicine (BIBM). IEEE. doi:10.1109/bibm.2017.8217802

Manson, N. J. (2006). Is operations research really research? *ORiON, 2*(22), 2224-0004. doi:10.5784/22-2-40

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 4*(15), 251-266. doi:10.1016/0167-9236(94)00041-2

Martins, J. A. (1997). Avaliação de diferentes tecnicas para reconhecimento da fala. *Tese (doutorado) - Universidade Estadual de Campinas. Faculdade de Engenharia Eletrica e de Computação*, (p. 161). Campinas. SP. Fonte: http://www.repositorio.unicamp.br/handle/REPOSIP/260759

Martins, R. M. (2014). Análise comparativa entre os métodos HMM e GMM-UBM na busca pelo α-ótimo dos locutores crianças para utilização da técnica VTLN. *Dissertação de mestrado, Instituto Nacional de Telecomunicações*. Santa Rita do Sapucaí, Minas Gerais, Brasil.

Miguel, F. K. (2015). Psicologia das emoções: uma proposta integrativa para compreender a expressão emocional. *Psico-USF, 1*(20), 153-162. doi:10.1590/1413-82712015200114

Milner, B., & Shao, X. (2006). Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end. *Speech Communication, 48*(6), 697-715. doi:10.1016/j.specom.2005.10.004

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017, 3 5-9). Automatic speech emotion recognition using recurrent neural networks with local attention. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2227-2231. doi:10.1109/ICASSP.2017.7952552

Motlícek, P. (2003). Feature extraction in speech coding and recognition. *Report of PhD research internship in ASP Group, OGI-OHSU*, 46.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication, 41*(4), 603-623. doi:10.1016/s0167-6393(03)00099-2

Petry, A. (2002). Reconhecimento automatico de locutor utilizando medidas de invariantes dinâmicas não-lineares. *Tese (doutorado) - Universidade Federal do Rio Grande do Sul. Instituto de Informática Programa de Pós-Graduação em Computação*, (p. 155). Porto Alegre. RS. Fonte: https://lume.ufrgs.br/handle/10183/3111

Pichora-Fuller, M. K., & Dupuis, K. (2020). Toronto emotional speech set (TESS). *DRAFT VERSION*. doi:10.5683/SP2/E8H2MF

Plutchik, R. (2002). *Emotions and Life: Perspectives from Psychology, Biology, and Evolution* (1 ed.). Washington, Estados Unidos: Amer Psychological Assn.

Prinz, J. J. (2007). Emotion: Competing theories and philosophical issues. In *Philosophy of Psychology and Cognitive Science* (pp. 247-266). Elsevier. doi:10.1016/b978-044451540-7/50025-6

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. *IEEE International Conference on Acoustics Speech and Signal Processing, 85*(9), 1437-1462. doi:10.1109/icassp.2002.5745552

Rong, J., Li, G., & Chen, Y.-P. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing & Management, 45*(3), 315-328. doi:10.1016/j.ipm.2008.09.003

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 5*(69), 379-399. doi:10.1037/h0046234

Schröder, M. (2001). Emotional speech synthesis: A review. *EUROSPEECH - 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event* (pp. 561-564). Aalborg, Denmark: EUROSPEECH.

Schuller, B. (2018). Speech emotion recognition. *Communications of the ACM, 61*(5), 90-99. doi:10.1145/3129340

Schuller, B., Arsić, D., Wallhoff, F., & Rigoll, G. (2006). Emotion recognition in the noise applying large acoustic feature sets. *Speech Prosody, Dresden*, 276-289.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., & Wagner, J. (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. *INTERSPEECH*, 2253-2256.

Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, (pp. 1520-6149). doi:10.1109/icassp.2004.1326051

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*, 312-315.

Singh, V., Jain, V. K., & Tripath, N. (2014). A comparative study on feature extraction techniques for language identification. *International Journal of Engineering Research and General Science, 2*, 286-291.

Subramaniam, L. V., Faruquie, T. A., Ikbal, S., Godbole, S., & Mohania, M. K. (2009). Business intelligence from voice of customer. *IEEE 25th International Conference on Data Engineering.* doi:10.1109/icde.2009.41

Subramaniam, L. V., Faruquie, T. A., Ikbal, S., Godbole, S., & Mohania, M. K. (2009). Business Intelligence from Voice of Customer. *IEEE 25th International Conference on Data Engineering.* doi:10.1109/icde.2009.41

Tahon, M., & Devillers, L. (2016). Towards a small set of robust acoustic features for emotion recognition: Challenges. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(1), 16-28. doi:10.1109/taslp.2015.2487051

Takeuchi, H., Subramaniam, L. V., Nasukawa, T., & Roy, S. (2007). Automatic identification of important segments and expressions for mining of business-oriented conversations at contact centers. *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical*, (pp. 458-467). Prague, República Tcheca.

Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies, 1*(1), 19-22.

Ververidis, D., & Kotropoulos, C. (2006). A state of the art review on emotional speech databases. *In Eleventh Australasian international.* Auckland, New Zealand,.

Wieringa, R. (2009). Design science as nested problem solving. *Association for Computing Machinery*, 12. doi:10.1145/1555619.1555630

Worren, N. A., Moore, K., & Elliott, R. (2002). When theories become tools: Toward a framework for pragmatic validity. *SAGE Publications, 10*(55), 1227-1250. doi:10.1177/a028082

Wu, C.-H., & Liang, W.-B. (2011). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *Institute of Electrical and Electronics Engineers (IEEE), 1*(2), 10-21. doi:10.1109/t-affc.2010.16

Wu, S., Falk, T. H., & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication, 53*(5), 768-785. doi:10.1016/j.specom.2010.08.013

Yang, B., & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal Processing, 90*(5), 1415-1423. doi:10.1016/j.sigpro.2009.09.009

Yeh, J.-H., Pao, T.-L., Lin, C.-Y., Tsai, Y.-W., & Chen, Y.-T. (2011). Segment-based emotion recognition from continuous Mandarin Chinese speech. *Computers in Human Behavior, 5*(27), 1545-1552. doi:10.1016/j.chb.2010.10.027

Yu, D., & Deng, L. (2014). *Automatic speech recognition a deep learning approach.* London: Springer. doi:10.1007/978-1-4471-5779-3

APÊNDICE A

Nessa seção será disponibilizado o código fonte do artefato produzido. A linguagem de programação utilizada para desenvolvimento foi a Python 3. Para reprodução do código é necessário instalar as seguintes bibliotecas:

- audiosegment==0.23.0
- matplotlib==3.3.3
- numpy==1.19.4
- PyAudio==0.2.11
- pydub==0.24.1
- scikit-learn==0.24.0
- spafe==0.1.2

A estrutura de diretórios do projeto é:

- src
  - modules
    - extract_features.py
  - main.py

Arquivo principal (main.py)

```
#!/usr/bin/python
# -*- coding: UTF-8 -*-

from __future__ import print_function, unicode_literals

import warnings
warnings.filterwarnings('ignore')

import sys
import os, glob, pickle
import numpy as np
from sklearn import svm, mixture
from sklearn.metrics import accuracy_score, classification_report,
confusion_matrix, f1_score

from modules import extract_features

#Emotions dataset
emotions={
```

```
    'a':'alegria',
    'c':'calma',
    't':'tristeza',
    'r':'raiva',
    's':'surpresa',
}

# Emotions to observe
observed_emotions=['alegria', 'calma', 'raiva', 'surpresa', 'tristeza',]

def create_dataset_train(feature):
    x,y=[],[]

    print("Create training dataset - feature %s" % (feature))

    for emotion in observed_emotions:
        for file in glob.glob("../data/treinamento/%s/*.wav" % emotion):
            filename = os.path.basename(file)

            if emotions.get(filename.split("_")[0]) is None:
                continue

            if feature=='mfcc':
                mfcc = extract_features.MFCC()
                dataset = mfcc.mfcc(file)
            elif feature=='plp':
                plp = extract_features.PLP()
                dataset = plp.plp(file)

            x.append(dataset)
            y.append(emotion)

            print('.', end='', flush=True)

    print('')

    return (np.array(x), y)

def create_dataset_test(feature):
    x,y=[],[]

    print("Create testing dataset - feature %s" % (feature))

    for file in glob.glob("../data/teste/*.wav"):
        filename = os.path.basename(file)

        if emotions.get(filename.split("_")[0]) is None:
```

```
            continue

        emotion = emotions[filename.split("_")[0]]

        if feature=='mfcc':
            mfcc = extract_features.MFCC()
            dataset = mfcc.mfcc(file)
        elif feature=='plp':
            plp = extract_features.PLP()
            dataset = plp.plp(file)

        x.append(dataset)
        y.append(emotion)

        print('.', end='', flush=True)

    return (np.array(x), y)

def create_dataset(feature, force=False):
    if os.path.isdir("../data/models") is False:
        os.mkdir("../data/models")

    dataset_file = ("../data/models/dataset_%s.dump" % feature)

    if force==True:
        if os.path.exists(dataset_file):
            os.remove(dataset_file)

    if os.path.exists(dataset_file) is False:

        x_train, y_train = create_dataset_train(feature)
        x_test, y_test = create_dataset_test(feature)

        #Get the shape of the training and testing datasets
        print((f'Files train: {x_train.shape[0]}'))
        print((f'Files test: {x_test.shape[0]}'))

        #Get the number of features extracted
        print(f'Features train extracted: {x_train.shape[1]}')
        print(f'Features test extracted: {x_test.shape[1]}')

        file = [x_train, y_train, x_test, y_test]

        with open(dataset_file, "wb") as f:
            pickle.dump(file, f)
```

```
def classifier_svm(feature):
    dataset_file = ("../data/models/dataset_%s.dump" % feature)

    if os.path.exists(dataset_file) is False:
        print("File %s not exists." % dataset_file)
        exit()

    with open(dataset_file, "rb") as f:
        data = pickle.load(f)

        x_train, y_train, x_test, y_test = data

        #Create a svm Classifier
        model = svm.SVC(kernel="poly", C=1, gamma=0.2, break_ties=True,
decision_function_shape='ovr')

        #Train the model using the training sets
        model.fit(x_train, y_train)

        #Predict the response for test dataset
        y_pred = model.predict(x_test)

        print("F1 Score: %f" % (f1_score(y_test, y_pred, labels=np.unique(y_pred),
average='micro')))

        print("Confusion Matrix:")
        print(confusion_matrix(y_test, y_pred))

        print("Classification Report:")
        print(classification_report(y_test, y_pred))

        #Calculate the accuracy of our model
        accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)

        #Print the accuracy
        print("Accuracy SVM: {:.2f}%".format(accuracy*100))

def classifier_gmm(feature):
    _y_pred = []
    dataset_file = ("../data/models/dataset_%s.dump" % feature)

    if os.path.exists(dataset_file) is False:
        print("File %s not exists." % dataset_file)
        exit()

    with open(dataset_file, "rb") as f:
        data = pickle.load(f)
```

```
        x_train, y_train, x_test, y_test = data

        n_classes = len(np.unique(y_train))

        # Try GMMs using different types of covariances.
        model          =          mixture.GaussianMixture(n_components=n_classes,
covariance_type='full', max_iter=500, n_init=3)

        # Train the other parameters using the EM algorithm.
        model.fit(x_train)

        y_pred = model.predict(x_test)

        for value in y_pred:
            if observed_emotions[value] is None:
                continue

            _y_pred.append(observed_emotions[value])

        f1score    =    f1_score(y_test,    _y_pred,    labels=np.unique(_y_pred),
average='micro')

        print("F1 Score: %f" % f1score)

        print("Confusion Matrix:")
        print(confusion_matrix(y_test, _y_pred))

        print("Classification Report:")
        print(classification_report(y_test, _y_pred))

        #Calculate the accuracy of our model
        accuracy=accuracy_score(y_true=y_test, y_pred=_y_pred)

        #Print the accuracy
        print("Accuracy GMM: {:.2f}%".format(accuracy*100))

def main():
    print("Create dataset MFCC")
    create_dataset("mfcc", False)

    print("Create dataset PLP")
    create_dataset("plp", False)

   print("Classifier SVM - MFCC")
    classifier_svm("mfcc")
```

```
    print("Classifier GMM - MFCC")
    classifier_gmm("mfcc")

    print("Classifier SVM - PLP")
    classifier_svm("plp")

    print("Classifier GMM - PLP")
    classifier_gmm("plp")

if __name__ == '__main__':
    main()
```

Classes de extração de características (extract_features.py)

```
#!/usr/bin/python
# -*- coding: UTF-8 -*-

from __future__ import print_function, unicode_literals

import logging
import audiosegment
import numpy as np
from spafe.features import mfcc, rplp
from sklearn.decomposition import PCA

class MFCC(object):
    def mfcc(self, filename):
        audio  =  audiosegment.from_file(filename).resample(sample_rate_Hz=16000,
sample_width=1, channels=1)
        framerate, signal = audio.frame_rate, audio.to_numpy_array()

        result=np.array([])
        pca = PCA(n_components=25)

        _mfcc   =   mfcc.mfcc(signal,   framerate,   num_ceps=25,   pre_emph=1,
pre_emph_coeff=0.97, win_len=0.025, win_hop=0.01, win_type='hamming', nfilts=25,
nfft=514,   low_freq=None,   high_freq=None,   scale='constant',   dct_type=2,
use_energy=False, lifter=25, normalize=1)

        mfccs = np.nanmean((pca.fit_transform(_mfcc)[:150]).T, axis=0)
        result = np.hstack((result, mfccs))

        return result


class PLP(object):
```

```
    def plp(self, filename):
        audio = audiosegment.from_file(filename).resample(sample_rate_Hz=16000,
sample_width=1, channels=1)
        framerate, signal = audio.frame_rate, audio.to_numpy_array()

        result=np.array([])
        pca = PCA(n_components=13)

        _plp = rplp.plp(signal, framerate, num_ceps=13, pre_emph=1,
pre_emph_coeff=0.97, win_len=0.025, win_hop=0.01, modelorder=13, normalize=1)

        plps = np.nanmean((pca.fit_transform(_plp)[:150]).T, axis=0)
        result = np.hstack((result, plps))

        return result
```