Neste trabalho foi desenvolvido um método para a criação de um banco de dados de voz para estudo do reconhecimento de emoções. Foram coletados áudios em um estúdio onde foram induzidas por meio de vídeos e atividades, seis emoções: felicidade, tristeza, raiva, nojo, medo e surpresa. Além disso foram coletados áudios de vídeos da internet e de filmes brasileiros e estrangeiros dublados em português brasileiro. Dos áudios das emoções induzidas foram selecionadas 200 frases para avaliação por 20 pessoas, tendo um desempenho de identificação positiva acima de 79%. Após a criação deste banco de dados para ilustrar a aplicabilidade do trabalho foram testados algoritmos de extração de parâmetros como a frequência fundamental, formantes e coeficientes mel-cepstrais, que foram utilizados em dois métodos de aprendizagem de máquinas, k-vizinhos mais próximos (KNN) e máquina de vetores de suporte (SVM).

Orientador: Dr. Aleksander Sade Paterno

JOINVILLE, 2019

ANO 2019 RAFAEL KINGESKI

DESENVOLVIMENTO DE UM BANCO DE DADOS DE VOZ COM EMOÇÕES EM IDIOMA PORTUGUÊS BRASILEIRO

UDESC

UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE CIÊNCIAS TECNOLÓGICAS – CCT
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA – PPGEEL

DISSERTAÇÃO DE MESTRADO

DESENVOLVIMENTO DE UM BANCO DE DADOS DE VOZ COM EMOÇÕES EM IDIOMA PORTUGUÊS BRASILEIRO

RAFAEL KINGESKI

JOINVILLE, 2019

RAFAEL KINGESKI

DESENVOLVIMENTO DE UM BANCO DE DADOS DE VOZ COM EMOÇÕES EM IDIOMA PORTUGUÊS BRASILEIRO

Dissertação submetida ao Curso de Pós-Graduação em Engenharia Elétrica, do Centro de Ciências Tecnológicas da Universidade do Estado de Santa Catarina, para a obtenção do Grau de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Aleksander Sade Paterno

JOINVILLE

2019

Ficha catalográfica elaborada pelo programa de geração automática da

Biblioteca Setorial do CCT/UDESC,

com os dados fornecidos pelo(a) autor(a)

Kingeski, Rafael
Desenvolvimento de um Banco de Dados de Voz com Emoções em Idioma Português Brasileiro / Rafael Kingeski. -- 2019.
100 p.

Orientador: Aleksander Sade Paterno
Dissertação (mestrado) -- Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação , Joinville, 2019.

1. Voz. 2. Banco. 3. Emoções. 4. Reconhecimento. I. Paterno, Aleksander Sade. II. Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação . III. Título.

Desenvolvimento de um Banco de Dados de Voz com Emoções em Idioma Português Brasileiro

por

Rafael Kingeski

Esta dissertação foi julgada adequada para obtenção do título de

MESTRE EM ENGENHARIA ELÉTRICA

Área de concentração em "Sistemas Eletroeletrônicos"
e aprovada em sua forma final pelo

CURSO DE MESTRADO ACADÊMICO EM ENGENHARIA ELÉTRICA
DO CENTRO DE CIÊNCIAS TECNOLÓGICAS DA
UNIVERSIDADE DO ESTADO DE SANTA CATARINA.

Banca Examinadora:

Prof. Dr. Aleksander Sade Paterno
CCT/UDESC (Orientador/Presidente)

Por videoconferência

Prof. Dr. Eduardo Furtado de Simas Filho
UFBA

Prof. Dr. Cristiano Mauro Assis Gomes
UFMG

Joinville,SC, 27 de setembro de 2019.

Dedico este trabalho aos meus pais: Benta e Luis Paulo

AGRADECIMENTOS

Agradeço ao professor Aleksander Sade Paterno pela orientação, a disposição, a paciência e a compreensão durante todo o processo de elaboração desta pesquisa.

À Larissa Aparecida Schueda pela dedicação e companhia durante os experimentos para coleta dos dados desta pesquisa.

À Thalia Cazarin por ter me ajudado durante todo este trabalho.

Aos professores e colegas do curso de pós graduação em engenharia elétrica (PPGEEL) pela troca de conhecimento e discussões.

Aos voluntários que dedicaram seu tempo e sua voz para este projeto.

À CAPES pelo subsídio financeiro na forma de bolsas de estudo.

À minha irmã Mônica, pela companhia durante estes anos de estudo.

Aos meus pais pelo carinho e apoio.

RESUMO

KINGESKI, Rafael. **Estudo de Reconhecimento de Emoções de um Banco de Dados de Voz em Idioma Português Brasileiro**. Dissertação (Mestrado Acadêmico em Engenharia Elétrica) – Universidade do Estado de Santa Catarina. Joinville, 2019.

Neste trabalho foi desenvolvido um método para a criação de um banco de dados de voz para estudo do reconhecimento de emoções. Foram coletados áudios em um estúdio onde foram induzidas por meio de vídeos e atividades, seis emoções: felicidade, tristeza, raiva, nojo, medo e surpresa. Além disso foram coletados áudios de vídeos da *internet* e de filmes brasileiros e estrangeiros dublados em português brasileiro. Dos áudios das emoções induzidas foram selecionadas 200 frases para avaliação por 20 pessoas, tendo um desempenho de identificação positiva acima de 79%. Após a criação deste banco de dados para ilustrar a aplicabilidade do trabalho foram testados algoritmos de extração de parâmetros como a frequência fundamental, formantes e coeficientes mel-cepstrais, que foram utilizados em dois métodos de aprendizagem de máquinas, k-vizinhos mais próximos (KNN) e máquina de vetores de suporte (SVM).

Palavras-chave: Voz, Áudio, Banco, Emoções, Indução de Emoções.

ABSTRACT

KINGESKI, Rafael. **Development of an Emotional Speech Database in Brazilian Portuguese**. (Mestrado Acadêmico em Engenharia Elétrica) – Universidade do Estado de Santa Catarina. Joinville, 2019.

In this work a method has been developed for the creation of a speech database for the study of emotion recognition. The audio recordings of vocal emotional expressions have been collected in a studio where where the vocal expression of emotions were induced while the subject watched videos or was required to perform specific tasks, six emotions: happiness, sadness, anger, anger, fear and surprise. Also was collect speechs from internet personal videos and from Brazilian movies and dubbed movie in Brazilian Portuguese. A number of six induced emotions were selected through a set of 200 sentences being evaluated by 20 people, with an identification performance of 79%. After the creation of this database, algorithms of features extraction such as fundamental frequency, formants, energy and mel-cepstral coefficients were tested, which were used in two machine learning methods, k-nearest neighbors (KNN) and Support Vector Machine (SVM).

Keywords: Voice, Audio, Database, Emotions, Emotion Induction.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A comunicação homem máquina tem evoluído cada vez mais com novas tecnologias. Com isso, surge a ideia de uma nova forma de comunicação que possua características humanas em máquinas. O processamento de sinais emitidos pelo corpo humano é capaz de reconhecer emoções, estas podem ser identificadas por sinais cardíacos, temperatura do corpo, tamanho da pupila, atividades musculares (DOUGLAS-COWIE; COWIE; SCHRÖDER, 2000), expressão facial e voz (EKMAN, 2011). Atualmente, há algoritmos para que computadores e máquinas possam reconhecer o que foi dito e por quem foi dito, porém como isto foi dito, ou seja, a identificação de qual a emoção expressada na fala não é algo tão popular como o reconhecimento da fala. A identificação de emoções pode ser aplicada em diversas tecnologias já conhecidas como computadores, *tablets*, *smartphones* e outros. Podendo ser aplicado em: detectores de mentiras, pesquisas sobre satisfação de clientes em um teleatendimento, robôs humanoides com reconhecimento de emoções por voz, jogos virtuais. Além disso, pode-se aplicar este tipo de tecnologia em sistemas de computação vestível, que além de monitorar estados emocionais pode prover um instrumento de suporte adicional no diagnostico e tratamento de transtornos mentais, tais como: ansiedade, síndrome do pânico, depressão, transtorno bipolar entre outros (MASSEY et al., 2009; TACCONI et al., 2008).

Alguns autores propõem dar emoções aos computadores. Baseado em teorias neurológicas é possível afirmar que emoções tem um grande papel, não apenas na criatividade e inteligência, mas também na racionalidade humana para tomar decisões. Sendo assim, uma forma de dar estas características a um computador seria dar-lhe emoções, ou pelo menos, a habilidade de detectar emoções (PICARD, 1997). Com isso, poder-se-ia dar ao computador uma forma de escuta empática, entendendo o que as pessoas sentem e comunicando este reconhecimento ao usuário (PICARD; KLEIN, 2002).

Através do processamento de sinais de voz, existem diversas técnicas em estudo para que se possa fazer a identificação de emoções. Para que se possa fazer um estudo baseado em emoções identificadas pela voz, faz-se necessário a criação de um banco de dados de sinais de voz que seja validado por humanos.

Os bancos de dados de voz utilizados para reconhecimento de emoções podem ser feitos com atores, indução de emoções, coleta de vídeos e filmes ou coleta de emoções em situações cotidianas. Todos estes já foram utilizados para este tipo de estudo (BURKHARDT et al., 2005; BARRA-CHICOTE et al., 2008; DOUGLAS-COWIE; COWIE; SCHRÖDER, 2000; SNEDDON et al., 2012).

Neste estudo o objetivo foi criar um banco de dados de voz, com resultado de desempenho da percepção humana e do desempenho de métodos de aprendizado de máquinas.

1.1 JUSTIFICATIVA

A inexistência de um banco de dados de voz similar no idioma Português Brasileiro foi a principal justificativa para este trabalho. Isso traz uma novidade a área de pesquisa relacionada a emoções em áudios de voz e processamento de sinais de voz. Sendo algo importante para estudos sobre computação afetiva, que é uma área de interface humano-computador que combina campos de conhecimento de psicologia, computação, engenharia, educação, sociologia entre outros.

1.2 OBJETIVOS

O objetivo deste trabalho é a criação de uma base de dados de voz para estudo de reconhecimento de emoções. Além disso outros objetivos são:

- Desenvolver um método para indução de emoções.
- Validação dos dados obtidos.
- Teste em algoritmos de extração de parâmetros.
- Teste em classificadores.

1.3 ESTRUTURA DO TRABALHO

Este trabalho é dividido em cinco capítulos. O primeiro traz uma breve introdução ao que se pretende realizar neste trabalho com algumas possíveis aplicações.

O segundo capítulo trata-se da revisão bibliográfica sobre a teoria das emoções e uma breve descrição de alguns dos bancos de dados de voz já existentes, a teoria dos sinais de voz, os parâmetros de voz e os classificadores que são utilizados para este estudo.

No terceiro, é descrito como foram coletados os dados, os equipamentos utilizados e o local onde foram feitas as gravações.

O quarto capítulo relatam-se os experimentos feitos com os classificadores, análise e discussão dos resultados.

No quinto capítulo são feitas as considerações finais, discussões gerais e possíveis trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo são apresentadas as teorias sobre emoções, com o objetivo de prover suporte técnico e teórico à escolha do respectivo conjunto de emoções para criação do banco de dados, os fundamentos do processamento de sinais de voz e os métodos de classificação utilizados para reconhecimento de emoções.

2.1 TEORIA DAS EMOÇÕES

Existem teorias sobre emoções desde a Grécia antiga, Aristóteles descreve em seu livro *Ética a Nicômaco* como as emoções podem influenciar as ações humanas (ARISTÓTELES, 1991; SCHMITTER, 2016). No século XIX, Charles Darwin estudou expressões causadas por emoções nos humanos e teve novos pareceres que o sustentavam pelo estudo de expressões que eram universais e não culturais. Mais tarde, a teoria de Darwin foi comprovada pelo estudo de expressões faciais de Paul Ekman. Ainda segundo EKMAN, P. (2011) existem gatilhos que nos causam emoções e mudanças em partes do nosso cérebro. Para lidarmos com a mudança que deflagrou a emoção o corpo humano sofre mudanças no sistema nervoso autônomo, tais como, o batimento cardíaco, a respiração, transpiração. Estes gatilhos causam também mudanças nas expressões, na face, na voz e na postura (EKMAN, 2011). A seguir serão apresentadas algumas teorias sobre emoções, apesar de serem diferentes todas elas tem um fator comum: as emoções são uma reação fisiológica causada por um fator externo.

2.1.1 As Emoções Segundo a Filosofia Moderna

Alguns filósofos do século XVII, criaram teorias sobre emoções que ganharam destaque e até hoje são utilizadas para estudo na neurociência e na psicologia. Os primeiros conceitos de emoções primárias surgiram nessa época. René Descartes assumiu que existem seis emoções, ou como ele cita em seu livro, paixões primárias. A admiração, o amor, o ódio, o desejo, a alegria e a tristeza eram as seis paixões primárias, as outras todas, segundo ele, eram uma mistura ou variação destas (DESCARTES, 2018). SPINOZA (2009) usava o termo afeto,e dizia que: "Afeto são as afecções do corpo, pelas quais sua potência de agir é aumentada ou diminuída, estimulada ou refreada, e, ao mesmo tempo, as ideias dessas afecções". A definição de afeto para SPINOZA (2009) é algo que pode ser comparada a definição de emoções, estas também provocam reações no corpo, no agir e no pensar. Para ele existiam apenas três afetos primários: a alegria, a tristeza e o desejo. Por exemplo, o amor seria a alegria, acompanhada da

ideia de uma causa exterior, e o ódio, a tristeza, acompanhada da ideia de uma causa exterior (SPINOZA, 2009). Já para Thomas Hobbes, existiam apenas sete paixões simples: apetite, desejo, amor, aversão, ódio, alegria e tristeza (HOBBES, 1983).

2.1.2 Teoria das Emoções de James-Lange

Os primeiros estudos no campo da psicologia sobre emoções foram de Willian James e de Carl Lange. Ambos tinham a mesma visão sobre emoções e por isso ela é chamada de teoria das emoções de James-Lange (DALGLEISH, 2004). Segundo eles, as emoções ocorrem em uma sequência: estímulo, reação física e então a emoção. Por exemplo, se você encontra um urso, seu coração acelera, você sai correndo e por consequência de ter fugido e o coração ter acelerado, você sente medo. Outro exemplo, seria a morte de um familiar ou alguém muito próximo, o comum seria pensar que a pessoa que está chorando, está triste porque perdeu alguém mas na verdade segundo a teoria de James-Lange a pessoa chora, então sente a tristeza porque está chorando e está chorando porque perdeu alguém. Isto é, segundo essa teoria a reação fisiológica ocorre antes do que a emoção (JAMES, 1884).

2.1.3 Teoria das Emoções de Cannon-Bard

Em 1927, um fisiologista chamado Walter Cannon, em oposição a teoria de James-Lange, publicou uma nova teoria. Nesta, ele afirmou que as atividades fisiológicas não podem provocar emoções, mais tarde Philip Bard complementou o estudo de Cannon. Com base em um estudo feito em animais, Cannon e Bard argumentaram que se as emoções fossem causadas pela percepção de mudanças do corpo, elas deveriam ser totalmente dependentes, precisando de seu cortéx sensorial e motor intactos. Eles propuseram que o fato de que a remoção do córtex, em gatos, não ter eliminado as emoções, significava que James e Lange estavam errados. A teoria de Cannon-Bard afirmava que as emoções e as reações fisiológicas eram simultâneas (CANNON, 1987; DALGLEISH, 2004).

2.1.4 Teoria das Emoções de Schachter-Singer

Stanley Schachter e Jerome E. Singer, ambos psicólogos americanos, publicaram em seu estudo sobre emoções que as elas dependiam de dois componentes: o cognitivo e o fisiológico. Por isso também é conhecida por teoria dos dois fatores da emoção.

Essa teoria fundiu as teorias de James-Lange e Cannon-Bard, ela afirma que não basta apenas uma resposta fisiológica ou uma interpretação cognitiva para desencadear e identificar

uma emoção. Por exemplo, se um indivíduo sente que seu coração acelerou ao realizar uma prova, é possível que ele interprete como uma emoção negativa, como medo ou angustia. Porém, se o seu coração acelerou ao saber que tirou a nota máxima na prova é bem provável que a emoção que ele está sentindo seja felicidade. Para uma mesma resposta fisiológica existem diversas emoções, ou seja, para identificar a emoção correta é necessário a composição da resposta fisiológica com a cognição (SCHACHTER; SINGER, 1962).

2.1.5 Os Postulados de Robert Plutchik

Robert Plutchik, psicólogo estadunidense, participou de um projeto de pesquisa relacionado com a registro de alterações fisiológicas em pacientes com transtornos mentais durante entrevistas psiquiátricas. Durante este trabalho, notou uma grande dificuldade: a incapacidade de avaliar as emoções. De particular importância foi o fato de que emoções eram sempre misturadas e difíceis de especificar ou desvendar. Essa observação sugeriu um possível paralelo entre mistura de emoções e mistura de cores. Fazendo uma analogia com cores, percebe-se que certos tons, como vermelho e verde ou amarelo e azul são complementares ou opostos. Isso também é verdade com emoções: alegria e tristeza, amor e ódio, aceitação e rejeição. Plutchik também concluiu que além da analogia pode-se fazer análise e síntese com as emoções. Neste caso, a análise é identificar elementos básicos de um conjunto, ou seja as emoções básicas de um conjunto e a síntese seria misturar este as emoções básicas para formar novas. (PLUTCHIK, 1962).

Para Pluthick a definição de emoção é:

> "Uma emoção pode ser definida como uma reação corporal padronizada de destruição, reprodução, incorporação, orientação, proteção, privação, rejeição ou exploração, ou alguma combinação destes, que é provocada por um estímulo" (PLUTCHIK, 1962, p. 176, tradução livre).

Plutchik criou a sua teoria das emoções e seus postulados. Nestes o o autor definiu as emoções básicas e algumas importantes definições sobre as emoções nos seres humanos. São eles:

- Postulado 1. Há um pequeno número de emoções puras ou emoções primárias.
- Postulado 2. Todas as outras emoções são uma composição das primárias, isto é, podem ser sintetizadas por várias combinações das emoções primárias.
- Postulado 3. As emoções primárias diferem umas das outras em relação à fisiologia e ao comportamento.

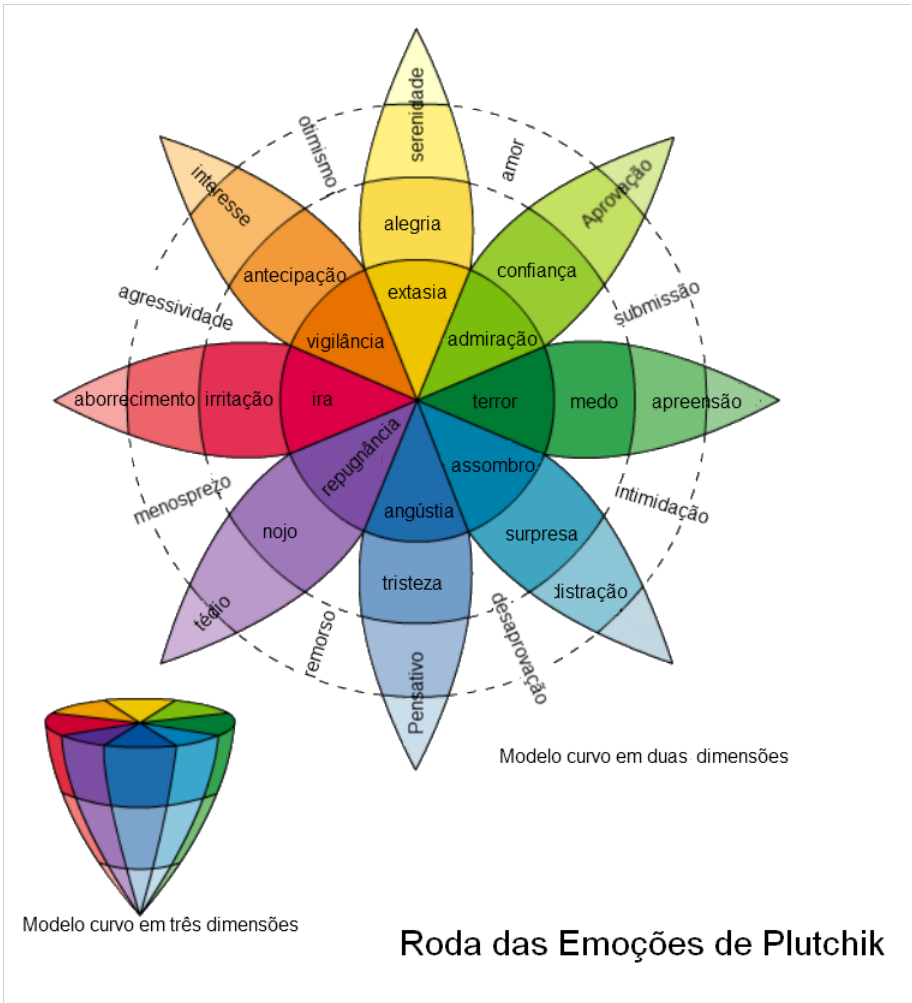
- Postulado 4. Emoções primárias em sua forma pura são construções hipotéticas ou estados idealizados cujas propriedades podem ser inferidas por vários tipos de evidências.

- Postulado 5. As emoções primárias podem ser conceitualizadas em termos de pares de opostos polares.

- Postulado 6. Cada emoção pode existir em graus variados de intensidade ou níveis de excitação.

Com base nestes postulados e na teoria das emoções, Plutchik define um diagrama de emoções onde se têm as emoções separadas por oito dimensões, ou oito padrões de comportamento separadas por pares de polos opostos. Uma representação deste diagrama pode ser observado na Figura 2.1 neste diagrama as emoções centrais são as mais fortes e as externas são mais fracas, no centro de cada divisão temos as emoções primárias: alegria, antecipação, confiança, irritação, nojo, tristeza, surpresa, medo, confiança.

Figura 2.1 – Diagrama de Estados Emocionais



Fonte: Página Robert Plutchik Wikipédia [1]

2.1.6 Emoções Segundo Paul Ekman

Segundo EKMAN, P. e FRIESEN, W. (1975), no estudo sobre emoções em expressões faciais, mostra-se que as principais emoções e mais fáceis de serem identificadas por expressões faciais são: medo, raiva, felicidade, tristeza, surpresa, e nojo. Ele cita também que em um diálogo existem duas formas de se identificarem emoções:

- Ao ouvir, você coleta informações de pelo menos três fontes no canal auditivo: as palavras reais usadas, o som da voz, e coisas como a rapidez com que as palavras são faladas, quando há pausas, quanto o discurso é interrompido por interjeições como *"aah"* ou *"ummh"*.

- Ao olhar, você coleta informações de pelo menos quatro fontes no canal visual: o rosto, as inclinações da cabeça, a postura corporal total e os movimentos musculares esqueletais dos braços, mãos, pernas e pés.

Cada uma dessas fontes no canal auditivo e visual, pode dizer algo sobre emoção (EKMAN; FRIESEN, 1975). Este trabalho, inspirou-se em um dos canais indicados por Ekman, para coletas das informações, processando a voz vinculada à expressão das emoções.

2.1.7 Além das Emoções

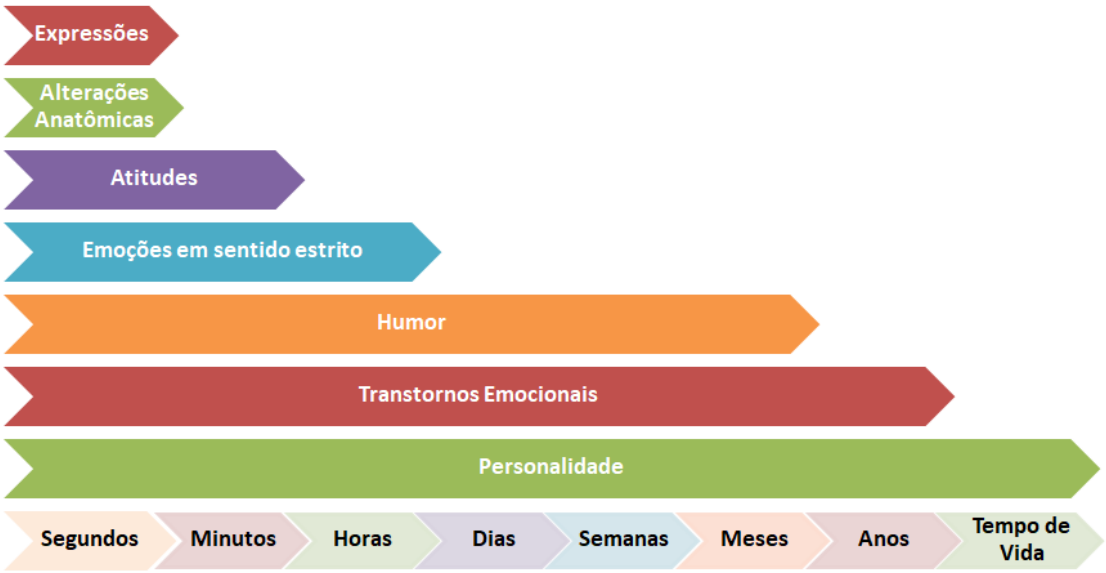
Nem todo estado emocional pode ser considerado como uma emoção isolada, sendo assim, alguns critérios utilizados para diferenciar emoções de humores pode ser o tempo, a intensidade e a presença ou ausência de um objetivo (COWIE et al., 2001).

Ainda segundo COWIE et al., (2001), estados emocionais são divididos em categorias que estão relacionados ao tempo. Emoção em seu sentido estrito é geralmente de curta duração e intenso. Humor descreve um estado emocional que é subjacente e relativamente prolongada. Traços emocionais são disposições mais ou menos permanentes para entrar em certos estados emocionais. Transtornos emocionais como depressão ou ansiedade patológica, também se enquadram na categoria dos estados emocionais que pode ser considerados uma emoção prolongada. E por fim as emoções expressas por toda a vida que descrevem a personalidade do individuo.

A Figura 2.2 ilustra onde cada estado emocional está, segundo a linha do tempo de vida de um ser humano. Pode-se perceber que emoções tem um tempo mais curto se comparado a outros estados emocionais.

[1]Disponível em: <https://pt.wikipedia.org/wiki/Robert_Plutchik> acesso em 10 de janeiro de 2019

Figura 2.2 – Diagrama de Estados Emocionais em Função do Tempo



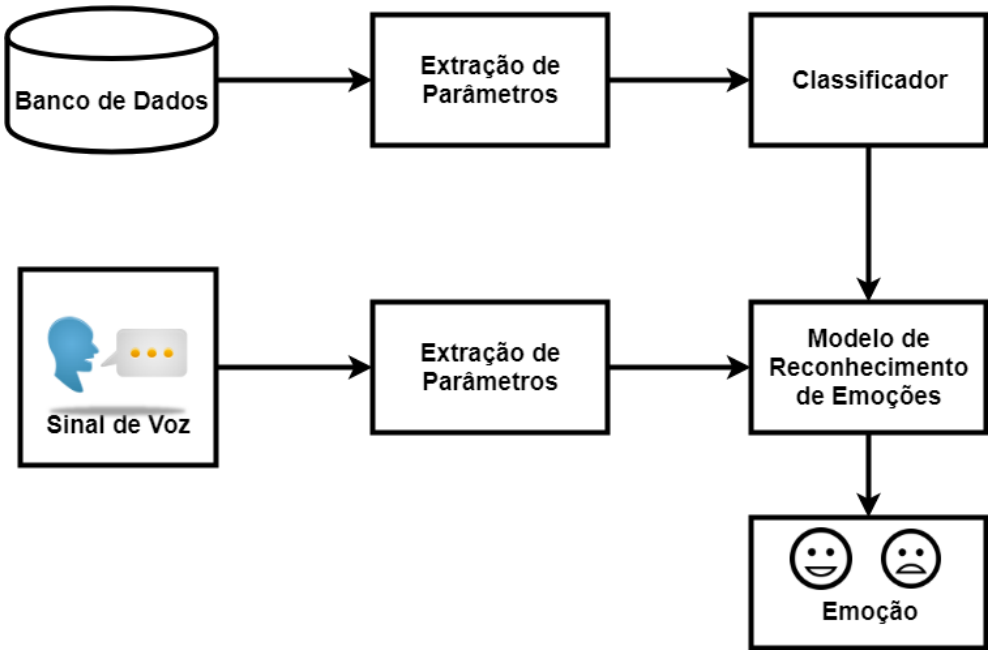
Fonte: Adaptado de (COWIE et al., 2001)

Para se observarem e coletarem alterações anatômicas de uma emoção deve ater-se então aos primeiros segundo depois do estímulo que causou a emoção, é nesse curto tempo que algumas alterações, como a voz, podem ser percebidas. Isso deve ser levado em consideração ao selecionar a janela de tempo dos áudios que serão analisados para criação do banco de dados.

2.2 BANCOS DE VOZ COM EMOÇÕES

Sabendo da necessidade da disponibilidade do material de análise obtido de forma padronizada e sistemática, a saber, sinais de voz avaliados como expressando as emoções, encaminha-se o problema a ser tratado neste trabalho. Existem diversos bancos de voz com emoções em outras línguas, em que foram utilizados diferentes metodologias próximas a proposta nesta pesquisa. Entre os já existentes, foram selecionados alguns para estudo a fim de se ter uma base para criação de um novo em português brasileiro. Na Figura 2.3 é dado o diagrama de um sistema de reconhecimento de emoções por voz, partindo do banco de dados desenvolvido, seguindo para onde são extraídos parâmetros e criado modelos de reconhecimento.

Figura 2.3 – Diagrama de um sistema para reconhecimento de emoções



Fonte: Próprio Autor

2.2.1 Banco de dados em Alemão - *Berlin Database of Emotional Speech*

Criado em 1999 o *Berlin Database of Emotional Speech*, banco de dados de voz em alemão foi construído como parte de uma pesquisa da *Technical University Berlin*. Gravado em um câmara anecoica, os áudios foram produzidos por 10 atores, 5 do sexo feminino e 5 do sexo masculino. Foram gravadas 5 frases curtas e 5 frases longas em 7 emoções, totalizando aproximadamente 800 frases. Este banco de dados está disponível na *internet*. As emoções que foram utilizadas neste banco foram: neutra, raiva, medo, alegria, tristeza, nojo e tédio. Os áudios gravados possuíam em seu conteúdo, material de texto sem sentido, como por exemplo, séries

de figuras ou letras aleatórias, ou palavras de fantasia. E também frases normais usadas na vida cotidiana. Baseado em um teste de percepção, 20 participantes colaboraram. Foram apresentados os enunciados em ordem aleatória na frente de um monitor de computador. Os participantes foram autorizados a ouvir cada amostra apenas uma vez antes de decidir, em qual estado emocional o locutor tinham simulado e como o desempenho era convincente (BURKHARDT et al., 2005).

2.2.2 Banco de Voz em Espanhol - *Spanish Expressive Voices*

Criado em 2008, este banco de dados contém gravações de fala e video de um ator e uma atriz que simulam as emoções: neutra, felicidade, tristeza, raiva, surpresa, medo e nojo. Foram gravados mais de 100 minutos por emoção, as gravações foram feitas em um estúdio, também foi gravado sinais de eletroglotografia[2]. e de vídeo para cada enunciado com o objetivo principal de permitir a pesquisa sobre detecção de emoção usando informações visuais, tais como, estudos de rastreamento facial, a possibilidade de estudo específico de linguagem corporal que poderia estar relacionado a características como nível de intensidade nos sinais de fala gravados ou dar informações relevantes de cada emoção. Além disso, fusão de sensores audiovisuais para identificação de emoções e mesmo o reconhecimento de fala afetiva é considerado como uma potencial aplicação. Todo o banco de dados foi validado por exames objetivos e perceptivos, atingindo uma pontuação de 89%. Foi avaliado usando uma *interface web*. Seis avaliadores para cada voz participaram da avaliação. Eles puderam ouvir cada enunciado quantas vezes eles desejassem (BARRA-CHICOTE et al., 2008).

2.2.3 Banco de Voz em Inglês - *The Belfast Data-base*

Este banco de dados foi criado em 2000, na cidade de Belfast na Irlanda do Norte, os áudios foram coletados das seguintes formas:

- Emoções livres - Foram utilizados alunos de pós graduação que possuíam uma intimidade, de forma que escolheram um assunto que causava uma emoção e então o discutiram, em um estúdio de gravação.

- Emoções induzidas por entrevista - Gravando vídeo com áudio, as técnicas de trabalho de campo foram baseadas em procedimentos padrão em sociolinguística. Em particular, o cuidado assumiu três questões. Primeiro, a configuração física foi feita como o mais informal possível (por meio de montagem de parede discreta câmeras, acessórios físicos,

[2] A eletroglotografia, ou EGG, é o sinal da vibração das cordas vocais medido com um instrumento não invasivo.

como mesa de café, etc.). Segundo, as gravações foram longas e em terceiro lugar, o entrevistador usou conhecimento prévio de cada assunto para adaptar a conversa. Cada sessão de entrevista seguiu o mesmo padrão amplo. O entrevistador começou com temas neutros (perguntas sobre a família, descrição de trabalho), em seguida mudou para tópicos positivos e, finalmente, para tópicos negativos. Tópicos positivos geralmente incluem feriados, sucessos infantis, nascimento de filhos e netos, relembrando momentos felizes e eventos. Tópicos negativos eram tipicamente problemas políticos na Irlanda do Norte, falecimento e problemas no trabalho.

- Programas de Televisão - A televisão foi a principal fonte de material envolvendo emoção relativamente forte. Após assistir uma série de programas ao longo de um período de vários meses, identificou-se alguns tipos de programas que eram potencialmente úteis. Todos lidaram com interações reais em vez de material de atuação. Os tipos de programas foram (i) conversar mostra, (ii) programas religiosos (iii) programas que traçam o vida de pessoas reais ao longo do tempo (iv) programas de assuntos atuais. Foram excluídos programas onde acreditou-se que havia um elemento de "encenação".

Os autores descrevem também que, deste material, de um total de 20 gravações de estúdio, 9 foram identificados como material utilizável. Cada um, em média, continha 3 ou 4 episódios que foram considerados como sendo emocionalmente marcantes em um período de quatro meses, 45 transmissões de televisão foram identificados como material utilizável. Dentro de cada um desses em média, houve 2 episódios que poderiam ser descritos como contendo fortes emoções (DOUGLAS-COWIE; COWIE; SCHRÖDER, 2000).

2.2.4 Banco de Vídeos em Inglês - *The Belfast Induced Natural Emotion Data-base*

O Banco de Dados de Emoções Induzidas Naturalmente de Belfast foi criado para fornecer exemplos de respostas emocionais leve a moderadamente forte a uma série de tarefas baseadas em laboratório. Os vídeos são curtos (5 a 60 segundos) com som estéreo. Cada tarefa foi escolhida para fornecer aos participantes um contexto fixo concebido, provocando um estado emocional forte o suficiente para revelar diferenças individuais, mas não tão forte a ponto de causar preocupações éticas. Embora as tarefas ocorram de forma relativamente artificial em um laboratório, descreve-se a emoção induzida como natural. Tarefas como assistir vídeos, e atividades com objetos em laboratórios, foram utilizadas para induzir estas emoções. Possui um total de 1400 vídeos, feitas por homens e mulheres da Irlanda do Norte e do Peru (SNEDDON et al., 2012). O banco de dados está disponível em: http://www.psych.qub.ac.uk/BINED/Default.aspx.

2.2.5 Banco de Voz em Português

Criado em 2010, possui um conjunto de frases semanticamente neutras e frases compostas por pseudopalavras (palavras sem significado). Foram produzidas por dois nativos europeus portugueses, variando as emoções para retratar raiva, nojo, medo, felicidade, tristeza, surpresa e neutralidade. Ao todo são 16 frases e 16 pseudofrases (frases compostas por pseudopalavras). Também foi feita a validação por taxas de precisão e tempos de reação em uma identificação dessas emoções, bem como os julgamentos de intensidade que foi coletado de 80 participantes. Foi atinginda uma precisão satisfatória de 190 frases e 178 pseudofrases. Classificando-se então como uma alta precisão, o acerto de 75% para frases e 71% para pseudofrases. Reconhecimento rápido e julgamentos de alta intensidade foram obtidos para todas as qualidades emocionais retratadas. Segundo os autores este banco de dados é útil ferramenta de pesquisa sobre prosódia emocional, incluindo estudos cruzados que envolvem o idioma Português Europeu, pode ser útil para fins clínicos na avaliação de pacientes com danos cerebrais (CASTRO; LIMA, 2010).

2.2.6 *DES - Danish Emotional Speech Database*

Banco de voz Dinamarquês, criado em 1995 na Universidade de Aalborg, composto por 4 atores (2 masculinos e 2 femininos) expressão 5 emoções (neutra, surpresa, felicidade, tristeza e raiva), cada uma com duração de 30 segundos totalizando 10 minutos de gravação. A avaliação humana dos áudios foi feita por 20 pessoas, o resultado das avaliações teve 67% de acerto na identificação das emoções (HANSEN, 1996).

2.2.7 *eNTERFACE'05 Audio-Visual Emotion database*

Criada em 2005 essa base de dados é composta por vídeos no idioma inglês, tem um total 1166 frases com 6 emoções (raiva, nojo, medo, felicidade, tristeza e surpresa). As emoções foram obtidas com a intenção de ser o mais próximo da reação emocional e não apenas uma atuação, para isto o autor utilizou pequenas histórias que poderiam induzir emoções, os locutores liam e depois eram gravadas 5 frases padrões para cada historia que correspondiam a uma emoção (MARTIN et al., 2006). Está disponível para *download* no site: http://www.enterface.net/results/

2.3 TEORIA DOS SINAIS DE VOZ

A linguagem verbal transmitida pela fala é utilizada para comunicação entre locutores e ouvintes. Por comando cerebral, manda-se informações para o aparelho vocal, que a transforma em ondas mecânicas, assim podem ser propagadas pelo ar até o aparelho auditivo, que envia estas informações ao cérebro dos ouvintes. Por ser um tipo de comunicação de preferência da maioria das pessoas, a comunicação verbal deixou de ser apenas humano-humano, atualmente, existem *interfaces* humano-maquina (HUANG; ACERO; HON, 2001). Estas possuem tanto o reconhecimento de fala, onde o computador reconhece o que foi dito, quanto a síntese de fala, onde o computador emite sinais de voz para o ser humano.

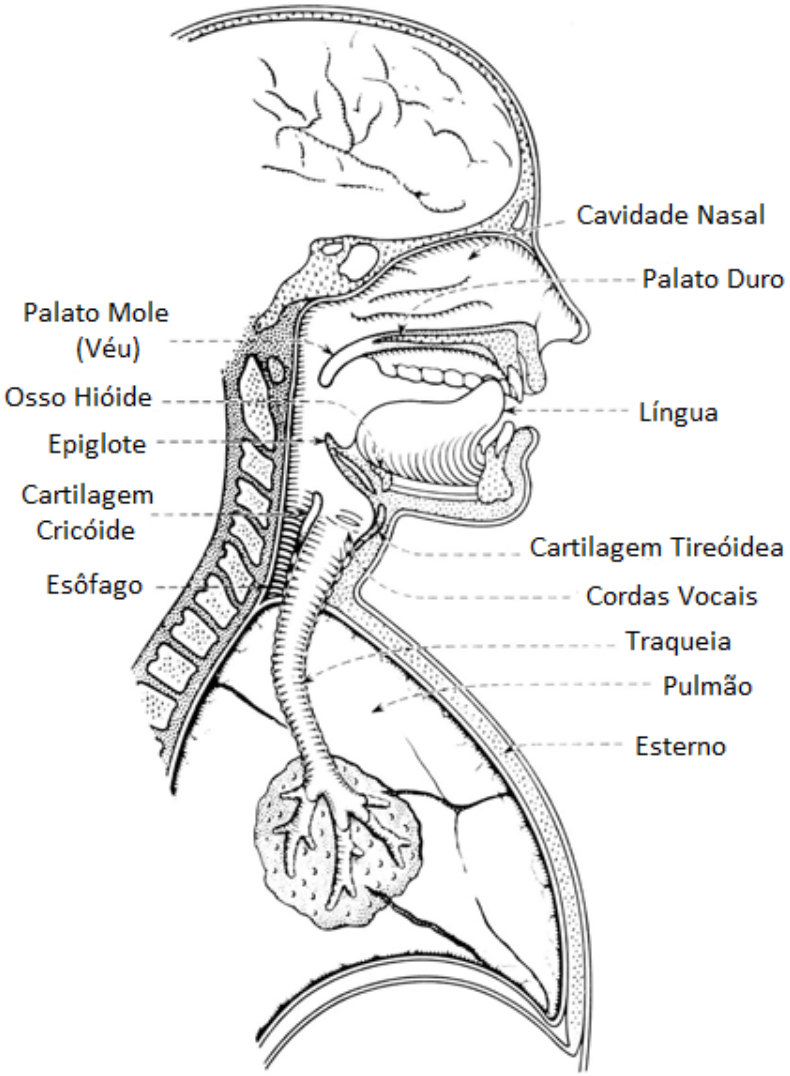
2.3.1 O Sinal de Voz

Os sons da fala de um ser humano são propagados por ondas acústicas que são produto de movimentos no aparelho respiratório e mastigatório. Na Figura 2.4 está ilustrado o aparelho vocal, que possui um tubo acústico não uniforme que se estende da glote (abertura entre as cordas vocais) aos lábios. A forma do aparelho vocal varia no tempo conforme o movimento dos lábios, língua e úvula. O aparelho nasal se estende das narinas à úvula, o controle do aparelho vocal e nasal é controlado pela úvula. A musculatura do tórax é a fonte de energia que por meio de contrações impulsiona o ar dos pulmões ao aparelho vocal para produção da fala (ALCAIM; OLIVEIRA, 2011).

Dependendo dos sons da fala a serem gerados, três mecanismos básicos de excitação podem estar envolvidos:

- Sons Sonoros: Como o som de /u/ em uva, as vibrações nas corda vocais são quase periódicas.
- Sons Fricativos Surdos: como som de s em sala, é criado por uma fonte de ruído contínuo com um espectro amplo e uniforme.
- Sons Oclusivos: como o som de /p/ em pato, consiste de um súbito desprendimento de um excesso de pressão gerado em alguma parte do aparelho vocal.

Figura 2.4 – Aparelho Vocal



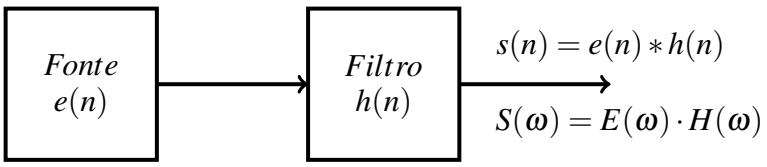
Fonte: (FLANAGAN, 1972)

2.3.2 Modelo de Produção da Fala

Para fazer uma representação matemática do processo de fala, pode-se utilizar o modelo de um sistema linear que é uma aproximação do modelo real, onde o sinal de voz é a resposta ao sistema, filtro (aparelho vocal) e fonte de excitação (RABINER; SCHAFER, 1980).

Na Figura 2.5 esta o diagrama de blocos deste modelo de produção da fala, onde o primeiro bloco, representa a fonte $e(n)$ e o segundo bloco o filtro $h(n)$, que representa o aparelho vocal. Isto significa que o sinal possui duas componentes, a fonte $e(n)$ relacionada a uma estru-

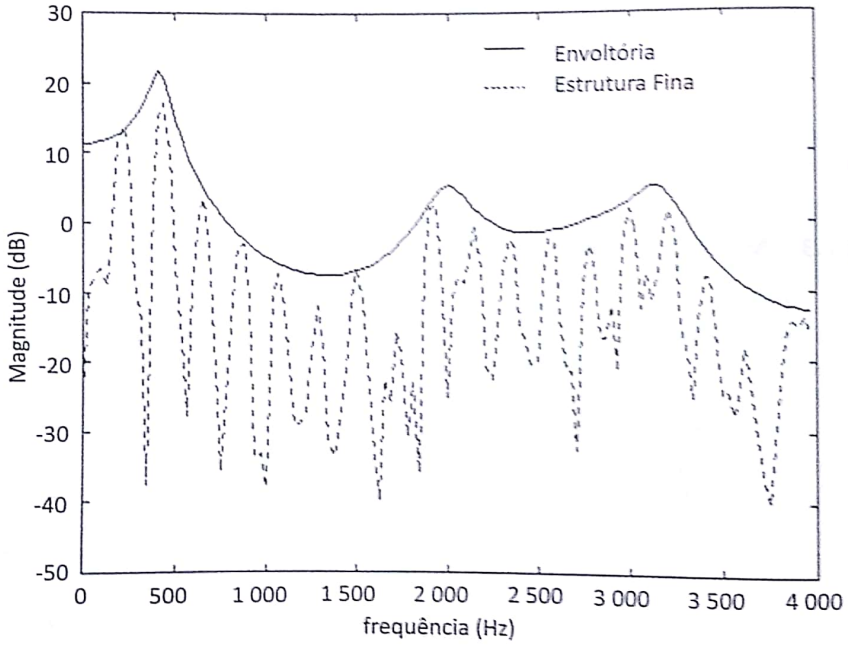
Figura 2.5 – Diagrama de Blocos do modelo da Produção de Fala



Fonte: Adaptado de (ALCAIM; OLIVEIRA, 2011)

tura fina (sinal pontilhado no gráfico da Figura 2.6) e o filtro $h(n)$ relacionado a uma envoltória espectral suave (linha contínua no gráfico da Figura 2.6). A estrutura fina da imagem possui um espaçamento aproximadamente equidistante, isto se da pela frequência fundamental do sinal de voz, já os picos que aparecem na envoltória suave estão relacionados as frequências de ressonância do filtro, que também são chamadas de formantes.

Figura 2.6 – Envoltória e estrutura fina em um intervalo curto de sinal sonoro.



Fonte: (ALCAIM; OLIVEIRA, 2011)

2.3.3 Parâmetros de Voz

Efeitos fisiológicos causados pelas emoções afetam o Sistema Nervoso Autônomo (SNA) e, por consequência, causam alterações na respiração, na salivação e na musculatura do peito,

da garganta e da cabeça, estas alterações acabam afetando diretamente as características do trato vocal (SCHERER, 1979). Alguns autores citam padrões de alteração nos parâmetros da voz, como consequência destes efeitos fisiológicos.

Na Tabela 2.1, MURRAY,I R; ARNOTT,J.L (1993) fez uma seleção de alterações mais comuns em alguns parâmetros conforme as emoções indicadas, essas alterações observadas foram relativas a fala de pessoas em estado neutro, ou seja, sem emoções.

Tabela 2.1 – Comparação de alterações na voz para cinco emoções diferentes

	Raiva	**Felicidade**	**Tristeza**	**Medo**	**Nojo**
Velocidade da Fala	mais rápida	rápida ou lenta	mais devagar	muito rápido	muito devagar
Média F_0	muito alta	muito alta	mais baixa	altíssima	baixíssima
Variação do F_0	muito alta	muito alta	menor	muito grande	mais larga
Intensidade	alta	alta	baixa	normal	baixa
Qualidade da Voz	voz soprada, voz de peito	voz soprada, estridente	ressonante	voz irregular	resmungada, voz de peito
Mudanças no F_0	abrupta	suave, inflexões para cima	inflexões para baixo	normal	grande, inflexões para baixo
Articulação	tensa	normal	arrastada	precisa	normal

F_0 - Frequência Fundamental
Fonte: (MURRAY; ARNOTT, 1993, p. 176, tradução livre)

Para SCHERER (1979) considerando a análise de efeitos fisiológicos, as emoções podem ser divididas em agradáveis e desagradáveis.

Verificação Intrínseca de Agradabilidade: é a forma mais básica de avaliação em todo o organismo, critério de agradável ou desagradável pode ser esperado tanto de características inatas quanto por associação prévia ou instruída (SCHERER, 1986).

Os efeitos de emoções desagradáveis parecem ser bem mais intensos: constrição e tensionamento da faringe, encurtamento do trato vocal, que levam a maior ressonância na região de alta frequência, estreitamento na largura de banda dos formantes, aumento do primeiro formante, queda do segundo formante e possivelmente do terceiro, causando assim um efeito de "voz estreita". Já os efeitos de emoções agradáveis são mais difíceis de prever, a expansão e o relaxamento faríngeo causa uma queda da frequência do primeiro formante e um amortecimento em altas frequências, porém uma laringe encurtada pode compensar este efeito do relaxamento faríngeo causando um equilíbrio em toda a faixa de frequência e produzindo uma estrutura harmônica clara, esse padrão pode ser chamado de "voz ampla" (SCHERER, 1986).

Alguns parâmetros de voz que já foram utilizados para teste de reconhecimento de emoções são:

- Frequência fundamental (VERVERIDIS; KOTROPOULOS; PITAS, 2004; MCGILLOWAY et al., 2000; HUANG; SONG; ZHAO, 2016; COWIE; DOUGLAS-COWIE,).
- Energia (VERVERIDIS; KOTROPOULOS; PITAS, 2004; MCGILLOWAY et al., 2000; HUANG; SONG; ZHAO, 2016; COWIE; DOUGLAS-COWIE,).
- Formantes (VERVERIDIS; KOTROPOULOS; PITAS, 2004; MCGILLOWAY et al., 2000; HUANG; SONG; ZHAO, 2016).
- Coeficientes mel-cepstrais (HUANG; SONG; ZHAO, 2016).

2.3.4 Frequência Fundamental

O *Pitch* ou frequência fundamental, é um dos parâmetros mais importantes para processamento de sinais de voz (RABINER; SCHAFER, 1980). Durante a fala, as cordas vocais podem estar em dois estados: o ritmo abrindo para um sons vozeados (com vibração) ou totalmente aberto para sons não vozeados (sem vibração). Se o som for vozeados, então a resposta expressa do trato vocal é um sinal periódico, que consiste em uma resposta a uma séries de impulsos. A distância entre os impulsos é a duração e o inverso é a frequência fundamental F_0 (KAMÍNSKA; PELIKANT, 2012). Existem diversas formas de se obter o *pitch*, tais como, autocorrelação, análise cepstral, coeficientes LPC, entre outras (CHENG, 1975; RABINER; SCHAFER, 1980). Para a classificação dos sinais obtêm-se paramêtros caracteristicos do *pitch*, tais como: máximo valor, mínimo valor, média, desvio padrão e variação (SHARMA; NEUMANN; KIM, 2002).

O conceito principal do *pitch*, independentemente de qual algoritmo é utilizado para a sua extração, é obter o valor da frequência fundamental de cada *frame*, ou a frequência de maior amplitude, deste *frame* (HUANG; ACERO; HON, 2001).

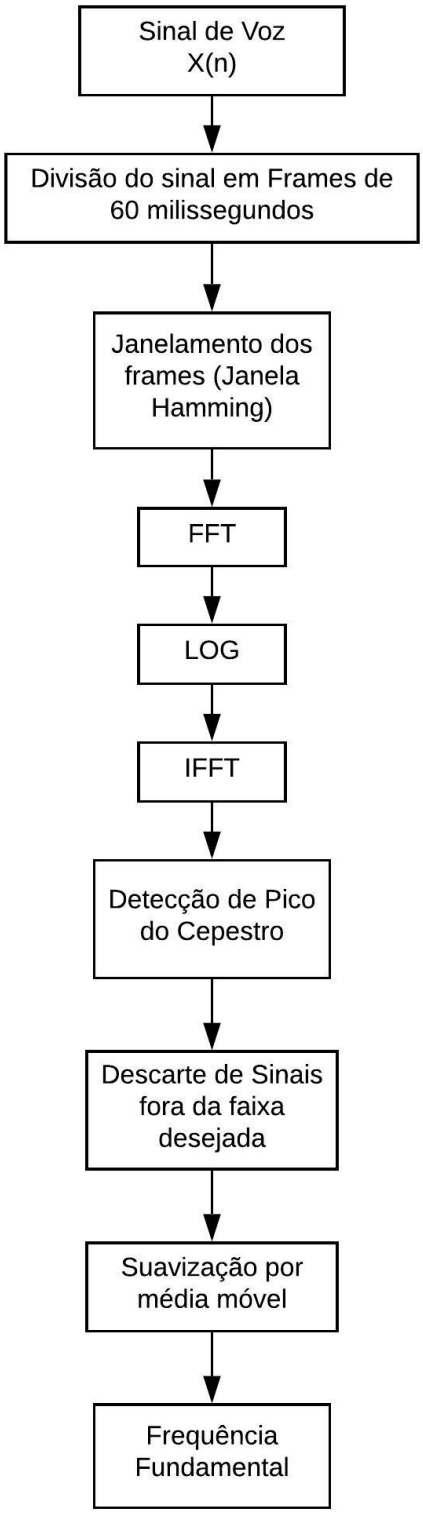
Segundo (HUANG; ACERO; HON, 2001), são típicos as seguintes dificuldades na extração deste parâmetro:

- Erros de sub-harmônicos: Se o sinal é periódico, com o período T, ele é também periódico em $2T$, $3T$... e assim por diante. Então, espera-se que os valores destas harmônicas nos algoritmos para a extração do *pitch* também seja alta.
- Erros de harmônicos: Se o harmônico com energia total do sinal dada por M for dominante, o valor no período T_0/M será maior, isso pode acontecer se o harmônico coincidir com a mesma frequência de algum dos formantes, tornando o sinal neste ponto muito maior, o que pode causar um erro na identificação do período fundamental do sinal.
- Ruídos: quando a relação sinal-ruído é baixa a extração do *pitch* é quase impossível.

- Laringealização da voz: Neste caso, a frequência fundamental varia bruscamente, nesta situação o *pitch* não é bem definido, e impor suavização de curva pode causar uma perda de informação. Em inglês o termo para definir a laringealização da voz é *Vocal Fry*.
- O *pitch* alterar bruscamente uma oitava acima ou abaixo.
- Voz soprada também prejudica a distinção entre o ruído e a frequência fundamental do sinal.
- Filtragem de banda estreita, causada por configurações do trato vocal pode fazer com que o sinal torne-se periódico.

A frequência fundamental, foi extraída baseado no algoritmo de determinação de pitch por análise cepstral proposta por (NOLL, 1967). O diagrama da Figura 2.7 mostra como o algoritmo foi implementado. Após fazer o janelamento dos *frames* e calcular o cepestro é feito uma detecção do pico do sinal que corresponde a frequência fundamental do sinal caso ele seja vozeado. O janelamento utilizado para os *frames* foi de 60 milissegundos e 50 milissegundos de sobreposição (BA et al., 2012). Esse janelamento corresponde a 83,33% de sobreposição e foi escolhida para melhorar a resolução em função do tempo, quanto maior a porcentagem de sobreposição mais janelas são utilizadas para separar o sinal em função do tempo, 60 milissegundos de janelamento corresponde uma resolução de 16,67 Hz no domínio da frequência, sendo útil neste caso já que fica dentro da faixa da frequência fundamental. A janela escolhida para este cálculo foi a janela de Hamming para redução dos lobos laterais e melhora da seletividade das frequências, neste caso um janelamento retangular não teria este mesmo desempenho, além disso a janela de Hamming é muito utilizada em processamento de sinais de voz pois além de melhorar a seletividade das frequências ele permite a reconstrução do sinal original (RABINER, 2010).

Figura 2.7 – Diagrama de extração da Frequência Fundamental



Fonte: Próprio Autor

2.3.5 Energia de Curto Termo

Em um sinal de voz a amplitude varia em função do tempo, ao analisar visualmente um gráfico da amplitude deste sinal em função do tempo onde há pausas na fala ou intervalos sem voz é possível notar que os sinais possuem um valor de amplitude muito inferior as demais partes, na região de menor amplitude em geral não há sinal de voz propriamente, podendo ser um ruído do ambiente ou do próprio equipamento de captura do sinal. A energia de curto termo representa as variações de amplitude de um sinal em função do tempo (RABINER; SCHAFER, 1980).

Para o cálculo da energia do sinal foi utilizado a equação 2.2. A Energia de curto termo é uma forma de representar a variação da amplitude do sinal em função do tempo (RABINER; SCHAFER, 1980). É definida por:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \tag{2.1}$$

E que pode ser reescrita da forma:

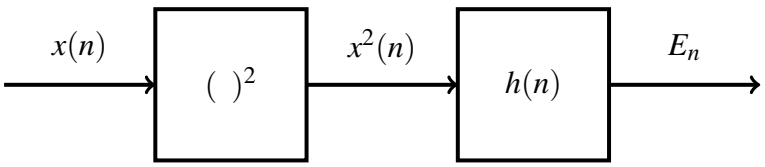
$$E_n = \sum_{m=-\infty}^{\infty} x(m)^2 \cdot h(n-m) \tag{2.2}$$

Onde:

$$h(n) = w^2(n) \tag{2.3}$$

A equação 2.2 pode ser interpretada como um sinal elevado ao quadrado que passa por um filtro linear, com resposta ao impulso dado por 2.3, como pode ser visto no diagrama de blocos da Figura 2.8.

Figura 2.8 – Diagrama de Blocos da Energia de Curto Termo



Fonte: Adaptado de (RABINER; SCHAFER, 1980)

Caso $h(n)$ na equação 2.2 seja uma janela retangular com uma largura L, dada pela equação 2.4 quanto maior for o valor de L mais suave será o sinal de E_n. Caso L for muito pequeno em relação ao sinal, ou L=1, então teremos o próprio sinal $x^2(n)$, ou seja, não teremos um sinal

suave. Este é um conflito muito comum em estudos de representação de sinais de curto termo (RABINER; SCHAFER, 1980).

$$W_R\left(e^{j\omega n}\right) = \sum_{n=0}^{L-1} = \frac{\sin\left(\omega L/2\right)}{\sin\left(\omega/2\right)} e^{-j\omega(L-1)/2} \tag{2.4}$$

A largura de banda da janela W_R é dada por F_s/L, para um sinal com frequência de amostragem $F_s = 44100Hz$, definindo uma janela de 20 milissegundos a largura da janela será de L=882 amostras.

A energia de curto termo é um parâmetro útil para reconhecimento de emoções, pois descreve uma característica temporal de uma emoção (VERVERIDIS; KOTROPOULOS, 2006; IRIONDO et al., 2000).

Para os ensaios neste trabalho foi utilizado um janelamento de 20 milissegundos, tipicamente este janelamento é feito com janelas entre 10 e 30 milissegundos (RABINER, 2010). Este tipo de parâmetro é dado no domínio do tempo conforme o tamanho da janela utilizada, neste caso então foi escolhido o valor médio entre o usual utilizado pela literatura. A janela utilizada para o cálculo de energia de curto termo foi a de Hamming para reduzir os lobos temporais.

2.3.6 Formantes

As frequências dos formantes, ou somente formantes, são as ressonâncias no trato vocal. As frequências de maior amplitude no espectro do sinal de voz definem os formantes. Uma das formas de se obter os formantes é pela análise de codificação lineares preditiva, ou LPC do inglês *Linear Predictive Coding*.

As Frequências Formantes foram calculados por codificação linear preditiva, foi utilizada a função para extração de LPCs do programa MATLAB, sendo o algoritmo de Levinson-Durbin usado para esta função (RABINER; SCHAFER, 1980). O janelamento utilizado para os *frames* foram os mesmos utilizados na extração da frequência fundamental, LPC de ordem 46.

O número de coeficientes para representar qualquer segmento de fala adequadamente é determinado pelo número de ressonâncias e antirressonâncias do aparelho vocal na largura de banda de interesse. Ela deve ser duas vezes maior que o comprimento de onda que passa da glote para os lábios. Para um trato vocal de aproximadamente 17,5cm de comprimento, haverá uma ressonância a cada 1000 Hz (RABINER, 2010). Para um sinal com frequência de amostragem de 44,1 kHz, deverá ter no mínimo 22 pares de polos para representar estas ressonâncias. Em geral é utilizado entre 1 a 2 polos a mais do mínimo para garantir a representação do sinal corretamente (HARRINGTON JONATHAN; CASSIDY, 2000).

2.3.6.1 Codificação de Predição Linear - LPC

Codificação Linear Preditiva ou LPC do inglês (Linear Prection Coding) é um tipo de codificação de sinais de voz, baseado em modelo, o modelo por sua vez é o modelo fisiológico do aparelho vocal descrito no item 3.1.1. Esse modelo divide a voz em duas categorias, sons sonoros e sons surdos, todas as vogais e algumas consoantes são consideradas sonoras pois possuem um frequência bem definida quando falamos, como por exemplo a pronúncia do *g* em gota e do *b* em boa. Varias consoantes são consideradas surdas, como por exemplo o *r*, na pronúncia da palavra roda e *p* em pulo (LATHI, 2012).

Sons sonoros tem em seu sinal de excitação uma frequência bem definida, esta é a frequência fundamental do sinal de voz, já os sons surdos possuem uma excitação que se assemelha a um ruído de banda larga.

O LPC foi desenvolvido como uma forma de melhorar a taxa de transmissão dos sinais de voz, ele é um transmissor muito mais eficiente se comparado a codificadores de forma de onda, como por exemplo a PCM (modulador por codificação de pulso) do inglês *pulse code modulation*, que necessita de uma taxa de transmissão de 64 kbits/s e a ADPCM (PCM diferencial adaptativa) com uma taxa de 32 kbits (LATHI, 2012). O LPC pode chegar a uma taxa de 2,4 kbits/s, obviamente a qualidade do sinal transmitido é reduzida (ELIAS, 1955; LATHI, 2012). Baseado no LPC foram desenvolvidos outras formas de codificação que melhoram essa perda de qualidade e que é utilizado em telecomunicações(ALCAIM; OLIVEIRA, 2011; LATHI, 2012).

O modelo LPC é baseado em prever uma amostra de voz pela análise de amostras anteriores. Isto pode ser descrito pela equação:

$$\hat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{2.5}$$

E o filtro que representa o modelo de produção da fala pode ser representado por uma equação de transferência com apenas polos, dada por:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{2.6}$$

Representando no domínio da transformada-z o sinal de excitação sendo $E(z)$ e o sinal de voz $S(z)$, que no tempo discreto é dado por $s(n)$. Assim tem-se:

$$E(z) = S(z) \cdot A(z) \tag{2.7}$$

logo, no domínio do tempo discreto podemos escrever:

$$e(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{2.8}$$

substituindo a equação 2.5 na equação, pode-se reescrever como:

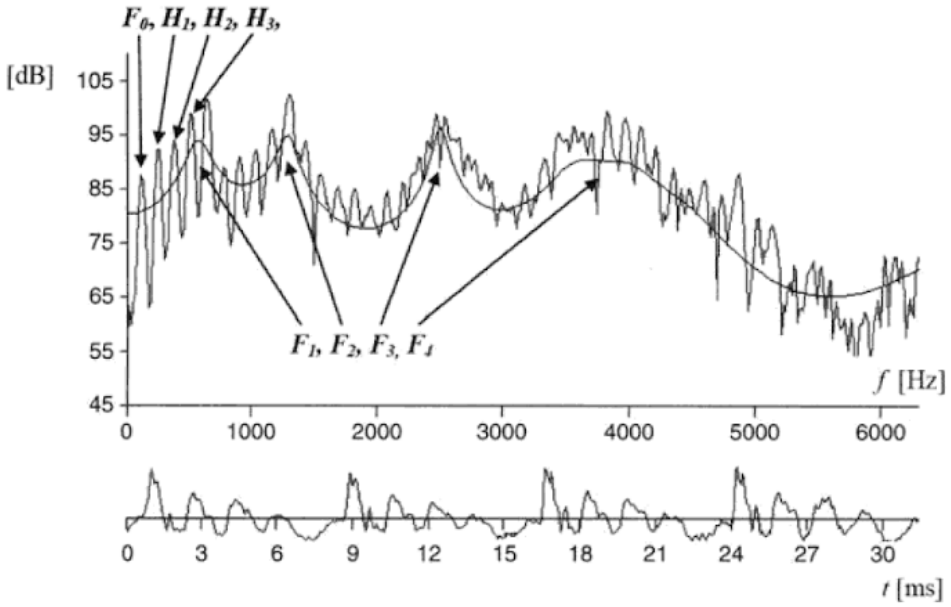
$$e(n) = s(n) - \hat{s}(n) \tag{2.9}$$

Logo, o sinal de excitação $e(n)$ é o erro da predição. Devido a necessidade de otimização do erro $e(n)$ para diminuir a taxa de transmissão, para o caso de uso em comunicações, foram desenvolvidos algoritmos que visam minimizar o valor médio quadrático do erro. Pode-se escrever esse valor médio quadrático como uma função dada por:

$$\alpha = \sum_{n=n_0}^{n_1} e^2(n) \tag{2.10}$$

onde os limites do intervalo são dados por n_0 e n_1.

Figura 2.9 – *Frame* de voz de 30 ms com envoltória espectral suave identificando os 4 primeiros formantes(F_1,F_2,F_3,F_4) a frequência fundamental F_0 e os harmônicos da frequência fundamental.



Fonte:(DUTOIT; MARQUES, 2010)

2.3.7 Coeficientes Mel-Cepstrais - MFCC

Os Coeficientes Mel-Cepstrais MFCC do inglês *Mel-Frequency Cepstrum Coefficients*, são parâmetros que primeiramente foram utilizados para o reconhecimento de voz, eles são baseados na resposta em frequência do ouvido humano.

O ouvido humano, representado na Figura 2.10 é divido em três partes:

- O ouvido externo, que é composto pelo pavilhão auricular, que reúne som e o conduz através do canal ao ouvido médio;
- O ouvido médio, começando no tímpano, e incluindo três pequenos ossos, o martelo, a bigorna e o estribo, que realiza uma transdução a partir de ondas mecânicas de pressão;
- O ouvido interno, que consiste na cóclea e no conjunto de conexões neurais ao nervo auditivo nervo, que conduz os sinais neurais para o cérebro.

Figura 2.10 – Ouvido humano

Fonte: Página do CTS [3].

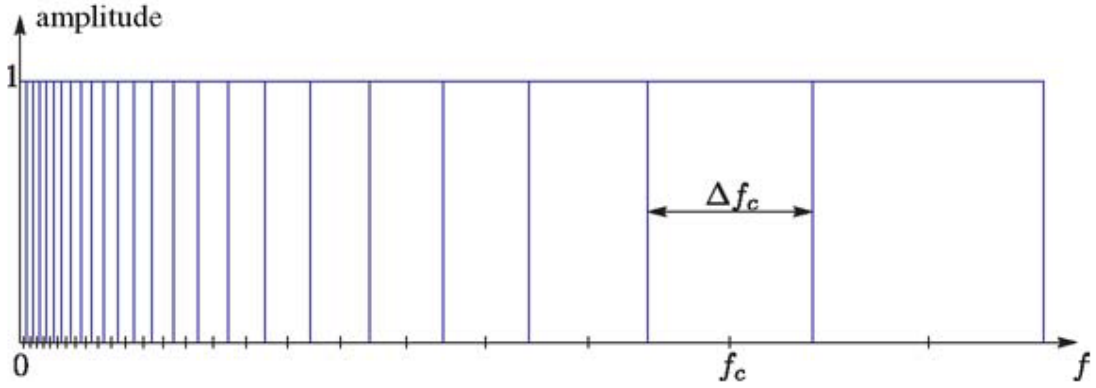
Entre a cóclea e o tímpano encontra-se a membrana basilar, responsável pela transferência de informação entre células ciliares em atividades neurais. A resposta não uniforme no domínio da frequência ocorre nesta membrana, suas respostas de frequência se sobrepõem já que os pontos na membrana basilar não podem vibrar independentemente. Mesmo assim, o conceito de análise de filtro passa-faixa na cóclea é bem estabelecida, e as larguras de bandas críticas foram definidas e medidas usando uma variedade de métodos, mostrando que as larguras de banda efetivas são constantes a cerca de 100 Hz para as frequências centrais abaixo de 500 Hz e com uma largura de banda relativa de cerca de 20% da frequência central acima de 500 Hz. Possui aproximadamente 25 filtros de banda crítica entre 0 Hz e 20 kHz (RABINER; SCHAFER, 2007).

[3]Disponível em: <https://www.cstsegurancadotrabalho.com/2016/11/como-funciona-o-sistema-auditivo.html> acesso em 10 de janeiro de 2019

Uma forma de representar esta resposta pode ser vista na Figura 2.11, a fórmula para a largura de banda de cada espaçamento Δf, é encontrada de forma empírica:

$$\Delta f_c = 25 + 75\left[1 + 1,4(f_c/1000)^2\right]^{0.69} \tag{2.11}$$

Figura 2.11 – Representação da largura de banda baseado na teoria de banda crítica do ouvido



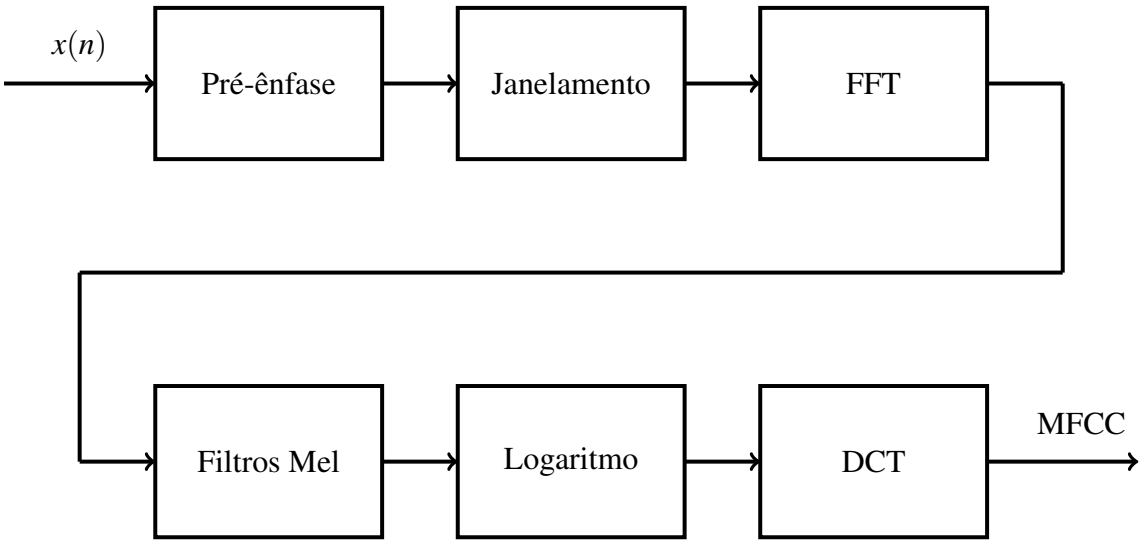
Fonte:(RABINER; SCHAFER, 2007)

Baseado na teoria de banda crítica do ouvido, Davis e Mermelstein (1980), criaram um novo método de representação cepstral, o MFCC.

Utilizando filtros triangulares, com espaçamento baseado na escala mel, à saída dos filtros foram aplicadas em uma função logarítmica e em seguida é extraído os valor dos coeficientes por uma Transformada Cosseno Discreta (DTC). (DAVIS; MERMELSTEIN, 1980). Diferente da Figura 2.11, que representa a largura de banda com a amplitude iguais e de valor 1, os filtros utilizados nesta técnica tem um formato triangular, e suas amplitudes não são iguais em todo o espectro. Como pode ser visto na figura 2.13, eles tem uma amplitude constante de 0 Hz à 1000 Hz e entre 1000 Hz à 4000 Hz os filtros tem um decaimento exponencial na sua amplitude. Além disso a largura dos filtros também só aumenta após os 1000 Hz.

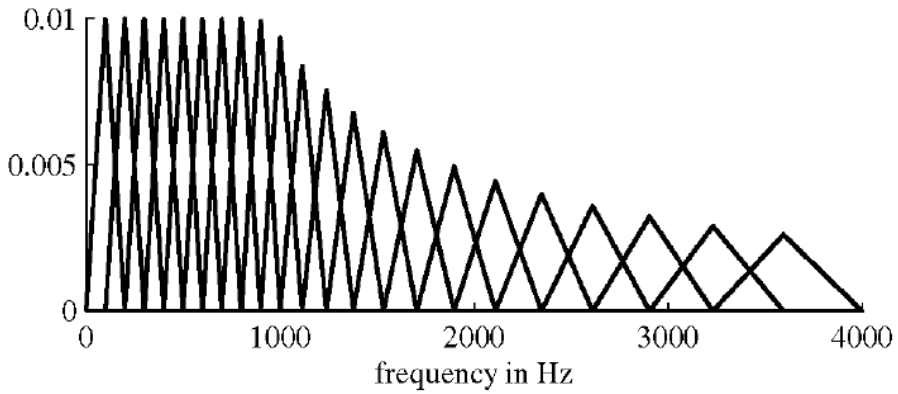
Foi utilizado a função de extração de coeficientes Mel-cepstrais -*mfcc* do programa MATLAB, o programa utiliza a estrutura baseada na rotina de Dan Ellis *rastamat* em seu algoritmo (ELLIS, 2005).

Figura 2.12 – Diagrama de Blocos da extração dos MFCCs



Fonte: Próprio Autor

Figura 2.13 – Representação dos filtros em escala Mel



Fonte:(RABINER; SCHAFER, 2007)

3 PARTE EXPERIMENTAL

3.1 COLETA DOS DADOS - BANCO DE VOZ

Após fazer um levantamento bibliográfico sobre bancos de voz já existentes e ainda algumas técnicas para criação destes, foi desenvolvido um método para criar o banco de dados de voz em português brasileiro. Utilizou-se três fontes para coleta: filmes, vídeos disponíveis na *internet* e indução de emoções em pessoas num estúdio de gravação.

3.1.1 Filmes

Para primeira coleta de dados e testes, foram utilizados filmes nacionais e filmes estrangeiros dublados em português brasileiro. Selecionou-se então 5 emoções básicas. São elas: felicidade, medo, neutra, raiva e tristeza. O motivo de ter escolhido apenas 5 emoções básicas foi por questão de quantidade de dados para análise, estas 5 emoções foram mais fáceis de selecionar, restringir um pouco das emoções é comum quando se tem uma fonte limitada. Com um total 141 áudios, sendo 85 femininos e 56 masculinos, extraídos de 1 seriado nacional, 14 filmes nacionais e 5 filmes dublados.

3.1.2 Vídeos da *Internet*

Como uma segunda forma de obter dados foram coletados áudios de vídeos da *internet* onde pessoas comuns que filmam seu cotidiano e publicam vídeos em diversas situações tais como: um desabafo sobre a perda de alguém, uma reação a algum vídeo nojento, em atividades como jogos virtuais entre outras. Nestes vídeos fez-se a seleção de 5 pessoas, do sexo masculino, e coletado 206 áudios. Coletou-se 6 emoções e um estado neutro.

3.1.3 Indução de Emoções

Geralmente em outros estudos a indução de emoções é feita por vídeos ou imagens (SNEDDON et al., 2012). Outras formas, sendo uma delas agressões físicas, como choques ou frio extremo, também são utilizadas em alguns experimentos, porém os efeitos da agressões psicológicas e físicas podem se sobrepor, mas eles não são idênticos. É portanto, mais importante separar processos psicológicos de fisiológicos para fins de análise (LAZARUS et al., 1962).

James Gross (1995) utilizou vídeos para indução de emoções. Estes, foram escolhidos depois de uma pesquisa, entre colegas sobre quais filmes causaram diferentes emoções, alguns vídeos de outros pesquisadores que também tinham o mesmo objetivo, incluídos para que um grupo pudesse ver e qualificar. Destes selecionou-se 16 melhores, ou seja, os mais emocionantes segundo o relato de cada participante, na grande maioria vídeos curtos retirados de filmes, onde a cena causava uma emoção de forma bem definida (GROSS; LEVENSON, 1995).

Outra forma de induzir emoções é por tarefas diversas, como feito no banco de dados de Belfast (SNEDDON et al., 2012). Onde foram utilizados caixas pretas, vídeos e um jogo de Nervo-Teste, que consistem em passar uma argola por um arame tortuoso sem encostar a argola no arame.

Figura 3.1 – Imagem de uma participante durante os experimentos de coleta de áudios em estúdio.

Fonte: Próprio Autor

Para construção deste banco de dados, foram escolhidas seis emoções para serem induzidas através de métodos adaptados de estudos similares, além de um método novo que teve o resultado esperado, a indução do medo. A escolha destas seis emoções foi baseada nas teorias de emoções básicas ou primárias de Paul Ekman (EKMAN; FRIESEN, 1975). Esta seleção se justifica pelo aspecto acessível de indução e identificação. Foi gravado também um estado neutro, ou seja, um estado onde a pessoa não estava expressando nenhuma emoção.

Este trabalho foi aprovado pelo comitê de ética com CAAE:83106617.5.0000.0118 Estudo de Reconhecimento de Emoções de um Banco de Dados de Voz em Idioma Português Brasileiro RAFAEL KINGESKI - FUNDAÇÃO UNIVERSIDADE DO ESTADO DE SC - UDESC.

As gravações foram feitas no estúdio da Rádio UDESC de Joinville. Na Figura 3.1 uma locutora do sexo feminino assistindo aos vídeos selecionados para induzir as emoções.

O banco de dados chama-se Emossônico. Os dados foram divulgados no site: <https://www.udesc.br/cct/geb/projetos/emossonico>. Neste há uma breve apresentação sobre o projeto, com os scripts utilizados para extrair os parâmetros e modelar os classificadores.

3.1.3.1 Neutro

O estado neutro foi coletado das conversas iniciais, supondo que a pessoa não estava sentindo nenhuma emoção. Para confirmar isto, também foi questionado, sobre estar sentindo algo que pudesse considerar como emoção, caso a resposta fosse negativa então era considerado neutro.

3.1.3.2 Felicidade

A felicidade foi induzida por meio de vídeos de comediantes contando piadas e vídeos de quedas engraçadas. Similar ao vídeo proposto em (GROSS; LEVENSON, 1995; SNEDDON et al., 2012). Os vídeos assistidos foram:

- Vídeo com o título: *"PARKOUR FAILS OF JULY 2018 | HE ALMOST DIED | PARKOUR FAIL COMPILATION"* (PARKOUR..., 2018) mostra uma série de quedas de pessoas.
- Vídeo de *Stand up comedy*, um comediante em um palco conta piadas e histórias engraçadas. Foram apresentados 3 vídeos dessa categoria. Com os títulos: "Michelly Summer - Terça Insana - 15/10/2013 (HD - By Alan Junior)", "FABIANO CAMBOTA - *Comedy Central* #2 (Paraguai)", "*Low Comedy Central* Apresenta Angela Dip e Paulo Vieira"(MICHELLY..., 2013; FABIANO..., 2016; LOW..., 2016).

3.1.3.3 Tristeza

Foram selecionado alguns vídeos para que deixasse as pessoas tristes, baseando se no que já foi proposto por (GROSS; LEVENSON, 1995; SNEDDON et al., 2012). Usou-se animações e filmes com morte de animais e familiares de personagens. Sendo eles:

- Um cachorro sendo sacrificado por um veterinário enquanto o dono conversava e se despedia de seu cachorro (VIGILANTE..., 2014).
- Curta metragem de animação escrito e dirigido por Pedro Solís García, com o título "Cordas", a história é sobre a amizade de duas crianças, sendo uma invalida, no final a criança inválida falece. Recebeu o Prêmio Goya de melhor curta-metragem de animação em 2014 (CUERDAS, 2013).

- Vídeo que mostra uma história feita em tira de quadrinhos pelo autor Gearboy com o título *"A Little Gaming Moment..."*. Conta a história de uma família que passa por diversas perdas, animais de estimação, dificuldade financeira e falecimento de algumas pessoas. Possui uma música de fundo, não possui fala nem legenda, apenas mostra cada quadrinho conforme passa o tempo do vídeo (LIFE. . . , 2014).

- Filme "O Presente" (PRESENT. . . , 2014) Desenho animado dirigido por Jacob Frey que conta a história de um menino sem a perna que ganha um cachorro que também não possui uma pata, o vídeo mostra a rejeição e aceitação do menino com o cachorro. A história é baseada no livro "Perfeição" do quadrinista brasileiro Fábio Coala. O autor disponibilizou o filme em: https://vimeo.com/152985022.

3.1.3.4 Nojo

Para induzir nojo, escolheu-se um video composto por um compilado de outros pequenos vídeos nojentos com duração de 15 minutos. Título do vídeo: *"Disgusting! Parasites zits insects in people's ears more prepare to lose your lunch - TomoNews."*

Descrição dos vídeos:

1. Extração de piolhos com um pente fino da cabeça de uma criança.
2. Pessoas espremendo um abcessos (2 vídeos).
3. Remoção de um rato pelo umbigo de um homem.
4. Remoção de um inseto gigante da orelha de uma pessoa (2 vídeos).
5. Minhoca expelindo um excremento.
6. Verme parasita saindo de uma barata.

3.1.3.5 Medo

Para o medo foram utilizados duas formas de indução. Uma baseada no banco de dados de Belfast (SNEDDON et al., 2012), composta por uma caixa preta com imagem de duas aranhas e um aviso de cuidado, além de avisos de choque, com um orifício tapado por um tecido, onde pedia-se então para colocar o braço, tatear e descrever verbalmente o que tinha na caixa. Dentro foi inserido pedaços de barbante e papeis colados para provocar uma sensação de ter algo dentro, provocando assim medo de encontrar uma aranha ou levar um choque na caixa.

Para o outro método de indução de medo, utilizou-se um óculos de realidade virtual para celular, um controle sem fio e fones de ouvido. O participante era desafiado por um jogo de terror com estes equipamentos e descrevia verbalmente detalhes sobre o jogo.

Figura 3.2 – Imagem da caixa utilizada para indução de medo

Fonte: Próprio Autor

Figura 3.3 – Óculos de Realidade Virtual

Fonte: Próprio Autor

Figura 3.4 – Controle Bluetooth

Fonte: Próprio Autor

3.1.3.6 Surpresa

Os vídeos utilizados para indução de surpresa foram:

- Sete apresentações no programa de televisão *"Got Talent"* que causaram surpresa ao público do auditório do programa. Título do vídeo: *"Most SURPRISING AUDITIONS on Got Talent! Got Talent Global"* (MOST..., 2017).
- Uma compilação de 22 vídeos com situações inesperadas diversas, como um carro partindo ao meio em movimento, um cachorro brincando com um tigre entre outros, o tempo total destes video é de 4 minutos e 46 segundos. Título: *"Unexpected Compilation"* (UNEXPECTED..., 2016).

3.1.3.7 Raiva

A atividade para indução de raiva tratou-se de um jogo eletrônico, *I Wanna Be the Guy*, também conhecido como IWBTG. É um jogo eletrônico de plataforma 2D independente lançado em 2007, desenvolvido pelo designer Michael Kayin O'Reilly. O jogador controla um personagem num mapa cheio de armadilhas que tornam o jogo difícil (O'REILLY, 2007).

Foi escolhido para induzir raiva por ser um jogo de nível alto e também, por possuir diversas situações inesperadas, tornando-o cansativo e irritante.

3.1.4 Resultados do experimento

Na Tabela 3.1 é descrito cada atividade e a respectiva emoção que cada participante mencionou ter durante estas. As atividades foram distribuídas, as que estão marcadas com "x"não foram realizadas por aquele participante. Nem todos tiveram o resultado esperado, sendo que

alguns descreveram outras emoções ou nenhuma emoção, como no caso do participante 1 (P1) que se sentiu surpresa e com agonia durante a atividade de nojo.

Tabela 3.1 – Descrição pessoal do locutor sobre sua emoção durante o experimento de indução

	Sexo	**Vídeo de Nojo**	**Vídeos Tristeza**	**Vídeos Felicidade**	**Video Surpresa**	**Jogo Medo**	**Caixa medo**	**Jogo Raiva**
P 1	**Feminino**	Surpresa, Agonia	Tristeza	Felicidade	Sem emoção	x	x	x
P 2	**Feminino**	nojo	Tristeza	Felicidade	x	x	x	x
P 3	**Feminino**	nojo	Sem emoção	Felicidade	x	Medo	x	x
P 4	**Masculino**	nojo	Sem emoção	Felicidade	x	Sem emoção	x	x
P 5	**Masculino**	nojo	Sem emoção	Felicidade	x	Sem emoção	x	x
P 6	**Feminino**	nojo	Tristeza	Felicidade	x	x	x	x
P 7	**Masculino**	nojo	Tristeza	Felicidade	x	Medo	x	x
P 8	**Masculino**	nojo	Tristeza	Felicidade	x	Medo	x	x
P 9	**Masculino**	nojo	Tristeza	Felicidade	Surpresa	Medo	x	x
P 10	**Masculino**	nojo	Tristeza	Felicidade	Surpresa	x	x	x
P 11	**Masculino**	x	Sem emoção	Felicidade	x	x	x	Raiva
P 12	**Masculino**	x	Tristeza	Felicidade	x	x	x	Raiva
P 13	**Masculino**	x	Sem emoção	Sem emoção	x	x	x	Raiva
P 14	**Feminino**	x	Tristeza	Felicidade	Surpresa	x	x	x
P 15	**Feminino**	x	Tristeza	Felicidade	Surpresa	x	x	Raiva

3.1.5 Validação dos dados

Para medir a qualidade dos áudios coletados foi realizado testes de percepção das emoções. Vinte pessoas escutaram aleatoriamente e depois de ouvir cada áudio o número de vezes que considerasse necessário, puderam avaliar qual a emoção do locutor.

Na Tabela 3.2, é dada a matriz de confusão da avaliação dos áudios femininos, a menor taxa de acerto foram para estado neutro e para felicidade. A melhor taxa de acerto foi para tristeza, com 93%.

Na tabela 3.3, é dado a matriz de confusão da avaliação dos áudios masculinos, neutro e surpresa ficaram com pior desempenho e felicidade o maior com 93%, Comparando o desempenho dos áudios femininos e masculinos, o neutro possui uma das menores taxas de acerto em ambos. A raiva ficou com valores iguais em ambos os sexos com 78%.

Tabela 3.2 – Matriz de confusão - Teste de Percepção - Feminino

Emoções induzidas em estúdio - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist.
fel.	**71**	14	14	0	0	0	0
medo	7	**79**	0	7	0	7	0
neutro	22	11	**67**	0	0	0	0
nojo	0	0	0	**75**	0	15	10
raiva	0	0	0	0	**78**	17	6
surp.	0	0	0	4	4	**88**	4
trist.	0	0	0	0	7	0	**93**

Tabela 3.3 – Matriz de confusão - Teste de Percepção - Masculino

Emoções induzidas em estúdio - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist.
fel.	**93**	4	0	0	4	0	0
medo	11	**85**	0	0	4	0	0
neutro	0	0	**71**	0	29	0	0
nojo	0	0	0	**73**	0	20	7
raiva	0	4	0	0	**78**	4	13
surp.	5	0	0	16	5	**74**	0
trist.	6	0	0	12	0	0	**82**

Tabela 3.4 – Matriz de confusão - Teste de Percepção - Masculino

Masculino - Videos *Internet*							
%	fel.	medo	neutro	nojo	raiva	surp.	trist.
fel.	**92**	0	0	0	4	0	4
medo	0	**95**	0	0	0	5	0
neutro	8	0	**71**	0	17	0	4
nojo	0	0	0	**96**	0	0	4
raiva	4	0	8	0	**75**	0	13
surp.	0	0	0	4	0	**96**	0
trist.	0	0	0	0	4	0	**96**

3.2 EXTRAÇÃO DE PARÂMETROS

Selecionou-se para utilizar no classificador a média, o mínimo, o máximo e o desvio padrão para cada um dos parâmetros. Utilizou-se a média, o mínimo, o máximo e o desvio padrão da primeira e segunda derivada de alguns parâmetros.

Para garantir a independência de locutor, foi feita uma normalização dos parâmetros. Esta normalização foi feita com base na fala neutra de cada locutor. Na equação:

$$P_j = \frac{P_i}{P_n} \tag{3.1}$$

Onde, P_j é o vetor parâmetro normalizado, P_i é o vetor parâmetro sem normalização e P_n é o vetor de parâmetros do neutro de cada locutor.

3.2.1 Seleção dos parâmetros

Para aplicar os dados em um modelo de classificação, convém selecionar quais são os melhores parâmetros, isto é, os parâmetros que pertencem a um subgrupo que otimiza uma função objetivo. Esta seleção pode ser feita de duas formas: por transformação do conjunto de observações de variáveis ou por seleção de um subconjunto de variáveis.

Para a seleção de parâmetros temos a seguinte definição: Dado um conjunto de parâmetros $X = \{x_i | i = 1...N\}$ deve-se encontrar o subconjunto Y_M, com $M < N$ que maximiza a função objetivo $J(Y)$:

$$Y_M = \{x_{i1}, x_{i2}, ..., x_{iM}\} = arg\,max_{M,i_M} J\{x_i | i = 1..N\} \tag{3.2}$$

A seleção dos parâmetros eliminando dados irrelevantes é um dos problemas principais no aprendizado de máquina (BLUM; LANGLEY, 1997).

Existem diversos métodos que podem ser utilizados para fazer uma seleção de parâmetros, neste trabalho será utilizado a Análise de Componentes Principais, além de testes em grupos pré selecionados de parâmetros para avaliar o desempenho de cada grupo.

Tabela 3.5 – Tabela de Parâmetros Extraídos dos Sinais de Voz

Índice	**Parâmetros (média, máximo, mínimo, desvio padrão)**
1-4	Frequência Fundamental (F0)
5-8	1a Derivada da Frequência Fundamental (F0)
9-12	2a Derivada da Frequência Fundamental (F0)
13-16	1º Formante
17-20	2º Formante
21-24	3º Formante
25-28	1ª Derivada do 1º Formante
29-32	1ª Derivada do 2º Formante
33-36	1ª Derivada do 3º Formante
37-40	2ª Derivada do 1º Formante
41-44	2ª Derivada do 2º Formante
45-48	2ª Derivada do 3º Formante
49-52	1º MFCC
53-56	2º MFCC
57-60	3º MFCC
61-64	4º MFCC
65-68	5º MFCC
69-72	6º MFCC
73-76	7º MFCC
77-80	8º MFCC
81-84	9º MFCC
85-88	10º MFCC
89-92	11º MFCC
93-96	12º MFCC
97-100	13º MFCC
101-104	1ª Derivada 1º MFCC
105-108	1ª Derivada 2º MFCC
109-112	1ª Derivada 3º MFCC
113-116	1ª Derivada 4º MFCC
117-120	1ª Derivada 5º MFCC
121-124	1ª Derivada 6º MFCC
125-128	1ª Derivada 7º MFCC
129-132	1ª Derivada 8º MFCC
133-136	1ª Derivada 9º MFCC
137-140	1ª Derivada 10º MFCC
141-144	1ª Derivada 11º MFCC
145-148	1ª Derivada 12º MFCC
149-152	1ª Derivada 13º MFCC
153-156	2ª Derivada 1º MFCC

Tabela 3.6 – Tabela de Parâmetros Extraídos dos Sinais de Voz (Cont.)

Índice	Parâmetros (média, máximo, mínimo, desvio padrão)
157-160	2ª Derivada 2º MFCC
161-164	2ª Derivada 3º MFCC
165-168	2ª Derivada 4º MFCC
169-172	2ª Derivada 5º MFCC
173-176	2ª Derivada 6º MFCC
177-180	2ª Derivada 7º MFCC
181-184	2ª Derivada 8º MFCC
185-188	2ª Derivada 9º MFCC
189-192	2ª Derivada 10º MFCC
193-196	2ª Derivada 11º MFCC
197-200	2ª Derivada 12º MFCC
201-204	2ª Derivada 13º MFCC
205-208	Energia de curto termo

3.2.1.1 Análise de Componentes Principais *Principal Component Analysis* (PCA)

Esta técnica é amplamente utilizada em muitas pesquisas (JOLLIFFE, 1986). Tem como objetivo gerar um subespaço vetorial computacionalmente manipulável descartando componentes que possuam menores variâncias. A ideia do PCA é procurar uma base vetorial Λ que maximize a orientação na direção da máxima variância das amostras, podendo ser equacionado da seguinte forma:

$$\Lambda = arg\,max(A^T \Sigma A) \tag{3.3}$$

onde $\vec{A}$ são os autovetores da matriz de covariância Σ do conjunto dos das amostras.

Utilizado em (QUAN et al., 2013) para reconhecimento de emoções em sinais de voz, mostrou um exatidão superior na classificação. Em outro estudo similar foi reduzido 43 parâmetros a 11 por PCA e obteve-se uma melhora na exatidão da classificação (WANG et al., 2010).

3.2.2 Métodos de Classificação

Classificadores são utilizados para reconhecer padrões a partir de um conjunto de dados ou características. Para fazer a escolha de qual classificador utilizar, pode-se testá-los e aplicar algum método de escolha baseado no desempenho e no custo computacional, um destes métodos é o da Navalha de Occam.

Navalha de Occam é uma teoria que defende encontrar um caminho mais simples para resolver um problema. DOMINGOS (1999) explica que esta teoria aplicada em classificadores pode ser interpretada da seguinte maneira: entre dois ou mais métodos de classificação com os mesmos resultados convém escolher o mais simples .

Alguns problemas de generalização que podem ocorrer ao utilizar um classificador:

- Sobre-ajuste - é quando um classificador gera mais superfícies de decisão do que o necessário, causando assim um aumento computacional e tornando o classificador mais suscetível a ruídos, pois devido ao fato de ser mais seletivo acaba se atendo a variações pequenas que só prejudicam a taxa de acerto (JAIN; DUIN; MAO, 2000).
- Sobre-treinamento - mais comum em redes neurais do que em classificadores estatísticos, ocorre quando um conjunto muito grande de padrões com pequena variação intra-classe ou muitas iterações de treinamento (JAIN; DUIN; MAO, 2000).
- Maldição da dimensionalidade - um número maior de parâmetros aumenta a complexidade do classificador de forma exponencial (THEODORIDIS et al., 2006).

3.2.2.1 Máquina de Vetores de Suporte (*Support Vector Machine* - SVM)

Máquina de Vetores de Suporte, do inglês *Support Vector Machine* é um método de classificação de parâmetros para reconhecimento de padrões que implementa a seguinte ideia: mapear os vetores de entrada num espaço dimensional superior através de uma transformação não linear, neste espaço cria-se uma superfície de decisão linear chamado de hiperplano que separa as classes (BOSER; GUYON; VAPNIK, 2004).

Para explicar a teoria deste classificador, pode-se partir de um caso linear com separação de duas classes e depois generalizar para todos os outros casos.

Dado um vetor x_i de parâmetros para o treinamento do grupo P. Pertencentes a duas classes A e B, lineares e separáveis.

$$(x_1, y_1), (x_2, y_2) ... (x_p, y_p) \tag{3.4}$$

$$\begin{cases} y_k = 1, \quad se \quad x_k \in A \\ y_k = -1, \quad se \quad x_k \in B \end{cases} \tag{3.5}$$

O objetivo é achar um hiperplano:

$$D(x) = \sum_{i=1}^{N} \vec{w}_i \vec{\varphi}_i(x) + b \tag{3.6}$$

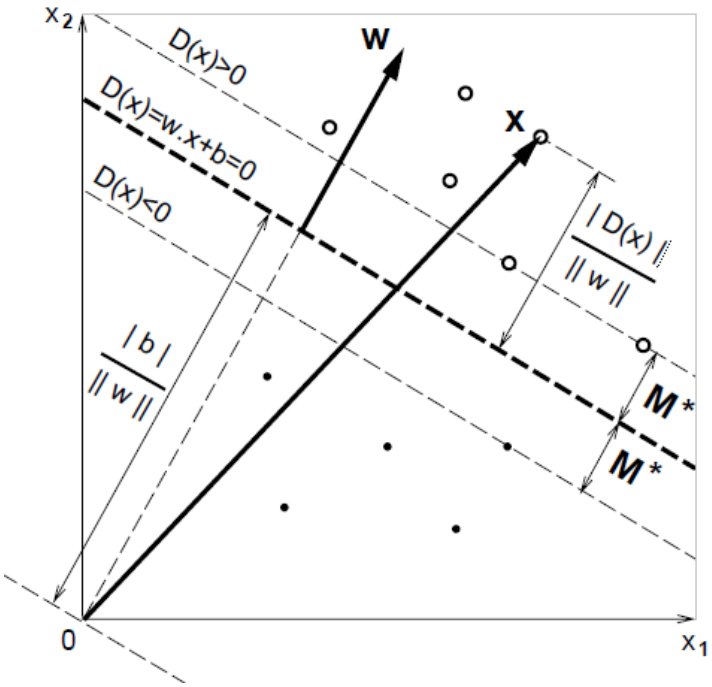
onde w e b são os parâmetros ajustáveis e $\varphi(x)$ é uma função pré-definida.

Obviamente que ao separar dois grupos tal como na Figura 3.5, podemos encontrar vários hiperplanos, porém para otimizar essa ideia deve-se achar aquele que possui a maior distância entre dois grupos (BOSER; GUYON; VAPNIK, 2004).

com uma margem $\frac{1}{||\vec{w}||} + \frac{1}{||\vec{w}||} = \frac{2}{||\vec{w}||}$, requer que:

$$\begin{cases} \vec{w} \cdot \vec{x} + b \geq 1, & \forall x \in A \\ \vec{w} \cdot \vec{x} + b \leq -1, & \forall x \in B \end{cases} \tag{3.7}$$

Figura 3.5 – Margem linear máxima entre dois grupos dada pela função de decisão $D(x)$, onde $\varphi(x) = x$



Fonte: (BOSER; GUYON; VAPNIK, 1992)

No espaço dual a função de decisão é dada por:

$$D(x) = \sum_{k=1}^{p} \alpha_k K(\vec{x}_k, \vec{x}) + b \tag{3.8}$$

Os coeficientes α_k são os parâmetros a serem ajustados e x_k são os parâmetros de treinamento, a função K é uma função de Kernel predefinida, podendo ser uma função polinomial:

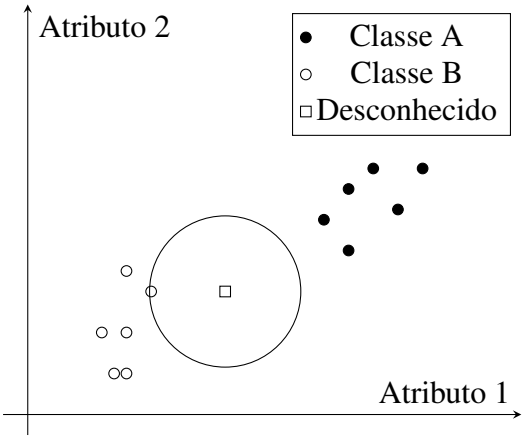
$$K(\vec{x}, \vec{x'}) = (\vec{x} \cdot \vec{x'} - 1)^q \tag{3.9}$$

Que corresponde a uma função de Kernel polinomial de ordem q.

3.2.2.2 k-Vizinhos mais próximos (*k-nearest neighbors* k-NN)

K- Vizinhos mais próximos é um método não paramétrico usado para classificação e regressão. Em ambos os casos, a entrada consiste nos k exemplos de treinamento mais próximos no espaço de recursos. A saída do k-NN usado para classificação é uma associação de classe, um objeto é classificado pela distância entre seus vizinhos, com o objeto sendo atribuído à classe mais comum entre seus k vizinhos mais próximos (k é um inteiro positivo, tipicamente pequeno), se k = 1, então o objeto é simplesmente atribuído à classe daquele único vizinho mais próximo (COVER; HART, 1967).

Figura 3.6 – Exemplo gráfico do método de classificação k-NN



Fonte: Próprio Autor

No caso do exemplo da Figura 3.6, dado duas classes, A e B, e um objeto desconhecido, para k=1 o objeto seria considerado da classe B, pois seu vizinho mais próximo, considerando a distância euclidiana é de um objeto da classe B.

4 TESTES DE RECONHECIMENTO DE EMOÇÕES NA BASE DE DADOS

Neste Capítulo são apresentados resultados da aplicação dos dados coletados nos sistemas de classificação escolhidos, com algumas formas de seleção de parâmetros.

Os classificadores tiveram seu desempenho avaliado pelo método de validação cruzada k-fold divido em 5 grupos.

4.1 SELEÇÃO DE PARÂMETROS

Foi decidido, primeiramente, fazer uma seleção de parâmetros para realizar os ensaios nos classificadores, com o objetivo de avaliar os grupos de parâmetros selecionados e seus respectivos desempenhos. Sendo eles:

- Grupo 1 - Todos os Parâmetros
- Grupo 2 - Media, máxima, minima e desvio padrão do F0, F1, F2, F3, Energia e da primeira e segunda derivadas de cada um destes (na Tabela 3.5 de índices do 1 ao 48 e do 205 ao 208).
- Grupo MFCC - Coefficientes mel-cepstrais (MFCC).
- Análise de Componente Principal.

4.1.1 KNN - Grupo 1

Para o método de classificação k-NN, realizou-se testes com diversos valores k e escolhido dois sendo um deles o menor valor de k e um deles um valor em que o classificador obteve um desempenho melhor em relação aos outros testados, o valor escolhido foi k=10. Nas Tabelas 4.1 e 4.2 esta dispostos os valores da matriz de confusão para k=1, que foi o melhor método com desempenho geral de $41,4\%$ para os áudios masculinos e $21,78\%$ para os áudios femininos. Nas Tabelas 4.3 e 4.4 a matriz de confusão para k=10, que obteve um desempenho geral de $37,3\%$ para os áudios masculinos e de $20,8\%$ para os áudios femininos.

Tabela 4.1 – Matriz Confusão KNN 208 parâmetros

KNN k=1 - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**36,0**	4,0	0,0	8,0	16,0	20,0	16,0
medo	25,0	**25,0**	25,0	0,0	0,0	25,0	0,0
neutro	0,0	0,0	**75,0**	0,0	25,0	0,0	0,0
nojo	7,7	7,7	7,7	**30,8**	15,4	15,4	15,4
raiva	5,6	0,0	11,1	5,6	**66,7**	5,6	5,6
surp.	18,8	6,3	0,0	37,5	6,3	**18,8**	12,5
trist	26,7	0,0	6,7	13,3	13,3	0,0	**40,0**

Tabela 4.2 – Matriz Confusão KNN 208 parâmetros

KNN k=1 - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**10,0**	20,0	30,0	10,0	20,0	0,0	10,0
medo	0,0	**50,0**	11,1	11,1	0,0	16,7	11,1
neutro	8,3	0,0	**25,0**	8,3	16,7	16,7	25,0
nojo	0,0	30,8	7,7	**7,7**	15,4	15,4	23,1
raiva	25,0	0,0	15,0	10,0	**15,0**	10,0	25,0
surp.	0,0	16,7	25,0	16,7	8,3	**16,7**	16,7
trist	0,0	6,3	18,8	31,3	25,0	0,0	**18,8**

Tabela 4.3 – Matriz Confusão KNN 208 parâmetros

KNN k=10 - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**28,0**	8,0	4,0	8,0	12,0	16,0	24,0
medo	0,0	**25,0**	25,0	0,0	0,0	25,0	25,0
neutro	0,0	0,0	**75,0**	0,0	25,0	0,0	0,0
nojo	7,7	0,0	7,7	**30,8**	15,4	15,4	23,1
raiva	16,7	0,0	11,1	11,1	**55,6**	5,6	0,0
surp.	25,0	6,3	6,3	12,5	6,3	**37,5**	6,3
trist	33,3	0,0	20,0	6,7	13,3	0,0	**26,7**

Tabela 4.4 – Matriz Confusão KNN 208 parâmetros

KNN k=10 - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**10,0**	20,0	30,0	0,0	10,0	10,0	20,0
medo	0,0	**33,3**	22,2	22,2	0,0	5,6	16,7
neutro	8,3	8,3	**33,3**	25,0	8,3	0,0	16,7
nojo	0,0	23,1	7,7	**7,7**	15,4	23,1	23,1
raiva	25,0	0,0	15,0	10,0	**20,0**	5,0	25,0
surp.	16,7	16,7	16,7	25,0	16,7	**0,0**	8,3
trist	0,0	6,3	6,3	31,3	25,0	0,0	**31,3**

4.1.2 SVM - Grupo 1

Para o classificador SVM foi utilizado função de kernel gaussiana e polinomial de grau 1,2 e 3. A escala de kernel gaussiana foi variada de 3 a 60, sendo por fim escolhido o valor 15, por ser um valor que mostrou um melhor desempenho. Para as funções polinomiais foram utilizados escala igual a 1.

Na Tabela 4.5 podemos ver o desempenho geral de cada um dos métodos testados. Apesar de terem um desempenho melhor que os do k-NN algumas emoções acabaram ficando com uma taxa de acerto de 0%, algo que ocorre muito mais no SVM. Já no kNN apenas uma emoção na matriz feminina para k=10 obteve essa taxa de acerto. Em geral estas emoções são as que tem o menor número de frases na base de dados utilizada.

Tabela 4.5 – Desempenho Geral do Classificador SVM para o Grupo 1

Classificador SVM		
	Masculino	Feminino
Linear	52,3 %	50,5 %
Quadrático	47,5%	48,5 %
Cúbico	45,5%	43,6 %
Gaussiano	49,5%	46,5 %

Tabela 4.6 – Matriz confusão SVM - Grupo 1

SVM Linear - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**72,0**	0,0	0,0	8,0	4,0	12,0	4,0
medo	50,0	**0,0**	0,0	0,0	0,0	50,0	0,0
neutro	12,5	0,0	**75,0**	0,0	12,5	0,0	0,0
nojo	38,5	0,0	0,0	**46,2**	0,0	7,7	7,7
raiva	22,2	0,0	0,0	0,0	**77,8**	0,0	0,0
surp.	43,8	0,0	6,3	18,8	0,0	**6,3**	25,0
trist	20,0	0,0	0,0	13,3	13,3	6,7	**46,7**

Tabela 4.7 – Matriz confusão SVM - Grupo 1

SVM Linear - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	40,0	0,0	0,0	60,0	0,0	0,0
medo	0,0	**77,8**	16,7	0,0	0,0	5,6	0,0
neutro	0,0	16,7	**33,3**	8,3	33,3	0,0	8,3
nojo	0,0	23,1	7,7	**23,1**	0,0	0,0	46,2
raiva	0,0	0,0	5,0	0,0	**95,0**	0,0	0,0
surp.	0,0	25,0	8,3	16,7	25,0	**0,0**	25,0
trist	0,0	12,5	0,0	0,0	18,8	0,0	**68,8**

Tabela 4.8 – Matriz confusão SVM - Grupo 1

SVM Quadrático - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**48,0**	0,0	4,0	12,0	8,0	8,0	20,0
medo	75,0	**0,0**	0,0	25,0	0,0	0,0	0,0
neutro	0,0	0,0	**75,0**	0,0	12,5	0,0	12,5
nojo	23,1	7,7	7,7	**46,2**	0,0	0,0	15,4
raiva	27,8	0,0	0,0	0,0	**66,7**	0,0	5,6
surp.	25,0	6,3	6,3	18,8	0,0	**25,0**	18,8
trist	26,7	0,0	0,0	0,0	13,3	13,3	**46,7**

Tabela 4.9 – Matriz confusão SVM - Grupo 1

SVM Quadrático - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**10,0**	20,0	0,0	10,0	60,0	0,0	0,0
medo	0,0	**77,8**	16,7	0,0	0,0	5,6	0,0
neutro	0,0	8,3	**41,7**	8,3	33,3	0,0	8,3
nojo	0,0	23,1	7,7	**46,2**	0,0	15,4	7,7
raiva	0,0	0,0	5,0	5,0	**85,0**	0,0	5,0
surp.	0,0	8,3	8,3	58,3	8,3	**0,0**	16,7
trist	0,0	31,3	0,0	12,5	18,8	0,0	**37,5**

Tabela 4.10 – Matriz confusão SVM - Grupo 1

SVM Cubico - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**84,0**	0,0	0,0	0,0	8,0	0,0	8,0
medo	75,0	**0,0**	0,0	0,0	0,0	25,0	0,0
neutro	12,5	0,0	**75,0**	0,0	12,5	0,0	0,0
nojo	84,6	0,0	0,0	**0,0**	0,0	7,7	7,7
raiva	22,2	0,0	0,0	0,0	**77,8**	0,0	0,0
surp.	81,3	0,0	0,0	0,0	0,0	**0,0**	18,8
trist	60,0	0,0	0,0	0,0	13,3	0,0	**26,7**

Tabela 4.11 – Matriz confusão SVM - Grupo 1

SVM Cubico - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	40,0	0,0	0,0	60,0	0,0	0,0
medo	0,0	**83,3**	0,0	0,0	0,0	11,1	5,6
neutro	0,0	16,7	**25,0**	8,3	33,3	0,0	16,7
nojo	0,0	38,5	7,7	**7,7**	0,0	7,7	38,5
raiva	0,0	0,0	5,0	0,0	**90,0**	0,0	5,0
surp.	0,0	33,3	0,0	16,7	16,7	**8,3**	25,0
trist	0,0	18,8	0,0	0,0	25,0	18,8	**37,5**

Tabela 4.12 – Matriz confusão SVM - Grupo 1

SVM Gaussiana - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**80,0**	0,0	0,0	0,0	4,0	4,0	12,0
medo	75,0	**0,0**	0,0	0,0	0,0	25,0	0,0
neutro	12,5	0,0	**75,0**	0,0	12,5	0,0	0,0
nojo	69,2	0,0	0,0	**30,8**	0,0	0,0	0,0
raiva	38,9	0,0	0,0	0,0	**61,1**	0,0	0,0
surp.	68,8	0,0	0,0	0,0	0,0	**18,8**	12,5
trist	53,3	0,0	0,0	0,0	13,3	0,0	**33,3**

Tabela 4.13 – Matriz confusão SVM - Grupo 1

SVM Gaussiana - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	40,0	0,0	0,0	60,0	0,0	0,0
medo	0,0	**83,3**	0,0	0,0	0,0	16,7	0,0
neutro	0,0	8,3	**25,0**	8,3	33,3	16,7	8,3
nojo	0,0	38,5	7,7	**0,0**	7,7	0,0	46,2
raiva	0,0	0,0	5,0	0,0	**95,0**	0,0	0,0
surp.	0,0	25,0	8,3	16,7	33,3	**0,0**	16,7
trist	0,0	18,8	0,0	0,0	18,8	0,0	**62,5**

4.1.3 KNN - Grupo 2

Para o segundo grupo de parâmetros selecionados também foram utilizados os valores de k=1 e k=10. Ambos mostraram um desempenho baixo, e com uma taxa de acerto baixa na escolha das emoções, maior frequência de taxa de 0%.

Tabela 4.14 – Desempenho Geral do Classificador KNN para o Grupo 2

Classificador KNN		
k	Masculino	Feminino
1	35,3%	26,7%
10	34,4%	28,7%

Tabela 4.15 – Matriz confusão KNN - Grupo 2

KNN k=1 - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**44,0**	8,0	4,0	0,0	8,0	24,0	12,0
medo	25,0	**0,0**	0,0	0,0	0,0	50,0	25,0
neutro	12,5	0,0	**75,0**	0,0	12,5	0,0	0,0
nojo	23,1	0,0	7,7	**30,8**	0,0	7,7	30,8
raiva	16,7	5,6	0,0	0,0	**55,6**	22,2	0,0
surp.	37,5	0,0	6,3	25,0	12,5	**0,0**	18,8
trist	20,0	6,7	6,7	13,3	6,7	20,0	**26,7**

Tabela 4.16 – Matriz confusão KNN - Grupo 2

KNN k=10 - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**28,0**	8,0	0,0	12,0	12,0	24,0	16,0
medo	25,0	**0,0**	0,0	0,0	0,0	50,0	25,0
neutro	12,5	0,0	**75,0**	0,0	0,0	0,0	12,5
nojo	23,1	0,0	7,7	**53,8**	0,0	0,0	15,4
raiva	11,1	0,0	0,0	0,0	**55,6**	27,8	5,6
surp.	25,0	0,0	6,3	25,0	12,5	**6,3**	25,0
trist	33,3	13,3	6,7	6,7	6,7	13,3	**20,0**

Tabela 4.17 – Matriz confusão KNN - Grupo 2

KNN k=1 - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	10,0	10,0	10,0	30,0	0,0	40,0
medo	0,0	**44,4**	5,6	22,2	0,0	22,2	5,6
neutro	0,0	0,0	**33,3**	0,0	16,7	16,7	33,3
nojo	23,1	7,7	0,0	**23,1**	15,4	23,1	7,7
raiva	0,0	0,0	20,0	15,0	**40,0**	10,0	15,0
surp.	0,0	16,7	16,7	33,3	16,7	**8,3**	8,3
trist	12,5	12,5	25,0	18,8	6,3	6,3	**18,8**

Tabela 4.18 – Matriz confusão KNN - Grupo 2

KNN k=10 - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**20,0**	10,0	10,0	10,0	30,0	0,0	20,0
medo	0,0	**44,4**	5,6	22,2	0,0	22,2	5,6
neutro	0,0	0,0	**33,3**	0,0	16,7	16,7	33,3
nojo	15,4	7,7	0,0	**23,1**	7,7	23,1	23,1
raiva	0,0	5,0	15,0	15,0	**40,0**	10,0	15,0
surp.	0,0	25,0	8,3	33,3	16,7	**16,7**	0,0
trist	12,5	18,8	18,8	25,0	6,3	6,3	**12,5**

4.1.4 SVM - Grupo 2

Na Tabela 4.19 estão dispostos os valores do desempenho para o grupo 2. Apesar de um queda nos SVM linear e quadrático, o SVM com a função kernel cúbica teve um melhor desempenho em relação a todos os outros do grupo 1 de parâmetros para os áudios masculinos. Contudo os áudios femininos tiveram uma piora no desempenho para este grupo.

Tabela 4.19 – Desempenho Geral do Classificador SVM para o Grupo 2

Classificador SVM		
	Masculino	Feminino
Linear	51,5%	39,6%
Quadrático	53,5%	44,6%
Cúbico	55,6%	41,6%
Gaussiano	41,4%	36,6%

Tabela 4.20 – Matriz confusão SVM - Grupo 2

SVM Linear - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**80,0**	0,0	0,0	4,0	8,0	0,0	8,0
medo	25,0	**0,0**	0,0	0,0	0,0	50,0	25,0
neutro	25,0	0,0	**50,0**	0,0	25,0	0,0	0,0
nojo	30,8	0,0	0,0	**38,5**	0,0	7,7	23,1
raiva	27,8	0,0	0,0	0,0	**61,1**	5,6	5,6
surp.	31,3	0,0	0,0	12,5	6,3	**37,5**	12,5
trist	40,0	0,0	0,0	6,7	13,3	6,7	**33,3**

Tabela 4.21 – Matriz confusão SVM - Grupo 2

SVM Linear - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	0,0	10,0	10,0	50,0	10,0	20,0
medo	0,0	**61,1**	16,7	11,1	0,0	11,1	0,0
neutro	0,0	16,7	**25,0**	8,3	33,3	8,3	8,3
nojo	0,0	23,1	15,4	**23,1**	0,0	7,7	30,8
raiva	0,0	0,0	5,0	0,0	**80,0**	0,0	15,0
surp.	0,0	25,0	16,7	25,0	16,7	**8,3**	8,3
trist	0,0	25,0	6,3	0,0	25,0	6,3	**37,5**

Tabela 4.22 – Matriz confusão SVM - Grupo 2

SVM Quadrático - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**52,0**	0,0	4,0	4,0	16,0	8,0	16,0
medo	0,0	**0,0**	0,0	25,0	0,0	50,0	25,0
neutro	0,0	0,0	**75,0**	0,0	25,0	0,0	0,0
nojo	7,7	7,7	0,0	**53,8**	7,7	0,0	23,1
raiva	16,7	0,0	0,0	0,0	**77,8**	5,6	0,0
surp.	25,0	0,0	0,0	6,3	6,3	**37,5**	25,0
trist	13,3	0,0	6,7	6,7	13,3	13,3	**46,7**

Tabela 4.23 – Matriz confusão SVM - Grupo 2

SVM Quadrático - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	10,0	0,0	10,0	60,0	20,0	0,0
medo	5,6	**61,1**	11,1	11,1	0,0	5,6	5,6
neutro	0,0	0,0	**33,3**	8,3	25,0	16,7	16,7
nojo	7,7	30,8	7,7	**23,1**	0,0	15,4	15,4
raiva	0,0	0,0	0,0	0,0	**90,0**	5,0	5,0
surp.	8,3	8,3	16,7	16,7	41,7	**0,0**	8,3
trist	6,3	6,3	0,0	0,0	18,8	12,5	**56,3**

Tabela 4.24 – Matriz confusão SVM - Grupo 2

SVM Cúbico - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**80,0**	0,0	0,0	4,0	4,0	4,0	8,0
medo	25,0	**0,0**	0,0	0,0	0,0	75,0	0,0
neutro	0,0	0,0	**75,0**	0,0	12,5	12,5	0,0
nojo	23,1	0,0	0,0	**38,5**	0,0	23,1	15,4
raiva	22,2	0,0	0,0	0,0	**77,8**	0,0	0,0
surp.	43,8	0,0	0,0	6,3	0,0	**25,0**	25,0
trist	40,0	0,0	0,0	0,0	13,3	6,7	**40,0**

Tabela 4.25 – Matriz confusão SVM - Grupo 2

SVM Cúbico - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	30,0	0,0	0,0	50,0	10,0	10,0
medo	0,0	**66,7**	22,2	11,1	0,0	0,0	0,0
neutro	0,0	8,3	**25,0**	16,7	33,3	8,3	8,3
nojo	0,0	30,8	7,7	**15,4**	0,0	15,4	30,8
raiva	0,0	0,0	0,0	0,0	**85,0**	0,0	15,0
surp.	0,0	25,0	16,7	16,7	33,3	**0,0**	8,3
trist	0,0	0,0	12,5	6,3	18,8	12,5	**50,0**

Tabela 4.26 – Matriz confusão SVM - Grupo 2

SVM Gaussiano - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**96,0**	0,0	0,0	0,0	4,0	0,0	0,0
medo	50,0	**0,0**	0,0	0,0	0,0	50,0	0,0
neutro	75,0	0,0	**0,0**	0,0	25,0	0,0	0,0
nojo	76,9	0,0	0,0	**7,7**	0,0	15,4	0,0
raiva	33,3	0,0	0,0	0,0	**66,7**	0,0	0,0
surp.	87,5	0,0	0,0	0,0	0,0	**12,5**	0,0
trist	73,3	0,0	0,0	0,0	6,7	6,7	**13,3**

Tabela 4.27 – Matriz confusão SVM - Grupo 2

SVM Gaussiano- Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	20,0	0,0	0,0	60,0	0,0	20,0
medo	0,0	**66,7**	5,6	11,1	5,6	0,0	11,1
neutro	0,0	8,3	**0,0**	8,3	50,0	0,0	33,3
nojo	0,0	38,5	0,0	**0,0**	7,7	0,0	53,8
raiva	0,0	0,0	0,0	0,0	**90,0**	0,0	10,0
surp.	0,0	25,0	8,3	8,3	41,7	**0,0**	16,7
trist	0,0	18,8	0,0	0,0	37,5	0,0	**43,8**

4.1.5 KNN MFCC

Para o grupo de parâmetros MFCC, o classificador k-NN obteve um desempenho melhor em relação aos outros dois grupos, porém, ainda é baixo, pode-se notar nas Tabelas 4.29, 4.30, 4.31 e 4.32, onde estão os dados das matrizes de confusão. Nota-se que existem algumas emoções com 0% de acerto. Para k=1, os melhores resultados foram para raiva, tanto para os áudios masculinos com 72,2%, quanto para os femininos 80%, e neutro apenas para os masculinos com 75%. Já para k=10, apenas os femininos apresentaram raiva, com o melhor acerto 75% e para os masculinos foi o estado neutro, com 75% de acerto.

Tabela 4.28 – Desempenho Geral do Classificador KNN para o MFCC

Classificador KNN		
k	Masculino	Feminino
1	42,4%	35,6%
10	46,5%	33,7%

Tabela 4.29 – Matriz confusão KNN - MFCC

KNN k=1 - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**68,0**	0,0	0,0	4,0	4,0	12,0	12,0
medo	25,0	**0,0**	25,0	0,0	0,0	25,0	25,0
neutro	12,5	0,0	**75,0**	0,0	12,5	0,0	0,0
nojo	46,2	0,0	7,7	**7,7**	0,0	7,7	30,8
raiva	16,7	0,0	5,6	5,6	**72,2**	0,0	0,0
surp.	43,8	0,0	6,3	12,5	0,0	**25,0**	12,5
trist	20,0	0,0	0,0	13,3	13,3	13,3	**40,0**

Tabela 4.30 – Matriz confusão KNN - MFCC

KNN k=1 - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	40,0	0,0	0,0	60,0	0,0	0,0
medo	0,0	**77,8**	0,0	0,0	0,0	11,1	11,1
neutro	0,0	8,3	**25,0**	8,3	50,0	8,3	0,0
nojo	0,0	15,4	7,7	**38,5**	0,0	15,4	23,1
raiva	0,0	0,0	10,0	0,0	**80,0**	0,0	10,0
surp.	8,3	8,3	16,7	41,7	16,7	**0,0**	8,3
trist	0,0	12,5	6,3	0,0	18,8	25,0	**37,5**

Tabela 4.31 – Matriz confusão KNN - MFCC

KNN k=10 - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**60,0**	0,0	0,0	4,0	4,0	12,0	20,0
medo	50,0	**0,0**	25,0	0,0	0,0	25,0	0,0
neutro	0,0	0,0	**75,0**	0,0	12,5	0,0	12,5
nojo	15,4	7,7	7,7	**30,8**	7,7	7,7	23,1
raiva	27,8	0,0	5,6	5,6	**44,4**	5,6	11,1
surp.	50,0	0,0	0,0	0,0	0,0	**18,8**	31,3
trist	20,0	0,0	0,0	13,3	13,3	20,0	**33,3**

Tabela 4.32 – Matriz confusão KNN - MFCC

KNN k=10 - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	40,0	10,0	0,0	40,0	0,0	10,0
medo	0,0	**66,7**	0,0	5,6	0,0	16,7	11,1
neutro	0,0	8,3	**33,3**	0,0	41,7	16,7	0,0
nojo	0,0	15,4	0,0	**7,7**	7,7	38,5	30,8
raiva	5,0	0,0	15,0	0,0	**75,0**	0,0	5,0
surp.	0,0	16,7	16,7	33,3	16,7	**8,3**	8,3
trist	0,0	12,5	0,0	12,5	18,8	6,3	**50,0**

4.1.6 SVM MFCC

Para o grupo MFCC, o classificador SVM mostrou um resultado inferior ao outro grupo, sendo o com melhor desempenho o com polinômio de kernel cúbico para ambos os sexos. Observando as tabelas pode-se perceber que alguns SVM podem possuir um valor geral de desempenho maior, porém eles acabam confundindo emoções que não estão relacionadas, isto é, não possuem variações de parâmetros parecidos. Um exemplo pode ser a raiva, que é facilmente confundida por classificadores com felicidade e surpresa, estas três possuem uma resposta fisiológica mais ativa e portanto podem apresentar variações nos seus parâmetros parecidas, isto pode ser observado na Tabela 2.1 onde as emoções raiva e felicidade possuem características alterações da frequência fundamental similares.

Tabela 4.33 – Desempenho Geral do Classificador SVM para o MFCC

Classificador SVM		
	Masculino	Feminino
Linear	47,5%	43,6%
Quadrático	41,4%	40,6%
Cúbico	49,5%	39,6%
Gaussiano	36,4%	42,6%

Tabela 4.34 – Matriz confusão SVM - MFCC

SVM Linear - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**72,0**	0,0	0,0	4,0	4,0	0,0	20,0
medo	50,0	**0,0**	0,0	0,0	0,0	50,0	0,0
neutro	0,0	0,0	**75,0**	0,0	25,0	0,0	0,0
nojo	46,2	0,0	0,0	**15,4**	7,7	0,0	30,8
raiva	16,7	0,0	0,0	0,0	**83,3**	0,0	0,0
surp.	56,3	0,0	0,0	6,3	0,0	**12,5**	25,0
trist	33,3	0,0	0,0	6,7	13,3	6,7	**40,0**

Tabela 4.35 – Matriz confusão SVM - MFCC

SVM Linear - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	40,0	0,0	0,0	50,0	0,0	10,0
medo	0,0	**77,8**	0,0	0,0	0,0	16,7	5,6
neutro	0,0	8,3	**25,0**	8,3	50,0	8,3	0,0
nojo	0,0	38,5	7,7	**15,4**	15,4	15,4	7,7
raiva	10,0	0,0	10,0	0,0	**75,0**	0,0	5,0
surp.	0,0	16,7	8,3	33,3	8,3	**0,0**	33,3
trist	0,0	12,5	0,0	12,5	31,3	6,3	**37,5**

Tabela 4.36 – Matriz confusão SVM - MFCC

SVM Quadrático - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**88,0**	0,0	0,0	0,0	4,0	0,0	8,0
medo	75,0	**0,0**	0,0	0,0	0,0	25,0	0,0
neutro	75,0	0,0	**0,0**	0,0	12,5	0,0	12,5
nojo	84,6	0,0	0,0	**0,0**	7,7	7,7	0,0
raiva	33,3	0,0	0,0	0,0	**66,7**	0,0	0,0
surp.	87,5	0,0	0,0	0,0	0,0	**0,0**	12,5
trist	73,3	0,0	0,0	0,0	13,3	0,0	**13,3**

Tabela 4.37 – Matriz confusão SVM - MFCC

SVM Quadrático - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	40,0	0,0	0,0	60,0	0,0	0,0
medo	0,0	**83,3**	0,0	0,0	0,0	5,6	11,1
neutro	0,0	8,3	**0,0**	0,0	58,3	8,3	25,0
nojo	0,0	7,7	0,0	**15,4**	30,8	0,0	46,2
raiva	0,0	0,0	0,0	0,0	**95,0**	0,0	5,0
surp.	0,0	16,7	0,0	0,0	33,3	**0,0**	50,0
trist	0,0	6,3	0,0	6,3	43,8	0,0	**43,8**

Tabela 4.38 – Matriz confusão SVM - MFCC

SVM Cúbico - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**52,0**	4,0	0,0	16,0	8,0	16,0	4,0
medo	25,0	**25,0**	0,0	25,0	0,0	25,0	0,0
neutro	0,0	0,0	**75,0**	0,0	12,5	0,0	12,5
nojo	7,7	7,7	0,0	**30,8**	15,4	23,1	15,4
raiva	22,2	0,0	0,0	11,1	**50,0**	5,6	11,1
surp.	37,5	6,3	0,0	12,5	6,3	**25,0**	12,5
trist	20,0	6,7	0,0	6,7	13,3	20,0	**33,3**

Tabela 4.39 – Matriz confusão SVM - MFCC

SVM Cúbico - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	20,0	10,0	10,0	50,0	0,0	10,0
medo	5,6	**50,0**	5,6	11,1	0,0	22,2	5,6
neutro	8,3	0,0	**33,3**	25,0	25,0	8,3	0,0
nojo	0,0	30,8	15,4	**30,8**	0,0	15,4	7,7
raiva	10,0	0,0	20,0	10,0	**55,0**	0,0	5,0
surp.	0,0	8,3	25,0	33,3	8,3	**8,3**	16,7
trist	0,0	12,5	6,3	12,5	18,8	6,3	**43,8**

Tabela 4.40 – Matriz confusão SVM - MFCC

SVM Gaussiano - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**52,0**	0,0	0,0	20,0	12,0	12,0	4,0
medo	25,0	**25,0**	0,0	25,0	0,0	25,0	0,0
neutro	0,0	0,0	**75,0**	0,0	12,5	0,0	12,5
nojo	7,7	7,7	7,7	**38,5**	15,4	15,4	7,7
raiva	22,2	5,6	0,0	11,1	**50,0**	11,1	0,0
surp.	37,5	0,0	0,0	25,0	6,3	**31,3**	0,0
trist	26,7	6,7	6,7	0,0	6,7	6,7	**46,7**

Tabela 4.41 – Matriz confusão SVM - MFCC

SVM Gaussiano- Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	20,0	10,0	0,0	50,0	0,0	20,0
medo	11,1	**44,4**	11,1	11,1	0,0	16,7	5,6
neutro	8,3	0,0	**33,3**	25,0	25,0	8,3	0,0
nojo	0,0	30,8	7,7	**30,8**	7,7	15,4	7,7
raiva	15,0	0,0	25,0	5,0	**50,0**	0,0	5,0
surp.	0,0	8,3	25,0	41,7	8,3	**0,0**	16,7
trist	6,3	12,5	6,3	6,3	12,5	6,3	**50,0**

4.1.7 Análise de Componente Principal

Para análise de componente principal, foram feitos diversos ensaios com grupos de parâmetros diferentes. Inicialmente utilizou-se todos e conforme foi diminuindo o grupo de parâmetros obteve-se desempenhos melhores, por fim foi selecionado um grupo de 20 parâmetros do índice 1 ao índice 20 da Tabela 3.5. A escolha foi excluir primeiro os MFCC depois os formantes, a energia e por último a frequência fundamental, após um exaustivo teste de parâmetros foi escolhidos este grupo.

Tabela 4.42 – Desempenho Classificadores - Masculino

Desempenho Masculino - ACP	
KNN k=1	27,7%
KNN k=10	26,2%
SVM Linear	34,3%
SVM Quadrático	34,3%
SVM Cúbico	45,4%
SVM Gaussiano	25,5%

Tabela 4.43 – Desempenho Classificadores - Feminino

Desempenho Feminino - ACP	
KNN k=1	28,7%
KNN k=10	26,7%
SVM Linear	20,7%
SVM Quadrático	28,7%
SVM Cúbico	41,6%
SVM Gaussiano	19,8%

Tabela 4.44 – Matriz de confusão com o melhor desempenho ACP

SVM Cúbico - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**80,0**	0,0	0,0	0,0	4,0	8,0	8,0
medo	75,0	**0,0**	0,0	25,0	0,0	0,0	0,0
neutro	12,5	0,0	**75,0**	0,0	12,5	0,0	0,0
nojo	76,9	0,0	0,0	**7,7**	0,0	15,4	0,0
raiva	27,8	0,0	0,0	0,0	**50,0**	5,6	16,7
surp.	50,0	0,0	0,0	6,3	0,0	**37,5**	6,3
trist	73,3	0,0	0,0	0,0	6,7	0,0	**20,0**

Tabela 4.45 – Matriz de confusão com o melhor desempenho ACP

SVM Cúbico - Feminino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**0,0**	10,0	0,0	0,0	60,0	10,0	20,0
medo	0,0	**61,1**	5,6	0,0	11,1	11,1	11,1
neutro	0,0	8,3	**8,3**	8,3	41,7	16,7	16,7
nojo	0,0	15,4	7,7	**30,8**	7,7	0,0	38,5
raiva	5,0	0,0	0,0	0,0	**80,0**	0,0	15,0
surp.	0,0	16,7	0,0	0,0	33,3	**16,7**	33,3
trist	0,0	6,3	0,0	0,0	43,8	0,0	**50,0**

4.1.8 Removendo emoções

Algumas emoções tiveram seu desempenho muito baixo para a maioria dos classificadores. De modo a melhorar os resultados das outras emoções excluiu-se uma e foram refeitos os ensaios.

Inicialmente o medo foi removido, pois é a que possui menor número de áudios masculinos e menor desempenho. Para os femininos a felicidade, que foi a qual menos obteve-se dados. Estas duas mostram uma taxa de acerto muito baixa para quase todos os classificadores utilizados.

Assim obteve-se um aumento no desempenho dos classificadores de um modo geral. Contudo o feminino ainda ficou com uma média baixa e com um erro grande no reconhecimento da surpresa.

Tabela 4.46 – Desempenho Geral dos Classificadores excluindo a emoção Medo

	Classificador SVM - Sem Medo		Classificador KNN - Sem Medo	
		Masculino	k	Masculino
Grupo 1	**Linear**	**57,9%**	**1**	**44,2%**
	Quadrático	54,7%	10	43,2%
	Cúbico	26,3%		
	Gaussiano	49,5%		
Grupo 2	**Linear**	**53,7%**	**1**	**40,0%**
	Quadrático	51,6%	10	38,9%
	Cúbico	34,7%		
	Gaussiano	40,0%		
MFCC	Linear	43,2%		
	Quadrático	**48,4%**	**1**	**45,3%**
	Cúbico	27,4%	10	42,1%
	Gaussiano	42,1%		
ACP	Linear	33,7%	**1**	**29,5%**
	Quadrático	37,9%	10	24,2%
	Cúbico	**47,4%**		
	Gaussiano	26,3%		

Tabela 4.47 – Matriz confusão SVM - Melhor Desempenho

SVM Linear - Masculino						
%	fel.	neutro	nojo	raiva	surp.	trist
fel.	**72,0**	0,0	4,0	4,0	8,0	12,0
neutro	12,5	**75,0**	0,0	12,5	0,0	0,0
nojo	38,5	0,0	**38,5**	0,0	15,4	7,7
raiva	22,2	0,0	0,0	**77,8**	0,0	0,0
surp.	25,0	6,3	12,5	0,0	**25,0**	31,3
trist	13,3	0,0	6,7	13,3	13,3	**53,3**

Tabela 4.48 – Desempenho Geral dos Classificadores excluindo a emoção Felicidade

	Classificador SVM - Sem Felicidade		Classificador KNN - Sem Felicidade	
		Feminino	k	Feminino
Grupo 1	**Linear**	**52,7%**	**1**	**26,4%**
	Quadrático	58,2%	10	24,2%
	Cúbico	22,0%		
	Gaussiano	49,5%		
Grupo 2	Linear	44,0%	**1**	**36,3%**
	Quadrático	**52,7%**	10	32,5%
	Cúbico	33,0%		
	Gaussiano	42,9%		
MFCC	Linear	53,8%	1	20,9%
	Quadrático	**53,8%**	**10**	**27,5%**
	Cúbico	22,0%		
	Gaussiano	44,0%		
ACP	Linear	29,7%	1	31,9%
	Quadrático	31,9%	**10**	**34,1%**
	Cúbico	**48,4%**		
	Gaussiano	22,0%		

Tabela 4.49 – Matriz confusão SVM - Melhor Desempenho

SVM Quadrático - Feminino sem Felicidade						
%	medo	neutro	nojo	raiva	surp.	trist
medo	**72,2**	16,7	5,6	0,0	5,6	0,0
neutro	8,3	**41,7**	16,7	25,0	0,0	8,3
nojo	15,4	7,7	**46,2**	0,0	15,4	15,4
raiva	0,0	10,0	0,0	**90,0**	0,0	0,0
surp.	25,0	8,3	41,7	16,7	**0,0**	8,3
trist	12,5	0,0	6,3	6,3	6,3	**68,8**

4.1.9 Videos da *Internet*

Para os vídeos da *internet* foram escolhidos apenas os áudios masculinos. O melhor desempenho foi do grupo 1 de parâmetros com o classificador SVM Quadrático, em destaque na Tabela 4.50. Estes resultados, mostraram-se melhores das emoções induzas. A justificativa para isso poderá ser a grande disponibilidade de material por meio de vídeos na *internet*, comparada ao produzido em estúdio, e sua efetividade na geração de emoções devido até mesmo à atuação.

Na Tabela 4.51 é dada a matriz de confusão do melhor resultado de classificação. Bem como nos áudios induzidos, o medo teve uma baixa taxa de acerto comparada às outras. Ademais, a felicidade manteve sua alta taxa de acerto tanto para emoções induzidas quanto para os vídeos.

Para análise de componente principal, foi utilizado os 208 parâmetros e para classificação foram feitos ensaios de 3 a 12 primeiras componentes tendo sido escolhido o valor 12 para dispor os resultados. A partir da 12^a os resultados se mantiveram com desempenho similar ou desempenho pior do que com 12 componentes.

Tabela 4.50 – Desempenho Geral dos Classificadores

	Classificador SVM		Classificador KNN	
		Masculino	k	Masculino
Grupo 1	Linear	66,1%	1	35,5%
	Quadrático	**71,8%**	10	34,7%
	Cúbico	15,9%		
	Gaussiano	69,8%		
Grupo 2	Linear	64,9%	1	35,9%
	Quadrático	65,3%	10	40,0%
	Cúbico	61,2%		
	Gaussiano	63,3%		
MFCC	Linear	61,2%	1	33,9%
	Quadrático	65,3%	10	32,7%
	Cúbico	23,3%		
	Gaussiano	65,3%		
ACP	Linear	33,9%	1	30,2%
	Quadrático	32,7%	10	28,6%
	Cúbico	33,9%		
	Gaussiano	32,7%		

Tabela 4.51 – Matriz de confusão para áudios de vídeos da *internet*

SVM Quadrático - Masculino							
%	fel.	medo	neutro	nojo	raiva	surp.	trist
fel.	**79,4**	0,0	8,8	11,8	0,0	0,0	0,0
medo	8,1	**51,4**	5,4	8,1	21,6	0,0	5,4
neutro	2,9	2,9	**79,4**	5,9	5,9	0,0	2,9
nojo	5,9	2,9	2,9	**73,5**	8,8	5,9	0,0
raiva	2,9	8,8	0,0	5,9	**64,7**	14,7	2,9
surp.	8,1	0,0	0,0	8,1	2,7	**78,4**	2,7
trist	2,9	0,0	14,3	5,7	0,0	0,0	**77,1**

4.1.10 Filmes

Diferente de todos os outros resultados, para a classificação de filmes o SVM teve um menor desempenho em relação ao KNN para a maioria dos casos. Nestes também obteve-se um resultado incomum na aplicação da análise de componente principal, que teve um desempenho melhor para os áudios femininos e próximo para os resultados masculinos em relação aos outros grupos. Neste caso, o desempenho geral foi maior porém estes possuem apenas 4 emoções e um estado neutro.

Para a ACP dos áudios femininos, foram utilizados as 6 primeiras componentes que representam 91,3% da variância dos dados. Nos masculinos foram utilizados as 12 primeiras componentes principais que representam 98,1% de variância dos dados.

Por ter sido o melhor desempenho o classificador KNN com k=10 para os áudios femininos, foi feito uma análise gráfica das 3 primeiras componentes principais como mostra a Figura 4.1.

Tabela 4.52 – Desempenho Geral dos Classificadores

		Classificador SVM		Classificador KNN		
		Masculino	Feminino	k	Masculino	Feminino
Grupo 1	Linear	57,9%	54,1%	1	**63,2%**	70,6%
	Quadrático	50,9%	54,1%	10	64,9%	**71,8%**
	Cúbico	56,1%	29,4%			
	Gaussiano	45,6%	52,9%			
Grupo 2	Linear	56,1%	69,4%	1	63,2%	71,8%
	Quadrático	45,6%	61,2%	10	61,4%	**75,3%**
	Cúbico	21,5%	41,2%			
	Gaussiano	31,6%	56,5%			
MFCC	Linear	42,1%	41,2%	1	45,6%	35,3%
	Quadrático	40,4%	38,8%	10	38,6%	36,5%
	Cúbico	42,1%	24,7%			
	Gaussiano	38,6%	40,0%			
ACP	Linear	54,4%	60,0%	1	**63,2%**	74,1%
	Quadrático	50,9%	61,2%	10	59,6%	**75,3%**
	Cúbico	54,4%	34,1%			
	Gaussiano	31,6%	28,2%			

Tabela 4.53 – Matriz de Confusão áudios de filmes - Feminino

KNN k=10 - Feminino					
%	fel.	medo	neutro	raiva	trist.
fel.	**93,3**	0,0	0,0	0,0	6,7
medo	7,7	**76,9**	0,0	7,7	7,7
neutro	0,0	0,0	**60,0**	20,0	20,0
raiva	0,0	9,5	4,8	**57,1**	28,6
trist.	0,0	4,8	14,3	4,8	**76,2**

Tabela 4.54 – Matriz de Confusão áudios de filmes - Feminino

SVM Quadrático - Feminino					
%	fel.	medo	neutro	raiva	trist.
fel.	**93,3**	0,0	0,0	0,0	6,7
medo	7,7	**76,9**	0,0	7,7	7,7
neutro	0,0	0,0	**66,7**	13,3	20,0
raiva	0,0	9,5	14,3	**61,9**	14,3
trist.	0,0	0,0	19,0	0,0	**81,0**

Tabela 4.55 – Matriz de Confusão áudios de filmes - Feminino

ACP KNN k=10 - Feminino					
%	fel.	medo	neutro	raiva	trist.
fel.	**93,3**	0,0	0,0	0,0	6,7
medo	7,7	**76,9**	0,0	7,7	7,7
neutro	0,0	0,0	**60,0**	13,3	26,7
raiva	0,0	9,5	9,5	**66,7**	14,3
trist.	0,0	9,5	9,5	0,0	**81,0**

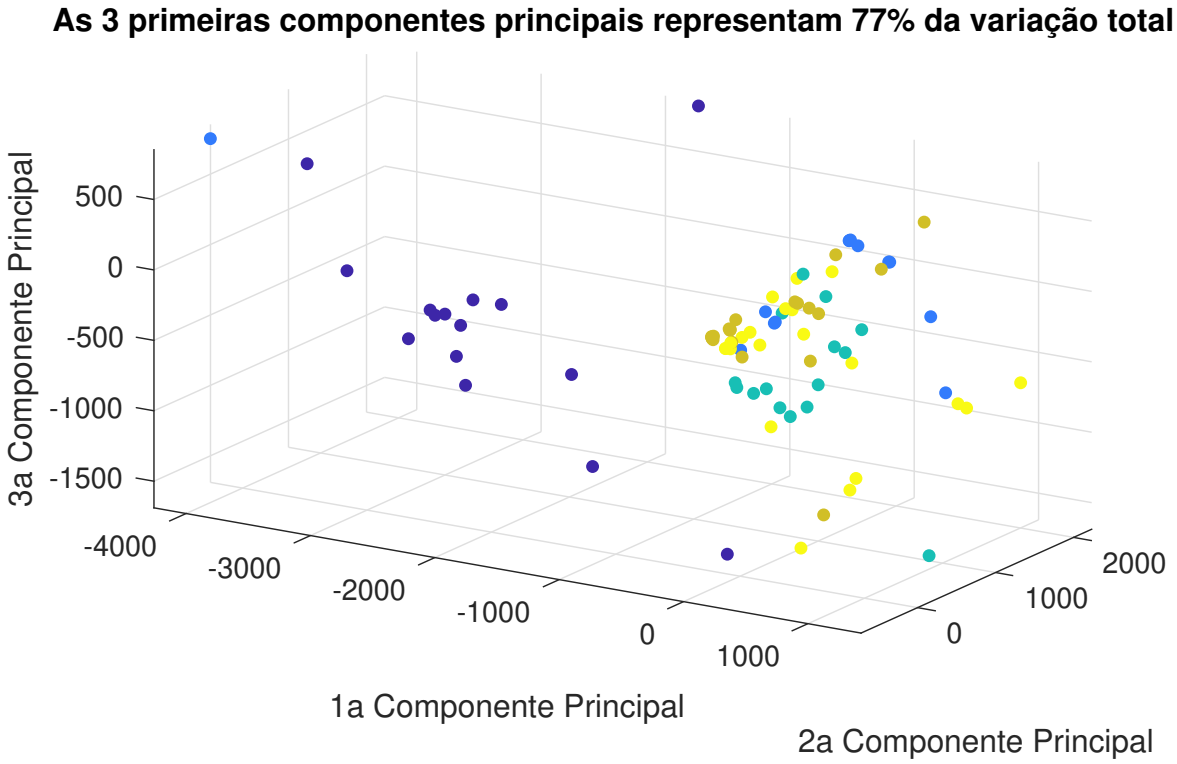
Tabela 4.56 – Matriz de Confusão áudios de filmes - Masculino

ACP KNN k=1 - Masculino					
%	fel.	medo	neutro	raiva	trist.
fel.	**45,5**	0,0	18,2	27,3	9,1
medo	0,0	**50,0**	33,3	16,7	0,0
neutro	7,7	0,0	**84,6**	7,7	0,0
raiva	16,7	5,6	0,0	**66,7**	11,1
trist.	11,1	11,1	0,0	11,1	**66,7**

Tabela 4.57 – Matriz de Confusão áudios de filmes - Masculino

KNN k=10 - Masculino					
%	fel.	medo	neutro	raiva	trist.
fel.	**45,5**	9,1	9,1	27,3	9,1
medo	16,7	**50,0**	16,7	16,7	0,0
neutro	7,7	7,7	**76,9**	7,7	0,0
raiva	16,7	5,6	0,0	**72,2**	5,6
trist.	11,1	11,1	0,0	11,1	**66,7**

Figura 4.1 – Representação gráfica da ACP para os áudios femininos



Fonte: Próprio Autor

4.2 DISCUSSÃO DOS RESULTADOS

Os resultados das classificações para o banco de voz com emoções induzidas ficou com o pior desempenho e a classificação utilizando os áudios de filmes teve o melhor desempenho, chegando a 75,3% para o sexo feminino e 63,2% para o sexo masculino. Comparando a estudos similares pode-se perceber alguns resultados parecidos para estas classificações como em (VERVERIDIS; KOTROPOULOS; PITAS, 2004; HUANG; SONG; ZHAO, 2016; SCHULLER et al., 2010; KAMÍNSKA; PELIKANT, 2012).

Para emoções induzidas em estúdio, o desempenho ficou entre 20% e 60% em todos os casos, mas para a maioria dos testes o classificador SVM obteve melhores resultados. Uma provável causa deste baixo desempenho pode ser pela quantidade dos dados utilizados, pela intensidade das emoções induzidas, por erros nas técnicas de indução de emoções, que pode ter baixa intensidade ou pela escolha dos parâmetros que foram utilizados nos classificadores.

No estudo feito por (HUANG; SONG; ZHAO, 2016) utilizando o banco de dados áudio-visual eNTERFACE (MARTIN et al., 2006) com o método de classificação modelo de misturas

de gaussianas, sigla GMM (do inglês *Gaussian mixture model*). Alguns dos parâmetros utilizados neste estudo: frequência fundamental, formantes, energia e MFCCs. Na matriz de confusão da Tabela 4.58 é possível observar que alguns resultados como do reconhecimento de emoção medo ficaram com um desempenho baixíssimo comparado aos outras emoções. O que também ocorre neste trabalho com os áudios coletados em estúdio por indução de emoções, sendo em muitos casos um acerto de 0%.

Tabela 4.58 – Matriz de Confusão - Banco de dados eNTERFACE05

Classificador GMM						
%	raiva	nojo	medo	fel.	trist.	surp.
raiva	**48,8**	12,2	14,6	4,9	7,3	12,2
nojo	9,8	**39,0**	4,9	9,8	19,5	17,1
medo	14,6	7,3	**9,8**	14,6	4,9	22,0
fel.	9,8	9,8	14,6	**48,8**	2,4	17,1
trist.	4,9	17,1	48,8	2,4	**53,7**	17,1
surp.	4,9	17,1	2,4	14,6	19,5	**29,3**

Fonte: (HUANG; SONG; ZHAO, 2016)

Em (SCHULLER et al., 2010) é feito um estudo comparativo entre os bancos de dados de voz já existentes, sendo alguns deles já citados neste trabalho. Na tabela 4.59 estão os resultados obtidos por dois tipos de classificadores, Cadeias de Markov Escondidas (HMM) e SVM que também foi utilizado neste trabalho. Em questão de desempenho geral foram obtidos resultados melhores para o EMO-DB, que é um banco de dados feito por atores.

Tabela 4.59 – Resultados obtidos para a classificação em outros bancos de dados

Base	HMM (%)	SVM (%)
DES	45,3	59,9
EMO-DB	73,2	84,6
eNTERFACE	67,1	72,5

Fonte: (SCHULLER et al., 2010)

Neste trabalho foram obtidos resultados melhores para os áudios coletados de filmes, que tem em comum com o EMO-DB serem emoções atuadas e não genuínas. Porém áudios dos filmes coletados não possuem uma validação humana com vários participantes, como no caso dos áudios com emoções induzidas em estúdio, apenas a validação do autor do trabalho, que coletou os dados e os avaliou conforme as cenas assistidas.

Os resultados dos classificadores de aprendizado de máquina, para os áudios com emoções induzidas deste trabalho, não tiveram resultados satisfatórios a ponto de utiliza-los para

possíveis aplicações. A classificação por seres humanos foi muito melhor e necessária para que pudesse utilizar a identificação humana nos classificadores, tanto pra treinamento quanto para avaliação de desempenho. Um dos fatores que pode ter influenciado a melhora na classificação humana, pode ter sido pela informação semântica das frases, ou seja, o avaliador acabou avaliando mais pelo contexto do que pela entonação da fala.

O Grupo 1 e o Grupo 2 de parâmetros mostraram um resultado melhor, os coeficientes Mel-Cepstrais não apresentaram uma melhora expressiva do desempenho dos classificadores nem foram melhores utilizados isoladamente. A separação dos grupos não foi feita de forma a otimizar os classificadores, mas sim de forma a avaliar alguns grupos. Em especial foi separado o grupo de MFCC por que além de ser um grupo de tamanho expressivo, eles já são utilizados para reconhecimento de fala. Os mais comuns para classificação de emoções são os formantes e a frequência fundamental, de fato eles tiveram bons resultados e se mostraram superiores aos MFCCs.

A Análise de Componentes Principais teve um desempenho muito ruim para os áudios de estúdio e de vídeos da internet, porém um bom resultado para os áudios coletados nos filmes, melhorando o desempenho dos classificadores. Não houve uma investigação sobre o motivo dessa diferença, porém pode estar relacionada a qualidade dos sinais, tanto em questões do equipamento de coleta e do tratamento do sinal feita pelos estúdios dos filmes quando em qualidade de expressar as emoções, visto que sem ACP os resultados também foram melhores.

Os classificadores aqui treinados com os dados de emoções induzidas não podem ser aplicados em reconhecimento de emoções visto que tem um baixo desempenho, o ideal seria ter uma taxa de acerto maior, o que pode ser melhorada tanto na hora da coleta dos dados quanto na hora de utiliza-los nos classificadores. Outra opção para melhora destes classificadores seria testar outros parâmetros e analisar grupos menores que otimizem o classificador.

Outro ponto a se destacar nos resultados para os áudios extraídos de filmes é que o método de k-vizinhos mais próximos apresentou um melhor desempenho, tanto para os grupos de parâmetros quanto para análise de componentes principais, na Figura 4.1 é possível ver alguns grupos separados pela análise de componentes principais para os áudios femininos, com um grupo em destaque por estar bem distante de todos os outros e com pontos bem distribuídos.

5 CONCLUSÃO

A criação de um banco de dados de emoções pode ser feito com diferentes metodologias, inclusive, as propostas neste trabalho, que mostrou ter um bom resultado para classificação humana, porém não foram encontrados resultados satisfatórios em classificadores de aprendizado de máquinas.

Para se obter um banco de dados por emoções induzidas, deve-se inicialmente trabalhar com métodos que provoquem emoções intensas, disponibilizando também, espaço para o locutor interagir. Este pode ser o método mais difícil de se obter este tipo de dado, porém o mais próximo da realidade, os métodos propostos neste trabalho foram de baixa complexidade, por meio de vídeos e atividades os locutores puderam sentir e expressar o que estavam sentindo.

Trata-se de um processo de obtenção de dados incerto, pois nem sempre as atividades vão provocar as emoções desejadas. Além disso, as emoções presenciadas pelos locutores podem divergir de um para outro, comprometendo assim a classificação da mesma.

Para tal, fez-se necessário o uso de três estágios:

1. O locutor descreve a emoção sentida.
2. Seleção de frases.
3. Avaliação humana, onde um grupo de pessoas escutam e avaliam as emoções dos áudios selecionados.

Uma novidade apresentada neste trabalho foi a forma de indução de algumas emoções. Para o medo que foi utilizado um óculos de realidade virtual e um *joystick* onde os locutores participaram um jogo de terror. As outras atividades foram similares as já propostas na literatura, porém com algumas adaptações em diferentes vídeos e também o uso de outros jogos virtuais.

Além dos áudios com emoções induzidas foram coletados áudios de outras duas fontes: videos disponíveis na *internet*, filmes brasileiros e dublados em português brasileiro. Estes mostraram um melhor desempenho nos classificadores.

Conclui-se então, que fazer uma boa obtenção de dados é preciso um bom método e um número expressivo de locutores e de áudios. Além disso deve-se rever a alguns pontos desse método de indução de emoções, caso se trabalhe com essa forma de obtenção de dados, de forma a melhorar a intensidade das emoções e a facilitar o locutor a se expressar de forma mais natural possível.

5.1 TRABALHOS FUTUROS

Um trabalho a ser realizado futuramente com estes dados, seria a aplicação de outros métodos para seleção de parâmetros, de modo a melhorar o desempenho dos classificadores. Também testes dos dados em outros classificadores além dos que foram utilizados.

Além disso, pode-se desenvolver uma aplicação dos modelos de classificação, tais como: interfaces afetivas, auxílio de diagnóstico de transtornos mentais, avaliador de satisfação de clientes, jogos entre outros.

Uma comparação entre idiomas, analisando se é possível aplicar modelos de classificação treinados em um idioma, para reconhecer emoções em outros idiomas.

Analisar se uso de pseudopalavras tem os mesmos resultados em estudo de reconhecimento de emoções em sinais de voz.

Outro possível estudo, envolvendo testes de voz e sinais fisiológicos direcionados a melhora da taxa de reconhecimento das emoções poderá ser desenvolvido. Podendo combinar estes métodos com um sistema vestível para obtenção e análise de dados.

6 PUBLICAÇÕES

Neste capítulos estão descritas as publicações feitas durante o período de realização deste mestrado.

O primeiro artigo com o título: *Classificação de Parâmetros de Sinais de Voz de Filmes Para Reconhecimento de Emoções.* (KINGESKI; SCHUEDA; PATERNO, 2018a). Foi publicado no Simpósio Brasileiro de Telecomunicações e Processamento de Sinais. Neste trabalho foi criado um banco com áudios de filmes e testados em alguns algoritmos de extração de parâmetros e classificadores.

O segundo trabalho publicado foi uma comunicação científica com o título: *Reconhecimento de Emoções por Voz para Auxílio ao diagnóstico de Transtornos Mentais* (KINGESKI; SCHUEDA; PATERNO, 2018b). Publicado e apresentado em forma de poster no Congresso Brasileiro de Engenharia Biomédica. Neste trabalho foi feito uma coleta de dados por vídeos disponíveis na *internet* que foram testados em classificadores para análise de desempenho, além disso foi feito uma breve revisão bibliográfica de trabalhos similares que utilizam reconhecimento de emoções para auxílio de diagnóstico e tratamento de pessoas com transtornos mentais.

REFERÊNCIAS BIBLIOGRÁFICAS

AELURI, P.; VIJAYARAJAN, V. Extraction of emotions from speech–a survey. **International Journal of Applied Engineering Research**, 2017. v. 12, p. 5760–5767, 01 2017.

ALCAIM, A.; OLIVEIRA, C. A. S. **Fundamentos do Processamento de Sinais de Voz e Imagem**. Rio de Janeiro – RJ,: Interciência, PUC Rio, 2011.

ARISTÓTELES. **Ética a Nicômaco**. São Paulo,: Editora Nova Cultural, 1991.

BA, H.; YANG, N.; DEMIRKOL, I.; HEINZELMAN, W. Bana: A hybrid approach for noise resilient pitch detection. In: . [S.l.: s.n.], 2012.

BARRA-CHICOTE, R.; MONTERO, J. M.; MACIAS-GUARASA, J.; LUFTI, S. L.; LUCAS, J. M.; DHARO, L. F.; SAN-SEGUNDO, R.; FERREIROS, J.; CORDOBA, R.; PARDO, M. Spanish expressive voices: Corpus for emotion research in spanish. In: **In Proc. of 6th International Conference on Language Resources and Evaluation LREC**. [S.l.: s.n.], 2008.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artifical Intelligence**, 1997. v. 97, p. 245–271, 1997.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the Fifth Annual Workshop on Computational Learning Theory**. New York, NY, USA: ACM, 1992. p. 144–152.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: . [S.l.: s.n.], 2004.

BURKHARDT, F.; PAESCHKE, A.; ROLFES, M.; SENDLMEIER, W.; WEISS, B. A database of german emotional speech. In: **Proceedings of Interspeech, Lissabon**. [S.l.: s.n.], 2005. p. 1517–1520.

CANNON, W. B. The james-lange theory of emotions: A critical examination and an alternative theory. **The American Journal of Psicology Vol. 100**, 1987. p. 567–586, 1987. Disponível em: <https://www.jstor.org/stable/1422695>.

CASTRO, S. L.; LIMA, C. F. Recognizing emotions in spoken language: A validated set of portuguese sentences and pseudosentences for research on emotional prosody. **Behavior Research Methods**, 2010. v. 42, n. 1, p. 74–81, Feb 2010. Disponível em: <https://doi.org/10.3758/BRM.42.1.74>.

CHENG, M. J. Comparative performance study of several pitch detection algorithms. 1975. v. 58, p. 61, 07 1975.

COVER, T. M.; HART, P. E. Nearest Neighbor Pattern Classification. **IEEE Transactions on Information Theory**, 1967. 1967.

COWIE, R.; COWIE, E.; TSAPATSOULIS, N.; AL et. Emotion recognition in human-computer interaction. **IEEE Signal Processing Magazine**, 2001. v. 18, n. 1, p. 32–80, 2001.

COWIE, R.; DOUGLAS-COWIE, E. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. **Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP 96**.

CUERDAS. Canal Youtube: Cuerdas Cortometraje Oficial, 2013. Disponível em: <https://www.youtube.com/watch?v=4INwx_tmTKw>. Acesso em: 28 de fevereiro de 2019.

DALGLEISH, T. The emotional brain. **Nature Reviews Neuroscience**, 2004. p. 583–589, April 2004. Disponível em: <10.1038/nrn1432>.

DAVIS, S. B.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **ACOUSTICS, SPEECH AND SIGNAL PROCESSING, IEEE TRANSACTIONS ON**, 1980. p. 357–366, 1980.

DESCARTES, R. **As Paixões da Alma**. [S.l.]: KTTK, 2018.

DOMINGOS, P. The role of occam's razor in knowledge discovery. **Data Mining and Knowledge Discovery**, 1999. p. 409–425, 1999. Disponível em: <https://doi.org/10.1023/A:1009868929893>.

DOUGLAS-COWIE, E.; COWIE, R.; SCHRÖDER, M. A new emotion database: Considerations, sources and scope. In: **Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research**. Belfast: Textflow, 2000. p. 39–44.

DUTOIT, T.; MARQUES, F. **Applied Signal Processing: A MATLAB Based Proof of Concept**. [S.l.]: Springer US, 2010.

EKMAN, P. **A Linguagem das Emoções**. Rio de Janeiro,: Editora Leya, 2011.

EKMAN, P.; FRIESEN, W. V. **Unmasking the face: A guide to recognizing emotions from facial clues**. Oxford: Prentice-Hall, 1975.

ELIAS, P. Predictive coding – I. **IEEE Transactions on Information Theory**, 1955. v. 1, p. 16–24, 1955.

ELLIS, D. P. W. **PLP and RASTA (and MFCC, and inversion) in Matlab**. 2005. Online web resource. Disponível em: <http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/>.

FABIANO CAMBOTA - Comedy Central #2 (Paraguai). Canal Youtube: Fabiano Cambota, 2016. Disponível em: <https://www.youtube.com/watch?v=ezGGhn4Hykw>. Acesso em: 28 de fevereiro de 2019.

FLANAGAN, J. **Speech Analysis Synthesis and Perception**. [S.l.]: Springer – Verlag, 1972.

GROSS, J. J.; LEVENSON, R. W. Emotion elicitation using films. **Cognition and Emotion**, 1995. v. 9, p. 87–108, 1995.

HANSEN, I. S. E. A. V. Documentation of the danish emotional speech database. 1996. Aalborg University, 1996. Disponível em: <http://kom.aau.dk/~tb/speech/Emotions/des.pdf>. Acesso em: 21 de novembro de 2019.

HARRINGTON JONATHAN; CASSIDY, S. Techniques in speech acoustics. **Computational Linguistics**, 2000. MIT Press, v. 26, p. 294–295, 2000.

HOBBES, T. **Leviatã. Matéria, forma e poder de um Estado eclesiástico e civil**. [S.l.]: Abril Cultural, 1983.

HUANG, C.; SONG, B.; ZHAO, L. Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering. **International Journal of Speech Technology**, 2016. v. 19, n. 4, p. 805–816, 2016.

HUANG, X.; ACERO, A.; HON, H.-W. **Spoken Language Processing: A Guide to Theory, Algorithm, and System Development**. 1st. ed. Upper Saddle River, NJ, USA,: Prentice Hall PTR, 2001.

IRIONDO, I.; GUAUS, R.; RODRÍGUEZ, A.; LÁZARO, P.; MONTOYA, N.; BLANCO, J. M.; BERNADAS, D.; OLIVER, J. M.; TENA, D.; LONGHI, L.; AL. et. Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques. **Proc. ISCA Workshop Speech and Emotion**, 2000. p. 161–166, Set 2000.

JAIN, A. K.; DUIN, R. P.; MAO, J. Statistical pattern recognition: A review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2000. 2000.

JAMES, W. What is an emotion? **Mind – Oxford Academy**, 1884. p. 188–205, April 1884.

JOLLIFFE, I. **Principal Component Analysis**. [S.l.]: Springer Verlag, 1986.

KAMÍNSKA, D.; PELIKANT, A. Recognition of human emotion from a speech signal based on plutchik's model. **INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS**, 2012. v. 58, p. 165–170, 8 2012. Disponível em: <DOI:10.2478/v10177-012-0024-4>.

KINGESKI, R.; SCHUEDA, L. A.; PATERNO, A. S. Classificação de parâmetros de sinais de voz de filmes para reconhecimento de emoções. **XXXVI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais**, 2018. 2018.

KINGESKI, R.; SCHUEDA, L. A.; PATERNO, A. S. Reconhecimento de emoções por voz para auxílio ao diagnóstico de transtornos mentais. **Anais do XXVI Congresso Brasileiro de Engenharia Biomédica – CBEB: Sociedade Brasileira de Engenharia Biomédica**, 2018. 2018.

LATHI, B. **Sistemas De Comunicações Analógicos E Digitais Modernos**. Rio de Janeiro – RJ,: LTC, 2012.

LAZARUS, R. S.; SPEISMAN, J. C.; MORDKOFF, A. M.; DAVISON, L. A. A laboratory study of psychological stress produced by a motion picture film. **Psychological Monographs General and Applied**, 1962. v. 76, n. 34, p. 1–35, 1962.

LIFE ISN'T PERFECT – Sad Story. Canal Youtube:dixieSTARLET, 2014. Disponível em: <https://www.youtube.com/watch?v=stdvLavp5Js>. Acesso em: 28 de fevereiro de 2019.

LOW Comedy Central Apresenta Angela Dip e Paulo Vieira. Canal Youtube: Paulo Vieira, 2016. Disponível em: <https://www.youtube.com/watch?v=kQjUi64cztc>. Acesso em: 28 de fevereiro de 2019.

MARTIN, O.; KOTSIA, I.; MACQ, B.; PITAS, I. The enterface'05 audio-visual emotion database. In: **Proceedings of the 22Nd International Conference on Data Engineering Workshops**. Washington, DC, USA: IEEE Computer Society, 2006. (ICDEW '06).

MASSEY, T.; MARFIA, G.; POTKONJAK, M.; SARRAFZADEH, M. Experimental analysis of a mobile health system for mood disorders. **IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society**, 2009. v. 14, p. 241–7, 10 2009.

MCGILLOWAY, S.; COWIE, R.; DOUGLAS-COWIE, E.; GIELEN, S. C. A. M.; WESTERDIJK, M.; STROEVE, S. H. Approaching automatic recognition of emotion from voice: a rough benchmark. In: . [S.l.: s.n.], 2000.

MICHELLY Summer - Terça Insana - 15/10/2013 (HD - By Alan Junior). Canal Youtube: fotografo20001, 2013. Disponível em: <https://www.youtube.com/watch?v=GZjjDWOqAf4>. Acesso em: 28 de fevereiro de 2019.

MOST SURPRISING AUDITIONS on Got Talent! | Got Talent Global. Canal Youtube: Got Talent Global, 2017. Disponível em: <https://www.youtube.com/watch?v=W04vSVGO8PE>. Acesso em: 28 de fevereiro de 2019.

MURRAY, I. R.; ARNOTT, J. L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. **The Journal of the Acoustical Society of America**, 1993. v. 93, n. 2, p. 1097–1108, 1993.

NOLL, A. M. **The Journal of the Acoustical Society of America**, 1967. v. 41, n. 2, p. 293–309, 1967.

O'REILLY, M. K. **I Wanna Be The Guy**. 2007. Disponível em: <http://kayin.moe/iwbtg/>. Acesso em: 28 de fevereiro de 2019.

PARKOUR FAILS OF JULY 2018 | HE ALMOST DIED | PARKOUR FAIL COMPILATION. Canal Youtube: Wins and Fails, 2018. Disponível em: <https://www.youtube.com/watch?v=xN8zxjXdsRE>. Acesso em: 28 de fevereiro de 2019.

PICARD, R.; KLEIN, J. 2002. 2002.

PICARD, R. W. **Affective Computing**. Cambridge, MA, USA: MIT Press, 1997.

PLUTCHIK, R. **The Emotions: Facts and Theories, and a New Model**. [S.l.]: Random House, 1962. (Random House studies in psychology).

PRESENT, The. Dir. Jacob Frey, Ludwigsburg, Germany: Institute of Animation, Visual Effects and Digital Postproduction at the Filmakademie Baden-Wuerttemberg, 2014. Disponível em: <https://vimeo.com/152985022>. Acesso em: 28 de fevereiro de 2019.

QUAN, C.; WAN, D.; ZHANG, B.; REN, F. Reduce the dimensions of emotional features by principal component analysis for speech emotion recognition. **Proceedings of the 2013 IEEE/SICE International Symposium on System Integration**, 2013. 2013.

RABINER, L.; SCHAFER, R. **Digital Processing of Speech Signals**. [S.l.]: Prentice-Hall, 1980. (Prentice-Hall Signal Processing Series: Advanced monographs).

RABINER, L. R.; SCHAFER, R. W. **Introduction to Digital Speech Processing**. Hanover, MA, USA: Now Publishers Inc., 2007.

RABINER, R. W. S. L. R. **Theory and Applications of Digital Speech Processing**. [S.l.]: Pearson, 2010.

SCHACHTER, S.; SINGER, J. E. Cognitive, social, and physiological determinants of emotional state. **Psychological Review**, 1962. v. 69, p. 379–399, 1962.

SCHERER, K. R. Nonlinguistic vocal indicators of emotion and psychopathology. **Emotions in Personality and Psychopathology**, 1979. p. 493–529, 1979.

SCHERER, K. R. Vocal affect expression: A review and a model for future research. **Psychological Bulletin**, 1986. v. 99, p. 143–165, 1986.

SCHMITTER, A. M. **17th and 18th Century Theories of Emotions, The Stanford Encyclopedia of Philosophy**. 2016. Disponível em: <https://plato.stanford.edu/entries/emotions-17th18th/LD1Background.html>. Acesso em 07 de Setembro de 2017.

SCHULLER, B.; VLASENKO, B.; EYBEN, F.; RIGOLL, G.; WENDEMUTH, A. Acoustic emotion recognition: A benchmark comparison of performances. In: . [S.l.: s.n.], 2010. p. 552 – 557.

SHARMA, R.; NEUMANN, U.; KIM, C. Emotion recognition in spontaneous emotional utterances from movie sequences. **Proceedings of the WSEAS International Conference on Electronics, Control & Signal Processing**, 2002. 2002.

SNEDDON, I.; MCRORIE, M.; MCKEOWN, G.; HANRATTY, J. The belfast induced natural emotion database. **IEEE Transactions on Affective Computing**, 2012. Institute of Electrical and Electronics Engineers, v. 3, p. 32–41, 1 2012.

SPINOZA. **Ética**. [S.l.]: Editora Autêntica, 2009.

TACCONI, D.; MAYORA, O.; LUKOWICZ, P.; ARNRICH, B.; SETZ, C.; TROSTER, G.; HARING, C. Activity and emotion recognition to support early diagnosis of psychiatric diseases. In: . [S.l.: s.n.], 2008. p. 100 – 102.

THEODORIDIS, S.; KOUTROUMBAS, K.; NOS, K.; BAS, K. M. **Pattern Recognition – Second Edition**. [S.l.: s.n.], 2006.

UNEXPECTED Compilation. Canal Youtube: Cipwreck, 2016. Disponível em: <https://www.youtube.com/watch?v=xjk7hL8keuI>. Acesso em: 28 de fevereiro de 2019.

VERVERIDIS, D.; KOTROPOULOS, C. Emotional speech recognition: Resources, features, and methods. **Speech Communication**, 2006. v. 48, n. 9, p. 1162–1181, 2006.

VERVERIDIS, D.; KOTROPOULOS, C.; PITAS, I. Automatic emotional speech classification. **IEEE International Conference on Acoustics, Speech, and Signal Processing**, 2004. 2004.

VIGILANTE se despede do seu Cachorro Pastor Alemão. Canal Youtube: Videos 1000, 2014. Disponível em: <https://www.youtube.com/watch?v=FazK02PLA00>. Acesso em: 28 de fevereiro de 2019.

WANG, S.; LING, X.; ZHANG, F.; TONG, J. Speech emotion recognition based on principal component analysis and back propagation neural network. **2010 International Conference on Measuring Technology and Mechatronics Automation**, 2010. 2010.