BRUNO ANDREETTO
DANIEL TRIGO RIZZUTO
MAURÍCIO IMAEDA

ANÁLISE EMOCIONAL DERIVADA DO PROCESSAMENTO DE VOZ

São Paulo
2018

BRUNO ANDREETTO
DANIEL TRIGO RIZZUTO
MAURÍCIO IMAEDA

ANÁLISE EMOCIONAL DERIVADA DO PROCESSAMENTO DE VOZ

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do título de Graduação em Engenharia.

Área de Concentração:

Engenharia de Computação

Orientador:

Prof. Dr. Ricardo Luis de Azevedo da Rocha

São Paulo
2018

RESUMO

A análise emocional da voz humana é um dos maiores desafios no processamento da fala. Muitos sistemas foram desenvolvidos para realizar a detecção de emoções a partir da linguagem natural falada.

Este estudo apresenta a pesquisa, elaboração e os resultados de um sistema voltado à identificação de emoções presentes na fala de um interlocutor através do algoritmo de classificação Naive Bayes aplicado sobre dimensões extraídas através do processamento e análise das faixas de áudio.

Palavras-Chave – Naive Bayes, Aprendizado de Máquina, Reconhecimento Emocional, Processamento da Fala.

ABSTRACT

The emotional analysis of the human voice is one of the greatest challenges in speech processing. Many systems have been developed to perform the detection of emotions from the human natural language.

This study presents the research, development and results of a system aimed at identifying the emotions present in the speech of an interlocutor through the Naive Bayes classification algorithm applied on dimensions extracted through the processing and analysis of the audio tracks.

Keywords – Naive Bayes, Machine Learning, Emotional Recognition, Speech processing.

LISTA DE FIGURAS

1 Fluxograma da metodologia 10

2 Roda das emoções de Plutchik 12

3 FFT aplicada à energia. Felicidade (vermelho) x Raiva (verde) 23

4 Estrutura de pastas 33

5 Correlação taxa de acerto pela variância explicada 34

6 Taxa de acerto utilizando o desvio padrão 35

7 Taxa de acerto utilizando a média 36

8 Taxa de acerto utilizando a frequência natural 37

9 Taxa de acerto utilizando a mediana 38

10 Taxa de acerto utilizando a soma 39

LISTA DE TABELAS

1 *Features* extraídas 19

2 Componentes principais da ACP obtidas a partir da média, desvio padrão dos dados 21

3 Componentes principais da ACP obtidas a partir da média, desvio padrão e frequência natural dos dados 22

4 Resultados obtidos utilizando Raiva, Medo, Felicidade e Tristeza 25

5 Resultados obtidos na classificação de todas as emoções 26

SUMÁRIO

1 Introdução **8**

1.1 Motivação 9

1.2 Objetivo 9

1.3 Metodologia 10

2 Fundamentos **11**

2.1 Conceitos sobre emoções 11

2.2 Análise de componentes principais (ACP) 13

2.3 Naive Bayes 14

2.4 Tipos de erros 15

3 Procedimentos **16**

3.1 Organização da equipe 16

3.2 Seleção do banco de dados 17

3.3 Escolha da linguagem de programação 18

3.4 Extração das features 18

3.5 Observação inicial 20

3.6 Aplicação da análise de componentes principais (ACP) 20

3.7 Transformada rápida de Fourier (FFT) 23

3.8 Desenvolvimento do classificador 24

3.9 Observação dos resultados 25

4 Considerações finais **28**

4.1 Trabalhos futuros 28

4.1.1 Aprimorar a predição 28

4.1.2 Aplicação como produto 28

5 Anexos **31**

5.1 Arquitetura de software 31

5.1.1 Banco de dados de áudios 31

5.1.2 Extração de *features* 32

5.1.3 Tratamento das *features* 33

5.1.4 Seleção das dimensões 34

5.2 Resultado da classificação emocional 34

Referências **40**

1 INTRODUÇÃO

A linguagem verbal (linguagem falada) é a forma de expressão que acompanha a humanidade desde a pré-história. Ela, como a humanidade, passou por uma série de modificações e aperfeiçoamentos para chegar ao patamar que está atualmente. Diferente da pré-história, tempo em que a comunicação oral era feita principalmente por gemidos e grunhidos. Hoje, a fala é complexa e carrega muito mais informações do que uma simples mensagem de texto. Com ela, é possível extrair várias informações que um texto por si só não contém.

A partir de um dialogo, é possível obter várias informações, tais como: variações no volume, variações no timbre, oscilações e falhas (gagueira). Esses dados depois de processados se tornam muito valiosos para diversas aplicações.

Tratando essas informações, pode-se detectar a emoção presente na voz do interlocutor. Uma empresa de *call centers e telemarketings*, por exemplo, poderia se beneficiar do uso desses dados extraídos dos telefonemas realizados para traçar perfis e padrões de comportamento de seus clientes e assim otimizar o atendimento. Tendo em mãos o estado emocional do cliente do outro lado da linha, o operador consegue conduzir o telefonema de uma forma mais empática e coerente com o emocional do interlocutor.

1.1 Motivação

Os algoritmos de detecção de fala existentes atualmente estão em um estágio bem avançado, conseguindo capturar e traduzir com perfeição falas de diferentes interlocutores e em diferentes linguagens. Essa tecnologia já está madura o suficiente para se observar várias aplicações práticas no nosso dia a dia, como por exemplo, a maioria dos *smartphones* atuais possui algum tipo de assistente pessoal que utiliza o reconhecimento de fala como interface.

Porém, um aspecto que ainda não está maduro é o reconhecimento de emoções presentes na fala, e com o amadurecimento do reconhecimento da fala, as atenções estão se voltando para este campo ainda pouco explorado.

Com o avanço de pesquisas de inteligencia artificial e o surgimento de interfaces baseadas na fala, a capacidade de um computador de inferir a emoção de um interlocutor pela sua fala vem se tornando cada vez mais importante, e as aplicações decorrentes disso são inesgotáveis, podendo ser desde prover um *feedback* instantâneo sobre o sentimento do usuário para tecnologias interativas (em um jogo de computador por exemplo) ou até construir uma inteligência artificial capaz de se adaptar ao estado emocional de um ser humano.

Dito isto, consideramos um campo de pesquisa extremamente importante para o avanço das tecnologias atuais e fizemos deste trabalho uma prova de conceito do que pode ser feito com os algoritmos e com o poder computacional atual no campo referente ao reconhecimento emocional a partir da fala humana.

1.2 Objetivo

O propósito desse projeto é desenvolver um sistema que analisa e classifica o estado emocional de uma pessoa a partir de características apresentadas em sua fala.

Para atingir esse objetivo, o projeto consiste em um sistema que receberá um arquivo de áudio como parâmetro, analisará a entrada e fornecerá a emoção que fora detectada na fala presente no áudio de entrada.

1.3 Metodologia

Para uma aplicação de machine learning, fazem-se necessários os dados para treinamento do modelo, as características extraídas dos dados (*features*) e um classificador.

Neste estudo, os dados para o treinamento foram obtidos do banco de dados *The Ryerson Audio-Visual Database of Emotional Speech and Sound (RAVDESS)*. A extração das *features* dos áudios foi realizada utilizando a biblioteca em python *pyAudioAnalysis*. A classificação foi realizada através do algoritmo de aprendizado de máquina supervisionado *Naïve Bayes*.

O fluxograma exibido na imagem 5.1 indica de maneira visual o funcionamento geral do sistema.

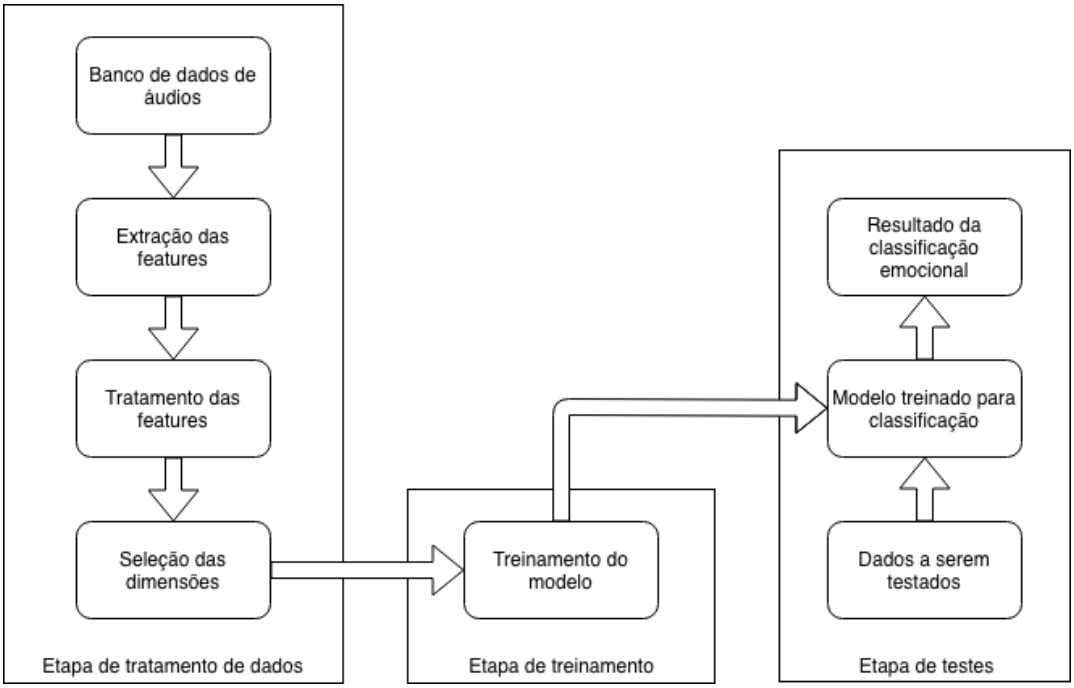


Figura 1: Fluxograma da metodologia

Uma análise mais detalhada da arquitetura interna do sistema pode ser encontrada na seção de anexos deste documento.

2 FUNDAMENTOS

2.1 Conceitos sobre emoções

A teoria Psicoevolucionária das Emoções de Robert Plutchik define que existem oito emoções biologicamente primitivas, denominadas emoções primárias: irritação, medo, tristeza, nojo, surpresa, antecipação, confiança e felicidade.

Essas emoções se desenvolveram ao longo das gerações por incrementar a aptidão reprodutiva animal. Isto é, indivíduos capazes de sentir cada emoção e, consequentemente, reagir de acordo com o comportamento provocado possuem maior probabilidade de sobrevivência e reprodução.

A própria palavra emoção, derivada do latim *emovere*, energia em movimento, carrega em si o significado de inspirar reações, no indivíduo que a sente, capazes de alterar o seu estado atual. O medo por exemplo, dispara a reação de fuga ou luta. A antecipação, por outro lado, desperta o interesse de desbravar e descobrir o desconhecido.

Outras emoções, como tédio, aprovação, interesse e distração, são composições das oito emoções primárias.

Plutchik realizou o mapeamento das relações entre as emoções existentes em um figura conhecida como Roda das emoções que é apresentada na figura 2:

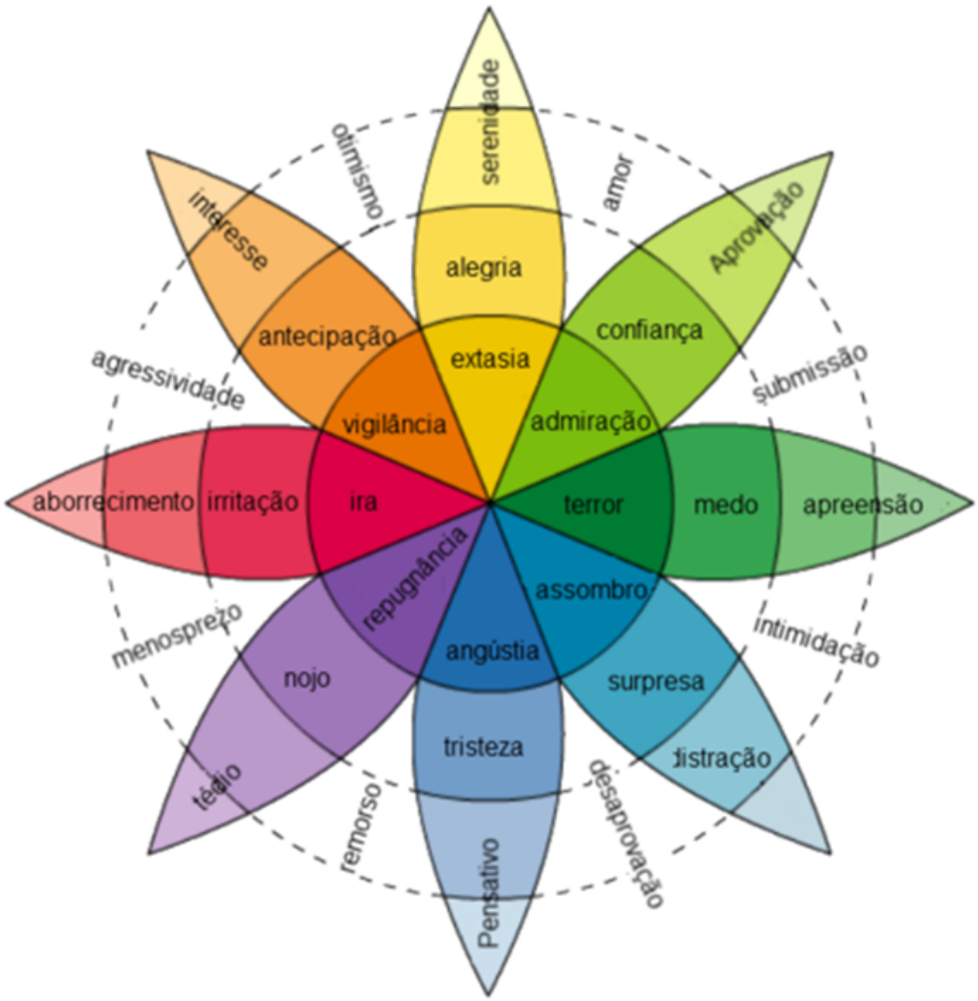


Figura 2: Roda das emoções de Plutchik

Na Roda das emoções, quanto mais distantes estão duas emoções, mais antagônica são entre si. Assim, para as emoções primárias, podemos destacar quatro pares de emoções totalmente adversas: felicidade e tristeza, nojo e confiança, irritação e medo e antecipação e surpresa.

2.2 Análise de componentes principais (ACP)

A análise de componentes principais - ACP, do inglês *Principal component analysis*, é um método que tem a finalidade principal, a análise de dados, visando a eliminação de dados redundantes e escolher uma base vetorial de referência que maximiza a variância entre os dados originais.

O método originalmente apresentado por Karl Pearson (1857 - 1936), em 1901, em seu *paper*: *"On lines and planes of closest fit to systems of points in space"* [1], demonstra um algoritmo que, dada uma matriz de dados X de dimensão $m*n$, contendo n valores representados em m dimensões, o ACP encontra um novo conjunto de k dimensões (ou um novo referencial) onde todas as dimensões são ortogonais entre si, *i.e.* linearmente independentes, ordenadas por variância.

Os passos para calcular os componentes principais são simples: calcular a matriz de covariância C_x da matriz de dados X; calcular os auto-vetores e seus respectivos auto-valores; arranjar em ordem decrescente os auto-vetores em função de seus auto-valores; escolher um novo numero k de dimensões, dado que $k \leq m$, onde m é o numero de dimensões dos dados originais; transformar os dados originais da antiga base vetorial M para a base K.

Analisando um pouco melhor o que essas operações significam, dada que a matriz de covariância pode ser calculada por:

$$C_x = \frac{1}{n-1}(X - \overline{X})(X - \overline{X})^T \tag{2.1}$$

O resultado dessa operação será uma matriz contendo em sua diagonal a variância dos dados e na resto da matriz a covariância, como representado no exemplo a seguir de uma matriz *3*3*:

$$C_x = \begin{bmatrix} V_a & C_{a,b} & C_{a,c} \\ C_{a,b} & V_b & C_{b,c} \\ C_{a,c} & C_{b,c} & V_c \end{bmatrix} \tag{2.2}$$

Ao calcular os auto-vetores basicamente está se realizando a diagonalização da matriz, ou seja, o novo referencial será um referencial em que os dados possuem covariância zero e variância máxima. Ordenando os auto-vetores por seus auto-valores (que são agora as novas variâncias dos dados) estamos ordenando por explicação da variância, então

podemos analisar e escolher uma quantidade menor de dimensões que explicam uma certa quantidade da variância que seja suficiente para o treinamento do nosso modelo, e.g.: ao escolher os k, com $k \leq m$, com primeiros componentes da ACP, estaríamos explicando 90% da variância do modelo m-dimensional, podendo ser suficiente para a análise em questão.

Se tratando de uma aplicação de *machine learning* este fato é muito importante, ainda mais quando poucos dados estão disponíveis para o treinamento, pois pode-se diminuir a quantidade de dados mas aproveita-se ao máximo a quantidade de informação contida dentro deles. Esta técnica pode ser crucial no desempenho de alguns modelos.

2.3 Naive Bayes

O Naive Bayes é um algoritmo probabilístico simples baseado no teorema de Bayes. O teorema, criado pelo pastor e matemático Thomas Bayes (1701 - 1761), descreve a probabilidade de um evento ocorrer dado o conhecimento prévio relacionado ao seu acontecimento.

O algoritmo de Naive Bayes se utiliza da probabilidade condicional do teorema de Bayes para realizar a classificação. Ele recebe o nome de Naive, do inglês ingênuo, devido ao fato de não considerar a existência de correlações entre as variáveis (*features*) e tratá-las como independentes.

Atualmente, o algoritmo tem sido altamente utilizado em aplicações de inteligência artificial e *machine learning* para categorizar textos baseado na frequência das palavras e, assim, classificar se um e-mail se trata de um *spam* ou não.

A escolha do algoritmo para esta aplicação se deu principalmente pela sua velocidade de execução, que permite rápidas análises e classificação de um conjunto de características fornecidos. Dessa forma, aplicações em tempo real podem ser futuramente derivadas deste estudo com bom desempenho prático.

2.4 Tipos de erros

A análise dos erros cometidos pelo algoritmo de classificação dado um modelo e um conjunto de testes é fundamental para que se possa realizar a comparação entre as alterações realizadas no modelo e a consequente determinação se as mudanças surtiram de fato efeitos positivos ou negativos na assertividade da classificação.

Serão avaliados dois tipos diferentes de erros neste estudo:

- Tipo I: também chamado de falso-positivo, se dá quando a hipótese nula é rejeitada quando, em realidade, é verdadeira;

- Tipo II: também denominado falso-negativo, ocorre quando se dá a aceitação da hipótese nula, sendo que a hipótese é falsa.

Dada a definição de cada tipo de erro, para a classificação de um conjunto de faixas de áudio dentre as emoções, teremos a taxa de cada um dos dois tipos de erro para cada emoção. A esta taxa, é dada o nome de medida de evidência (*p-valor).*

Para a emoção felicidade por exemplo, ocorrerá um erro tipo I no caso em que uma faixa referente à qualquer emoção diferente de felicidade for classificada pelo algoritmo como felicidade. Um erro tipo II ocorrerá no caso em que uma faixa que sabidamente representa a emoção felicidade for classificada como qualquer outra emoção diferente de felicidade.

3 PROCEDIMENTOS

3.1 Organização da equipe

Durante o desenvolvimento de projetos de tecnologia, muitas vezes o gerenciamento do projeto em si constitui um imenso desafio a ser superado para que os objetivos propostos sejam alcançados.

A gestão do tempo, pessoas, recursos disponíveis, a priorização de quais tarefas serão executadas, assim como a ordem em que serão executadas, são elementos intrínsecos em qualquer projeto e que afetam diretamente seu resultado final.

A fim de garantir uma gestão eficiente no decorrer do desenvolvimento deste projeto, escolheu-se utilizar uma metodologia que vem ganhando espaço imenso em projetos de tecnologia e é utilizada por gigantes empresas do setor como *Google*, *Facebook* e *Yahoo*!. Trata-se da metodologia ágil [6].

Para este projeto, foi escolhida uma metodologia baseada no modelo ágil *SCRUM* [5], com algumas adaptações que fazem sentido para os integrantes da equipe. As reuniões para alinhamento foram estabelecidas entre os integrantes toda segunda-feira. Essas reuniões, juntamente com serviços de trocas de mensagem substituia as *dailys* do SCRUM. Definiu-se que toda sexta-feira seria realizada uma sessão de orientação entre os integrantes da equipe e o orientador e que, nesta reunião seriam definidos os próximos objetivos a serem atingidos, juntamente com a elaboração e atribuição das tarefas a serem executadas por cada integrante do time na próxima iteração de desenvolvimento de uma semana. A sessão de orientação desempenha o mesmo papel que a reunião de *planning* possui no SCRUM.

O modelo de iterações, cada uma denominada um *sprint*, com duração de uma semana, permite que a equipe mantenha-se alinhada e que possa se observar o resultado do que se desenvolveu na iteração em relação aos objetivos definidos de forma rápida. Se há algo que está sendo desenvolvido e não está ajudando a equipe a alcançar os objetivos,

a implementação dessa funcionalidade do sistema pode ser abortada em detrimento de outra solução ou adaptada para que seu resultado faça mais sentido para que se atinja os objetivos pré-estabelecidos.

A organização dos *sprints*, assim como a gerência do andamento das tarefas durante as fases de planejamento, execução e análise de resultados foi realizada utilizando a ferramenta *Asana*.

3.2 Seleção do banco de dados

A construção ou seleção de um banco de dados é o primeiro passo para qualquer aplicação de *machine learning* supervisionada. É necessário que se tenha dados com boa qualidade e classificados, que representem de forma similar as situações reais nas quais o classificador deverá atuar, uma vez que estes dados serão a base para o treinamento do modelo do classificador.

Por estas razões, a elaboração de uma base de dados em que se possa confiar é um processo extremamente custoso e demorado. No caso de uma base composta por áudios, é necessário um instrumental incluindo microfones de alta qualidade, ambiente silencioso e com ótima acústica para as gravações, tratamento de ruídos, tratamento dos arquivos de áudio e um processo rigoroso quanto à classificação para que se siga um padrão bem definido.

Tratando-se de áudios que representem seres humanos expressando emoções, adiciona-se outra camada de complexidade ao processo, pois necessita-se validar a classificação da emoção transmitida por cada falante em cada faixa de áudio. Este processo acaba por ser subjetivo e passível de erros.

Atualmente, existem poucos centros de pesquisas que dedicaram esforços para desenvolver uma base de dados de áudios classificados por emoção.

Como exemplo de caso de bases bem fundamentadas e com grande quantidade de áudios, tem-se a base *Emo-DB* [2]. Patrocinada pela Fundação Alemã de Pesquisas, a base apresenta áudios gravados em estúdio por atores que representaram sete emoções, através de dez diferentes frases. A base é toda no idioma alemão.

Outro exemplo é a base no idioma inglês *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)* [3]. Esta base foi composta por áudios gravados por 24 atores prossionais, sendo 12 homnens e 12 mulheres. Cada ator foi responsável por

gravar duas frases em cada uma das oito emoções em duas intensidades diferentes.

Infelizmente, em português não há nenhuma base de dados sólida o suficiente que possa ser utilizada com confiança.

Dessa forma, como a semelhança fonética entre português e inglês é superior à semelhança fonética entre português e alemão, decidiu-se utilizar a base de dados *RAVDESS* para este estudo.

3.3 Escolha da linguagem de programação

Para a realização dessa pesquisa a linguagem de desenvolvimento escolhida pelo grupo foi o *Python*. Essa linguagem de programação muito poderosa foi escolhida por dois principais motivos:

O primeiro motivo não está relacionado à linguagem em si, mas sim às ferramentas para ela disponíveis. O *Python* apresenta uma vasta lista de bibliotecas muito úteis em aplicações que envolvem cálculos estatísticos e também *machine learning*, como as bibliotecas: *Numpy*, utilizada para trabalhar e manipular matrizes multidimensionais e o *Scikit learning*, utilizado para criar, treinar e testar o classificador utilizado na pesquisa. A existência dessas bibliotecas em uma única linguagem foi um divisor de águas para a decisão.

O segundo e último motivo para essa escolha, deve-se a simplicidade e facilidade que se tem para desenvolver uma aplicação nessa linguagem. O *Python* por ser uma linguagem de alto nível apresenta uma sintaxe muito simples e próxima da linguagem humana, fato esse que contribui para redução da curva de aprendizado e do tempo para o desenvolvimento da aplicação.

3.4 Extração das features

Os áudios correspondentes a cada emoção de cada ator do banco de dados RAVDESS foram tratados a fim de extrair as características (features) do som presentes em cada arquivo.

Para tanto, definiu-se uma faixa de 50ms de intervalo e 25ms de sobreposição entre a captação das características presentes. Sendo assim, para cada áudio, foram analisadas as características do som presentes em intervalos de 50ms com uma distância de 25ms para

cada intervalo.

A fim de realizar a análise, foi utilizada a biblioteca em python *pyAudioAnalysis* [4], capaz de realizar a extração de trinta e quatro features do som. As características extraídas pela biblioteca estão descritas na tabela 1:

Nº Feature	Nome	Descrição
1	Zero Crossing Rate	A taxa de mudança de sinal durante a duração do frame
2	Energy	A soma dos quadrados dos valores dos sinais, normalizado pelo tamanho do frame
3	Entropy of Energy	A entropia das energias normalizadas dos subframes. É uma medida que quantifica mudanças bruscas
4	Spectral Centroid	O centro de gravidade do espectro
5	Spectral Spread	O segundo momento central do espectro
6	Spectral Entropy	A entropia da energia do espectro normalizada
7	Spectral Flux	O quadrado da diferença entre as magnitudes normalizadas de dois frames seguidos
8	Spectral Rolloff	A frequência sobre a qual 90% da distribuição de magnitude do sinal está concentrada abaixo
9-21	Mel Frequency Cepstral Coefficients	Mel Frequency Cepstral Coefficients
22-33	Chroma Vector	Lisa de 12 elementos representando a energia espectral
34	Chroma Deviation	O desvio padrão dos 12 coeficientes Chroma

Tabela 1: *Features* extraídas

Por se tratar de um banco de dados com vinte e quatro atores, cada um atuando sobre oito emoções em duas frases diferentes cada, há um total de 384 faixas de áudio para realizar a extração das características. Para isto, foi desenvolvido um *script* para realizar a extração de forma automática através da biblioteca *pyAudioAnalysis* e organizar os arquivos em pastas de uma maneira que seja fácil encontrar qualquer informação sobre

qualquer ator, emoção e frase.

3.5 Observação inicial

Ao realizar a extração das features através do *pyAudioAnalysis*, foi desenvolvido uma função em *python* que recebe três parâmetros: um número inteiro entre 0 e 7, indicando a emoção, um número inteiro de 0 à 24, indicando o ator e um número inteiro de 0 à 33, indicando a *feature* a ser analisada. A função então realiza a leitura dos dados relacionados aos parâmetros fornecidos e apresenta o gráfico da curva relacionada.

Como forma de generalização, a função foi desenvolvida de modo que se algum dos parâmetros for omitido em sua chamada, todos os valores associados aos outros parâmetros serão exibidos. Por exemplo: se a chamada omitir o parâmetro representando a emoção, porém especificar o ator e a *feature*, serão exibidos, de forma conjunta, as curvas de todas as features relacionadas às frases do ator especificado para todas as emoções. Se todos os parâmetros forem omitidos, serão exibidos, no mesmo gráfico, as curvas de todas as features, para todos os atores e para todas as emoções.

Dessa maneira, verificar em gráficos os dados obtidos pela extração ficou simples e pudemos realizar diversas observações iniciais a olho nu.

O objetivo dessas análises iniciais, seria observar como emoções antagônicas divergem entre cada uma das *features* coletadas.

Assim, foram analisados os gráficos de Felicidade x Tristeza e de Ira x Medo para cada uma das características extraídas pelo *pyAudioAnalysis*.

3.6 Aplicação da análise de componentes principais (ACP)

A aplicação direta das 34 *features* obtidas pelo tratamento dos áudios apresentou um resultado pouco satisfatório.

Algumas das características utilizadas no treinamento do classificador pouco variavam entre as diversas emoções. Estas *features* com baixa variabilidade acabam por ofuscar o verdadeiro potencial classificador de outras características que realmente representam diferenças maiores entre as emoções.

Dessa forma, a fim de reduzir ao máximo dados redundantes e obter apenas as com-

ponentes que apresentam a maior variância possível para classificação das emoções, foi aplicado o método de Análise de Componentes Principais (ACP).

Para tanto, aplicou-se a média e o desvio padrão para cada uma das trinta e quatro *features* a fim de obter-se os dados de entrada para a ACP.

A aplicação da ACP sobre os 68 dados de entrada (média e desvio padrão de cada característica) forneceu as componentes ordenadas em forma decrescente de variância. Todavia, o que se observou é que a primeira componente obteve variância de 96.291% e a segundo componente 1.835%. A tabela 2 indica as 5 primeiras componentes com suas respectivas variâncias.

Nº Componente	Variância
1	96.291%
2	1.835%
3	1.152%
4	0.0272%
5	0.090%

Tabela 2: Componentes principais da ACP obtidas a partir da média, desvio padrão dos dados

Este resultado mostra que a variância entre as componentes está mal distribuída e isto não é positivo para a análise, pois a maioria das componentes só confunde o classificador, invés de fornecer dados relevantes para a classificação.

Assim, buscou-se maneiras de tratar os dados de maneira mais eficiente para que a distribuição de variância das componentes fizesse maior sentido para a classificação dos dados.

Foi então, aplicado sobre os dados a Transformada Rápida de Fourier (FFT), como explicado na seção 3.7 para que se obtivesse a frequência natural de cada uma das 34 *features* e utilizou-se este valor, em conjunto com a média e o desvio padrão de cada feature, como entrada para a ACP. As 25 primeiras componentes obtidas pelo resultado desta nova aplicação está descrito na tabela 3.

Nº Feature	Variância
1	12.548%
2	7.075%
3	5.793%
4	5.663%
5	5.084%
6	4.938%
7	4.686%
8	4.311%
9	4.291%
10	4.075%
11	3.716%
12	3.452%
13	3.313%
14	3.089%
15	2.698%
16	2.533%
17	2.202%
18	2.172%
19	2.053%
20	1.658%
21	1.626%
22	1.523%
23	1.312%
24	1.299%
25	1.212%

Tabela 3: Componentes principais da ACP obtidas a partir da média, desvio padrão e frequência natural dos dados

Nota-se que a distribuição da variância entre as componentes está bem mais distribuída. Isto significa que cada uma das componentes agrega muito mais valor na classificação do que as componentes obtidas antes da utilização da frequência natural.

Dessa forma, dispõe-se de dados muito mais refinados e coerentes para a aplicação no classificador Naive Bayes.

3.7 Transformada rápida de Fourier (FFT)

Todos os dados extraídos dos áudios são obtidos de pequenos *frames* divididos na escala do tempo. Dessa forma, para cada áudio, têm-se 34 características extraídas e, para cada uma dessas características, têm-se um gráfico de como aquela *feature* evolui no decorrer do tempo.

A aplicação da transformada rápida de Fourier a cada uma das 34 *features* permite uma análise dos dados no domínio da frequência.

Através da análise visual dos gráficos das transformadas obtidas, pôde-se observar que, para emoções antagônicas, as componentes de frequência que compunham os gráficos divergia de maneira considerável. Um exemplo dessa observação pode ser vista no gráfico representado na figura 3, onde é exibida a curva da FFT aplicada à característica energia para as emoções felicidade e raiva.

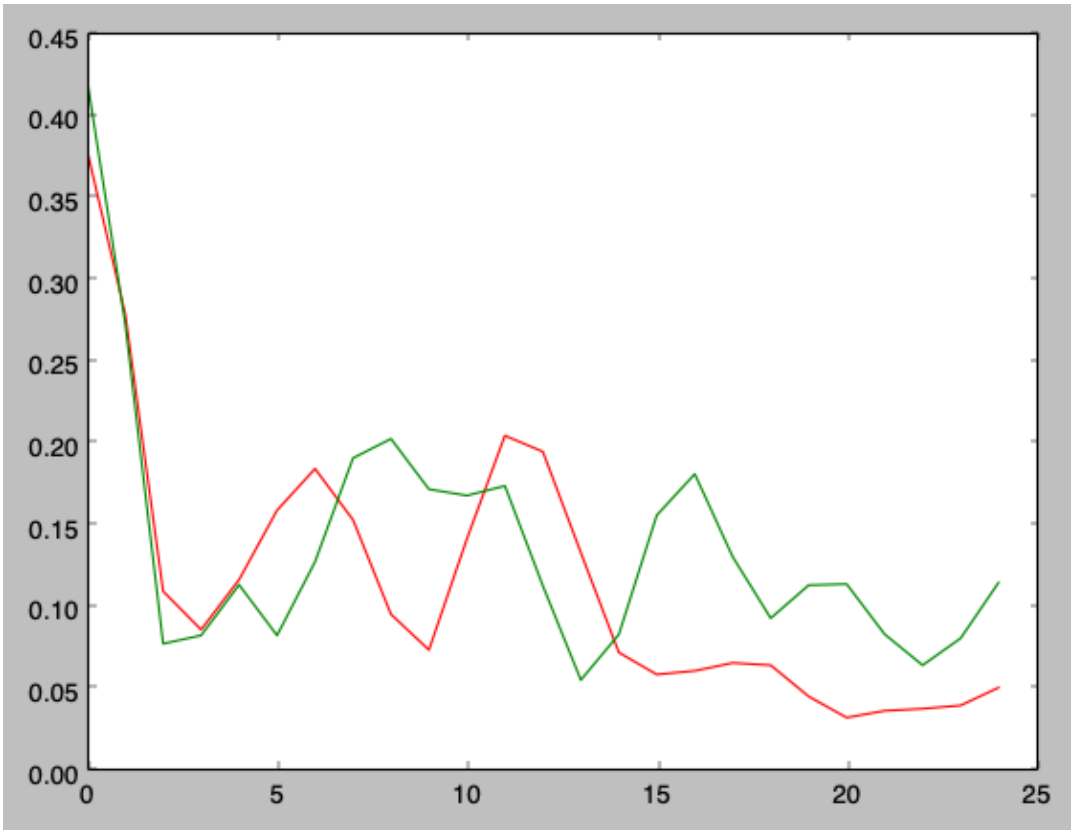


Figura 3: FFT aplicada à energia. Felicidade (vermelho) x Raiva (verde)

A fim de incorporar essa característica dos sinais, no domínio da frequência, no treinamento do modelo, obteve-se a frequência natural correspondentes ao gráfico de cada uma das 34 características após aplicar a técnica da transformada rápida de Fourier.

Dessa forma, a frequência natural obtida pode ser utilizada como entrada para o treinamento do modelo a ser utilizado pelo classificador. O desenvolvimento do classificador será descrito a seguir na seção 3.8.

3.8 Desenvolvimento do classificador

Como descrito no capítulo *Fundamentos*, o classificador utilizado na pesquisa é o *Naive Bayes*. Esse classificador está inbutido na biblioteca escrita em *python* voltada à *machine learning*, denominada *scikit learn*.

Como entrada para o treino, o classificador recebe duas matrizes, uma de dados e outra composta pelos identificadores dos respectivos dados.

Nessa pesquisa, os dados de entrada são os componentes gerados depois da aplicação da ACP para cada áudio e os identificadores são os nomes das emoções presentes em cada áudio fornecido na entrada.

Para o teste, o classificador recebe a entrada contendo os dados obtidos através do tratamento de cada áudio a ser testado. O tratamento consiste na extração das características do áudio, obtenção da transformada de Fourrier para extrair a frequência natural, reduz-se as curvas das características em valores únicos que as descrevem e aplica-se a ACP a fim de obter as componentes com a maior variância possível.

É interessante notar que como os dados de entrada são dados contínuos, surgiu a hipótese de que se os dados fossem separados e classificados por faixas de desvio, poderiam ser obtidos melhores resultados com o classificador *Naive Bayes*.

Para isso, foi desenvolvida uma função que, dado um parâmetro indicando o valor do desvio desejado, realiza a categorização dos dados em grupos classificados. Isto é, dados dois conjuntos de dados que, para a mesma componente obtidas pela ACP, possuem os valores de *3.23* e *3.25* respectivamente, se for definido o desvio de *0.02*, ambos serão classificados no mesmo grupo para aquela componente da ACP. Esta função é aplicada a todos os dados correspondentes aos áudios de entrada no classificador.

Quando o classificador recebe estes dados, retorna a predição da emoção com a maior probabilidade de estar correta, seguindo o que o modelo treinado previamente permitiu calcular.

Para a análise do classificador, foi empregado uma razão de treino e teste de 80/20, ou seja, 80% dos dados são aplicados para treinar e 20% utilizados para testar o modelo

treinado.

Além disso, foi empregada uma rotação entre treino e teste, de forma que parte dos dados utilizados para treinar o algoritmo são separados e aplicados na fase de testes. De forma semelhante, parte dos dados utilizados na etapa de testes são separados e usados para treinar o algoritmo. Com essa rotação, espera-se não perder dados que poderiam ser cruciais na classificação das emoções durante o treinamento do modelo.

3.9 Observação dos resultados

Após o treino do modelo e a realização de vários testes, conseguimos obter os resultados, utilizando 4 emoções Raiva, Medo, Felicidade e Tristeza. Os resultados obtidos são exibidos na tabela 4:

	Raiva	Medo	Felicidade	Tristeza
Raiva	**56.41%**	12.88%	23.25%	7.46%
Medo	10.38%	**52.33%**	21.98%	15.31%
Felicidade	10.94%	15.12%	**64.70%**	9.23%
Tristeza	9.17%	24.69%	14.21%	**51.92%**

Tabela 4: Resultados obtidos utilizando Raiva, Medo, Felicidade e Tristeza

Estes resultados foram obtidos utilizando 80% dos dados para treino do modelo e 20% para testes, nos quais os acertos acima são a média de 5 iterações, alternando a fatia usada para testes a cada iteração, testando efetivamente 100% da base.

Para atingir esses resultados, foram utilizados 51 componentes dos 204 extraídos pela ACP, com os quais é possível explicar 80% da variância de toda a base de dados.

O motivo para utilizar apenas 80% da variância da base se deu por um fator de super-adequação (*overfitting*). Ao treinar o modelo com 100% da variância ele se tornava extremamente específico para classificar a apenas os dados com os quais foi treinado, diminuindo consideravelmente a acuidade para dados novos (no caso a fatia dos dados que não foram usados para o treino). A explicação de 80% da variância pela ACP foi o ponto encontrado que resultava no melhor equilíbrio entre acuidade para dados de treino/acuidade para dados de teste, sem sacrificar muito a capacidade de classificação do modelo por poucos dados.

Antes de aplicar a ACP, foi feita a normalização dos dados, trazendo a média para 0 e variância para 1, o que fez grande diferença na classificação e nos resultados da ACP.

Foi observado também que a tentativa de definir faixas para os valores a serem aplicados no modelo (para tentar agrupar valores próximos em uma categoria a fim de evidenciar melhor características de uma emoção) não resultou em nenhuma melhoria na acuidade do modelo, na prática observou-se uma variação negativa na acuidade com o aumento dessa faixa, então essas faixas foram removidas do modelo final utilizando os dados originais para o treinamento.

A emoção com maior acurácia foi a Felicidade, com 64.70% de acerto, seguida da Raiva com 56.41% e Medo com 52.33%. Como pior resultado está a Tristeza com 51.92% de acuidade.

Dado que são quatro categorias a serem classificadas, um chute corresponderia a uma chance de acerto de 25%. Observa-se com o modelo que as taxas de acerto possuem valores que variam de 207,68% a 262,8% superiores a um chute.

A justificativa detalhada desses resultados pode ser encontrada em anexo no final deste documento.

Após a obtenção dos resultados no escopo simplificado de quatro emoções, foi aplicada a técnica novamente para a categorização de todas as oito emoções. O resultado está exposto na tabela 5.

	Raiva	Calma	Nojo	Medo	Felicidade	Neutro	Tristeza	Surpresa
Raiva	**36.58%**	7.48%	17.98%	8.37%	10.67%	3%	4.05%	11.87%
Calma	5.67%	**28.36%**	10.12%	11.47%	2.07%	17.84%	22.39%	2.07%
Nojo	20.86%	8.24%	**34.57%**	7.56%	8.45%	15.31%	2.90%	2.11%
Medo	5.32%	4.68%	5.32%	**44.96%**	16.49%	6.81%	7.06%	9.35%
Felicidade	7.72%	4.34%	7.44%	10.98%	**42.35%**	8.74%	4.33%	14.09%
Neutro	5.93%	16.58%	4.40%	1.86%	14.96%	**34.10%**	9.33%	12.83%
Tristeza	5.28%	21.30%	7.02%	15.17%	2.50%	19.79%	**26.44%**	2.50%
Surpresa	12.26%	1.00%	6.68%	11.41%	18.56%	9.16%	3.91%	**37.02%**

Tabela 5: Resultados obtidos na classificação de todas as emoções

Nota-se que com a utilização de mais categorias a serem classificadas, a acuidade para

cada emoção diminui de maneira significativa. Todavia, este resultado é esperado, uma vez que o classificador possui uma gama maior de categorias nas quais pode classificar uma determinada entrada. Neste caso, uma classificação por chute corresponderia a uma taxa de acerto de 12,5% e observamos resultados que superam o chute em 211% a 359,6%.

4 CONSIDERAÇÕES FINAIS

4.1 Trabalhos futuros

Os trabalhos futuros poderiam seguir dois diferentes caminhos: aprimorar a predição ou desenvolver uma aplicação com a finalidade de disponibilizar a ferramenta ao o público.

4.1.1 Aprimorar a predição

Para aprimorar a predição e aumentar sua acuidade novas abordagens devem ser consideradas.

Uma das abordagens a serem consideradas é explorar outros algoritmos e classificadores. Nesse caso, seria importante considerar pesquisas em torno de algoritmos mais complexos que o *Naive Bayes* ou que apresentem características de predição distintas da dele.

Outra abordagem que deve ser levada em consideração é a utilização de redes neurais para realizar a predição de emoções. As redes neurais podem ser uma alternativa aos métodos estatísticos tradicionais, porém para obter um resultado satisfatório seria necessário trocar a base de dados ou adicionar mais dados, devido ao fato de redes neurais necessitarem de mais dados para atingir uma melhor predição.

4.1.2 Aplicação como produto

Jeb Blount, CEO da Sales Gravy, uma gigante plataforma online de vendas, diz que "O que separa vendedores que atingem performances excelentes daqueles que performam dentro da média é a habilidade do primeiro grupo de casar seu entendimento sobre processos de vendas com a aspectos inteligencia emocional". Isto é, vendedores capazes de identificar em que momento, dentro do processo de vendas, devem buscar conduzir o possível cliente a uma determinada emoção, aplicando alguma técnica pré estabelecida

dadas as regras de negociação serão aqueles que atingirão resultados impressionantes de vendas.

O entrave, é que o treinamento da inteligência emocional de vendedores não é nada simples. A subjetividade no julgamento de como conduzir a persuasão do possível cliente estará sempre presente em processos de vendas. Porém, mesmo diante de diversas dificuldades, empresas de *call centers* e *telemarketing* buscam aprimorar, tanto quanto o entendimento dos processos de vendas, a inteligência emocional de seus funcionários. Para tanto, são fornecidos cursos, *workshops* e até uma técnica muito comum de fornecer *scripts* para que os vendedores sigam passos pré definidos quando se deparam com situações em que suas atitudes possam influenciar o resultado da negociação.

Nota-se dentro deste universo, que a tecnologia pouco tem contribuído para alavancar o potencial da inteligência emocional de vendedores. Os processos utilizados hoje em dia são manuais e bem semelhantes aos utilizados desde o princípio dessas empresas.

Em contraste, quando analisamos o avanço da indústria de anúncios em mídias digitais [7], percebemos um crescimento gigantesco na última década da importância das emoções para o direcionamento de anúncios à públicos específicos. O *Facebook*, por exemplo, utiliza-se das reações emocionais através dos botões Curtir, Tristeza, Engraçado, Raiva, Espanto disponíveis para os usuários nas publicações da rede social a fim de determinar que tipo de conteúdo midiático exibir ao usuário futuramente. Assim, dependendo de como um grupo de usuários reage a um tipo de anúncio, o anunciante pode optar por alterar o público-alvo ou a mensagem exibida àquele tipo de usuário.

Dessa forma, o uso de um classificador emocional pode ser de extremo valor para ser um aliado de seus funcionários no processo de vendas. Uma interface visual para os agentes de vendas, durante uma negociação por telefone, onde é exibida a emoção que o cliente está sentindo no momento e, dadas as regras de negócio pré estabelecidas, fornece uma lista de possíveis ações a serem tomadas pelo agente a fim de conduzir a negociação de forma ótima, baseando-se nas emoções presentes no diálogo e no histórico de negociações da empresa pode vir a transformar de forma extremamente positiva a maneira como as empresas estão se relacionando com seus clientes por telefone.

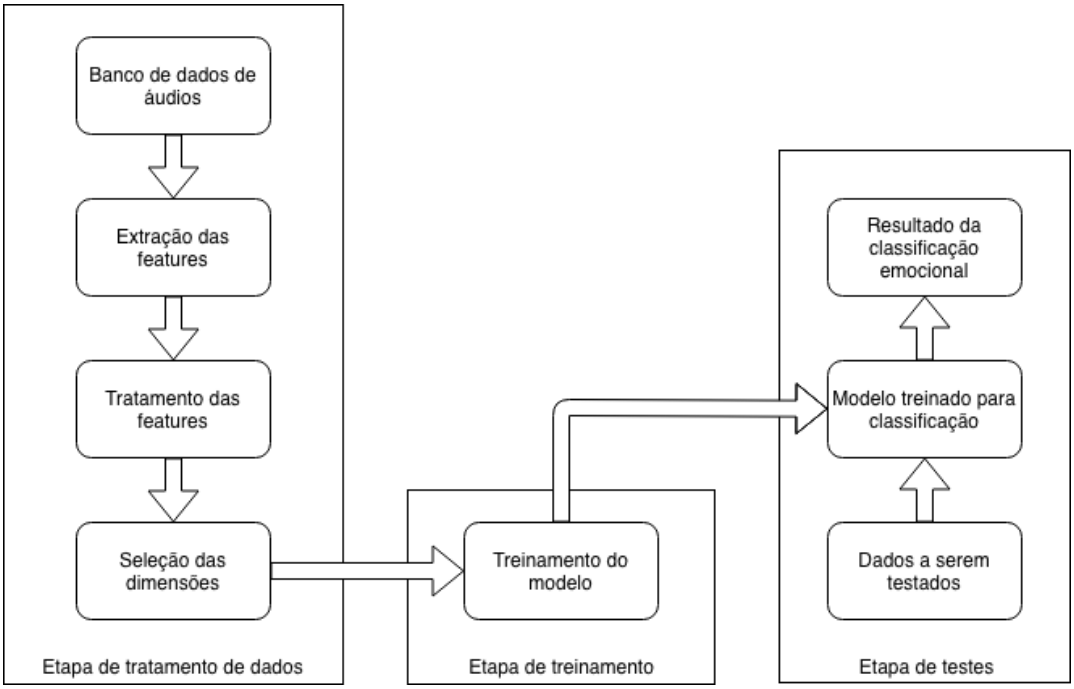
A classificação emocional a partir da voz humana pode ser grande aliada às ferramentas de reconhecimento de fala em aplicações com foco na interface humano-computador. Conforme novas interfaces são desenvolvidas, tendo como foco o reconhecimento da fala humana, se torna crucial para determinadas aplicações que ocorra a humanização do relacionamento entre o homem e a máquina, como é o caso de robôs babás, enfermeiras,

garçons, entre outros. Além de interfaces como assistentes virtuais em celulares e computadores. As possibilidades em se atrelar o reconhecimento emocional nestas ferramentas que utilizam este tipo de interface para embasar a tomada de decisões da máquina são enormes.

5 ANEXOS

5.1 Arquitetura de software

Como apresentado na seção de metodologia, o a arquitetura aplicada para desenvolver o nosso reconhecedor segue o seguinte modelo:



Nas sub-sessões abaixo, serão explicadas melhor as etapas de tratamento dos dados, que são as que apresentam maior complexidade, as demais etapas são procedimentos padrões que já foram explicados anteriormente.

5.1.1 Banco de dados de áudios

Os áudios de treino codificados no formato *.wav* foram salvos em uma única pasta, obedecendo a seguinte nomenclatura:

<Tipo>-<Entonação>-<Emoção>-<Intensidade>-<Frase>-<Repetição>-<Ator>

Cada código apresentado acima é composto por 2 dígitos e indica uma característica

do áudio.

O *Tipo* representa o formato do áudio, a base utilizada possuía as gravações em vídeo e áudio, como neste estudo foi usado apenas áudio, este código era sempre 03.

A *Entonação* se refere à entonação que áudio foi gravado. Na base de dados utilizada possuía dois tipos de entonação, a cantada e a falada. Como nessa pesquisar era necessário apenas a fala o código era sempre 01.

A *Emoção* refere-se a emoção expressa no áudio. A base utilizada contemplava no total 8 emoções, sendo: 01 - Neutro, 02 - Calma, 03 - Felicidade, 04 - Tristeza, 05 - Raiva, 06 - Medo, 07 - Nojo, 08 - Surpreso.

A *Intensidade* representa a intensidade com que a respectiva emoção é expressada, sendo 01 - intensidade normal e 02 - intensidade elevada.

A *Frase* refere-se a frase dita no áudio gravado, sendo 01 - a frase *"Kids are talking by the door"* e 02 - a frase *"Dogs are sitting by the door"*.

A *Repetição* se refere ao número da repetição da frase gravada, sendo que 01 representa a primeira repetição e 02 a segunda repetição.

Por fim o *Ator* se refere a qual ator gravou o áudio. No total existem 24 atores, sendo 12 homens e 12 mulheres. No código os números ímpares referem-se aos atores e os números pares referem-se às atrizes.

5.1.2 Extração de *features*

A extração de *features* foi realizado por meio da biblioteca de *Python* chamada *pyAudioAnalysis*. Para utilizar essa biblioteca foi desenvolvido uma rotina em *bash* que automatizava o processo de extração das *features*. Esse processo, criava como resultado uma estrutura de pastas da seguinte forma:

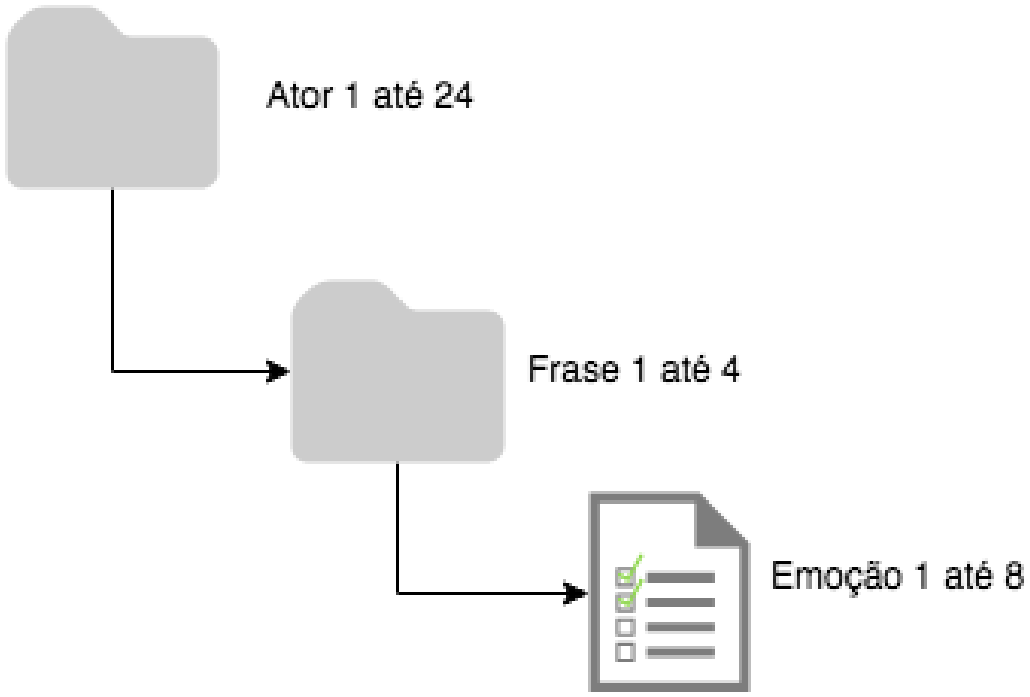


Figura 4: Estrutura de pastas

5.1.3 Tratamento das *features*

Neste módulo foi realizado o tratamento das features para o posterior treino do modelo. Basicamente este módulo agrupa todos os áudios em uma matriz e reduz cada *feature* em um único valor através de diversas funções pré-definidas, são elas: somatória, média, mediana, desvio padrão, primeira frequência natural, segunda frequência natural e terceira frequência natural. Cada função cria uma nova coluna na matriz contendo o valor calculado para o áudio em questão, onde cada linha da matriz é um áudio diferente e as colunas os valores calculados a partir das features (e.g. *Somatória da Energia, Média do Centro Espectral*).

Com as features de cada áudio reduzidas a um valor, aplica-se uma normalização trazendo a média para 0 e a variância para 1 em cima de cada coluna de features. Isso é extremamente importante para a posterior aplicação da ACP. Sem a normalização foi obtido da ACP que 92% da variância estava contida em apenas 1 componente, enquanto que com a normalização foi necessário usar 51 componentes para explicar 80% da variância.

Isso ocorre pois sem a normalização os dados podem estar em escalas diferentes e isso pode dar a impressão de uma maior variância onde não há ao mesmo tempo que suprime variância de dados em escalas menores.

5.1.4 Seleção das dimensões

Recebendo como entrada os dados devidamente tratados, é realizada a ACP, que retorna uma lista de novos componentes em ordem de peso na explicação da variância. A partir disso são escolhidos os k primeiros componentes tal que k componentes resultam em uma explicação da variância maior que uma percentagem pré-definida.

Como apontado na seção de observação de resultados, a percentagem de variância ótima encontrada foi 80% resultando em 51 componentes, como podemos observar no gráfico a seguir:

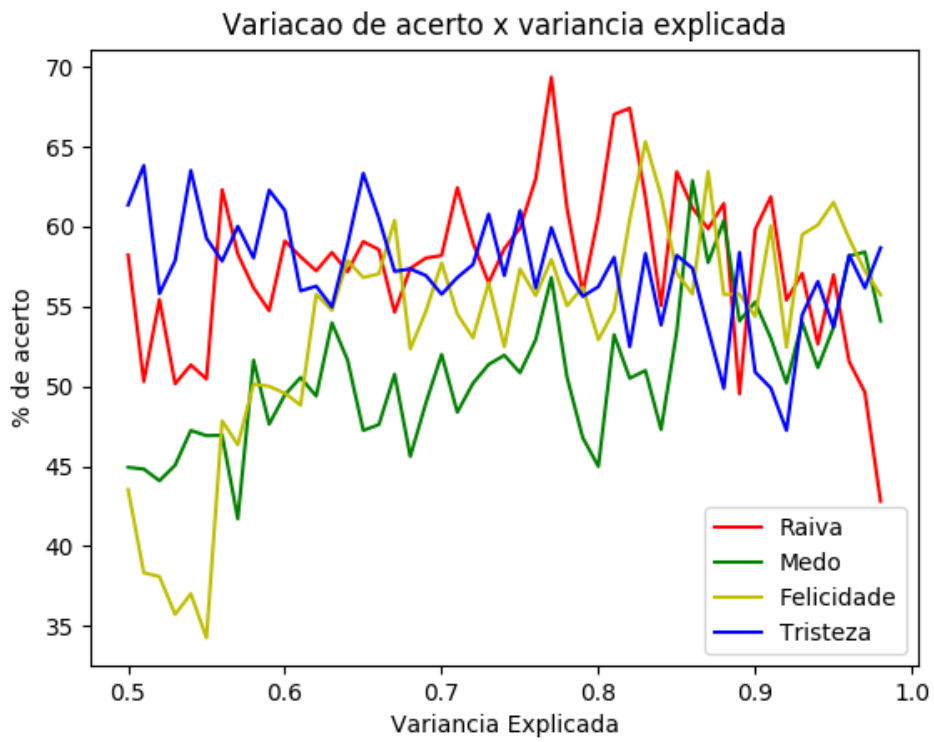


Figura 5: Correlação taxa de acerto pela variância explicada

5.2 Resultado da classificação emocional

Afim de analisar a influência que algumas formas de medidas, como: o desvio padrão, a média, a mediana, a soma e a frequência natural, realizam na acurácia do modelo foi desenvolvido uma série de testes.

Os testes resumiram-se em realizar uma rodada de predição na razão treino/teste de 80/20 com uma taxa variância de 80% e com uma variação na taxa de erro utilizando

apenas uma dessas formas de medidas para avaliar individualmente sua influência.

No total foi realizado cinco testes que produziram os seguintes resultados:

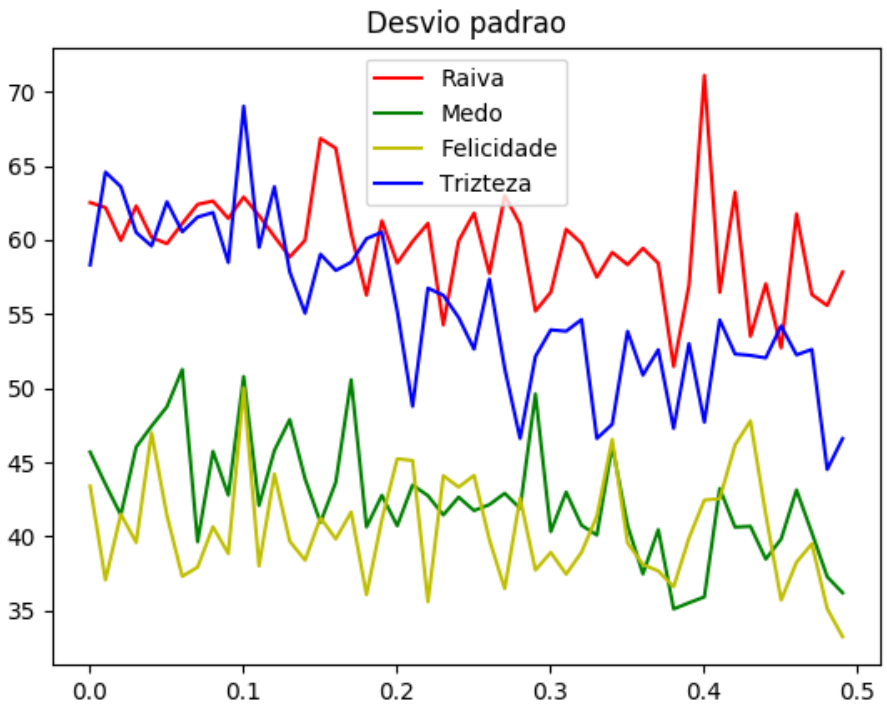


Figura 6: Taxa de acerto utilizando o desvio padrão

A imagem acima deixa claro como o desvio padrão privilegia as emoções raiva (em vermelho) e a tristeza (em azul) em detrimento do medo (em verde) e felicidade (em amarelo).

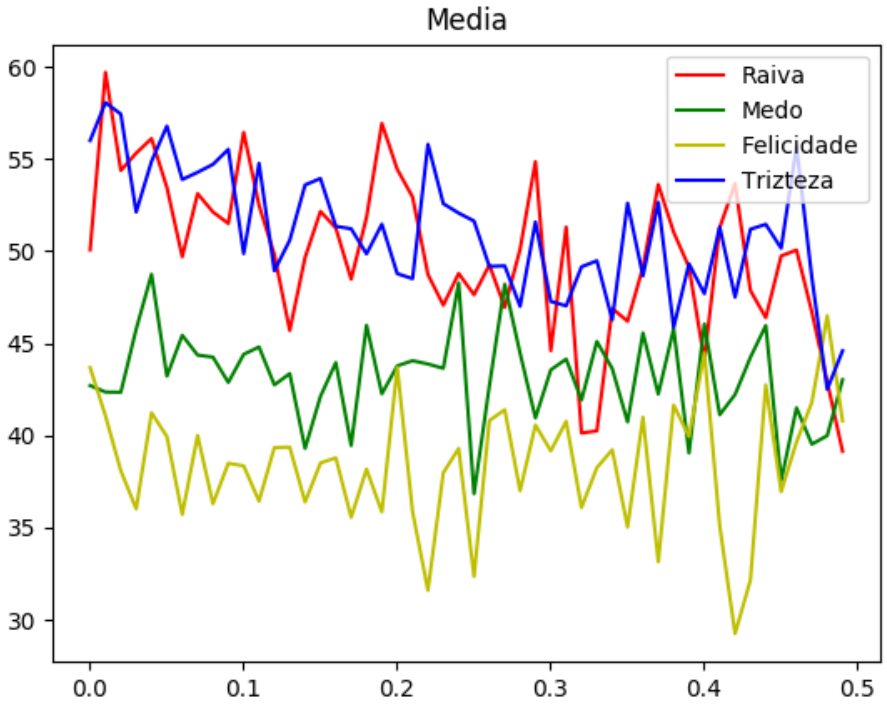


Figura 7: Taxa de acerto utilizando a média

A imagem acima mostra que igualmente ao desvio padrão a média, apesar de uma maneira mais suave, também privilegia as emoções raiva (em vermelho) e a tristeza (em azul) em detrimento do medo (em verde) e felicidade (em amarelo).

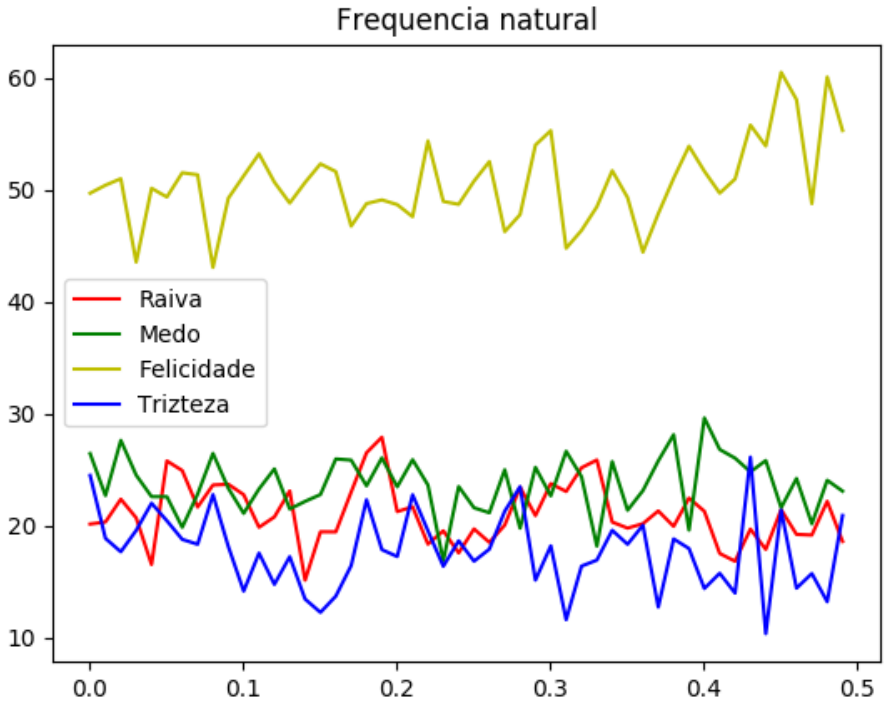


Figura 8: Taxa de acerto utilizando a frequência natural

A imagem acima mostra como a utilização da frequência natural discrimina a emoção de felicidade (em amarelo) destoando de maneira muito significativas das demais emoções. Esse um exemplo nítido de como uma forma de medida pode influenciar na categorização das emoções.

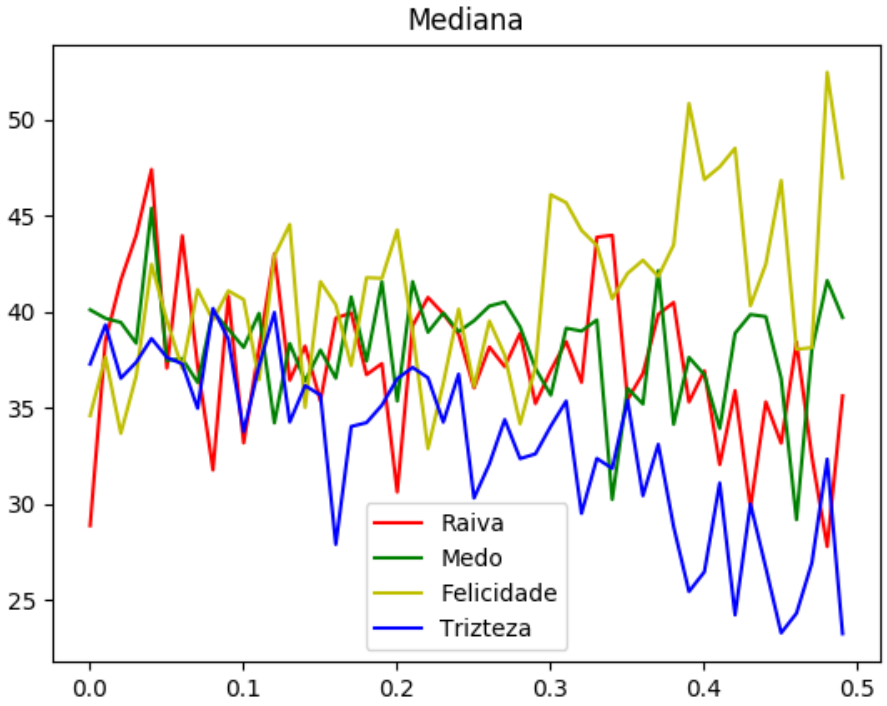


Figura 9: Taxa de acerto utilizando a mediana

Utilizando a mediana pode-se notar que não há discriminação relevante entre as emoções, observando o gráfico pode-se notar uma faixa padrão de acerto para todas as emoções.

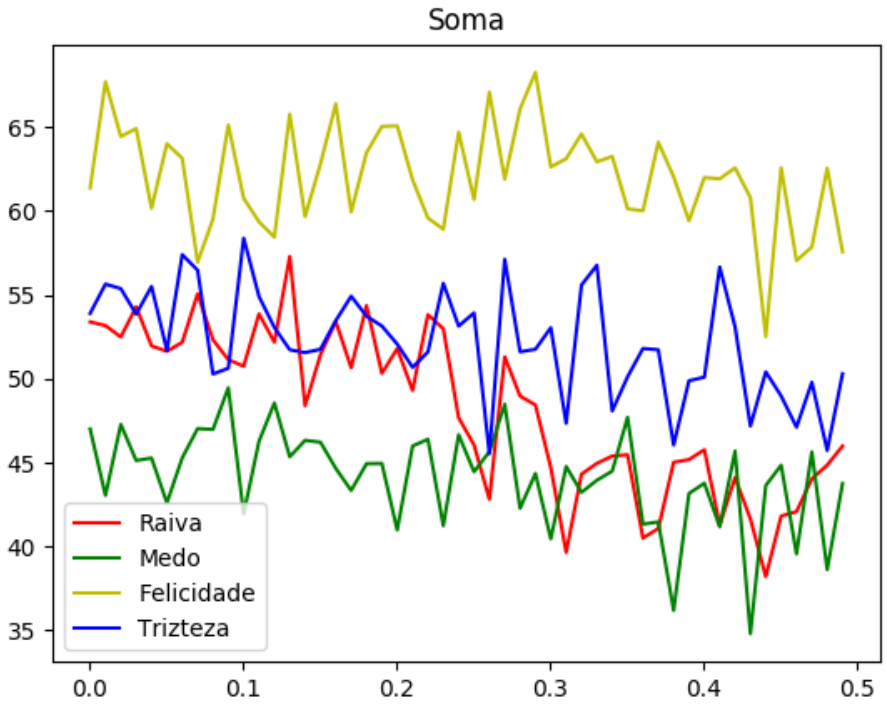


Figura 10: Taxa de acerto utilizando a soma

A soma por sua vez, convergiu para um resultado interessante. Ela conseguiu discriminar as emoções em três faixas de acerto. A primeira faixa entre 65 e 60%, a segunda faixa de 60 a 50% e a terceira faixa de 50 a 40%.

Tendo em posse esses resultados, pode-se chegar a conclusão que a variação da taxa de erro não ofereceu nenhuma melhora na acuidade do modelo. Na prática, a variação da taxa de erro manteve a acuidade constante, em alguns casos, o aumento da taxa de erro diminuiu a acurácia do modelo.

REFERÊNCIAS

[1] Karl Pearson, 1901 K. Pearson, "On lines and planes of closest fit to systems of points in space", The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series, 2, pp. 559-572 (1901) (1857-1936)

[2] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walther F. Sendlmeier and Benjamin Weiss, "A Database of German Emotional Speech", Proc. Interspeech (2005)

[3] Livingstone SR, Russo FA (2018), "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English", PLoS ONE 13(5): e0196391

[4] Giannakopoulos T (2015) pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. PLoS ONE 10(12): e0144610

[5] Schwaber, Ken. "Scrum development process." Business Object Design and Implementation. Springer London, 1997.

[6] Cockburn, Alistair, and Jim Highsmith. "Agile software development: The people factor." Computer 34.11 (2001)

[7] Webber, Richard. (2013). The evolution of direct, data and digital marketing. Direct. 14. 10.1057/dddmp.2013.20.