

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Eduardo Costa de Paiva

**Reconhecimento de emoção através da voz
para integração em uma aplicação web**

Uberlândia, Brasil

2017

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Eduardo Costa de Paiva

**Reconhecimento de emoção através da voz para
integração em uma aplicação web**

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, Minas Gerais, como
requisito exigido parcial à obtenção do grau
de Bacharel em Ciência da Computação.

Orientador: Márcia Aparecida Fernandes

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2017

Eduardo Costa de Paiva

Reconhecimento de emoção através da voz para integração em uma aplicação web

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, Minas Gerais, como
requisito exigido parcial à obtenção do grau
de Bacharel em Ciência da Computação.

Trabalho aprovado. Uberlândia, Brasil, 01 de Agosto de 2017:

Márcia Aparecida Fernandes
Orientador

Cláudio Camargo Rodrigues

Maria Adriana Vidigal de Lima

Uberlândia, Brasil
2017

A todos professores que contribuíram para minha formação.

A minha mãe, Divina, que sempre apoiou minhas decisões e ao meu pai, Sérgio, onde quer que ele esteja.

Agradecimentos

Agradeço a minha família e amigos por todo apoio e carinho. A minha orientadora, Márcia Fernandes, pela paciência, dedicação e ensinamentos não só neste trabalho, mas durante minha vida acadêmica. Aos doutorandos Fabiano e Luciane pelo incentivo e grande ajuda com o fornecimento de material que possibilitaram que eu realizasse este trabalho. Agradeço à minha namorada Gabriela, cujo companheirismo foi primordial em todas as dificuldades encontradas durante o curso.

*“Nunca se esqueça quem você é, o resto do mundo não esquecerá.
Faça disso sua força. Assim, não poderá ser nunca a sua fraqueza.
Arme-se com esta lembrança, e ela nunca poderá ser usada para lhe magoar.”*

George R. R. Martin

Resumo

O ensino nunca foi algo imutável, com o passar dos anos professores das mais diversas localidades buscam aprimorar a maneira com que o conhecimento é transmitido aos alunos, isso ocorre devido ao fato de que os estudantes aprendem de maneira diferente, visto que cada indivíduo possui características distintas. Dessa forma, com o avanço da tecnologia, várias alternativas surgiram e uma de grande importância é o reconhecimento de emoções através da voz, pois caso um sistema seja capaz de extrair informações da voz a ponto de saber a emoção que uma pessoa está sentindo, tais informações poderiam ser usadas para mudar o comportamento de uma plataforma de ensino, de tal forma que esta se adeque às necessidades do usuário. O objetivo deste trabalho é a criação de uma aplicação web capaz de classificar emoções através da voz que posteriormente poderá ser utilizada para aprimorar uma plataforma de ensino já existente, como por exemplo a plataforma *Moodle*. Foram utilizadas duas bases de dados: uma em Alemão *Berlin Database of Emotional Speech (EmoDB)* e a outra em Inglês *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. A primeira contém as seguintes emoções: Raiva, Tédio, Nojo, Medo, Felicidade, Tristeza e Neutra; enquanto a segunda é composta por: Raiva, Calma, Nojo, Medo, Felicidade, Tristeza, Neutra e Surpresa. Para os testes foram analisados dois algoritmos de classificação: J48 e SMO. Os resultados obtidos mostram que o SMO obteve melhor desempenho. Para a base RAVDESS o melhor resultado, utilizando SMO, apresentou acurácia de 79.30%, enquanto para a base EmoDB a acurácia foi de 87.47%. Por fim, foi desenvolvida a aplicação web capaz de gravar a voz e classificar a emoção do áudio gravado, que poderá ser facilmente integrada a um ambiente virtual de aprendizagem posteriormente.

Palavras-chave: Classificação de Emoções, Processamento de Voz, Algoritmos de Aprendizagem.

Lista de ilustrações

Figura 1 – Características mais utilizadas de um sinal de voz (SCHULLER et al., 2011)	17
Figura 2 – Diagrama do algoritmo MFCC	19
Figura 3 – Validação Cruzada 10-fold	22
Figura 4 – Quantidade de arquivos de cada emoção	23
Figura 5 – Construção do modelo utilizando a base de dados em inglês RAVDESS	24
Figura 6 – Construção do modelo com o algoritmo J48	24
Figura 7 – Estatísticas do treinamento com J48	25
Figura 8 – Estatísticas do treinamento com J48 e Validação Cruzada 10-fold	25
Figura 9 – Seleção do parâmetro emotion do arquivo emobase.arff através da interface	26
Figura 10 – Detalhes da acurácia do teste realizado utilizando J48 com todos os arquivos da base RAVDESS	27
Figura 11 – Plotagem ROC para a emoção raiva ($\text{Área} = 0.664$)	29
Figura 12 – Visão geral da aplicação.	39
Figura 13 – Shell Script responsável pela análise e classificação da emoção	40
Figura 14 – Tela inicial da aplicação	41
Figura 15 – Saída da execução do script	41
Figura 16 – Resultado da classificação na aplicação web	43

Lista de tabelas

Tabela 1 – Emoções e parâmetros da voz (Murray e Arnott, 1993)	19
Tabela 2 – J48-8-emoções-RAVDESS: Estatísticas gerais de desempenho	27
Tabela 3 – J48-8-emoções-RAVDESS - Detalhes de acurácia para cada uma das classes	28
Tabela 4 – J48-8-emoções-RAVDESS - Matriz de confusão	29
Tabela 5 – SMO-8-emoções-RAVDESS: Estatísticas gerais de desempenho	30
Tabela 6 – SMO-8-emoções-RAVDESS - Detalhes de acurácia para cada uma das classes	30
Tabela 7 – SMO-8-emoções-RAVDESS - Matriz de confusão	30
Tabela 8 – J48-4-emoções-RAVDESS: Estatísticas gerais de desempenho	31
Tabela 9 – J48-4-emoções-RAVDESS - Detalhes de acurácia para cada uma das classes	31
Tabela 10 – J48-4-emoções-RAVDESS - Matriz de confusão	32
Tabela 11 – J48-4-emoções-50-arquivos-RAVDESS: Estatísticas gerais de desempenho	32
Tabela 12 – J48-4-emoções-50-arquivos-RAVDESS - Detalhes de acurácia para cada uma das classes	32
Tabela 13 – J48-4-emoções-50-arquivos-RAVDESS - Matriz de confusão	33
Tabela 14 – SMO-4-emoções-RAVDESS: Estatísticas gerais de desempenho	33
Tabela 15 – SMO-4-emoções-RAVDESS - Detalhes de acurácia para cada uma das classes	33
Tabela 16 – SMO-4-emoções-RAVDESS - Matriz de confusão	34
Tabela 17 – SMO-4-emoções-50arquivos-RAVDESS: Estatísticas gerais de desempenho	34
Tabela 18 – SMO-4-emoções-50-arquivos-RAVDESS - Detalhes de acurácia para cada uma das classes	34
Tabela 19 – SMO-4-emoções-50-arquivos-RAVDESS - Matriz de confusão	35
Tabela 20 – J48-7-emoções-EmoDB: Estatísticas gerais de desempenho	35
Tabela 21 – J48-7-emoções-EMODB - Detalhes de acurácia para cada uma das classes	36
Tabela 22 – J48-7-emoções-EmoDB - Matriz de confusão	36
Tabela 23 – SMO-7-emoções-EmoDB: Estatísticas gerais de desempenho	37
Tabela 24 – SMO-7-emoções-EMODB - Detalhes de acurácia para cada uma das classes	37
Tabela 25 – SMO-7-emoções-EmoDB - Matriz de confusão	37

Lista de abreviaturas e siglas

API	Application Programming Interface
MFCC	Mel-Frequency Cepstral Coefficients
AJAX	Asynchronous JavaScript And XML
SMO	Sequential Minimal Optimization
QP	Quadratic Problem
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
ROC	Receiver Operating Characteristic
VP	Verdadeiro Positivo
FP	Falso Positivo

Sumário

1	INTRODUÇÃO	12
1.1	Objetivos	13
1.1.1	Objetivo Geral	13
1.1.2	Objetivos específicos	13
1.2	Justificativas	13
1.3	Organização do trabalho	14
1.4	Metodologia	14
2	ESTADO DA ARTE	16
3	FUNDAMENTAÇÃO TEÓRICA	18
3.1	Características de um sinal de voz	18
3.1.1	Ferramenta para extração de dados da voz - openSMILE	19
3.2	Bases de Dados de Voz	20
3.3	Algoritmos e <i>Softwares</i> para classificação	20
3.4	Método de Treinamento	21
4	EXPERIMENTOS E RESULTADOS	23
4.1	Construção do modelo de treinamento	23
4.2	Experimento 1: Algoritmo J48 e todas emoções da base RAVDESS	24
4.3	Experimento 2: Algoritmo SMO e todas emoções da base de dados RAVDESS	30
4.4	Experimento 3: Algoritmo J48 e apenas 4 emoções da base de dados RAVDESS	31
4.4.1	Experimento 3.1: Todos arquivos de cada classe de emoção	31
4.4.2	Experimento 3.2: 50 arquivos em cada classe de emoção	32
4.5	Experimento 4: Algoritmo SMO e apenas 4 emoções da base de dados RAVDESS	33
4.5.1	Experimento 4.1: Todos arquivos de cada classe de emoção	33
4.5.2	Experimento 4.2: 50 arquivos em cada classe de emoção	34
4.6	Experimento 5: Algoritmo J48 e todas emoções da base de dados EmoDB	35
4.7	Experimento 6: Algoritmo SMO com todas emoções da base de dados EmoDB	36
5	APLICAÇÃO WEB PARA CLASSIFICAÇÃO DE EMOÇÕES	39

5.1	Visão geral do projeto	39
5.2	Script para análise e classificação de emoções	39
5.3	Desenvolvimento da aplicação Web	40
5.3.1	Javascript	40
5.3.2	NodeJS	40
5.3.3	Biblioteca RecordRTC	40
6	CONCLUSÃO	44
6.1	Dificuldades encontradas	45
6.2	Trabalhos futuros	45
	REFERÊNCIAS	47

1 Introdução

A comunicação através da voz fez parte do processo evolutivo da humanidade, as primeiras tentativas de expressar sentimentos e opiniões foi através da fala. Os seres humanos se comunicam e interagem a todo momento variando suas emoções, são as emoções que nos guiam muitas das vezes. O que não acontece ao manusear um computador, pois só um lado da interação possui essa variação.

Cada vez mais a tecnologia faz com que nossas relações com as máquinas se assemelhem com as relações humanas. Porém a diferença entre os humanos e as máquinas são tremendas. Afinal um computador entende apenas instruções previamente programadas.

Com o avanço frequente de técnicas como Inteligência Artificial e Reconhecimento de Padrões, a interação humano-computador pôde se aproximar das relações entre humanos. Se o computador é capaz de extrair características que identificam a emoção de um usuário, poderá utilizá-las para providenciar uma experiência melhor para o indivíduo.

Para a construção de uma aplicação capaz de reconhecer emoções, se faz necessário saber o que é uma emoção de fato. Existe uma grande pesquisa que especifica detalhadamente a natureza desse estado, chamada de a "Tradição Psicológica" (*Psychological Tradition*) (COWIE et al., 2001, p. 35) concebida por grandes nomes de diferentes áreas como Filosofia (Rene Descartes), Biologia (Charles Darwin) e Psicologia (William James).

Para Descartes, algumas emoções básicas constituíam toda a vida emocional. Darwin introduziu a ideia de que emoções são inseparáveis de alguns determinados hábitos selecionados por meio da evolução, seus maiores exemplos foram: expressões faciais e movimentos corporais. William James focou sua pesquisa na relação entre as emoções e a excitação somática¹. (COWIE et al., 2001).

Uma problemática é que não existe exatamente um conjunto de emoções básicas. O psicólogo Paul Ekman (EKMAN, 1992) definiu um conjunto de seis emoções universais, dentre elas: Raiva, Nojo, Medo, Felicidade, Tristeza e Surpresa. Todas as demais emoções seriam variações ou uma mistura das demais. Robert Plutchik (PLUTCHIK, 1980) propôs um aprimoramento à ideia de Ekman, criando a roda das emoções, que contém 8 emoções básicas e as possíveis variações destas.

Como existem diferentes abordagens de quais emoções devem ser consideradas básicas ou não, deve-se fazer escolhas pragmáticas para o tipo de problema em questão. Neste trabalho foram analisadas duas base de dados, uma em alemão contendo as seguintes emoções: Raiva, Tédio, Nojo, Medo, Felicidade, Tristeza e Neutra. E uma base em

¹ Físico ou corporal, que tem em conta o aspecto físico do corpo humano

inglês que possui as emoções: Raiva, Calma, Nojo, Medo, Felicidade, Tristeza, Neutra e Surpresa.

A fim de executar o reconhecimento de emoções através da voz, deve-se extrair características deste sinal de áudio. Muito ainda se discute quais parâmetros devem ser utilizados, afinal não se sabe como nosso cérebro consegue discernir as emoções. Deve-se então criar uma aproximação através de modelos matemáticos. Muitos autores fazem uso de parâmetros como intensidade, frequência fundamental, assim como as propriedades espectrais da voz e coeficientes cepstrais.

Neste trabalho foi realizado um estudo sobre características relevantes que possam ser extraídas de um sinal de voz, para se construir uma aplicação capaz de reconhecer emoções através desses sinais.

1.1 Objetivos

1.1.1 Objetivo Geral

Reconhecer emoções através da voz em uma aplicação *web*. Não sendo necessária a gravação prévia de um arquivo de voz para que o sistema analise-o.

1.1.2 Objetivos específicos

Os objetivos específicos desse trabalho são:

- Estudo de características e ferramentas de reconhecimento de emoções através da voz.
- Reconhecer emoções através da voz utilizando uma ferramenta que tenha bom desempenho.
- Integrar reconhecedor de emoções à uma aplicação *web*.
- Identificar a viabilidade de integração do reconhecedor de emoções através da voz em ambientes virtuais de aprendizagem.

1.2 Justificativas

A percepção do aluno atual não é focada apenas em exercícios e avaliações, é preciso notar o comportamento do mesmo em diferentes situações do cotidiano de uma sala de aula, como suas emoções e seus movimentos faciais. Além disso, a maneira que as aulas são ministradas dificilmente agrada todos os alunos de uma determinada turma. Diante disso, é impraticável que um professor consiga desenvolver uma metodologia para

se adequar a cada aluno em específico. Portanto, fica a cargo de um sistema computacional verificar cada indivíduo em particular e buscar uma maneira de melhorar seu aprendizado.

A aplicação *web* desenvolvida neste trabalho poderá ser integrada a um sistema virtual de aprendizagem. Esse sistema trará grandes vantagens, além de ser um grande avanço para os estudantes em geral, uma pessoa com deficiência poderá ter um ensino totalmente adaptado para suas necessidades.

1.3 Organização do trabalho

Este trabalho possui a seguinte organização textual:

O Capítulo 2 descreve as características relevantes para análise de um sinal de voz e o estado da arte.

O Capítulo 3 descreve as bases de dados, *softwares* e método de treinamento, utilizados.

O Capítulo 4 apresenta os experimentos e resultados obtidos utilizando os classificadores propostos.

O Capítulo 5 apresenta o processo de desenvolvimento da aplicação, descrevendo linguagens e tecnologias utilizadas, bem como o uso do classificador de melhor desempenho obtido no capítulo anterior.

No Capítulo 6, por fim, serão apresentadas as conclusões e trabalhos futuros, bem como as dificuldades encontradas.

1.4 Metodologia

Neste trabalho foi feita uma pesquisa descritiva com uma abordagem quantitativa. Afinal, se faz necessário um grande levantamento bibliográfico das técnicas de classificação de emoção e das análises estatísticas desta classificação, já que não se pode garantir acurácia total neste tipo de problema. O intuito foi obter um reconhecedor de emoções através da voz com um bom desempenho, unificá-lo com uma aplicação *web* e verificar a viabilidade da integração e aprimoramento de um sistema virtual de aprendizagem.

Após o levantamento bibliográfico, algumas ferramentas *open source* foram estudadas para verificar a possibilidade da integração proposta. Com estas ferramentas, foram verificadas usabilidade, comunidade e documentação, optando pelo uso da ferramenta que se mostrasse melhor nessas características. Se nenhuma ferramenta obtivesse um bom desempenho, seria necessário o desenvolvimento de uma aplicação que atendesse as necessidades previamente estabelecidas. Seguem as etapas que foram executadas na seguinte ordem:

- 1) Revisão bibliográfica sobre ferramentas para reconhecimento de emoções através da voz.
- 2) Revisão bibliográfica sobre as principais técnicas e algoritmos utilizados para classificação de emoções.
- 3) Revisão bibliográfica sobre as abordagens aplicadas à classificação de emoções.
- 4) Seleção dos aspectos a serem observados na voz para classificação das emoções.
- 5) Estudo de algumas ferramentas de reconhecimento de emoções através da voz.
- 6) Estudo de classificadores já existentes para o tipo de dado em questão.
- 7) Testes utilizando algumas bases de dados disponíveis.
- 8) Verificação de desempenho das ferramentas escolhidas.
- 9) Verificação da viabilidade e do desempenho da integração com uma aplicação *web*.
- 10) Desenvolvimento de uma aplicação *web*, que faça a integração com a ferramenta escolhida.

2 Estado da Arte

Atualmente, os estudos referentes à extração de dados para classificação de emoções tendem a se focar no âmbito visual, mais especificamente em expressões faciais, afinal esta técnica se mostrou muito eficaz. Porém existem diversos trabalhos atuais que mostram que a análise de sinais de áudio possui uma acurácia considerável.

Por exemplo, o trabalho de R. Iriya ([IRIYA, 2014](#)) fez um comparativo dos métodos de classificação como: K-vizinhos mais próximos (KNN), Máquinas de Vetores de Suporte (SVM), Modelos de Misturas Gaussianas (GMM) e Modelos Ocultos de Markov (HMM).

No trabalho de Marc Escalona ([MENA, 2012](#)), são comparados os seguintes classificadores: Sequential Minimal Optimization (SMO), Naive Bayes e Logistic Model Tree (LMT), com o SMO sendo o melhor entre eles.

Roeland ([ROELAND, 2014](#)) utilizou ambas as técnicas de reconhecimento de emoções, fazendo uma análise separada entre reconhecimento de emoções através da face e através da voz, depois foram atribuídos pesos e uma análise foi feita para determinar a emoção mais relevante. Roeland analisou apenas o classificador J48, diferentemente do foco deste trabalho que comparou o algoritmo J48 com o algoritmo SMO.

Em alguns trabalhos, pesquisadores optaram por implementar seus próprios algoritmos para treinamento, mas muitos outros utilizaram a ferramenta openSMILE ([EYBEN et al., 2013](#))([ROELAND, 2014](#)) para facilitar a extração de características relevantes da voz, que são apresentadas na tabela da Figura 1.

No âmbito empresarial, uma ferramenta com uma relevância considerável é o *EMOSpeech* ([EMOSPEECH](#),), que se trata de um software de análise de fala, que permite *call centers* analisarem ligações gravadas. É baseado num modelo que reconhece diferentes emoções em um espectro de emoções acústicas.

Empath ([EMPATH](#),) é um programa desenvolvido pela *Smartmedical Corp.* Não ficava evidente como, mas o algoritmo identifica a emoção utilizando algumas propriedades da voz. Baseado em várias amostras, o programa foi capaz de detectar raiva, alegria, tristeza, tranquilidade e energia. Além disso, a empresa provê uma API para desenvolvedores.

Acoustics	Intonation (F0 or pitch modelling)	Deriving (raw LLD, deltas, regression coefficients, auto- and cross-correlation coefficients, cross-LLD, LDA, PCA, ...)	Filtering (smoothing, normalising, ...)	Chunking (absolute, relative, syntactic, semantic, emotional)	Extremes (min, max, range, ...)	Deriving (raw functionals, hierarchical, cross-functionals, cross-chunking, contextual, LDA, PCA, ...)	Filtering (smoothing, normalising, ...)
	Intensity (energy, Teager, ...)				Mean (arithmetic, absolute, ...)		
	Linear Prediction (LPCC, PLP, ...)				Percentiles (quartiles, ranges, ...)		
	Cepstral Coefficients (MFCC, ...)				Higher Moments (std. dev., kurtosis, ...)		
	Formants (amplitude, position, ...)				Peaks (number, distances, ...)		
	Spectrum (MFB, NMF, roll-off, ...)				Segments (number, duration, ...)		
	TF-Transformation (Wavelets, Gabor, ...)				Regression (coefficients, error, ...)		
	Harmonicity (HNR, spectral tilt, ...)				Spectral (DCT coefficients, ...)		
	Perturbation (jitter, shimmer, ...)				Temporal (durations, positions, ...)		
	Linguistics				Linguistics (phonemes, words, ...)		
Para-Linguistics (laughter, sighs, ...)		Look-Up (word lists, concepts, ...)					
Disfluencies (pauses, ...)		Statistical (salience, info gain, ...)					
Low-Level-Descriptors					Functionals		

Figura 1 – Características mais utilizadas de um sinal de voz (SCHULLER et al., 2011)

3 Fundamentação Teórica

Este capítulo descreve as ferramentas utilizadas para extração das características da voz, as bases de dados utilizadas para testes, bem como as técnicas e *softwares* utilizados para classificação de emoções através da voz.

3.1 Características de um sinal de voz

Como dito anteriormente, há algumas características que são consideradas relevantes para análise de um sinal de voz. Dentre estas estão o *pitch*, intensidade e os coeficientes mel-cepstrais.

A palavra *pitch* não possui uma tradução, mas pode ser entendida como a frequência medida em um sinal de áudio. Em muitos casos o *pitch* pode ser igual a frequência fundamental conhecida por f_0 , que é definida como sendo a menor frequência de um sinal de onda. Na música, a fundamental, é a mais baixa e a mais forte frequência componente da série harmônica de um som. Mas nem sempre essas duas propriedades serão iguais, a frequência é objetiva, um atributo científico que pode ser medido, já o *pitch* é subjetivo e diz respeito a como o ouvido humano é capaz de diferenciar sons graves e agudos. O *pitch* é relevante para a análise da voz, pois ele pode facilitar a separação das emoções. Onde um *pitch* alto com muitas variações indica emoções de alta excitação e um *pitch* baixo com poucas variações reflete emoções de baixa excitação.

A intensidade também é outro parâmetro importante, dependente de muitos fatores como: locutor, idioma, cultura, o tipo da expressão. Como o *pitch*, uma intensidade alta indica emoções de alta excitação.

Um estudo realizado por Murray e Arnott (MURRAY; ARNOTT, 1993) descreve alguns parâmetros estudados extensivamente na área de psicologia, dentre eles *pitch* e intensidade, a Tabela 1 adaptada de (COWIE et al., 2001, p.46)(IRIYA, 2014, p.42) resume esse estudo.

Tabela 1 – Emoções e parâmetros da voz (Murray e Arnott, 1993)

	Raiva	Felicidade	Tristeza	Medo	Nojo
Pitch	maior média, maior extensão, mudanças abruptas	maior média, maior extensão	menor média, menor extensão	maior média, maior extensão	menor média, menor extensão
Intensidade	mais alta	mais alta	mais baixa	normal	mais baixa
Velocidade da Fala	ligeiramente mais rápida	mais rápida ou mais lenta	mais lenta	mais rápida	mais lenta
Qualidade da voz	ofegante	ofegante, estridente	ressonante	irregular	resmungante

Outra característica muito utilizada no reconhecimento de emoções na voz são os coeficientes mel cepstrais (*Mel Frequency Cepstral Coefficients*) (LUTTER, 2014) (MEL... ,). Foram introduzidos por Davis e Mermelstein em 1980 (DAVIS; MERMELSTEIN, 1980) e se tornaram uma características do estado da arte desde então. Tais coeficientes possuem a forma apresentada na Figura 2.

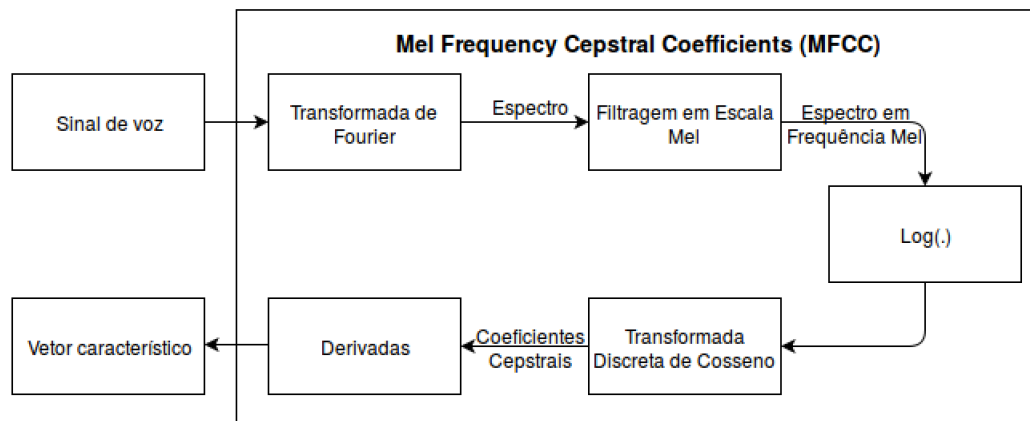


Figura 2 – Diagrama do algoritmo MFCC

(LUTTER, 2014)

3.1.1 Ferramenta para extração de dados da voz - openSMILE

The Munich open-Source Media Interpretation by Large feature-space Extraction (openSMILE) (EYBEN *et al.*, 2013) é uma ferramenta escrita em C++, que tem como principal objetivo a extração de características em processamento de sinais e em aplicações de aprendizado de máquina. O foco primário desta, é nos sinais de áudio. Neste trabalho, a ferramenta *openSMILE* foi utilizada para extrair os parâmetros mencionados na seção 3.1 para a classificação de emoções através da voz. Um guia para sua instalação pode ser encontrado no artigo *Opensmile Toolkit Quick Setup* (SHAH, 2015)

3.2 Bases de Dados de Voz

O *Berlin Database of Emotional Speech* (BURKHARDT et al., 2005) é uma base de áudios em alemão. Tais áudios são gravações de 10 atores (5 homens e 5 mulheres) simulando 10 sentenças (5 curtas e 5 longas) que poderiam ser usadas em conversas do cotidiano. As emoções simuladas são as 7 emoções básicas: alegria, nojo, tédio, medo, neutra, raiva, tristeza.

As gravações foram feitas com equipamentos de qualidade. O material gravado compreende cerca de 800 sentenças (dez atores * sete emoções * dez sentenças + algumas versões reserva).

O *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS) (LIVINGSTONE; PECK; RUSSO, 2012) é uma base de dados de vozes e músicas composta por 24 atores (12 homens e 12 mulheres) em inglês e está sob a licença *Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International*¹.

Essa base oferece áudios para as seguintes emoções: neutra, calma, felicidade, tristeza, raiva, medo, surpresa e nojo.

3.3 Algoritmos e Softwares para classificação

Weka (HALL et al., 2009) é um software *open-source* com diversos algoritmos de aprendizado de máquina para mineração de dados. Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamados através de um código Java. Neste trabalho, o Weka foi utilizado para verificar a partir de um conjunto de algoritmos de classificação, qual possuiria melhor eficácia para reconhecer as emoções através das características extraídas da voz.

Árvore de Decisão J48 foi um dos algoritmos utilizados para os experimentos devido ao fato de já ter sido aplicado em trabalho (ROELAND, 2014) semelhante a este. O algoritmo J48 é uma implementação *open-source* em Java do algoritmo C4.5 (QUINLAN, 1993) no software Weka. De acordo com Quinlan (QUINLAN, 1993), uma árvore de decisão pode ser utilizada para classificar um determinado caso da seguinte forma: Primeiramente se inicia na raiz, os testes vão ramificando as possibilidades da árvore até que uma folha seja encontrada. A árvore resultante do algoritmo C4.5 possui a seguinte estrutura:

- Uma folha, indicando uma classe.
- Um nó de decisão que especifica algum teste a ser executado em um atributo, com uma ramificação e uma subárvore para cada saída do teste.

¹ <http://creativecommons.org/licenses/by-nc-sa/4.0>

O C4.5 é um algoritmo guloso, isto é, executa sempre o melhor passo avaliado localmente, sem se preocupar se isso produzirá a melhor solução final. Além disso, este algoritmo segue a abordagem "dividir para conquistar", a partir da raiz são produzidas sub-árvores que dividem os múltiplos problemas de decisão de tal forma a simplificar o problema original.

O SMO (*Sequential Minimal Optimization*) (PLATT, 1998) é um algoritmo que foi concebido para ser uma solução ao grande problema de programação quadrática (QP) que acontece nos demais algoritmos SVM (*Support Vector Machines*). O funcionamento de uma Máquina de Vetores de Suporte (SVM) pode ser descrito basicamente pela divisão das classes através de uma linha de separação chamada *hiperplano*. Tal linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classes. Para solucionar o problema quadrático presente nas SVM o algoritmo SMO divide este problema em uma série de outros problemas QP de menor tamanho possível, desse modo os problemas são resolvidos de maneira analítica. O SMO possui complexidade de tempo linear para o tamanho do conjunto de treinamento. Nesse trabalho, esse algoritmo foi utilizado como comparativo ao J48.

3.4 Método de Treinamento

Inicialmente, foi feito um teste utilizando como treinamento todos os arquivos de áudio da base em alemão *Berlin Database of Emotional Speech* (BURKHARDT et al., 2005). Ao utilizar os arquivos de áudio em inglês da base RAVDESS (LIVINGSTONE; PECK; RUSSO, 2012) como entrada, os resultados não foram como esperados obtendo uma acurácia extremamente baixa, de tal forma que não houvesse a necessidade de documentar este teste. Fazendo o inverso, utilizando a base em inglês como treinamento e o alemão como entrada, os resultados continuam ruins. Por mais que a escrita das duas linguagens sejam semelhantes, a diferença da pronúncia e fonética entre elas gerou péssimos resultados. Dessa forma, foi escolhida uma outra abordagem que é o método de treinamento por Validação Cruzada e o foco em apenas um dos idiomas por vez.

A Validação Cruzada (KOHAVI et al., 1995, p.2-3) é uma técnica para avaliar a generalização de um conjunto de dados. Esta técnica é utilizada, principalmente, em configurações onde o objetivo é a predição e deseja-se saber quão preciso um modelo preditivo será.

Neste trabalho foi utilizada a validação 10-fold. Em uma validação cruzada k -fold, a amostra de dados original é dividida em k sub-amostras iguais, uma das k sub-amostras é escolhida para teste e as $k - 1$ restantes são utilizadas para treinamento, esse processo é repetido k vezes. A Figura 3 descreve esse processo.

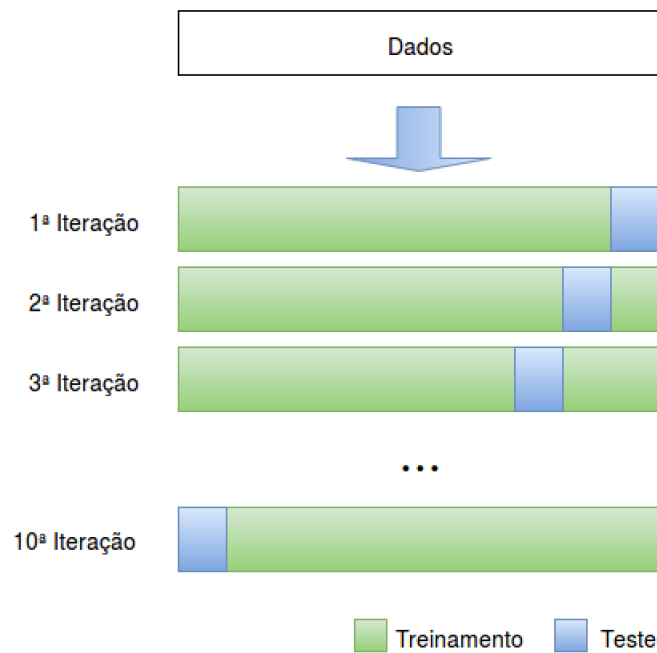


Figura 3 – Validação Cruzada 10-fold

Neste treinamento foram utilizadas ambas bases de áudio (RAVDESS e EmoDB), porém o foco foi na base que utiliza a língua inglesa (RAVDESS). A escolha de focar no uso da língua inglesa se dá por ser considerada uma língua universal e de grande popularidade. Porém, a base *RAVDESS* possui uma leve diferença em relação as emoções oferecidas. Enquanto a base em alemão oferece sete emoções, dentre elas: Raiva, Tédio, Nojo, Medo, Felicidade, Neutra e Tristeza; a base de dados em inglês oferece outras, sendo elas: Raiva, Calma, Nojo, Medo, Felicidade, Neutra, Tristeza e Surpresa. As emoções foram divididas em diretórios em que cada diretório representa uma emoção específica.

4 Experimentos e Resultados

Esta seção descreve os experimentos realizados, utilizando as duas bases de dados (EmoDB e RAVDESS) e os algoritmos de classificação Árvore de Decisão J48 e SMO.

4.1 Construção do modelo de treinamento

Para iniciar os experimentos, se faz necessária a construção do modelo de treinamento. Primeiramente, foram separados os arquivos de áudio de cada emoção com seus respectivos nomes. De forma que todos os arquivos da emoção raiva (*anger*) se encontrem na pasta *anger*, os arquivos da emoção tristeza (*sadness*) se encontrem na pasta *sadness* e, assim por diante. Tanto na base de dados RAVDESS, como no EmoDB, os autores descreveram como os nomes de arquivos são organizados e por isso foi possível separá-los em pastas. Os gráficos da Figura 4 ilustram a quantidade de arquivos nas duas bases de dados.

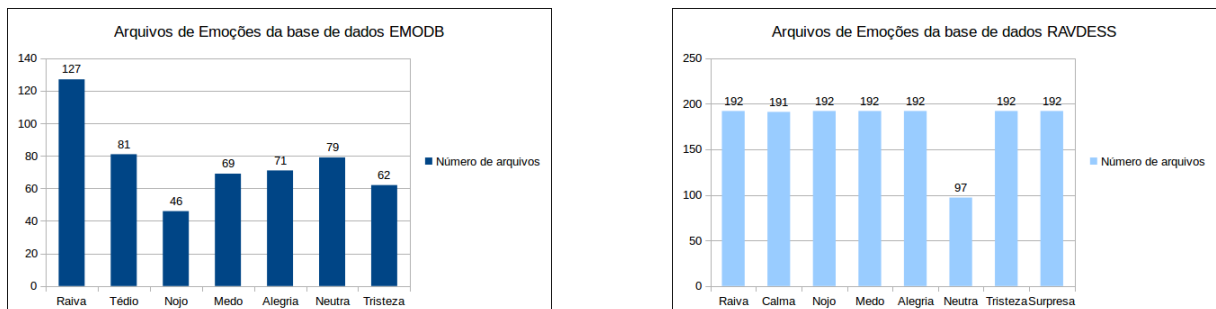


Figura 4 – Quantidade de arquivos de cada emoção

Após realizada a separação dos arquivos, o diretório raiz do *openSMILE* foi acessado e então direcionado para o diretório `/scripts/modeltrain`. O script escrito em *Perl* `makemodel.pl` foi executado, tendo como parâmetro a pasta onde se encontram todos os arquivos de áudio e o arquivo de configuração `emobase.conf`, disposto na pasta `/config`.

Finalizada a execução deste comando, um diretório `/work` foi gerado, contendo um arquivo denominado `emobase` e de extensão `ARFF`, sendo esta a extensão utilizada para arquivos de entrada do *software Weka*. Esse arquivo e o *Weka* foram utilizados para gerar o modelo de treinamento. Esse passo pode ser realizado diretamente pela linha de comando, conforme Figura 5 ou utilizando-se a interface gráfica do *software Weka*. A seção a seguir descreve essa etapa para o classificador J48 e a base RAVDESS. Para as

```

eduardo@eduardopc:~/Documentos/Faculdade/TCC/Projeto/openSMILE-2.1.0/scripts/modeltrain$ perl makemodel.pl ../../Integracao/BancoDeAudios-In
gles/Treinamento/
emobase.conf
-- Corpus mode --
Running feature extraction on corpus '../../Integracao/BancoDeAudios-Ingles/Treinamento/' ...
label: angry
label: calm
label: disgust
label: fear
label: happiness
label: neutral
label: sadness
label: surprised
angry
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-02.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-03.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-04.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-05.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-06.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-07.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-10.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-11.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-12.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-13.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-14.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-16.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-18.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-19.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-20.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-21.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-23.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-01-24.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-02.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-03.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-04.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-06.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-07.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-08.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-10.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-11.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-13.wav... OK.
../../Integracao/BancoDeAudios-Ingles/Treinamento/angry/03-01-05-01-01-02-14.wav... OK.

```

Figura 5 – Construção do modelo utilizando a base de dados em inglês RAVDESS

demais seções de experimento, esse mesmo comando foi executado, fazendo as devidas alterações necessárias.

4.2 Experimento 1: Algoritmo J48 e todas emoções da base RAVDESS

Esta seção detalha o processo para realizar o treinamento tanto via linha de comando como pela interface do *software Weka*, bem como o primeiro experimento que utilizou o algoritmo J48 e todas as emoções da base RAVDESS.

```

eduardo@eduardopc: ~/Documentos/Faculdade/TCC/Testes/RAVDESS-AllFiles
eduardo@eduardopc:~/Documentos/Faculdade/TCC/Testes/RAVDESS-AllFiles$ java -classpath weka.jar weka.classifiers.trees.J48 -t emobase.arff -d emobase.model

```

Figura 6 – Construção do modelo com o algoritmo J48

A Figura 6 mostra a execução do treinamento com classificador J48, salvando o resultado deste em um arquivo modelo `emobase.model`, que foi usado posteriormente na fase de classificação, onde o *openSMILE* extraiu as características relevantes em um arquivo `output.arff`, que foi utilizado pelo Weka para classificação. Para todos os experimentos, esse comando foi executado, alterando o classificador e o arquivo `emobase.arff`

A Figura 7 mostra as estatísticas do treinamento com J48, através da linha de comando e a Figura 8 mostra as estatísticas referentes a este treinamento, porém considerando validação cruzada 10-Fold.

```

Number of Leaves :      239

Size of the tree :      477

Time taken to build model: 16.21 seconds
Time taken to test model on training data: 0.61 seconds

=== Error on training data ===

Correctly Classified Instances      1346      93.4722 %
Incorrectly Classified Instances    94      6.5278 %
Kappa statistic                    0.925
Mean absolute error                 0.0235
Root mean squared error             0.1084
Relative absolute error             10.7846 %
Root relative squared error         32.8402 %
Total Number of Instances          1440

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  <-- classified as
184  0  1  0  3  1  0  3 | a = angry
  0 183  2  2  2  0  1  1 | b = calm
  6  2 176  1  1  1  4  1 | c = disgust
  2  3  0 183  1  0  2  1 | d = fear
  2  2  3  2 180  0  1  2 | e = happiness
  5  0  2  0  0  87  2  1 | f = neutral
  4  3  2  3  3  1 174  2 | g = sadness
  1  1  2  3  2  0  4 179 | h = surprised

```

Figura 7 – Estatísticas do treinamento com J48

```

=== Stratified cross-validation ===

Correctly Classified Instances      503      34.9306 %
Incorrectly Classified Instances    937      65.0694 %
Kappa statistic                    0.2527
Mean absolute error                 0.1652
Root mean squared error             0.3895
Relative absolute error             75.8244 %
Root relative squared error         118.0289 %
Total Number of Instances          1440

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  <-- classified as
75  6 23 21 31  2 18 16 | a = angry
  7 84 25  9  9 18 34  5 | b = calm
23 12 67 22 16  7 23 22 | c = disgust
23  7 17 76 29  2 20 18 | d = fear
33  1 21 24 61 10 18 24 | e = happiness
  5 15 12  7  8 21 20  9 | f = neutral
15 29 25 26 14 21 48 14 | g = sadness
20 10 15 26 26  6 18 71 | h = surprised

```

Figura 8 – Estatísticas do treinamento com J48 e Validação Cruzada 10-fold

As Figuras 9 e 10 mostram o mesmo experimento realizado acima, porém utilizando a interface *Weka Explorer*. Através da interface, é possível ter uma visão melhor dos resultados. A mesma construção do modelo descrito na Figura 5 foi executada, porém utilizando a interface. Após abrir o Weka e selecionar o arquivo no diretório desejado, selecionou-se o parâmetro *emotion* para realizar o treinamento.

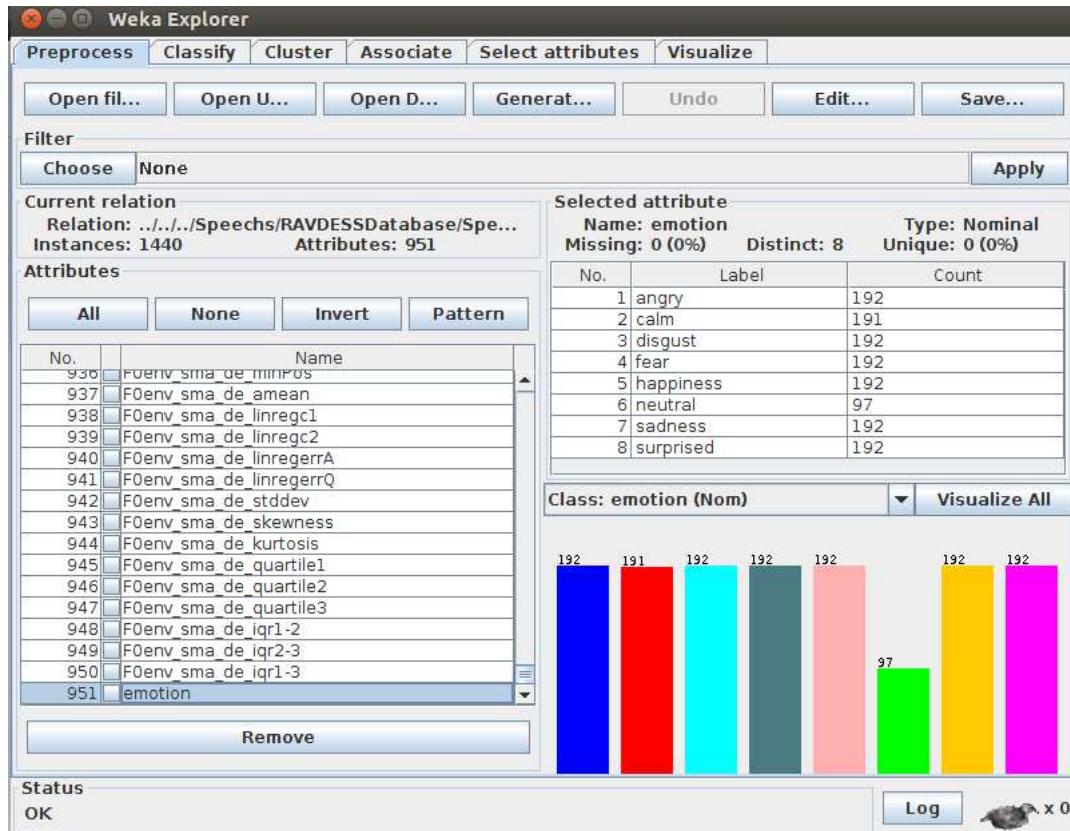


Figura 9 – Seleção do parâmetro *emotion* do arquivo *emobase.arff* através da interface

No próximo passo, a aba *Classify* e o classificador a ser utilizado foram selecionados.

Esse experimento foi feito com a árvore de decisão J48 e Validação Cruzada 10-fold. A validação 10-fold permitiu uma boa aproximação da acurácia média do classificador, conforme pode ser visto na Figura 10.

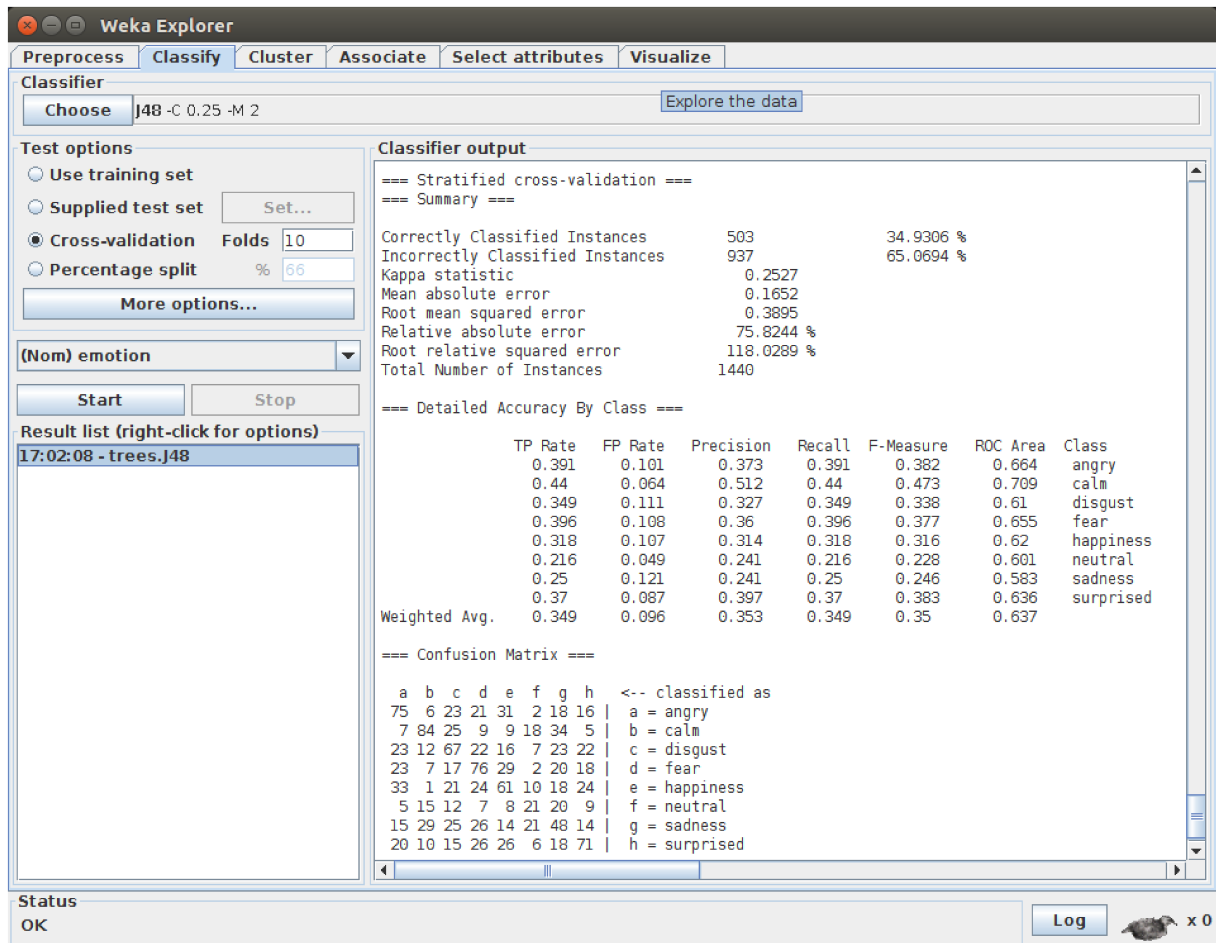


Figura 10 – Detalhes da acurácia do teste realizado utilizando J48 com todos os arquivos da base RAVDESS

Os resultados textuais da execução do treinamento no *Weka* foram divididos em três tabelas para obter uma visualização dos resultados. Assim, as estatísticas gerais de desempenho, os detalhes da acurácia e a matriz de confusão são apresentados, respectivamente, pelas Tabelas 2, 3 e 4, que dizem respeito ao treinamento envolvendo o algoritmo J48.

Tabela 2 – J48-8-emoções-RAVDESS: Estatísticas gerais de desempenho

Instâncias classificadas corretamente	503 (34.93%)
Instâncias classificadas incorretamente	937 (65.07%)
Estatística Kappa	0.2527
Erro absoluto médio	0.1652
Raiz do erro quadrático médio	0.3895
Erro absoluto relativo	75.82%
Raiz do erro médio relativo	118.03%
Número total de instâncias	1440

Tabela 3 – J48-8-emoções-RAVDESS - Detalhes de acurácia para cada uma das classes

Classe	Taxa VP	Taxa FP	Precisão	Revocação	F-Measure	ROC Area
Raiva (angry)	0.391	0.101	0.373	0.391	0.382	0.664
Calma (calm)	0.44	0.064	0.0512	0.44	0.473	0.708
Nojo (disgust)	0.349	0.111	0.327	0.349	0.338	0.61
Medo (fear)	0.396	0.108	0.36	0.396	0.377	0.655
Felicidade (happiness)	0.318	0.107	0.314	0.318	0.316	0.62
Neutra (neutral)	0.216	0.049	0.241	0.216	0.228	0.601
Tristeza (sadness)	0.25	0.121	0.241	0.25	0.246	0.583
Surpresa (surprised)	0.37	0.087	0.397	0.37	0.383	0.636

Cada um dos itens da Tabela 3 são interpretados da seguinte maneira:

1. A **Taxa VP** (Verdadeiro Positivo) significa o número de instâncias de uma classe que foram classificadas corretamente. Por exemplo, emoções classificadas com a classe raiva e que realmente pertencem ao conjunto de emoções dessa classe. A taxa de verdadeiros positivos pode ser calculada da seguinte maneira:

$$TaxaVP = \frac{VP}{VP + FN}$$

Esses valores podem ser observados de maneira muito simples na matriz de confusão. VP é o elemento da diagonal principal e FN são todos os demais elementos dessa mesma linha. Para a classe **Calma** na linha 2 da tabela 4 a posição da diagonal principal contém o valor 84, $VP = 84$. Somando os demais elementos desta linha temos que $FN = 107$

$$TaxaVP = \frac{84}{191} = 0.44$$

2. A **Taxa FP** (Falso Positivo) significa a quantidade de instâncias classificadas com uma determinada classe, mas que originalmente pertenciam a outra classe. Pode ser calculada da seguinte maneira: Somando todos os elementos da coluna da classe, com exceção do elemento da diagonal principal, dividido pela soma das linhas de todas as demais classes. Para a classe **Raiva** temos a seguinte taxa de falsos positivos:

$$TaxaFP = \frac{7 + 23 + 23 + 33 + 5 + 15 + 20}{(192 \times 5) + 191 + 97} = 0.101$$

3. A **Precisão** é a proporção das instâncias que realmente possuem uma determinada classe em relação a todas as instâncias que foram classificadas com essa classe. Pode ser calculada da seguinte maneira:

$$Precisão = \frac{VP}{VP + \Sigma FP}$$

4. A **Revocação** é o equivalente à taxa de verdadeiros positivos.

5. A **F-Measure** é uma média ponderada entre Precisão e Revocação. Geralmente, quanto mais próxima de 1, melhor o resultado. Calculada da seguinte forma:

$$F - Measure = \frac{2 \times (Precisão \times Revocação)}{Precisão + Revocação}$$

6. A **Área ROC** (Receiver Operating Characteristic)(DAVIS; GOADRICH, 2006) é a área sob a curva ROC para cada classe classificada. Essa curva se encontra em um plano cartesiano onde o eixo Y representa a taxa de verdadeiros positivos e o eixo X representa a taxa de falsos positivos. Um bom modelo se encontra o mais próximo possível do ponto (1,0). Modelos localizados na diagonal são modelos aleatórios, onde a *Taxa VP* = *Taxa FP*. A Figura 11 ilustra o gráfico da Área ROC para a emoção raiva nesse teste em específico, pode-se notar que este não se encontra próximo ao ponto (1,0), sendo um modelo não muito bom.

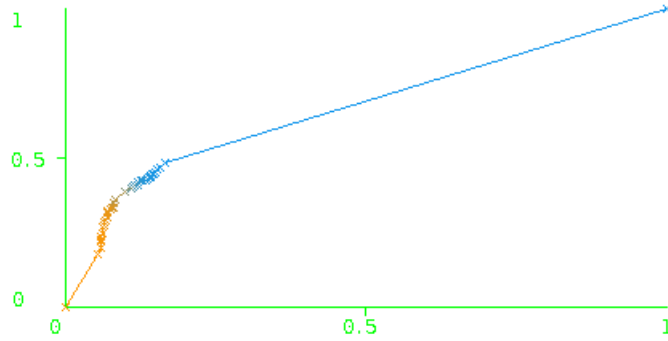


Figura 11 – Plotagem ROC para a emoção raiva (Área = 0.664)

Tabela 4 – J48-8-emoções-RAVDESS - Matriz de confusão

Raiva	Calma	Nojo	Medo	Felicidade	Neutra	Tristeza	Surpresa	
75	6	23	21	31	2	18	16	Raiva
7	84	25	9	9	18	34	5	Calma
23	12	67	22	16	7	23	22	Nojo
23	7	17	76	29	2	20	18	Medo
33	1	21	24	61	10	18	24	Felicidade
5	15	12	7	8	21	20	9	Neutra
15	29	25	26	14	21	48	14	Tristeza
20	10	15	26	26	6	18	71	Surpresa

Através dos dados das Tabelas 2, 3, 4 podemos notar que o classificador J48 não se comportou muito bem para esse conjunto de dados, classificando apenas 34,9306% das instâncias corretamente. A apresentação dos resultados dos demais experimentos segue o mesmo formato deste primeiro experimento. Sendo assim, a interpretação da segunda tabela de cada experimento é a mesma conforme descrição da Tabela 3.

4.3 Experimento 2: Algoritmo SMO e todas emoções da base de dados RAVDESS

Nesta seção o experimento realizado utiliza o algoritmo SMO e todas as emoções existentes na base de dados RAVDESS.

Tabela 5 – SMO-8-emoções-RAVDESS: Estatísticas gerais de desempenho

Instâncias classificadas corretamente	976 (67.78%)
Instâncias classificadas incorretamente	464 (32.22%)
Estatística Kappa	0.63
Erro absoluto médio	0.1932
Raiz do erro quadrático médio	0.3001
Erro absoluto relativo	88.72%
Raiz do erro médio relativo	90.93%
Número total de instâncias	1440

Tabela 6 – SMO-8-emoções-RAVDESS - Detalhes de acurácia para cada uma das classes

Classe	Taxa VP	Taxa FP	Precisão	Revocação	F-Measure	ROC Area
Raiva (angry)	0.75	0.047	0.709	0.75	0.729	0.917
Calma (calm)	0.806	0.054	0.697	0.806	0.748	0.965
Nojo (disgust)	0.703	0.052	0.675	0.703	0.689	0.922
Medo (fear)	0.719	0.044	0.715	0.719	0.717	0.929
Felicidade (happiness)	0.661	0.05	0.668	0.661	0.665	0.89
Neutra (neutral)	0.536	0.028	0.584	0.536	0.559	0.941
Tristeza (sadness)	0.479	0.065	0.532	0.479	0.504	0.821
Surpresa (surprised)	0.698	0.03	0.784	0.698	0.738	0.933

Tabela 7 – SMO-8-emoções-RAVDESS - Matriz de confusão

Raiva	Calma	Nojo	Medo	Felicidade	Neutra	Tristeza	Surpresa	
144	2	14	4	13	1	6	8	Raiva
0	154	3	1	1	12	20	0	Calma
13	5	135	6	7	4	17	5	Nojo
13	4	1	138	12	1	16	7	Medo
15	5	6	16	127	4	7	12	Felicidade
3	19	7	1	4	52	10	1	Neutra
6	32	22	16	11	9	92	4	Tristeza
9	0	12	11	15	6	5	134	Surpresa

Nesse experimento, em relação ao Experimento 1, houve um ganho considerável em várias estatísticas como a acurácia de 67.78% extraída da Tabela 5 e em outros parâmetros tais como Taxa VP, Precisão, F-Measure a Área-ROC extraídos da Tabela 6. Pode-se notar que todos esses parâmetros ficaram mais próximos de 1, que é o resultado ideal que se almeja alcançar. Na matriz de confusão representada pela Tabela 7 nota-se um aumento de instâncias na diagonal principal, afinal a taxa de verdadeiros positivos aumentou.

4.4 Experimento 3: Algoritmo J48 e apenas 4 emoções da base de dados RAVDESS

Este experimento foi dividido em duas etapas, a primeira etapa utilizou todos os arquivos de 4 emoções escolhidas e para a segunda, foram escolhidos 50 arquivos de cada emoção. O intuito foi aumentar a acurácia com a redução da quantidade de classes de emoções, bem como a quantidade de arquivos. Afinal existem emoções que possuem características semelhantes entre si, por exemplo emoções de baixa excitação como: tristeza, tédio e neutra podem ser confundidas. Dessa forma foram separadas apenas 4 emoções, sendo elas: **Raiva, Felicidade, Tristeza e Surpresa**, isto é duas emoções negativa (Raiva e Tristeza), uma consideravelmente neutra (Surpresa) e uma positiva (Felicidade). Para o objetivo futuro desse trabalho, que é a integração com sistema virtual de aprendizagem, essas emoções possuem grande relevância.

4.4.1 Experimento 3.1: Todos arquivos de cada classe de emoção

As Tabelas 8, 9 e 10 descrevem a primeira etapa do experimento. Note que se trata do mesmo classificador utilizado no Experimento 1 e que ao diminuir a quantidade de emoções houve um ganho na acurácia, porém não muito considerável.

Tabela 8 – J48-4-emoções-RAVDESS: Estatísticas gerais de desempenho

Instâncias classificadas corretamente	413 (53.78%)
Instâncias classificadas incorretamente	355 (46.22%)
Estatística Kappa	0.3837
Erro absoluto médio	0.2346
Raiz do erro quadrático médio	0.4689
Erro absoluto relativo	62.57%
Raiz do erro médio relativo	108.28%
Número total de instâncias	768

Tabela 9 – J48-4-emoções-RAVDESS - Detalhes de acurácia para cada uma das classes

Classe	Taxa VP	Taxa FP	Precisão	Revocação	F-Measure	ROC Area
Raiva (angry)	0.557	0.16	0.538	0.557	0.547	0.725
Felicidade (happiness)	0.516	0.215	0.444	0.516	0.477	0.647
Tristeza (sadness)	0.589	0.116	0.628	0.589	0.608	0.74
Surpresa (surprised)	0.49	0.125	0.566	0.49	0.525	0.678

Tabela 10 – J48-4-emoções-RAVDESS - Matriz de confusão

Raiva	Felicidade	Tristeza	Surpresa	
107	39	18	28	Raiva
44	99	23	26	Felicidade
14	47	113	18	Tristeza
34	38	26	94	Surpresa

Observa-se na Tabela 8 que houve um ganho na acurácia em relação ao Experimento 1, atingindo 53.78%. A Tabela 9 possui valores um pouco melhores em relação a Tabela 3. Na Tabela 10, vemos que a taxa de verdadeiros positivos é ruim, tomando como exemplo a linha da classe **Raiva**, onde 107 foram classificadas corretamente e 85 classificadas incorretamente. Portanto, mesmo diminuindo o número de instâncias, o ganho na acurácia foi relativamente baixo e se trata de um resultado ruim.

4.4.2 Experimento 3.2: 50 arquivos em cada classe de emoção

Como dito anteriormente, este experimento reduziu-se a quantidade de arquivos em cada emoção, na expectativa de que, ao diminuir o espaço do problema, se alcançasse maior precisão nos resultados.

As Tabelas 11, 12, 13 descrevem a segunda etapa do experimento, que utiliza o algoritmo J48 e 50 arquivos por emoção.

Tabela 11 – J48-4-emoções-50-arquivos-RAVDESS: Estatísticas gerais de desempenho

Instâncias classificadas corretamente	82 (41%)
Instâncias classificadas incorretamente	118 (59%)
Estatística Kappa	0.2133
Erro absoluto médio	0.2973
Raiz do erro quadrático médio	0.531
Erro absoluto relativo	79.28%
Raiz do erro médio relativo	122.63%
Número total de instâncias	200

Tabela 12 – J48-4-emoções-50-arquivos-RAVDESS - Detalhes de acurácia para cada uma das classes

Classe	Taxa VP	Taxa FP	Precisão	Revocação	F-Measure	ROC Area
Raiva (angry)	0.46	0.22	0.411	0.46	0.434	0.649
Felicidade (happiness)	0.38	0.167	0.432	0.38	0.404	0.567
Tristeza (sadness)	0.36	0.207	0.367	0.36	0.364	0.598
Surpresa (surprised)	0.44	0.193	0.431	0.44	0.436	0.619

Tabela 13 – J48-4-emoções-50-arquivos-RAVDESS - Matriz de confusão

Raiva	Felicidade	Tristeza	Surpresa	
23	8	11	8	Raiva
11	19	10	10	Felicidade
8	13	18	11	Tristeza
14	4	10	22	Surpresa

Conforme observado na Tabela 11 ao diminuir a quantidade de arquivos em relação ao experimento anterior, o desempenho obtido na Tabela 8 caiu de 53.78% para 41%. Além disso, os parâmetros da Tabela 12 também decaíram em relação a Tabela 9. A quantidade de verdadeiros positivos também ficou muito baixa, como pode ser visto na Tabela 13, o elemento da diagonal principal é menor que a soma dos demais elementos da linha em questão. Isso nos leva a crer que diminuir o espaço do problema, nem sempre leva a melhorias na acurácia.

4.5 Experimento 4: Algoritmo SMO e apenas 4 emoções da base de dados RAVDESS

Nesta seção o experimento também foi dividido em duas etapas para realizar um comparativo com o uso do algoritmo J48 na seção anterior.

4.5.1 Experimento 4.1: Todos arquivos de cada classe de emoção

As Tabelas 14, 15 e 16 descrevem a primeira etapa do experimento, utilizando SMO e todos os arquivos das 4 emoções previamente escolhidas.

Tabela 14 – SMO-4-emoções-RAVDESS: Estatísticas gerais de desempenho

Instâncias classificadas corretamente	609 (79.2969%)
Instâncias classificadas incorretamente	159 (20.70%)
Estatística Kappa	0.724
Erro absoluto médio	0.2756
Raiz do erro quadrático médio	0.3482
Erro absoluto relativo	73.4943%
Raiz do erro médio relativo	80.4165%
Número total de instâncias	768

Tabela 15 – SMO-4-emoções-RAVDESS - Detalhes de acurácia para cada uma das classes

Classe	Taxa VP	Taxa FP	Precisão	Revocação	F-Measure	ROC Area
Raiva (angry)	0.818	0.075	0.785	0.818	0.801	0.91
Felicidade (happiness)	0.755	0.082	0.755	0.755	0.755	0.864
Tristeza (sadness)	0.813	0.061	0.817	0.813	0.815	0.926
Surpresa (surprised)	0.786	0.059	0.816	0.786	0.801	0.914

Tabela 16 – SMO-4-emoções-RAVDESS - Matriz de confusão

Raiva	Felicidade	Tristeza	Surpresa	
157	16	10	9	Raiva
18	145	14	15	Felicidade
9	17	156	10	Tristeza
16	14	11	151	Surpresa

Nesse teste, a acurácia aumentou consideravelmente atingindo quase 80%, conforme pode ser visto na Tabela 14. O que é uma melhoria tremenda comparada aos 53% do classificador J48 utilizado na seção anterior para o mesmo conjunto de dados. Nota-se também que os parâmetros da Tabela 15 foram muito bons atingindo valores muito próximos de 1. A Tabela 16 que representa a matriz de confusão possui valores muito bons. Esse experimento também foi melhor em relação ao Experimento 2, descrito na seção 4.3, que foi a primeira execução do SMO considerando todos as emoções da base, obtendo 67.78%.

4.5.2 Experimento 4.2: 50 arquivos em cada classe de emoção

Nesta seção o experimento realizado visou verificar se o fato de diminuir a quantidade de arquivos em cada emoção realmente piora os resultados.

Tabela 17 – SMO-4-emoções-50arquivos-RAVDESS: Estatísticas gerais de desempenho

Instâncias classificadas corretamente	143 (71.5%)
Instâncias classificadas incorretamente	57 (28.5%)
Estatística Kappa	0.62
Erro absoluto médio	0.2858
Raiz do erro quadrático médio	0.3629
Erro absoluto relativo	76.22%
Raiz do erro médio relativo	83.7987%
Número total de instâncias	200

Tabela 18 – SMO-4-emoções-50-arquivos-RAVDESS - Detalhes de acurácia para cada uma das classes

Classe	Taxa VP	Taxa FP	Precisão	Revocação	F-Measure	ROC Area
Raiva (angry)	0.84	0.08	0.778	0.84	0.808	0.934
Felicidade (happiness)	0.66	0.107	0.673	0.66	0.667	0.826
Tristeza (sadness)	0.62	0.08	0.721	0.62	0.667	0.847
Surpresa (surprised)	0.74	0.113	0.685	0.74	0.712	0.838

Tabela 19 – SMO-4-emoções-50-arquivos-RAVDESS - Matriz de confusão

Raiva	Felicidade	Tristeza	Surpresa	
42	3	1	4	Raiva
6	33	5	6	Felicidade
4	8	31	7	Tristeza
2	5	6	37	Surpresa

Por mais que a Tabela 17 mostre um resultado de 71% de acurácia, a Tabela 18 mostre bons resultados para os parâmetros Taxa VP, Precisão, F-Measure e Área ROC e a matriz de confusão descrita na Tabela 19 mostre uma boa quantidade de verdadeiros positivos, esse experimento ainda é pior em relação ao experimento 4.1 descrito na subseção 4.5.1. Portanto, verificou-se que diminuir o espaço do problema não foi garantia para aumentar a acurácia. Na realidade, pelos experimentos realizados, esta diminuiu em ambos os casos.

Com os testes realizados na base RAVDESS, pôde-se perceber que ao selecionar apenas 4 emoções, a acurácia aumentou consideravelmente alcançando a taxa de 79.29% utilizando o classificador SMO e todos os arquivos destas emoções. Ao diminuir a quantidade de arquivos para 50 por emoção, o resultado não foi conforme esperado, a acurácia decaiu em ambos os classificadores. Foram realizados mais alguns testes com a base de dados EmoDB para selecionar o método de treinamento a ser utilizado na aplicação *web*.

4.6 Experimento 5: Algoritmo J48 e todas emoções da base de dados EmoDB

Nesta seção, foi observado o desempenho do algoritmo J48 em uma base de dados diferente, a base EmoDB, que é considerada estado da arte e frequentemente utilizada em trabalhos tais como (ROELAND, 2014) (MENA, 2012) (IRIYA, 2014).

Tabela 20 – J48-7-emoções-EmoDB: Estatísticas gerais de desempenho

Instâncias classificadas corretamente	296 (55.3271%)
Instâncias classificadas incorretamente	239 (44.6729%)
Estatística Kappa	0.472
Erro absoluto médio	0.1297
Raiz do erro quadrático médio	0.3426
Erro absoluto relativo	53.7606%
Raiz do erro médio relativo	98.6662%
Número total de instâncias	535

Tabela 21 – J48-7-emoções-EMODB - Detalhes de acurácia para cada uma das classes

Classe	Taxa VP	Taxa FP	Precisão	Revocação	F-Measure	ROC Area
Raiva (anger)	0.685	0.074	0.744	0.685	0.713	0.826
Tédio (boredom)	0.543	0.081	0.543	0.543	0.0543	0.776
Nojo (disgust)	0.239	0.067	0.25	0.239	0.244	0.591
Medo (fear)	0.464	0.099	0.41	0.464	0.435	0.714
Felicidade (happiness)	0.493	0.099	0.432	0.493	0.461	0.707
Neutra (neutral)	0.557	0.053	0.647	0.557	0.599	0.8
Tristeza (sadness)	0.694	0.049	0.652	0.694	0.672	0.836

Tabela 22 – J48-7-emoções-EmoDB - Matriz de confusão

Raiva	Tédio	Nojo	Medo	Felicidade	Neutra	Tristeza	
87	1	5	6	28	0	0	Raiva
1	44	8	3	1	13	11	Tédio
2	6	11	14	5	1	7	Nojo
8	5	8	32	9	6	1	Medo
18	1	4	10	35	0	2	Felicidade
1	17	3	10	2	44	2	Neutra
0	7	4	3	1	4	43	Tristeza

A acurácia para este experimento atingiu 55.35%, como pode ser visto na Tabela 20. Ao observar a Tabela 21 notamos que os valores para os parâmetros Área ROC e F-Measure foram relativamente bons, enquanto que a Taxa VP foi mediana, que reflete diretamente na matriz de confusão da Tabela 22, onde podemos ver que ainda existe bastante instâncias classificadas erroneamente.

Foi um resultado ruim, porém deve-se levar em conta que todos os arquivos e emoções da base EmoDB foram utilizados e, portanto, este resultado é melhor comparado ao Experimento 1, que também utiliza o algoritmo J48, todos os arquivos da base RAVDESS e atingiu a acurácia de apenas 34.93%. Esse ganho pode referir-se a grande variação na fonética da língua alemã, o que leva maior diferença entre as classes de emoção.

4.7 Experimento 6: Algoritmo SMO com todas emoções da base de dados EmoDB

Este experimento utilizou o algoritmo SMO e a base de dados EmoDB, no intuito de comprovar que o SMO obteve o melhor desempenho em ambas as bases.

Tabela 23 – SMO-7-emoções-EmoDB: Estatísticas gerais de desempenho

Instâncias classificadas corretamente	468 (87.4766%)
Instâncias classificadas incorretamente	67 (12.5234%)
Estatística Kappa	0.8512
Erro absoluto médio	0.2062
Raiz do erro quadrático médio	0.3043
Erro absoluto relativo	85.3983%
Raiz do erro médio relativo	87.6486%
Número total de instâncias	535

Tabela 24 – SMO-7-emoções-EMODB - Detalhes de acurácia para cada uma das classes

Classe	Taxa VP	Taxa FP	Precisão	Revocação	F-Measure	ROC Area
Raiva (anger)	0.913	0.049	0.853	0.913	0.882	0.969
Tédio (boredom)	0.914	0.02	0.892	0.914	0.902	0.981
Nojo (disgust)	0.848	0.012	0.867	0.848	0.857	0.98
Medo (fear)	0.826	0.015	0.891	0.826	0.857	0.969
Felicidade (happiness)	0.746	0.024	0.828	0.746	0.785	0.946
Neutra (neutral)	0.937	0.018	0.902	0.937	0.919	0.985
Tristeza (sadness)	0.887	0.013	0.902	0.887	0.894	0.987

Tabela 25 – SMO-7-emoções-EmoDB - Matriz de confusão

Raiva	Tédio	Nojo	Medo	Felicidade	Neutra	Tristeza	
116	0	0	2	8	1	0	Raiva
0	74	1	0	0	3	3	Tédio
1	1	39	2	0	2	1	Nojo
5	0	1	57	3	1	2	Medo
14	0	1	3	53	0	0	Felicidade
0	4	1	0	0	74	0	Neutra
0	4	2	0	0	1	55	Tristeza

Como pode ser visto, a Tabela 23 deste experimento mostra a acurácia de 87.47%, que foi a melhor entre todos os experimentos. A Tabela 24 mostra valores muito bons para os parâmetros Taxa VP, Precisão, F-Measure e Área ROC. A matriz de confusão, descrita na Tabela 25, nos mostra que houve poucas instâncias classificadas de maneira incorreta, por exemplo para a linha da classe **Raiva**, que possui 127 instâncias, apenas 11 foram classificadas incorretamente.

De acordo com os experimentos, notou-se que o classificador de melhor desempenho foi o SMO. Além disso, o melhor experimento foi o Experimento 6, utilizando todos os arquivos da base de dados EmoDB que obteve 87.47%, enquanto que na base RAVDESS, foi necessário filtrá-la deixando apenas 4 emoções, com isso atingindo acurácia de 79.30% que também é consideravelmente boa para o estilo de problema sendo estudado.

Embora a acurácia para a base em língua inglesa tenha sido um pouco menor, é bastante válido utilizá-la para aplicação *web* que será construída, porque se trata de uma

língua é universal, há mais aplicações possíveis e um maior público poderá ser beneficiado com essa escolha. Desta forma, o modelo de treinamento com esta base foi escolhido e utilizado para a criação do *script*, que será executado para a classificar a emoção de entrada.

Um experimento com apenas 4 emoções na base EmoDB também foi realizado, a acurácia subiu para 90%. Houve um ganho de 3% em relação ao melhor resultado obtido na seção 4.7, porém com a remoção de 3 emoções (Medo, Nojo e Neutra) da classificação. Desse modo, optou-se considerar o melhor resultado o Experimento 6 descrito na seção 4.7 com acurácia de 87.47%. Afinal, 3% é um ganho negligível para perdermos 3 emoções na classificação.

5 Aplicação Web para Classificação de Emoções

Este capítulo apresenta o desenvolvimento da aplicação capaz de gravar áudios através do navegador, analisá-los e apresentar a emoção correspondente.

5.1 Visão geral do projeto

Nesse trabalho foi desenvolvida uma aplicação *Web*, que conta com o auxílio de *softwares* como *openSMILE* (EYBEN et al., 2013) e *Weka* (HALL et al., 2009) para a classificação de emoções através da voz, conforme pode ser visto na Figura 12.

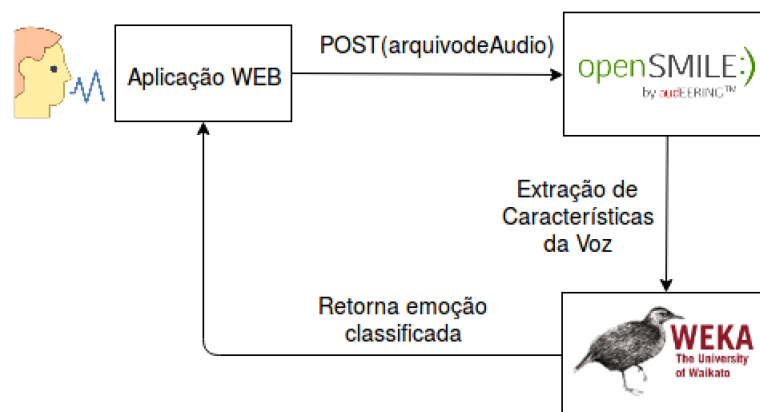


Figura 12 – Visão geral da aplicação.

5.2 Script para análise e classificação de emoções

O *script* funciona da seguinte forma: Inicialmente o arquivo *output.arff* é removido (arquivo de saída do openSMILE com as propriedades extraídas), isso deve ser feito por que se não os resultados iriam sendo concatenados ao arquivo anterior. Na segunda etapa o arquivo SMILExtract é executado recebendo o áudio como parâmetro e finalmente o classificador é executado na ferramenta Weka recebendo o modelo criado e o arquivo *output.arff* que foi gerado pelo *openSMILE*. Sua implementação está descrita na Figura 13

```
#!/bin/bash
echo "Executando reconhecedor de emoções"

# $1 is the first argument, in this case is the audio file
file="output.arff"
if [ -f "$file" ]
then
    echo "output.arff removido"
    rm output.arff
fi

./SMILExtract -C emobase.conf -I $1 -classes \"{angry,happiness,sadness,surprised}\"
java -classpath weka.jar weka.classifiers.functions.SMO -l emobase.model -T output.arff -p 0
```

Figura 13 – Shell Script responsável pela análise e classificação da emoção

5.3 Desenvolvimento da aplicação Web

5.3.1 Javascript

Como linguagem de programação *front-end* foi utilizada JavaScript, para tratar o comportamento referente a gravação do áudio e envio desta gravação para o *back-end* via AJAX.

5.3.2 NodeJS

Para a linguagem *back-end* foi utilizado o NodeJS, que executa código javascript no lado do servidor. A escolha do NodeJS se deu pelo fato da manutenção e legibilidade do código. Com apenas uma linguagem de programação foi possível desenvolver tanto o *back-end* como o *front-end*.

5.3.3 Biblioteca RecordRTC

Para a gravação do áudio na página web e realizar *upload* para o lado do servidor foi utilizada a biblioteca RecordRTC(KHAN,) como auxiliadora. É uma biblioteca que está sob *MIT License* e oferece recursos para a gravação de áudio e vídeo no navegador. Neste trabalho foi utilizado apenas o recurso de áudio.

Ao clicar no botão "Gravar" representado na Figura 14 o RecordRTC irá disparar um evento para a gravação do áudio guardando este em um *buffer*. Ao clicar em "Parar Gravação" será realizada uma requisição do tipo POST enviando o arquivo que foi gravado. No lado do servidor será executado o script, que possui *output* conforme a Figura 15.

Como visto a saída contém informações desnecessárias para um usuário final, que precisará saber apenas a emoção classificada. Para resolver isso, ao término da execução do script será executada uma função para processar a saída obtida.

A função `runScript` é responsável por executar o *script* fazendo o uso de um processo filho, e retornar a saída da função `processResult`. A função `processResult` verifica



Figura 14 – Tela inicial da aplicação

```
(MSG) [2] in SMILEExtract : openSMILE starting!
(MSG) [2] in SMILEExtract : config file is: emobase.conf
(MSG) [2] in cComponentManager : successfully registered 95 component types.
(MSG) [2] in instance 'arffsink' : writing ARFF header (950 features)...
(MSG) [2] in cComponentManager : successfully finished createInstances
                                (23 component instances were finalised, 1 data
memories were finalised)
(MSG) [2] in cComponentManager : starting single thread processing loop
(MSG) [2] in cComponentManager : Processing finished! System ran for 258 ticks.
Executando reconhecedor de emoções
output.arff removido

=== Predictions on test data ===

inst#    actual  predicted error prediction
    1      1:?    1:angry    0.333
```

Figura 15 – Saída da execução do script

a ocorrência do padrão da expressão regular `/[0-9]:[a-z]+/` no resultado recebido como parâmetro, e retorna a string que se encontra depois do caractere `':'`.

As funções que estão descritas abaixo, foram implementadas usando Node.js no lado do *back-end*.

```
1      function processResult(result){
2          var emotionRegex = /[0-9]:[a-z]+/,
3              onlyEmotion = '',
4              splittedString;
5
6          splittedString = result.match(emotionRegex)[0].split(':');
7          onlyEmotion = splittedString[1];
```

```
8         return onlyEmotion;
9     }
10
11     function runScript(response, fileName){
12
13         //Execute the emotionScript
14         var execSync = require('child_process').execSync;
15         stdoutOfChild = execSync("sh emotionScript.sh uploads/" +
16             fileName, function (error, stdout, stderr) {
17             if (error) {
18                 console.log('exec error: ' + error);
19             }
20         });
21
22         return processResult( stdoutOfChild.toString());
23     }
24
25     // this function uploads files
26     function upload(response, postData) {
27         var files = JSON.parse(postData),
28             responseObject = {};
29
30         // writing audio file to disk
31         _upload(response, files.audio);
32
33         responseObject.emotion = runScript(response, files.audio.name);
34         responseObject.audioName = files.audio.name;
35
36         if (files.uploadOnlyAudio) {
37             response.statusCode = 200;
38             response.writeHead(200, { 'Content-Type': 'application/json' });
39             response.end(JSON.stringify(responseObject));
40         }
41
42         if (!files.uploadOnlyAudio) {
43             // writing video file to disk
44             _upload(response, files.video);
45
46             merge(response, files);
47         }
48     }
```

O retorno da função `runScript` será utilizado para montar o objeto `responseObject` que será enviado no sucesso da requisição para o *front-end*, contendo a emoção e o nome do arquivo de áudio gerado.

A emoção será mostrada no *front-end* e será gerado um *playback* do arquivo de áudio para poder ser ouvido no próprio navegador. A Figura 16 ilustra o resultado.

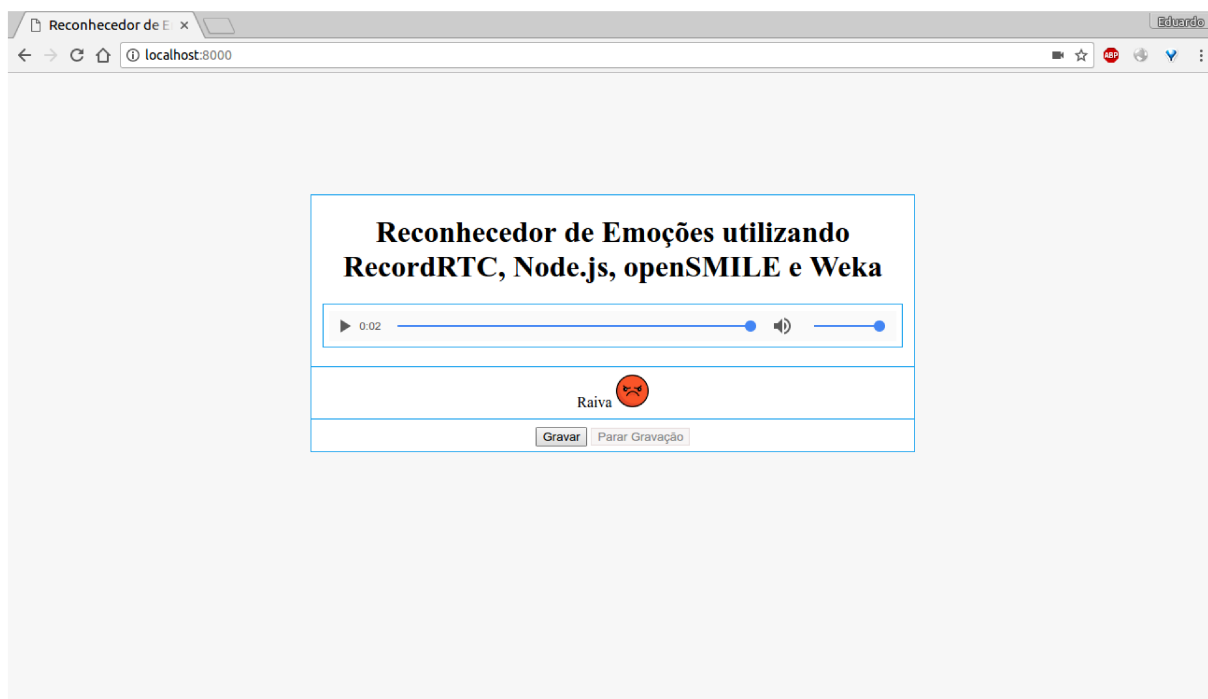


Figura 16 – Resultado da classificação na aplicação web

A aplicação respondeu razoavelmente bem aos testes executados, o grande problema é a qualidade de gravação e os possíveis ruídos que podem interferir no sinal da voz. São muitas variáveis envolvidas, como qualidade do microfone que está sendo usado para realizar a gravação, o local em que o indivíduo se encontra, a perda de qualidade ao enviar o áudio para o servidor, entre outras.

Ao verificar que o desempenho ao integrar a aplicação com a base de dados RAVDESS foi razoável, foi realizado um teste integrando com a base EmoDB, o que é uma troca bem simples de ser feita. Pôde ser observado um aumento na acurácia e por esse fato fica claro o por que dessa base ser considerada estado da arte. Afinal, os arquivos de áudio possuem qualidade melhor e maior variação entre eles, impulsionando a classificação.

Para os trabalhos futuros deve-se verificar possíveis melhorias em classificadores envolvendo a base RAVDESS ou outra base em inglês. Deve ser avaliado também a concepção de uma base em português, para realização de outros possíveis testes.

6 Conclusão

Neste trabalho foi desenvolvida uma aplicação *web* capaz de gravar áudios e reconhecer emoções através deste. Para a extração de características do áudio a ferramenta *openSMILE* foi utilizada e para a classificação foi utilizado o classificador SMO providenciado pela ferramenta *Weka*. Foi alcançado 79.29% de acurácia ao utilizar a base de dados RAVDESS e 87.47% ao utilizar a base EmoDB. Os resultados foram bons, porém essa acurácia pode variar devido a ruídos em sinais de áudio.

O resultado do Experimento 4 serviu para contribuição a um artigo em parceria com dois doutorandos: Fabiano Silvério Ribeiro Alves e Luciane de Fátima Silva, submetido ao SBIE 2017 com o título: *"Reconhecimento Dinâmico de Emoções da Fala e da Face em Ambiente Virtual de Aprendizagem com foco na Análise Paraconsistente Anotada para Recomendações"*. O modelo de treinamento deste experimento foi selecionado devido ao fato de ter sido o melhor dos testes utilizando a base RAVDESS, que foi escolhida por ser uma base em inglês e poder abranger mais aplicações e um maior público.

O *software openSMILE* foi utilizado para a extração das características de voz, dentre elas as mais importantes foram os coeficientes mel-ceptrais (MFCC), a intensidade e o *pitch*. O *openSMILE* mostrou ser uma solução viável à integração com uma aplicação *web*.

A integração entre a aplicação e os *softwares openSMILE* e *Weka* pôde ser realizada devido ao fato de ser possível utilizar as funções necessárias de cada *software* em linha de comando. Dessa forma um *script bash* foi criado, responsável por executar tais funções. Além disso com o auxílio do Node.js foi possível isolar *back-end* e *front-end* da aplicação utilizando basicamente a mesma linguagem em ambos e com isso implementar código Javascript no *backend*.

Essa aplicação foi concebida com o objetivo de aprimorar futuramente um ambiente virtual de aprendizagem para melhorar a experiência de um aluno. Mas as suas aplicações são tremendas, como por exemplo esse reconhecimento de emoção através da voz poderia ser utilizado para ligações telefônicas de *call centers* no intuito de coletar dados sobre como os clientes se comportam; Em reuniões do *Skype* ou *Hangouts*; Ambientes de aprendizado voltados a pessoas com deficiências motoras e/ou visuais; entre outras.

Pode-se ver que a tecnologia caminha para uma direção onde ficará cada vez mais próxima das relações humanas. De tal forma que sistemas computacionais fiquem responsáveis por reconhecerem emoções e se adaptem para melhor prover o usuário.

6.1 Dificuldades encontradas

A maior das dificuldades ao início do trabalho foi pensar em como enviar o áudio para o servidor de maneira contínua, isto é, o usuário iniciaria sua fala e aos poucos seriam feitas requisições ao servidor enviando partes do áudio gravado. Várias dúvidas surgiram: em qual momento o arquivo deveria ser enviado, quantos segundos cada parte deveria possuir, como garantir que as partes seguiriam uma ordem sendo que chamadas assíncronas estavam sendo realizadas. Esses contra-tempos foram resolvidos com o controle de botões na parte do *front-end*, dessa forma o arquivo só será enviado ao término da gravação, que será controlada pelo o usuário. Uma solução momentânea que deverá ser melhorada em trabalhos futuros.

A escassez de bases de áudio de emoções também foi um problema. Foram encontradas apenas a base RAVDESS e EmoDB, não sendo encontrada nenhuma base em português.

Pode-se perceber também que a área de reconhecimento facial atualmente possui mais materiais, ocasionando em menos referências bibliográficas para o reconhecimento de voz. Provavelmente devido ao fato da análise da face tratar de parâmetros discretos, enquanto a voz trata de parâmetros contínuos. Além disso, a face pode ser considerada mais padronizada e universal do que a voz, pois a diferença de idiomas e regiões gera bastante mudanças na maneira como falamos.

Por fim, foi difícil encontrar *softwares open-source* que extraíssem características da voz. O *software openSMILE* foi a única solução que atendeu as necessidades desse trabalho, porém devido aos poucos materiais, foi uma tarefa árdua realizar a instalação e usar as funcionalidades que ele proporciona. Pelo estilo de arquivo de saída não foi possível realizar os testes utilizando os classificadores propostos por (IRIYA, 2014). Além do fato de ter conseguido realizar a instalação apenas em sistemas Unix, no caso deste trabalho o sistema operacional utilizado foi o Ubuntu 16.04.

6.2 Trabalhos futuros

Nesta seção serão listados alguns pontos que podem ser melhorados futuramente:

1. Visto que não foram encontradas bases de áudio em português, seria interessante realizar um projeto onde atores gravariam as falas representando emoções, para virem a ser utilizadas no treinamento.
2. As gravações de áudio podem vir a ser contínuas, de tal forma que o usuário inicia uma sessão, como por exemplo uma aula e durante toda a aula será gravado seu áudio e de tempos em tempos uma requisição será enviada para o servidor.

3. A aplicação *web* por trabalhar com *Javascript* poderá ser integrada a um ambiente virtual de aprendizagem como *Moodle*.
4. Para a integração com o *Moodle*, uma ideia inicial seria salvar o registro das emoções e o tempo de cada arquivo em uma tabela em um banco de dados, para dessa forma ser fácil consultá-la e fazer inferências a partir desta.
5. Nessa integração o reconhecedor de emoções pela voz será unificado com um reconhecedor facial. Dessa forma, ambos irão inferir emoções e fazendo o uso da lógica paraconsistente será decidido qual emoção deverá ser considerada como correta.

Referências

- BURKHARDT, F. et al. A database of german emotional speech. In: *Interspeech*. [S.l.: s.n.], 2005. v. 5, p. 1517–1520. Citado 2 vezes nas páginas 20 e 21.
- COWIE, R. et al. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, IEEE, v. 18, n. 1, p. 32–80, 2001. Citado 2 vezes nas páginas 12 e 18.
- DAVIS, J.; GOADRIC, M. The relationship between precision-recall and roc curves. In: ACM. *Proceedings of the 23rd international conference on Machine learning*. [S.l.], 2006. p. 233–240. Citado na página 29.
- DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, IEEE, v. 28, n. 4, p. 357–366, 1980. Citado na página 19.
- EKMAN, P. An argument for basic emotions. *Cognition & emotion*, Taylor & Francis, v. 6, n. 3-4, p. 169–200, 1992. Citado na página 12.
- EMOSPEECH. <<https://emospeech.net/>>. Acessado em: 21/11/2016. Citado na página 16.
- EMPATH. <<https://webempath.net/lp.eng>>. Acessado em: 21/11/2016. Citado na página 16.
- EYBEN, F. et al. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In: *Proc. of ACM Multimedia 2013*. Barcelona, Spain: ACM, 2013. p. 835–838. Citado 3 vezes nas páginas 16, 19 e 39.
- HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009. Citado 2 vezes nas páginas 20 e 39.
- IRIYA, R. *Análise de sinais de voz para reconhecimento de emoções*. Tese (Doutorado) — Universidade de São Paulo, 2014. Citado 4 vezes nas páginas 16, 18, 35 e 45.
- KHAN, M. *RecordRTC*. <<https://github.com/muaz-khan/RecordRTC>>. Acessado em: 20/07/2017. Citado na página 40.
- KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: STANFORD, CA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145. Citado na página 21.
- LIVINGSTONE, S. R.; PECK, K.; RUSSO, F. A. Ravdess: The ryerson audio-visual database of emotional speech and song. In: *Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science*. [S.l.: s.n.], 2012. Citado 2 vezes nas páginas 20 e 21.
- LUTTER, M. *Mel-Frequency Cepstral Coefficients*. 2014. <<http://recognize-speech.com/feature-extraction/mfcc>>. Acessado em: 21/11/2016. Citado na página 19.

- MEL Frequency Cepstral Coefficients Tutorial. <<http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/>>. Acessado: 04/12/2016. Citado na página 19.
- MENA, M. E. Emotion recognition from speech signals. Universitat Politècnica de Catalunya, 2012. Citado 2 vezes nas páginas 16 e 35.
- 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*, ASA, v. 93, n. 2, p. 1097–1108, 1993. Citado na página 18.
- PLATT, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. [S.l.], 1998. 21 p. Disponível em: <<https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/>>. Citado na página 21.
- PLUTCHIK, R. A general psychoevolutionary theory of emotion. *Theories of emotion*, Academic Press New York, v. 1, n. 3-31, p. 4, 1980. Citado na página 12.
- QUINLAN, R. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993. Citado na página 20.
- ROELAND, J. P. Multisensory emotion recognition with speech and facial expression. 2014. Citado 3 vezes nas páginas 16, 20 e 35.
- SCHULLER, B. et al. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, Elsevier, v. 53, n. 9, p. 1062–1087, 2011. Citado 2 vezes nas páginas 7 e 17.
- SHAH, A. *Opensmile Toolkit Quick Setup*. 2015. <<https://www.linkedin.com/pulse/opensmile-toolkit-quick-setup-ankit-shah>>. Acessado em: 24/06/2017. Citado na página 19.