Projeto de Pesquisa: Identificação de emoções na fala através da análise do timbre da voz

Aluno: Jardel Oliveira Santos RA 121029

Orientador: Tiago F. Tavares

Co-orientador: Paula D. P. Costa

14 de novembro de 2017

Sumário

Resumo

O estudo em computação afetiva, ou *affective computing*, desenvolve sistemas capazes de interagir com emoções humanas, com a habilidade de detectá-las e respondê-las apropriadamente. O objetivo do projeto é implementar um método de identificação e classificação de emoções a partir da análise do timbre da voz, por meio da aplicação de técnicas de processamento digital de sinais e técnicas de aprendizado de máquina relacionando problema de classificação de vetores em espaços vetorais. Uma base de dados rotulada em diferentes emoções foi criada e, a partir dela extraídas informações para classificação. Foram testados diversos métodos de extração e classificação e os resultados obtidos foram uma acurácia em torno de 18.52% para um modelo de 22 emoções e de 37.24% para um modelo com 7 grupos de emoções. A limitação de desempenho dos classificadores se devem à dificuldade na obtenção de melhores e maiores quantidades de amostras para o treino.

1 Introdução

A computação afetiva, ou *affective computing*, é um campo da computação que busca desenvolver sistemas que detectam e respondem apropriadamente a emoções humanas. As pesquisas nessa área relacionam engenharia e ciência da computação com psicologia, ciência cognitiva, neurociência e sociologia.

As emoções estão presentes na vida humana e influenciam suas ações, seja como forma de motivação para realizar determinada tarefa, no aprendizado, na forma em que se interage socialmente, nas tomadas de decisões e no uso da intuição [8]. Com o crescimento do número de pessoas interagindo com computadores, é interessante considerar o uso de sistemas de reconhecimento da emoção no desenvolvimento de novas interfaces tecnológicas e o impacto que isso possa gerar.

A comunicação inter-humana pode ser feita tanto pelo uso do corpo e gestos como pela fala. Além da informação linguística carregada na fala, ela também pode conter informações a respeito do estado emocional do orador. Por esse motivo, uma mesma mensagem pode ter diferentes significados se falada com diferentes expressões de emoção. O reconhecimento da emoção pela fala, ou Speech Emotion Recognition(SER), visa identificar automaticamente o estado emocional de uma pessoa a partir de sua voz [1].

As emoções humanas podem ser classificadas considerando diferentes aspectos e suas expressões podem variar significativamente entre culturas e ao longo de gerações [6]. Existem na literatura diversos modelos emoções. Em particular, os modelos categóricos de emoções classificam-nas segundo um pequeno vocabulário de emoções. É o caso, por exemplo, das seis emoções básicas de Ekman [4], sendo elas: felicidade, tristeza, surpresa, nojo, medo e raiva. Este modelo possui aplicação limitada na modelagem de interações típicas do cotidiano [3]. No contexto deste trabalho, adotaremos o modelo proposto por Ortony, Clore e Collins, também denominado modelo OCC de emoções, que adota um vocabulário mais abrangente, de 22 emoções [7]. Nesse modelo, as emoções são consideradas a partir de uma hierarquia de acontecimentos, onde, por exemplo, o alívio, uma das emoções adotadas, é sentido devido a não confirmação de um evento que originou um medo, outra emoç-ão.

Esse trabalho visa implementar um método de identificação e classificação de emoções a partir da análise do timbre da voz, por meio da aplicação de técnicas de processamento digital de sinais e técnicas de aprendizado de máquina relacionando problema de classificação de vetores em espaços vetorais.

2 Objetivos

1. Identificação e separação de amostras para base de dados;
2. Extração de atríbutos relacionados ao timbre por meio de técnicas de processamento digital de sinais;
3. Classificação de vetores em espaços vetoriais utilizando técnicas de inteligência computacional;
4. Avaliar resultados obtidos por meio de distintas métricas de aprendizado de máquina.

3 Metodologia

Como primeira etapa da metodologia, foi feito um estudo de técnicas de processamento digital de sinais relacionados ao problema de identificação de timbre, e o estudo de técnicas de inteligência computacional relacionados ao problema de classificação de vetores em espaços vetoriais. Esta etapa foi necessária para familiarizar o aluno com os conceitos do campo em estudo, bem como proporcionar embasamento teórico para correta aplicação das técnicas que serão exploradas.

Foi criada uma base de dados para o projeto. Nesta base de dados foi necessário separar as amostras para cada emocão, e rotulá-las devidamente para o processo de classicacão.

A base de dados foi construída a partir de uma seleção de vídeos de 7 atores e atrizes do Instituto de Artes da Universidade Estadual de Campinas - UNICAMP, já capturados previamente a este trabalho [3]. Cada um dos atores e atrizes receberam scripts com 4 a 6 falas sendo algumas falas neutras, onde o ator/atriz apenas faria a leitura sem expressar emoção, e as mesmas falas com emoção, onde o ator/atriz representaria a emoção em si; essas falas contém uma grande quantidade de fonemas variados da língua portuguesa. Para criação de rótulos, dividiu-se as falas em 22 emoções baseadas no modelo de emoções OCC [7], sendo elas: feliz-por, alegria, esperança, satisfação, alívio, orgulho, gratificação, gratidão, admiração, amor, pena, tristeza, medo, ressentimento, medos confirmados, vergonha, censura, remorso, regojizo, desapontamento, nojo e raiva. Os vídeos foram gravados no estúdio da RTV Unicamp, em ambiente controlado, com fundo do tipo *chroma-key*.

Após criação e rotulação das amostras, foram aplicadas as técnicas de processamento digital de sinais estudadas para extrair atributos do sinal de áudio provido das falas dos atores. Posterior à essa extração, foram utilizados os algoritmos de inteligência computacional estudados para realizar a classificação, e como etapa final, utilizou-se métricas para análise dos resultados obtidos.

3.1 Extração dos atributos de áudio

Os arquivos de áudio utilizado nesse trabalho estão específicados no formato MIREX. Esse formato envolve arquivos com a extensão *.wav* com taxa de amostragem de 16 bits/amostra.

Para o processo de extração dos atributos, as amostras de áudio foram mapeadas em um espaço vetorial onde cada dimensão corresponde a um diferente atributo que descreve o sinal de áudio. A principal hipótese por trás dos atributos extraídos é de que sons relacionados à percepção são associados a atributos de nível baixo, como por exemplo a distância entre os vetores, no qual os de mesma classe devem possuir uma distância menor que de classes distintas. Assim, a extração dos atributos de cada amostra foi feita da seguinte maneira.

Inicialmente a amostra é normalizada para variância unitária com média zero, de forma a evitar efeitos de ganho do sinal nas etapas seguintes. Em seguida, a amostra é dividida em quadros de 46,3 milisegundos, com uma sobreposição de 50% entre quadros subsequentes.

Cada quadro da amostra é multiplicado pela função da janela de Hanning. As funções em janela são funções matemáticas cujo valor é não-nulo apenas em um intervalo especificado, e a multiplicação pela janela de Hanning, será utilizada para estimar a magnitude espectral do sinal. A janela de Hanning, definida como a Expressão 1, é mostrada na Figura 1.

$$w(n) = 0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] \tag{1}$$

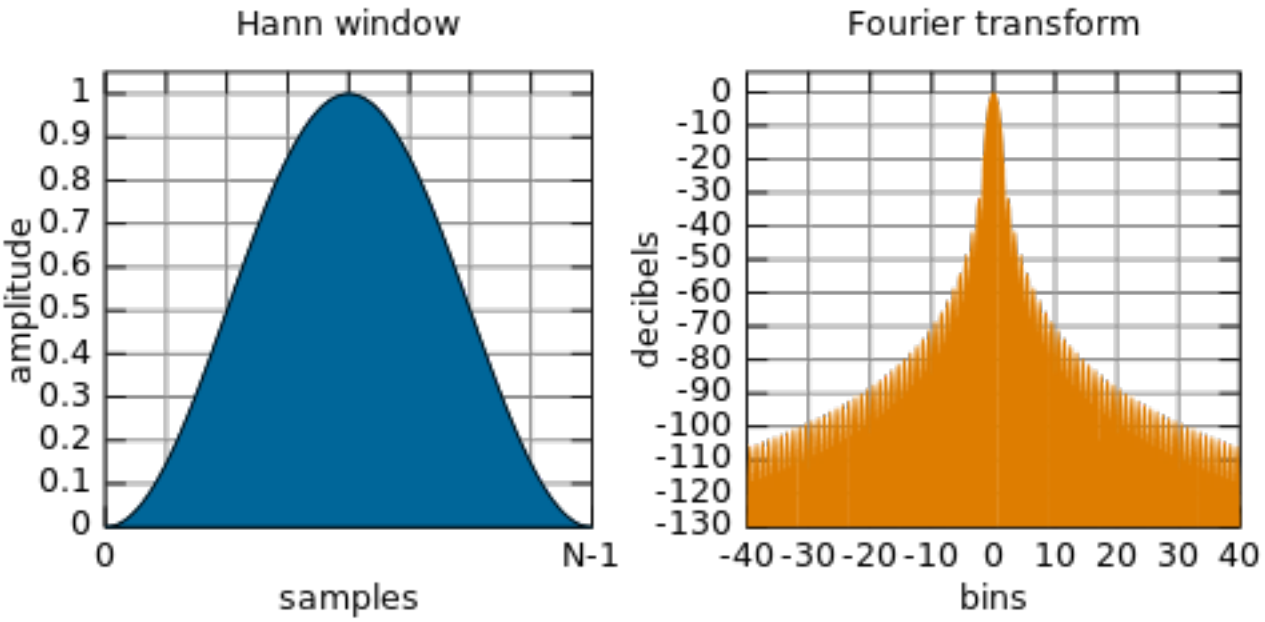

Figura 1: Janela de Hanning. Imagem extraída de *en.wikipedia.org/wiki/Window_function*

O espectro é, então, parametrizado como uma série de descritores (features), a saber:

1. Centroide do espectro: indica a posição no qual o centro de massa do espectro se encontra.
2. Roll-off do espectro: valor (normalmente dado em dB/decada) de queda ou ganho de potência pela frequência do espectro quando esse deixa de ser plano.
3. Planicidade do espectro: valor indicando a planitude do sinal, utilizado para quantificar a ruides do som.
4. Fluxo do espectro: mede a rapidez da mudança de magnitude no espectro, através da comparação da magnitude entre os quadros subsequentes.
5. Coeficientes Cepstrais de frequências melódicas (Mel-Frequency Ceptral Coefficients): tipo de representação cepstral do áudio que se relaciona ao timbre, muito utilizado em reconhecimentos de voz.

As primeiras e segundas derivadas dos atributos também são calculadas, pois contém informação da variação do conteúdo do áudio ao longo do tempo.

3.2 Classificação de emoções

Para a classificação das emoções foram utilizados 3 algoritmos de aprendizado de máquina: K-Nearest Neighbors (K-NN), Support Vector Machines (SVM) e Random Forest (RF), que são introduzidos a seguir.

O princípio por trás dos métodos de classificação está na divisão dos dados de um banco em treino e teste. Em nosso caso cada dado equivale a uma fala, contendo variáveis que servirão para caracterizá-la. Essas variáveis são os atributos extraídos do sinal de áudio da fala e o rótulo indicando a que emoção pertence.

Após a divisão dos dados, os de treino são utilizados pelo classificador para otimizar os parâmetros do modelo que, por sua vez, tentará prever a emoção dos dados de teste com base em seus atributos e no algoritmo de classificação utilizado. Em seguida, as emoções preditas pelo modelo são comparadas com os rótulos originais, estimando a eficácia do modelo.

3.2.1 K-Nearest Neighbors (K-NN)

K-NN é um algoritmo simples que guarda todos os dados de treino e os utilizam para classificar novos dados baseado na medida de similaridade, como por exemplo a distância euclidiana utilizada em nosso modelo.

Dado dois vetores $\vec{x}$ e $\vec{y}$ com n dimensões, a distância euclidiana entre eles é dada pela seguinte fórmula:

$$E_d = \sqrt[2]{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

A palavra k é um parâmetro desse algoritmo que indica o número de vizinhos próximos, ou os k vetores treinados com as menores distâncias de um novo vetor teste. Com base no rótulo que mais aparece nos k vizinhos realiza-se a predição da classe.

A Figura 2 mostra as diferentes predições de um novo vetor para diferentes valores de k.

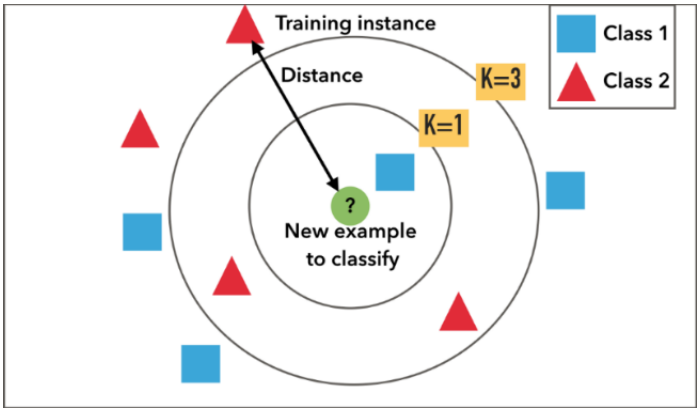

Figura 2: Modelo KNN, para diferentes valores de k. Imagem extraída de *https* : *//sflscientific.com/data – science – blog/2016/6/4/time – series – analysis – fitbit – using – dtw – and – knn*

3.2.2 Support Vector Machines (SVM)

O classificador SVM, primeiramente introduzido por Vapnik [9], realiza uma classificação binária, ou seja, ele separa um conjunto de vetores de treino para duas classes distintas (x_1, y_1), (x_2, y_2), ..., (x_m, y_m), onde $x_i \in R^d$, denota vetores em um espaço com dimensão d, e $y_i \in \{-1, 1\}$ é o rótulo da classe. O modelo SVM é gerado através do mapeamento dos vetores de entrada em um novo espaço com maior dimensão, denotado como $\phi : R^d \rightarrow H^f$ onde $d < f$. Depois, um hiperplano de separação ótima é construído no novo espaço dimensional através de uma função de núcleo $\mathrm{K}(x_i, x_j)$, sendo o produto dos vetores de entradas x_i e x_j mapeados pela transformação $\phi(x)$ [5].

A Figura 3 ilustra o procedimento de um SVM baseado em um núcleo linear, o qual mapea um espaço de entrada não linear em um novo espaço separável. Em particular, todos os vetores de um lado do hiperplano são rotulados como -1, e os vetores do outro lado como +1. As instâncias de treino que se situam próximas ao hiperplano no espaço transformado são chamadas de vetores de suporte. O número de vetores de suporte normalmente são pequenos comparados com o tamanho do banco de dados e eles determinam a margem do hiperplano e, portanto, a superfície de decisão

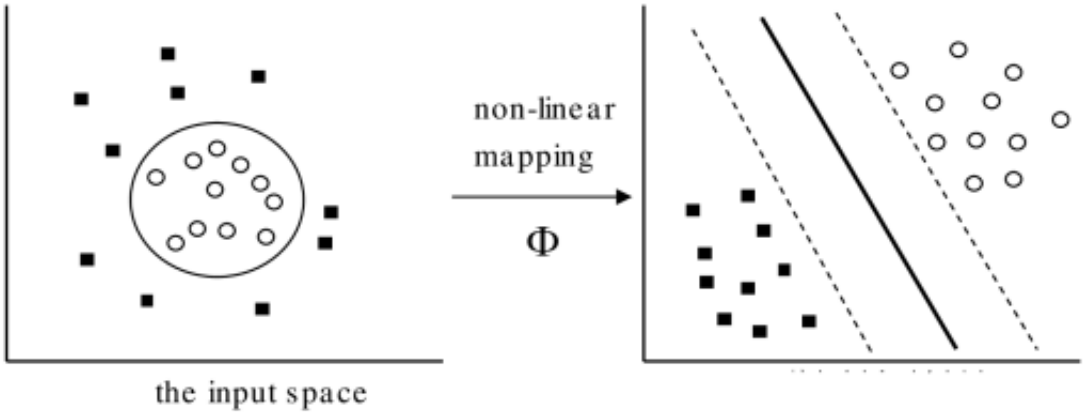

Figura 3: Modelo SVM. Imagem extraída de [5].

[5]. Ao retornar para o espaço não linear, o hiperplano toma uma forma não linear que separa as classes, como o círculo observado.

3.2.3 Random Forest (RF)

Random Forest se refere a um conjunto de classificadores em árvore de decisão [2] onde cada árvore é gerada a partir de um subconjunto das amostras de treino. Dado um novo vetor x_i, sua predição final é baseada na classe mais prevista pelas árvores do modelo. A Figura 4 mostra as fases de treinamento e classificação do modelo RF.

Nesse método, cada subconjunto utilizado pode conter amostras existentes em outros subconjuntos, o que facilita na estimação do erro do modelo, no qual as próprias amostras de treino são testadas separadamente, dispondo das árvores cujo subconjunto gerador não contém a amostra de treino a ser testada, método conhecido como *out of bag error*.

Para um bom modelo RF, considerando os vetores de treino possuíndo M dimensões representando seus atributos, cada árvore utiliza $m << M$ atributos em seus nós, de forma que mesmo com uma menor capacidade invidual de predição de uma arvóre, a correlação entre as árvores também são menores, assim se espera uma menor variância na resposta média do conjunto das árvores em relação à árvores individuais que utilizam mais ou todos os M atributos.

3.3 Treinamento dos algoritmos - Validação Cruzada

A separação das falas em dados de treino e teste serão feitas por meio de uma validação cruzada entre os atores.

A validação cruzada é uma técnica amplamente empregada em modelos de predição. O con-

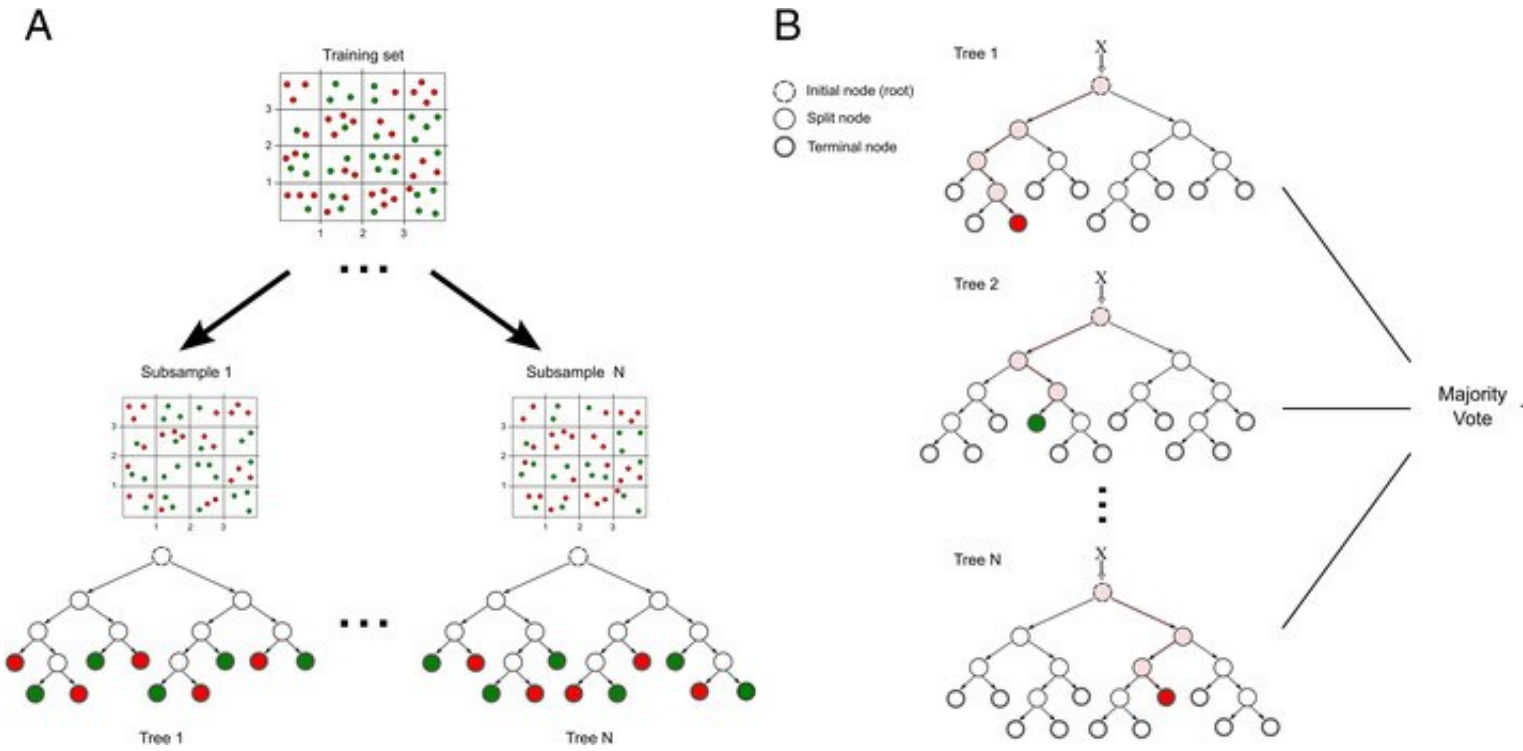

Figura 4: Modelo RF. A) Treinamento das árvores. B) Classificação de um vetor X. Imagem extraída de $https://www.slideshare.net/F789GH/customer-linguistic-profiling$

ceito central dessa técnica está no particionamento de um conjunto de dados em subconjuntos mutualmente exclusivos para, posteriormente, utilizar alguns desses subconjuntos na estimação dos parâmetros do modelo (dados de treinamento), e o restante dos subconjuntos (dados de teste) são empregados na validação do modelo.

Em nosso caso, os subconjuntos são os atores. Nossos modelos de classificação serão treinados com as falas de 6 atores e realizarão a predição no ator restante, sendo esse processo realizado sete vezes, uma para cada ator servindo de teste.

A validação cruzada entre os atores tem o cuidado de garantir que os atores presentes no conjunto de teste não estejam presentes no conjunto de treino. Assim, é possível medir a variância do modelo, ou seja, qual a performance do modelo para cada ator e o quanto ela varia entre os atores, criando uma melhor percepção de sua perfomance em prever novas emoções.

3.4 Métricas para avaliação dos classificadores

As métricas utilizadas nesse trabalho são bastante conhecidas no campo de aprendizado de máquina e mais específico para classificações estastíticas. Essas métricas são calculadas a partir de uma matriz, ou tabela, de confusão criada pelo modelo de predição.

A matriz de confusão é uma tabela no qual cada coluna representa os valores previstos pelo modelo, e cada linha representa os valores reais dos dados de teste. Se basearmos em uma matriz binária, como a da Figura 5, os valores de True Positive (TP) indicam a quantidade de dados que pertencem àquela classe e foram classificados corretamente, enquanto os valores de False Positive apontam o número de dados que não pertencem a classe mas foram previstos como pertencentes. De forma analoga, True Negative (TN) e False Negative (FN) designam os dados classificados corretamente como não sendo da classe e os dados que são da classe mas foram previstos como não sendo, respectivamente.

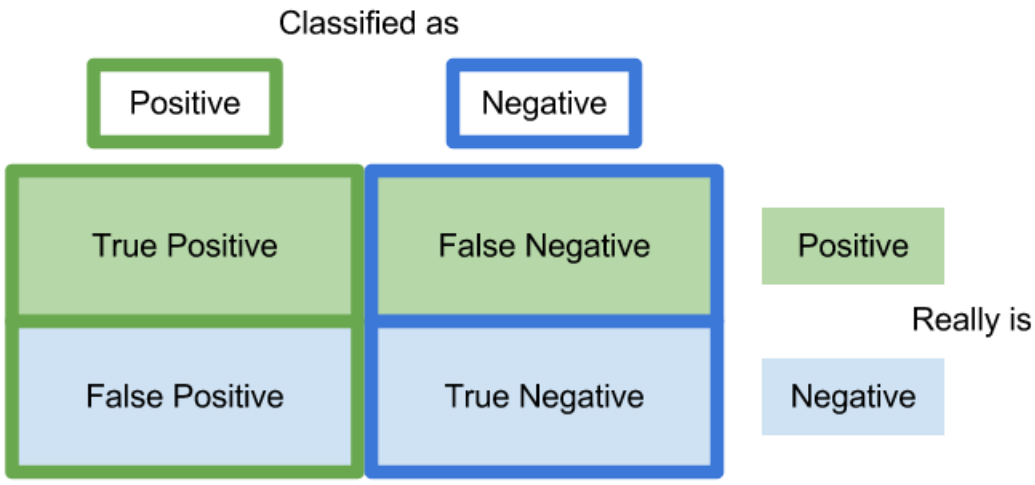

Figura 5: Matriz de confusão binária. Imagem extraída de $http://yuvalg.com/blog/2012/01/01/precision-recall-sensitivity-and-specificity/$

Com base na matriz de confusão, as métricas utilizadas foram:

1. Acurácia ⇒ quantidade de acertos do modelo, dividido pela quantidade total de dados:

$$\mathrm{A} = \frac{TP+TN}{TP+FP+TN+FN}$$

2. Precisão ⇒ para a predição de uma classe, indica a razão entre os dados previstos corretamentes e o total de dados classificados como sendo dessa classe:

$$\mathrm{P} = \frac{TP}{TP+FP}$$

3. Revocação (Recall) ⇒ para os dados pertencentes a uma classe, indica a razão entre os dados previstos corretamente e o total de dados pertencentes à classe:

$$\mathrm{R} = \frac{TP}{TP+FN}$$

4. F1-Score ⇒ é a média harmônica entre os resultados de precisão e revocação, dado da seguinte forma:

$$\text{F1} = 2 \cdot \frac{P \cdot R}{P + R}$$

Devido ao modelo de validação cruzada desse trabalho, para cada ator foram calculadas as métricas acima e, como resultado final será dado a média das métricas dos atores e o desvio padrão correspondente.

4 Resultados

4.1 Configuração 1 - Modelo OCC de 22 emoções

A Tabela 1 mostra resultados de classificação considerando o modelo de emoções OCC.

Modelo	Acurácia (%)	Precisão (%)	Revocação (%)	F1-score (%)
KNN	18.52 +/- 6.42	19.77 +/- 11.18	18.52 +/- 6.42	16.10 +/- 7.74
SVM	18.70 +/- 9.03	16.83 +/- 9.89	18.70 +/- 9.03	14.26 +/- 9.35
Random Forest	18.49 +/- 10.66	15.66 +/- 12.22	18.49 +/- 10.66	14.32 +/- 11.52

Tabela 1: Configuração 1 - Validação Cruzada por Ator - Métricas

Como pode-se perceber na Tabela 1, de uma maneira geral, os modelos obtiveram resultados parecidos. O modelo K-NN, obteve menos desvios nas métricas, assim, foi impressa uma matriz de confusão para esse modelo, mostrado na Figura 6.

A matriz de confusão possui 22 linhas e 22 colunas, no qual cada linha representa uma das emoções e as colunas referentes à linha representam a emoção em que foi prevista. Para facilitar tome como exemplo o ponto (0,5) cujo valor é 4. Esse valor representa a quantidade de falas cujo emoção expressa é a "Emoção 0" (Linha 0), mas o classificador K-NN preveu como sendo a "Emoção 5" (Coluna 5).

A partir da matriz de confusão foi calculada a taxa de acerto na predição de cada emoção, utilizando a métrica de revocação. As taxas de acerto por emoção, bem como uma legenda de qual linha/coluna da matriz ela corresponde estão mostradas na Tabela 2.

Linha/Coluna	**Emoção**	**Acertos (Revocação)**
0	Feliz-por	0%
1	Alegria	20%
2	Esperança	6.67%
3	Satisfação	40%
4	Alívio	20%
5	Orgulho	50%
6	Gratificação	10%
7	Gratidão	10%
8	Admiração	13.33%
9	Amor	25%
10	Pena	18.75%
11	Tristeza	0%
12	Medo	0%
13	Ressentimento	30%
14	Medos confirmados	6.67%
15	Vergonha	6.67%
16	Censura	40%
17	Remorso	13.33%
18	Regojizo	7.14%
19	Desapontamento	40%
20	Nojo	0%
21	Raiva	13.33%

Tabela 2: 22 emoções da matriz de confusão.

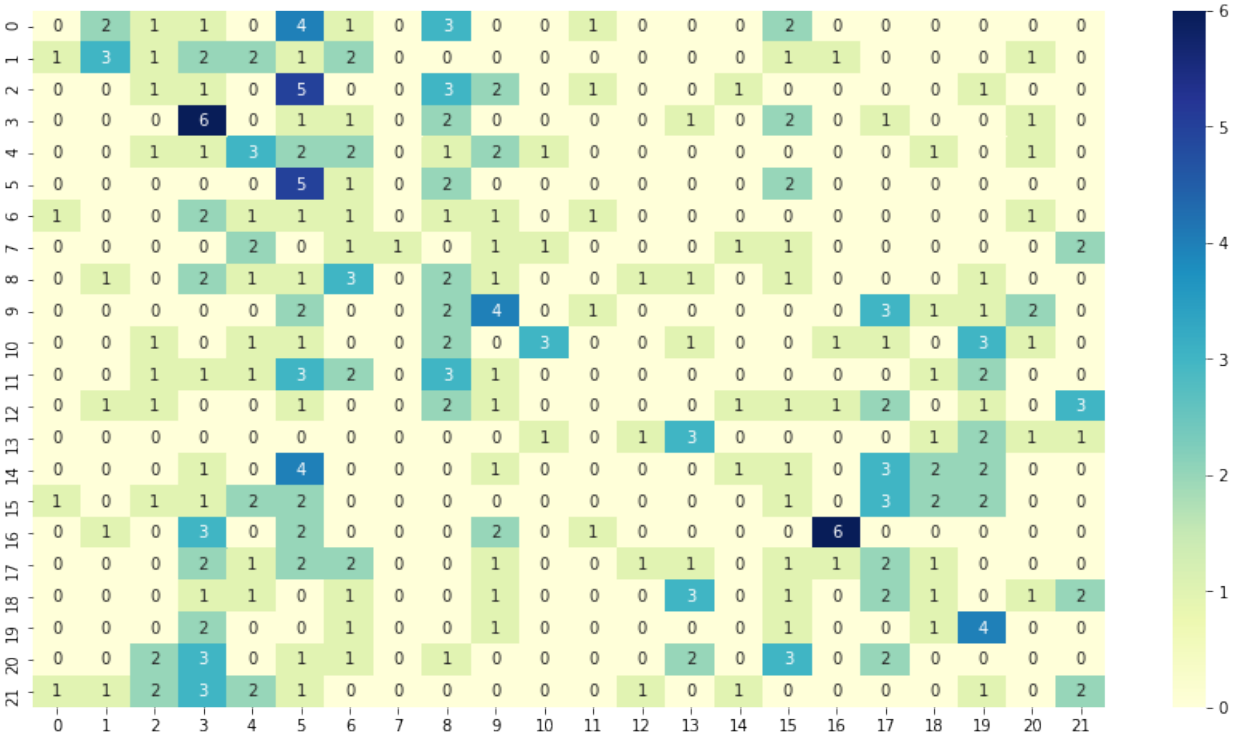

Figura 6: Matriz de confusão modelo KNN de 22 emoções.

4.2 Configuração 2 - Redução do modelo OCC em 7 emoções

Dado que as métricas obtidas na Configuração 1 foram pequenas, se tornou interessante reduzir o número de emoções e analisar como os classificadores utilizados se comportam nesse novo evento. Para isso foram utilizados as emoções reduzidas conforme trabalho de Dornhofer [3], cujo banco de emoções é o mesmo utilizado nesse trabalho.

O trabalho de Dornhofer reduz as 22 emoções para apenas 7 utilizando técnicas de clusterização em hierarquia. Nessa técnica, emoções com características semelhantes passam a se agrupar em um único cluster. As emoções reduzidas podem ser vistas conforme o dendrograma da Figura 7.

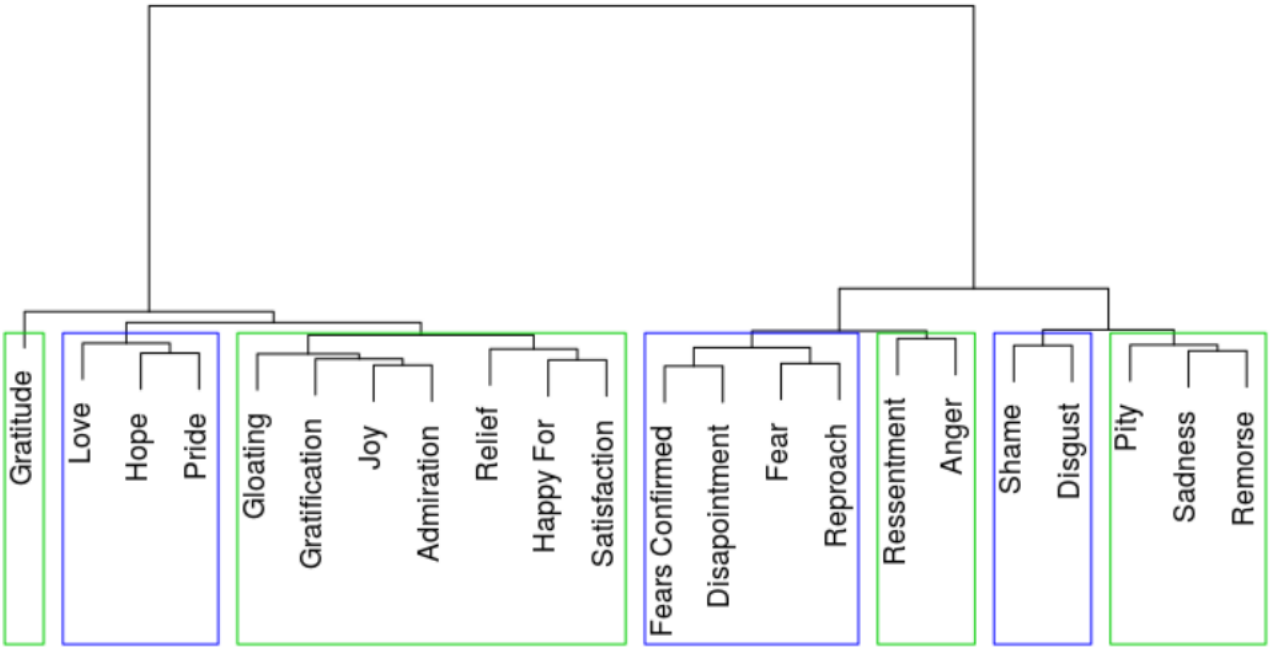

Figura 7: Dendrograma, reduzindo 22 emoções em 7. Extraída de [3].

O interessante em utilizar a redução das emoções nesse trabalho é que além de testar a hipótese de que um número menor de emoções com características mais distintas entre elas podem levar a uma melhor predição, o agrupamento das emoções foi realizado utilizando atributos extraídos das imagens dos atores. Assim, uma melhora nos resultados pode também indicar uma relação entre as características expressas na imagem e no áudio.

A partir dessa redução, aplicando os algoritmos de classificação no banco de emoções, foram obtidas as métricas da Tabela 3. Da mesma forma que anteriormente, também foi criada a matriz de confusão para o algoritmo de classificação K-NN, e criada uma legenda das emoções com suas taxas de acerto.

Modelo	Acurácia (%)	Precisão (%)	Revocação (%)	F1-score (%)
KNN	37.24 +/- 7.64	39.40 +/- 11.27	37.24 +/- 7.64	33.81 +/- 8.23
SVM	34.81 +/- 8.04	41.48 +/- 20.09	34.81 +/- 8.04	30.20 +/- 10.92
Random Forest	36.91 +/- 6.59	27.63 +/- 11.08	36.91 +/- 6.59	27.02 +/- 8.71

Tabela 3: Configuração 2 - Validação Cruzada por Ator - Métricas

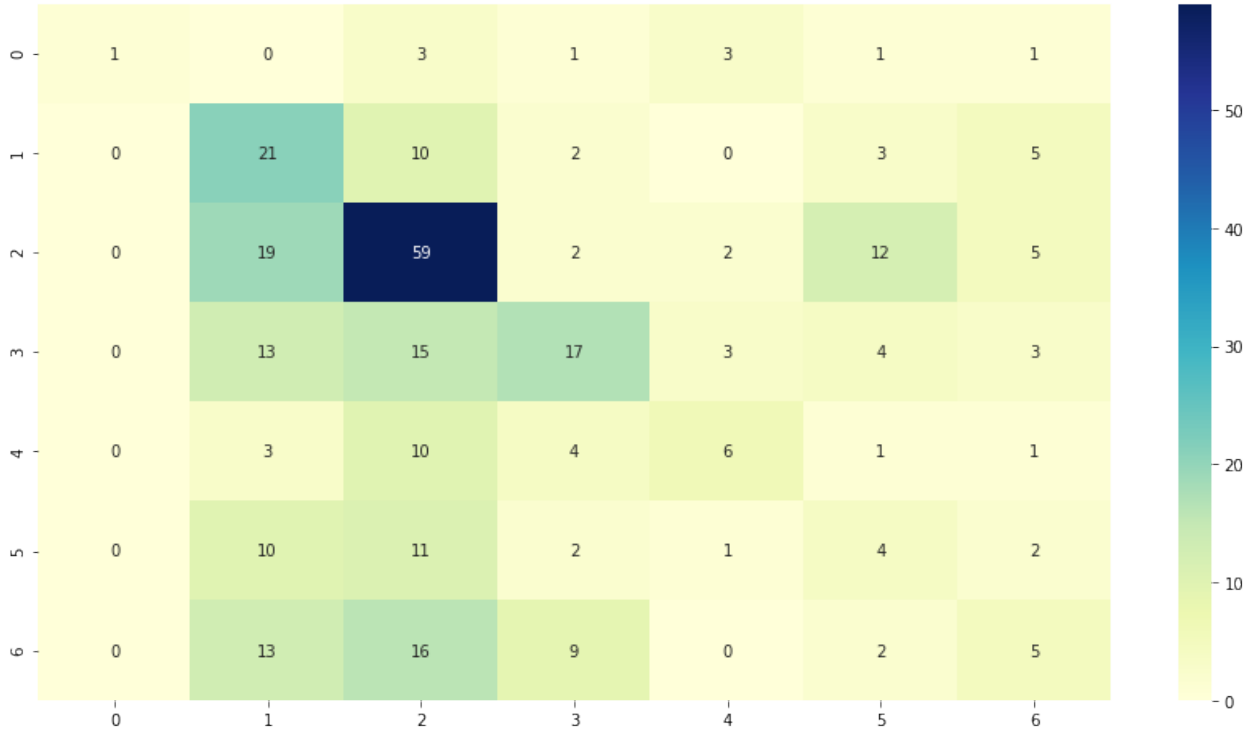

Figura 8: Matriz de confusão modelo KNN de 7 emoções.

Linha/Coluna	**Emoção**	**Acertos (Revocação)**
0	Gratidão	10%
1	Amor/Esperança/Orgulho	51.22%
2	Regojizo/Gratificação/Alegria/Admiração/Alívio/Feliz-por/Satisfação	59.60%
3	Medos confirmados/Desapontamento/Medo/Censura	30.91%
4	Ressentimento/Raiva	24%
5	Vergonha/Nojo	13.33%
6	Pena/Tristeza/Remorso	11.11%

Tabela 4: Legenda 7 emoções da matriz de confusão.

5 Discussão

Os resultados obtidos nesse trabalho trazem informações relevantes sobre o identificação da emoção através da fala.

A príncipio percebe-se na primeira configuração, que as métricas das predições para as 22 emoções foram baixas. Todavia, ao considerar um preditor que classifica aleatóriamente as emoções, sua taxa de acerto estaria em torno de $1 \div 22$, ou seja, 4,55%. Assim, podemos inferir a existência de uma relação entre as emoções e as características retiradas da fala.

Ainda, a quantidade de dados utilizados nesse trabalho foram pequenas, apenas 389 falas para classificar 22 emoções, podendo não ter sido o suficiente para o treino dos classificadores. Em situações semelhantes em aprendizado de máquinas, uma maior quantidade de dados resultam em melhores predições e menor variância.

Além da necessidade de uma maior quantidade, os dados coletados foram feitos por atores expressando as emoções baseado em seus próprios entendimentos e experiência de como elas se manifestam, sem necessariamente estarem sentido-as. Isso fez com que a mesma emoção fosse expressa de maneiras distintas, dificultando os classificadores de encontrarem características semelhantes para a mesma emoção.

A matriz de confusão da primeira configuração mostra que as emoções que geraram mais acertos foram as de orgulho, satisfação, censura e desapontamento, com taxas maiores ou iguais a 40%, enquanto nas emoções de feliz-por, tristeza, medo e nojo o classificador não obteve nenhum acerto.

Embora há variações no desempenho do classificador para cada emoção, ao olhar a matriz de confusão percebe-se uma certa dispersão entre as predições erradas. Logo, os atributos escolhidos na extração possuem características distintas e não há sobreposição de um atributo sobre os demais, sendo uma consequência positiva da normalização dos atributos para variância unitária e média zero.

Analogamente, a redução das 22 emoções em 7 agrupamentos de emoções, baseadas no artigo da Dornhofer [3], tiveram resultados distintos. A Tabela 2 indica um aumento considerável nas predições para apenas 7 emoções. Entretanto, vale alientar que uma classificação aleatória nesse caso obteria uma precisão de acertos em torno de $1 \div 7$, ou seja, 14.3%, taxa maior que a para 22 emoções. Se compararmos cada configuração com o classificador aleatório correspondente, a primeira configuração com 22 emoções possui uma acurácia 4.07 vezes maior, enquanto a segunda configuração com 7 grupos de emoções possui uma acurácia 2.61 vezes maior.

Pela matriz de confusão da configuração 2, percebe-se que os agrupamentos com maiores quantidades de emoção obtiveram um aumento considerável de acertos, o que mostra que as emoções

agrupadas através da imagem também possuem atributos semelhantes na fala, não descartando a possibilidade de uma relação entre a fala e a imagem de uma pessoa ao expressar a emoção.

6 Conclusão

O trabalho de fim de curso teve como objetivo a implementação de um método de identificação e classificação de emoções a partir da análise do timbre da voz, tendo como base estudos em computação afetiva. Para isso, foram aplicadas técnicas de processamento digital de sinais e de aprendizado de máquina relacionando problema de classicação de vetores em espaços vetoriais.

Como base de dados para as emoções, um total de 7 atores e atrizes gravaram vídeos expressando 22 emoções diferentes, baseadas no modelo OCC de emoções, sendo elas: feliz-por, alegria, esperança, satisfação, alívio, orgulho, gratificação, gratidão, admiração, amor, pena, tristeza, medo, ressentimento, medos confirmados, vergonha, censura, remorso, regojizo, desapontamento, nojo e raiva.

Técnicas de processamento digital de sinais foram utilizadas para extrair atributos dos áudios e foram utilizados três modelos de classificação supervisionada em aprendizado de máquina para identificar as emoções, sendo eles o K-Nearest Neighbors, Suport Vector Machine e o Random Forest. Esses modelos de classificação foram treinados e testados por meio de uma validação cruzada entre os atores, no qual classificou-se as emoções das diferentes falas de cada ator, utilizando como treino para a predição os dados dos atores restantes.

Os resultados do trabalho mostraram a presença de relação entre a emoção e a fala de uma pessoa, onde os classificadores para as 22 emoções obtiveram uma acurácia em torno de 18.52%, 4.07 vezes maior que ao se tentar classificá-las aleatoriamente, e para uma redução das 22 emoções em 7 agrupamentos de emoções obteve-se acurácia próxima de 37.24%, 2.61 vezes maior que um classificador aleatório.

Referências

[1] Rajneet Kaur Aastha Joshi. A Study of Speech Emotion Recognition Methods. *International Journal of Computer Science and Mobile Computing*, 2:28–31, 2013.

[2] Leo Breiman. Random forests. *Machine Learning*, 2001.

[3] Paula Dornhofer Paro Costa. *Two-dimensional Expressive Speech Animation*. PhD thesis, School of Electrical and Computer Engineering, University of Campinas (Unicamp), February 2015.

[4] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[5] Min-Wei Huang, Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, and Chih-Fong Tsai. SVM and SVM Ensembles in Breast Cancer Prediction. 2017.

[6] Nangyeon Lim. Cultural differences in emotion: differences in emotional arousal level between the East and the West. *Integrative Medicine Research*, pages 105–109, 2016.

[7] A. Ortony, G. Clore, and A. Collins. *Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[8] Rosalind Wright Picard. Affective computing. 1995.

[9] V. Vapnik. Statistical Learning Theory. 1998.