INSTITUTO FEDERAL DE EDUCAÇÃO CIÊNCIA E TECNOLOGIA DA PARAÍBA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

IURY ANDERSON FERNANDES COELHO

Análise da relevância dos recursos de sinal de voz para classificação de sentimento usando Redes Neurais Multilayer Perceptron (MLP)

CAMPINA GRANDE - PB
2021

IURY ANDERSON FERNANDES COELHO

Análise da relevância dos recursos de sinal de voz para classificação de sentimento usando Redes Neurais Multilayer Perceptron (MLP)

Artigo apresentado ao Instituto Federal de Educação Ciência e Tecnologia da Paraíba, como parte dos requisitos para obtenção do título de bacharel em Engenharia de Computação.

Orientador: Prof. Paulo Ribeiro Lins Júnior, Dr.

CAMPINA GRANDE - PB

2021

C672a Coelho, Iury Anderson Fernandes.
Análise da relevância dos recursos de sinal de voz para classificação de sentimento usando Redes Neurais Multilayer Perceptron (MLP). / Iury Anderson Fernandes Coelho. - Campina Grande, 2022.
18f. : il.

Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Engenharia de Computação) - Instituto Federal da Paraíba, 2022.
Orientador: Prof. Dr. Paulo Ribeiro Lins Júnior.

1. Engenharia de computação. 2. Rede neural - MultiLayer perceptron (MLP). 3. Reconhecimento de voz - Classificação de emoções. I. Lins Júnior, Paulo Ribeiro. II. Título.

CDU 004.5

RESUMO

As emoções são aspectos de grande impacto no entendimento das relações humanas e a pesquisa moderna está tentando encontrar métodos que possam dar à máquina a capacidade de reconhecer as emoções. Neste artigo, é descrita uma comparação da classificação de emoções realizadas por um modelo de rede neural do tipo MultiLayer Perceptron utilizando recursos de voz extraídos de áudios gravados no idioma português.

Palavras-chave: Classificação de emoções: SER, Reconhecimento automático da fala, idioma português; Técnicas de Machine Learning: Desempenho, MLP; Extração de recursos: MFCC, Zero Cross Rate, Cromograma.

SUMÁRIO

LISTA DE FIGURAS

Projeto de Engenharia de Computação II (2021), **2**, 1–15

Instituto Federal de Tecnologia da Paraíba

ARTICLE

Análise da relevância dos recursos de sinal de voz para classificação de sentimento usando Redes Neurais Multilayer Perceptron (MLP)

Paulo Junior, Iury Coelho*

Grupo de Pesquisa em Comunicações e Processamento de Informação - GComPI
*Correspondencia do Autor. Email: iury.fernandes@academico.ifpb.edu.br

(Iniciado em Julho de 2021; Primeira a submissão online 12 de Setembro de 2021)

Abstract

Neste artigo, é descrita uma comparação da classificação de emoções realizadas por um modelo de rede neural do tipo MultiLayer Perceptron (MLP), utilizado recursos de voz extraídos de áudios gravados no idioma português e disponíveis no banco de dados VERBO. É proposto o índice FRI4SA, que mede a relevância a partir da variabilidade da similaridade de uma determinada medida para os diversos sentimentos analisados. Observou-se, de acordo com a métrica considerada, que o gênero do interlocutor da amostra de voz usada para analisar a relevância de uma característica impacta de forma muito sensível os valores obtidos. Um modelo de classificação usando MLP é desenvolvido a partir da entrada de diferentes recursos de voz combinados entre si de maneira que a atuação do modelo de classificação procede em duas etapas, a etapa de treino, que usa as diferentes combinações de recursos como entrada, e a fase de testes onde é comparado o desempenho de classificação em termos de acurácia e precisão. O aumento da acurácia se dá pela quantidade e combinação de recursos utilizados. Observou-se que a classificação usando a combinação dos recursos de MFCC e ZCR tem uma pontuação da acurácia de 88% e é maior 9% do que a acurácia para recursos que usam apenas MFCC. A principal contribuição do trabalho é analisar relevância de recursos que podem prover melhor compreensão da máquina no reconhecimento automático de emoções na voz.

Keywords: Classificação de emoções: SER, Reconhecimento automático da fala, idioma português; Banco de dados: VERBO; Técnicas de Machine Learning: Desempenho, MLP; Extração de recursos: MFCC, Zero Cross Rate, Cromograma.

1. Introdução

As emoções são aspectos de grande impacto no entendimento das relações humanas, sendo consideradas também como um dos principais fatores psicológicos para o sucesso das organizações. A pesquisa moderna está tentando encontrar métodos que possam dar a máquina capacidade de reconhecer emoções. Essa área de pesquisa está concentrada no ramo da computação afetiva, uma tentativa de propiciar uma relação mais próxima entre computadores e humanos (Langari, Marvi e Zahedi 2020). Reconhecer emoções na voz de forma automática, tem vários benefícios como por exemplo compreender a interação entre médicos e pacientes (Li et al. 2021), criação de agentes inteligentes para avaliar a satisfação de clientes em ambiantes de call-center (Parra-Gallego e Orozco-Arroyave 2021), detecção de estresse através de análise de sinais de voz (Hilmy et al. 2021).

A computação afetiva, que tem motivado um número significativo de trabalhos nos últimos anos, sendo trabalhado com as mais diferentes fontes de informação, como textos, sinais de áudio, de voz, de vídeo ou arranjos desses (Kaur e Kautish abril de 2019). O SER (*Speech Recognize Emotion* - Reconhecimento de Emoções na Fala) é uma área especifica dentro da computação afetiva que usa sinal de voz apra o reconhecimento de emoções. Segundo (Ho, Vuong et al. 2021) um número total de 1.646 trabalhos realcionados a computação efetiva foram produzidos de 1995 a 2020 sendo esse tópico um dos temas mais interdisciplinares e que tem mostrado uma taxa global de crescimento anual da produção científica de 12,5%. Estudos recentes de revisão mostram que as principais tendências para o futuro, no que diz respeito a classificação computacional da emoção na fala, seguem as seguintes vertentes: concepção de banco de dados que melhor se adequem aos descritores para classificação, relevância de recursos de sinais de áudio que podem prover melhor compreensão da máquina e métodos de aprendizagem de máquina para o desenvolvimento de algorítimos mais eficientes (Schuller e Schuller 2021).

Em pesquisas recentes usou-se áudios no idioma inglês britânico para extraído recursos de Coeficientes Cepstrais de Frequência de Mel (MFCC) a fim de provocar o melhoramento da classificação modelos de redes neurais. Já (Koduru, Valiveti e Budati 2020), utilza-se de coeficientes cepstral e Zero Crossing Rate (ZCR) para aperfeiçoar o reconhecimento de emoções de fala de um sistema usando os diferentes algoritmos de extração de características. Este artigo traz contribuições ao problema de seleção de características para a análise de sentimentos em sinais de voz, importante para a construção de métodos de classificação mais eficientes e mais precisos, uma vez que a quantidadede características consideradas impacta diretamente no custo computacional e na eficiência do modelo (Sharma, Umapathy e Krishnan 2020) (Farooq et al. 2020).

Embora muitos estudos tenham sido feitos no campo do SER, não foram encontrados pesquisas que utilizam sinais de voz no idioma português brasileiro. Neste estudo, utilizou-se sinais de voz oriundos do conjunto de dados VERBO, um banco de dados de áudios feito no idioma português, composto por amostras de sinais de voz associados a diferentes emoções: felicidade, desgosto, medo, neutralidade, raiva, surpresa e tristeza. Ainda, é proposto o Índice de Relevância de Característica para Análise de Sentimentos (*Relevance Index for Sentiment Analysis – FRI4SA*), uma nova métrica de seleção que mensura a relevância de uma característica, com base na sua similaridade quando usada para todos os sentimentos considerados no estudo. A motivação para o uso de similaridade como forma de avaliar a relevância de características de sinais de voz vem do emprego recorrente desse tipo de medida na avaliação da relevância de termo em documentos textuais(Chandrasekaran e Mago 2021). Depois de calculado o índice FRI4SA para alguns dados do verbo, validou-se as relevâncias de recursos aplicando-os como entrada de redes neurais MLP. Os recursos de coeficientes cepstrais de frequência de Mel (MFCCs), Zero Cross e Cromagrama (Cg) são usados para pré-processamento dos dados de treinamento, testes e validação de experimentos de aprendizagem supervisionada com um modelo de classificação. Na etapa de treinamento do modelo usou-se diferentes combinações de recursos e foi feito a comparação do desempenho do classificador em termos de acurácia e precisão para cada combinação de recursos.

2. Desenvolvimento

A estratégia de desenvolvimento para este estudo pode ser visualizada em três etapas através do fluxograma detalhado na Figura 1.

A primeira etapa: extração de recursos, diz respeito à coleta e preprocessamento de dados de sinais de aúdios da fala retirados do banco de dados VERBO. Após a extração de recursos, as informações preprocessadas são combinadas e armazenados em data sets. Na segunda etapa: treinamento e testes, os dados armazenados nos data sets são inseridos como entradas no modelo de rede neural MLP. A atuação do modelo procede usando diferentes combinações de recursos armazenados nos data sets como entrada para treinamento. Em seguida, são realizados testes de desempenho na classificação das

emoções. A terceira etapa: avaliação de resultados, consiste em avaliar a relevância de recursos que podem prover melhor compreensão da máquina no reconhecimento automático de emoções vocais usando redes neurais MLP.

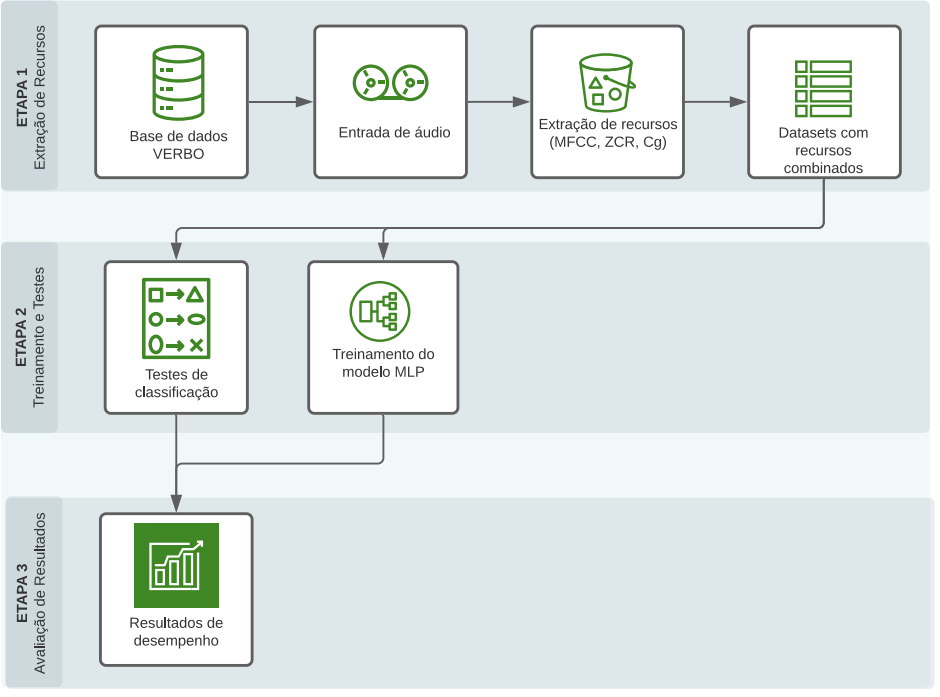

Figura 1. Fluxo detalhado de Desenvolvimento

3. Métodos

3.1 Banco de dados de emoções vocais

O progresso na área do SER está ligado ao desenvolvimento de bancos de dados apropriados. Os trabalhos voltados para este campo, estão essencialmente preocupados em fornecer dados confiáveis que possam descrever emoções na fala (Douglas-Cowie et al. 2003). Neste trabalho, o banco de dados VERBO (Neto et al. 2018) foi escolhido para compor os estudos de pesquisa. Essa base de dados é formada por 1176 gravações de áudio no formato .WAV consolidados em amostras de voz no idioma português brasileiro, referentes a sete diferentes emoções: alegria, desgosto, medo, neutro, raiva, surpresa e tristeza. São 14 sentenças, categorizadas como "longa", "sem sentido", "questão" e "curta", que podem ter de 2 à 5 segundos, lidas por 12 atores – 6 mulheres e 6 homens, para cada um dos sete emoções considerados, o que faz desse conjunto de dados parte importante do treinamento de um sistema de classificação que considere as peculiaridades idiomáticas do Brasil. Este estudo considerou, inicialmente, algumas amostras de áudios para analisar a relevância das características e em uma outra etapa, todos as amostras de áudios emotivos para treinar e testar o classificador de emoções. Na Figura 2 é possível perceber uma representação em porcentagem da quantidade de áudios emotivos usados neste trabalho.

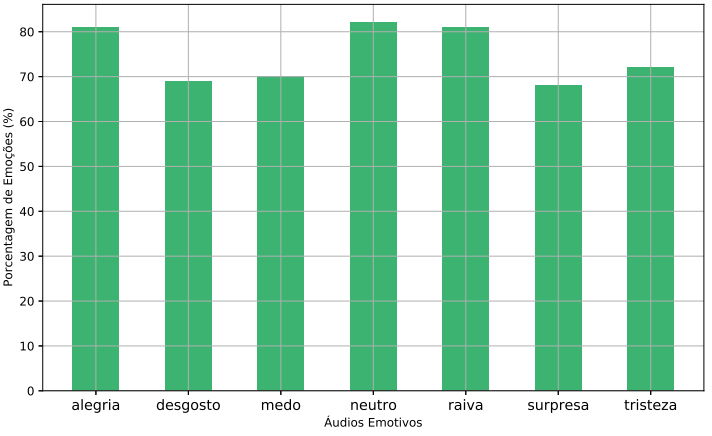

Figura 2. Porcentagem da quantidade de áudios emotivos no VERBO

3.2 Extração de Recursos

O Processo de extração de recursos, no âmbito do reconhecimento de emoções através da voz, é considerado um dos métodos mais importantes e tem o intuito de produzir parâmetros de sinais de voz que são mais adequados para os modelos de classificação (Langari, Marvi e Zahedi 2020). Basicamente esse procedimento tenta prover uma representação paramétrica que seja capazes de dar a máquina a imitação do aparelho auditivo humano (Tiwari 2010). Existem várias características de áudios que podem ser extraídas da voz, sendo que, algumas delas são mais relevantes no desenvolvimento para analisar os sinais de áudios de voz (Alıas, Socoró e Sevillano 2016). Como características que podem ser extraídas de sinais de áudios vocais podemos citar: taxa de cruzamento no zero (ZCR): mede o número de vezes que a forma de onda do sinal muda de sinal no decorrer da janela de tempo considerada; valor médio quadrático da Energia (rmse): raiz do valor médio quadrático da energia do sinal analisado; contraste (ctt):medida referente ao contraste do sinal de voz, sendo útil para destacar pequenas diferenças fonéticas, sendo uma medida indireta de inteligibilidade entre os interlocutores; centroide espectral (sc): medida do centro de massa do espectro do sinal, sendo bastante utilizado, no contexto de análise de áudio, como medida associada ao timbre; Coeficientes de Frequência Mel-Cepstral (MFCC): características motivadas pela percepção que fornecem uma representação compacta do envelope do espectro de tempo curto. cromagrama (cg): representação do croma de um sinal de áudio em função do tempo; rolloff espectral (sr): frequência sob a qual o corte da energia total do espectro está contido, podendo ser usado para distinguir entre sons harmônicos e barulhentos; O uso de MFCC's tem se tornado uma tendencia em muitos métodos implementados para reconhecimento de emoções na fala (Sharma, Umapathy e Krishnan 2020). Apesar disso, algoritmos de energia e taxa de cruzamento zero (ZCR) também são usados como parâmetros como uma alternativa de melhoramento no reconhecimento de emoções de fala (Koduru, Valiveti e Budati 2020). Algumas pesquisam estão mais focadas em descobrir dentre um conjunto de recursos de áudio, quais características apresentam uma melhor performance. Essa abordagem geralmente é feita calculando uma analogia ou similaridade dos seguimentos da fala, bem como experimentos que tentam descobrir qual a quantidade ideal de recursos característicos da fala, já que isso impacta diretamente no custo computacional e na eficiência do classificador (Ezz-Eldin et al. 2021). Conforme (Kumar e Mahajan 2019) normalmente 20 MFCC's são usados e esse número tem se mostraram suficientes para os classificadores. Nesta pesquisa usa-se um numero variado de

MFCC's combinados os recursos de Zero Cross e Cromograma.

A extração de características dessa pesquisa foi possível graças a biblioteca Librosa (McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, 2020), escrita em Python e especializada no processamento de música e sinais de áudio em geral. Neste trabalho é proposto o Índice de Relevância de Característica para Análise de Sentimentos (Feature Relevance Index for Sentiment Analysis – FRI4SA), como método para comparar a similaridade entre as medidas de uma dada característica para os diferentes sinais de áudios do VERBO. Basicamente, o índice proposto compara a similaridade entre as medidas de uma dada característica para os diferentes sentimentos analisados, pressupondo que a relevância dessa característica é diretamente proporcional a variabilidade da similaridade entre os diferentes sentimentos. Dessa forma, quanto maior a variabilidade da similaridade, mais interessante para o sucesso de um classificador na tarefa de predição de sentimentos do sinal analisado. Seja S um vetor contendo as medidas de similaridade par a par entre todas as emoções

$$\mathbf{S} = [s_1, s_2, s_3, \ldots, s_m] \tag{1}$$

em que $m = \frac{n(n-1)}{2}$ é o número de medidas obtidas para n sentimentos considerados. O Índice de Relevância de Característica para Análise de Sentimentos (*Feature Relevance Index for Sentiment Analysis* – FIR4SA) é definido pela função:

$$\mathrm{FIR4SA}(\mathbf{S}) = C_m \cdot \left| \frac{\underset{1 \leq i \leq m}{\mathrm{med}} \left(|s_i - \tilde{\mu}| \right)}{\tilde{\mu}} \right| \tag{2}$$

em que $\tilde{\mu}$ é a mediana dos elementos do vetor de similaridades **S**, med(·) é a função que calcula a mediana do argumento (a notação da mediana foi separada para facilitação do entendimento), C_m é um fator de escala, cujo valor depende do número de elementos do vetor considerado e é tabelado em alguns trabalhos na literatura (Park, Kim e Wang 2020).

O FRI4SA é baseado no cálculo de um coeficiente de variação, medida relativa que permite a comparação da variabilidade dos dados de dois ou mais conjuntos de dados diferentes. Tipicamente, coeficientes de variação são calculados com base na média e variância (ou desvio padrão), porém essas são medidas susceptíveis à valores espúrios. Para evitar essa limitação, optou-se por trabalhar com uma versão mais robusta, usando desvio mádio absoluto e mediana para o cálculo da variabilidade dos dados (Arachchige, Prendergast e Staudte 2020)(Park, Kim e Wang 2020) (Ospina e Marmolejo-Ramos 2019). Além disso, optou-se por trabalhar unicamente com a magnitude dessa variabilidade relativa, por atender melhor ao critério de decisão estabelecido pelo índice.

3.3 Classificação de emoções usando redes neurais MLP

A etapa de classificação de emoções é o intuito principal do modelo de aprendizado de máquina (MLP) desenvolvido neste artigo. O MLP ou Perceptron de Multicamadas é uma rede neural artificial que se propaga em direção única, ou seja, o erro ocorrido, é calculado na camada de saída e recalculado a partir do valor dos pesos da última camada de neurônios (Nazzal, El-Emary e Najim 2008). O MLP que é constituído por um conjunto de métodos de aprendizado supervisionado e é capaz de solucionar problemas de não linearidade de forma que é possível especificar dados que não tem um comportamento linear (Qiao, Khishe e Ravakhah 2021). Esse modelo de aprendizado de máquina, utiliza vários neurônios conectado dispostos em camadas. Existe uma camada de entrada e outra(s) oculta(s) interligadas a uma camada de saída. O MLP utiliza a técnica de *backpropagation* para treinamento, que permite a atualização de valores encontrados que produzem os melhores desempenho de classificação(Jin, Liu e Long 2021). Neste artigo, o processo de classificação é composto por duas

etapas essenciais: etapa de treino, que é feita para inicializar o modelo de classificação imputando diferentes combinações de recursos de voz, e a etapa de testes da performance das classificações, que compões os métodos avaliativos das métricas de desempenho do classificador.A separação dos conjuntos dados de teste e de treino foi feita de forma que 10% dos dados do VERBO foram separados aleatoriamente e utilizados para a fase de testes. Durante a fase de treinamento variou-se os recursos de entradas afim de analisar a relevância no desempenho para cada entrada especificada. O estágio de treinamento usa data sets contendo as combinações de recursos e produzir um bom padrão de classificação. A fase de teste é realizada par analisar a melhor configuração de entradas que seja capaz de fornecer os resultados de classificação mais precisos. Para a concepção do modelo MLP usou-se a biblioteca python Sklearn seguindo as configurações paramétricas geradas pela biblioteca e que podem ser vistas na Tabela 1

Tabela 1. Hiperparâmetros do classificador de emoção MLP

Parâmetro	Valor
hidden_layer_sizes	300
batch_size	256
alpha	0.01
learning_rate	0.3
max_iter	600

4. Resultados

Os resultados obtidos no estudo de avaliação do reconhecimento automático de emoções pode ser observado sob dois aspectos: avaliação da relevância de características de sinais de voz para classificação de sentimentos a partir do índice FRI4SA, que visa entender o quanto um recurso tem impacto na classificação de emoções usando dados do VERBO, e avaliação do desempenho de reconhecimento obtidos para a classificação de emoções udando MLP utilizando diferentes recursos de entrada: que tem avalia a combinação de recursos para produzir resultados de maior desempenho na classificação.

4.1 Relevância de Recursos

Como caso de uso para avaliar a funcionalidade da medida apresentada nesse trabalho, o presente trabalho estuda a relevância de características a partir da análise de amostras de voz disponíveis no conjunto de dados VERBO (Neto et al. 2018), composto por amostras de voz em português brasileiro, referentes a sete diferentes emoções: alegria, desgosto,medo, neutro, raiva, surpresa e tristeza. Para o experimento desse trabalho, foram analisadas amostras referentes a quatro locutores, sendo duas mulheres (identificadas como F1 e F2), e dois homens (identificados como M1 e M2), lendo três frases distintas: uma frase longa, "Agora vou pôr a camiseta e sair para uma caminhada", identificada como l3, uma questão, "Sábado à noite, o que vai fazer?", identificada como q1 e uma frase curta, "Os operários levantam cedo", identificada como s1. Essas frases foram selecionadas aleatoriamente de cada uma das categorias disponíveis no conjunto de dados. Para o cálculo do índice FRI4SA, adotou-se o valor do fator de escala C_m = 1.540681, obtido a partir das tabelas dispostas em (Park, Kim e Wang 2020), para m = 21. *l3* é a taxa de cruzamento por zero, para *q1* é o MFCC, enquanto para *s1* é o cromagrama. A única unanimidade entre todas as frases é que o contraste e o *rolloff* espectral são as medidas de menor relevância.Os valores dos índices FRI4SA para cada sentimento e para cada uma das amostras foram calculados e são sumarizados nos gráficos mostrados nas Figuras 3 à 6 e Tabelas de 2 à 5. Quando analisamos da segunda posição em diante desse ordenamento, todas diferem. Por exemplo, a segunda característica mais relevante para *l3* é a taxa de cruzamento por

zero, para *q1* é o MFCC, enquanto para *s1* é o cromagrama. A única unanimidade entre todas as frases é que o contraste e o *rolloff* espectral são as medidas de menor relevância.

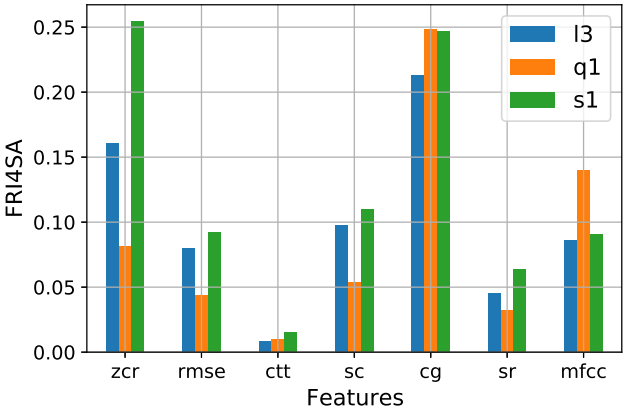

Figura 3. FRI4SA de cada característica analisada para a amostra de voz F1.

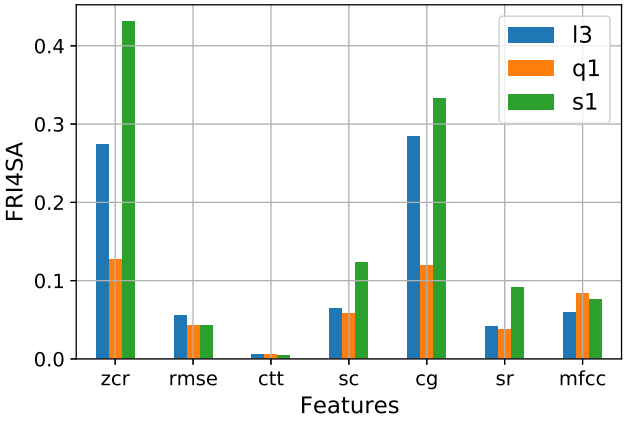

Figura 4. FRI4SA de cada característica analisada para a amostra de voz F2.

Tabela 2. Valores do índice FRI4SA da amostra F1 para os sentimentos analisados.

	l3	q1	s1
zcr	0.160330	0.081362	0.254490
rmse	0.079513	0.043800	0.092354
ctt	0.008418	0.009603	0.014913
sc	0.097692	0.054012	0.109633
cg	0.213042	0.248475	0.246658
sr	0.045314	0.031892	0.063908
mfcc	0.085914	0.139510	0.090950

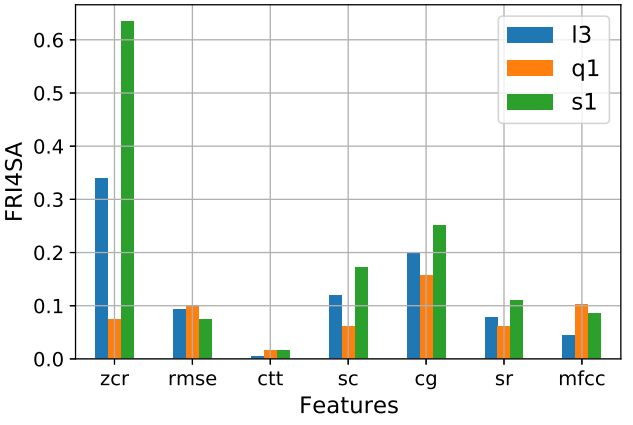

Figura 5. FRI4SA de cada característica analisada para a amostra de voz M1.

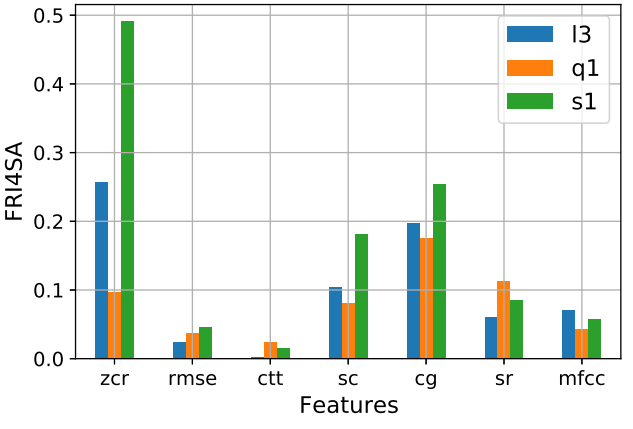

Figura 6. FRI4SA de cada característica analisada para a amostra de voz M2.

Tabela 3. Valores do índice FRI4SA da amostra F2 para os sentimentos analisados.

	l3	q1	s1
zcr	0.274616	0.127509	0.430843
rmse	0.055829	0.043310	0.042454
ctt	0.005268	0.005195	0.005026
sc	0.064682	0.057492	0.122867
cg	0.284090	0.119078	0.332169
sr	0.041314	0.038172	0.091309
mfcc	0.059059	0.083827	0.076636

Tabela 4. Valores do índice FRI4SA da amostra M1 para os sentimentos analisados.

	l3	q1	s1
zcr	0.339203	0.074843	0.634203
rmse	0.092956	0.101379	0.075183
ctt	0.004855	0.016010	0.015769
sc	0.118690	0.061530	0.171368
cg	0.200088	0.157480	0.250352
sr	0.077116	0.061766	0.109604
mfcc	0.044666	0.102101	0.086305

Tabela 5. Valores do índice FRI4SA da amostra M2 para os sentimentos analisados

	l3	q1	s1
zcr	0.257400	0.096296	0.491003
rmse	0.023936	0.036614	0.045568
ctt	0.002820	0.024593	0.015181
sc	0.103949	0.080595	0.181311
cg	0.197098	0.176078	0.253533
sr	0.060409	0.112176	0.085276
mfcc	0.070596	0.042513	0.056826

4.2 Classificação de Emoções

A pesquisa da classificação de emoções usa 1176 dados da fala emocional com 81% dos dados representados pela emoção alegria, 69% desgosto, 70% neutro, 82% medo, 81% raiva, 68 % tristeza e 72% surpresa. Cada áudio emotivo foi preprocessado usando os recursos de de MFCC, ZCR, Cg e armazenados em data sets enumerados de 1 à 6. Na tabela abaixo é possível visualizar os data sets e os recursos armazenados. Para entrada de dados de treinamento do modelo usou as diferentes combinações de recursos a fim de obter diferentes desempenhos. A partir do modelo treinado analisou-se os resultados de desempenho do classificador através dos testes t1 até t6 utilizando as métricas de acurácia, precisão, recall e f1-escore. Os resultados dos testes de classificação bem como os recursos de voz utilizados podem ser visualizados nas Tabelas de 7 à 13. Nos gráfico da Figura 7 e Figura 9 é mostrado respectivamente a acurácia e a precisão do reconhecimento de emoções. Usando recursos de MFCC(40) e ZCR é possível ter uma acurácia de 88%, sendo essa a maior taxa de reconhecimento dentre todos os recursos testados. Ainda, usando recursos MFCC apenas, é possível obter uma precisão de 100% para as emoções de tristeza e surpresa. A combinação dos três recursos, MFCC, ZCR e Cg obteve precisão máxima para a emoção medo. Na Figura 8 pode-se visualizar matriz de confusão para a melhor combinação de recursos.

Tabela 6. Valores do índice FRI4SA da amostra F1 para os sentimentos analisados.

Dataset	Valores Pré-processados
D1	40 MFCCs
D2	40 MFCCs e ZCR
D3	40 MFCCs, ZCR e Cg
D4	60 MFCCs
D5	60 MFCCs e ZCR
D6	60 MFCCs, ZCR e Cg

Tabela 7. Resultado de testes usando a combinação 40 MFCCs

	Precision	Recall	F1-Score	Support
ALEGRIA	0.52	1.00	0.69	12
DESGOSTO	0.70	0.58	0.64	12
MEDO	0.79	0.83	0.81	18
NEUTRO	0.90	0.86	0.88	22
RAIVA	0.67	0.77	0.71	13
SURPRESA	1.00	0.53	0.70	15
TRISTEZA	1.00	0.84	0.91	25

5. Conclusão

Para a relevância de recursos, considerou-se o conjunto de dados VERBO, um conjunto de amostras de voz rotuladas por sete sentimentos distintos em português brasileiro. Selecionou-se duas amostras masculinas e duas amostras femininas do conjunto, e três frases para cada amostra. Observou-se, de acordo com a métrica considerada, que o gênero do interlocutor da amostra de voz usada para analisar a relevância de uma característica impacta de forma muito sensível os val-ores obtidos. De fato, observaram-se destacadas diferenças entre as amostras masculinas e femininas, além, também, de diferenças significativas entre as amostras femininas, o que indica que ter uma diversidade de interlocutores, inclusive de gênero, é importante para o treinamento de classificadores eficientes.Além disso, observou-se também que o tipo de sentença usado na amostra de voz exerce grande influência

Tabela 8. Resultado de testes usando a combinação 40 MFCCs, ZCR

	Precision	Recall	F1-Score	Support
ALEGRIA	0.79	0.92	0.85	12
DESGOSTO	0.69	0.75	0.72	12
MEDO	0.87	0.72	0.79	18
NEUTRO	0.94	0.73	0.82	22
RAIVA	0.86	0.92	0.89	13
SURPRESA	0.87	0.87	0.87	15
TRISTEZA	0.79	0.92	0.85	25

Tabela 9. Resultado de testes usando a combinação 40 MFCCs, ZCR e Cg

	Precision	Recall	F1-Score	Support
ALEGRIA	0.90	0.75	0.82	12
DESGOSTO	0.82	0.75	0.78	12
MEDO	1.00	0.89	0.94	18
NEUTRO	0.86	0.86	0.86	22
RAIVA	0.71	0.92	0.80	13
SURPRESA	0.92	0.73	0.81	15
TRISTEZA	0.79	0.92	0.85	25

Tabela 10. Resultado de testes usando a combinação 60 MFCCs

	Precision	Recall	F1-Score	Support
ALEGRIA	0.85	0.92	0.88	12
DESGOSTO	0.64	0.75	0.69	12
MEDO	0.89	0.89	0.89	18
NEUTRO	0.90	0.86	0.88	22
RAIVA	0.73	0.85	0.79	13
SURPRESA	0.85	0.73	0.79	15
TRISTEZA	0.96	0.88	0.92	25

Tabela 11. Resultado de testes usando a combinação 60 MFCCs e ZCR

	Precision	Recall	F1-Score	Support
ALEGRIA	0.85	0.92	0.88	12
DESGOSTO	0.73	0.92	0.81	12
MEDO	0.94	0.89	0.91	18
NEUTRO	0.94	0.73	0.82	22
RAIVA	0.86	0.92	0.89	13
SURPRESA	0.93	0.93	0.93	15
TRISTEZA	0.88	0.92	0.90	25

Tabela 12. Resultado de testes usando a combinação 60 MFCCs, ZCR e Cg

	Precision	Recall	F1-Score	Support
ALEGRIA	0.92	0.92	0.92	12
DESGOSTO	0.75	1.00	0.86	12
MEDO	0.94	0.89	0.91	18
NEUTRO	0.94	0.73	0.82	22
RAIVA	0.85	0.85	0.85	13
SURPRESA	0.87	0.87	0.87	15
TRISTEZA	0.81	0.88	0.85	25

Tabela 13. Acuraccy do classificar usando os diferentes testes

Features	MFCC(40)	MFCC(40), ZCR	MFCC(40), ZCR, Cg	MFCC(60)	MFCC(60), ZCR	MFCC(60), ZCR, Cg
Datasets	d1	d2	d3	d4	d5	d6
Testes	t1	t2	t3	t4	t5	t6
ALEGRIA	0.52	0.79	0.90	0.85	0.85	0.92
DESGOSTO	0.70	0.69	0.82	0.64	0.73	0.75
MEDO	0.79	0.87	1.00	0.89	0.94	0.94
NEUTRO	0.90	0.94	0.86	0.90	0.94	0.94
RAIVA	0.67	0.86	0.71	0.73	0.86	0.85
SURPRESA	1.00	0.87	0.92	0.85	0.93	0.87
TRISTEZA	1.00	0.79	0.79	0.96	0.88	0.81
ACURRACY	0.79	0.83	0.84	0.85	0.88	0.86

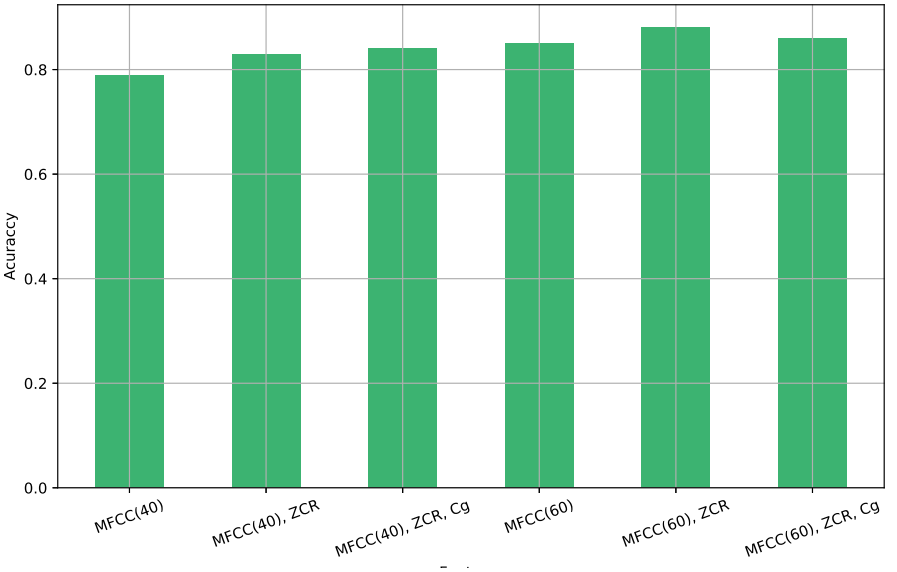

Figura 7. Acurácia do classificador usando diferentes combinações de recursos

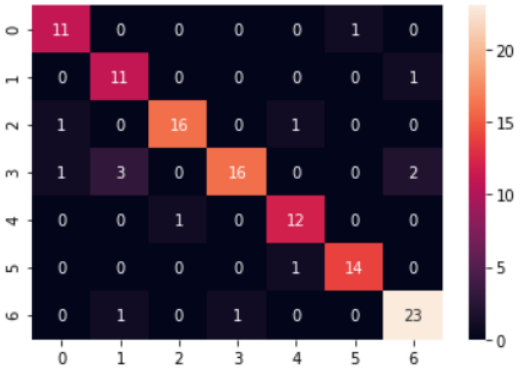

Figura 8. Matriz de confusão para a melhor combinação de recursos.

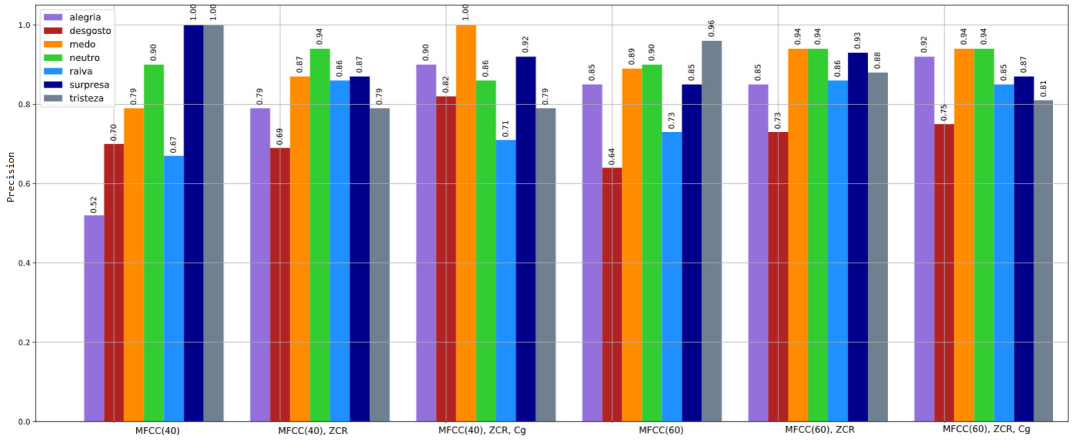

Figura 9. Precisão da combinação de recursos na classificação das emoções

na relevância de uma característica. O seu tamanho, se é uma pergunta ou uma afirmação, são fatores que se mostraram relevantes para as diferenças observadas.Com isso, a principal conclusão dessa análise é que é necessário considerar uma grande diversidade – de interlocutores, de gênero, de sentenças, para a seleção de características que possam impactar positivamente na eficiência de classificador de sentimentos a partir de sinais de voz. Para a classificação de emoções, testando o modelo MLP usando 300 camadas ocultas e recursos de voz combinados de MFCC's, ZCR e Cg , observou-se melhoria significativa em comparação ao uso de apenas um recurso. Utilizando MFCC e ZCR a pontuação da acurácia é de 88% e é maior 9% do que a precisão para o recurso que usa apenas MFCC. O aumento da acurácia também se dá pela quantidade de recursos utilizados. Nos testes que usamos 40 MFCC's a precisão foi de 79% já quando aumentamos o número de recursos de MFCC para 60%, a precisão é aumenta para 84%.

Referências

Alıas, Francesc, Joan Claudi Socoró e Xavier Sevillano. 2016. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences* 6 (5): 143.

Arachchige, Chandima N. P. G., Luke A. Prendergast e Robert G. Staudte. 2020. Robust analogs to the coefficient of variation. *Journal of Applied Statistics* 0 (0): 1–23.

Chandrasekaran, Dhivya, e Vijay Mago. 2021. Evolution of Semantic Similarity—A Survey. *ACM Computing Surveys (CSUR)* 54 (2): 1–37.

Douglas-Cowie, Ellen, Nick Campbell, Roddy Cowie e Peter Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech communication* 40 (1-2): 33–60.

Ezz-Eldin, Mai, Ashraf AM Khalaf, Hesham FA Hamed e Aziza I Hussein. 2021. Efficient Feature-Aware Hybrid Model of Deep Learning Architectures for Speech Emotion Recognition. *IEEE Access* 9:19999–20011.

Farooq, Misbah, Fawad Hussain, Naveed Khan Baloch, Fawad Riasat Raja, Heejung Yu e Yousaf Bin Zikria. 2020. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors* 20 (21): 6008.

Hilmy, Muhammad Syazani Hafiy, Ani Liza Asnawi, Ahmad Zamani Jusoh, Khaizuran Abdullah, Siti Noorjannah Ibrahim, Huda Adibah Mohd Ramli e Nor Fadhillah Mohamed Azmin. 2021. Stress Classification based on Speech Analysis of MFCC Feature via Machine Learning. Em *2021 8th International Conference on Computer and Communication Engineering (ICCCE),* 339–343. IEEE.

Ho, Manh-Tung, Quan-Hoang Vuong et al. 2021. Affective computing at the edge: A bibliometric analysis of the period 1995-2020.

Jin, Xin, Qian Liu e Huizhen Long. 2021. Impact of cost–benefit analysis on financial benefit evaluation of investment projects under back propagation neural network. *Journal of Computational and Applied Mathematics* 384:113172.

Kaur, Ramandeep, e Sandeep Kautish. Abril de 2019. Multimodal Sentiment Analysis: A Survey and Comparison. *International Journal of Service Science, Management, Engineering, and Technology* 10 (): 38–58.

Koduru, Anusha, Hima Bindu Valiveti e Anil Kumar Budati. 2020. Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology* 23 (1): 45–55.

Kumar, Yogesh, e Manish Mahajan. 2019. Machine learning based speech emotions recognition system. *Int. J. Sci. Technol. Res* 8 (7): 722–729.

Langari, Shadi, Hossein Marvi e Morteza Zahedi. 2020. Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked* 20:100424.

Li, Huan-Chung, Telung Pan, Man-Hua Lee e Hung-Wen Chiu. 2021. Make Patient Consultation Warmer: A Clinical Application for Speech Emotion Recognition. *Applied Sciences* 11 (11): 4782.

Nazzal, Jamal M, Ibrahim M El-Emary e Salam A Najim. 2008. Multilayer perceptron neural network (MLPs) for analyzing the properties of Jordan Oil Shale 1.

Neto, JRT, GP Filho, LY Mano e J Ueyama. 2018. Verbo: voice emotion recognition database in Portuguese language. *J Comput Sci* 14 (11): 1420–1430.

Ospina, Raydonal, e Fernando Marmolejo-Ramos. 2019. Performance of Some Estimators of Relative Variability. *Frontiers in Applied Mathematics and Statistics* 5:43.

Park, Chanseok, Haewon Kim e Min Wang. 2020. Investigation of finite-sample properties of robust location and scale estimators. *Communications in Statistics - Simulation and Computation* 0 (0): 1–27.

Parra-Gallego, Luis Felipe, e Juan Rafael Orozco-Arroyave. 2021. Classification of Emotions and Evaluation of Customer Satisfaction from Speech in Real World Acoustic Environments. *arXiv preprint arXiv:2108.11981.*

Qiao, Weibiao, Mohammad Khishe e Sajjad Ravakhah. 2021. Underwater targets classification using local wavelet acoustic pattern and Multi-Layer Perceptron neural network optimized by modified Whale Optimization Algorithm. *Ocean Engineering* 219:108415.

Schuller, Dagmar M, e Björn W Schuller. 2021. A review on five recent and near-future developments in computational processing of emotion in the human voice. *Emotion Review* 13 (1): 44–50.

Sharma, Garima, Kartikeyan Umapathy e Sridhar Krishnan. 2020. Trends in audio signal feature extraction methods. *Applied Acoustics* 158:107020.

Tiwari, Vibha. 2010. MFCC and its applications in speaker recognition. *International journal on emerging technologies* 1 (1): 19–22.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Ostensivo (Público)

Projeto de Engenharia de Computação

Assunto: Projeto de Engenharia de Computação
Assinado por: Iury Fernandes
Tipo do Documento: Anexo
Situação: Finalizado
Nível de Acesso: Ostensivo (Público)
Tipo do Conferência: Cópia Simples

Documento assinado eletronicamente por:

- **Iury Anderson Fernandes Coelho**, **ALUNO (201811250026) DE BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO - CAMPINA GRANDE**, em 29/09/2022 07:17:47.

Este documento foi armazenado no SUAP em 29/09/2022. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 637343
Código de Autenticação: 20e7b5c508