DESENVOLVIMENTO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA QUANTIFICAR ATORES PRESENTES EM CENAS ACÚSTICAS

Paulo Henrique de Sousa[1,2], Thiago Medeiros de Menezes[1,2],
Ana Paula Carvalho Cavalcanti Furtado [2] e Péricles Barbosa Cunha de Miranda[2]
[1]*Sidia*
Av. Avenida Darcy Vargas, 654, 69055-035, Manaus, AM, Brasil
[2]*CESAR School*
Rua Bione, Cais do Apolo, 220, 50030-390, Recife, PE, Brasil

RESUMO

Considerando a classificação de cenas acústicas (CCA) como um meio para identificar características do som ambiente através de aprendizado de máquina (AM) e, diante do alto índice de violência contra a mulher, objetiva-se criar um *dataset* sintético de áudios para aplicar cinco arquiteturas de redes neurais convolucionais (RNC) e tratar o problema de contagem de locutores presentes em cenas acústicas envolvendo discussões que podem evoluir para agressões. Para tanto, procedeu-se à metodologia CRISP-DM para construir um *dataset* com diálogos sintéticos para treinar os cinco modelos. Observa-se que foi possível obter resultados satisfatórios na avaliação dos modelos conforme os resultados para a métrica f1-score: 95% Modelo Sequencial; 88% MobileNet; 86% GoogleNet e 86% ResNet. O que permite concluir que a CCA por meio de RNC é adequada para o propósito deste trabalho de quantificar os envolvidos numa discussão entre possíveis vítimas e agressores.

PALAVRAS-CHAVE

Classificação de Cenas Acústicas, Redes Neurais, Aprendizado de Máquina, Inteligência Artificial

1. INTRODUÇÃO

A violência física e psicológica é um problema grave que impacta negativamente a vida de milhares de mulheres. Segundo O Estado de S.Paulo (2021), 7 em cada 10 ocorrências de violência contra mulher são praticadas por pessoas conhecidas da vítima. A pesquisa também mostra que 48% das agressões ocorreram na residência da vítima, 19,9% ocorreram na rua e 9,4% no trabalho. Como a maioria das vítimas não denuncia as agressões por medo de futura retaliação dos seus agressores, é importante que haja um mecanismo para detectar a violência e acionar a Polícia e/ou a Justiça a fim de tomar medidas e ações de proteção para a vítima.

Este trabalho foi idealizado em conjunto com o grupo HEAR (*Helping Everyone to Actively React*) buscando formas de utilizar a tecnologia para combater a violência contra a mulher. Uma das abordagens utilizadas pelo grupo consiste na análise e aplicação da CCA através de técnicas de AM para identificar a ocorrência de violência em uma cena acústica durante uma discussão entre duas ou mais pessoas (Agencia Brasil, 2020). A CCA é uma área de pesquisa dentro da inteligência artificial (IA) que tem como intuito categorizar cenas acústicas a partir de sons presentes num determinado ambiente e contexto (Barchiesi & al., 2015). Dentro dessa abordagem, a quantificação dos atores presentes numa cena acústica é muito importante para corroborar com classificadores que identifiquem a existência de violência contra a mulher.

Esta pesquisa visa aprimorar modelos de CCA utilizados para detectar violência com a inclusão da estimativa da quantidade de locutores envolvidos numa cena acústica, uma vez que se trata de um detalhe que pode enriquecer o método de identificação da agressão e evidenciar a presença de indivíduos que poderiam ser omitidos numa suposta denúncia, ora por falta de conhecimento do denunciante, ora por decisão da vítima.

2. REFERENCIAL TEÓRICO

A IA é uma área multidisciplinar que se tornou um assunto popular na última década e que tem proficiência na resolução de problemas lógicos bem definidos (Choi, 2020). Uma abordagem para solucionar problemas de AM é através da utilização de um grande volume de dados. Em complemento à AM, existe uma técnica conhecida como Aprendizado Profundo (AP) que consiste no uso de redes neurais artificiais em modelos que possuem várias camadas.

A voz e os sons presentes numa cena acústica são informações que podem ser processadas por sistemas de inteligência artificial. Esses sistemas fazem uso de modelos de CCA. Atualmente, assistentes pessoais como o Bixby, a Alexa, o Google Now e o Siri já interpretam a fala humana obedecendo a comandos de voz e atuando com respostas que se assemelham aos comportamentos dos humanos. Além dos assistentes pessoais, outras possibilidades relevantes podem ser exploradas pela IA através dos modelos de CCA, por exemplo, o monitoramento de ambientes para detecção de presença e detecção de anormalidades.

O uso de métodos tradicionais de AM para aplicação da CCA é uma opção viável. Para tanto, é necessário extrair os atributos necessários para a análise do áudio tais como *pitch*, energia, MFCC (*Mel Frequency Cepstral Coefficient*, ZCR (*Zero Rating Crossing*) e utilizá-los em algoritmos de classificação como o KNN (*K-nearest neighbors*). Embora estes métodos tradicionais tenham êxito na classificação, eles possuem uma limitação quanto à correta escolha e extração dos atributos utilizados. É difícil encontrar características que possam ser utilizadas para representar padrões em diferentes ambientes e situações (Mu, 2021).

Este problema pode ser resolvido por meio do uso de redes neurais profundas (RNP), em especial, as RNC que são capazes de extrair os atributos automaticamente e que provaram ter uma boa capacidade para capturar *time–frequency features*, o que torna as RNCs adequadas para a resolução de tarefas de CCA (Mu, 2021). As redes neurais são estruturas compostas por unidades neurais artificiais interconectadas que imitam um neurônio humano e que trocam informações. As redes neurais também podem ser utilizadas no AP onde ocorre a aplicação de redes neurais que possuem camadas profundas (IBM.Cloud, 2021).

As RNC tem sido amplamente utilizadas na tarefa de classificação, detecção e reconhecimentos em imagens e vídeos. Segundo (Vargas, Paes, & Vasconcelos, 2016), "uma Rede Neural Convolucional é uma variação das redes de Perceptrons de Múltiplas Camadas, tendo sido inspirada no processo biológico de processamentos de dados visuais".

3. METODOLOGIA

Nesta pesquisa foi utilizada a metodologia CRISP DM (*Cross Industry Standard Process for Data Mining*). Essa metodologia fundamenta-se no método experimental para padronizar a realização de experimentos de *Data Mining* provendo um *framework* para organizar os dados independente do setor da indústria e do tipo de tecnologia a ser utilizada (Wirth & Hipp, 2000), (Barros & Lehfeld, 2007). A metodologia CRISP DM divide-se em 6 etapas: (1) entendimento do negócio; (2) entendimento dos dados; (3) preparação dos dados; (4) modelagem; (5) avaliação e (6) implementação.

A 1ª etapa consistiu na identificação dos requisitos e das pessoas envolvidas no problema. Na 2ª etapa, foi necessário investigar as propriedades do conjunto de dados (*dataset*) a fim de prepará-lo para a 3ª etapa em que os dados foram normalizados e compatibilizados para atender aos modelos de AP. Na 4ª etapa, foram definidos os modelos e métricas de desempenho tais como a acurácia, a revocação (*recall*), a precisão e o f1-score, os quais foram utilizados na 5ª etapa para avaliação dos resultados obtidos pelos modelos. Na 6ª etapa, é realizada a implantação dos modelos desenvolvidos. Para a implantação, selecionamos os modelos cujos experimentos obtiveram os melhores desempenhos nos testes com amostras sintéticas.

O conjunto de dados utilizado na 2ª e 3ª etapa foi criado a partir da combinação de recortes de áudios para gerar amostras com um determinado número de locutores. A fonte de áudios original foi adaptada do *dataset* VoxCeleb que consiste em vozes extraídas de entrevistas publicadas no YouTube. Este *dataset* foi encontrado por meio de buscas na internet por *datasets* de áudios. O *dataset* VoxCeleb foi escolhido por conter áudios agrupados por locutor possibilitando a construção de um novo *dataset* com dados sintéticos rotulados com o número de locutores. O novo *dataset* utilizado nos experimentos foi gerado a partir da seleção de um conjunto de entrevistas realizadas com 40 pessoas. Os áudios foram selecionados de forma randômica gerando para cada classe 10.000 novos áudios de 6 segundos no formato WAV. Cada classe representando o número de pessoas

falantes no áudio. Os áudios foram gerados por meio da biblioteca *Scaper*. O *Scaper* é uma biblioteca em Python que fornece métodos para composição de arquivos de áudio de *background* e *foreground* entre outras funcionalidades (Salamon, 2017).

Foram programadas variações no tempo de início e duração das falas no intervalo fixo de 6 segundos. A partir destas configurações, foram gerados os arquivos WAV simulando os diálogos. Cada arquivo continha dados de série temporal com um determinado número de amostras por segundo. O número de amostras por segundo também é conhecido por *Sample Rate*. Foi utilizada a frequência de 16.000 amostras por segundo (16.000 Hz ou 16,0 kHz) que está dentro das frequências perceptíveis pela audição humana que é capaz de detectar frequências desde aproximadamente 20 Hz a 20.000 Hz (Apple, 2022).

Para adaptar o *dataset* para treinamento do modelo é necessário transformar os arquivos de áudio em sinais digitais e posteriormente em tensores. Para isto, foi utilizada a API do *TensorFlow* que possibilitou decodificar os arquivos WAV 16-bit PCM em tensores. Nesta conversão, os valores que representam o sinal são escalados entre -1.0 e 1.0 e retornados numa matriz. É utilizado o parâmetro *desired channels* para obter uma representação do sinal sonoro em apenas um canal de reprodução. Este formato em apenas um canal é conhecido como sistema monofônico. Esta conversão foi escolhida para diminuir o número de parâmetros do *dataset* e agilizar o treinamento.

Nos experimentos realizados nesta pesquisa, foi utilizado como entrada para o *dataset* o espectrograma obtido por meio da transformada de Fourier de tempo curto. Para realizar a transformada de Fourier foi utilizada uma função presente na API do TensorFlow para cálculo do STFT. Para gerar o espectrograma foi implementada em Python uma função que recebe o sinal de áudio como parâmetro, padroniza a duração do sinal, aplica a transformada de Fourier de tempo curto, obtém os valores absolutos da transformada e retorna o espectrograma para construir o *dataset*.

Ao final, foram criados 5 modelos utilizando as seguintes arquiteturas de RNC: arquitetura Sequencial, GoogLeNet, MobileNet, ResNet e DenseNet. Estes modelos foram avaliados utilizando as métricas de acurácia, precisão, revocação e *f1-score*

4. TRABALHOS RELACIONADOS

Este capítulo apresenta uma análise geral sobre os trabalhos relacionados a esta pesquisa que foram selecionados a partir de um conjunto de critérios de pesquisa e seleção.

Os trabalhos relacionados foram consultados por meio da seguinte *string* de busca: “speaker count” AND (“neural network” OR “machine learning” OR “deep learning”). A *string* de busca foi proposta com a intensão de encontrar trabalhos de pesquisa que tratassem da contagem de atores presentes em gravações de áudio para que fosse consultada a técnica utilizada e os resultados obtidos. Para diminuir a avaliação de estratégias tradicionais de AM, optou-se por limitar a consulta para os artigos publicados nos últimos 5 anos. Esta decisão foi também uma estratégia para selecionar trabalhos que estivessem utilizando técnicas modernas que pudessem representar o estado da arte sobre a contagem de locutores presentes em gravações de áudio.

Foram utilizados como fonte de pesquisa as plataformas IEEE Xplore e Scopus. Como resultado da consulta foram obtidos 22 artigos relacionados na base de dados do IEEE Xplore e 5 artigos relacionados na base de dados do Scopus. Foi realizado uma análise dos títulos e resumos dos trabalhos para selecionar apenas aqueles que abordavam a contagem de locutores presentes em áudios utilizando aprendizagem de máquina. A partir dos critérios definidos, foram selecionados 9 artigos no total. A Figura \ref{fig:fig_trabalhos_relacionados} apresenta uma visão cronológica dos mesmos onde pode ser observado a existência de trabalhos recentes incluindo 3 publicações no ano de 2021. Também é apresentada a lista com os títulos dos trabalhos e em seguida um resumo sobre a pesquisa.

Lista de trablhos selecionados:

1. Classification vs. Regression in Supervised Learning for Single Channel Speaker Count Estimation (STÖTER et al., 2018)
2. CountNet: Estimating the number of concurrent speakers using supervised learning (FABIAN-ROBERT et al., 2019)

3. Overlapped Speech Detection and Competing Speaker Counting - Humans Versus Deep Learning (ANDREI et al., 2019)
4. End-to-End Overlapped Speech Detection and Speaker Counting with Raw Waveform (ZHANG et al., 2019)
5. Competing speaker count estimation on the fusion of the spectral and spatial embedding space Layer (PENG; WU; QU, 2020)
6. Speaker Counting Model based on Transfer Learning from SincNet Bottleneck (WANG et al., 2020)
7. Count and Separate: Incorporating Speaker Counting for Continuous Speaker Separation (ZHONG-QIU; WANG, 2021)
8. A distributed approach to speaker count problem in an open-set scenario by clustering pitch features (PANDEY; BANERJEE, 2021)
9. High-Resolution Speaker Counting in Reverberant Rooms Using CRNN with Ambisonics Features (GRUMIAUX et al., 2021)

Embora tenha sido realizado o filtro por data de publicação, a maioria dos trabalhos encontrados a partir da *string* de busca eram recentes então a limitação pela data não removeu muitos trabalhos. Conforme demonstrado nos Quadros 1 e 2 os trabalhos analisados utilizaram diferentes técnicas para abordar a contagem de locutores. Os trabalhos de (FABIAN ROBERT et al., 2019), (ANDREI et al, 2019) e (ZHONG-QIU; WANG, 2021) utilizaram RNC assim como também foi utilizado neste trabalho. Os demais trabalhos realizam a extração de parâmetros para utilização de métodos tradicionais de aprendizagem de máquina.

O conjunto de dados utilizado nos trabalhos selecionados foi praticamente todo gerado a partir de fontes de áudios como Wall Street Journal corpus e LibriSpeech. A única exceção foi o trabalho de Grumiaux et al. (2021) que não utiliza um conjunto de dados. Neste trabalho foi usada a mesma estratégia dos trabalhos que usaram *dataset*, foi escolhida uma fonte de áudios, no caso a fonte de áudios do VoxCeleb e, então, gerado um conjunto sintético de áudios com sobreposição das falas.

Quadro 1. Quadro de análise de trabalhos relacionados extraídos da base do Scopus

Trabalho	**Técnica**	**Observações**
Stöter et al. (2018)	Rede neural profunda comparando classificação e regressão	Avalia o estado da arte do uso de rede neural profunda baseado numa solução utilizando a arquitetura Bi-directional Long Short-Term Memory
Fabian-Robert et al. (2019)	Uso de arquiteturas profundas e análise de RNC recorrentes	Proposta de paradigma probabilístico unificador
Andrei et al. (2019)	Análise de percepção de número de falantes por humanos e por RNC	Análise de parâmetros para melhor captação do áudio e experimentos usando RNC
Peng, Wu e Qu (2020)	Uso de deep spectral and spatial embedding fusion para contagem de falantes	Realização de experimentos para demonstrar os ganhos de acurácia através do uso de métodos embarcados
Zhong-Qiue Wang (2021)	Proposta de modulo de contagem de falantes (zero, um e dois falantes) para alternar entre um modo offline de separação contínua da fala e outro modo de aprimoramento da fala	Realiza experimentos utilizando o conjunto de dados LibriCSS para validação da proposta

Fonte: Elaborado pelo autor

Quadro 2. Quadro de análise de trabalhos relacionados extraídos da base do IEEE Xplore

Trabalho	**Técnica**	**Observações**
Zhang et al. (2019)	Proposta de estrutura de ponta a ponta para detecção de fala sobreposta e contagem de falantes	Realização de experimentos sobre a proposta para demonstrar o ganho de acurácia na tarefa de detecção de fala e contagem de falantes IEEE Xplore
Wang et al. (2020)	Proposta de metodologia baseada no framework de rede neural SincNet	Comparativo de performance entre a proposta e o uso de Mel-Frequency Cepstral Coefficients (MFCC)
Pandey e Banerjee (2021)	Abordagem de agrupamento distribuído para resolver o problema da contagem de falantes utilizando vários microfones disponíveis em smartphones emuma grande área geográfica para capturar e extrair características estatísticas de pitch dasamostras de áudio.	Realizada a avaliação de desempenho do algoritmo usando smartphones reais
Grumiaux et al. (2021)	Utiliza rede neural recorrente convolucional multicanal que produz uma estimativa em uma resolução de quadro de curto prazo	Treinamento de modelos para detecção de até cinco falantes simultâneos utilizando dados simulados, incluindo condições diferentes em termos de posiçõesde fonte e microfone, reverberação e ruído.

Fonte: Elaborado pelo autor

Pelo que foi observado pela leitura dos artigos selecionados, a contagem de locutores em ambientes abertos e não controlados ainda necessita de soluções mais abrangentes. Mas para ambientes controlados as RNC são bem empregadas conforme indentificado de acordo com as características dos trabalhos selecionados incluindo a técnica utilizada e resultados obtidos sobre a predição da quantidade de locutores presentes em gravações de áudio.

5. RESULTADOS

Nesta pesquisa, foram utilizados 5 modelos de RNC: (1) arquitetura sequencial do Keras com 10 camadas, (2) arquitetura baseada no GoogLeNet de 76 camadas, (3) arquitetura MobileNet com 28 camadas, (4) arquitetura ResNet com 50 camadas, e (5) arquitetura DenseNet com 121 camadas. Para todas as arquiteturas, foram executadas 25 épocas para treinamento dos modelos. Considerou-se como condição de parada do treinamento a execução de 10 épocas sem percepção de melhoria. Além disso, foi utilizada uma opção que permite a restauração da melhor versão obtida durante o treinamento.

A Tabela 1 apresenta as características de cada modelo. O número de épocas de treinamento estava limitado ao parâmetro "*patience*" que fez com que houvesse valores diferentes no decorrer dos experimentos realizados. Foram realizados 5 experimentos para cada modelo para então realizar testes estatísticos e verificar a consistência dos resultados. O tamanho do lote utilizado para os modelos foi de 128 imagens por etapa. Como mostrado na Tabela 2, todos os modelos obtiveram desempenhos de acurácia acima de 85%. O modelo (1) foi o que obteve a melhor acurácia.

Outros modelos também obtiveram resultados tão bons quanto o modelo Sequencial em determinados experimentos. Porém, estatisticamente, o modelo Sequencial se sobressaiu em relação aos demais em todas as métricas utilizadas. Para confirmar este resultado, foi aplicado o teste de Friedman. O cálculo da significância (também referenciado como *p-value*) foi 0,120 ficando acima do valor do nível de significância $\alpha = 0.050$. Portanto, pode-se afirmar que a diferença nas distribuições é estatisticamente significativa e não há evidências suficientes para rejeitar a hipótese nula. Para examinar este resultado, foram gerados diagramas de diferença crítica para cada métrica e a partir deles foi possível constatar que o modelo Sequencial foi de fato o classificador com melhor desempenho.

Tabela 1. Tabela de métricas dos modelos com a média dos valores obtidos nas repetições dos experimentos

Modelos	Camadas	Parâmetros	Patience	Épocas	Opitmizer
Sequencial	10	504.140.809		25, 25, 25, 25, 25	
GoogLeNet	76	1.661.240		23, 16, 21, 25, 24	
MobileNet	28	3.233.413	10	23, 20, 17, 17, 17	Adam
ResNet	50	23.591.685		17, 19, 17, 25, 25	
DenseNet	121	36.357		16, 25, 19, 16, 22	

Fonte: Elaborada pelo autor

Tabela 2. Tabela de métricas dos modelos com a média dos valores obtidos nas repetições dos experimentos

Métricas do Modelo Sequencial do Keras				
Classe	Precisão	Revocação	F1-Score	Acurácia
1	100%	100%	100%	
2	97%	99%	98%	
3	95%	94%	94%	95%
4	90%	88%	88%	
5	93%	94%	93%	
Métricas do Modelo GoogLeNet				
1	99%	90%	93%	
2	87%	97%	91%	
3	83%	86%	84%	87%
4	81%	75%	77%	
5	91%	86%	88%	
Métricas do Modelo MobileNet				
1	100%	97%	98%	
2	86%	97%	91%	
3	81%	76%	79%	86%
4	77%	77%	76%	
5	90%	84%	87%	
Métricas do Modelo ResNet				
1	97%	95%	96%	
2	88%	92%	90%	
3	87%	80%	83%	86%
4	77%	76%	76%	
5	85%	89%	86%	
Métricas do Modelo DenseNet				
1	92%	100%	96%	
2	96%	89%	93%	
3	93%	86%	89%	87%
4	81%	88%	85%	
5	93%	91%	92%	

Fonte: Elaborada pelo autor

6. DISCUSSÃO

Os experimentos deste trabalho possibilitaram avaliar as arquiteturas de CNN citadas na seção anterior através da criação do conjunto de áudios e da realização dos ciclos de experimentos utilizando a metodologia CRISP-DM a fim de estimar a quantidade de pessoas presentes nas cenas acústicas utilizadas.

Adaptar o *dataset* para que fique compatível com os modelos é uma tarefa árdua, mas, necessária para garantir bons resultados durante a avaliação. Além disto, o treinamento dos modelos é demorado podendo ultrapassar 10 horas de execução e existe a possibilidade de interrupção da execução.

Conforme explicado anteriormente, foi utilizada a métrica f1-score para avaliar os modelos dado que a acurácia não é uma métrica recomendada para avaliar *datasets* não simétricos. Assim, para melhorar a comparação dos modelos foi utilizada a média aritmética do f1-score dos cinco experimentos realizados em cada modelo.

A partir dos resultados obtidos, o modelo com melhor desempenho foi apresentado pela arquitetura sequencial do Keras com acurácia e f1-score de 95%. É possível aplicar esse modelo em conjunto com detectores de violência contra a mulher por meio de CCA. A quantidade de pessoas envolvidas numa cena acústica pode sugerir que não esteja ocorrendo uma agressão quando houver somente um locutor e que, numa cena com vários locutores, onde a possível vítima não está sozinha com o agressor, existem outras pessoas presentes que podem intervir.

De forma geral, a geração dos áudios sintéticos atendeu o propósito de criar uma base de áudios rotuladas com o número de locutores; o formato escolhido do áudio possibilitou obter o *Waveform* e o Espectrograma sem perda de informações importantes para a contagem e; os espectrogramas obtidos foram capazes de treinar os modelos. Também foi possível medir o desempenho dos classificadores e confirmar estatisticamente a consistência dos resultados.

Existe uma abordagem mais avançada de AM que consiste na identificação dos locutores presentes. Vale ressaltar que esta pesquisa não abordou a identificação dos locutores e, portanto, não consegue identificar os locutores presentes numa cena acústica. Este trabalho poderia ser melhorado para que fosse incluída essa opção para identificar uma voz conhecida, por exemplo.

7. CONCLUSÃO

Esta pesquisa teve como proposta o treinamento destes modelos para predizer a quantidade de atores presentes numa cena acústica. As cenas acústicas de treinamento consistiram em áudios modificados digitalmente com a inclusão de vozes para atingir um determinado número de locutores.

Para isso, criou-se um conjunto de dados de áudio e aplicou-se 5 modelos de RNC em experimentos para avaliar aqueles com melhor desempenho considerando as métricas de *f1-score*. O melhor resultado foi obtido pela arquitetura sequencial do Keras cujo f1-score foi de 95%.

Embora o trabalho tenha adotado essa métrica, existe também a métrica MAE, que pode vir a ter um resultado mais exato por levar em consideração a distância do erro, isto é, quanto mais distante a predição estiver do valor correto, maior é a penalização. Além disso, existe uma abordagem mais avançada de AM que consiste na identificação dos locutores presentes.

Seria muito relevante que a identificação dos locutores presentes fosse tema de pesquisa em trabalhos futuros já que o presente trabalho não abordou este tema. Isto seria muito útil para auxiliar na perícia de casos de agressão. Outra possibilidade que também não foi aplicada neste trabalho é a de além de identificar, vincular as vozes ao locutor. Está técnica é conhecida por "*Speaker Diarization*" e poderia ser aplicada em trabalhos futuros.

O tratamento das limitações apontadas aprimoraria os resultados deste trabalho com a extração de mais informações relevantes da cena acústica facilitando a predição e a análise de uma agressão por terceiros como órgãos competentes. Também, a inclusão de novas métricas tal como o MAE seria importante para confirmar o desempenho de cada classificador.

AGRADECIMENTOS

Esta pesquisa foi apoiada por Samsung Eletrônica da Amazônia Ltda de acordo com a Lei Federal do Brasil 8.387/1991

REFERÊNCIAS

Agencia Brasil, F. F. (January de 2020). *Equipe desenvolve plataforma para combater violência contra a mulher*. Fonte: Equipe desenvolve plataforma para combater violência contra a mulher: https://agenciabrasil.ebc.com.br/geral/noticia/2020-01/equipe-desenvolve-plataforma-para-combater-violencia-contra-mulher

Andrei, Valentin et al. Overlapped Speech Detection and Competing Speaker Counting—Humans Versus Deep Learning. IEEE Journal of Selected Topics in Signal Processing, v. 13, n. 4, p. 850–862, 2019.

Apple (2022). Som e Audição. Fonte: https://www.apple.com/pt/sound

Barchiesi, D., & al., e. (May de 2015). Acoustic Scene Classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine, 32*, 16–34.

Barros, A., & Lehfeld, S. (2007). *Fundamentos de metodologia: um guia para a iniciação científica.* Pearson Prentice Hall.

Choi, R. e. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology (2020) 9(2).*

Fabian-Robert et al. CountNet: Estimating the Number of Concurrent Speakers Using Supervised Learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, v. 27, n. 2, p. 268–282, 2019.

Grumiaux, Pierre-Amaury et al. High-Resolution Speaker Counting in Reverberant Rooms Using CRNN with Ambisonics Features. In: 2020 28th European Signal Processing Conference (EUSIPCO)., 2021. P. 71–75

IBM.Cloud. (2021). *Redes neurais*. Fonte: Redes neurais: https://www.ibm.com/br-pt/cloud/learn/neural-networks

Mu, e. a. (November de 2021). Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*.

O Estado de S.Paulo, F. R. (July de 2021). *Uma a cada quatro mulheres foi vítima de violência no último ano, aponta pesquisa*. Fonte: Uma a cada quatro mulheres foi vítima de violência no último ano, aponta pesquisa: https://brasil.estadao.com.br/noticias/geral,uma-a-cada-quatro-mulheres-foi-vitima-de-violencia-no-ultimo-ano-aponta-pesquisa,70003738858

Salamon, J. e. (October de 2017). Scaper: A library for soundscape synthesis and augmentation. *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 344–348). IEEE.

Pandey, S.; Banerjee, A. A. Distributed Approach to Speaker Count Problem in an Open-Set Scenario by Clustering Pitch Features. IEEE Signal Processing Magazine, abr. 2021

Peng, Chao; Wu, Xihong; Qu, Tianshu. Competing Speaker Count Estimation on the Fusion of the Spectral and Spatial Embedding Space. In: PROC. Interspeech 2020., 2020. P. 3077–3081.

Stöter, Fabian-Robert et al. Classification vs. Regression in Supervised Learning for Single Channel Speaker Count Estimation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2018. P. 436–440.

Vargas, A., Paes, A., & Vasconcelos, C. (2016). Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres. *Proceedings of the XXIX Conference on Graphics, Patterns and Images*, (pp. 1–4).

Wang, Wei et al. Speaker Counting Model based on Transfer Learning from SincNet Bottleneck Layer. In: 2020 IEEE International Conference on Pervasive Computing and

Communications (PerCom), 2020. P. 1–8.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (pp. 29–39).

Zhong-Qiu, Wang; Wang, DeLiang. Count And Separate: Incorporating Speaker Counting For Continuous Speaker Separation. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021. P. 11–15.