ESTUDO DE ATIVAÇÃO DE SISTEMA DE SEGURANÇA POR RECONHECIMENTO DE VOZ USANDO REDE NEURAL CONVOLUCIONAL

Marcos Paulo Rizzi dos Santos, Pedro Henrique Lucas, Daniel da Silva Caubianco, Amita Muralikrishna, Julio Cesar Serafim Casini, Carlos Eduardo Oliveira da Silva.

Instituto Federal de São Paulo, Campus São José dos Campos, Rodovia Presidente Dutra, km 145, Jardim Diamante - 12223-201 - São José dos Campos-SP, Brasil, marcos.rizzi@aluno.ifsp.edu.br, p.jose@aluno.ifsp.edu.br, daniel.caubianco@aluno.ifsp.edu.br, amita@ifsp.edu.br, julio.casini@ifsp.edu.br, carlossilva@ifsp.edu.br.

Resumo

Considerando o avanço tecnológico trazido pelo reconhecimento de voz para agilizar tarefas do dia a dia, entende-se o potencial que a integração entre acionamento de alarmes e a flexibilidade de reconhecer palavras chaves pode trazer benefícios para a sociedade. O objetivo deste projeto é a construção inicial de um sistema de segurança usando rede neural convolucional (RNC) para aplicação em sistemas de segurança para escolas, que seja capaz de ser ativado através de reconhecimento de voz em caso de emergência, como roubos, terrorismo etc. Propõe-se a identificação de algumas palavras, que foram selecionadas de forma empírica e de acordo com critérios de confiabilidade e de uso cotidiano, selecionadas propositalmente com intuito de avaliar o modelo perante a similaridade fonética, falsos positivos. A coleta de 15000 áudios foi realizada no ambiente escolar, em momentos variados de silêncio e perturbações. O processo de criação do modelo foi realizado através da transformada de Fourier de Tempo Curto, espectrogramas e RNC, inicialmente com uma camada de normalização, duas camadas convolucionais de 32 e 64 filtros, uma camada densa de 128 neurônios intermediária e uma camada densa de 20 neurônios na saída, uma camada *flatten*, uma camada de *pooling*, e duas camadas de *dropout* com taxa de 0,25 e 0,50 respectivamente, também a taxa de aprendizagem de 0,001 foi utilizada. O tamanho de dados para treinamento e validação foram separados em 30% e 70% respectivamente. Os resultados obtidos pela identificação de cada palavra foram inseridos em uma matriz de confusão, a fim de validar o modelo construído. O modelo de RNC demonstrou ser eficiente, com uma taxa de 6% de alarme falso, identificando palavras específicas com tempo de processamento médio de 30 milissegundos e com precisão de 89%, sendo testado nas mesmas condições de coleta de dados, mesmo considerando limitações e desafios encontradas no dia a dia como: alarmes falsos, efeitos da perturbação do ambiente, semelhança de fonemas entre palavras.

Área do Conhecimento: Engenharia de Computação, Engenharia de Controle e Automação.

Introdução

Atualmente a segurança é um desafio enfrentado diariamente em todos os locais. Seguindo esse raciocínio, em média no Brasil 41.844 pessoas são vítimas por ano de fatalidades ocasionadas por latrocínio, homicídio doloso e lesão corporal seguida de morte (Sinespjc, 2023). Para que um sistema de segurança seja eficiente é necessário que a ação de perigo seja detectada o quanto antes possível e que a reação seja rápida. A evolução tecnológica permitiu o avanço na área da segurança, como o exemplo de Rouhani (2019), em que a Inteligência Artificial (IA) é capaz de realizar reconhecimento facial para liberação de acesso, evitando que a perda de chaves de acesso coloque em risco a segurança de um local privado. Além disso, o reconhecimento de voz é uma tecnologia já conhecida; o *Amazon Echo Dot*, por exemplo, é um assistente virtual que demonstra alta capacidade de atender comandos de voz recebidos, que pode ser programada de várias formas, entre elas como uma ferramenta na área da saúde na qual a IA da *Amazon* lembra uma pessoa de idade avançada o horário e o tipo do medicamento que se deve ministrar conforme prescrito pelo médico (Godoy, Farina, Florian, 2021). Com isso, o reconhecimento de voz traz o acionamento de sistemas rapidamente, agilizando o processo de chamada de emergência, podendo ser aplicado a uma ampla variedade de casos como

sistemas de segurança em bancos, escolas, ambiente fabril (Passos *et al*, 2006). Contudo, sabe-se que palavras que possuem pronúncia parecida podem causar confusão até mesmo em sistemas de reconhecimento de voz.

O objetivo deste projeto é desenvolver um modelo inicial de RNC capaz de reconhecer palavras-chave específicas para uso em sistemas de segurança escolar. Esse sistema visa melhorar a resposta a emergências em escolas, onde situações críticas como invasões, ameaças de violência ou outros incidentes podem colocar em risco a vida de estudantes e funcionários. A rápida detecção e ativação de alarmes através do reconhecimento de voz pode ser crucial para salvar vidas nesses ambientes. O processo de construção de um sistema de reconhecimento de voz é crucial para a eficácia do projeto. Dessa forma, a coleta de dados precisa ser cuidadosamente planejada para possibilitar a criação de sistemas de segurança adaptados a situações específicas (Warden, 2018). Com o uso do reconhecimento de voz, podem surgir limitações conhecidas, como ruídos de fundo, sobreposição de fala, variações no tom de voz e a semelhança fonética entre palavras. Essas situações podem comprometer a precisão do modelo e resultar em falsos positivos ou negativos. Para mitigar esses desafios, o sistema foi projetado para reconhecer palavras-chave específicas, cuidadosamente selecionadas para evitar confusão com palavras comuns usadas no dia a dia escolar. Além disso, o uso de palavras-chave discretas garante que o sistema possa ser ativado sem chamar a atenção de potenciais criminosos, aumentando a segurança e a eficiência em situações críticas.

Uma maneira eficaz de superar essas limitações é utilizar espectrogramas, que representam visualmente como a intensidade de diferentes frequências de um sinal varia ao longo do tempo. Definidos pela Transformada de Fourier de Tempo Curto, os espectrogramas permitem analisar sinais não estacionários, como a voz. A definição das janelas no espectrograma é crucial para a qualidade da imagem, pois ao reduzir o tamanho da janela, pois influencia diretamente na capacidade de capturar detalhes temporais e espectrais dos sinais de áudio (Oppenheim, 2013).

Como explicado em Luger (2013), a compreensão da complexidade das habilidades humanas e a análise em grande escala de dados desafiam a automação computacional, para isso aplicam-se técnicas de aprendizado de máquinas, e uma delas é a RNC que é composta por várias camadas interconectadas (Choi *et al*, 2018) que realizam operações específicas, como reconhecimento e detecção de padrões em imagens, como espectrogramas (Cordeiro, 2023). Sendo assim, os tipos de camadas aplicadas ao projeto foram camada densa, que combina todas as entradas de maneira global para aprender padrões complexos e realizar a classificação ou regressão, camadas convolucionais que detectam padrões como bordas, texturas e formas em imagens, mantendo a relação espacial, a camada *pooling*, que diminui o tamanho da representação, mantendo as informações mais importantes e reduzindo a sobrecarga computacional, a camada densa onde cada neurônio está conectado a todos os neurônios da camada anterior, sendo que essa camada realiza uma transformação linear dos dados de entrada, seguida por uma função de ativação que introduz não-linearidades no modelo, e também a camada *flatten*, que conecta as camadas convolucionais, que trabalham com dados espaciais, às camadas densas (Goodfellow, Bengio, Courville, 2016).

Metodologia

Para garantir uma seleção criteriosa das palavras-chaves e comuns utilizadas no experimento, de forma empírica, os autores adotaram um sistema de avaliação ponderada, no qual cada integrante atribuiu notas com base em critérios definidos no Quadro 1. Essa abordagem permitiu que tanto as palavras-chaves quanto as comuns fossem avaliadas objetivamente, levando em consideração sua adequação ao propósito do sistema de segurança. Palavras-chaves foram selecionadas para serem discretas o suficiente para não alertar potenciais criminosos, enquanto palavras comuns precisavam evitar ativar o alarme de forma não intencional. Além disso, o processo de seleção considerou o contexto prático de uso, recriando um cenário realista de uma sala de aula com ruídos e perturbações, que poderiam afetar o desempenho do modelo de reconhecimento de voz. Essa preocupação visa simular um ambiente típico no qual o sistema seria empregado, tornando os resultados mais aplicáveis ao uso real. A pontuação máxima representou as palavras ideais para cada categoria, indicando sua adequação para minimizar falsos alarmes e garantir discrição. Ao todo, 36 palavras foram analisadas com base nesses critérios, garantindo uma abordagem mais objetiva e menos sujeita a decisões arbitrárias, uma vez que todos os autores avaliaram. Embora esse conjunto de palavras possa parecer restrito, ele é suficiente para testar a viabilidade e o desempenho do modelo em um cenário específico,

como a simulação de uma sala de aula, sendo ideal para adequado para um escopo inicial mais limitado do projeto.

Quadro 1 – Critérios de Seleção de Palavras

Tipo de Palavra	N° Critério	Nota	Critério
Comum	1	Máxima	Palavra que é muito comum no dia a dia
		Mínima	Palavra pouco usada em sala de aula
Chave	2	Máxima	Palavra difícil de perceber a tentativa de acionar o alarme
		Mínima	Palavra que alertaria os criminosos
	3	Máxima	Palavra pouco usada em sala de aula
		Mínima	Palavra que é muito comum no dia a dia

Fonte: O autor

Foi utilizado o microfone condensador *Blue Yeti*, com sensibilidade à faixa de frequência de 20 Hz a 20kHz na configuração omnidirecional, e os áudios foram gravados com a frequência de 16000 Hz com a duração de 1 segundo e armazenados diretamente em um notebook. Além disso, foi considerada a gravação de voz de 3 alunos. Foram gravados, no total, 15000 áudios de 20 palavras definidas pelo grupo selecionadas pontualmente por serem palavras comuns no dia a dia e potencial palavras-chave, também foi considerado cenários de silêncio absoluto na sala de aula para evitar um acionamento equivocado do alarme. Utilizou-se *Jupyter Lab* com linguagem *Python* e bibliotecas gratuitas disponíveis, dentre essas o *Seaborn*, que fornece uma interface para desenhar gráficos estatísticos, o *TensorFlow*, aplicado para computação numérica, aprendizado de máquina e gerar os espectrogramas, e o *Keras*, para criar modelos, construir camadas e treinar modelos de redes neurais.

Foi necessário dividir as palavras em dois diretórios para a estruturação dos áudios: 30% dos dados foram usados para treinamento do modelo e 70% dos dados utilizados para a validação e teste do modelo. A RNC projetada neste estudo recebe como entrada espectrogramas, que passam inicialmente por uma camada de redimensionamento para ajustar as dimensões em 32x32. Em seguida, é aplicada uma camada de normalização para padronizar os dados. A arquitetura inclui duas camadas convolucionais com 32 e 64 filtros, respectivamente, ambas com função de ativação *ReLU*. Após as convoluções, há uma camada de *pooling* para redução dimensional e uma camada de *dropout* com taxa de 0,25 para prevenção de *overfitting*. A saída das camadas convolucionais é achatada por uma camada *flatten*, seguida por uma camada densa intermediária de 128 neurônios com ativação *ReLU*. Outra camada de *dropout* com taxa de 0,5 é aplicada antes da camada final densa, que possui tantos neurônios quanto o número de rótulos do conjunto de dados. A arquitetura foi implementada utilizando uma taxa de aprendizado de 0,001 e a saída será a classificação dos áudios. Por fim, observa-se a precisão do modelo e o tempo de execução da predição, gerando uma matriz de confusão para identificar a taxa de alarmes falsos e a influência de palavras curtas ou com fonemas semelhantes.

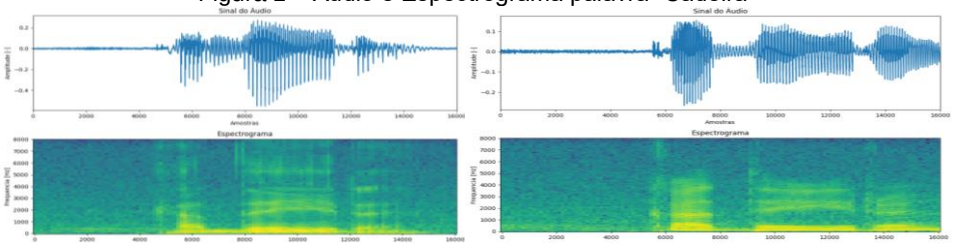
Resultados

Potenciais palavras-chave foram selecionadas com base nos critérios definidos e, entre elas, uma foi escolhida para o acionamento do sistema de segurança. Algumas das palavras escolhidas foram: conjuntura, arma, ninguém, cuidado. Do mesmo modo foram selecionadas, propositalmente, palavras comuns usadas no cotidiano por serem genéricas e populares no ambiente escolar, sendo elas: cadeira, mesa, professor, atenção, caderno, borracha, silêncio, caneta, lousa, senta, celular. Foi dada preferência por palavras curtas, com fonemas semelhantes e palavras que fossem adequadas em uma emergência, para que o sistema tivesse capacidade de diferenciá-las corretamente.

A Figura 1 apresenta dois áudios gravados da mesma palavra, em momentos diferentes de fundo, ou seja, ora com barulho de sala de aula, ora silêncio, e por pessoas distintas com sotaques e entonações diferentes durante os testes.

Observa-se um deslocamento temporal do espectrograma, fazendo com que o primeiro áudio forme uma imagem com espectrograma mais à esquerda.

Figura 1 – Áudio e Espectrograma palavra “Cadeira”

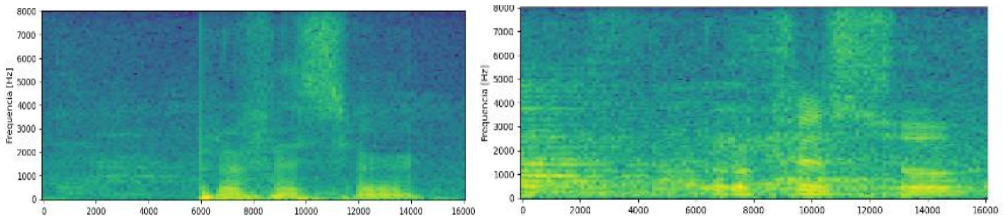


Fonte: O autor

Embora a olho nu as imagens possam não parecer iguais, cada pixel assegura um valor de amplitude em dB localizado em coordenadas de tempo e frequência. Ou seja, ao que parece, para a rede neural é indiferente o deslocamento em tempo, isso ocorre por causa das camadas convolucionais, que detecta padrões, independentemente da sua localização exata no espectrograma e de *pooling*, que agrupa informações próximas e torna o modelo menos sensível a pequenas variações no tempo de pronúncia, logo o que importa para o modelo é a sequência de amplitudes geradas na formação da imagem que vão se repetir como é mostrado na Figura 1.

Outro ponto identificado foi a presença de ruídos e perturbações na coleta de informações, lembrando que se trata de um evento previsto nos ensaios e não fora filtrado propositalmente para simular um ambiente mais hostil ou barulhento, onde a palavra-chave teria que ser identificada para acionar a emergência.

Figura 2 – Espectrograma palavra “Professor” (a) Sem Perturbação (b) Com Perturbação

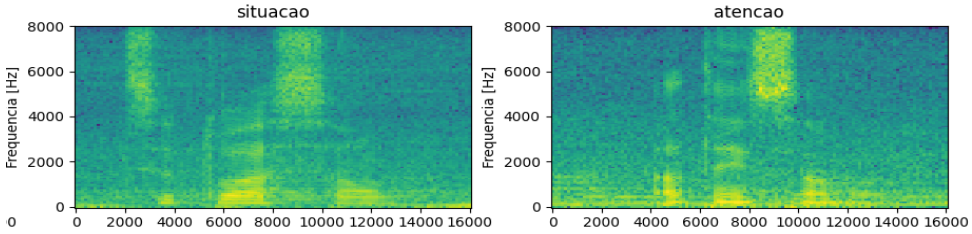


(a) (b)

Fonte: O autor

Como é mostrado na Figura 2, a perturbação é tida como vozes de fundo, movimentação de mesas, conversas etc. A capacidade da rede neural em detectar palavras-chave mesmo na presença desses distúrbios é um indicador importante de sua robustez e eficácia em condições adversas, demonstrando que o modelo RNC conseguiu ignorar perturbações e focar nas palavras-chave.

Figura 3 – Comparação entre palavras parecidas



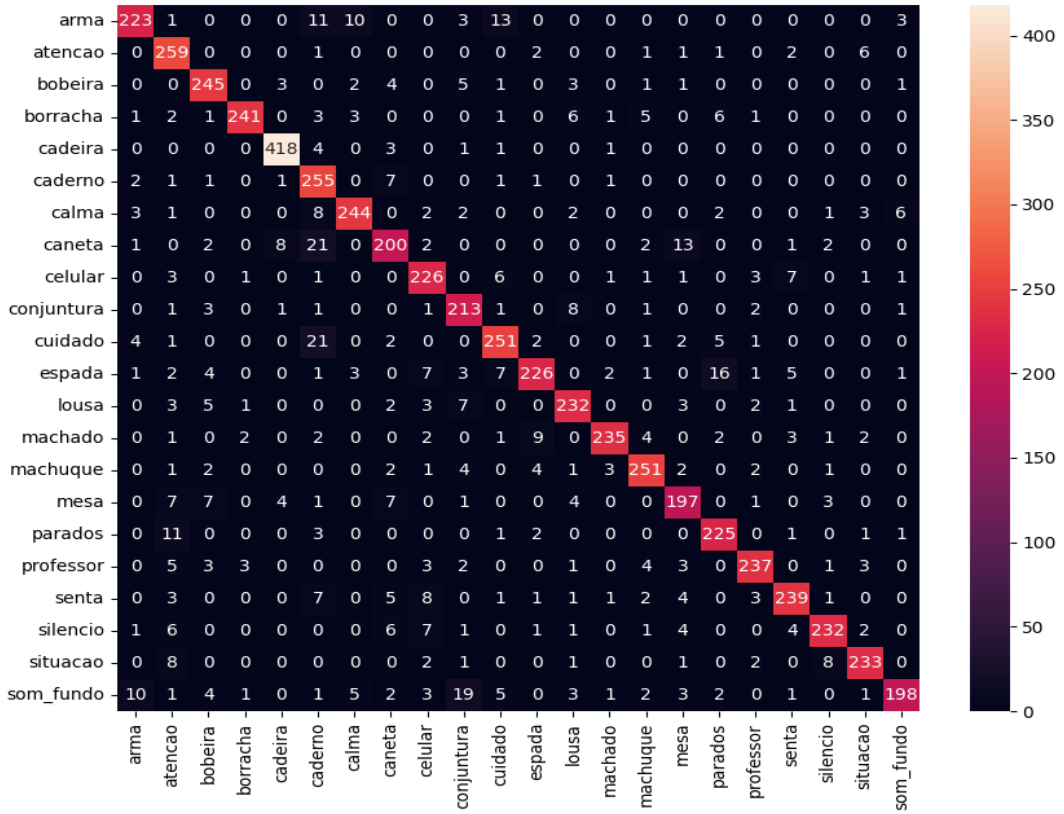
Fonte: O autor

Nos áudios coletados, também foi percebido que algumas palavras com fonéticas parecidas apresentaram espectrogramas visualmente idênticos, o que muitas vezes explica a identificação

errônea da palavra, podendo gerar alarmes falsos no sistema de segurança se o modelo não for treinado e validado devidamente.

A Figura 4 apresenta uma matriz de confusão, para ela é usada parte dos dados obtidos inicialmente feita para realizar o teste. Observou-se que 6% das palavras despertariam o alarme falso.

Figura 4 – Matriz de confusão das palavras selecionadas



Fonte: O autor

A validação é feita de forma que uma palavra seja testada com todas as outras palavras de forma aleatória e no final são apresentadas quantas vezes a palavra testada foi reconhecida como outra e como a palavra correta. Apesar das perturbações presentes no ambiente (Figura 2) e palavras com som parecido (Figura 3), a Figura 4 mostra uma diagonal precisa, em que a taxa de assertividade é de 89% de acerto na sua identificação. Na medição de tempo de processamento da classificação realizada pelo modelo, verificou-se que é necessário total de 30 milissegundos para obter a palavra a partir da voz. Além disso, a partir da Figura 4 não houve correlação entre palavras curtas e falsos alarmes positivos.

Discussão

A aplicação de redes neurais convolucionais na identificação de palavras específicas para a área de segurança em escolas demonstrou resultados promissores, conforme evidenciado pelos experimentos comparados às referências utilizadas. A escolha inicial das palavras-chave foi feita de forma empírica, mas planos futuros incluem a adoção de critérios mais fundamentados. Algumas métricas e informações adicionais solicitadas serão exploradas em etapas posteriores, dada a fase inicial do projeto.

Em comparação com trabalhos anteriores, que também utilizaram espectrogramas e RNC para reconhecimento de fala e sons específicos, observou-se que a precisão do modelo foi semelhante, sendo essa abordagem se destaca pela sua especificidade no contexto de segurança escolar, desde o ambiente de coleta de dados sujeito a perturbações, quanto a escolha das palavras. Ou seja, a personalização do modelo para identificar palavras críticas no contexto de segurança escolar apresenta

uma inovação relevante, contribuindo diretamente para a melhoria da segurança e prevenção de incidentes em instituições de ensino.

Ao analisar a sensibilidade do modelo gerado, verificou-se que o modelo não é sensível à variação de entonação da voz, nem a vozes mais graves ou mais agudas. Porém devido a vasta possibilidade de palavras com fonemas parecidos, percebeu-se que palavras comuns que não foram treinadas no modelo geraram alarmes falsos também.

Conclusão

Os resultados obtidos mostraram que a RNC foi capaz de identificar com precisão palavras específicas relacionadas à segurança, como "arma", "cuidado" e "machado". A análise dos espectrogramas revelou que características acústicas distintas dessas palavras foram corretamente reconhecidas pelo modelo, confirmando a eficácia do uso de espectrogramas na representação de dados de áudio para tarefas de classificação.

Apesar dos resultados positivos, alguns desafios e limitações foram identificados: a variabilidade na pronúncia das palavras por diferentes indivíduos (Figura 1) e a perturbação de fundo (Figura 2) presente em ambientes escolares representam desafios adicionais para a precisão do modelo. Além disso, a necessidade de um conjunto de dados abrangente e diversificado para treinamento é crucial para melhorar a robustez do modelo frente a diferentes situações e contextos.

Referências

CHOI, K.; FAZEKAS, G.; CHO, K.; SANDLER, M. **A Tutorial on Deep Learning for Music Information Retrieval**. 2018. arXiv. DOI: 10.48550/arXiv.1709.04396. Disponível em: https://doi.org/10.48550/arXiv.1709.04396. Acesso em: 20 jul. 2024.

CORDEIRO, M. **Detecção de edições em áudios baseada na análise tempo-frequência e em redes neurais convolucionais**. 2024. Dissertação (Mestrado em Engenharia Elétrica e Informática Industrial) - Universidade Tecnológica Federal do Paraná, Curitiba, 2023. Disponível em: http://repositorio.utfpr.edu.br/jspui/handle/1/33191. Acesso em: 04 abr. 2024.

GODOY, R.; FARINA, R. M.; FLORIAN, F. **INTELIGÊNCIA ARTIFICIAL ADAPTADA A IDOSOS**. 2021. Revista Interface Tecnológica, [S. l.], v. 18, n. 2, p. 208–218, 2021. DOI: 10.31510/infa.v18i2.1275. Disponível em: Revista FATEC. Acesso em: 3 abr. 2024.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. Disponível em: https://www.deeplearningbook.org/. Acesso em: 07 abr. 2024.

LUGER, G. F. **Inteligência artificial**. 6. ed. São Paulo: Pearson, 2013. E-book. Disponível em: Biblioteca Virtual. Acesso em: 04 abr. 2024.

OPPENHEIM, A. V.; SCHAFER, R. W. **Processamento em tempo discreto de sinais**. 3. ed. São Paulo, SP: Pearson, 2013. E-book. Disponível em: Biblioteca Virtual. Acesso em: 07 abr. 2024.

PASSOS, M.; LUCIENE, S.; AGUIAR, B. G.; FECHINE, J. M. et al. **Um ambiente para processamento digital de sinais aplicado à comunicação vocal homem-máquina**. Revista Principia - Divulgação Científica e Tecnológica do IFPB, João Pessoa, n. 14, p. 25-31, 2006. DOI: 10.18265/1517-03062015v1n14p25-31. Disponível em: Periódicos IFPB. Acesso em: 07 jul. 2024.

WARDEN, P. **Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition**. 2018. arXiv. DOI: 10.48550/arXiv.1804.03209. Disponível em: Coronell University. Acesso em: 20 jul. 2024.