Reconhecimento de Palavras-Chave com CNNs para Ajuda Silenciosa a Vítimas de Violência Doméstica

Diana Laura Fernández Duarte
National Institute of Telecommunication - Inatel
Santa Rita do Sapucaí, Brazil
diana.duarte@mtel.inatel.br

I. Introdução

A violência de gênero é um problema crítico que afeta diariamente mulheres de todas as classes sociais, raças ou religiões. De acordo com um relatório publicado pela Entidade das Nações Unidas para a Igualdade de Gênero e o Empoderamento das Mulheres em novembro de 2024, aproximadamente 736 milhões de mulheres já sofreram violência física ou sexual por parte de seus parceiros, o que representa quase um terço da população feminina mundial [1]. Especificamente no Brasil, no ano de 2022, segundo dados coletados pelo Fórum Brasileiro de Segurança Pública, 34.5% do total de homicídios contra mulheres ocorreram dentro de suas próprias residências, sendo cometidos, em sua maioria, por homens conhecidos das vítimas [2].

Diante desse panorama alarmante, torna-se necessário o desenvolvimento de sistemas que permitam às vítimas solicitar ajuda de forma rápida e discreta. Nesse contexto, as tecnologias de reconhecimento de palavras-chave (KWS) constituem uma alternativa promissora, ao possibilitar a identificação de comandos ou frases específicas em fluxos de áudio, sem a necessidade de interação física com o dispositivo. Trata-se de uma técnica amplamente utilizada para habilitar a interação por voz em dispositivos inteligentes, como os assistentes virtuais em smartphones (Siri, Alexa e Google Assistant) e os sistemas de automação residencial (Google Home e Amazon Echo) [3].

Para a implementação de sistemas de KWS, as técnicas de aprendizagem profunda são geralmente utilizadas, em particular as redes neurais convolucionais (CNN), devido à sua capacidade de modelar as correlações locais no tempo e na frequência dos sinais de áudio. Diferentemente das Redes Neurais Totalmente Conectadas, nas CNNs apenas algumas conexões são estabelecidas entre camadas convolucionais adjacentes, o que reduz significativamente o uso de memória. Além disso, as operações ponto a ponto realizadas entre os dados de entrada e os kernels são computacionalmente menos custosas do que as operações matriciais, o que melhora a eficiência do processamento [4], [5].

A. Trabalhos Relacionados

Estudos recentes têm explorado o uso de KWS em ambientes de emergência. É o caso do estudo [6], no qual se propõe um sistema automático para a detecção de pedidos de socorro em elevadores, utilizando uma combinação de KWS e análise paralinguística (PA). Para a detecção da palavra-chave definida "jiu ming", é utilizada uma rede neural composta por duas camadas Long Short-Term Memory (LSTM) empilhadas, uma camada intermediária de average pooling e uma camada final totalmente conectada. O estudo [7] também propõe um sistema de reconhecimento por KWS voltado a situações de emergência. O modelo foi treinado para detectar a palavra "help" em diferentes idiomas (inglês, árabe, curdo e malaio), utilizando a plataforma Edge Impulse. Posteriormente, foi quantizado para inteiros de 8 bits (int8), visando uma implementação eficiente no microcontrolador Arduino Nano 33 BLE Sense.

Com o objetivo de melhorar a capacidade de resposta rápida em situações de emergência sanitária, o artigo [8] propõe um sistema de KWS em áudios contínuos. O conjunto de dados utilizado foi composto por 42 mensagens oficiais da UNESCO relacionadas à saúde pública durante a pandemia da COVID-19. A fim de identificar um total de 62 palavras-chave, os autores treinaram e compararam dois modelos de redes neurais convolucionais: ResNet-18 e ResNet-152. No estudo [9], também é proposto um sistema de reconhecimento de comandos de voz baseado em CNN, utilizando o conjunto de dados Speech Commands do Google. O sistema foi projetado para uso em dispositivos móveis e aplicações como assistentes de voz ou situações de emergência. O modelo alcançou uma precisão de 94.5 %, com um total de aproximadamente 244.400 parâmetros, o que o torna adequado para ambientes com recursos computacionais limitados.

Até onde se tem conhecimento, são poucos os estudos que direcionam seus esforços para o auxílio de vítimas em situações de agressão, especificamente para ajudar mulheres que sofrem violência doméstica. Os autores do estudo [10] propuseram um modelo de aprendizado de máquina (ML) para reconhecer se uma porta foi fechada de forma agressiva ou normalmente, com o objetivo de contribuir para o monitoramento de comportamentos potencialmente agressivos e a detecção precoce de indícios de violência doméstica. Para isso, foram registradas múltiplas amostras de aceleração durante o fechamento da porta, bem como sinais de áudio, utilizando o acelerômetro e o microfone integrados na placa Arduino Nano 33 BLE Sense. Os sinais de áudio foram processados com a biblioteca Librosa em Python e, posteriormente, todas as amostras foram carregadas na plataforma Edge Impulse, onde foram utilizadas para treinar uma CNN capaz de classificar os

eventos em duas categorias: slam ou normal close.

Embora este estudo represente uma proposta inovadora para a detecção precoce de comportamentos agressivos, seu enfoque está centrado no monitoramento do ambiente físico, o que pode limitar sua capacidade de resposta imediata diante de situações críticas. Além disso, em muitos casos fatais, os atos de violência são cometidos por ex-parceiros que já não convivem com a vítima, o que reduz a efetividade de um sistema dependente do ambiente físico compartilhado.

B. Contribuições

Com o objetivo de oferecer uma solução em que a rapidez e a discrição sejam essenciais para salvar vidas, propõe-se um sistema de KWS baseado em CNN, ativado pela própria vítima e projetado para acionar um protocolo silencioso de ajuda em tempo real. Esse protocolo pode incluir ações como o compartilhamento da localização, o envio de mensagens de socorro para contatos de emergência ou até mesmo a notificação discreta às autoridades. O sistema será executado diretamente no telefone celular da vítima, a fim de garantir maior portabilidade. Essa proposta se destaca por sua discrição, pois não exige que a vítima interaja fisicamente com o celular, além de utilizar palavras-chave de uso cotidiano como gatilho de ativação, com o objetivo de reduzir o risco. Para o seu desenvolvimento, será utilizada a plataforma Edge Impulse, que oferece uma interface amigável para a criação de soluções baseadas em Aprendizado de Máquina, otimizadas para dispositivos embarcados.

II. METODOLOGIA

A seguir, são descritas as etapas que compõem o desenvolvimento deste projeto, desde a coleta de dados até sua implantação em dispositivos móveis.

A. Coleta de Dados

Para a construção da base de dados, serão coletadas amostras de áudio correspondentes às seguintes cinco classes:

- Morango_Diana
- Pipoca_Diana
- Morango_Outro
- Pipoca_Outro
- Ruído

Para maior robustez, as gravações serão realizadas em diferentes locais e sob diversos níveis de interferência acústica. A classe Ruído será composta por amostras de áudio provenientes da própria base de dados pública de Keyword Spotting da plataforma Edge Impulse, especificamente das classes noise e unknown, além da inclusão de amostras de silêncio. Todas as amostras terão duração de 1 segundo e frequência de amostragem de 16 kHz.

B. Pré-processamento dos Dados

Como os modelos de aprendizado de máquina não operam diretamente sobre dados brutos, é necessário extrair as informações linguísticas mais relevantes e suprimir elementos irrelevantes, como o ruído de fundo ou os silêncios da gravação. Esses componentes representativos do sinal de áudio são conhecidos como características, e constituem a base para o treinamento do modelo.

Entre os blocos de processamento oferecidos pela plataforma Edge Impulse, o mais adequado para o tratamento da fala humana é o de Mel-Frequency Cepstral Coefficients (MFCC). A escala Mel, de natureza não linear, baseia-se na percepção humana do som, que é mais sensível a variações em frequências baixas do que em frequências altas. Para o cálculo dos MFCC, o sinal de áudio é dividido em janelas temporais, sobre as quais se aplica a Transformada Discreta de Fourier (DFT). Em seguida, são aplicados filtros passa-faixa, espaçados linearmente nas frequências baixas e logaritmicamente nas frequências altas, com o objetivo de associar a frequência percebida pelo ouvido humano à frequência real medida. Finalmente, aplica-se a Transformada Discreta do Cosseno (DCT) aos valores logarítmicos das energias obtidas a partir do banco de filtros Mel, para selecionar os coeficientes mais representativos do conteúdo fonético do sinal [11].

C. Projeção do Modelo

Quanto à arquitetura da CNN, é necessário definir o número de camadas convolucionais, bem como o número de neurónios e de kernels que compõem cada uma dessas camadas. Essa configuração também deve ser definida para a camada final densamente conectada, responsável pela tarefa de classificação. Normalmente, camadas de pooling são adicionadas para reduzir o tamanho dos mapas de caraterísticas, de modo a manter apenas os mais representativos. A utilização da técnica de dropout, que consiste em desativar aleatoriamente uma percentagem de neurónios durante o treino, de forma a evitar o sobreajuste, pode ser relevante. No contexto deste projeto, é essencial encontrar um equilíbrio entre precisão e complexidade, de forma a não comprometer os recursos do dispositivo móvel onde será implementado.

D. Treinamento do Modelo

É importante definir corretamente o número de épocas, o fator de aprendizagem e o limiar mínimo de confiança para que uma previsão seja considerada válida, uma vez que estes parâmetros influenciam diretamente o processo de aprendizagem da rede neuronal. O número de épocas definido deve ser suficiente para que o modelo atinja a convergência, mas sem levar a um sobreajuste. No caso do fator de aprendizagem, este deve acelerar a convergência, mas sem gerar oscilações em torno do ponto de mínimo. Estes parâmetros serão ajustados através de um processo de tentativa-e-erro, após vários ensaios experimentais. Uma alternativa para evitar o sobreajuste é o aumento de dados, que modifica aleatoriamente os dados durante cada ciclo de treino, adicionando ruído ou mascarando bandas temporais ou de frequência. O modelo de aprendizagem automática será treinado com 80% das amostras, enquanto os restantes 20% serão utilizados para testar o modelo.

E. Avaliação do Modelo e Otimização

Para avaliar o desempenho do modelo, será medido o nível de acurácia, ou seja, a percentagem de dados que

foram corretamente classificados. A matriz de confusão, que compara, para cada classe, os rótulos reais com as previsões, também será analisada. Outra métrica importante a ter em conta é a perda ou erro, cujo valor esperado é o mínimo possível, uma vez que fornece uma medida da distância entre as previsões do modelo e os resultados reais. Tanto o erro do conjunto de treino como o erro de validação devem diminuir ao longo do treino até se tornarem praticamente constantes e pequenos. Se o modelo não convergir corretamente, é necessário ajustá-lo manualmente ou através de técnicas de otimização paramétrica.

F. Inferência

Finalmente, o modelo treinado e avaliado será utilizado para reconhecer, em tempo real, se o áudio captado contém uma das palavras-chave para a ativação do protocolo silencioso de assistência à vítima.

III. Conclusões

Este sistema de KWS destinado ao atendimento de vítimas de violência doméstica representa uma proposta de grande relevância, contribuindo para a preservação da vida das vítimas e para a tentativa de redução dos alarmantes números actuais de violência sexual e psicológica contra as mulheres. Embora este projeto tenha sido inicialmente concebido para assistir as mulheres vítimas de maus-tratos por parte de seus parceiros, a sua abordagem pode ser alargada a outras situações de risco enfrentadas pelas mulheres no dia a dia, graças à sua natureza discreta e à sua capacidade de operar autonomamente em dispositivos móveis.

References

[1] UN Women, "Facts and figures: Ending violence against women," November 2024.

[2] Fórum Brasileiro de Segurança Pública, "Anuário brasileiro de segurança pública," 2022.

[3] S. Rai, T. Li, and B. Lyu, "Keyword spotting – detecting commands in speech using deep learning," 2023.

[4] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, p. 53, 2021.

[5] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge: Keyword spotting on microcontrollers," 2018.

[6] H. Chu, Y. Wang, R. Ju, Y. Jia, H. Wang, M. Li, and Q. Deng, "Call for help detection in emergent situations using keyword spotting and paralinguistic analysis," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, ICMI '21 Companion, (New York, NY, USA), p. 104–111, Association for Computing Machinery, 2021.

[7] B. V. Nived, K. Jamal, G. Mahesh, and R. M. Kumar, "Design of custom keyword recognition using edge impulse on arduino nano 33 ble sense," in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 1522–1529, 2023.

[8] S. Jaballi, M. J. Hazar, S. Zrigui, H. Nicolas, and M. Zrigui, "Resnet-based pandemic keyword spotting in continuous multilingual speech: A study in unesco's audio messages for rapid health response," 2025.

[9] X. Li and Z. Zhou, "Speech command recognition with convolutional neural network." https://cs229.stanford.edu/proj2017/final-reports/5244201.pdf, 2017.

[10] O. Morgan, H. Kayan, and C. Perera, "Poster abstract: Feasibility on detecting door slamming towards monitoring early signs of domestic violence," in *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pp. 141–142, 2022.

[11] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022.