On-Device KWS and Speaker Recognition System for Discreet Emergency Assistance on Android

Diana Laura Fernández Duarte, Felipe A. P. de Figueiredo, and Victoria D. P. Souto
National Institute of Telecommunications
Santa Rita do Sapucaí, Brazil
diana.duarte@mtel.inatel.br, [felipe.figueiredo, victoria.souto]@inatel.br

***Abstract*—This paper presents an Android-compatible system for discreet emergency assistance, specifically targeting victims of domestic violence. We propose a hierarchical Convolutional Neural Network (CNN) architecture combining Keyword Spotting (KWS) and Speaker Recognition (SR) to minimize false alarms. The KWS module detects innocuous Portuguese keywords, "Morango" (strawberry) and "Pipoca" (popcorn), in 1-second audio clips, while the Speaker Verification module confirms the victim's identity (enrolled as "Diana"). Models were trained using MFCC spectrograms (40 Mel filters, 13 coefficients per window) and optimized via hyperparameter tuning. Implemented on-device through a web application, the system achieves more than 98% accuracy on test sets (KWS: 98.31% accuracy, SR: 98.84% accuracy) with low loss ($\leq$ 0.038). Developed using Edge Impulse, this solution demonstrates robust real-time capability on mobile devices, offering a privacy-preserving alternative for emergency intervention. The system's source code is publicly available[1].**

***Index Terms*—KWS, Speaker Recognition, CNN, Domestic violence**

I. Introduction

Gender-based violence is a critical problem that affects women of all social classes, races, and religions daily. According to a report published by the United Nations Entity for Gender Equality and the Empowerment of Women in November 2024, approximately 736 million women have suffered physical or sexual violence from their partners, which represents almost a third of the world's female population [1]. Specifically in Brazil, in 2022, according to data collected by the Brazilian Public Security Forum, 34.5% of all homicides against women occurred inside their own homes, most of which were committed by men known to the victims [2].

Faced with this alarming scenario, it is necessary to develop systems that allow victims to request help quickly and discreetly. In this context, Keyword Spotting (KWS) technologies offer a promising alternative, enabling the identification of specific commands or phrases in audio streams without requiring physical interaction with the device. It is a technique widely used to enable voice interaction in smart devices, such as virtual assistants in smartphones (Siri and Alexa) [3] and home automation systems (Google Home and Amazon Echo) [4].

This work was partially funded by CNPq (Grant Nos. 403612/2020-9, 311470/2021-1, 403827/2021-3, and 306199/2025-4), by FAPEMIG (Grant Nos. PPE-00124-23, APQ-04523-23, APQ-05305-23, and APQ-03162-24), and Brasil 6G project (01245.020548/2021-07), supported by RNP and MCTI.

[1]https://github.com/Diana9908/TP557

For the implementation of KWS systems, deep learning (DL) techniques are generally employed, particularly convolutional neural networks (CNNs), due to their ability to model the local correlations in time and frequency of audio signals [5]. Unlike fully connected neural networks (DNNs), in CNNs, only a few connections are established between adjacent convolutional layers, which significantly reduces memory usage. Additionally, the point-to-point operations performed between the input data and the kernels are computationally less expensive than matrix operations, thereby improving processing efficiency [6].

However, systems based exclusively on KWS are vulnerable to false positives, since anyone in the environment, intentionally or accidentally, can activate the system. An alternative method to ensure that only the victim can call for help is to use SR models, which allow for the identification of the speaker based on their biometric voice characteristics [7].

A. Related Work

Recent studies have investigated the application of KWS in emergency settings. This is the case of [8], which proposes a system for detecting calls for help in elevators, using a combination of KWS and paralinguistic analysis (PA). In [9], the authors propose a KWS recognition system for emergencies. The model was trained to detect the word "help" in different languages (English, Arabic, Kurdish, and Malay) using the Edge Impulse platform and was implemented on the Arduino Nano 33 BLE Sense microcontroller.

To improve rapid response capacity in health emergencies, [10] proposes a continuous audio KWS system. The dataset used consisted of 42 official UNESCO messages related to public health during the COVID-19 pandemic. To identify a total of 62 keywords, the authors evaluated two CNN models: ResNet-18 and ResNet-152.

SR was investigated in [11], which presents a CNN-based classifier for identifying speakers in stressful environments. The system achieved an average accuracy of 81.6% on the Emirati-accented corpus. The authors of [12] propose an innovative system for performing SR directly on TinyML devices. The Adaptive Speaker Verification (ASV) module extracts d-vectors. These low-dimensional representations capture the distinctive characteristics of a speaker's voice, using a pre-trained CNN to build the enrollment set during the registration phase. During the inference phase, a new extracted d-vector is compared with the stored vectors using cosine similarity. The model was tested on an Infineon PSoC 62S2 Wi-Fi BT Pioneer

TABLE I: Work related to KWS and SR.

Reference	Data Preprocessing	KWS	Speaker Recognition	PA	On-Device Implementation
[8]	Fbank, MFCC	LSTM	-	ResNet-18	-
[9]	MFCC	CNN	-	-	Arduino Nano 33 BLE Sense
[10]	MFCC + M-BERT Embedding	ResNet-18, ResNet-152	-	-	-
[11]	MFCC, MFCC-delta, MFCC-delta-delta	-	CNN	-	-
[12]	MFCC	CNN	d-vector extractor (CNN) + adaptive instance-based model	-	Infineon PSoC 62S2 Wi-Fi BT
This work	MFCC	CNN	CNN	-	smartphone

Kit. Table I establishes a comparison between the previously presented works and our proposal.

Few studies aim to support people in aggressive situations, especially women facing domestic violence. In [13], the authors built a model to detect if a door was slammed or closed normally, helping identify aggressive behavior. They used the Arduino Nano 33 BLE Sense to collect audio and motion data, then trained a neural network in Edge Impulse to classify the door events.

B. Contributions

To provide a quick and discreet solution, a CNN-based help system is proposed, featuring a hierarchical architecture comprising two cascaded modules. The first module is responsible for detecting two everyday keywords: ”Morango” and ”Pipoca”. The second module verifies that the spoken word matches the victim’s voice, thereby reducing the likelihood of false positives. Specifically, this module was trained to distinguish between the speaker ”Diana” and ”Others”. This system was integrated into an Android environment via a web application, allowing it to be run directly on the victim’s cell phone. The models were trained using Edge Impulse, the latter providing an intuitive environment for developing machine learning (ML) solutions on embedded devices. The overall functioning of the system is illustrated in Figure 1.

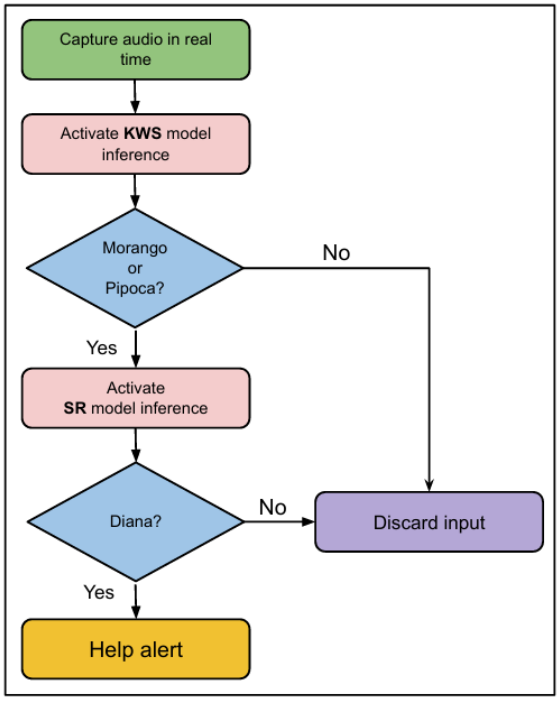

Fig. 1: System Flow Diagram.

II. Methodology

Below are the steps that make up the development of this project, from data collection through to its deployment on mobile devices.

A. Data collection

To build the database, samples of different voices, both female and male, pronouncing the keywords ”Morango” and ”Pipoca” were taken. Noise and silence samples were also collected using Edge Impulse’s own public Keyword Spotting database. The samples were organized into different classes according to the specific objectives of each model. For the KWS module, 1630 samples were collected from each of the classes: ”Morango”, ”Pipoca”, and ”Noise”, totaling 4890 samples. The choice of these keywords is based on their phonetic diversity, which enables the construction of a more comprehensive vocal profile of the user. For the SR module, 2160 samples were collected from each of the classes, ”Diana” and ”Others”, totaling 4320 samples.

In addition, a reasonable balance was maintained between the classes in the training, validation, and test sets: 60% of the total samples comprise the training set, 20% comprise the validation set, and the remaining 20% comprise the test set. All samples have a duration of 1 second and a sampling frequency of 16 kHz.

B. Data preprocessing

The audio samples, before being used as inputs for both models, are pre-processed and transformed into Mel-frequency cepstral coefficient (MFCC) spectrograms to extract the most relevant linguistic information. The Mel scale, which is non-linear, is ideal for speech processing because it is based on

the human perception of sound, which is more sensitive to variations at low frequencies than at high frequencies.

To calculate the MFCC, the audio signal is divided into time windows, to which the Fast Fourier Transform (FFT) is applied. Next, band-pass filters are applied, spaced linearly at low frequencies and logarithmically at high frequencies, to associate the frequency perceived by the human ear with the actual frequency measured. Finally, the Discrete Cosine Transform (DCT) is applied to the logarithmic values of the energies obtained from the Mel filter bank, to select the coefficients most representative of the phonetic content of the signal [14].

The result is a spectrogram of dimension $i \times j$, where i represents the number of frequency bands extracted per window and j is the number of windows that make up an audio sample. The value of j is given by $j = \left\lfloor \frac{W-(\lambda-\phi)}{\phi} \right\rfloor$, where W represents the total duration of the audio sample, λ represents the duration of each window and ϕ is the step between windows. Finally, the spectrogram is flattened to be used as input for the classification models.

In this project, each audio sample was segmented into windows of 32 ms duration, with no overlap between consecutive windows. A 512-point FFT was applied to each window, followed by a bank of 40 Mel filters, distributed over the frequency range from 0 Hz to 8 kHz. The first 13 coefficients per window, which contain the most relevant linguistic information, were extracted. This results in spectrograms of dimension 31×13, which are flattened to a vector of dimension 1×403. Finally, each MFCC spectrogram was normalized using the Local Normalization of Mean and Cepstral Variance technique with a sliding window covering 101 consecutive frames.

C. KWS module

The module consists of a CNN trained to solve a multi-class classification task, being able to distinguish whether the input audio sample belongs to one of the following classes: "Morango", "Pipoca", or "Noise".

The model is composed of a first layer that corrupts the input data by adding additive Gaussian noise to prevent overfitting. Next, a Reshape layer reorganizes the MFCC vectors into a three-dimensional tensor with the format $31 \times 13 \times 1$ suitable for processing by 2D convolutional layers. The model also consists of two blocks of Conv2D, MaxPooling2D, and Dropout layers. After these blocks, a GlobalAveragePooling2D layer is added, which reduces the dimensions of the feature maps. The resulting vector is processed by a Dense layer with a ReLU activation function, followed by an additional Dropout layer. Finally, the output Dense layer with a softmax activation function classifies the audio sample into one of the three defined classes.

The model was trained on the Edge Impulse platform for up to 100 epochs. To prevent overfitting, the Early Stopping technique was applied with a patience of 15 epochs. The training data was artificially augmented using the SpecAugment technique proposed in [15], which involves applying random masks in both the time and frequency domains. The optimizer used was the Adaptive Moments (Adam), with an initial learning rate of 0.0007, which was progressively reduced every 10 epochs using Keras Tuner's LearningRateScheduler class. Table II shows the best hyperparameters found.

TABLE II: Hyperparameters of the KWS model.

Layers	Hyperparameters
GaussianNoise	stddev = 0.2
Reshape	-
Conv2D	filters = 64, kernel = 3x3, stride = 1
MaxPooling2D	pool_size = 2x2, stride = 2
Dropout	rate = 0.15
Conv2D	filters = 128, kernel = 3x3, stride = 1
MaxPooling2D	pool_size = 2x2, stride = 2
Dropout	rate = 0.15
GlobalAveragePooling2D	–
Dense	units = 256
Dropout	rate = 0.35
Dense	units = 3

D. SR Module

The SR module corresponds to a binary classifier that distinguishes between the "Diana" class and the "Others" class. The model begins with a Reshape layer, followed by five blocks composed of Conv2D, BatchNormalization, ReLU, MaxPooling2D, and Dropout layers. This is followed by a GlobalAveragePooling2D layer. The classification stage includes the following sequence of layers: Dense, BatchNormalization, ReLU, Dropout and a final output Dense layer.

This model was also trained on the Edge Impulse platform using the Early Stopping technique with a patience of 15 epochs. The adopted optimizer was Adam, with a learning rate of 0.007. Table III presents the architecture of this model along with the corresponding hyperparameters for each layer.

E. On-device implementation

Both models have been integrated into a web application that captures audio in real time. The audio is initially processed by the KWS model, which is responsible for detecting the defined keywords. If any of them are recognized, the audio is sent to the second model, which checks whether the voice corresponds to "Diana". If so, the system issues an emergency alert. Finally, the application was deployed in an Android environment.

III. Results and Discussions

The performance of the proposed KWS and SR models was evaluated using the validation and test sets, with metrics including accuracy, precision, recall, F1-score, and loss. The detailed results are presented in Table IV.

These results demonstrate strong generalization capability for both models. The KWS module achieved a test accuracy of 98.31%, with a particularly high recall of 99.80%, indicating its ability to reliably detect the defined emergency keywords even in unseen audio samples. Such high recall is crucial in safety-oriented applications, where missed detections could have serious consequences.

The SR module also delivered high performance, with a test accuracy of 98.84% and balanced precision and recall. These

TABLE III: Hyperparameters of the SR model.

Layer	Hyperparameters
Reshape	-
Conv2D	filters = 32, kernel = 6x6
BatchNormalization	-
ReLU	-
MaxPooling2D	pool_size = 2x2, strides = 2
Dropout	rate = 0.1
Conv2D	filters = 64, kernel = 4x4
BatchNormalization	-
ReLU	-
MaxPooling2D	pool_size = 2x2, strides = 2
Dropout	rate = 0.1
Conv2D	filters = 64, kernel = 4x4
BatchNormalization	-
ReLU	-
MaxPooling2D	pool_size = 2x2, strides = 2
Dropout	rate = 0.1
Conv2D	filters = 128, kernel = 2x2
BatchNormalization	-
ReLU	-
MaxPooling2D	pool_size = 2x2, strides = 2
Dropout	rate = 0.1
Conv2D	filters = 128, kernel = 2x2
BatchNormalization	-
ReLU	-
MaxPooling2D	pool_size = 2x2, strides = 2
Dropout	rate = 0.1
GlobalAveragePooling2D	-
Dense	units = 256
BatchNormalization	-
ReLU	-
Dropout	rate = 0.12
Dense	units = 2

TABLE IV: Performance of the Models

Model	Dataset	Accuracy (%)	Precision (%)	F1-Score (%)	Recall (%)	Loss
KWS	Validation	99.40	99.41	99.41	99.41	0.020
KWS	Test	98.31	98.81	98.80	99.80	0.038
SR	Validation	99.9	99.88	99.88	99.88	0.007
SR	Test	98.84	99.19	99.19	99.19	0.018

results suggest the model is effective in correctly verifying the speaker's identity and reducing false positives triggered by others speaking the keywords.

Importantly, both models maintained low loss values ($\leq$ 0.038), indicating proper convergence and stable training. The consistent performance between validation and test sets further confirms the robustness of the models and the effectiveness of the preprocessing pipeline based on MFCCs and data augmentation with SpecAugment.

Compared to previous works, which either focused on single-module systems or required dedicated hardware, such as Arduino Nano BLE or Infineon PSoC boards, the proposed dual-module solution achieves high accuracy while remaining fully functional in a smartphone environment. This enhances the feasibility and scalability of the system, especially in contexts where user privacy and device ubiquity are essential.

IV. Conclusions

This work presents a high-accuracy ($> 98\%$) and privacy-focused emergency system for Android that empowers domestic violence victims through discreet voice-activated alerts. By cascading CNN-based Keyword Spotting (detecting innocuous Portuguese keywords) and Speaker Verification (authenticating enrolled users), the solution minimizes false alarms while operating entirely on-device. Implemented via Edge Impulse with real-time capability, it offers a scalable, privacy-preserving alternative for urgent assistance. Future work will expand speaker and keyword diversity and integrate emotion recognition. The system's open-source availability and low-resource deployment potential highlight its life-saving applicability in critical scenarios.

References

[1] UN Women, "Facts and figures: Ending violence against women," November 2024. [Online]. Available: https://www.unwomen.org/en/articles/facts-and-figures/facts-and-figures-ending-violence-against-women

[2] Fórum Brasileiro de Segurança Pública, "Anuário brasileiro de segurança pública," 2022. [Online]. Available: https://forumseguranca.org.br/wp-content/uploads/2022/06/anuario-2022.pdf

[3] S. Rai, T. Li, and B. Lyu, "Keyword spotting – detecting commands in speech using deep learning," 2023. [Online]. Available: https://arxiv.org/abs/2312.05640

[4] J. Mishra, T. Malche, and A. Hirawat, "Embedded intelligence for smart home using tinyml approach to keyword spotting," *Engineering Proceedings*, vol. 82, no. 1, 2024. [Online]. Available: https://www.mdpi.com/2673-4591/82/1/30

[5] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge: Keyword spotting on microcontrollers," 2018. [Online]. Available: https://arxiv.org/abs/1711.07128

[6] 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. [Online]. Available: https://doi.org/10.1186/s40537-021-00444-8

[7] A. Q. Ohi, M. F. Mridha, M. A. Hamid, and M. M. Monowar, "Deep speaker recognition: Process, progress, and challenges," *IEEE Access*, vol. 9, pp. 89 619–89 643, 2021.

[8] 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*, ser. ICMI '21 Companion. New York, NY, USA: Association for Computing Machinery, 2021, p. 104–111. [Online]. Available: https://doi.org/10.1145/3461615.3491111

[9] D. Nabaz, N. Shafie, and A. Azizan, "Design of emergency keyword recognition using arduino nano ble sense 33 and edge impulse," *Open International Journal of Informatics*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:270890933

[10] 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. [Online]. Available: https://www.researchgate.net/publication/389396172_ResNet-Based_Pandemic_Keyword_Spotting_in_Continuous_Multilingual_Speech_A_Study_in_UNESCO's_Audio_Messages_for_Rapid_Health_Response

[11] I. Shahin, A. B. Nassif, and N. Hindawi, "Speaker identification in stressful talking environments based on convolutional neural network," *International Journal of Speech Technology*, vol. 24, no. 4, pp. 1055–1066, 2021.

[12] M. Pavan, G. Mombelli, F. Sinacori, and M. Roveri, "Tinysv: Speaker verification in tinyml with on-device learning," in *Proceedings of the 4th International Conference on AI-ML Systems*, ser. AIMLSystems 2024. ACM, Oct. 2024, p. 1–10. [Online]. Available: http://dx.doi.org/10.1145/3703412.3703415

[13] 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)*, 2022, pp. 141–142.

[14] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, ser. interspeech_2019. ISCA, Sep. 2019.