AI-Driven Real-Time Distress Detection Through Speech Recognition for Emergency Response Systems

Malvina Halilaj[1†], **Elektra Myrto**[2*,†] **and Aldo Franco Dragoni**[3†]

[1] *Department of Information Engineering, Polytechnic University of Marche Ancona, Italy m.halilaj@pm.univpm.it*

[2] *Department of Computer Science, University of Tirana, Albania elektra.myrto@fshn.edu.al*

[3] *Department of Information Engineering, Polytechnic University of Marche Ancona, Italy a.f.dragoni@univpm.it*

Abstract

Violence against women and children remains a critical global issue, requiring immediate and innovative interventions. Traditional emergency response systems heavily rely on manual reporting, which may not be feasible in life-threatening situations. This paper introduces an AI-driven voice recognition model designed to detect distress signals in real time. The proposed system leverages deep learning techniques, specifically trained on emotionally labeled speech datasets, to classify distress calls and trigger emergency alerts when necessary.

The system consists of a real-time audio capture module, a feature extraction component that processes speech signals, and a deep learning model trained to recognize distress speech patterns. It compares multiple feature extraction methods, including MFCCs and spectrogram-based approaches, and evaluates the performance of convolutional neural networks (CNNs) against state-of-the-art architectures such as Wav2Vec2 and Whisper. Results indicate that transformer-based models significantly outperform traditional CNNs, particularly in handling noisy environments and multilingual speech. The model has been successfully trained and evaluated, and an API has been developed to support real-time classification of audio input. While full mobile integration is still under development, these efforts demonstrate the feasibility of future deployment into mobile applications and IoT security devices for real-time emergency response.

Keywords

Voice recognition, AI, speech processing, deep learning, emotional cues, machine learning

1. Introduction

Timely intervention in cases of violence, particularly against women and children, is crucial. Conventional emergency communication methods, such as phone calls to emergency services, may not always be viable in high-risk situations. Hands-free, voice-activated devices capable of identifying distress signals can be life-saving. This study explores the implementation of an AI-powered violence detection model that automatically categorizes distress speech and facilitates rapid emergency response [1][2].

With the widespread adoption of mobile technologies, integrating this AI-based distress recognition system into a mobile application provides a seamless and practical approach to emergency response. A mobile-based implementation ensures accessibility, enabling users to discreetly trigger emergency alerts without needing manual intervention. The application is designed to capture and process real-time audio, extract key speech features, and employ deep learning models to classify distress speech. While the final mobile deployment is still under development, the system has already been trained and tested, and a functional API has been built to enable real-time classification.

Using TensorFlow Lite, models such as Wav2Vec2 and Whisper are being prepared for on-device processing to reduce latency and ensure real-time response [3][4]. Additionally, integrating this solution into smartphones allows for low-power, real-time inference, ensuring continuous monitoring without excessive battery consumption. The app can be further enhanced with edge AI techniques,

6th International Conference Recent Trends and Applications in Computer Science and Information Technology

* Corresponding author.

† These authors contributed equally.

m.halilaj@pm.univpm.it(M.Halilaj); elektra.myrto@fshn.edu.al (E.Myrto); a.f.dragoni@univpm.it (A.F.Dragoni)

0009-0001-1981-128X (M.Halilaj); 0009-0001-1601-2327 (E.Myrto); 0000-0002-3013-3424 (A.F.Dragoni)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4044/paper03.pdf

leveraging embedded AI processing within mobile hardware. By embedding this technology in wearable devices or IoT-based security systems, a proactive emergency response mechanism can be established, automatically notifying emergency contacts or authorities in case of distress detection [5].

The main contributions of this research are:

1. Development of a real-time distress detection system.
2. Integration of deep learning-based voice recognition models.
3. Comparative analysis of feature extraction methods (spectrogram-based features vs. MFCCs).
4. Evaluation of CNN-based architectures versus state-of-the-art models such as Whisper and Wav2Vec2.
5. Expansion of multilingual datasets to improve generalization and accessibility.
6. Deployment considerations for mobile applications and edge computing devices for real-time detection.

Speech recognition is the process of converting spoken language into machine-readable data. This can be achieved through traditional rule-based approaches or modern machine learning techniques. Rule-based systems, in use since the 1960s, require manual tuning and are labor-intensive to maintain [6]. In contrast, machine learning approaches allow models to automatically learn from training data, reducing the need for ongoing manual intervention and offering greater scalability. While training such models can be computationally expensive, they prove far more efficient and adaptable in the long run [7].

Speech recognition enables a system to interpret human speech and convert it into formats like text or structured commands that machines can understand and act upon. Depending on the specific application, this output may be used for real-time classification, transcription, or triggering responses, such as emergency alerts in the context of this research.[8]

2. Methodology

The research approach of this paper is aimed at making it easy to design a highly accurate and efficient distress voice detection system. The approach involves various phases like system architecture design, data acquisition, feature extraction, and deployment of the deep learning model. The goal is to develop a deployable real-time system that can perform well in emergency scenarios with minimal latency. The following subsections explain the components of the proposed approach in detail.

2.1. System Architecture

1. The proposed system consists of three primary components:
2. Audio Capture Module: Captures live audio using a mobile device or wearable technology.
3. Feature Extraction and Voice Processing: Extracts relevant features such as pitch, tone, frequency patterns, and emotional cues.
4. Machine Learning Model: A deep learning model trained to detect distress signals and trigger an alert when necessary.

2.2. Data Collection

The model is trained using the following datasets:

- Speech Commands Dataset:

An audio dataset of spoken words designed to help train and evaluate keyword spotting systems. Its primary goal is to provide a way to build and test small models that detect when a single word is spoken, from a set of ten target words, with as few false positives as possible from background noise or unrelated speech. Note that in the train and validation set, the label “unknown” is much more prevalent than the labels of the target words or background noise. One difference from the release version is the handling of silent segments. While in the test set the silence segments are regular 1 second files, in the training they are provided as long segments under “background_noise” folder. The dataset consists of **over 100,000** audio files, which are split into training and testing sets, with multiple recordings per word.

- RAVDESS Dataset: Provides emotionally labeled speech samples.

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset is a collection of audio and visual data designed for emotional speech and facial expression recognition research. It was created to support the development of systems that can recognize and understand human emotions from speech and facial expressions. his portion of the RAVDESS contains 1012 files: 44 trials per actor x 23 actors = 1012. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Song emotions includes calm, happy, sad, angry, and fearful expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

- Custom Dataset: Contains real-world distress calls recorded in emergency situations.
- Future Expansion: Plans to include speech samples in Albanian and Italian to enhance linguistic diversity.

2.1. Feature Extraction

Mel Frequency Cepstral Coefficients (MFCCs) are a set of features developed at MIT in the late 1960s for seismic audio echo analysis and simulating human voice characteristics. [9]They are simple sound characteristics used in various applications, including in this project, obtained by taking a discrete Fourier transform of a signal, applying a logarithm, and then a Fourier inverse. Although MFCC is used for feature extraction from input data of various domains, it is faced with many problems, which have not been addressed extensively in the literature. The objective of this paper is to provide an extensive review of MFCC and its applications like speech recognition, speaker recognition, emotion recognition, bearing fault detection, gear fault detection, Electrocardiogram (ECG) and Electroencephalogram (EEG) classification. Feature extraction is a crucial step for improving model performance. The study evaluates multiple techniques, including:

- **Mel-Frequency Cepstral Coefficients (MFCCs):** Captures short-term power spectrum features of speech.
- **Mel Spectrograms:** Provides time-frequency representations useful for deep learning models.
- **Chroma Features:** Captures harmonic content essential for recognizing distress patterns.
- **Root Mean Square Energy (RMSE):** Identifies distress signals by measuring energy variations in speech.
- Experimental results show that spectrogram-based features outperform MFCCs in distress emotion recognition.

3. Machine learning models

Machine learning models play a crucial role in the detection and classification of distress signals. The selection of an appropriate model impacts both accuracy and computational efficiency. This study evaluates two primary architectures: convolutional neural networks (CNNs) and transformer-based models.

3.1. CNN-Based Model:

CNN uses a standard neural network to solve classification problems [10][11]. For the study, the traditional model of CNN was tested with various variations. The best results from the stan- dard model include four convolutional layers and two fully connected layers and a flatten layer. Each of the convolutional layers has a valid padding parameter. The activation function is the rectified linear unit (ReLU) and the stride number is 1. In research that seeks to tune CNNs for speech recognition tasks, only max pooling on the frequency axis is added after the first convolutional layer. Therefore, we have also added max-pooling by the kernel size and stride value. We have applied dropout after all convolutional layers and fully connected layers.

Strengths: CNNs are widely used for speech and audio classification tasks due to their ability

to capture local dependencies and extract spatial features from spectrograms.

Limitations: While CNNs perform well for structured speech classification, they struggle with capturing long-range dependencies and context-based recognition of distress speech, particularly in noisy environments.

3.2. Transformer-Based Models (Wav2Vec2 & Whisper)

Automatic speech recognition (ASR) is the process of converting audio signals to strings of words. Speech recognition allows one to maintain records and interpret voice commands.

In the domain of Automatic Speech Recognition (ASR), several challenges persist, such as limited training data, untranscribed data, and difficulty in low-resource languages and children's speech. Recent research efforts have addressed some of these issues, leading to impressive ASR performance for adult speech, even achieving human-level performance

Whisper represents a significant advancement in weakly supervised pre-training[12], extending its capabilities to encompass multilingual and multitask scenarios beyond English-only speech recognition.

wav2vec 2.0 is a speech recognition model based on self-supervised learning of speech representations through a two stage architecture for pretraining and fine tuning[13]. The architecture comprises three key components: a CNN feature extractor, a transformer-based encoder, and a quantization module .[14]

Wav2Vec2: A self-supervised learning model trained on raw waveform data, Wav2Vec2 leverages unsupervised pretraining followed by fine-tuning on labeled datasets. This approach improves its ability to capture subtle emotional cues and enhances speech recognition performance under noisy conditions.

Whisper: Developed by OpenAI, Whisper is a multilingual, multitask ASR model trained on large-scale datasets. It exhibits robustness against varying accents, dialects, and noise interference, making it highly suitable for real-world distress detection scenarios[15].

4. Approach

Our goal with this study is to create a highly efficient AI-powered system that can recognize distress signals in real-time and provide immediate assistance. To achieve this, we focused on building a system that is accurate, responsive, and deployable on everyday devices like smartphones and smart home systems. Here's how we approached it:

4.1. Data Preprocessing

To ensure our model performs well in real-world situations, we prepared our data carefully:

- Reducing Background Noise: We filtered out unwanted sounds using spectral subtraction to make sure the distress signals are clear.
- Normalizing Speech: By adjusting the volume levels of different audio clips, we made sure our model doesn't get confused by loud or soft voices.
- Segmenting Recordings: Longer audio files were broken down into smaller parts to help the model focus on key distress cues.
- Enhancing Diversity: We used techniques like pitch shifting and time stretching to expose the model to different ways people might call for help.

4.2. Model Training and Optimization

We tested two different types of models to find the best approach: traditional CNN-based models and more advanced transformer-based architectures like Wav2Vec2 and Whisper. Our training process involved:

- Extracting Key Speech Features: We used MFCCs, spectrograms, and chroma features to highlight important vocal patterns.
- Splitting Data: The dataset was divided into 80% for training and 20% for testing to ensure a fair evaluation.
- Fine-Tuning Performance: We adjusted learning rates, batch sizes, and optimization techniques

to maximize accuracy.

- Measuring Success: We evaluated the models based on accuracy, precision, recall, and F1-score to ensure reliability.

4.3. API Design and Deployment on IoT

As part of this research, we have developed a working API prototype built using Flask to demonstrate the real-time deployment capability of our AI model. The API allows users to upload short audio recordings, which are processed and classified into distress or non-distress categories using pre-trained TensorFlow models.

This web API architecture includes:

1. Three deep learning models trained on the RAVDESS, Speech Commands, and speaker datasets.
2. Preprocessing pipeline using Librosa to extract MFCC features from incoming audio.
3. Label encoders for mapping model predictions to human-readable classes.

The RESTful endpoint /predict supports POST requests, accepting audio files and classifying them using the appropriate model, depending on the selected dataset. This API forms the foundation for a planned mobile application and can also be integrated into IoT devices for in- field emergency detection.Real-Time Deployment -Future

Future work involves converting this trained model to TensorFlow Lite format, allowing it to run efficiently on edge devices like smartphones or microcontrollers. This conversion step is essential for embedding AI into mobile apps and IoT security systems, enabling real-time, on- device voice-based distress detection without internet dependency. This architecture not only supports low-latency emergency recognition but also ensures privacy by processing data locally. Key factors we are considering :

- Fast Response Times: By optimizing for low-latency inference, the system can recognize distress signals almost instantly.
- Edge AI Capabilities: The model can process speech directly on the device without needing an internet connection, improving privacy and speed.
- Seamless Integration: We designed it to work effortlessly with mobile applications, allowing users to trigger alerts hands-free when they need help.

5. Tables

Ravdess Model Performance

Category	Examples	Avg F1-Score	Remarks
Top Performing Classes	Class 4, 8, 12, 20	â‰ˆ 0.90 – 0.98	Highly accurate
Moderate Classes	Class 1, 5, 14, 17	â‰ˆ 0.80 - 0.85	Acceptable recognition
Low Performing Classes	Class 30, 41, 43	< 0.20	needs improvement

Table 1

The overall accuracy across all 60 classes was approximately **31%**, which is expected due to class imbalance and the presence of difficult-to-classify or ambiguous samples.

However, the **macro average F1-score was 0.56**, showing relatively strong performance across high-confidence categories.

Speech Commands Performance

Category	Examples	Avg F1-Score	Remarks
Top Performing Commands	yes, no, stop, go	â‰ˆ 0.90 – 0.95	Well-trained,
Moderate Commands	left, right, up, down	â‰ˆ 0.70 - 0.80	Moderate performance
Low Performing Commands	tree, bird, unknown	< 0.60	Needs data balancing

Table 2

The overall performance of the model is promising for frequently used, well-defined keywords. Confusion Matrix insights revealed that most errors come from acoustic similar or rarely used words.

6. Figures

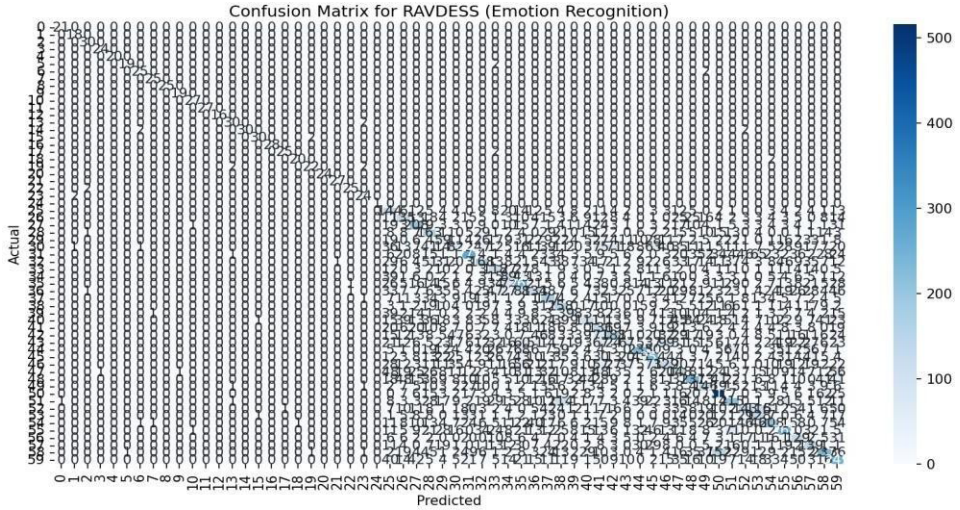

Figure 1: Ravdess Visualization

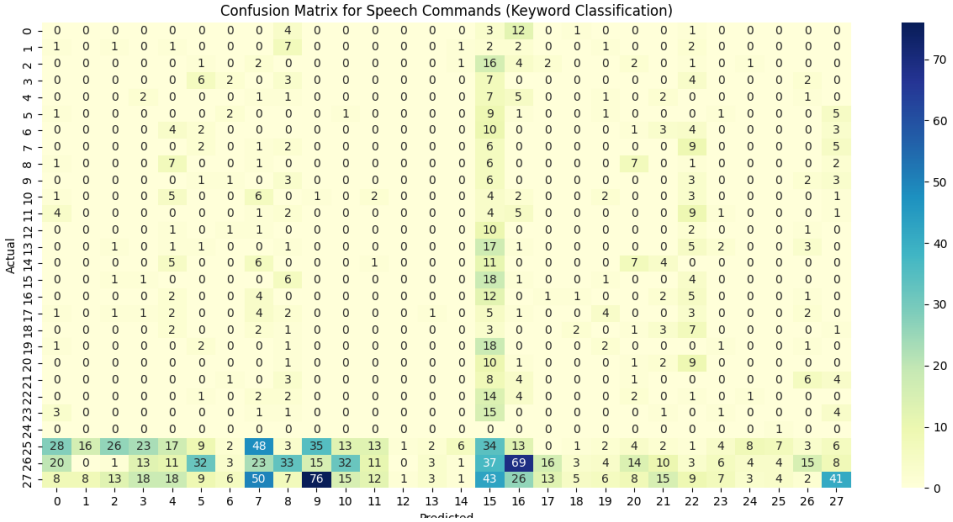

Figure 2: Speech Commands Visualization

7. Conclusion

This research introduces a powerful AI-driven voice recognition system designed to detect distress in real time, helping those in danger get the help they need faster. By leveraging cutting- edge deep learning models like Wav2Vec2 and Whisper, we have created a system that not only detects distress signals with high accuracy but also works effectively in noisy environments.

Key takeaways from our study:

- Advanced AI models outperform traditional approaches, making distress detection more reliable and accurate.
- Spectrogram-based features provide better insights into distress speech patterns compared to older MFCC methods.

- Mobile-friendly deployment makes real-time distress detection accessible, ensuring help is just a voice command away.Multilingual dataset expansion increases global usability, making the system effective across different languages and dialects.
- Looking ahead, we plan to:
- Expand our dataset to cover more languages and speech variations.
- Improve real-time detection using federated learning for more personalized and adaptive performance.
- Develop integration with wearable devices and smart security systems for automated emergency responses.

By making AI-driven distress detection widely accessible, this research contributes to the broader effort to enhance safety and security for vulnerable individuals, ensuring that no cry for help goes unheard.

Declaration on Generative AI

During the preparation of this work, the author(s) used X-GPT-4 and Gramby in order to: Grammar and spelling check. After using these tools/services, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1]. World Health Organization. "Violence against women prevalence estimates, 2018." WHO

[2]. United Nations Women. "The Shadow Pandemic: Violence against women during COVID-19." UN Women, 2020.

[3]. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Info Processing Systems.

[4]. Radford, A., et al. (2022). Whisper: Robust speech recognition via large-scale weak supervision. OpenAI.

[5]. Lane, N. D., Bhattacharya, S., Georgiev, P., Forlivesi, C., Kawsar, F., & Lymberopoulos, D. (2015). DeepEar: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning. ACM Conference on Embedded Networked Sensor Systems (SenSys).

[6]. Rabiner, L., & Juang, B. H. (1993). Fundamentals of speech recognition. Prentice-Hall.

[7]. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 29(6), 82-97.

[8]. M.Sharma, J.BinongAssistant & P.Kumar "Application of Artificial Intelligence for Voice Recognition" Feb 2023

[9]. H.Ilgaza, B.Akkoyuna, Ö.Alpaya, and M.Akcayol"CNN Based Automatic SpeechRecognition: A Comparative Study" Aug 2024

[10]. LeCun, Y., Bengio, Y., & Hinton, G., "Deep learning," Nature, vol. 521, no. 7553, 2015.

[11]. Cevik, K., Ozkan, S., & Kara, K., "Tuning CNNs for Speech Recognition Tasks," IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 5, pp. 1684-1695, 2020.

[12]. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I., "Robust Speech Recognition via Large-Scale Weak Supervision," OpenAI Research Paper, 2022.

[13]. Baevski, A., Zhou, H., Mohamed, A., & Auli, M., "wav2vec 2.0: A Framework for Self- Supervised Learning of Speech Representations," NeurIPS, 2020.

[14]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I., "Attention Is All You Need," NeurIPS, 2017.

[15]. Olston, C., Najork, M., "Web Crawling," Foundations and Trends in Information Retrieval, vol. 4, no. 3, pp. 175-246, 2010.