# SPEECH HELP

UCS749

## CONVO AI LAB EVALUATION

**Submitted by:**

**102117045 Shiya Mer**

**BE Fourth Year, COPC**

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

## 1. Introduction

This project aims to develop a specialized speech recognition system capable of identifying specific keywords ("yes", "no", "one", "two") to activate and confirm emergency calls to the 112 helpline. The system employs a two-stage approach: initial training on a public dataset followed by fine-tuning on personal voice recordings.

The choice of keywords is deliberate: "yes" and "no" provide clear confirmation options, while "one" and "two" could be used for additional functionality or verification steps. This limited vocabulary approach draws inspiration from research demonstrating the effectiveness of focused datasets for specific speech recognition tasks.

## 2. Dataset Preparation and Processing

### 2.1 Public Dataset

- Based on the "Speech Commands" dataset, focusing on selected keywords.
- Consists of audio files for "yes", "no", "one", "two", each approximately 1 second long.
- Provides a solid foundation for training, offering a variety of speakers and recording conditions.

### 2.2 Personal Dataset

- Created using personal voice recordings of the same keywords.
- Recorded using a Laptop to simulate real-world conditions.
- Allows the model to adapt to the specific voice characteristics of the user.

### 2.3 Data Preprocessing

- **Audio Loading**: Efficiently loads audio files and their corresponding labels.
- **Resampling**: Ensures consistent sample rate across all audio files.
- **Spectrogram Conversion**: Transforms audio into a 2D time-frequency representation.

### 2.4 Dataset Splitting

- Training set: 80% of data
- Validation set: 10% of data
- Test set: 10% of data

This split ensures enough data for training while reserving separate sets for validation and testing.

## 3. Model Architecture

The model uses a Convolutional Neural Network (CNN) architecture, designed to process the 2D spectrogram inputs efficiently. CNNs are particularly well-suited for this task due to their ability to capture local patterns and hierarchical features in the spectrograms.

```
Input shape: (124, 129, 1)
Model: "sequential"
```

| Layer (type) | Output Shape | Param # |
|---|---|---|
| resizing (Resizing) | (None, 32, 32, 1) | 0 |
| normalization (Normalization) | (None, 32, 32, 1) | 3 |
| conv2d (Conv2D) | (None, 30, 30, 32) | 320 |
| conv2d_1 (Conv2D) | (None, 28, 28, 64) | 18,496 |
| conv2d_2 (Conv2D) | (None, 26, 26, 128) | 73,856 |
| max_pooling2d (MaxPooling2D) | (None, 13, 13, 128) | 0 |
| dropout (Dropout) | (None, 13, 13, 128) | 0 |
| flatten (Flatten) | (None, 21632) | 0 |
| dense (Dense) | (None, 64) | 1,384,512 |
| dense_1 (Dense) | (None, 128) | 8,320 |
| dropout_1 (Dropout) | (None, 128) | 0 |
| dense_2 (Dense) | (None, 4) | 516 |

```
Total params: 1,486,023 (5.67 MB)
Trainable params: 1,486,020 (5.67 MB)
Non-trainable params: 3 (16.00 B)
```

### 3.1 Model Structure

1. **Input Layer**: Accepts spectrogram representation of audio.
2. **Preprocessing Layers**: Includes resizing and normalization for efficient processing.
3. **Convolutional Layers**: Multiple layers with increasing numbers of filters to capture complex patterns.

4. **Pooling and Regularization**: Reduces spatial dimensions and helps prevent overfitting.
5. **Fully Connected Layers**: Processes the extracted features for final classification.
6. **Output Layer**: Produces probabilities for each keyword.

### 3.2 Model Compilation

- Uses adaptive learning rate optimization for efficient training.
- Employs a loss function suitable for multi-class classification.
- Tracks accuracy as the primary performance met

## 4. Training Process

### 4.1 Initial Training

- Utilizes the public dataset to train the base model.
- Employs early stopping to prevent overfitting.
- Monitors validation performance to ensure generalization.

### 4.2 Fine-tuning

- Loads the pre-trained model and adapts it to personal voice data.

## 5. Model Evaluation

### 5.1 Metrics

- Accuracy: Measures overall correctness of predictions.

## 6. Conclusion

The project demonstrates significant progress in developing a specialized speech recognition system for emergency call activation. The current model shows promise in recognizing key phrases, with room for improvement through fine-tuning and additional optimizations. The next crucial steps involve completing the fine-tuning process, rigorous testing, and integrating the model into a user-friendly emergency call system.

The potential impact of this system is substantial, as it could provide a hands-free way to call for help in emergencies. However, the critical nature of its application necessitates thorough testing and careful consideration of ethical implications before deployment.

The combination of advanced machine learning techniques with practical, life-saving applications showcases the potential of AI to make a meaningful impact on public safety and individual well-being.