TinyCNN: An Embedded CNN Model for Speaker Identification Using ESP32

Hatem Zehir, Toufik Hafs, and Sara Daas

L.E.R.I.C.A Laboratory, Badji Mokhtar University, Annaba, Algeria

November 21-23, 2023

جامعة باجي مختار عنابة

Université BADJI Mokhtar Annaba

Table of Contents

What is biometrics?

Biometrics are unique physical or behavioural characteristics that can be used to identify an individual. The term ”biometrics” comes from the Greek words ”bios,” meaning life, and ”metron,” meaning measure.

What is Speaker recognition?

Speaker recognition is the process of automatically recognizing a user based on specific features of his voice.

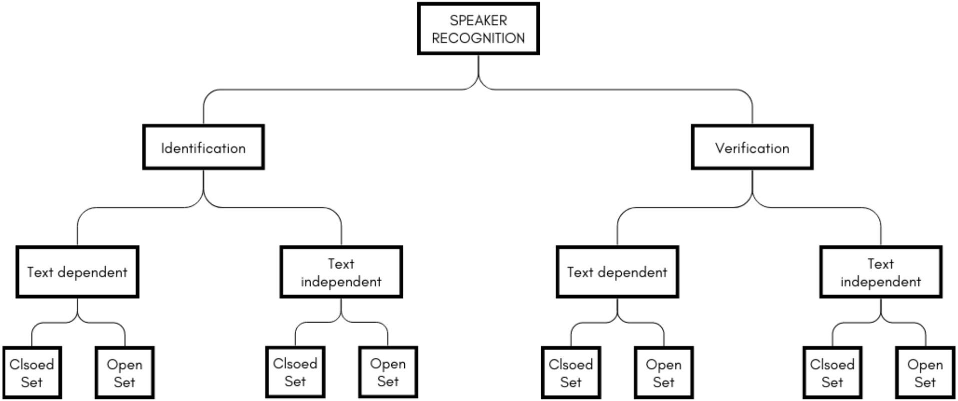

Figure: Schematic Representation of the Different Speaker Recognition Types

Table of Contents

2 Motivation

Motivation

- To overcome the challenges of deploying deep learning models on edge devices, such as the large size, the heavy computational power, and the memory requirements.

- To achieve state-of-the-art results for speaker identification on resource-constrained devices.

Table of Contents

State of the Art

Authors	Databases	Approach	Results
Prachi et al.	TIMIT and LibriSpeech	MFCC and DNN	97.85%
Bunrit et al.	YouTube's audio	Spectrograms + CNN	95.83%
Ye and Yang	Aishell-1	CNN + GRU	98.96%

Table of Contents

LibriSpeech

- Corpus of approximately 1000 hours of read English speech.
- Sampling rate of 16 kHz.
- Derived from read audiobooks from the LibriVox project.

Pre-Processing

- Silence Removal
  - Threshold for silence determined by:

$$threshold_{db} = mean_{db} - std_{db} \tag{1}$$

  - Sounds below threshold considered as silent.
- Window Segmentation
  - Audio segmented into 1-second windows.
  - Each window overlaps with the previous by 500 ms.

Feature Extraction

- MFCC is used to extract the spectrograms of the audio files.
- MFCCs are a powerful tool for speech recognition, speaker identification, and other audio processing tasks.

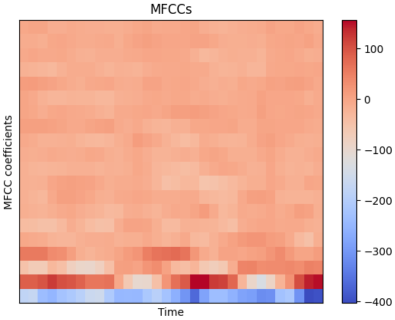

Figure: MFCC coefficients of a randomly selected audio segment.

Deep Learning Architecture

Table: The Architecture of the Neural Network Used

#	Layer	Output Shape	Param. #
1st	Conv2D	12, 32, 8	80
2nd	Dropout	12, 32, 8	0
3rd	Conv2D	12, 32, 8	584
4th	Dropout	12, 32, 8	0
5th	Flatten	3072	0
6th	Dense	10	30730

INT8 Quantization

- Deep learning model inference requires significant memory and computational resources.
- Deploying models on resource-constrained devices is challenging.
- Quantization reduces model size and improves inference time by simplifying calculations.
- For example, converting from 32-bit floating point to 8-bit integers can reduce the model size by a factor of 4.

Deployment on ESP32

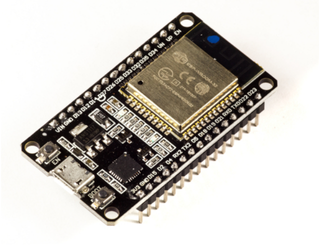

Figure: ESP32 Dev Board

- Quantized models are typically converted to the TFLite format.
- The Edge Impulse Framework is used to generate an Arduino library for deploying the quantized model on the ESP32.
- The ESP32 board is utilized without additional components, and the model is tested with data from the LibriSpeech database to assess its performance.

Table of Contents

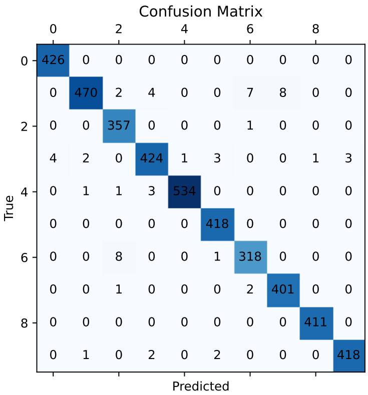

Figure: Confusion Matrix of the Unoptimized Mode

Table: A performance comparison between the quantized and unoptimized models.

	Inference time (ms)	Model Size (KB)	RAM Usage (KB)	Accuracy
Quantized (int8)	30	34.33	24.55	98.58%
Unoptimized (float32)	64	125.16	24.73	98.63%

Table of Contents

6 Conclusion

Conclusion

- A novel optimized CNN architecture was introduced for speaker recognition.
- Evaluation was conducted on a subset of 10 speakers from the LibriSpeech database.
- The process involved audio signal preprocessing, MFCC coefficient extraction, and CNN model training.
- The trained model was quantized, converted to TFLite format, and transferred to an Arduino-compatible library.
- Deployment onto an ESP32 board was successful, resulting in a quantized model with 98.58% accuracy, 53% faster performance, and 73% smaller size compared to the original model.
- Future work will explore incorporating additional modalities for a multimodal biometric system.

Thank you for listening!

Any questions?