EFFICIENT SPEECH RECOGNITION ON IOT DEVICES

PURDUE UNIVERSITY FORT WAYNE

Hira Memon
Advisors: Dr. Chao Chen, Dr. Zesheng Chen, and Dr. Jack Li
Department of Electrical and Computer Engineering

EDGE IMPULSE
TensorFlow Lite

INTRODUCTION

Recent advancements in Tiny Machine Learning (TinyML) enable low-power edge devices in Internet of Things (IoT) applications to efficiently process deep learning models, revolutionizing real-time data analysis in resource-constrained environments. With 250 billion microcontrollers and 15 billion IoT devices globally, TinyML adoption is accelerating, projected to reach 11 billion installations by 2027 [1].

This study explores the application of TinyML in audio processing, focusing on speaker verification—a key technology for enhancing IoT security and privacy. By enabling voice authentication on edge devices, such as smart assistants, TinyML reduces reliance on cloud services and mitigates privacy risks. As an initial step, a speech recognition system is implemented and deployed at a resource-constrained IoT device. This not only showcases the feasibility of TinyML but also leads to a more secure and private IoT applications in the future.

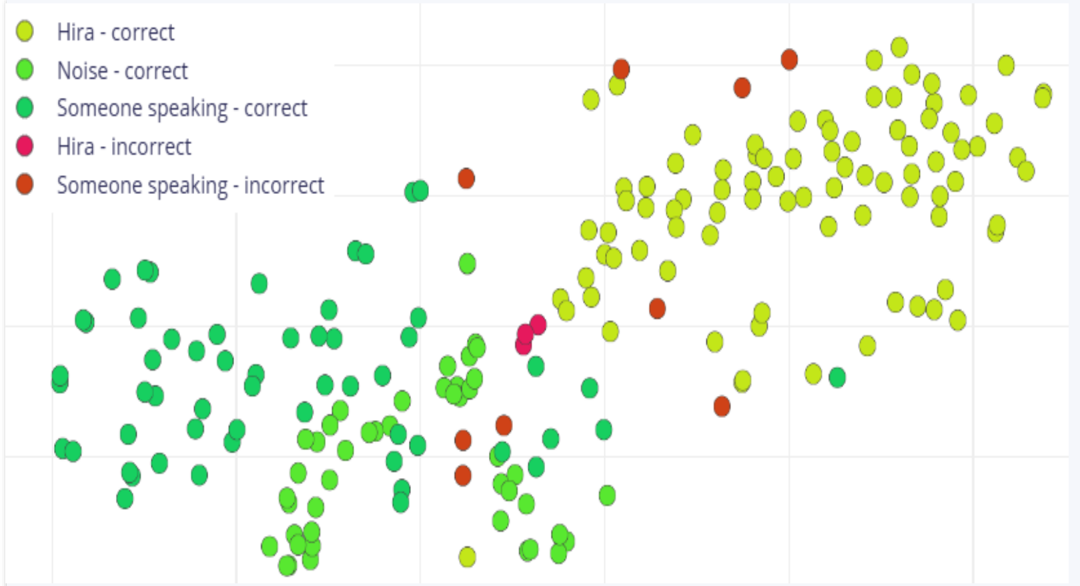

Figure 1. Dataset collected

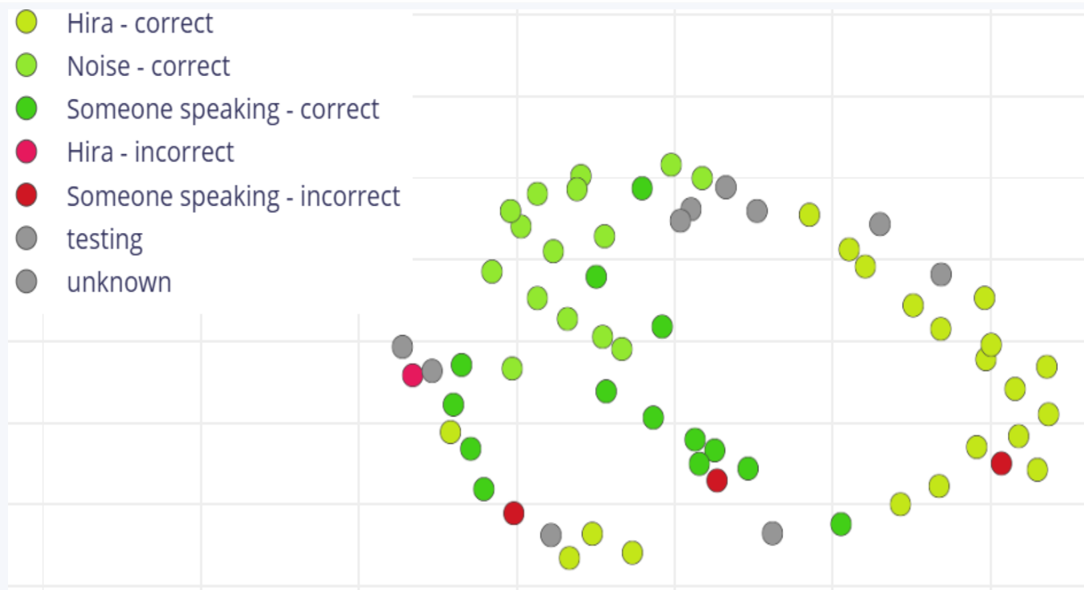

Figure 2. Test data after inference

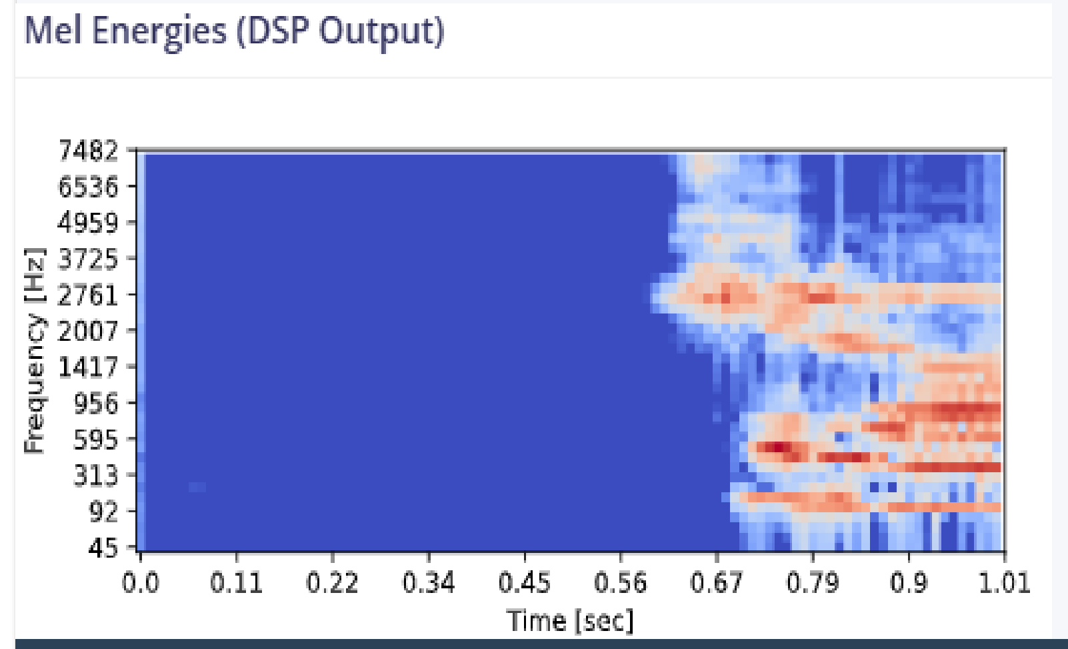

Figure 3. Mel Energies

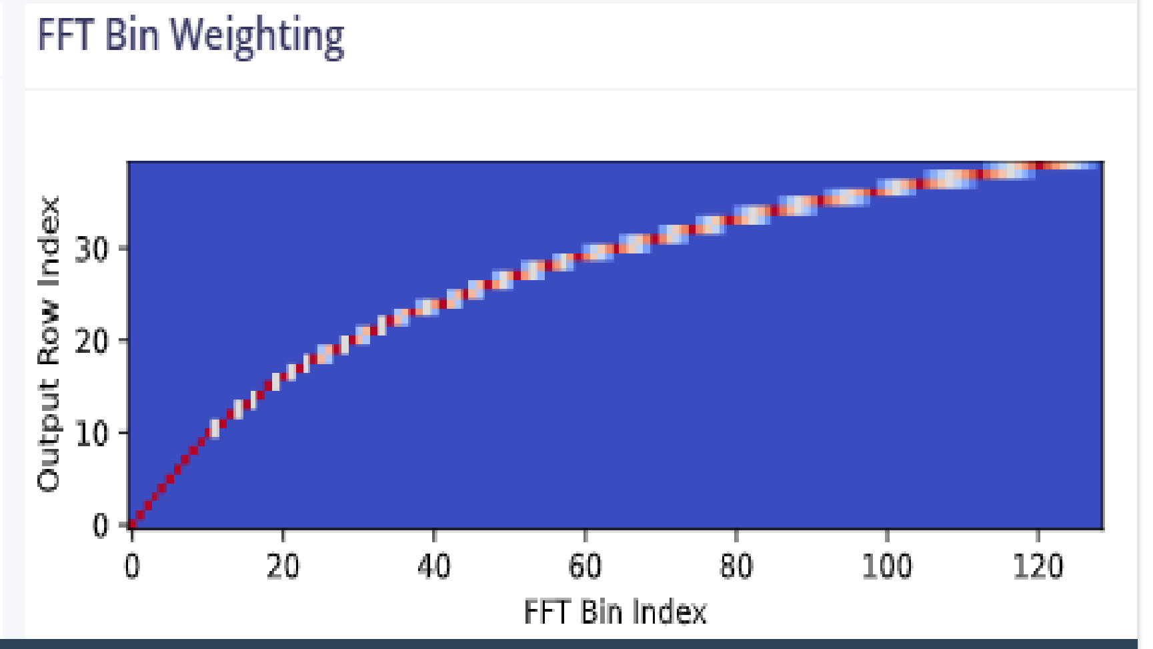

Figure 4. FFT Bin Weighting

Enabling efficient and privacy-preserving speaker verification on edge devices through tiny machine learning (TinyML)

METHODOLOGY

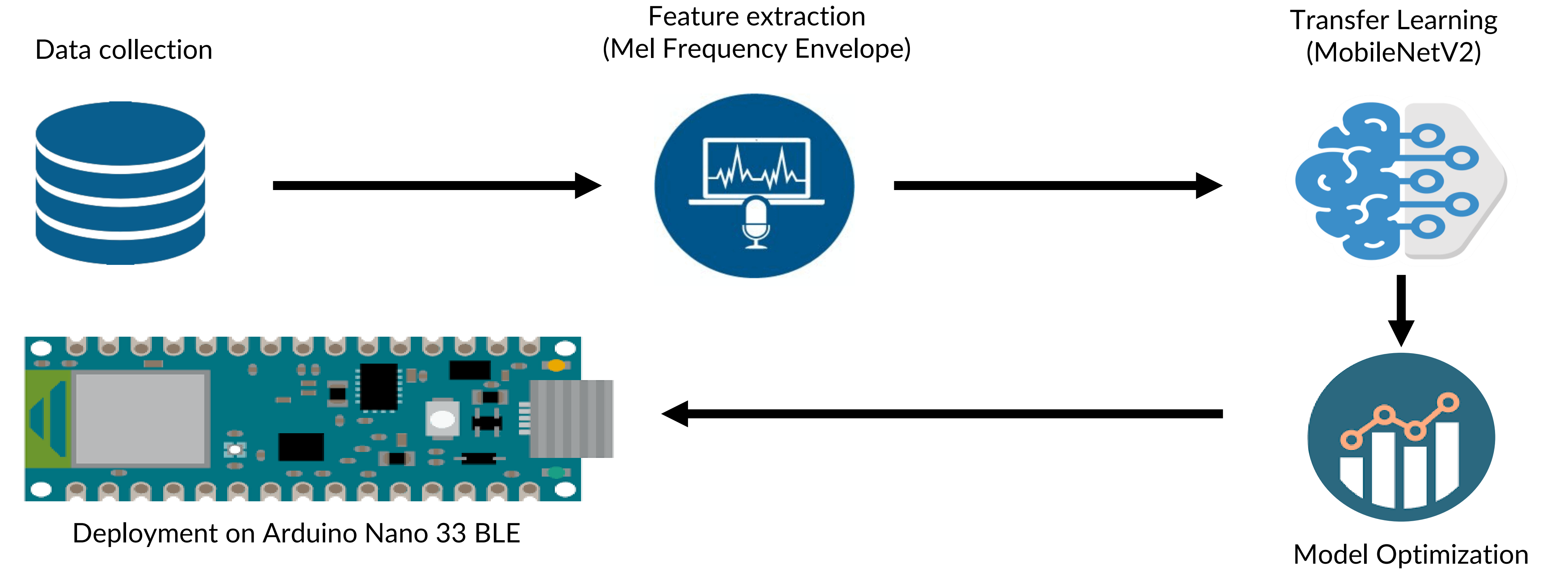

Figure 5. Methodology Workflow

A TinyML-based speech recognition system enables a device to recognize a spoken name and activate. It is part of a cascading architecture where a machine learning model at the edge device detects the name, triggering a more sophisticated model at a gateway or in the cloud to verify the speaker's identity.

The methodology starts with data collection using Edge Impulse [2], a development platform for machine learning on edge devices. Feature extraction leverages the Mel Frequency Envelope, which captures the power spectrum of a sound signal to enhance spectrogram components on the Mel Scale, as shown in Figures 3 and 4. A pretrained MobileNetV2 model is then optimized via quantization to balance size and accuracy, running TensorFlow Lite on the backend [3].

Test results are summarized in Table 1, demonstrating the performance of the optimized model, requiring approximately 168.5 KB RAM. This model was successfully deployed on an Arduino Nano 33 BLE Sense [4], proving its feasibility for low-power edge applications.

RESULTS

Table 1. Confusion Matrix

	Hira	Noise	Someone Speaking	Uncertain
Hira	**95.2%**	0%	0%	4.8%
Noise	0%	**100%**	0%	0%
Someone Speaking	5.9%	0%	**82.4%**	11.7%
F1 Score	0.95	1	0.90	

FUTURE IMPLEMENTATION

The study employs a cascading TinyML architecture, where a lightweight edge model detects a spoken name before triggering a cloud-based model for speaker verification.

Future work aims to enhance security by deploying a more advanced model directly on the edge using the Arduino Nicla Voice board [5] instead of Arduino Nano BLE 33 [4]. This hardware's superior computational power and audio processing capabilities will enable improved security while maintaining real-time performance in resource-constrained environments.

This research has key implications for advancing edge AI. As the TinyML market grows, it will reshape industries by enabling smart functions on small, low-power devices. By improving speaker verification on edge devices, this study aims to enhance security and privacy in IoT applications.

REFERENCES AND ACKNOWLEDGEMENTS

[1] "11 billion TinyML device installs, 481 million 5G advanced devices in 2027, and 35 other transformative technology stats you need to know," PR Newswire, Feb. 28, 2023. Available: https://www.prnewswire.com/news-releases/11-billion-tinyml-device-installs-481-million-5g-advanced-devices-in-2027-and-35-other-transformative-technology-stats-you-need-to-know-301742527.html, [accessed March 5, 2025]

[2] S. Hymel, et al., "Edge Impulse: An MLOps Platform for Tiny Machine Learning," arXiv preprint, arXiv:2212:03332, 2022.

[3] R. David, et al., "TensorFlow Lite Micro: Embedded Machine Learning on TinyML Systems," Proceedings of 4th Machine Learning and Systems (MLSys) Conference, 3, pp. 800-811, 2021.

[4] Nano 33 BLE Sense, https://docs.arduino.cc/hardware/nano-33-ble-sense/ [accessed March 5, 2025]

[5] Nicla Voice, https://docs.arduino.cc/hardware/nicla-voice/ [accessed March 5, 2025)

This research was funded by 2024-2025 ETCS Faculty Research Seed Grant.