Demo Abstract: SPIDER: Lightweight Speaker Identification on Resource-Constrained Embedded Devices

Markus Gallacher*, Tobias Koenig*, Carlo Alberto Boano*, Willian T. Lunardi†, Michael Baddeley†, M. S. Arun Sankar‡, and Utz Roedig§
*Institute of Technical Informatics, Graz University of Technology, Austria
†Technology Innovation Institute, United Arab Emirates
‡South East Technological University Carlow, Ireland
§University College Cork, Ireland

***Abstract*—In this demo, we showcase SPIDER, a solution that allows to perform voice-based speaker identification directly on resource-constrained embedded devices, running live on the open-source OpenEarable 2 earbud, which embeds an nRF5340 SoC with only 512 kB of RAM and 1 MB of flash memory. Voice-based speaker identification models can be categorized into two main groups: (i) closed-set speaker identification (CSSI), in which the model encounters only previously enrolled speakers, and (ii) open-set speaker identification (OSSI), in which the model additionally encounters unknown speakers that must be rejected. By running CSSI and OSSI on a smart earbud, SPIDER enables new use cases: from activating user-specific profiles to authenticating a user without requiring additional bulky pin-pads, or fingerprint scanners, or relaying information to more powerful hardware. Our demo uses the `xResNetSmall`, a shrunk machine learning model based on the xResNet18 that requires only 60 kB of RAM and 548 kB of flash memory. The model is trained with Log Mel spectrograms that are computed from the recorded audio samples.**

***Index Terms*—Machine Learning, Speaker Identification, Resource-Constraints, Embedded Systems.**

I. Introduction

Voice is a convenient and unobtrusive way to operate devices. Together with the ever growing Internet of Things (IoT), more and more resource-constrained devices with built-in microphones enable applications to rely on voice, such as smart speakers (e.g., Amazon Echo, Apple HomePod, and Google Home), smart TVs, and smart earbuds (such as the OpenEarable 2 [1]). Most applications, however, use keyword spotting to activate the device, and then relay speech to a large model operating in the cloud to provide more functionality. Voice-based speaker identification enables an application to not only activate, but already start user-specific actions or profiles while the user speaks freely. For applications that are only accessed by known speakers, closed-set speaker identification (CSSI) models identify a speaker based on the highest matching probability. If unknown speakers can access the application, the CSSI model would still select the highest matching speaker, which is a problem for applications that need to reject unknown speakers for security reasons, or for participation analytics that only evaluate a specific target group. In this case, open-set speaker identification (OSSI) adds a rejection mechanism to only identify speakers that have a high probability and otherwise rejects the speaker as unknown, as depicted in Figure 1. While this could be done in the cloud, relaying speech has implications on the reliability, privacy, availability, and timeliness of the system. Providing on-device solutions solves these implications, but requires the device to deploy more sophisticated models than simple keyword spotting models. The smallest common models used for CSSI tasks, AM-MobileNet1D [2] and xResNet18 [3], require up to 12MB of flash memory, while the IoT often deploys more constrained devices, such as the nRF5340 SoC with only 1 MB of flash memory. In SPIDER [4], we show that we can shrink these models with reverse compound scaling. Using the `xResNetSmall` with Log Mel spectrogram (LMS[1]) pre-processing, we can shrink the model by a factor of 16 and achieve 94.6 % CSSI accuracy and 91.8 % OSSI accuracy on a self-recorded real-world dataset using the nRF5340 SoC built into the nRFThingy:53. Additionally, SPIDER introduces two tuning knobs (probability threshold τ and consensus threshold) to adapt the OSSI model to various applications, see §II.

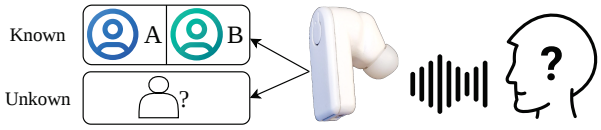

Fig. 1. Open-set speaker identification on an OpenEarable 2 [1] housing an nRF5340 SoC. Speaker A or B of the known pool are only identified if the model is confident enough, otherwise the speaker is rejected as unknown.

In this demo, we showcase SPIDER running *live* on the off-the-shelf OpenEarable 2 [1] to demonstrate its feasibility, the impact of τ and consensus on OSSI performance, and share challenges one faces during live deployment (e.g., impact of noise, and sensitivity and placement of the microphone).

[1]LMS are the log amplitudes of the fast Fourier Transform converted to the Mel scale to represent human hearing more accurately.

II. SPIDER in a Nutshell

In order to identify or reject a speaker by voice, we need to solve a classification problem for identification, and a binary problem for rejection.

CSSI tackles the classification problem, where the predicted speaker is marked by the model output with the highest probability. To handle arbitrary lengths of audio, individual segments of predefined size of the LMS are fed into the model sequentially. To get the final speaker prediction, a majority vote across multiple segments is performed.

OSSI solves the binary problem by introducing a confidence threshold (τ) and a consensus threshold. τ decides if the output probability of a segment is confident enough to be used in the majority vote. For example, if the model predicts speaker A with 80 % probability, but our τ is set to only allow probabilities $> 90\,\%$, this segment does not contribute to the final majority vote. The consensus threshold defines the required number of segments required for a valid majority vote. For example, if we calculate the majority vote over 10 segments and the consensus threshold is set to 50 %, then speaker A can only be identified if speaker A appears in at least 5 of the 10 segment-level predictions.

These OSSI settings (τ and consensus) allow SPIDER to adapt to various needs of an application. For instance, strict OSSI settings allow SPIDER to reduce the false acceptance rate (FAR), where an unknown speaker is falsely accepted as a known speaker. Unfortunately, reducing the FAR also increases the false rejection rate (FRR), where a known speaker is falsely rejected as an unknown speaker, as stricter OSSI settings require the model to also predict known speakers with higher confidence. Lenient OSSI settings allow SPIDER to reduce the FRR, but inadvertently increase the FAR by making it easier for an unknown speaker to be mistaken for a known speaker. This trade-off between FAR and FRR needs to be considered when designing applications, as a high FRR will lead to more inconvenience, where speakers often need to repeat themselves in order to be accepted, while a high FAR will allow unknown people to more easily be accepted.

III. Demo

In this demo, we showcase SPIDER running live on the open-source OpenEarable 2 earbud and show the impact of τ and consensus thresholds on OSSI accuracy. Two of the authors are present during the demo, and are part of the known pool of speakers the model was trained with, while willing participants act as unknown speakers that need to be rejected. The OpenEarable 2 is connected via Bluetooth Low Energy (BLE) to a graphical user interface (GUI) on a laptop, where the GUI can subscribe to a Nordic UART Service (NUS) to receive information, as shown in Figure 2. The OpenEarable 2 has an inward and an outward facing microphone, but we only use the inward facing microphone to block out as much of the surrounding noise as possible. The ear tip is cleaned between users as it needs to be placed in the ear of the tested speaker. With a button press, the OpenEarable 2 starts recording 2.9 s of audio, computes the LMS, and sequentially runs inference

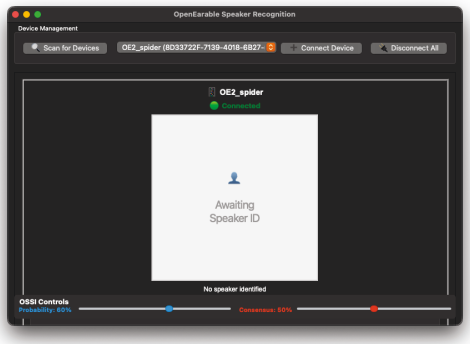

Fig. 2. GUI awaiting model predictions transmitted from the OpenEarable 2 via BLE. τ and consensus thresholds can be set at runtime to define when a speaker is rejected and shows a corresponding image in the GUI.

on individual segments of the LMS. During the states (i) audio recording, (ii) feature calculation, and (ii) inference, the GUI receives the current status of the OpenEarable 2 via BLE and displays it. Instead of using hardcoded τ and consensus thresholds, the OpenEarable 2 sends the model's output to the GUI, where the thresholds can be changed at runtime using two sliders. When a speaker is identified with high enough confidence as one of the known speakers, the GUI shows a picture of the identified speaker. If the speaker is not identified with a high enough confidence, the speaker is rejected and the GUI shows an image of an anonymous speaker. When τ and consensus are set to zero, SPIDER behaves like CSSI and always shows the identified speaker. The recorded audio remains on-device and is retained only temporarily in volatile memory for LMS computation; no raw audio is stored persistently or transmitted beyond the device.

Acknowledgments

This work was conducted within the SPIDR2 project (Secure, Performant, Intelligent, Dependable, Reliable, and Resilient Wireless Systems) financed by the Technology Innovation Institute. This publication has also emanated from research conducted with the financial support of Research Ireland under Grant number 19/FFP/6775. The authors would also like to thank Markus Schuss for technical support.

References

[1] T. Röddiger, M. Küttner, P. Lepold, T. King, D. Moschina, O. Bagge, J. A. Paradiso, C. Clarke, and M. Beigl, "OpenEarable 2.0: Open-Source Earphone Platform for Physiological Ear Sensing," *in Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 9, no. 1, 2025.

[2] J. A. C. Nunes, D. Macêdo, and C. Zanchettin, "AM-MobileNet1D: A Portable Model for Speaker Recognition," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2020.

[3] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of Tricks for Image Classification with Convolutional Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] M. Gallacher, C. A. Boano, W. T. Lunardi, M. Baddeley, M. S. A. Sankar, and U. Roedig, "SPIDER: Lightweight Speaker Identification on Resource-Constrained Embedded Devices," in *Proceedings of the ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems (SenSys)*, 2026.