

Demonstration Abstract: Automatic Speech Recognition for Resource-Constrained Embedded Systems

Felix Sutton, Reto Da Forno, Roman Lim, Marco Zimmerling, Lothar Thiele
Computer Engineering and Networks Laboratory, ETH Zurich, Switzerland
{fsutton,rdaforo,lim,zimmerling,thiele}@tik.ee.ethz.ch

Abstract—We demonstrate the design and implementation of a prototype hardware/software architecture for automatic single-word speech recognition on resource-constrained embedded devices. Designed as a voice-activated extension of an existing wireless nurse call system, our prototype device continually listens for a pre-recorded keyword, and uses speech recognition techniques to trigger an alert upon detecting a match. Preliminary experiments show that our prototype achieves a high average detection rate of 96%, while only dissipating 28.5 mW for continuous audio sampling and duty-cycled speech recognition.

Keywords—Speech recognition, single-word detection, embedded system, voice-activated nurse call, ARM Cortex-M4.

I. INTRODUCTION

Low-power wireless networks for healthcare are largely being used for monitoring patient activity and physiological parameters [1]. Instead, we recently developed a wireless nurse call system, providing bidirectional interactions between patient and caregiver, based on push-buttons and off-the-shelf low-power wireless devices [2]. We successfully deployed the system in a summer camp for boys with Duchenne muscular dystrophy (DMD).¹ Patient devices were installed at the beds so the boys could alert a caregiver for help during the night. Using the Low-Power Wireless Bus (LWB) [3], our system provides *highly reliable and timely bidirectional interactions* between wireless patient devices and a centralized graphical user interface through which caregivers acknowledge alerts.

Since boys with an advanced stage of DMD experience difficulty pressing a button, we have developed a voice-activated extension to our wireless nurse call system. To make it convenient for a boy to request help, the new patient device must automatically detect a single word with high accuracy. Furthermore, to simplify deployment and ensure a system lifetime of several weeks, the device must be battery-operated, exhibit a compact form factor, and have a low total system power dissipation. Contrary to most existing systems [4], we seek a high single-word detection accuracy, while striving to keep the energy footprint of each patient device to a minimum.

In this demonstration proposal, we outline a hardware/software architecture for single-word speech recognition on an embedded device exhibiting severe processing and memory constraints. Despite these resource constraints, our prototype achieves a high recognition rate of 96%, while only

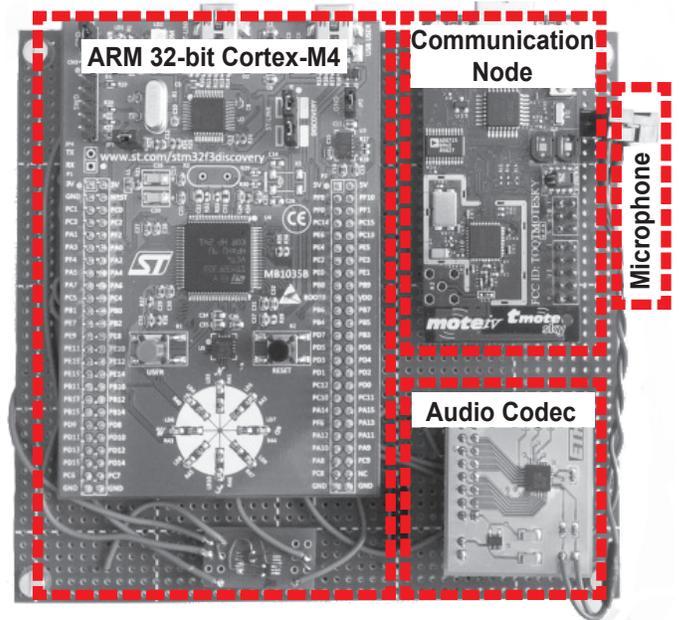


Fig. 1. Prototype speech recognition device.

dissipating an average of 28.5 mW for continuous hardware-based audio sampling and duty-cycled software-based speech recognition. Finally, we describe in detail how we plan to demonstrate our prototype at IPSN.

II. ARCHITECTURE OVERVIEW

Hardware. Figure 1 illustrates our current prototype, which features an ARM 32-bit Cortex-M4 microcontroller. Audio signal acquisition is performed using a dedicated audio codec connected to a microphone. The analog signal from the microphone is sampled at a rate of 8 kHz with 16-bit resolution. When the pre-recorded keyword is detected by the speech processing chain, an attached communication node sends an alert via LWB to the graphical interface just as if a button had been pressed.

Software. In order to achieve our application requirements, we implemented a state-of-the-art speech recognition algorithm [5], as shown in Figure 2, coupled with several power management techniques.

¹DMD occurs almost exclusively in males.

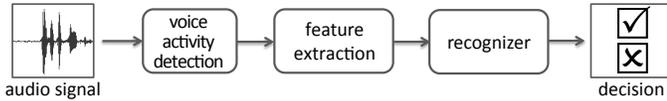


Fig. 2. Speech recognition processing chain.

Voice activity detection (VAD) uses an adaptive signal energy threshold to separate speech from background noise. When voice activity is detected, features of the digital audio signal are extracted. We use mel-frequency cepstral coefficients (MFCCs) as the set of features. The recognizer compares these features to the stored keyword features and decides if the keyword had been spoken or not. Using dynamic time warping (DTW), we account for the keyword being spoken at different speeds.

In order to reduce the power dissipation, we use direct memory access (DMA) transfers for storing the acquired audio signals into memory. As a result, the speech processing chain only needs to execute once every 128 ms, thus allowing aggressive duty cycling of the microcontroller. Additionally, we use dynamic frequency scaling to further reduce the system's power dissipation.

Performance. The two key performance metrics in our application scenario are power dissipation and keyword detection rate. Extensive measurements show that our prototype, without wireless communications, achieves continuous signal processing with a microcontroller duty cycle of 4.6% and an average system power dissipation of 28.5 mW.

To evaluate the keyword detection rate, we perform a 5-fold cross validation using audio samples from five different speakers and four different languages. In order to mimic the target environment (*i.e.*, a night-time dormitory room), we performed the experiments in a silent room. Each speaker chose a single keyword, which was recorded 20 times. A further 30 non-keywords were recorded once by each speaker. The audio samples were transferred from the prototype hardware into Matlab, where cross validation was performed. We find that our prototype achieves an average detection rate of 96%, while limiting the false accept rate to 1%. In summary, our system achieves a higher detection accuracy and a lower energy footprint than existing embedded speech recognition systems [4].

III. DEMONSTRATION DETAILS

Demo setup. We will present our hardware/software prototype of a voice-activated wireless nurse call system and demonstrate its ability to detect a keyword, while giving insights into the speech processing pipeline. The demonstration exhibits two main components:

- *Prototype Device:* We will showcase our prototype speech recognition device. Users will be able to configure it with a keyword of their choice and test the speech recognition by speaking the chosen keyword.
- *Graphical Visualization:* We will extract debug information from the prototype in real-time and provide a graphical visualization of the captured audio samples, the voice activity detection, the extracted feature set, and the decision made by the speech recognition algorithm. Users will be able to observe the displayed information in real-time as well as pause the visualization for detailed inspection.

User experience. Imagine a situation in which a boy with DMD is attending a summer camp. Due to progressive muscle weakness, he is not able to press a push-button and decides to use the voice-activated device.

A caregiver assists the boy in configuring the device with a personalized keyword that he is comfortable speaking when in need of assistance. The choice of keyword is entirely up to him; any word in any language or dialect will do. The only requirement is that he can repeatedly utter the word in less than a second and with similar emphasis. The configuration procedure involves speaking the keyword three times; flashing LEDs indicate when the device is ready to start each recording, and also provide feedback to the caregiver if the recording is admissible or must be repeated. Once configuration is complete, the boy tests the detection performance by speaking the keyword. In case the boy wants to change the keyword, he can do so by re-configuring the device.

IPSN attendees will be able to test our voice-activated prototype device following the procedure outlined above, as well as gain insights into the operation of the speech processing pipeline through the graphical visualization.

IV. ACKNOWLEDGMENTS

We thank Tofigh Naghibi for his advice on speech recognition. This work was supported by nano-tera.ch.

REFERENCES

- [1] J. Ko *et al.*, "Wireless sensor networks for healthcare," *Proc. IEEE*, vol. 98, no. 11, 2010.
- [2] M. Zimmerling *et al.*, "A reliable wireless nurse call system: Overview and pilot results from a summer camp for teenagers with duchenne muscular dystrophy," in *SenSys*, 2013.
- [3] F. Ferrari, M. Zimmerling, L. Mottola, and L. Thiele, "Low-power wireless bus," in *SenSys*, 2012.
- [4] O. Cheng, W. Abdulla, and Z. Salcic, "Hardware-software codesign of automatic speech recognition system for embedded real-time applications," *IEEE Trans. Ind. Electron.*, vol. 58, no. 3, 2011.
- [5] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice Hall, 2008.