A Low-cost keyword spotting architecture based on wavelet packets feature extraction for edge devices

***Abstract*— This paper proposes a novel voice keyword spotting (KWS) architecture that uses wavelet packets to reduce the implementation cost of its feature extraction component. The approach achieves a 54% reduction in latency and a 32% decrease in memory compared to conventional Fourier-based KWS architectures.**

I. Introduction

Voice keyword spotting (KWS) is a task that involves identifying a specific set of keywords within an audio input and holds significant importance in the era of smart devices, where voice interaction has become a prevalent mode of communication with these devices. The task is called binary KWS if the target set includes only two keywords. Typically, it operates continuously, activating speech recognition upon detecting a designated wake-up word like 'Alexa' or 'Hey Siri'. This paper, for the first time, uses wavelet packet decomposition (WPD) [1] to design and implement a KWS architecture with reduced hardware and computational complexity - an essential feature for real-time applications on edge devices and resource-constrained environments. Note that reducing computational complexity leads to a decrease in latency.

II. Proposed KWS Architecture

As shown in Fig. 1(a), the proposed KWS architecture consists of a data acquisition block that uses a microphone and analog to digital converter for collecting single channel 8-bit digital audio samples, a feature extraction block that transforms the samples into 2-dimensional (2D) feature maps, and a neural network classification block that classifies the feature maps. Instead of conventional approach using short-time Fourier transform (STFT) to extract time-frequency features in KWS [2], [3], we propose the use of WPD, which applies high-pass (G) and low-pass (H) frequency filters, followed by down sampling by 2, to generate time-frequency features. Replacing STFT with WPD provides two advantages. First, WPD eliminates the need for windowing operations, while still generates 2D feature maps with adjustable time-frequency precision. For example, Fig. 1(b) shows a 2-level WPD can extract 4x4 feature maps from 16-sample inputs. Second, in contrast to the Fourier transform, wavelet transforms encompass various families, some characterized by minimal computational complexity and exclusively integer (floating-point free) operations.

To minimize the computational complexity of our KWS, we employed the Haar wavelet [1], calculating the average and difference of consecutive pairs of input values as low and high-frequency elements. Additionally, we designed an 8-bit quantized neural network with three convolutional and one fully connected layers for the classification.

III. Evaluation

To evaluate the performance and hardware implementation of our KWS architecture, we designed, trained, 8-bit quantized, and deployed it on the TM4C ARM microcontroller for both STFT and WPD feature extractors.

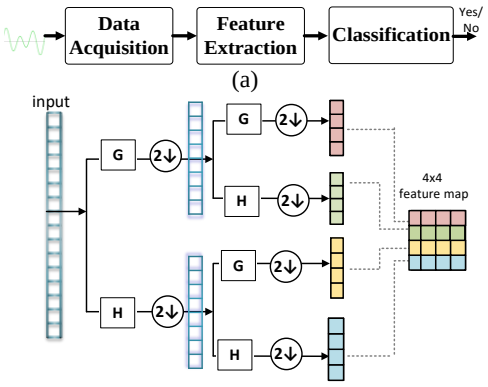

Fig. 1 (a) Components of the proposed KWS architecture, (b) Using 2 levels of WPD to generate a 4x4 time-frequency feature map from a 16-sample input.

TABLE I: MEMORY FOOTPRINT (KB) AND LATENCY (ms) FOR STFT AND WPD-BASED KWS ARCHITECTURES

STFT-based			WPD-based		
RAM	FLASH	Latency	RAM	FLASH	Latency
9.71	13.60	46.80	4.13	11.73	21.73

TABLE II: ACCURACY COMPARISON OF STFT AND WPD-BASED KWS FOR FOUR COMMON KEYWORD PAIRS

Keywords	STFT-based	WPD-based
Yes/No	93.89	93.50
Up/Down	93.52	93.14
Left/Right	86.69	84.41
Stop/Go	86.59	88.40

We used audio samples from Google Speech Dataset v1 and converted them into 8-bit single channel audio samples. We utilized 8-bit integer operations for all calculations, except for the accumulation, which was performed in 32-bit integer. As Table I reports, the memory footprint for the entire binary KWS architecture has been reduced by 57% for RAM part, 14% for Flash part, and 32% in total. Additionally, the latency has decreased from 46.80 ms to 21.73 ms, i.e., 54% reduction. The change in accuracy, however, is not significant, as indicated in Table II for four different pairs of common keywords.

IV. Conclusion

KWS plays a vital role in many voice communication systems connecting humans and electronic devices. This paper introduces a new KWS architecture employing WPD to reduce computational complexity and memory footprint. Experimental results demonstrate a reduction in latency and memory size compared to an STFT-based KWS architecture without compromising accuracy.

References

[1] Daubechies, I. (1992), Ten lectures on wavelets, SIAM.

[2] T. Sainath and C. Parada, “Convolutional Neural Networks for Small-Footprint Keyword Spotting,” in Interspeech, 2015.

[3] Y. Zhang, N. Suda, L. Lai, and V. Chandra, “Hello Edge: Keyword Spotting on Microcontrollers.” arXiv, Feb. 14, 2018. doi: 10.48550/arXiv.1711.07128.