18µW SoC for near-microphone Keyword Spotting and Speaker Verification

J.S.P Giraldo, Steven Lauwereins, Komail Badami, Hugo Van Hamme, Marian Verhelst
Department of Electrical Engineering ESAT-MICAS, University of Leuven, Belgium

Abstract

The first fully-integrated near-microphone Keyword Spotting (KWS) and Speaker Verification (SV) solution, directly interfacing with a passive or active analog microphone and not requiring any external memory. The 65nm SoC realizes speaker-specific keyword triggering while only consuming 10.6µW average or 18.3µW peak for real-time operation, or 10x below speaker-agnostic keyword spotting SotA. Low cost, power and good accuracy (99.5% SV TIMIT and 98.5% KWS TIDIGITS) are jointly achieved through a.) a fully integrated single-chip standalone solution; b.) optimized accelerators; and c.) HW-aware algorithmic tuning and task scheduling.

Introduction

Mobile and wearable devices are increasingly equipped with speech-triggered interfaces, which need to be always-on, always listening. Yet, the energy constraints of these systems conflict with the push towards more complex machine learning methods to improve the recognition accuracy. State-of-the-art (SotA) ASICs for KWS [1-2] lack efficient <100µW implementations, are unable to interface directly with microphones and miss speaker selectivity (Fig. 1). This work is the first to overcome these bottlenecks, with a single-chip, technology-scaling friendly, mixed-signal solution, optimized for power consumption.

Key innovations for optimal functionality at low-power

a.) Scaling-friendly single-chip solution: Fig. 2 shows the architecture of the SoC, which operates on analog microphone data, and does not need any other external IO or memory to generate the keyword and speaker classification labels. The SoC integrates a digital-technology friendly analog front-end (AFE) feeding the digital back-end. The amplifier in the AFE achieves a high linearity by encoding the input voltage to the time-domain through pulse-density modulation and then integrating back into the voltage-domain using a charge-pump. The amplifier output is digitized with a 10-bit oversampling SAR ADC. The AFE allows to obtain 54dB SNR, 0.18% THD at 75% rail-to-rail output at 0.6V, consuming only 1.8µW [3]. The AFE feeds the digitized data to the sound detector (SD), which can activate the FEx to generate a set of Mel Frequency Cepstral Coefficients (MFCCs) and their derivatives (Δ, ΔΔ). These are used as features for the two ML engines: 1.) a Long Short Term Memory (LSTM) accelerator for KWS and 2.) a Gaussian Mixture Model (GMM) accelerator for SV.

b.) Power- and memory-optimized HW-acceleration: Each block of the system is thoroughly optimized algorithmically and architecturally to strike the best balance between flexibility, task accuracy, memory needs and power cost.

SD+FEx: Both modules (Fig. 3) operate on frames of 32ms with 16ms step size. The SD is always-on at very low power consumption. It computes the average input amplitude across each frame and triggers when this crosses a predefined threshold, tuned to avoid miss detects at the expense of false alarms and hence not impacting KWS and SV accuracies (Fig. 3, top right). Upon wake-up, the FEx computes 20 MFCC's, Δ's and ΔΔ's, at programmable center frequencies. Required DFT and DCT steps are executed on a common, time-multiplexed, butterfly accelerator through an engineered input memory mapping reducing leakage by 12% and area by 14%. In nominal operation, a 1024pt real-DFT and 32pt DCT are executed per frame. After the DFT, mel-shaped bandpass filtering is performed. To enable storage of all twiddle factors and mel weights on chip, the sparse mel weights are compressed with position encoding (Fig. 3), enabling to reduce parameter storage memory by 170X. DFT and DCT depth, # mel filters and features are all configurable.

KWS: KWS computes an output keyword classification label for every incoming frame (Fig. 2). This is done through LSTM instead of CNN as it enables lower latency and lower MAC count for similar accuracy [5]. The implemented LSTM accelerator, based on [4], supports 1 to 2 LSTM layers with up to 64 neurons, and 1 to 2 FC layers, offering an accuracy vs cost trade-off (Fig 8). The model weights are fully stored on-chip in 32kB SRAM and can be stored both in 8b linear format and in 4b nonlinear encoding bringing 50% memory compression, in which case they are online decoded by a LUT.

GMM: SV is a computationally expensive task as it requires >0.5 sec of speech to obtain good accuracy [6]. This is achieved efficiently through a >99% accurate GMM with tunable number of Gaussians. The GMM accelerator, Fig. 4, evaluates the Log Likelihood Ratio (Λ) between a background GMM and a speaker-specific GMM. Power consumption is minimized through 1.) transforming the probability density function of a Gaussian to use base 2 instead of *e*, simplifying the exponentiation and accumulation of GMMs to floating point (FP) additions using its exponent input and simplifying the log calculation for Λ by only extracting the exponent of the FP adder (Fig. 4) removing the need for a CORDIC; 2.) early termination of the computation of Gaussians when encountering a dimension in which the distance of the feature to the mean of the Gaussian normalized by its variance is larger than a programmable threshold (σth). This reduces the computations by 50% without accuracy loss (Fig. 4, top right); 3.) evaluating 8 feature vectors in parallel, resulting in model parameter reuse and 8X model memory bandwidth reduction; 4.) storing the model weights fully on-chip in a 65kB SRAM for a maximum of 512 60-Dim Gaussians (2*nDim+1 8b parameters per Gaussian). A controller allows configurability regarding the number of GMMs, Gaussians/GMM and dimensions/Gaussian.

c.) HW-aware algorithmic tuning and scheduling: The configurability of the FEx, KWS, and SV allows to tune the system for minimal power while maintaining good accuracy. Fig 8 shows this trade-off across several parameters. It is important to note that used MFCCs are engineered for joint KWS and SV (Fig. 3, right). Next, for selected configuration parameters the optimal wake-up sequence and thresholds can be explored, rendering an additional 45% full system power savings with SD triggering KWS, triggering SV (Fig. 5).

Measurement Results

The integrated mixed-signal SoC was implemented in 65nm CMOS (die shot Fig. 5). Fig. 6 shows the frequency-power curves of the SoC during KWS and SV mode, allowing real-time operation (250kHz, 0.6V logic, 0.8V memory) with 5.5µW in SD mode, 16.1µW in KWS mode (1 hidden LSTM layer/64 neurons), 14.6µW in SV mode (256-Gaussians), and

18.3 µW with all blocks active. A full system SD-FEx-KWS-SV power trace is shown in Fig. 7, illustrating the task-dependent power consumption, averaging to 10.6µW when 50% SD, 40% KWS and 10% SV active.

As shown in Fig. 8, KWS achieves 98.5% accuracy on the TIDIGITS dataset (10 keywords), and 90.87% on the Google Speech Command Dataset (GSCD, 10 keywords plus unknown and silence), matching other small footprint deep learning solutions, like [1][2] (TIDIGITS) and [5] (GSCD), with at least an order of magnitude lower power (Table 2). For SV, 99.5% EER (equal error rate, balancing miss detects and false alarm) was achieved for 168 speakers from the TIMIT dataset with 256 60-D Gaussians per GMM, again showing SotA [6] accuracy without known prior ASIC implementations in this regard.

References

[1] M. Price, et al. IEEE ISSCC pp 244-245, 2017.
[2] S. Yin, et al. IEEE VLSI pp 139-140, 2018.
[3] K. Badami, et al. IEEE VLSI pp 241-242, 2018.
[4] J. Giraldo, et al. IEEE ESSCIRC pp 166-169, 2018.
[5] J. Fernandez-Marques, et al. Proc. EMDL'18, pp 13-18, 2018.
[6] F. Meriem, et al. IEEE SITIS pp 99-103, 2014.

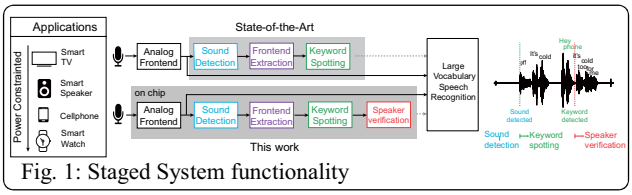


Fig. 1: Staged System functionality

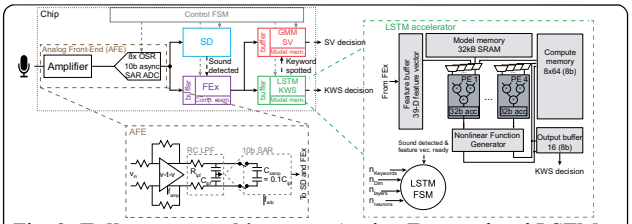


Fig. 2: **Full-system architecture**. Analog Frontend and LSTM accelerator specifications

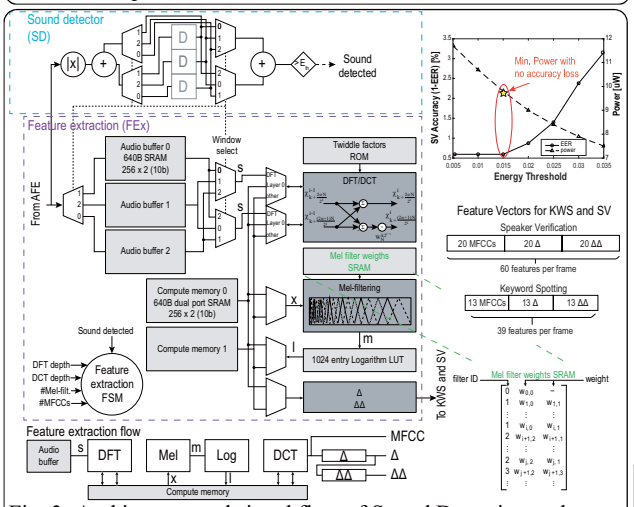


Fig. 3: Architecture and signal flow of Sound Detection and Feature Extraction

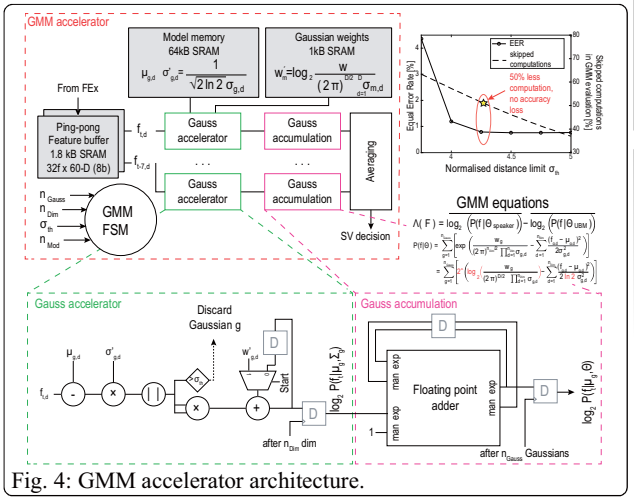


Fig. 4: GMM accelerator architecture.

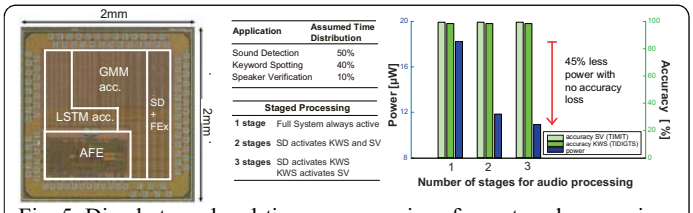


Fig. 5: Die photo and real-time power savings from staged processing.

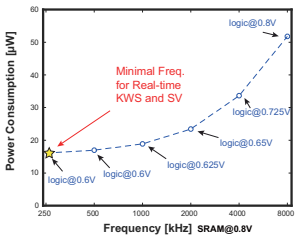


Fig. 6: Power consumption for KWS and SV depending on frequency.

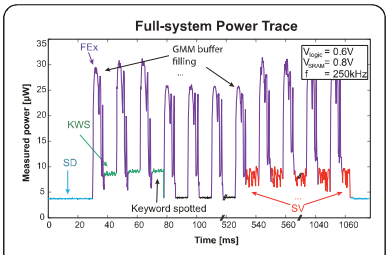
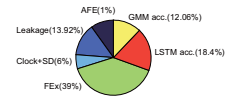


Fig. 7: **Power trace** of full-system execution (SD, KWS and SV)

Characteristics	Value
System	
Technology	65nm
Supply Voltage	0.6-1.2V
Frequency	250 kHz - 12.5 MHz
Core Size	1.6mm x 1.6mm
Die Size	2mm x 2mm
Total SRAM	105 kB
Analog Frontend @0.6V	
ADC frequency	4kHz-16kHz
ADC resolution	6b-10b
THD	0.18%
SNR	54 dB
Feature extraction	
Resolution	10b
DFT-size	1-1024
Mel-filters	1-256
DCT-size	1-256
Keyword Spotting	
Model quantization	4b, 8b
LSTM hidden layers	1-2
LSTM units per layer	1-64
Latency	16ms
Real-time power	16.11µW
Speaker Verification	
Model quantization	8b
GMM Model Parameters (# of GMMs X # of Gaussians/GMM X # of features/Gaussian)	Up to 65k
Latency	0.5s
Real-time power	14.95µW
Full-system power breakdown (@0.6V, 250 kHz)	



Total Power: 18.32 µW

Table 1: **System parameters** for full-system execution

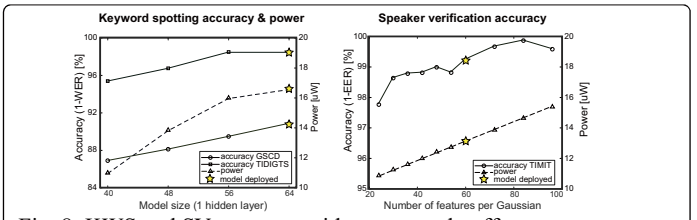


Fig. 8: KWS and SV accuracy with power trade-offs

	ISSCC[1]	VLSI [2]	**This work**
Tech.	65nm	28nm	65nm
Area	13.17mm^2	1.29mm^2	2.56mm^2
Voltage	0.6V-1.2V	0.57V-0.9V	0.6V-1.2V
AFE	NA	NA	✓
Sound Detection	✓	✓	✓
Feature Extraction	✓	✓	✓
Keyword Spotting	✓	✓	✓
Speaker Verification	NA	NA	✓
Latency	Real-time	0.5ms-25ms	16ms (KWS) 0.5s (SV)
Accelerators	VAD MFCC DNN+HMM	VAD MFCC BNN	VAD MFCC LSTM GMM/RBF-SVM
Stages	2 stages	2 stages	3 stages
KWS accuracy	98.35% TIDIGITS 96.88% WSJ	96.11% TIDIGITS	98.5% TIDIGITS 90.87% GSCD
SV accuracy	NA	NA	99.5% TIMIT
Power	**172µw @3MHz**	**141µW @2.5MHz**	**18.3µW @250kHz**

Table 2: State-of-the-Art comparison