Vocell: A 65-nm Speech-Triggered Wake-Up SoC for 10- μ W Keyword Spotting

and Speaker Verification (about MFCC)

Abstract

A complete mixed-signal system-on-chip, capable of directly interfacing to an analog microphone and performing keyword spotting (KWS) and speaker verification (SV), without any need for further external accesses.

Feature Extraction

The FEx block receives samples from the AFE and generates a set of vectors that compress the redundant information present in the speech signal and, which are used by the ML classifiers.

Mel frequency cepstral coefficients (MFCCs) are used extensively as features for speech processing applications due to their high reliability as encoders of the human voice spectrum. Generally, every 16 ms, a 32-ms window of consecutive audio samples is processed to compute a feature vector of MFCCs (between 13 and 20 coefficients) [11].

Its processing pipeline is composed of:

1. an fast Fourier transform (FFT) for spectral analysis;
2. mel filtering using triangular filters;
3. Logarithmic compression; and
4. a discrete Fourier transform (DFT) over the intermediate values [11].

Due to its complex dataflow, its implementation using general-purpose compute architectures presents a large overhead in terms of power consumption [12]. This creates a need for hardware acceleration, yet without giving up all flexibility in the feature parameters.

HARDWARE-AWARE ALGORITHMIC OPTIMIZATIONS

The DCT transformation is based on similar compute kernels as the DFT transformation. This allows **reusing optimized hardware resources for the DCT/DFT implementation**, saving area, and consequently leakage energy.

Real-Point DFT: To reduce the number of computations and memory accesses, the real-point DFT is computed as a complex DFT of half the size (N/2) with two changes:

1. the input samples are restructured

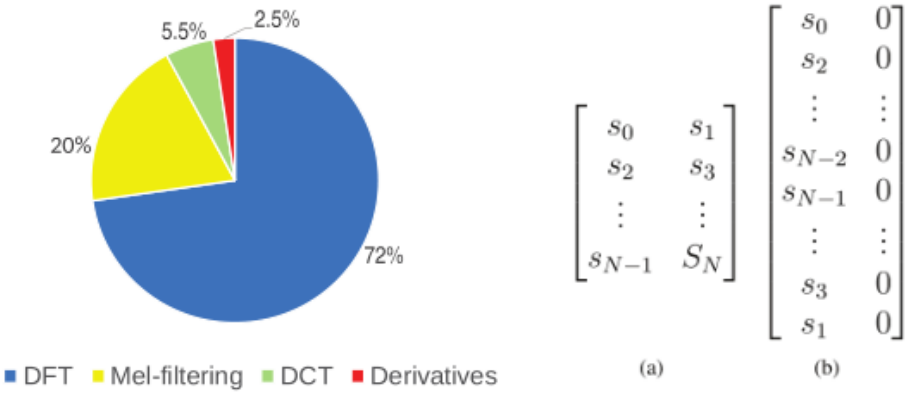

Fig. 4. (Left) MFCC Computation Breakdown in terms of total number of sums and multiplications. (Right) Input sample ordering for (a) real-point DFT and (b) DCT.

2. A final correction step, (4), is performed.

$$\chi_k^{\log_2 N} = \frac{1}{2}\left[\left(\chi_k^{\log_2 \frac{N}{2}} + \chi_{\frac{N}{2}-k}^{*\log_2 \frac{N}{2}}\right) - \left(\chi_k^{\log_2 \frac{N}{2}} - \chi_{\frac{N}{2}-k}^{*\log_2 \frac{N}{2}}\right)\left(jW_N^{kN/2}\right)\right] \quad (4)$$

DCT: compute a DCT via a DFT transformation, through three main steps:

1. reshuffling of the input; (see fig.4(b))
2. computation of a complex DFT;
3. a final correction step.

In the correction step (step 3) the final DCT is extracted from the complex DFT by applying the following transformation to the intermediate result:

$$X_k = \chi_k^{\log_2 N} \exp^{-j\frac{2\pi k}{N}} \frac{2}{\sqrt{2N}} \qquad \forall k \neq 0$$

$$X_k = \chi_k^{\log_2 N} \exp^{-j\frac{2\pi k}{N}} \frac{\sqrt{2}}{\sqrt{2N}} \qquad \forall k = 0. \quad (5)$$

IC ARCHITECTURE

Feature Extraction

The FEx module computes **20 MFCCs** with their respective **delta and delta–deltas** at programmable center frequencies.

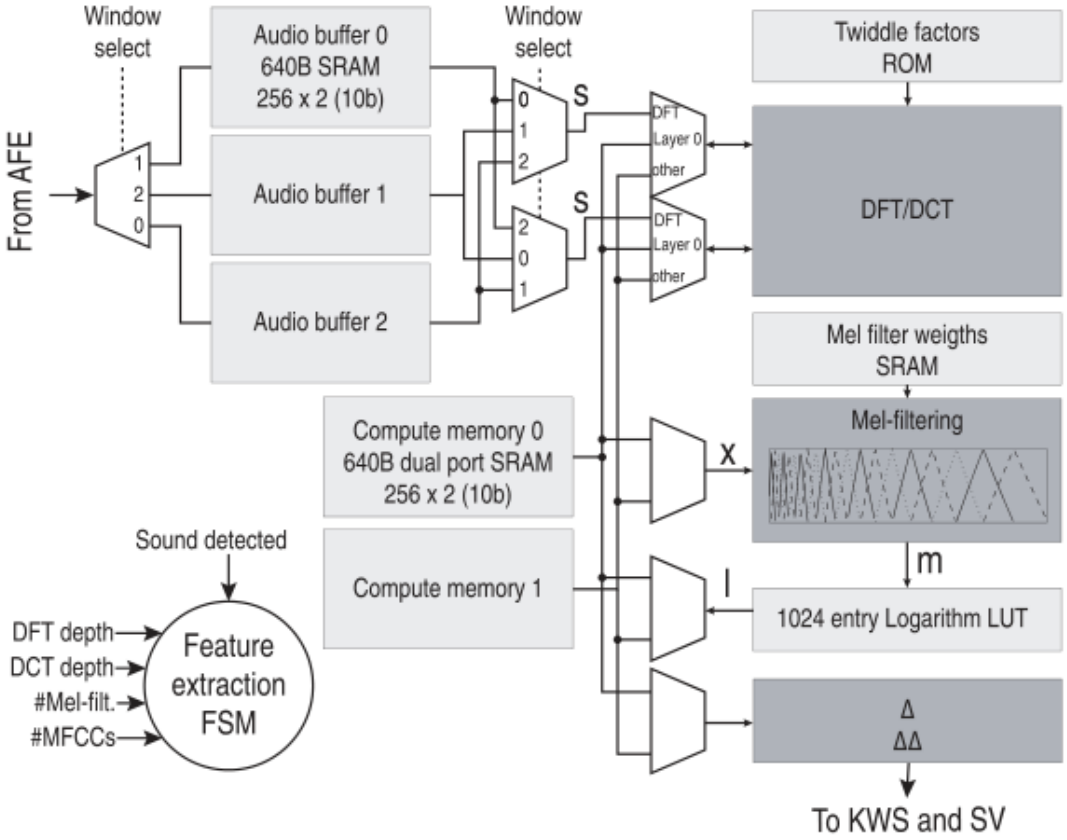

Fig. 11. FEx architecture.

Audio Buffering

Three audio buffers are implemented to interface with the ADC **due to the overlap between audio windows.** When one buffer is being filled with new data, the other two are used for computing the current DFT. The capacity of each buffer is configurable up to **512 10-b** samples.

DFT

A single butterfly stage [Fig. 12 (top)] was implemented in **fixed-point** hardware allowing to compute two
intermediate results per cycle.

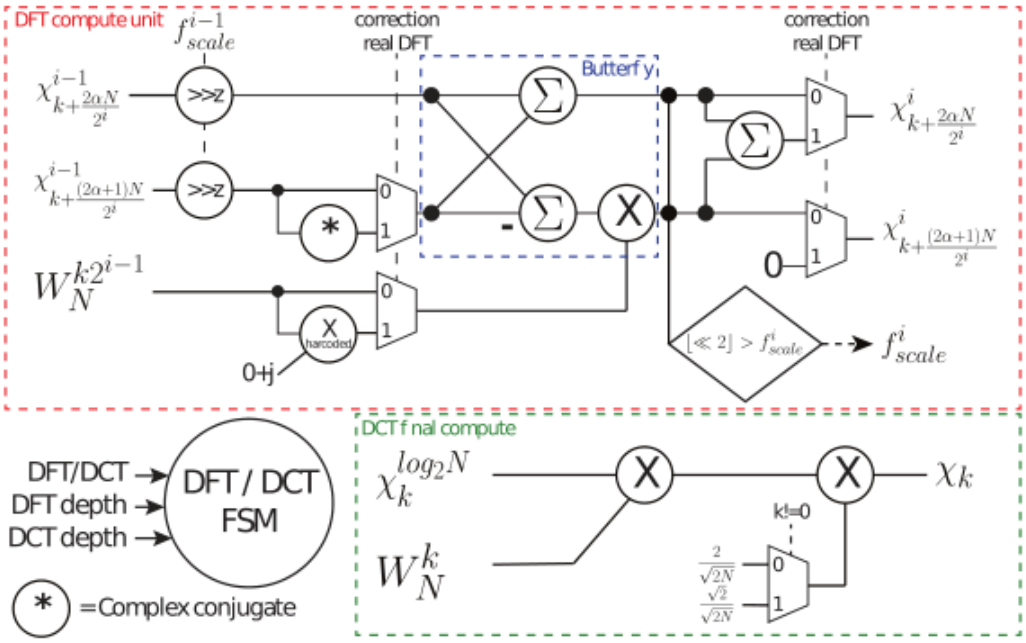

Fig. 12. DFT/DCT dataflow. (Top) Butterfly stage for DFT computation. During correction step [(4)] *correction real DFT* is set to 1. (Bottom) Final DCT computation.

1. In the first step, two values are read from each one of the input audio buffers, and the result is written to two dual-port SRAM computer memories.
2. For the following even cycles, two data points are read from the first dual-port SRAM and the twiddle factor is retrieved from the twiddle ROM, to perform a butterfly computation.
3. At the same time, the butterfly results of the previous clock cycle are saved to the second dual-port SRAM.
4. For odd cycles, the two input values are retrieved from the second dual-port SRAM for the butterfly computation, while previous cycle results are saved to the first dual-port SRAM.

In order to compute a 512-point complex DFT, 2048 cycles are required.

To avoid computation overflow, each cycle the outputs are evaluated on their magnitude. When they exceed |0.5|, the inputs in the next cycle are shifted down such that no value exceeds such maximum value.

Mel Filtering and Logarithm

Since each DFT output value contributes only up to two filters, a special memory format is proposed :

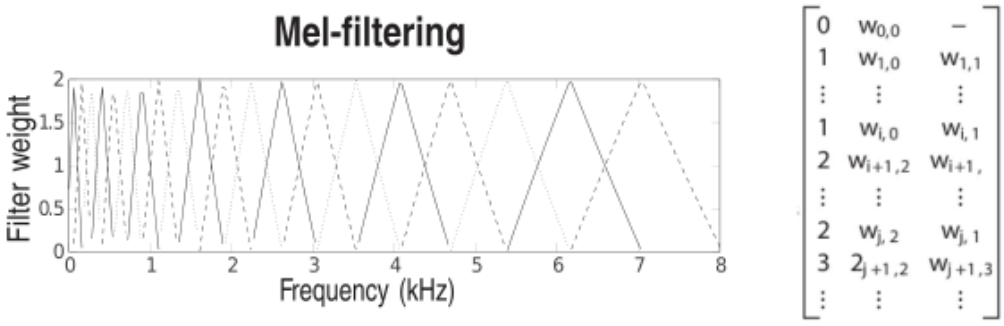

Fig. 13. (Left): Triangular mel filters for frequencies up to 8 kHz. (Right): Storage format for the mel-filter bank weights.

The memory contains one row for each DFT output value, composed of three parts: **highest filter index, weight 0 and weight 1**.

The first item refers to the highest Mel filter index to which the specific DFT value contributes.

Weight 0 and Weight 1 are the filter weights for the even and odd filter band to which the DFT value contributes.

These optimizations allow **saving M filter bands into 3M words**, with **5b and 12b for the filter index and the weights** respectively.

A **logarithm operation** works over the mel filtering output, implemented through **a lookup table**. Using the 10 b of the mel filtering output as the address of a 1024-entries LUT, a CORDIC implementation is avoided.

DCT + Time Derivatives

To perform the final transformation of a DCT from a complex DFT, a dedicated circuit [Fig. 12 (bottom)] implements the multiplications detailed in (5).

The required dimensionality for **KWS** and **SV** is **13** and **20** MFCC values, respectively.

For KWS, this means only the first 13 MFCC values are selected out of the 32 MFCC values computed.

For SV, the last 24 MFCC values are pairwise averaged to 12 new MFCC values. That is MFCC value 8 and 9 are combined to form MFCC value 8 for SV. The same combining is performed for time derivatives.

EVALUATION

input audio sampled at 16 kHz

Just 5.45 µW is consumed in idle mode mainly by leakage and the continuous operation of the AFE and SD.

If only KWS or SV mode is active, 16.11 µW and resp. 14.95 µW are consumed.

The FEx is the main source of energy dissipation for those states accounting for almost 50%.

The system operating at logic voltages from **0.6 V (250 kHz)** to 0.8 V (8 MHz) at SRAM voltage of 0.8 V

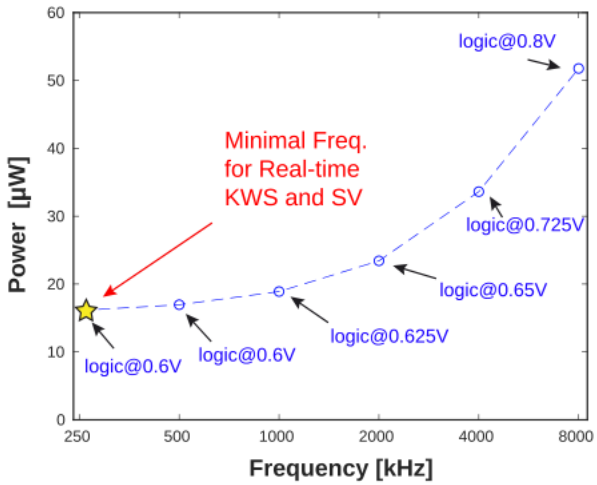

Fig. 18. Voltage-frequency scaling for full system active. SRAM voltage fixed at 0.8 V.

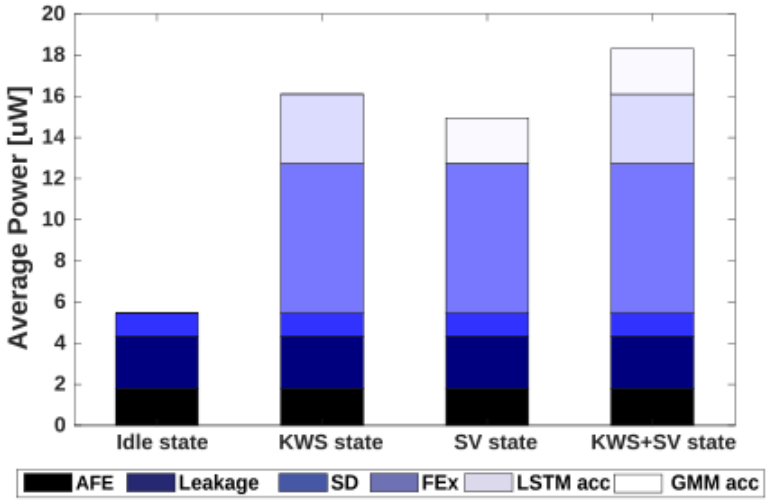

Fig. 19. Power breakdown for continuous operation in each state of the control unit FSM.

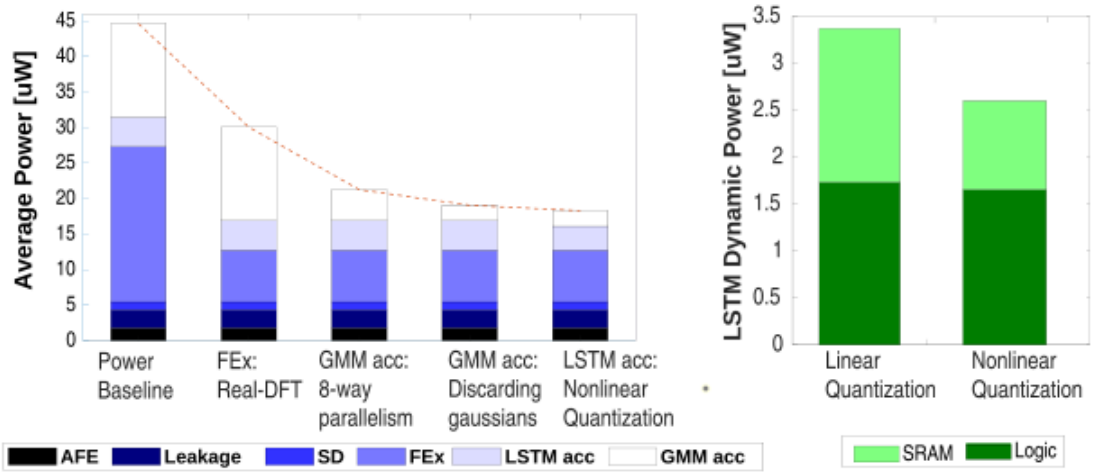

Fig. 22. (Left) Average power consumption for KWS+SV, showing the impact of the introduced optimizations. Each point corresponds to the average power after incrementally applying the corresponding technique. (Right) Dynamic power consumption of LSTM accelerator for LQ (8 bits) and NLQ (4 bits) for an LSTM network with 64 neurons.

TABLE II
STATE-OF-THE-ART COMPARISON

Reference	ISSCC'2017 [1]	VLSI'2018 [6]	ISSCC'2017 [24]	TVLSI'2014 [25]	This work
Technology	65nm	28nm	40nm	90nm	65nm
Area	13.17mm^2	1.29mm^2	7.1mm^2	19.53mm^2	2.56mm^2
Voltage	0.6V-1.2V	0.57V-0.9V	0.62V-0.9V	0.9V	0.6V-1.2V
AFE	NA	NA	NA	NA	✓
SD	✓	✓	NA	NA	✓
FE	✓	✓	✓	✓	✓
KWS	✓	✓	✓	NA	✓
SV	NA	NA	NA	✓	✓
Latency	Real-time	0.5ms-25ms	7ms	8ms(per feature vector)	KWS: 0.5-16ms(250kHz-8MHz) SV:515ms-1s(250kHz-8MHz)
Accelerators	VAD, MFCC DNN+HMM	VAD, MFCC BNN	FFT, DNN	LPC, SVM	VAD, MFCC LSTM, GMM
Stages	2 stages	2 stages	1 stage	1 stage	3 stages
KWS Accuracy	98.35% (TIDIGITS) 96.88%(WSJ)	96.11% (TIDIGITS)	NA	Not Reported	98.50% (TIDIGITS) 90.87%(GSCD)
SV Accuracy	NA	NA	NA	92.49% (NIST SRE)	99.5%(TIMIT)
Real-time Power	172μW@3MHz	141μW@2.5MHz	288μW@3.9MHz	8.12mW@100MHz	10.6μW@250kHz