Event-Driven Continuous-Time Feature Extraction for Ultra Low-Power Audio Keyword Spotting

S. Mourrane, Benoit Larras, A. Cathelin, Antoine Frappé

Event-Driven Continuous-Time Feature Extraction for Ultra Low-Power Audio Keyword Spotting

Soufiane Mourrane[1,2], Benoit Larras[2], Andreia Cathelin[1], Antoine Frappé[2]
[1]STMicroelectronics, Crolles, France
[2]Univ. Lille, CNRS, Centrale Lille, Junia, Univ. Polytechnique Hauts-de-France, UMR 8520 - IEMN, F-59000 Lille

***Abstract*—In the context of autonomous keyword spotting and sound detection, this paper proposes a low power feature extraction unit generating spectrograms that represent a unique signature allowing the classification of audio signals. This system is composed of a continuous-time digital signal processing feature extractor combined with a convolutional neural network engine. The study evaluates the hardware requirements to implement the feature extraction unit using an advanced CMOS process. Furthermore, a simulation of the complete system using *Matlab*® reveals that the recognition accuracy remains higher than 90% while offering a power consumption 4000X lower than a conventional discrete time system.**

I. Introduction

The increasing need of online sensing and autonomous recognition Internet-of-Things (IoT) devices pushes towards the development of embedded near-sensor processing units for applications such as sound detection and keyword spotting. However, dealing with real-world audio signals requires always-on power-hungry operations that cannot withstand low power and long battery lifetime constraints. In that regard, it is preferable to use a low-power selective audio processing, activating high-performance units only when a relevant signal is present at the input. Fig.1 shows an architecture where an audio signal is firstly analyzed by a keyword spotting unit, that generates a wake-up signal responsible of triggering the main processing elements. Inside the latter unit, a feature extractor performs the spectral division of the input signal into multiple frequency bands and extracts the per-band energies to be able to construct the spectrograms *i.e.* graphs displaying the energy in every band as function of time. Then, a classifier carries out the analysis by mapping the input spectrograms to some predefined classes, to decide or not to wake up the next processing blocks.

In such systems, the problem appears to be the implementation of the feature extraction unit with low hardware complexity level, ultra-low energy consumption as well as a reduced number of extracted features to simplify the classification mechanism [1]. In literature, several methods have been proposed to design the feature extractor. The filter bank can be built fully analog as in [2] and [3], where the audio signal is parallelly filtered using a 16 Gm-C bandpass filter bank, but this approach led to an architecture unable to scale with technologies or to ensure a certain configurable architecture to be adaptable for other audio applications. The spectrograms can also be generated using integrated FFT algorithms, as in [4] and [5]. This method appears to resolve the configurability issue, but still generates an excessive number of features that induce an increase in the latency, which will affect the system speed *i.e.* number of classifications per second.

In order to merge the advantages of both previous strategies, this work proposes an event-driven feature extraction unit based on continuous-time digital signal processing (CT-DSP) to implement the filtering in digital-friendly processes, as introduced in [6]. The event-driven operation builds a proportionality between dynamic power consumption and the activity of the input signal, what makes the system suitable for the recognition of keywords present in audio signals with long silence periods. The principal hardware-oriented parameters (such as number of bands, ADC resolution and energy quantizer resolution), are optimally chosen to scale down the system complexity while maintaining a recognition accuracy of 90%, as proposed in [1] for a discete-time digital system. Since in CT-DSP systems, only finite impulse response (FIR) filters with a limited number of taps are feasible [6], this paper proposes to study the implementation of a filter bank composed of 8 CT FIR filters, with a maximum of 16 delay taps each. In this configuration, low center frequencies are hardly reachable and the target frequency range is then reduced to 840Hz-6.25kHz. Moreover, the out-of-band attenuation of the filters will be limited. Even under these severe constraints, this study reveals that a recognition accuracy larger than 90%

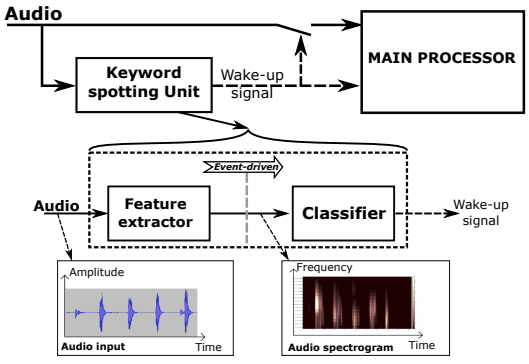

Figure 1. Structure of the pre-processing unit in the context of audio pattern recognition. The feature extraction block generates the spectrograms from the audio input signal.

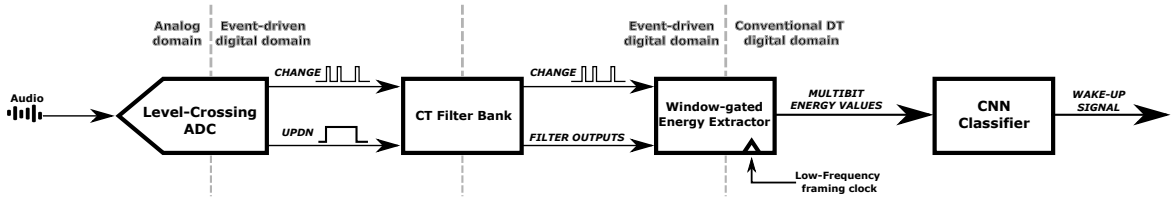

Figure 2. Structure of the continuous-time feature extraction block.

for 10-keyword spotting is still achievable when the produced spectrograms are classified with a state-of-the-art classifier relying on convolutional neural networks (CNN). Besides achieving high recognition accuracy, the system improves the power efficiency by a gain factor that can increase up to 4000 compared to a discrete-time system.

The remainder of this paper is organized as follows. Section II presents an overview of the entire CT processing chain. Section III details the operation and hardware implementation of the different components of the system. Section IV presents the results from simulations coupling the feature extraction with a CNN classifier and Section V concludes the paper.

II. CT PROCESSING CHAIN FOR AUDIO RECOGNITION

A. Audio recognition context

Conventional DT systems generate a constant number of events set by the sampling frequency, even if the input signal corresponds to a recording of a silent environment without any interesting information. However, CT systems avoid blind processing of the audio signals and generates samples only with the existence of a significant information-holding input. Thus, the advantage of CT processing system is that its performance changes with different audio scenarios. This work proposes the study of 3 different scenarios. Scenario #1 corresponds to a background noise inside a private room with no spoken keywords to test the system performance when no significant information is present at the input. Scenario #2 is about the recognition of 3 keywords per minute in the same background noise as before. Finally, the third scenario correspond to the recognition of 150 keywords per minute corresponding to conversational speech inside a noiseless environment.

B. CT processing chain architecture

Continuous-time digital signal processing (CT-DSP) systems provide the highest flexibility and scalability margins for digital systems, keeping the architecture clockless. The CT DSP is triggered to start the processing whenever a relevant event is present at the input, thus the power consumption will be strongly dependent on the input activity. Consequently, for signals with long silence periods and few sparse events, the consumption of CT DSP systems is drastically reduced compared to conventional sampled systems. The continuous-time processing chain starts with a level-crossing ADC (LC-ADC) that encodes asynchronously the input into a pair of bits (*CHANGE*, *UPDN*) indicating the crossing of some predefined threshold levels *i.e.* quantization levels, and the crossing direction, respectively. In that context, the events are defined as the crossings, thus the number of generated events directly depends on the number of threshold levels *i.e.* ADC resolution.

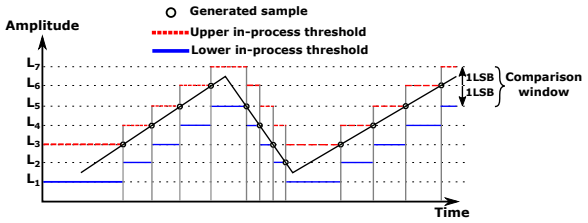

Figure 3. Illustration of the sample generation from an analog input signal in the LC-ADC. After each crossing, the comparison window is updated by ± 1 LSB.

Afterwards, a CT digital filter bank receives the pair (*CHANGE*, *UPDN*) at its input to decompose the spectral content of the audio signal into the frequency bands of interest. The filters outputs are then propagated along with *CHANGE* pulses towards the next block that performs energy estimation. This energy extractor block operates by squaring the values of the outputs of each filter, then accumulates these squared values over a precise time window defined by the period of a low frequency clock, as detailed in Fig.2. The window or frame duration is set to 25ms. This block is the boundary between the CT and the low-frequency DT domains.

III. HARDWARE IMPLEMENTATION OF CT-DSP

A. Level-crossing ADC

The LC-ADC is an asynchronous delta encoder composed of two analog comparators continuously comparing the input signal to a predefined set of high and low amplitude thresholds, to generate the *CHANGE* and *UPDN* signals. Then, the generated pair of bits is injected into a digital unit that refreshes the values of the low and high thresholds defining the comparison window [6], [7]. The operation principle of the LC-ADC is shown in Fig.3, where a sample is taken whenever the input signal fulfills the following conditions: $input \geq Upper\ threshold$ or $input \leq lower\ threshold$. The thresholds are updated by $\pm LSB$, depending on the direction of the change. Thus, signals with low amplitude generates low number of events, which will decrease the system power consumption. An inconvenient of such ADCs is the high sensitivity to noisy signals, that generate a large number of insignificant events, that degrades the power efficiency. A

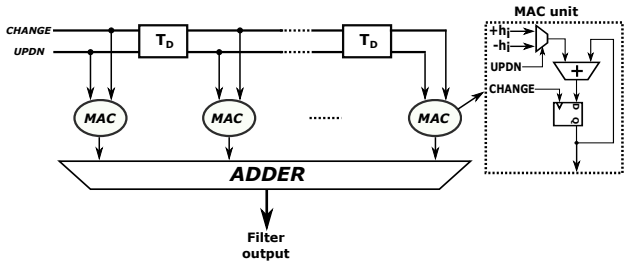

Figure 4. Schematic of a digital CT 16-tap FIR filter. For each *CHANGE* bit, the coefficients are accumulated depending on the *UPDN* polarity, and replace the multiplication in conventional FIR filters. A delay element T_D propagates the *CHANGE* and *UPDN* bits from one tap to the next one.

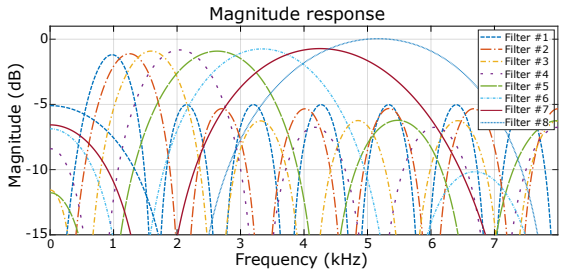

Figure 5. Magnitude response of the proposed bank of 8 digital CT 16-tap FIR filters. The stop-band attenuation of -5dB releases the design constraints to minimize the number of taps in each filter.

solution for this problem has been presented in [7], where a single sample is generated per each comparison window. Consequently, the upper and lower thresholds are separated by $2LSB$, as depicted in Fig.3.

B. Digital CT FIR filter bank

Thanks to the delta encoding of audio input, the CT FIR filter structure is simplified compared to a conventional structure, as depicted in Fig.4. The signals *CHANGE* and *UPDN* are delayed using analog delay taps (T_D) to maintain the clockless operation. Then, the signal pair is fed into the Multiplier/Accumulator (MAC) unit to perform successive additions of the filter's coefficients. During our investigations, we found that 6-bit coefficients are suitable for providing a decent selectivity and recognition accuracy. Finally, a multi-bit adder receives the MAC results and generates the multibit FIR output.

The main challenge faced in CT filtering is that low frequency and high selective filters need to have high orders *i.e.* high number of delay taps, which will increase the hardware complexity as well as the number of generated combinatorial events. The reduction of the frequency range as well as the degradation of the selectivity *i.e.* out-of-band attenuation, relaxes the complexity constraints. For speech recognition, it has been demonstrated in [1] that 8 filters logarithmically distributed between 75Hz and 7kHz can extract sufficient features to recognize 10 keywords with 90% accuracy. Since the human voice does not have a significant spectral content below 800Hz (only information on voice intonation and accent are present) [8], we propose to first restrict the frequency range of interest from 75Hz-7kHz to 840Hz-6.25kHz to reduce hardware complexity. This range is similarly divided into 8 logarithmically-distributed bands, using CT FIR filters. Reducing bandwidth results in a degradation in the accuracy lower than 2% compared to our initial setup in the 10-keyword detection case. In a second step, decreasing the out-of-band attenuation from 20dB to 5dB, reduces the accuracy from 93.1% down to 90.6%, when using a 6-bit LC-ADC. This accuracy loss is balanced by the gain in hardware complexity, enabling the implementation of the CT filters with a maximum of 16 taps. The magnitude response of the proposed filter bank is presented in Fig. 5.

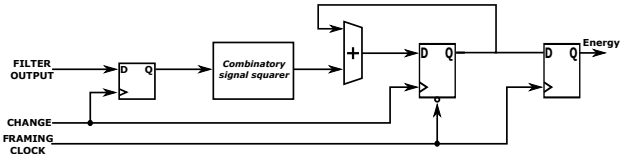

Figure 6. Schematic of an energy extractor. A combinatory signal squarer outputs the squared value of the input signal, which is accumulated during a 25-ms framing clock cycle.

C. Energy extraction

The filters outputs are squared and accumulated to extract the energy in each band by the circuit shown in Fig. 6. The combinatory squarer and the following accumulator are clock-gated by the *CHANGE* signal, thus an energy value is extracted only with the existence of an event at the input. Then, the squared results are accumulated during the framing clock period *i.e.* frame. Finally, the frames, lasting 25ms, are concatenated to form the spectrograms.

IV. Simulation Results

The CT processing chain has been simulated with the Google *Tensorflow*® speech commands dataset (GSCD) [9] and *Librispeech* dataset [10], using *Matlab*®.

Table I displays the number of generated events by the LC-ADC for the three scenarios for an ADC resolution varying from 5 to 9 bits. The number of generated samples scales with signal activity . The values of generated events are compared to a reference value of 960,000 events that corresponds to the number of generated events in the conventional discrete-time case using an ADC working with a sampling frequency equal to 16kHz during 1 minute.

To validate the benefits of the system applied on low activity signals, another simulation is proposed in Table I describing the evolution of the power consumption with the LC-ADC resolution. The model used to estimate the consumption calculates the number of logic gates and multiplies by a typical power for each event, extracted using Spectre simulator considering 28nm FDSOI technology. In this technology, a delay tap consumes 15 fJ/event, as stated in [11], thus the delay line dominates the CT system power consumption. For the DT system, the MAC units are dominating the power consumption,

	LC-ADC Resolution	5 bits	*6 bits*	7 bits	8 bits	9 bits
Scenario #1 in CT system	No of events	12	*198*	1,707	19,326	88,431
	Average power (nW)	0.003	*0.071*	0.96	34.19	301
Scenario #2 in CT system	No of events	5,436	*15,357*	38,022	105,660	267,705
	Average power (nW)	1.96	*8.62*	36.7	301	912.5
Scenario #3 in CT system	No of events	7,044	*29,092*	118,240	783,340	1,762,600
	Average power (nW)	10.5	*66.4*	306.1	1,390	6,000
DT system	No of events	960,000				
	Average power (nW)	311.5				

Table I
NUMBER OF EVENT AND POWER CONSUMPTION IN CT CASE FOR THE DIFFERENT SCENARIOS COMPARED TO DT CASE.

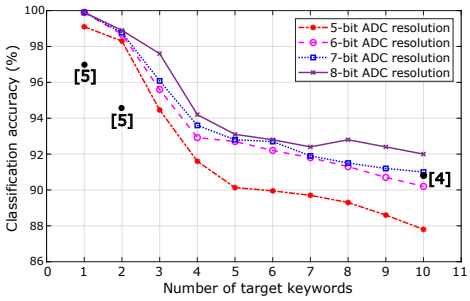

Figure 7. Evolution of recognition accuracy as function of number of target keywords for different ADC resolution. Comparison with accuracies from state-of-the-art works, using GSCD dataset.

considering that the MAC is composed from a ripple carry array multiplier. The reference power consumption is evaluated as 311.5nW in the case of a DT system. From Table I, we can observe that the CT system stays advantageous in terms of power consumption for all the scenarios up to 7-bit resolution.

To evaluate the trade-off between power consumption and system's performance, the generated spectrograms, with hardware-oriented optimized parameters, are classified using a state-of-the-art CNN consisting of 6 filtering steps composed of a 2-dimensional convolution layer and a rectifier liner unit (ReLU) layer. Between 2 successive filtering stages, an additional max-pooling layer is inserted. At the end, a fully connected layer, with a size matching the number of target keywords, is set after the last filtering step [1]. This kind of CNN classifiers is compatible with low-power embedded implementation [12]. To verify the relevance of the extracted features, a simulation describing the variation of the system recognition accuracy as a function of the number of target words and LC-ADC resolution is presented in Fig.7. The system naturally becomes more accurate with low number of target keywords. The recognition of 1 and 2 keywords is achieved with an accuracy higher than 99.1% and 98.1%, respectively, compared to an accuracy of 97.1% @1word and 94.6% @2words (black dots in Fig.7), with an average consumption of 336.6nW, for feature extractor, in [5], which is 5X higher than the CT system. Moreover, the system stays accurate even for a higher number of keywords, as shown in Fig.7, where the recognition of 10 keywords is achieved with an accuracy varying from 88.6% up to 92%, depending on the LC-ADC resolution, compared to an accuracy of 90.87% in [4] for GSCD dataset. The grey column in Table I, corresponding to 6-bit resolution, represents the best compromise between power consumption and recognition accuracy, because the CT system maintain an accuracy higher than 90% with reducing the power consumption by a factor of 4000, 40 and 5 compared to DT system for scenarios #1, 2 and 3, respectively.

V. CONCLUSION

An audio low-power continuous time feature extraction unit targeting the recognition of up to 10 keywords is simulated with different speech scenarios. This event-driven CT system allows the recognition of 10 keywords with an accuracy higher than 90% while being up to 4000X more power efficient than an equivalent discrete-time system. The simulated classification accuracy is at par or better than state of the art.

VI. ACKNOWLEDGEMENT

This work was supported in part by the French National Research Agency under Grant ANR-18-CE24-0006-01 LEOPAR.

REFERENCES

[1] S. Lecoq et al., "Low-complexity feature extraction unit for "Wake-on-Feature" speech processing," 2018 25th IEEE ICECS, Bordeaux, 2018, pp. 677-680.
[2] K. M. H. Badami et al., "A 90 nm CMOS, 6 µW Power-Proportional Acoustic Sensing Frontend for Voice Activity Detection," in IEEE Journal of Solid-State Circuits, vol. 51, no. 1, pp. 291-302, Jan. 2016.
[3] M. Yang et al., "A 1 µW voice activity detector using analog feature extraction and digital deep neural network," 2018 IEEE ISSCC, San Francisco, CA, 2018, pp. 346-348.
[4] J. S. P. Giraldo et al., "18 µW SoC for near-microphone Keyword Spotting and Speaker Verification," 2019 Symposium on VLSI Circuits, Kyoto, Japan, 2019, pp. C52-C53.
[5] W. Shan et al., "A 510-nW Wake-Up Keyword-Spotting Chip Using Serial-FFT-Based MFCC and Binarized Depthwise Separable CNN in 28-nm CMOS," in IEEE Journal of Solid-State Circuits, vol. 56, no. 1, pp. 151-164, Jan. 2021.
[6] B. Schell and Y. Tsividis, "A Continuous-Time ADC/DSP/DAC System With No Clock and With Activity-Dependent Power Dissipation," in IEEE Journal of Solid-State Circuits, vol. 43, no. 11, pp. 2472-2481, Nov. 2008.
[7] T. Marisa et al., "Pseudo Asynchronous Level Crossing adc for ecg Signal Acquisition," in IEEE Transactions on Biomedical Circuits and Systems, vol. 11, no. 2, pp. 267-278, April 2017.
[8] L.Rabiner and B.Juang, *Fundamentals of speech reecognition*. PTR Prentice-Hall. Inc., New Jersey, 1993.
[9] P.Warden, "Speech commands: A public dataset for single-word speech recognition." *Dataset available from http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz*,2017.
[10] V. Panayotovet al., "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE ICASSP, Brisbane, QLD, 2015, pp. 5206-5210. [Online]. Available: http//www.openslr.org/12/
[11] A. González et al., "A Wide Tuning Range Delay Element for Event-Driven Processing of Low-Frequency Signals in 28-nm FD-SOI CMOS," in IEEE Solid-State Circuits Letters, vol. 3, pp. 198-201, 2020.
[12] S. Zheng et al., "An Ultra-Low Power Binarized Convolutional Neural Network-Based Speech Recognition Processor With On-Chip Self-Learning," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 12, pp. 4648-4661, Dec. 2019.