Shen, Y., Straeussnigg, D. & Gutierrez, E. (21-25 May 2023). Towards Ultra-Low Power Consumption VAD Architectures with Mixed Signal Circuits [proceedings]. 56th Edition IEEE ISCAS 2023: IEEE International Symposium on Circuits and Systems (ISCAS), Monterey, CA, USA.

DOI: 10.1109/ISCAS46773.2023.10181669

Towards Ultra-Low Power Consumption VAD Architectures with Mixed Signal Circuits

Yukai Shen
Electronic Technology Department
University Carlos III of Madrid
Madrid, Spain
yshen@ing.uc3m.es

Dietmar Straeussnigg
Power and Sensor Systems
Infineon Technologies
Villach, Austria
Dietmar.Straeussnigg@infineon.com

Eric Gutierrez
Electronic Technology Department
University Carlos III of Madrid
Madrid, Spain
eric.gutierrez@uc3m.es

***Abstract*—A voice activity detector architecture based on an analog feature extractor and a mixed signal classification stage is proposed for ultra-low power activity. The feature extraction stage is composed of a set of analog band-pass filters and frame energy estimators. The classification stage has a fully connected first layer built with ultra-low power consumption ring oscillators, followed by gated recurrent unit layers. The ring oscillator based layer consumes nWs according to transient simulations performed in a low power 65 nm CMOS technology. Additionally it features the ability to perform the analog-to-digital conversion required to handle subsequent GRU layers, as well as the possibility of computing a non-linear function like sigmoid seizing the intrinsic non-linearity of the ring oscillator. Training and testing operations are made proving competitive classification performance between a baseline model and our proposed architecture. In light of this, proper features for deployment on power-restricted edge-computing applications are shown.**

***Index Terms*—Voice activity detection (VAD), analog feature extraction, recurrent neural network (RNN), gated recurrent unit (GRU), ring oscillator (RO).**

I. Introduction

The huge development of computing capabilities has enabled efficient implementations of artificial intelligence tasks over portable devices, such as cell-phones or wearables. However, due to battery life restrictions, limiting the power consumption is crucial, requiring the design of ultra-low power architectures, in the range of hundred of nWs [1], [2]. Regarding smart portable applications, speech recognition has become one of the topics that has received more attention, highlighting voice activity detectors (VADs). VADs are able to detect human voice within noisy environments during an always-on operation. The interest of this task relies on directly making use of the human voice as a command for other purposes, such as waking up a device or as the first processing stage of another more complex task like full-audio conversion or keyword spotting.

VADs require continuous monitoring of the input audio stream, typically sensed by a microphone. The conventional way to proceed is turning the analog input raw data into digital data, and then performing intensive digital computation (windowing, FFT, filtering and energy estimations), to extract the features within different spectral bands and looking for data patterns compatible with human voice. This approach leads to accurate classifications but consumes much power [3]. Another more energy efficient approach consists of taking the intensive digital computation into the analog domain to make at least the feature extraction stage by means of analog band-pass filters and energy estimators [1], [4]. Additionally contextual information is crucial to improve the performance of VADs [5], [6] where long- and short-term time dependencies between the input features become a meaningful parameter to achieve excellent accuracy in the classification [7], [8].

This paper was supported by program H2020-MSCA-ITN-2020 grant Nr.956601.

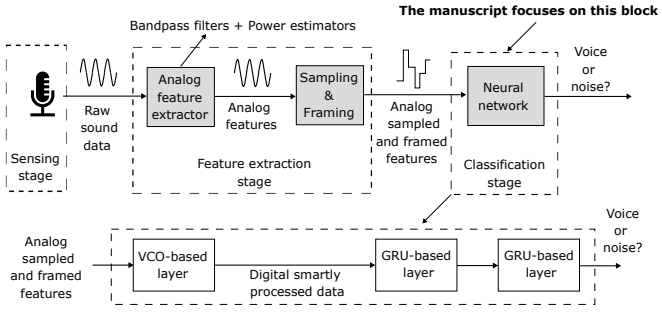

Fig. 1. General scheme of the proposed VAD architecture.

In this light our manuscript goes towards the proposal of new architectures for ultra-low power and excellent classification accuracy VAD tasks by means of combining mixed-signal circuits and recurrent neural networks. A general scheme of the proposal is depicted in Figure 1. The raw sound data are measured from a microphone and connected to the feature extraction stage. The features are extracted from a set of analog band-pass filters. The outputs of the filters are sampled, framed, and the energy for each frame is computed to feed the classification stage. The classification stage is composed of a fully-connected first layer built with ring oscillators (ROs) based multiply-accumulate (MAC) units, leading to ultra-low power consumption, followed by two digital unidirectional gated recurrent unit (GRU) layers, a fully connected (FC) layer and a softmax layer. This manuscript focuses on the design and validation of the classification stage.

The outline of the manuscript is as follows. Section II

describes in detail the proposed architecture for VAD tasks. Section III shows the training dataset and the validation results. Section IV provides some first data regarding the expected power consumption of the first layer in the neural network. Finally Section V concludes the manuscript.

II. Proposal for feature extraction and classification stages in VAD tasks

A. Analog based feature extraction stage

A conventional way to extract the audio features in the analog domain is shown in Figure 2. A set of analog band-pass filters with different center frequencies is firstly applied to the raw audio input data, then wave rectifiers are used, and the energy within a frame can be estimated by integrating the rectified signal over the frame time step.

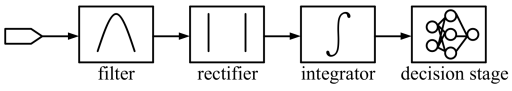

Fig. 2. General scheme of the analog feature extraction stage.

In our work, this kind of analog feature extraction stage is simulated in MATLAB. The band-pass filter is modeled by connecting a high-pass filter and a low-pass filter in cascade, leading to the following band-pass transfer function:

$$H(s) = \frac{s \cdot w_H}{s^2 + (w_H + w_L) \cdot s + w_H \cdot w_L}, \tag{1}$$

where w_H and w_L denote the upper and lower cut-off frequency, respectively. Considering that this work is in cooperation with an industrial partner, the details of the filters bands are not provided.

The outputs of the filters are framed. A frame length of 10 ms is used with a 7-ms frame shift. A frame is viewed as the basic decision unit [8]. To ease the computation, we use the short-time magnitude (STM) quantity to estimate the frame-wise energy. The STM quantity E_k for a given *k*-th frame containing *n* sampling points *x[i]*, is calculated by adding up all the absolute sampling values, as shown in Equation (2). The STM is not sensitive to large signal levels since it does not include squaring.

$$E_k = \sum_{i=1}^{n} |x[i]| \,. \tag{2}$$

B. Mixed-signal based classification stage

A VAD task can be treated as a sequence-to-sequence binary classification task. To consider contextual information the classification stage is built with a unidirectional recursive neural network (RNN) implemented with GRU units. The RNN classifier takes the sequential frame-wise features and predicts a sequence of outputs that has the same length as the input sequence.

To benchmark our RO-based RNN classifier, we firstly train and test a baseline RNN model that consists of 2 GRU layers composed of 10 hidden units each, and a FC layer with 2 neurons followed by a softmax layer. Secondly we insert an extra FC layer composed of 8 units before the GRU layers, and train the neural network keeping the rest of the architecture the same. This extra layer will become later on the ultra-low power RO-based layer. Both models are trained and tested in TensorFlow. The second one is then imported into a Simulink environment where the simulation of RO-based layers is much user-friendly (see Figure 3).

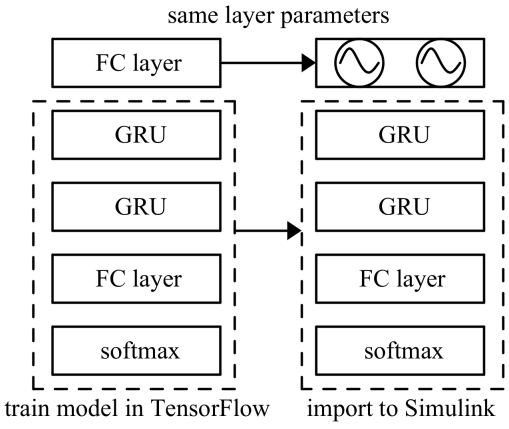

Fig. 3. Design flow of the neural network that includes a RO-based FC layer.

Figure 4 illustrates the block diagram of how the first FC layer is built with ROs, where the inputs $x_1, x_2, ..., x_n$ are the analog features coming from the previous stage. For each channel, the computed frame feature is sampled at every T_{frame}, and multiplied by the corresponding pre-trained weight, then the weighted results are summed up over all the channels and the bias value is added (not shown here) leading to an analog signal at node *A*.

The oscillation frequency of a RO can be described as follows:

$$f(t) = f_o + K_{VCO} \cdot A(t), \tag{3}$$

where f_0 denotes the rest oscillation frequency, K_{VCO} is the gain, and $A(t)$ is the analog signal at node *A*. If we look at the output of the RO in the phase domain, it can be treated as a phase integrator and Equation (3) can be rewritten as:

$$\phi(t) = 2\pi \int_0^t (f_0 + K_{VCO} \cdot A(\tau))d\tau = 2\pi f_0 t + 2\pi K_{VCO} \int_0^t A(\tau))d\tau, \tag{4}$$

where $\phi(t)$ is the instantaneous phase of the RO.

Assuming a square output signal in the RO an edge (either rising or falling) occurs whenever the phase $\phi(t)$ is a multiple of π. Asynchronous counters can be connected to the output signal of the RO to quantify the phase. By connecting a parallel RO always oscillating at the rest oscillation frequency and subtracting both outputs the resulting signal at node *B* is:

$$B(t) = 2\pi K_{VCO} \int_0^t A(\tau)d\tau, \tag{5}$$

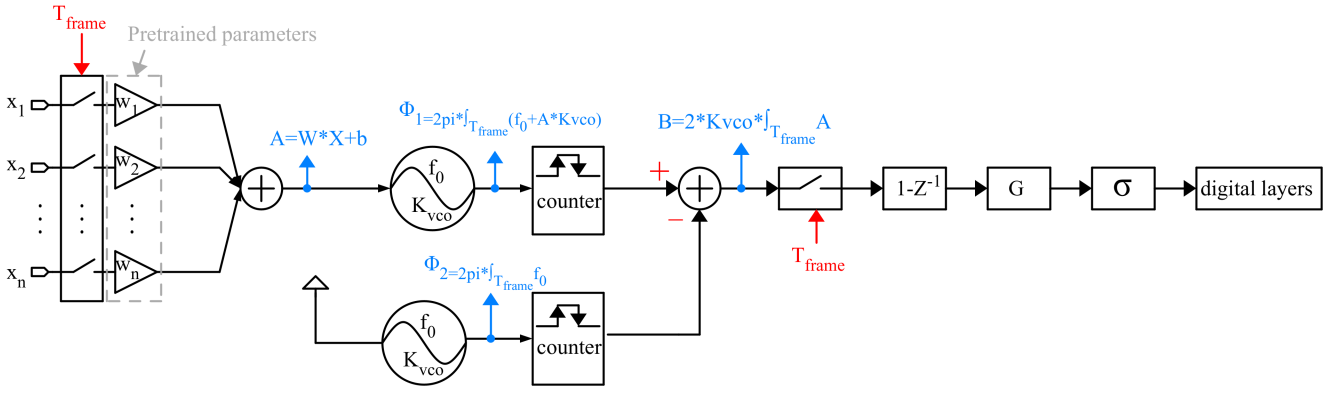

Fig. 4. Block diagram of the units in the first FC layer built with ROs.

which is the same operation performed by a conventional MAC unit inserted into a FC layer. Notice the advantages of this approach. On the one hand, the RO-based layer performs continuously in time with f_0 and K_{VCO} parameters that can be tuned for ultra-low power operation. On the other hand the RO is making the analog-to-digital conversion itself at the same time of performing the first stage of the classification, saving much power and area.

After the RO the remaining operations (sampling, first difference and gain) are performed in the digital domain. The gain G is $1/(2K_{vco}T_{frame})$ for a counter sensitive to both edges, and is used to get a final integrated expression whose gain equals 1 and compatible with the training previously performed in TensorFlow. Regarding the oscillation parameters of the RO, the higher K_{VCO} the higher the resolution achieved in the analog-to-digital conversion, but also the higher the power consumption. In this manuscript, the rest oscillation frequency f_0 is set to 1 MHz, and the oscillator gain to 0.1 MHz/V. Finally the activation function is applied (a sigmoid function) and the result is fed to the following GRU layers. Note that the activation function could be integrated into the RO operation seizing the non-linear voltage/current-frequency transfer function [9] saving even more power.

III. Simulation Results

A. Dataset

The model training and evaluation data are synthesized based on the clean speech from TIMIT corpus [10] and the noise from DEMAND database [11]. The TIMIT corpus consists of speech data from 630 speakers. Each speaker reads 10 sentences. The DEMAND noise database contains multi-channel recordings of acoustic noise in diverse environments, including 6 categories with 18 scenarios. Both the clean speech and noise data are sampled at 16 kHz. The TIMIT corpus has already been divided into training and test sets. To build the training data, we randomly select 500 clean utterances from 500 different speakers in the training set, concatenate them, and add silence segments with random length (maximum 3 seconds) in-between. We also add silence at the beginning of the concatenated data, for a better balanced speech and non-speech class. The evaluation data is designed in the same way with another 50 clean utterances from the test set. The overall duration of the training utterances is about 38 minutes with the speech class accounting for 58.8%, and around 4.1 minutes and 57.6% for the evaluation utterances. Different types of noise are added to the clean speech at different signal-to-noise ratio (SNR) levels, ranging from -10 dB to 20 dB with a step of 5 dB.

Here, 4 types of noise, *kitchen*, *park*, *office room* and *restaurant* noises, are individually added to the clean training utterances at random SNRs. The clean and noisy training utterances are then concatenated, constructing a total training data with a length of about 5.7 hours and the percentage of the speech class stays the same. The clean evaluation utterances are overlaid with noises at different SNRs as well, including 6 different background noises from a busy *cafeteria*, a *sports field*, a *hallway* inside an office building, a *washing room*, a *subway*, and a *living room* with background music. Note that the evaluation data contains different noise types from the training data.

In this manuscript, the VAD decision is on a frame-by-frame basis, a frame is treated as an example for training and testing, it can only be categorized as *speech* (denoted by *'1'*) or *non-speech* (denoted by *'0'*). Given a frame containing a number of samples, if the samples labeled as speech are more than a half, this frame will be assigned to speech class, otherwise, it will be labeled as non-speech. We use the word boundary information in the TIMIT corpus as the ground truth, and build the training targets in the way described above.

B. Training settings

To train the neural networks, the input features are first logarithmic scaled and normalized to have zero mean and unity standard deviation, and then split into a set of sequences with a time step of 800 with 75% overlap, and the ground-truth labels are also divided accordingly. The training epoch is set to 20 to prevent overfitting, and the batch size is 64. In the GRU cells, we use the sigmoid function to compute the gates, and the hyperbolic tangent function (tanh) to update the cell

and hidden states. In the FC layer, the sigmoid function is used. The initial learning rate is 0.001. The data is shuffled before each training epoch. And the *Adam* optimizer [12] is used for updating the weights and biases.

C. Results

Receiver operating characteristic (ROC) curve is commonly used as a metric of VAD models for its insensitivity to the skewed class distributions in the classification problems [13], and the area under an ROC curve (AUC) is a quantitative measure of the ROC curve. Here we use AUC as the metric to evaluate our work.

TABLE I
AVG. AUC COMPARISON BETWEEN THE BASELINE MODEL, THE BASELINE MODEL PLUS A FC LAYER, AND THE LATTER IMPLEMENTED WITH ROS

		Baseline model	**Proposed model**	**Proposed model w/. RO**
Noise	**SNR**	**N1**	**N2**	**N3**
cafeteria field hallway washing metro living room	20 dB	99.51%	99.53%	99.41%
	15 dB	99.37%	99.47%	99.35%
	10 dB	98.99%	99.32%	99.19%
	5 dB	97.98%	98.94%	98.79%
	0 dB	95.76%	97.74%	97.59%
	-5 dB	90.77%	94.36%	94.23%
	-10 dB	80.25%	88.89%	88.77%
Clean speech		99.63%	99.54%	99.46%

Table I lists the average AUC results per SNRs on test data with different background noises, for the baseline model N1, the baseline model N1 plus the additional FC (N2), and the proposed model with the first FC layer built with ROs (N3). By inserting an FC layer before the GRU layer, the proposed model N2 has a boosted performance against the baseline model N1, showing a significant improvement on the signals with poor SNRs. When the first FC layer is built with ROs, quantization noise is introduced into the proposed model due to the analog-to-digital conversion, and a slight AUC degradation can be observed. However, the RO based model N3 still outperforms the baseline model N1. The model parameter size is increased from 1102 for the baseline model to 1262 for the proposed model with 1 extra FC layer, increased by 14.5%.

Moreover, our proposed model with ROs based first FC layer (N3) exhibits an average speech and non-speech hitting rate of 97.23% and 95.22%, respectively, on the picked 10dB SNR test data. Compared to the state-of-art silicon measurement results, 90.1% and 94% speech and non-speech hitting rate on a 20 minutes, 10dB SNR audio test data, as reported in [14], our model provides an enhanced VAD performance, especially on the speech hitting rate.

IV. POWER ESTIMATION OF THE RO-BASED FC LAYER

To estimate the power consumption of a RO with the parameters we defined in the Simulink model, a RO with 3 taps is designed in a low power 65 nm CMOS technology and simulated. Figure 5 shows the circuit diagram of our 3-tap RO. All the transistor are set to the minimum size. The supply voltage is 0.4 V. The input voltage can be used to control the current flowing into the RO and thus tune the output frequency by means of transconductors. Several transconductors may be placed in parallel for different input signals. As a first approach the RO is designed according to the parameters used in the Simulink model, getting an average current of 7.5 nA and leading to a power consumption of only 3 nW per unit in the layer. In our case with 8 units, this becomes a power of 24 nW excluding the counters.

Considering the rest oscillation frequency f_o, the sampling frequency, and that the counters are triggered by both the rising and the falling edges, the number of bits required for the counters is 15 bits. Less bits could be used if some classification performance degradation is admissible due to higher quantization noise. The digital counter may be implemented with binary coding, leading to important area savings and possible automatic place-and-route digital designs.

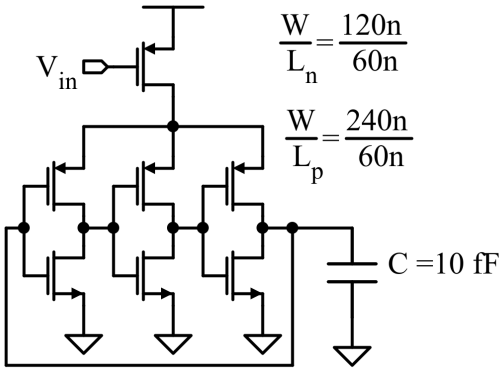

Fig. 5. Designed 3-tap RO for power estimation.

V. CONCLUSION

We propose an ultra-low power architecture for VAD tasks combining analog and digital circuitry for the feature extraction and classification stages. In this manuscript we focus on the proposal of the classification stage, composed of a RO-based FC layer and digital GRU layers. The power consumption of the RO-based layer is in the nWs range and performs the analog-to-digital conversion, always required for digital GRU-based layers. The neural network also includes the smart processing of the RO-based layer, showing relevant accuracy performance in comparison to simple neural networks not including the RO-based FC layer. The excellent power performance of the RO-based layer and the competitive classification performance result in a promising architecture to be used on battery-powered edge devices.

REFERENCES

[1] K. M. H. Badami, S. Lauwereins, W. Meert and M. Verhelst, "A 90 nm CMOS, 6μW Power-Proportional Acoustic Sensing Frontend for Voice Activity Detection," IEEE Journal of Solid-State Circuits, vol. 51, no. 1, pp. 291-302, Jan. 2016.

[2] M. Yang, C. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar and M. Seok, "Design of an Always-On Deep Neural Network-Based 1- μ W Voice Activity Detector Aided With a Customized Software Model for Analog Feature Extraction," IEEE Journal of Solid-State Circuits, vol. 54, no. 6, pp. 1764-1777, June 2019.

[3] A. Raychowdhury, C. Tokunaga, W. Beltman, M. Deisher, J. W. Tschanz and V. De, "A 2.3 nJ/Frame Voice Activity Detector-Based Audio FrontEnd for Context-Aware System-On-Chip Applications in 32-nm CMOS," IEEE Journal of Solid-State Circuits, vol. 48, no. 8, pp. 1963-1969, Aug. 2013.

[4] M. Croce, B. Friend, F. Nesta, L. Crespi, P. Malcovati and A. Baschirotto, "A 760-nW, 180-nm CMOS Fully Analog Voice Activity Detection System for Domestic Environment," IEEE Journal of Solid-State Circuits, vol. 56, no. 3, pp. 778-787, March 2021.

[5] X. -L. Zhang and D. Wang, "Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 2, pp. 252-264, Feb. 2016.

[6] V. Peddinti, D. Povey, S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," Proceedings of Interspeech, ISCA, 2015.

[7] S. Tong, H. Gu and K. Yu, "A comparative study of robustness of deep learning approaches for VAD," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[8] F. Eyben, F. Weninger, S. Squartini and B. Schuller, "Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[9] E. Gutierrez, C. Perez, S. Paton and L. Hernandez, "Low Power Phase-Encoded MAC Accelerator for Smart Sensors with VCO-based ADCs," 2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS), 2020

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgrena, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," 1993.

[11] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings," J. Acoust. Soc. Amer., vol. 133, no. 5, p. 3591, May 2013.

[12] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proc. Int. Conf. Learn. Represent., pp. 1–41, 2015.

[13] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, Jun. 2006.

[14] F. Chen, K. -F. Un, W. -H. Yu, P. -I. Mak and R. P. Martins, "A 108-nW 0.8-mm2 Analog Voice Activity Detector Featuring a Time-Domain CNN With Sparsity-Aware Computation and Sparsified Quantization in 28-nm CMOS," in IEEE Journal of Solid-State Circuits, vol. 57, no. 11, pp. 3288-3297, Nov. 2022, doi: 10.1109/JSSC.2022.3191008.