A 760 nW, 180 nm CMOS Fully-Analog Voice Activity Detection System for Domestic Environment

Marco Croce, Brian Friend, Francesco Nesta, Lorenzo Crespi, Piero Malcovati, *Senior Member, IEEE,* Andrea Baschirotto, *Fellow, IEEE*

***Abstract*—This paper presents a fully analog, signal-to-noise-ratio-based voice-activity detection circuit. The circuit is based on an energy-efficient analog implementation with continuous-time non-linear operation and fully-passive switched-capacitor signal processing. The overall chain is composed by a programmable-gain amplifier, a squarer, an integrator, a SC-based signal averaging circuit, and a periodic threshold update circuit for adaptability. This implementation allows the minimization of both power consumption and chip area. The VAD circuit prototype has been fabricated in a 180 nm CMOS technology and occupies an area of 0.14 mm^2. The device achieves 99.5 % classification accuracy in domestic environment in the presence of loud ambient noise, consuming 760 nW from a 1.2 V supply.**

***Index Terms*—Voice Activity Detection, low power analog circuit, audio signal processing, signal energy calculation, switched capacitor circuit.**

I. Introduction

Voice Activity Detection (VAD), also known as "*speech activity detection*" or "*speech detection*", is the function identifying the presence or absence of human speech in an audio signal. There are several applications where VAD can be applied as: speech recognition, speaker verification, speech enhancement, voice operation switch, and voice over internet protocol.

In the high-performance audio signal processing chain of Fig. 1, committed to process incoming voice, the VAD function is implemented in the DSP block at the end of the chain. Therefore, the full chain is always operating, independently of the presence of a voice signal, with consequently a large amount of power consumption, which is wasted during periods with no voice presence. In the system of Fig. 2, the VAD function is implemented in an additional low-power block, connected in parallel to the main signal processing chain. This architecture strongly reduces the power consumption. In fact, when no voice is present, only the low-power VAD block is active, saving a large amount of power. On the other hand, when the VAD block detects the presence of voice, it turns-on the high-performance signal processing chain. In this way the high-performance chain power is consumed only when necessary (i. e. when voice is present), while, otherwise, only the VAD block is active with very low power consumption.

Based on the above VAD function description, such a VAD block has to be implemented with the following features:

- *low power consumption*: as an always-on and real-time application, the VAD circuit power consumption must be extremely low;
- *accurate decision rule*: a physical property of the incoming audio signal frame is used to detect the presence or absence of speech, providing consistent and accurate classification;
- *adaptability*: the ability to handle non-stationary background noise variations improves robustness.

In the literature there are many techniques for detecting human voice, implemented either in the digital or in the analog domain. Most of them focus the analysis on one or more audio signal features, to get a robust indication on speech presence or absence, exploiting classification algorithms. Commonly used approaches for these algorithms are time-based or frequency-based, typically exploiting non-overlapping audio signal frames with duration between 10 ms and 20 ms [1]. These different techniques have to be compared in terms of the trade-off between detection accuracy and complexity/power consumption.

The Zero-Crossing (ZC) method [2] exploits signal frequency features to build up a VAD decision rule, assuming that voice components are located at low frequencies, whereas noise components at high frequencies. This method is based on the detection of the number of sign changes in the audio signal amplitude during the analyzed frame. If the number of zero crossings is low, the segment analyzed is classified as voice, whereas if it is high it is classified as noise. This algorithm is easy to implement, but features low detection accuracy with non-stationary noise.

The most popular and widely used techniques in speech detection are energy-based [3]–[6], since they require relatively low computational complexity. They exploit the comparison of the energy carried by the incoming signal with a threshold value. This method is based on the hypothesis that the energy of voiced speech segments is higher compared to unvoiced segments and voiced speech segments have most of their energy at low frequencies. The use of a fixed threshold is the main limitation of this technique.

Marco Croce, Piero Malcovati are with Department of Electrical, Computer, and Biomedical Engineering, University of Pavia, Pavia, Italy (marco.croce02@universitadipavia.it, piero.malcovati@unipv.it).

Brian Friend, Francesco Nesta, Lorenzo Crespi are with Synaptics, Irvine, CA, USA (brian.friend@synaptics.com, francesco.nesta@synaptics.com, lorenzo.crespi@synaptics.com).

Andrea Baschirotto is with Department of Physics "G. Occhialini", University of Molano-Bicocca, Milano, Italy (andrea.baschirotto@unimib.it).

The Single-Frequency Filtering (SFF) method [7], [8] is based on the assumption that noise energy is equally distributed over frequency, while speech energy has a non-uniform distribution. Therefore, the Signal-to-Noise Ratio (SNR) of the speech signal is higher in certain frequency bands. SNR variations due to non-stationary noise can be compensated by weighting the signal. The main drawback of this method is the increase of the computational complexity and, hence, of the power consumption.

Neural Networks (NN) in general can be defined as structures built to emulate human brain activity. The input signal fed to the network goes through different layers that emulate neural connections. After an input layer, there are a certain number of hidden layers, depending on the required complexity and precision of the prediction, and an output layer, which provides the desired result. An important and desirable characteristic of NN for VAD applications [9]–[12] is the ability to classify unstructured data based on their features in the frequency or time domain. Such sophisticated predictive models require a relatively high complexity and significant area.

Most of the above mentioned VAD algorithms are implemented in the digital domain, exploiting the main A/D Audio Signal Processing Chain (ASPC), whose block diagram is shown in Fig. 1, with full performance and large power consumption [13], [14]. Alternatively, a dedicated low-power ASPC with reduced performance in parallel to the main one can be used [15], turning on the main ASPC only upon audio signal detection. Since the main ASPC is operating only in the presence of voice, in this case, the power consumption is reduced but still significant. The proposed fully-analog implementation of the VAD algorithm (without any ADC) operates in parallel to the main ASPC, activating the main ASPC only when required, as shown in Fig. 2, thus further reducing the power consumption.

This paper presents the implementation of a fully-analog VAD circuit. The algorithm exploits the time-variant energy of the incoming audio signal as physical property for the decision rule, under the assumption that speech is more non-stationary than ambient noise [16]. The signal processing includes a Programmable-Gain Amplifier (PGA), a squarer, an integrator, a SC-based signal averaging circuit, and a periodic threshold update circuit for adaptability. The device is fabricated in a 180 nm CMOS technology and occupies an area of 0.14 mm^2. A classification accuracy in domestic environment in presence of loud ambient noise as large as 99.5 % is achieved, consuming 760 nW from a 1.2 V supply.

The paper is organized as follows. Section II presents the adopted VAD algorithm, Section III deals with the circuit implementation, and Section IV reports the experimental results. Section V concludes the paper.

II. VAD Algorithm

A commonly used feature to detect voice activity, is based on the extraction of the energy $E(i)$ of the input signal $y(t)$ in

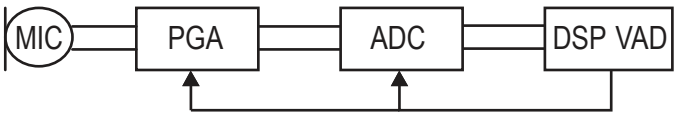

Fig. 1. Block diagram of the audio signal processing chain

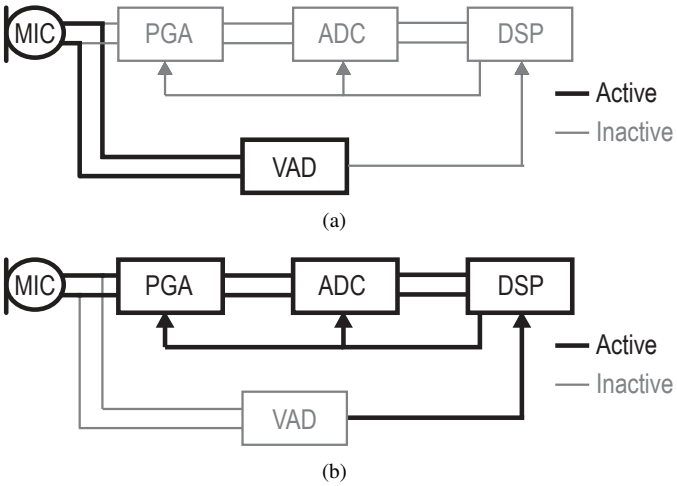

Fig. 2. Block diagram of the proposed VAD system in the absence of voice (a) and after voice is detected (b)

a certain time frame with duration T_{INT} (nominally $T_{INT} = 16$ ms), according to:

$$E(i) = \frac{1}{T_{INT}} \int_{(i-1)T_{INT}}^{iT_{INT}} |y(t)|^2 \, dt. \tag{1}$$

Considering a long period of time, it can be assumed that most of the input signal stream is composed by frames without voice, containing only background noise. The energy evaluated for each frame can then be considered as an estimation of the environment background noise, that in the following is defined as *Noise Level* (NL). The value of $NL(i)$ is updated in every frame (i) with a fraction of the total signal energy of the frame $E(i)$: for $E(i) > NL(i-1)$, $NL(i)$ is updated as

$$NL(i) = \beta_1 NL(i-1) + (1-\beta_1)\, E(i), \tag{2}$$

while for $E(i) \leq NL(i-1)$, $NL(i)$ is updated as

$$NL(i) = \beta_2 NL(i-1) + (1-\beta_2)\, E(i). \tag{3}$$

Parameters β_1 and β_2 range from 0.95 to 0.995 to optimize VAD operation. Therefore, $NL(i)$ is slowly following the instantaneous values of $E(i)$, to avoid sharp variations due to sudden audio signal changes. Furthermore, by tuning β_1 and β_2, the estimated noise level $NL(i)$ can be biased towards tracking its short-term maximum, in order to account for noise non-stationarity and produce less false detections.

Moreover, $E(i)$ is used at the end of each time frame for detecting voice presence, by evaluating $SNR(i)$ that is defined as:

$$SNR(i) = \frac{E(i) - NL(i)}{NL(i)}. \tag{4}$$

When $SNR(i) > TH_{SP}$, (with TH_{SP} ranging from 0.1 to 5), voice is assumed to be present in the frame and the VAD signal

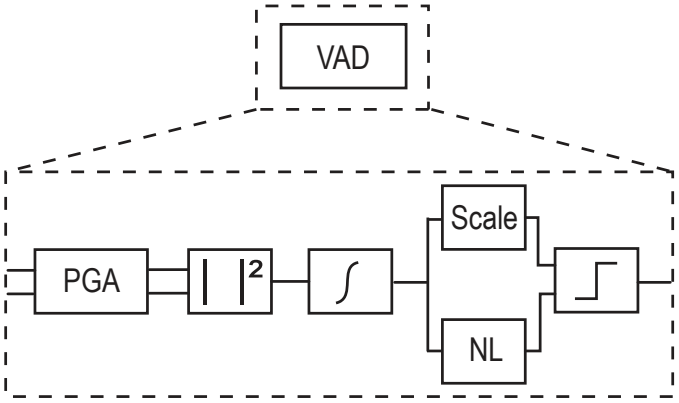

Fig. 3. Detailed block diagram of the proposed analog VAD circuit

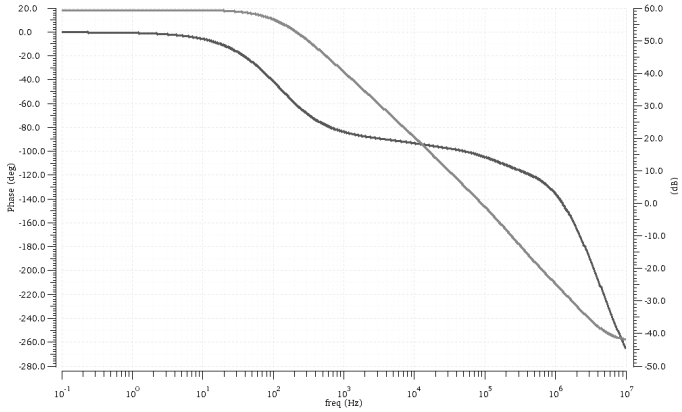

Fig. 4. PGA opne loop gain and phase of the proposed analog VAD circuit.

is generated. With this solution, the voice signal $E(i) - NL(i)$, if present, is detected even when the background noise $NL(i)$ is quite large (e. g. with air conditioning fan noise, road noise, etc.). Instead of evaluating $SNR(i) > TH_{SP}$, a simpler circuit implementation can be achieved by computing

$$\begin{aligned} &\frac{E(i) - NL(i)}{NL(i)} > TH_{SP} \\ &E(i) > NL(i)\,(1 + TH_{SP}) \\ &\frac{E(i)}{1 + TH_{SP}} > NL(i) \end{aligned} \tag{5}$$

and, hence, the scaled energy $E_{SC}(i)$ is defined as:

$$E_{SC}(i) = \gamma \cdot E(i) > NL(i), \tag{6}$$

with the scaling factor

$$\gamma = \frac{1}{1 + TH_{SP}}, \tag{7}$$

allowing the implementation of a simple averaging circuit, since the quantity γ is always lower than 1, as shown in detail in Section III.

In the proposed analog VAD implementation, whose block diagram is shown in Fig. 3, the microphone signal is band-pass filtered and amplified by the PGA, squared, and integrated for the desired time frame to obtain the signal energy $E(i)$, used to update the noise level $NL(i)$ (NL block) and to produce the scaled energy value $E_{SC}(i)$. The achieved values of $NL(i)$ and $E_{SC}(i)$ are compared to eventually generate the VAD signal.

TABLE I
PGA GAIN CONFIGURATIONS

A [dB]	C_I [pF]	C_F [pF]	R_I [MΩ]	R_F [MΩ]	$R_{F,SW}$ [MΩ]
12	16	4	1.46	2.18	133.8
6	16	8	1.46	1.09	66.9
0	8	8	2.93	1.09	66.9
−6	8	16	2.93	0.55	33.4
−12	4	16	5.85	0.55	33.4

III. VAD Circuit Implementation

The energy-efficient circuit implementation of the proposed VAD algorithm, whose circuit schematic is shown in Fig. 5, combines continuous-time non-linear operation and fully-passive switched-capacitor processing to minimize the power consumption.

A. Programmable Gain Amplifier

The first block of the VAD system is a PGA, whose programmable gain accommodates different microphone signal levels. The PGA is implemented with an Active-RC topology embedding a two-stage fully-differential Miller-compensated operational amplifier, whose schematic is shown in Fig. 7. A common-mode feedback circuit is used to set the common-mode output voltage $V_{b,cm}$ to the value required for properly biasing the following stage, with the aid of two 10 MΩ resistors implemented with long channel NMOS transistors connected between the amplifier outputs ($V_{o,n}$ and $V_{o,p}$) and $V_{b,cm}$ (six series-connected NMOS transistors with $W = 0.4$ µm, $L = 40$ µm and the gate connected to the power supply voltage). The amplifier, with an open-loop dc gain of 58 dB, a unity gain bandwidth of 100 kHz (Fig. 4), and an input-referred noise of 128 μV_{rms}, has been designed with a total current consumption of 350 nA, including the common-mode feedback circuit.

In a preliminary study [17] the possibility of integrating a low-frequency ac coupling capacitor was investigated, to deal with the possible common-mode voltage differences between microphone and amplifier. To achieve a flat frequency response at 20 Hz a large integrated resistance in the GΩ range is then required, even with an integrated capacitance as large as 80 pF. Two solutions were compared for implementing such a large resistance. The first one uses a feedback transistor with gate, source, and bulk connected together, the second one adopts a feedback switched resistor [18]. The frequency band of interest for the proposed VAD system is between 0.3 kHz and 6.8 kHz ([19]), in order to capture most of the speech power, while minimizing high-frequency noise. This relaxes the impedance requirement for the resistor from GΩ to hundreds of MΩ. Still, with the aim of reducing area consumption, the switched resistor topology is preferred, since the equivalent impedance variation over process, voltage and temperature (PVT) of the transistor-based implementation would be extremely wide [17]. The different gain configurations available in the PGA are

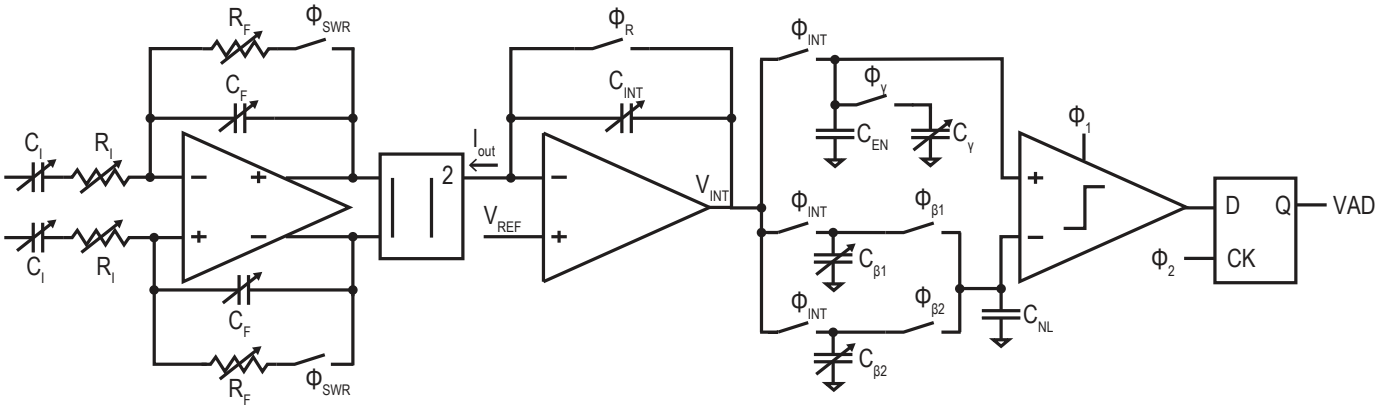

Fig. 5. Schematic of the proposed analog VAD circuit

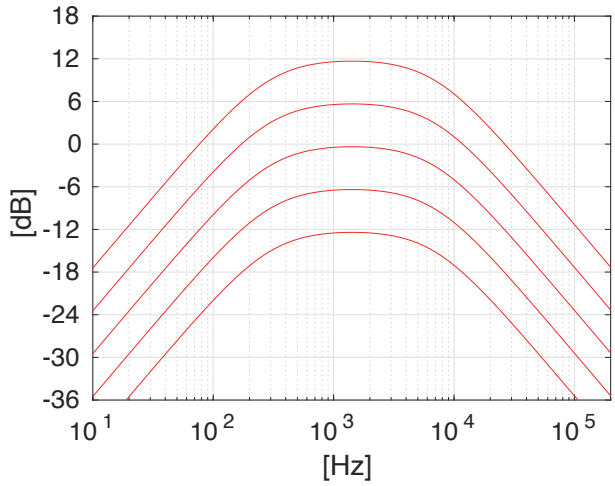

Fig. 6. PGA transfer function of the proposed analog VAD circuit

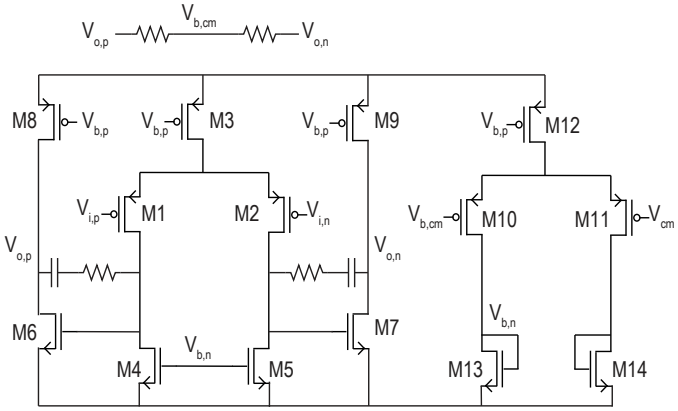

Fig. 7. Schematic of the operational amplifier used in the PGA

summarized in Table I. The maximum gain of 12 dB has been selected considering that in the VAD system the integrator is also amplifying the signal. Therefore, considering that the minimum integrator gain is constraint by the maximum allowed feedback capacitance value (limited by area), the overall gain has been optimized to achieve the maximum output range at the output of the integrator (constraint by power supply). The PGA gain is set by the ratio $A = C_I/C_F$, while the high-pass pole frequency is given by $f_{hp} = 1/\left(2\pi C_F R_{F,SW}\right)$ and the low-pass pole frequency by $f_{lp} = 1/\left(2\pi C_I R_I\right)$, as can be observed in Fig. 6. The total capacitance and resistance values per-branch are 32 pF and 6.4 MΩ, respectively. The equivalent resistance value $R_{F,SW}$ provided by the switching resistor is given by

$$R_{F,SW} = R_F \frac{S_C + S_O}{S_C}, \tag{8}$$

where S_C and S_O are the clock phases in which the switch SW is closed and open, respectively, as shown in Fig. 8.

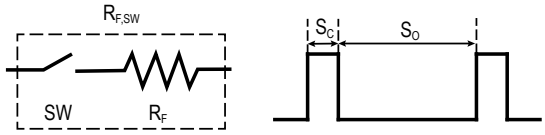

Fig. 8. Switched resistor schematic and time diagram

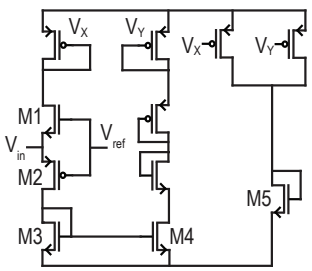

Fig. 9. Basic principle of the circuit for implementing the square operation [20]

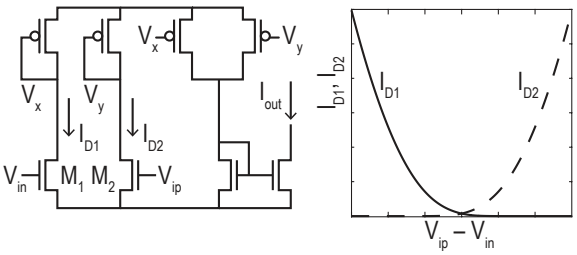

Fig. 10. Schematic and transfer characteristic of the circuit performing the square of the audio signal

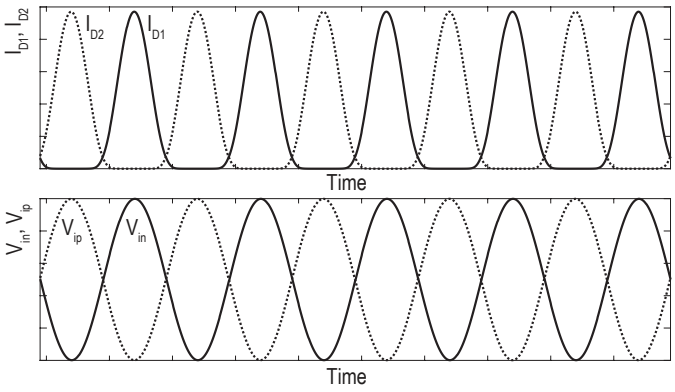

Fig. 11. Time response of the circuit performing the square of the audio signal

B. Energy Evaluation

The PGA differential output signal is connected to the second block of the VAD chain, which is devoted to compute the energy of the audio signal. First the square of the signal is evaluated, exploiting the quadratic relation between voltage and current in a MOS transistor

$$I_D = k\left(|V_G| - |V_S| - |V_{th}|\right)^2, \tag{9}$$

where I_D is the drain current, k represents a constant related to mobility, oxide capacitance and transistor dimensions, V_G, V_S and V_{th} are the gate, source and threshold voltages, respectively. The signal to be squared can be applied to the source or to the gate terminal, keeping the gate or the source terminal at a constant voltage which provides the correct biasing.

The basic principle of this circuit, originally proposed in [20], is illustrated in Fig. 9. The gate voltage of the input transistors is fixed at voltage V_{ref}, while the input signal V_{in} is applied to the source terminals of M_1 and M_2. If $V_{in} > V_{ref}$ M_2 provides the square of V_{in} to M_5 through the current mirrors, while, if $V_{in} < V_{ref}$, the square of V_{in} is computed by M_1 and the resulting current is mirrored to the output transistor M_5.

The solution adopted in the proposed VAD circuit, whose schematic is shown in Fig. 10, is based on the same principle. The input signals V_{ip} and V_{in} are biased by the PGA at the proper common-mode voltage to guarantee the desired quiescent current value for I_{D1} and I_{D2} (20 nA), thus achieving the desired transfer characteristic, also shown in Fig. 10. Fig. 11 illustrates the response of the circuit to a sinusoidal input signal. In order to achieve low power consumption the input transistors M_1 and M_2 have been designed with large L and small W.

The squared signal has then to be integrated to compute the energy. To this end, the output current of the squarer circuit, $I_{out} = I_{D1} + I_{D2}$, is fed to a resettable integrator with feedback capacitance C_{INT} (Fig. 5), thus achieving

$$V_{INT} = \int_{t_i}^{t_f} \frac{I_{out}}{C_{INT}} dt, \tag{10}$$

where V_{INT} is the integrator output voltage, which represents the signal energy $E(i)$ in the time frame $T_{INT} = t_f - t_i$. Both the integrator capacitance C_{INT} and time frame T_{INT} are programmable. The possible values of C_{INT} are 10 pF, 20 pF and 40 pF, while T_{INT} can be 8 ms, 16 ms and 32 ms. When a sinusoidal input signal at 1 kHz is applied to the squarer input, an output current I_{out} is obtained, always flowing out of the integrator virtual ground, of 1 nA, which, with $C_{INT} = 20$ pF and $T_{INT} = 16$ ms, leads to $V_{INT} = 400$ mV. To accommodate this voltage swing at the output of the integrator, also considering PVT variations, the input common-mode voltage of the integrator, V_{REF}, is fixed to 0.3 V. Simulations across corners and temperature have highlighted a variation of V_{REF} from 0.203 V to 0.313 V and a squarer current consumption variation from 15.76 nA to 20 nA, with negligible effects on the overall performance.

C. Energy Averaging

The signal energy $E(i)$ obtained at the integrator output has to be processed to generate the VAD signal. To this end, the energy signal is processed by two parallel paths: one devoted to the evaluation of the noise level (NL) and the other dedicated to compute the scaled energy signal $E_{SC}(i)$. The quantities involved in the adopted VAD algorithm are:

- $E(i)$ is the output of the integrator V_{INT} and represents the signal energy in the current integration period i with duration T_{INT};
- $NL(i)$ represents the noise level in the current integration period i.

Fig. 12(a) and Fig. 13 show the circuit used to update the NL value and the corresponding clock phases, respectively. During the integration time, the switches driven by Φ_{INT} are closed, storing the energy value $E(i)$ on capacitances $C_{\beta 1}$ and $C_{\beta 2}$, while the switches driven by $\Phi_{\beta 1}$ and $\Phi_{\beta 2}$ are open and, hence, C_{NL} holds the noise level value from the previous integration period, $NL(i-1)$. In this configuration the charge stored on the capacitors is

$$\begin{aligned} Q_{NL}(i-1) &= NL(i-1)C_{NL} \\ Q_{\beta 1,2}(i) &= E(i)C_{\beta 1,2} \end{aligned}, \tag{11}$$

where $C_{\beta 1}$ and $C_{\beta 2}$ are programmable from 50 fF to 500 fF with steps of 50 fF, while C_{NL} has a fixed value of 10 pF.

At the end of the integration period, the switches driven by Φ_{INT} are opened and either $\Phi_{\beta 1}$ and $\Phi_{\beta 2}$ is activated ($\Phi_{\beta 1,2}$), based on the values of $NL(i-1)$ and $E(i)$, closing

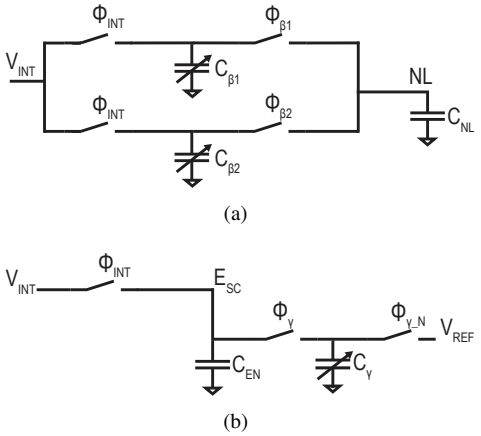

Fig. 12. Schematic of the circuits performing noise level update (a) and energy scaling (b)

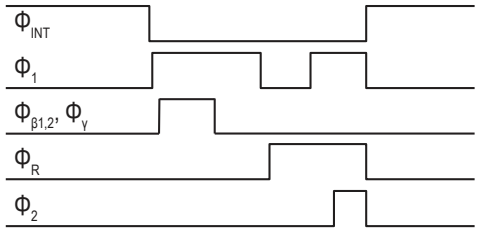

Fig. 13. Timing diagram of the clock phases

the corresponding switch. If $E(i) \geq NL(i-1)$, phase $\Phi_{\beta 1}$ is activated, connecting C_{NL} in parallel to $C_{\beta 1}$, while, if $E(i) < NL(i-1)$, phase $\Phi_{\beta 2}$ is activated, connecting C_{NL} in parallel to $C_{\beta 2}$. The charge stored on the capacitors, therefore, becomes

$$Q_{NL}(i) = NL(i)\left(C_{NL} + C_{\beta 1,2}\right). \tag{12}$$

Combining (11) and (12) the expression of the $NL(i)$ is obtained as

$$NL(i) = NL(i-1)\frac{C_{NL}}{C_{NL}+C_{\beta 1,2}} + E(i)\frac{C_{\beta 1,2}}{C_{NL}+C_{\beta 1,2}}. \tag{13}$$

Coefficients β_1 and β_2 from (2) and (3) are obtained through the capacitive divider $C_{NL}/\left(C_{NL}+C_{\beta 1,2}\right)$, with the programmability range and step required by the VAD algorithm.

The updated noise level value is then compared with the scaled energy $E_{SC}(i)$, given by (6), to determine the presence or absence of speech in the processed frame. Fig. 12(b) and Fig. 13 show the circuit used to compute $E_{SC}(i)$ and the corresponding clock phases, respectively. Phase Φ_1 represents the clock signal of the comparator, while Φ_2 is the phase in which the VAD decision is sampled, as highlighted in Fig. 5. During the integration time, the switches driven by Φ_{INT} and Φ_{γ_N} are closed, while the switch driven by Φ_γ is open. The charge stored on the capacitors is, therefore,

$$\begin{aligned} Q_{EN}(i) &= E(i)C_{EN} \\ Q_\gamma(i) &= V_{REF}C_\gamma \end{aligned}, \tag{14}$$

where the value of C_{EN} is fixed to 5 pF, while the value of C_γ is programmable from 0.5 pF to 24.4 pF. The present value of

TABLE II
POSSIBLE VALUES OF C_γ AND CORRESPONDING VALUES OF γ

C_γ [pF]	γ [%]	C_γ [pF]	γ [%]	C_γ [pF]	γ [%]
0.50	91	2.54	66	7.00	42
0.69	88	2.91	63	7.96	39
0.89	85	3.31	60	9.08	35
1.12	82	3.76	57	10.4	32
1.36	79	4.26	54	12.1	29
1.62	76	4.82	51	14.1	26
1.90	72	5.45	48	16.6	23
2.20	69	6.17	45	19.9	20
				24.4	17

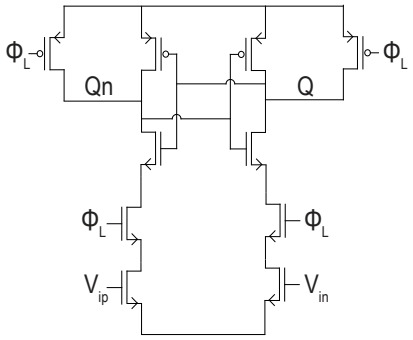

Fig. 14. Comparator schematic

$E(i)$ is compared with $NL(i-1)$ on the falling edge of Φ_{INT}. After the comparison, the switch driven by Φ_{γ_N} is opened and the switch driven by Φ_γ is closed. The charge stored on the capacitors, therefore, becomes

$$Q_{SC}(i) = E_{SC}(i)\left(C_{EN} + C_\gamma\right). \tag{15}$$

Combining (14) and (15), we obtain

$$E_{SC} = E(i)\frac{C_{EN}}{C_{EN}+C_\gamma} + V_{REF}\frac{C_\gamma}{C_{EN}+C_\gamma}. \tag{16}$$

Considering that $E(i) = V_{INT} = V_{REF} + E_{EFF}(i)$, where $E_{EFF}(i)$ is the actual energy value, (16) can be rewritten as

$$E_{SC} = E_{EFF}(i)\frac{C_{EN}}{C_{EN}+C_\gamma} + V_{REF}, \tag{17}$$

where the capacitive divider $C_{EN}/\left(C_{EN}+C_\gamma\right)$ implements the coefficient γ of the VAD algorithm and $E_{SC} = \gamma E_{EFF}(i)$ represents the scaled energy used to detect the presence of voice activity in the incoming signal. Capacitor C_γ is programmable with the values reported in Table II.

The comparator, whose schematic is shown in Fig. 14, is implemented with a dynamic latch topology to reduce the power consumption, since it is activated only for two comparisons per integration period. The comparator inputs V_{in} and V_{ip} are connected directly to C_{EN} and C_{NL}, while its outputs are reset to the supply voltage during the integration time and between the two comparisons. Considering an integration period of 16 ms, the average current consumption of the comparator is almost 0.2 nA. Over 100 Montecralo simulations, the comparator features an offset with mean value of 361 µV and standard deviation of 1 mV.

TABLE III
FEATURE AND PERFORMANCE SUMMARY OF THE PROPOSED VAD CIRCUIT AND COMPARISON WITH THE STATE OF THE ART

Parameter	**This Work**	**Yang, ISSCC 2018 [21]**	**Price, ISSCC 2017 [13]**	**Badami, ISSCC 2015 [15]**	**Raychowdhury, JSSC 2013 [14]**
Technology [nm]	180	180	65	90	32
Input Device	Passive Mic	Passive Mic	Digital Sound	Passive Mic	Digital Sound
Feature	Analog	Analog	Digital	Analog	Digital
Bandwidth [kHz]	0.3 – 6.8	0.1 – 5	N/A	0.075 – 5	11 – 62 000
Classifier	Analog SNR-Based Decision Rule	Digital Binarized Deep Neural Network	Digital Fixed-Point Deep Neural Network	Mixed-Signal Decision Tree	Digital Energy-Based Decision Rule
Power [µW]	0.76	1	8.5	6	300
Area [mm^2]	0.14	2.52	2.08	3	N/A
Rate [1/s]	31.25	100	100	N/A	32 600

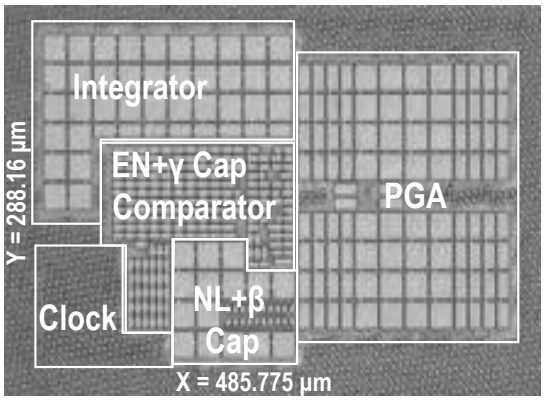

Fig. 15. Chip micrograph of the proposed VAD circuit

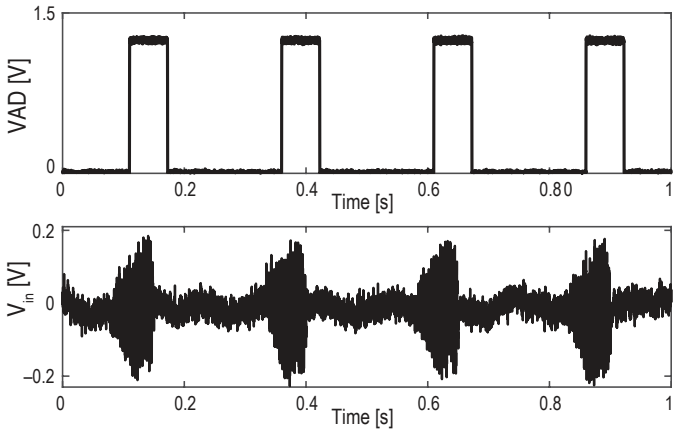

Fig. 16. Example of measured VAD signal

D. Clock Phases

All control signals (Fig. 13) are generated internally starting from a 3 MHz master clock. The master clock frequency is initially divided by a factor of 30 to obtain the 100 kHz signal required, together with the 3 MHz clock, to drive the switched resistor in the PGA ($S_C + S_O = 10\,\mu s$, while the minimum value of S_C is 333 ns), then it is divided down to 31.25 Hz to implement the 8 ms, 16 ms and 32 ms values of the integration period. During the integration period the switches driven by Φ_{INT} are closed to store $E(i)$ on $C_{\beta 1,2}$ and C_{EN}. During the evaluation phase, carried out after each integration period, two comparison are performed (phase Φ_1). After the first comparison and before the second comparison, control signals Φ_γ and $\Phi_{\beta 1}$ or $\Phi_{\beta 2}$ are activated, to compute $NL(i)$ and $E_{SC}(i)$, determining, with the second comparison, the presence or absence of voice in the evaluated frame. The result of the second comparison is sampled on the rising edge of phase Φ_2 and held until the next evaluation is performed.

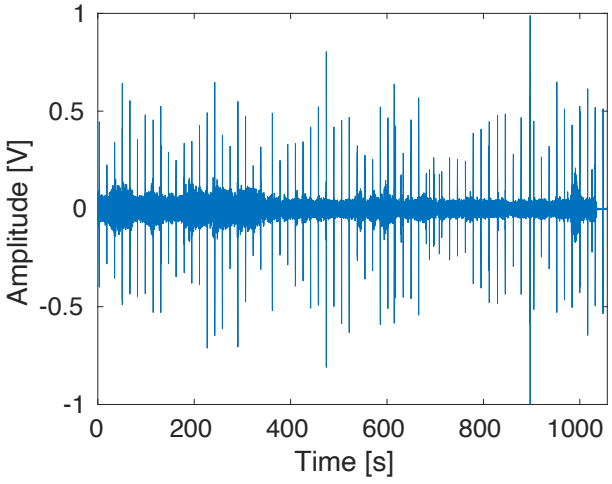

Fig. 17. Signal used for measurement

IV. EXPERIMENTAL RESULTS

The proposed VAD circuit has been implemented in a 180 nm CMOS technology with a die size of 0.14 mm^2, that is, thanks to the adopted choices, more than 10× smaller than any competitor. A micrograph of the chip is shown in Fig. 15.

Fig. 16 illustrates the VAD circuit operation for a 1 s input stimulus with a voice frame repeated 4 times and overlapped with noise.

The circuit has been fully characterized by applying at the input of the PGA the 18 min recorded audio signal shown in Fig. 17 (64 625 frames), containing different types of noise (microwave oven noise, air conditioning noise, office noise and steady background noise) and voice sources, for a total of 580 voice activity events with different SNR values. The scope of this circuit is to wake-up the main ASPC and the subsequent

DSP processing blocks in the full system, which will stay awake for a sufficient amount of time, e. g. 3 s to 5 s. Thus, the rate of correct detections when a new word is spoken is much more important that the frame-by-frame VAD performance for an entire utterance. As consequence, the VAD accuracy is evaluated only considering the initial on-set of speech rather than the full utterance (e.g. by considering approximately the first 100 ms). Performing significant tests of the proposed VAD circuit with a standard dataset would require having in place also the main ASPC, which is not available yet. A typical Detection Error Tradeoff (DET) operating curve for different values of the configuration parameters is shown in Fig. 18. With this input signal, the proposed VAD circuit achieves a classification accuracy of 99.5 % (0.5 % total errors) with 0.2 % of false positive (FP) errors and 0.3 % of false negative (FN) errors. The parameters used for the measurement are: $A = 12\,\text{dB}$, $T_{INT} = 16\,\text{ms}$, $C_{INT} = 40\,\text{pF}$, $\beta_1 = 0.95$, $\beta_2 = 0.98$ and $\gamma = 0.77$.

The total errors, as well as the FP and FN percentages obtained from measurements and simulations for different values of parameters β_1, β_2, A and C_{INT} are summarized in Table IV, while the total error percentage as a function of parameter γ is shown in Fig. 19. The achieved classification accuracy values demonstrate that the proposed VAD circuit is robust and insensitive to parameter variations. In the presence of supply noise ($50\,\text{mV}_\text{p}$ sinewave at 1 kHz and $150\,\text{mV}_\text{p}$ squarewave at 217 Hz) the VAD classification accuracy is basically unaffected.

The proposed VAD circuit consumes an average current of 633 nA from a 1.2 V power supply (760 nW). The breakdown of the current consumption contributions of the different blocks of the VAD circuit is reported in Table V. Table III summarizes and compares features and performances of the proposed analog VAD circuit with the state-of-the-art.

V. Conclusions

In this paper a fully analog voice-activity detection circuit is presented. The device achieves 99.5 % classification accuracy in a domestic environment in the presence of loud ambient noise. The circuit, exploiting an energy-efficient analog implementation with continuous-time non-linear operation and fully-passive switched-capacitor processing, consumes only 760 nW from a 1.2 V. The VAD circuit prototype, fabricated in a 180 nm CMOS technology, occupies $0.14\,\text{mm}^2$.

References

[1] Z. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798–807, 2010.

[2] T. H. Zaw and N. War, "The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection," in *Proceedings of IEEE International Conference of Computer and Information Technology (ICCIT)*, 2017, pp. 1–5.

[3] N. N. Lokhande, N. S. Nehe, and P. S. Vikhe, "Voice activity detection algorithm for speech recognition applications," in *IJCA Proceedings on International Conference in Computational Intelligence (ICCIA)*, 2012, pp. 1–4.

[4] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, "VAD techniques for real-time speech transmission on the internet," in *Proceedings of IEEE International Conference on High Speed Networks and Multimedia Communication (ICHSNMC)*, 2002, pp. 46–50.

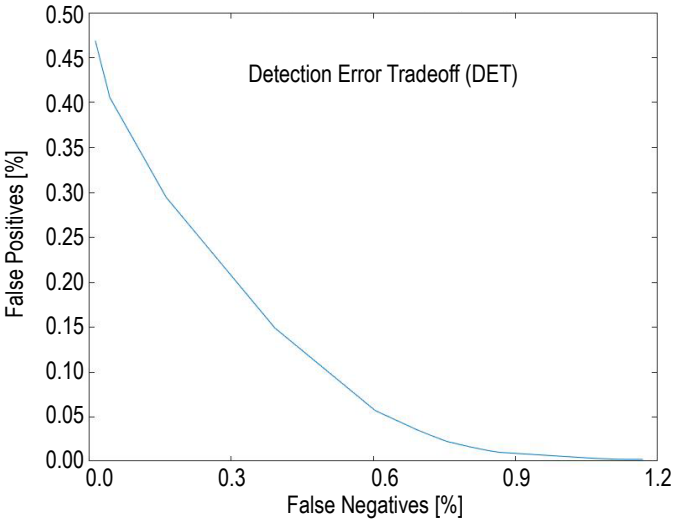

Fig. 18. Detection error tradeoff curve of the proposed VAD circuit

TABLE IV
Simulated and measured classification accuracy of the proposed VAD circuit with different values of parameters β_1, β_2, A and C_{INT}

$\beta_1 = 0.95$, $\beta_2 = 0.98$ $C_{INT} = 40\,\text{pF}$, $A = 12\,\text{dB}$			
Accuracy	**Total Errors [%]**	**FP [%]**	**FN [%]**
Measurement	0.4505	0.1528	0.2977
Simulation	0.2615	0.0712	0.1903
$\beta_1 = 0.99$, $\beta_2 = 0.95$ $C_{INT} = 40\,\text{pF}$, $A = 12\,\text{dB}$			
Accuracy	**Total Errors [%]**	**FP [%]**	**FN [%]**
Measurement	0.6415	0.4519	0.1896
Simulation	0.2615	0.0712	0.1903
$\beta_1 = 0.99$, $\beta_2 = 0.95$ $C_{INT} = 20\,\text{pF}$, $A = 6\,\text{dB}$			
Accuracy	**Total Errors [%]**	**FP [%]**	**FN [%]**
Measurement	0.7141	0.4104	0.3037
Simulation	0.2615	0.0712	0.1903
$\beta_1 = 0.99$, $\beta_2 = 0.95$ $C_{INT} = 10\,\text{pF}$, $A = 0\,\text{dB}$			
Accuracy	**Total Errors [%]**	**FP [%]**	**FN [%]**
Measurement	0.7245	0.3793	0.3452
Simulation	0.2615	0.0712	0.1903

TABLE V
Current consumption breakdown

Block	**Current [nA]**
PGA	350
Squarer	20
Integrator	200
Comparator + Digital	63
Total	**633**

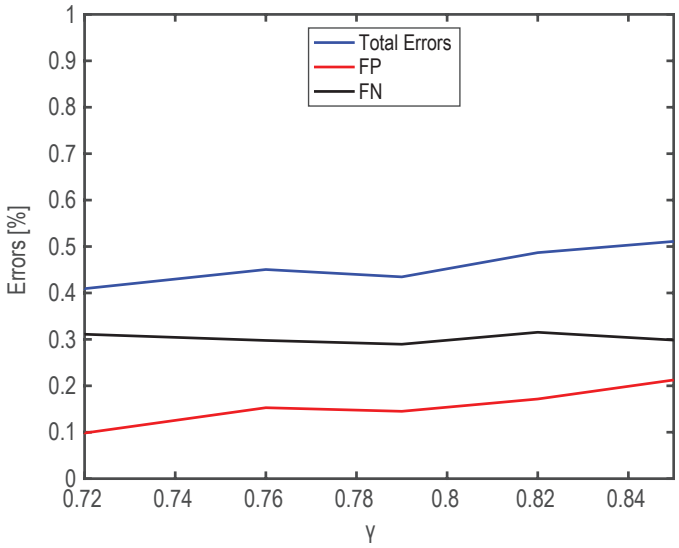

Fig. 19. Classification accuracy of the proposed VAD circuit with different values of parameter γ

[5] H. V. R. Vega, V. Molina, and L. Martinez, "VAD algorithms energy-based and spectral-domain applied in River Plate Castilian," in *Proceedings of IEEE Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*, 2016, pp. 1–5.

[6] X. Wang and L. Qu, "The self-adaptive voice activity detection algorithm based on time-frequency parameters," *The Open Automation and Control Systems Journal*, vol. 6, no. 1, pp. 1661–1668, 2014.

[7] M. T. Adiga and R. Bhandarkar, "Improving single frequency filtering based voice activity detection (VAD) using spectral subtraction based noise cancellation," in *Proceedings of IEEE International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 2016, pp. 18–23.

[8] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.

[9] Y. K. Bharath, S. Veena, K. V. Nagalakshmi, M. Darshan, and R. Nagapadma, "Development of robust VAD schemes for voice operated switch application in aircrafts: Comparison of real-time VAD schemes which are based on linear energy-based detector, fuzzy logic and artificial neural networks," in *Proceedings of IEEE International Conference on Applied and Theoretical Computing and Communication Technology (ICATCCT)*, 2016, pp. 191–195.

[10] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7378–7382.

[11] M. Price, J. Glass, and A. P. Chandrakasan, "A low-power speech recognizer and voice activity detector using deep neural networks," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 66–75, 2017.

[12] X. 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, 2016.

[13] M. Price, J. Glass, and A. P. Chandrakasan, "A scalable speech recognizer with deep-neural-network acoustic models and voice-activated power gating," in *IEEE International Solid-State Circuit Conference Digest of Technical Papers (ISSCC)*, 2017, pp. 244–245.

[14] A. Raychowdhury, C. Tokunaga, W. Beltman, M. Deisher, J. Tschanz, and V. De, "A 2.3 nJ/frame voice activity detector-based audio front-end for context-aware system-on-chip applications in 32 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 8, pp. 1963–1969, 2013.

[15] K. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "Context-aware hierarchical information-sensing in a 6 µW 90 nm CMOS voice activity detector," in *IEEE International Solid-State Circuit Conference Digest of Technical Papers (ISSCC)*, 2015, pp. 430–431.

[16] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[17] M. Croce, C. De Berti, L. Crespi, P. Malcovati, and A. Baschirotto, "MEMS microphone fully-integrated CMOS cap-less preamplifiers," in *Proceedings of IEEE Ph. D. Research in Microelectronics and Electronics (PRIME)*, 2017, pp. 37–40.

[18] H. Chandrakumar and D. Marković, "A 2 µW 40 mV_{pp} linear-input-range chopper-stabilized bio-signal amplifier with boosted input impedance of 300 MΩ and electrode-offset filtering," in *IEEE International Solid-State Circuit Conference Digest of Technical Papers (ISSCC)*, 2016, pp. 96–97.

[19] P. Jax and P. Vary, "Bandwidth extension of speech signals: a catalyst for the introduction of wideband speech coding?" *IEEE Communications Magazine*, vol. 44, no. 5, pp. 106–111, 2006.

[20] K. M. H. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "A 90 nm CMOS, 6 µW power-proportional acoustic sensing frontend for voice activity detection," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 291–302, 2016.

[21] M. Yang, C. H. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "A 1 µW voice activity detector using analog feature extraction and digital deep neural network," in *IEEE International Solid-State Circuit Conference Digest of Technical Papers (ISSCC)*, 2018, pp. 346–348.

Marco Croce received Bachelor of Science (2013) in Mathematical Engineering from Politecnico di Milano, Italy, Master of Science (2015) in Electronic Engineering from University of Pavia and the Ph.D. degree (2018) in Microelectronics from University of Pavia, Italy. His Ph.D. research was in collaboration with Conexant Systems Inc. and then Synaptics based in Irvine, California. He joined AMS AG in 2019 as analog design engineer in Pavia, Italy.

Brian Friend graduated in 1978 in Electrical Engineering at Pratt Institute Brooklyn, New York, USA. From 1978 to 1983 he was a CO-OP Student Intern at Airborne Instruments Labs/Div (1976-1977) Somerset, NJ USA, where he performed data analysis of the results of RF tests, using an HP9820 minicomputer. From 1978 to 1983 he joined General Dynamics Convair, San Diego, California, USA, as Senior Electronic Engineer working on detailed and system design engineering of Cruise Missile flight test Telemetry systems and data acquisition circuits. From 1983 to 1993 he joined TRW/LSI Products Inc, California USA, as Senior Analog Applications Engineer in high-speed Flash A/D and D/A Converters. From 1993 to 2000 he joined Conexant Systems Inc, California USA, as Senior Design EE (1993-1995) Staff Design EE (1995-1997) Manager (1998-2000) in Mixed Signal Design. From 2001 to 2004 he joined Entropic Communications San Diego, CA USA, in the field of Mixed-Signal Design. From 2004 to 2005 he joined Apexone Microelectronics ,San Diego, California USA, as VP Engineering he established a US design center for a Mixed Signal startup company. From 2004 to 2005 he joined Rapid Bridge LLC, San Diego, CA USA, as Director of Analog Design Engineering, design and development of 10-12 bit and 80-100MSPS pipeline ADC technology, several PLLs and DLLs in 65nm TSMC, 90nM TSMC process technologies. From 2007 to 2010 he joined IQ Analog in San Diego, CA USA, as VP of Analog Design working on A/D and D/A Converters. From 2010 to 2020 he joined Conexant Incorporated Irvine, CA USA, as Staff Analog Engineering (2010-2017) and as Senior Staff Analog Engineering (2017-2020). From 2020 he is a Senior Staff Design Engineer at Elevate Semiconductor San Diego, California USA.

Francesco Nesta graduated in Computer Engineering (2005) at Politecnico di Bari, Italy, and received the Ph.D. degree (2018) in ICT school in Information Technologies and Telecommunications, Trento, Italy. From 2006 to 2010 he joined Fondazione Bruno Kessler, Trento, Italy, as a researcher investigating on blind source separation and speech enhancement. From 2013 to 2015 he joined Conexant in Irvine, California USA, as Distinguished DSP engineer in the field of Speech/audio enhancement algorithms design. From 2017 he is a Technical Director Audio DSP/ML R&D engineering at Synaptics Incorporated, Irvine, California USA, in advanced input processing algorithm development for voice applications.

Lorenzo Crespi received the Laurea degree in electrical engineering from the University of Pavia, Pavia, Italy, in 1997. He joined the Mixed-Signal group of Rockwell Semiconductors, Newport Beach, CA, USA, in 1997, where he worked on design of analog circuits for telecommunications. From 2000 to 2017 he was with Conexant Systems, Irvine, CA, USA, working on data converters and analog front-ends with applications to ADSL, WLAN, audio, and imaging products. Since 2007 he has led design of audio and power management circuits. He is currently at Synaptics in Irvine, CA, as Director of Analog Mixed-Signal Design for the IoT division.

Piero Malcovati (M'95–SM'05) graduated in electronic engineering from the University of Pavia, Italy, in 1991. In 1992, he joined the Physical Electronics Laboratory (PEL) at the Federal Institute of Technology in Zurich (ETH Zurich), Switzerland, as a Ph. D. candidate. He received the Ph. D. degree in electrical engineering from ETH Zurich in 1996. From 1996 to 2001 he has been Assistant Professor and from 2002 to 2017 Associate Professor at the Department of Electrical, Computer, and Biomedical Engineering of the University of Pavia. From 2017 Piero Malcovati is Full Professor in the same institution. His research activities are focused on microsensor interface circuits, power electronics circuits, and high-performance data converters. He was and still is member of the Technical Program Committees for several International Conferences, including ISSCC, ESSCIRC, SENSORS, ICECS, and PRIME. He is Associate Editor of the IEEE Journal of Solid-State Circuits. He is an IEEE senior member.

Andrea Baschirotto graduated in Electronic Engineering (summa cum laude) from the University of Pavia in 1989. In 1994, he received the Ph.D. degree in electronics engineering from the University of Pavia. In 1994, he joined the Department of Electronics, University of Pavia, as a Researcher (Assistant Professor). In 1998, he joined the Department of Innovation Engineering, University of Lecce, Italy, as an Associate Professor. In 2007, he joined the Department of Physics, University of Milan-Bicocca, Italy, as an Associate Professor. Andrea Baschirotto has a long-term experience in microelectronics for teaching, researching, and industrial designing. He is teaching regular Academic courses since 1997. He uses to give advanced courses in companies and research center since 1996. He uses to give short courses and tutorial at the most important conferences (ISSCC, ISCAS, PRIME).About his research activity, he founded and directs the Microelectronics Group at University of Milan-Bicocca with collaboration with several companies and research institutions (Infineon, STMicroelectronics, Pirelli, IMEC, Univ. of Pavia, etc. . . .). His main research interests are in the design of CMOS mixed analog/digital integrated circuits, in particular for low-power and/or high-speed signal processing. He participated to several research collaborations, also funded by National and European projects. He is/has been responsible of some National and Regional projects for the design of ASIC.He has authored or co-authored more than 500 papers in international journals and presentations at international conferences, 6 book chapters, and holds almost 50 USA patents. In addition, he has co-authored more than 120 papers within research collaborations on high-energy physics experiments.Andrea Baschirotto has been Associate Editor IEEE Trans. Circuits Syst. – Part II, and of IEEE Trans. Circuits Syst. – Part I. He has been the Technical Program Committee Chairman for ESSCIRC 2002 and he was the Guest Editor for the IEEE JSSC for ESSCIRC 2003 and ESSCIRC2007. He is Associate Editor for the IEEE-JSSC since 2014. He was the General Chair of IEEE-PRIME2006, AACD2008, AACD2013, and IEEE-PRIME2013, AACD2016, and AACD19. He is (has been) the member of the Technical Program Committee of several international conferences (ISSCC, ESSCIRC, AACD, DATE, etc..). He is serving since several years the ESSCIRC TPC as Data Converter Subcommittee Chairman. He has been the secretary of the European Committee of ISSCC Technical Program Committee. He is an IEEE Fellow (2013). He is the founder and the Chairman of the IEEE Solid-State Circuit Society Italian Chapter that received the Best SSCS Chapter Award for 2020.