A 90 nm CMOS, 6 μW Power-Proportional Acoustic Sensing Frontend for Voice Activity Detection

Komail M. H. Badami, *Student Member, IEEE*, Steven Lauwereins, *Student Member, IEEE*, Wannes Meert, and Marian Verhelst, *Senior Member, IEEE*

***Abstract*—This work presents a sub-6 μW acoustic frontend for speech/non-speech classification in a voice activity detection (VAD) in 90 nm CMOS. Power consumption of the VAD system is minimized by architectural design around a new power-proportional sensing paradigm and the use of machine-learning-assisted moderate-precision analog analytics for classification. Power-proportional sensing allows for hierarchical and context-aware scaling of the frontend's power consumption depending on the complexity of the ongoing information extraction, while the use of analog analytics brings increased power efficiency through switching ON/OFF the computation of individual features depending on the features' usefulness in a particular context. The proposed VAD system reduces the power consumption by 10× as compared to state-of-the-art (SotA) systems and yet achieves an 89% average hit rate (HR) for a 12 dB signal-to-acoustic-noise ratio (SANR) in babble context, which is at par with software-based VAD systems.**

***Index Terms*—Acoustic frontend, analog machine learning, context-aware computing, hierarchical computing, scalable low power analog, voice activity detection (VAD).**

I. Introduction

TECHNOLOGICAL innovations are changing the way we interact with electronic devices. Interactions like voice control and gesture recognition are rapidly gaining popularity. Such natural interactive systems do need not only many integrated sensors but also always-awake, reactive sensor frontends. These frontends generate large amounts of raw signals that state-of-the-art (SotA) frontends immediately digitize for processing on a DSP. This very robust approach is not power efficient, as not all raw sensor signals are equally relevant. The net information content of a sensed signal is quite often significantly smaller than the Nyquist rate [1]–[7]. Existing works such as information-rate processing [1], [2], analog to information conversion [3]–[5], and compressed sensing [6], [7] show power savings by extracting or compressing the information from signals before digitizing the data. However, as these schemes operate in a static way, the compression or extraction parameters are set beforehand. Yet, the information content in raw signals and its application relevance dynamically varies depending on the operating context. Operating such systems efficiently thus requires a dynamic system adaptation depending on the context or signal information content. Existing systems do not perform such fine grain adaptive behavior, which severely limits their power savings as shown by solid line in Fig. 1.

We propose a self-scalable, power-proportional sensing paradigm, which gracefully scales the system's power consumption with the amount and complexity of extracted information, i.e., the power consumption for such a system increases only as the task of information extraction gets more complex. To this end, in this paper, we propose key enablers for power-proportionality and apply them to a proof of concept acoustic frontend for voice activity detection (VAD).

VAD systems distinguish speech from non-speech in different background noise contexts for varying signal-to-acoustic-noise ratios (SANR). SotA VAD systems [8]–[10] extract complex features like mel-frequency cepstral coefficients and DCT to differentiate speech from non-speech. The high computational complexity of such features results in large power consumption, typically about 50−100 μW [8]–[11] in addition to the power consumption of the required active microphone. Such a continuous large power consumption is unacceptable for battery powered always-on sensor frontends. This work exploits our new power-proportional sensing paradigm along with moderate-precision, computationally inexpensive, analog feature extraction, coupled with an embedded mixed-signal classifier to save more than 10× power consumption over SotA without compromising on the classification accuracy.

The outline of this paper is as follows. Section II discusses insights into the design principles for power-proportional sensing and explains the rationale behind the analog feature extraction instead of the commonly used digital scheme. Section III describes the architecture and specification set for VAD while the detailed implementation is discussed in Section IV. Measurement results for the chip and for the full VAD system are discussed in Section V.

II. Key Principles for Power-Efficient Sensing

This section details the two key principles that allow our always-on sensing system to scale its power consumption with the information extracted saving 10× power over SotA VAD systems.

A. Power-Proportional Sensing

The core premise for power-proportional sensing is that power consumption of the sensing system scales proportionally

Manuscript received May 04, 2015; revised August 28, 2015; accepted September 23, 2015. Date of publication November 02, 2015; date of current version December 30, 2015. This paper was approved by Guest Editor Yusuke Oike.

K. M. H. Badami, S. Lauwereins, and M. Verhelst are with KU Leuven, Departement Elektrotechniek ESAT-MICAS Kasteelpark, Leuven B-3001, Belgium.

W. Meert is with the Department of Computer Science, KU Leuven, Leuven 3001, Belgium.

Digital Object Identifier 10.1109/JSSC.2015.2487276

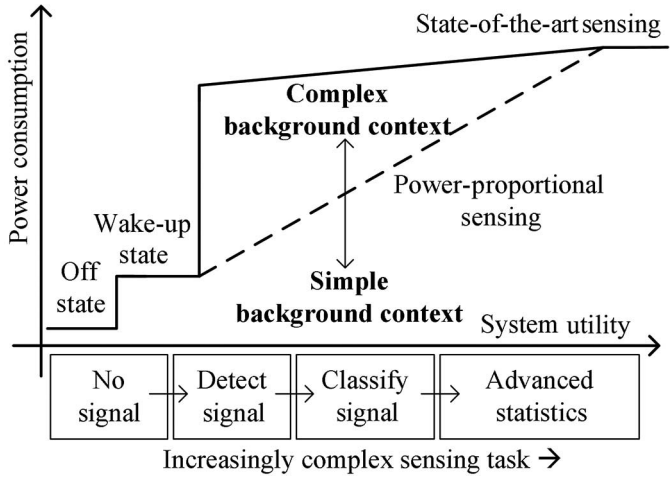

Fig. 1. Power-proportional sensing in contrast with SotA sensing systems.

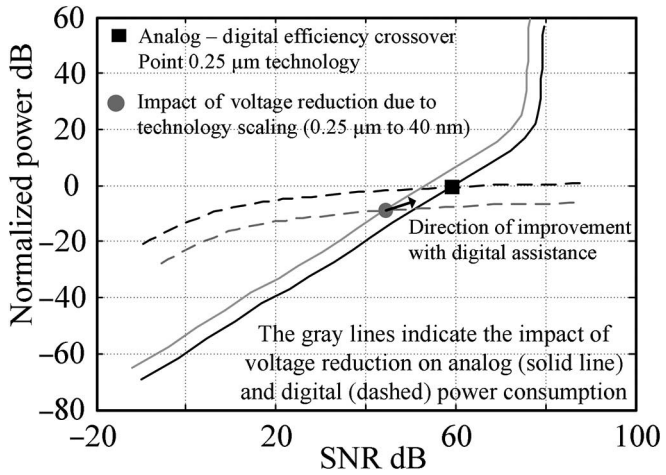

Fig. 2. Computation power scaling for analog (solid line) and digital (dashed line) implementations [12] and impact on efficiency cross over point due to voltage scaling and due to digital assistance by machine learning and/or calibration.

with the complexity of the sensing task. The sensing process with the target of information extraction can increase in complexity along two dimensions.

First, the amount of information extracted from the incoming signal can scale in complexity. Consider, e.g., the task of speaker identification versus speech detection. The former task entails the later as a prerequisite first step, thus justifying the increase in power consumption. Enabling hierarchical operation for tasks of increasing complexity allows scaling of power consumption with complexity of information extraction. In such an architecture, each processing stage extracts more complex information than the previous stage while consuming more power. This enables information extraction by necessity, as is shown in the horizontal axis in Fig. 1.

Second, even if the amount of extracted information remains the same, distinguishing the useful information from the background noise (the context) is subject to varying levels of difficulty. For this case, consider the complexity of speech detection in a quiet office, in contrast to a noisy street environment. The amount of information needed is same in both cases, but in the latter case, as the background noise maps directly onto the information spectrum, it creates in-band interference on the desired signal. As such, distinguishing speech from non-speech becomes more complex, hence justifying the increase in power consumption. Context-awareness enables power-proportional sensing to scale power as the background noise context scales the complexity of information extraction, as shown as bold in Fig. 1. For the example above, context-awareness allows to use a much smaller discriminating feature subset in a low-noise environment and a relatively larger subset for noisy background contexts, hence scaling power.

SotA sensing systems do not exploit the power scaling opportunity offered by the above scenarios, and typically operate constantly in full processing mode. This plateaus the on-state power consumption for SotA sensing systems independent of system utility as shown in Fig. 1.

B. Power Efficiency Through Analog Analytics

The power-proportional sensing paradigm as highlighted in previous paragraph needs complexity and precision-dependent power scalable hardware blocks. Such power scaling with precision is very different for analog and digital implementations. Analog power consumption scales gradually for thermal noise limited system with low-to-medium precision, while digital has a logarithmic power versus precision profile. As it has been shown in [12] and in Fig. 2, for a 0.25 μm CMOS technology, analog computation is not only more power efficient than digital for low-to-medium resolution processing but also exhibits better scalability.

Reduction in supply voltage due to technology scaling allows more power-efficient digital circuits and questions the beneficial analog behavior in advanced technologies. This is because with scaling, the cost of maintaining the same precision in analog increases as a larger bias current is needed to reduce the noise-floor compensating for reduction in signal swing. Assuming that the supply voltage has scaled from 2.5 V for 0.25 μm to 0.9 V for a 40 nm technology, the active digital power has scaled down by $10\log(2.5^2/0.9^2) \sim 9$ dB while analog power consumption goes up by 4.5 dB [12] for subthreshold design. Contrasting effects of reduction in average capacitance per node and increase in subthreshold-leakage on digital power consumption are not considered here. The above discussion implies that while analog keeps its favorable scalability, the analog–digital efficiency crossover point moves toward left by 2 bits. This renders analog computation cheaper than digital for up to 7 bits of precision as shown in Fig. 2.

Digital enhancements, such as machine learning and calibration, can restore some of the lost benefit of analog over digital computation for always-on sensing or classification tasks because these often do not need perfect signal reconstruction but only need error resilient processing such as detection or classification. Specifically, such tasks do not require accurate absolute computations but only relative comparisons of the computed feature values to on-chip trained thresholds, as we will show in the design presented in this paper. Hence, absolute precision requirements for such systems are rather modest, and mismatches and offset impairments are automatically taken care of and by the embedded trained classifier in the loop. As

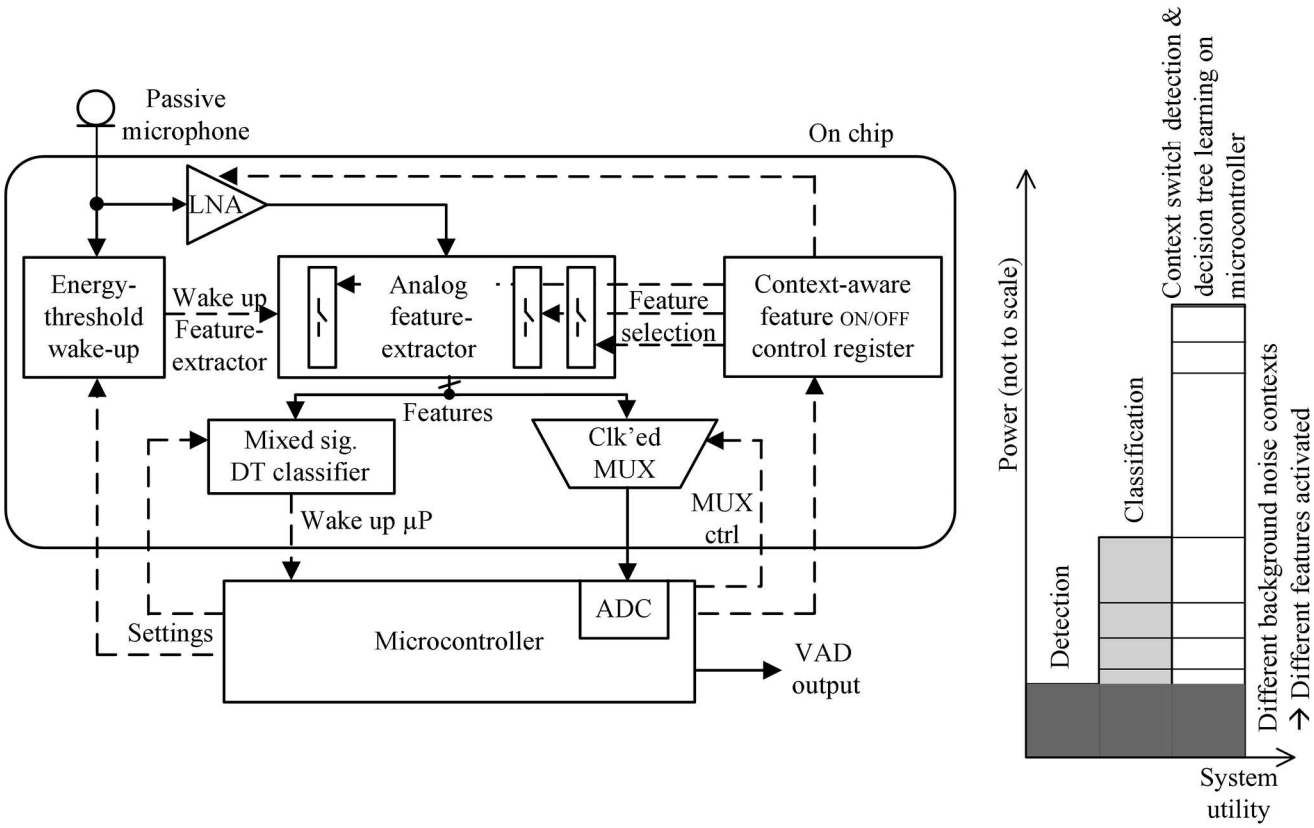

Fig. 3. System diagram of the power-proportional voice activity detector (left) and VAD power scaling with sensing complexity (right).

demonstrated by this work, as well as some existing works, machine learning assisted [13], [14] and/or digital calibration [15] can improve SNR by 6–10 dB for comparable power which pushes the efficiency crossover point in the rightward direction as shown in Fig. 2. These estimations support the use of analog computation for systems requiring scalability up to 8 bits of precision.

III. System Architecture and Specifications

This section highlights the use of the aforementioned key principles in the developed VAD architecture [16] and derives the specifications for the analog/mixed-signal building blocks.

A. VAD System Architecture

The top-level block diagram of the proposed power-proportional VAD system is shown in Fig. 3. The main sub-blocks of the system are the threshold-based wakeup block, the analog feature-extractor, the mixed-signal classifier, and the microcontroller, which operate in the described power-proportional sensing fashion as follows.

An always-awake threshold-based wakeup block keeps checking the passive microphone for sound activity. When any signal—not necessarily useful—is detected, it wakes up the analog feature-extractor that translates the input signal into a set of features. The on-chip classifier uses these computed features to classify the incoming signal as speech/non-speech. If the signal is speech, the classifier wakes up the microcontroller for more advanced processing.

Such hierarchical activation of information extraction hardware allows the VAD system to be in the lowest power-mode possible, while still able to compute the necessary information. This allows scaling the power with necessary information as outlined in Section II-A1. Also, as not all computed analog features carry information under all background noise contexts, machine-learning-based context-awareness allows dynamically disabling the computation of features that do not assist in classification. Such context-aware computing allows further power scaling depending on the number of useful features necessary as explained in Section II-A2. The control of feature activation and classifier configuration is done by the embedded microcontroller. This microcontroller periodically wakes up to check for background noise context changes and upon detecting a change, retrains the classifier, and activates the required features for the new context. As further modeled in Section III-B, considering that the analog feature-extraction blocks are in the loop during this training operation, all static analog impairments such as mismatch, gain errors, or offsets are absorbed in the trained feature thresholds and do not affect the classification accuracy. This justifies the usage of low-precision analog analytics for feature computation, as discussed in Section II-B. Before detailing the design of individual

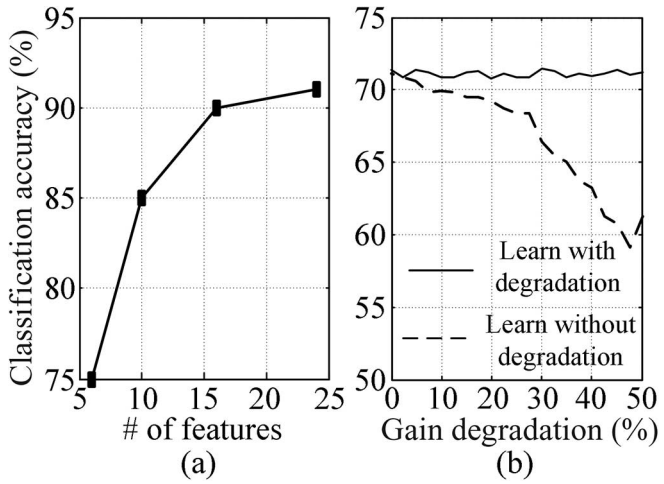

Fig. 4. (a) Impact of number of computed features. (b) Impact of gain degradation on classification accuracy. The results are for exhibition background noise with 12 and 0 dB SANR, respectively.

sub-blocks in Section IV, Section III-B derives specifications for the targeted VAD system.

B. Specifications for VAD System

This section first derives the system-level specifications and then the specifications for individual analog blocks. The system computes an analog feature-set for the acoustic signal by decomposing the signal into different frequency bands and then extracting the average value of the rectified signal in each frequency band. Mathematically, each analog feature af_i is defined as

$$af_i = \overline{\text{abs}\left[Ax(t) * h_i^{\text{BPF}}\right]} \tag{1}$$

where $\text{Ax}(t)$ is the amplified acoustic signal, h_i^{BPF} is the impulse response of bandpass filter (BPF) used to decompose the input signal into a smaller frequency band, abs, $*$, and $\bar{x}$ represent the absolute value, convolution, and averaging, respectively. The features thus represent the average power present in every frequency band. It is, therefore, important to determine the required frequency range, number of observed frequency bands, and the necessary precision, as these parameters will strongly influence the classification accuracy as well as the system's power consumption. Such system specifications are evaluated based on a MATLAB model of the analog feature-extractor of VAD system based on (1).

Along the frequency axis, the bulk of energy for speech and acoustic noise is concentrated in the frequency range 100 Hz to 4 kHz [17]. The MATLAB model varies the number of computed features in the above frequency range by scaling the Q factor of the BPFs. This ensures that the entire frequency range is always populated with filters, with an increasing frequency resolution as the number of computed features increases. The results of the above simulation are shown in Fig. 4(a). It can be seen that more features improve classification accuracy, yet accuracy gains diminish beyond 16 features allowing us to limit our design to a maximum of 16 [individually (dis)activated] features. Further, the model also evaluates the impact of static

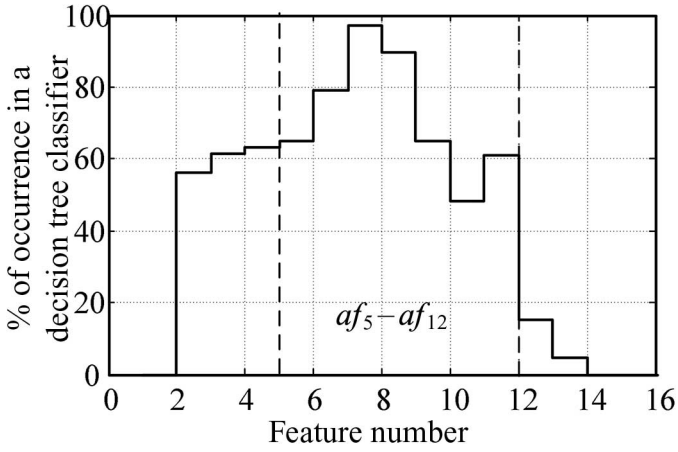

Fig. 5. Histogram depicting average usefulness of computed features in exhibition background noise context for SANR of 0 dB.

TABLE I
HIGHLIGHT OF IMPORTANT SPECIFICATIONS FOR TARGETED VAD SYSTEM

Frequency range	Maximum feature count	SANR	Microphone sensitivity	Output resolution
100 Hz to 5 kHz	16	0 to 12 dB	>−60 dBV	8 bits

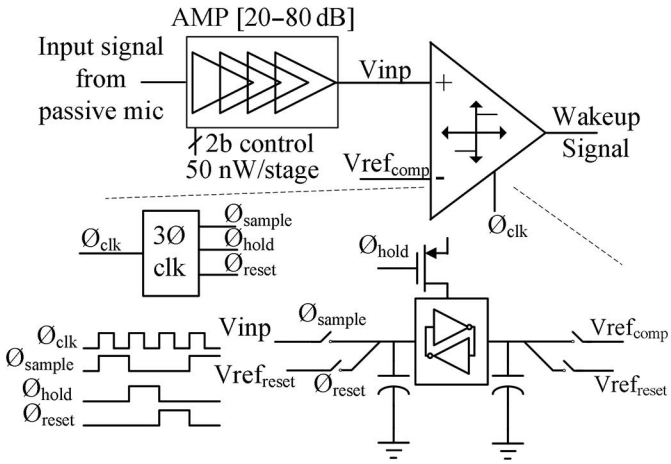

Fig. 6. Schematics for threshold-based wakeup detector.

analog impairments, e.g., by degrading the gain in the signal path, as seen in Fig. 4(b), as long as these occur within the training loop, they are absorbed in the thresholds learnt for classification and thus have no impact on classification accuracy.

Fig. 5 histogram shows the relative relevance of each of the 16 analog features in the speech versus non-speech classification for exhibition noise context with 0 dB SANR. It is clear that the middle-frequency features af_5 to af_{12} are more commonly used. Hence, we only pass these features to an on-chip classifier, while the full feature-set is passed on to a microcontroller only when needed for more complex tasks, such as context-change detection.

Another important group of parameters are the maximum input-referred noise and the necessary gain for the system. The specifications for input-referred noise and gain strongly depend

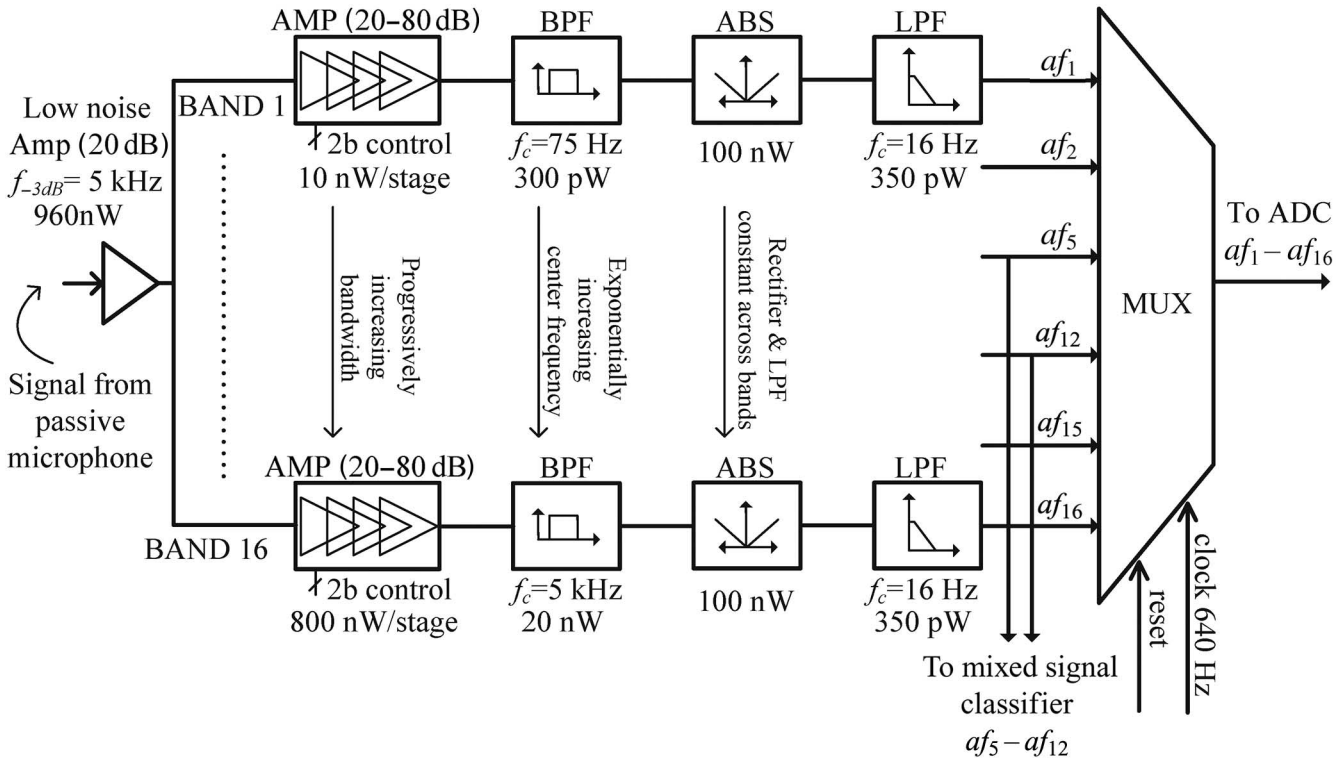

Fig. 7. Schematic and design parameters of the analog feature extraction block.

on the input signal level, which depend on the type and make of the microphones used in the system. The active microphones used by SotA VADs consume 20−50 μW [18], [19] in addition to the power consumption of the VAD circuitry itself. This is unacceptably high for always-on sensing acoustic systems. Such systems thus necessitate the use of passive microphones in low power budget applications. Such passive microphones typically have a sensitivity down to −60 dBV. This translates to an rms signal level of 30 μV at 65 dB sound pressure level (SPL) for a nominal conversation at 1 m distance [20]. This limits the maximum allowable noise-floor to less than 30 μVrms and also decides the minimum gain necessary in the amplifier depending on the LSB size, being 45 dB to achieve 8 bit precision over 1 V. This design has a gain-range from 20 to 80 dB in 20 dB steps to cover a wide range of input signals, although we anticipate that only up to 60 dB would be necessary. Also, the averaging time depends on the frequency of classification which in a typical VAD system is every 10–16 ms [8]–[10]. This averaging is implemented as LPF with a $f_{-3\text{ dB}}$ of 16 Hz. A summary of the VAD system specifications is highlighted in Table I.

IV. System Implementation

This section details the implementation nuances of the individual system blocks discussed in Section III-B, namely the wakeup detector, the analog feature-extractor, and the embedded mixed-signal classifier. A further section discusses system training for the complete VAD system before discussing one-time calibration and measurement results in Section V.

A. Wakeup Detector

The always-awake threshold-based wakeup detector acts as the system's watch-dog that wakes up the analog feature-extractor only when a signal of sufficient strength is detected. A single bit of information indicating the presence or absence of acoustic signal is needed. The wakeup detector is a low power three-phase comparator and its schematic is shown in Fig. 6. As the input signal level for this comparator can be as low as 30 μV and the comparator reference $\text{Vref}_{\text{comp}}$ is generated using 1.2 V, 8 bit DAC, at least 45 dB gain is necessary in the preamplifier to keep the signal swing greater than 1 LSB $\sim$ 4.5 mV.

The preamplifier is a cascade of four single-stage amplifiers. Each amplifier is a PMOS input source-coupled single-ended differential amplifier and can be turned ON/OFF individually to save power depending on the microphone's signal level and is designed to provide a midband gain of 20 dB. The $f_{-3\text{ dB}}$ of the amplifier is limited to 2 kHz as only the speech envelope needs to be detected. The comparator $\text{Vref}_{\text{comp}}$ can potentially vary as per the ambient noise-level, but this is beyond the scope of this work. Measured power consumption of this block is 700 nW when all four amplifier stages are turned ON, and excluding the external bias.

B. Analog Feature-Extractor

On receiving the wakeup signal from the threshold-based wakeup detector, the analog feature-extractor decomposes the input signal into the set of 16 features. The on-chip classifier evaluates whether the signal is potentially speech or background noise by comparing a feature subset to trained

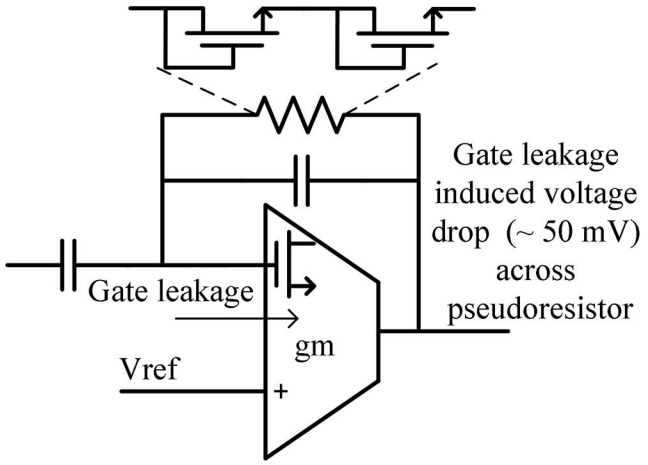

Fig. 8. Amplifier schematic highlighting gate leakage through the input pair.

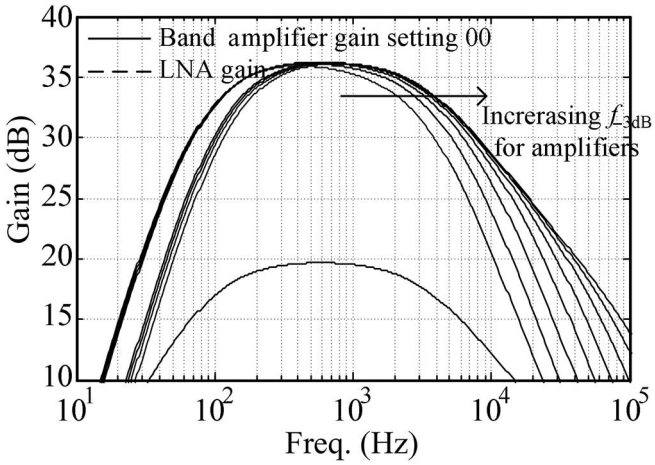

Fig. 9. Simulated frequency response for LNA and amplifiers in even bands showing increasing $f_{-3\ \mathrm{dB}}$.

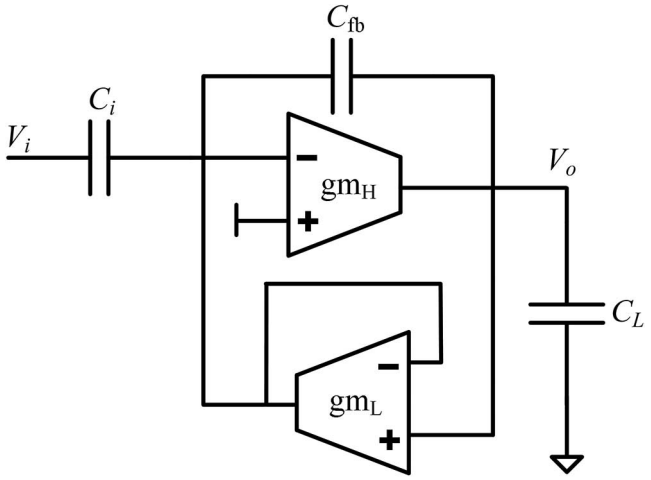

Fig. 10. First-order gm-C-based BPF topology.

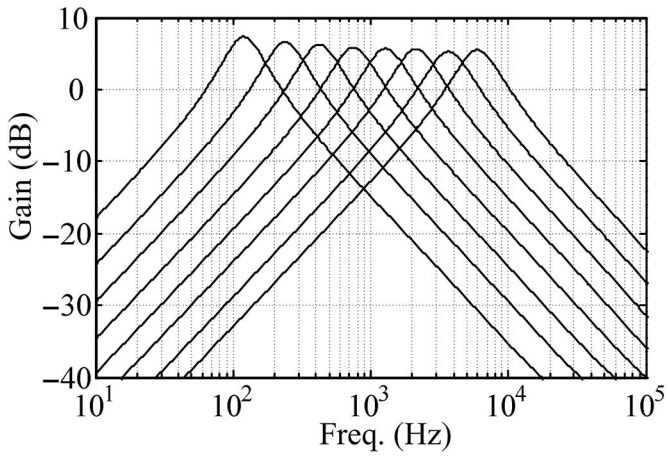

Fig. 11. Simulated frequency response for a constant $Q = 1.3$ BPF filters in even bands.

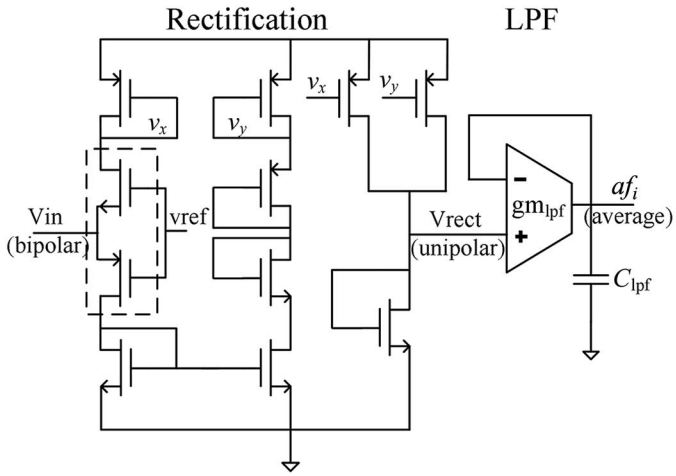

Fig. 12. Rectifier and LPF-based averaging circuit.

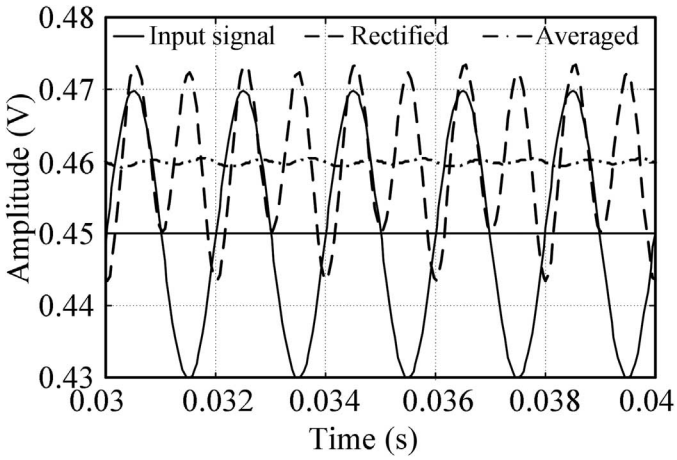

Fig. 13. Simulated response of the averaging circuit for a sinewave input of 20 mVpp amplitude and 500 Hz frequency.

thresholds in a decision tree (DT) topology (see Section IV-C). This section first describes the flow of the acoustic signal through the analog feature-extractor, followed by the implementation details of the individual blocks that participate in feature extraction.

Fig. 7 shows the detailed architecture for the analog feature-extractor. The signal from the passive microphone after low-noise amplification is fed to 16 bands. Each band allows further amplification and does a BPF operation with exponentially spaced f_c to mimic human hearing [21]. The output of each

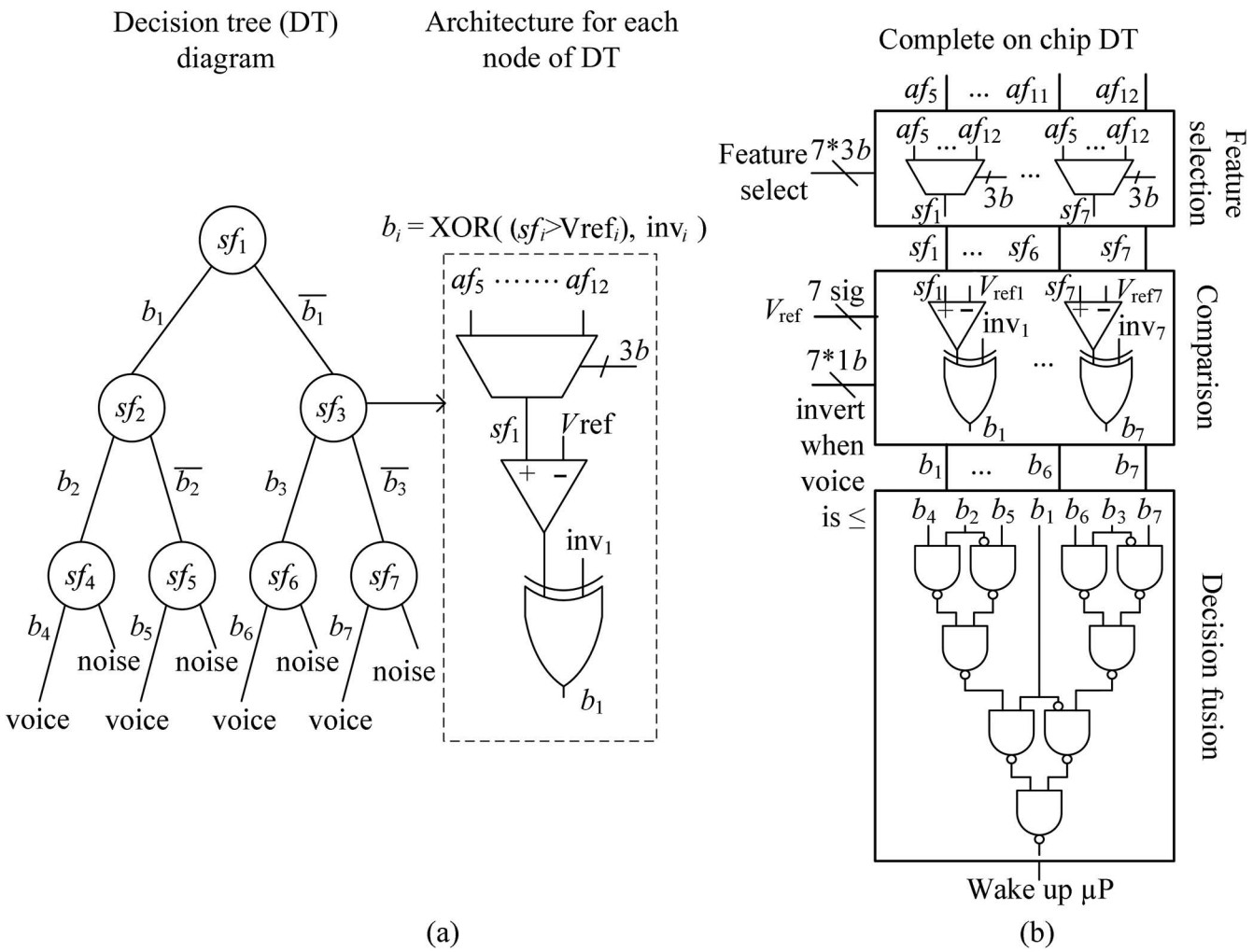

Fig. 14. Architecture of (a) one node of DT classifier and (b) complete classifier.

BPF filter is averaged by a rectification and LPF operation which results in 16 analog features $af_1 - af_{16}$, from which the subset $af_5 - af_{12}$ is used by the on-chip classifier.

The partitioning of the amplification between the shared LNA and the individual frequency bands allows a finer control over necessary amplification in each band. This contributes to power-proportional information extraction, as it allows turning OFF amplifier stages of unused features along with all other circuitry involved in individual feature computation. This enables context-aware power savings, as discussed in Section II-A2. The sub-blocks of the analog feature-extractor are now explained in more detail.

1) LNA and Amplifiers: The LNA is interfaced with a passive microphone and is designed to provide a midband gain of 20 dB up to a frequency range of 5 kHz while keeping the rms integrated input-referred noise smaller than 30 μV. The LNA is shared across all 16 bands as can be seen from Fig. 7. Further amplification in each band is done through a cascade of four individually controllable single-stage amplifiers with each stage designed to provide 20 dB gain as in Fig. 7. A single-stage amplifier topology was chosen for both LNA and in-band amplifiers for efficiency reasons, to avoid the power overhead of pushing nondominant pole(s) beyond the unity gain bandwidth. The closed-loop gain error introduced due to insufficient open-loop gain is a static error and is, as discussed, absorbed in the training phase.

The pseudoresistive feedback fixes the output bias point of the amplifier as shown in Fig. 8. As the area for the input transistors is large (80 μm × 10 μm) to reduce the flicker noise, gate leakage current up to 20 pA can shift the output bias point by as much as 50 mV due to voltage drop across the pseudoresistor. The interstage capacitive coupling, however, ensures that the bias point shift is not cascaded to next stage.

As will be discussed later, the BPFs across the bands have increasing center frequencies. To cover for this, the $f_{-3\text{ dB}}$ of the amplifiers in each band also increases progressively from band 1 to band 16. This is illustrated by the simulated magnitude response of the amplifiers in Fig. 9.

2) Bandpass Filters: The amplifier output in each of the 16 bands is passed through a BPF whose center frequency (f_c) increases exponentially from 75 Hz in band 1 to 5 kHz in band 16. The f_c for a second-order gm-C filter (see Fig. 10) is scaled by varying the bias current across the bands. From the BPF frequency response in Fig. 11, it can be seen that stop-band

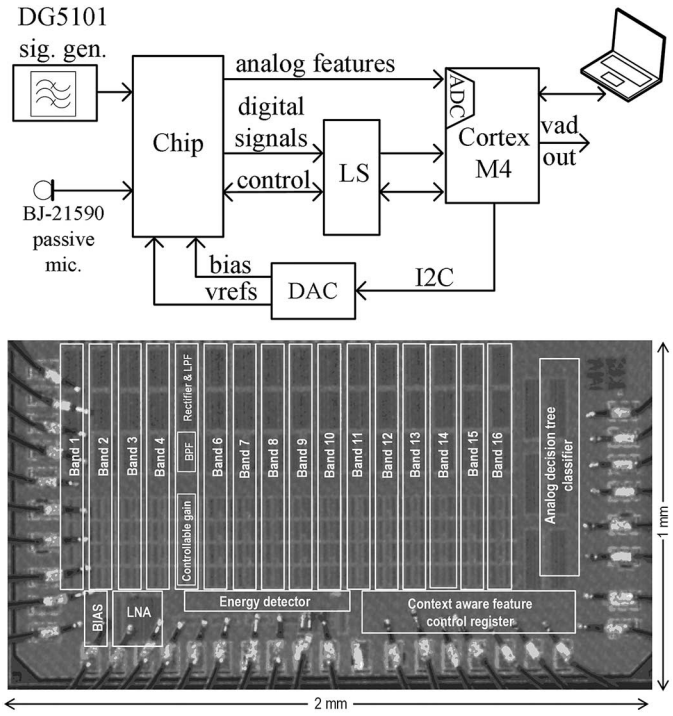

Fig. 15. Measurement setup (top) and chip micrograph (bottom) with important blocks highlighted.

attenuation for individual filters is better than -40 dB, but the adjacent band rejection is only -1.5 dB. This adds redundancy in the extracted features, leading to a high correlation between features of adjacent channels. This makes the system tolerant to shifts in the center frequency of BPFs.

3) Averaging Circuit: The output of each BPF is averaged individually by first rectifying and then low-pass filtering with an $f_{-3\text{ dB}}$ of 16 Hz to result in 16 analog features ($af_1 - af_{16}$). The architecture of the current-mode averaging is shown in Fig. 12. Normally-off transistors used for rectification (in dotted box) turn ON based on the direction of current from the BPF. The current steering network makes the current direction unipolar and is read across the gm-based resistors. A first-order gm-C LPF extracts the average value of this unipolar signal. Such normally-off transistors result in asymmetric rectification (dashed line) as in Fig. 13. This adds a dc-offset to the computed feature level shown by the averaged line (–dashed dot) in Fig. 13. Such offsets can be learnt during the training phase and do not affect classification accuracy.

C. DT-Based Classifier

The extracted feature subset $af_5 - af_{12}$ is passed on to the on-chip classifier (Fig. 5), while the complete feature-set $af_1 - af_{16}$ can be passed on to an off-chip ADC for more complex information extraction, such as context-change detection and retraining the classifier as in [22]. In these cases, the Nyquist sampling rate for the features is only $16 \times 2 \times 16 = 512$ Hz instead of 8 kHz for audio. The external ADC is *not* needed for embedded speech/non-speech classification.

The implementation of the on-chip 7 node 3 level mixed-signal DT classifier is shown in Fig. 14. Each node of the DT

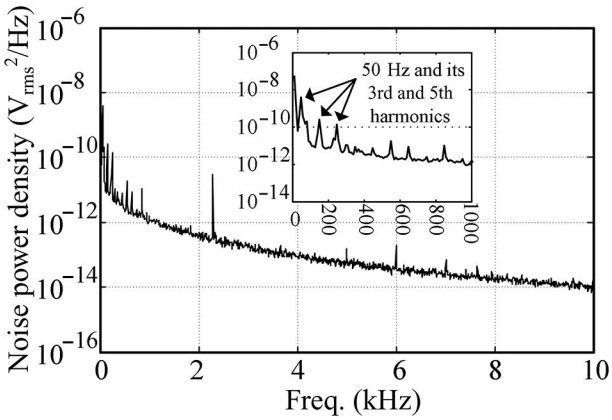

Fig. 16. Measured input-referred noise at the LNA output.

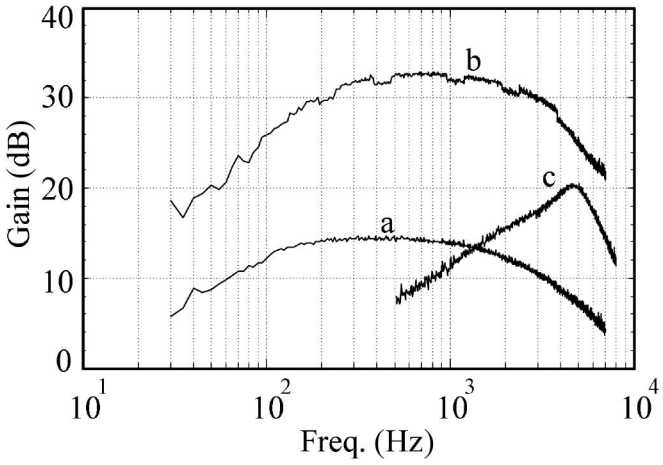

Fig. 17. (a) Measured small signal magnitude response for LNA, (b) amplifier with LNA, and (c) BPF with amplifier in 16th band.

can be configured to select one feature out of $af_5 - af_{12}$. The selected feature (sf_i) is then compared with a reference voltage (Vref_i) determined by a modified C4.5 machine-learning algorithm [22], generating the output decision b_i of each node

$$b_i = \text{xor}[(sf_i > \text{Vref}_i), \text{inv}_i] \tag{2}$$

where inv_i bit sets the comparison direction. The decision fusion logic shown in Fig. 14 combines the outputs of all DT nodes.

D. VAD System Training

The DT configuration and the individual feature activation are done using machine learning which selects the most discriminative features between speech and the current background noise context. To this end, the on-chip DT classifier is trained with our modified C4.5 algorithm with 160 s of labeled data from the standardized NOIZEUS database [23]. The traditional C4.5 algorithm selects a feature-set to maximize the total *information-gain*. Our modification to C4.5 maximizes the *information-gain/watt* and therefore outputs a *resource-efficient* model that maximizes the information capture while minimizing the power [22]. This is enabled as each feature extracts

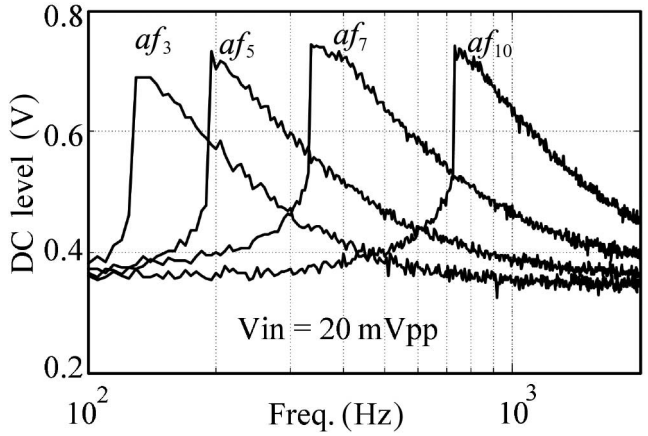

Fig. 18. Measured large signal frequency response of complete bands for bands 3, 5, 7, and 10.

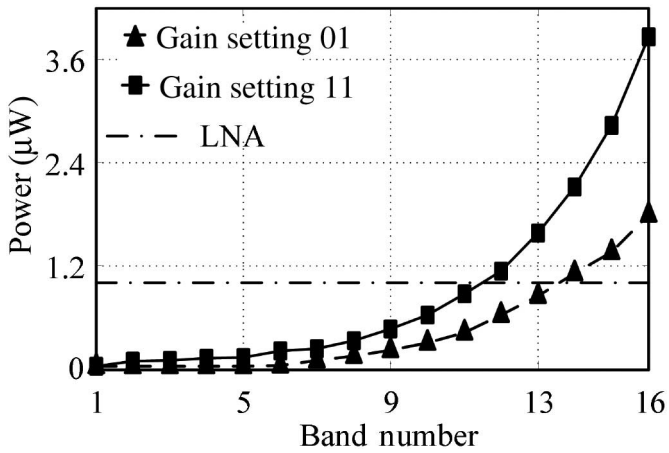

Fig. 19. Measured power consumption of LNA and of each band for gain setting of 01 and 11.

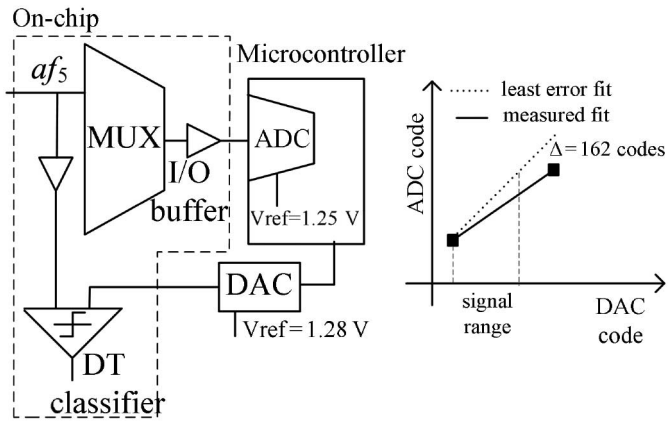

Fig. 20. Calibration scheme for ADC and DAC paths.

information from a higher frequency band so that the power cost increases from af_1 to af_{16}. This maximization of *information-gain/watt* furthers power-proportionality by increasing power consumption only for more (complex) information. The training runs on the microcontroller to generate a discriminating feature subset and reference levels for the comparison in the

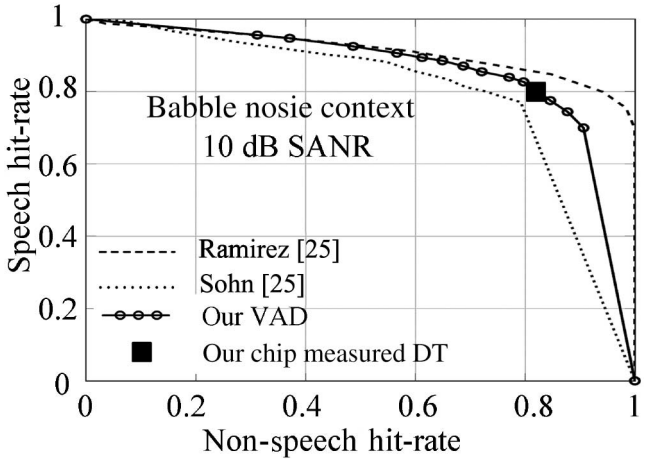

Fig. 21. Comparison of classification accuracy to STOA software VADs.

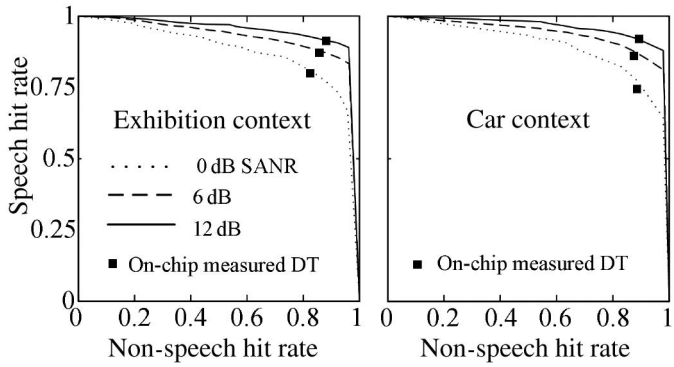

Fig. 22. Measured ROC curves depicting classification accuracy for multiple SANR in (a) exhibition and (b) car noise contexts.

TABLE II
MEASURED POWER CONSUMPTION VARIATION WITH CLASSIFICATION TASK COMPLEXITY ILLUSTRATING ACHIEVED POWER-PROPORTIONAL OPERATION

Task	Power
Signal detect	710 nW
Classify signal	2.6 μW—babble noise 1.1 μW—exhibition noise 1.05 μW—car noise
Detect context change and relearn DT [11]	57 μW
Voice activity detection (for babble noise context)	3.8 μW 80% signal detect 15% classification 5% detect context change and retrain

DT. The training results of the past context are not stored but dynamically learnt, as context change is detected [22].

V. MEASUREMENT SETUP AND RESULTS

The proposed system has been implemented on a 2 mm^2 chip in 90 nm CMOS as shown in Fig. 15. This section details the measurement results for the chip and for complete VAD system.

TABLE III
COMPARISON WITH STATE-OF-THE-ART VAD AND SIMILAR SYSTEMS

	This work	JETCAS'11 [27]	JSSC'13 [10]	[25]
Tech.	90 nm CMOS	0.5 μm CMOS	32 nm CMOS	Software only
Area	2 mm^2	2.25 mm^2	86 K gates	NA
Power (feature Extraction + classification)	6 μW worst case, all bands on	51 μW	< 50 μW	> 90 μW estimated [11]
Gain necessary for passive mic.	On chip	Off chip	Assumes digital mic.	NA
Feature type	Analog	Analog	Digital	Software
Classifier	On chip— mixed signal	Off chip— digital	On chip— digital	Software based
Context aware	Yes	NA	Yes	Yes
Feature-cost aware	Yes	NA	No	No
Latency	< 100 ms	100 ms	10 ms	10 ms
Classifier accuracy at 12 dB SNR	HR SP 89% HR non-SP 85% @ babble 12 dB SANR	90% car versus truck classification	97% unspecified SNR/context/ database	HR SP 89% HR non-SP 79% @ babble 12 dB SANR

A. Chip Performance Results

This section first discusses the measurement results for the LNA and some individual blocks in the 16th feature band in the chip followed by measurement results for complete bands.

The input-referred noise for the LNA is shown in Fig. 16. The noise has been measured at the LNA output over a frequency range of 10 Hz to 10 kHz. The rms input-referred integrated noise over the range of 75 Hz to 10 kHz is 32.5 μV. The total input-referred noise is expected to be 15% larger as this does not include the contributions from subsequent amplifier stages. For 3% and 5% THD, dynamic range is measured to be 40.2 and 45.4 dB, respectively, at 1 kHz.

Frequency responses of the individual blocks in the 16th feature band are shown in Fig. 17. Compared to simulation, the midband gain of the LNA is reduced by 4 dB, which is estimated to be due to insufficient open-loop gain. The large signal frequency response for the complete bands is shown in Fig. 18. As the f_c of the BPFs increases across bands, each band progressively processes higher frequency content to compute a feature; hence for a constant capacitive load, the power consumption increases from band 1 to band 16 as it can be seen from Fig. 19. As already mentioned in Section IV-D, this allows a power-aware learning to enable efficient classification. Finally, the measured rms noise at the output of each band is less than 2 mV. For an output signal range of 400 mV, this gives 7.5 bits of precision.

B. System Measurement Results

The chip is integrated with the microcontroller using external level-shifters and DACs, to form the complete VAD. Fig. 20 shows a one-time calibration to characterize for mismatch in the ADC and DAC paths. This section also displays the classification accuracy results for the complete VAD system and illustrates the achieved power-proportionality.

Receiver operating characteristic (ROC) curves characterize the classifier systems and depict hit rates (HR) for the variables under observation [24]. Fig. 21, ROC curve, shows that classification accuracy of our on-chip classifier is on-par with software-based VAD systems of [8], [9], and [25]. Further Fig. 22 validates the classification capacity over multiple contexts with different background noise conditions. Table II illustrates the power-proportional sensing in our VAD system by showing the gradual increase in system power consumption with the sensing task complexity. The power consumption for signal detection is measured to be below 1 μW, whereas power consumption for classification varies depending on the complexity of the operating context and has an upper bound of 6 μW. The power consumption for background context-change detection and relearning the DT is estimated to be 57 μW on a cortex M4 microcontroller. It is predicted that the VAD system will be 80% of the time in detection mode, 15% in classification mode, and about 5% of time performing complex tasks, such as relearning the context or DT training. The resulting duty-cycled power consumption is 3.8 μW for babble noise context. Further, the estimated power overhead for generating on-chip (currently off-chip) reference voltages for the comparators is leakage limited and is estimated to be less than 50 nW per reference value [26] as the reference voltage needs to drive only the gate nodes at near dc speed. Table III compares our work to SotA VADs [8]–[10], [25] and similar systems [27]. While maintaining the same classification accuracy as compared to software VADs, our system reduces the power consumption by 10×. Although hierarchical information extraction adds a maximum latency of 100 ms to the VAD decision task, this does not cause significant information loss as this latency is smaller than the average duration of a spoken vowel [28].

VI. Conclusion

This work demonstrates a power-efficient acoustic sensing frontend for speech/non-speech classification in a VAD system. The power efficiency is achieved by the use of machine-learning-assisted analog feature computation and by infusing the power-proportionality paradigm in various ways throughout the architecture. The use of analog features for information extraction allows individual turning ON/OFF of features depending on the usefulness of a feature in a particular context while the power-proportionality concept controls the hierarchical activation of different sub-blocks depending on the complexity of the information extraction task. The idea of power-proportional sensing is demonstrated for an acoustic sensing system and can be extended to other systems such as motion and image sensing systems.

References

[1] E.-Hung Chen *et al.*, "Adaptation of CDR and full scale range of ADC-based SerDes receiver," in *Proc. Symp. VLSI Circuits*, Jun. 16–18, 2009, pp. 12–13.

[2] B. Schell and Y. Tsividis, "A continuous-time ADC/DSP/DAC system with no clock and with activity-dependent power dissipation," *IEEE J. Solid-State Circuits*, vol. 43, no. 11, pp. 2472–2481, Nov. 2008.

[3] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 520–544, Jan. 2010.

[4] S. Pfetsch *et al.*, "On the feasibility of hardware implementation of sub-Nyquist random-sampling based analog-to-information conversion," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS'08)*, May 18–21, 2008, pp. 1480–1483.

[5] D. J. White, P. E. William, M. W. Hoffman, S. Balkir, and N. Schemm, "Analog sensing front-end system for harmonic signal classification," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS'12)*, May 20–23, 2012, pp. 1155–1158.

[6] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[7] F. Chen, A. P. Chandrakasan, and V. M. Stojanovic, "Design and analysis of a hardware-efficient compressed sensing architecture for data compression in wireless sensors," *IEEE J. Solid-State Circuits*, vol. 47, no. 3, pp. 744–756, Mar. 2012.

[8] J. Ramirez, J. M. Gorriz, J. C. Segura, C. G. Puntonet, and A. J. Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum LRT," *IEEE Signal Process. Lett.*, vol. 13, no. 8, pp. 497–500, Aug. 2006.

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

[10] A. Raychowdhury, C. Tokunaga, W. Beltman, M. Deisher, J. W. 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 J. Solid-State Circuits*, vol. 48, no. 8, pp. 1963–1969, Aug. 2013.

[11] S. Lauwereins, W. Meert, J. Gemmeke, and M. Verhelst, "Ultra-low-power voice-activity-detector through context-, and resource-cost-aware feature selection in decision trees," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP'14)*, Sep. 21–24, 2014, pp. 1–6.

[12] R. Sarpeshkar, "Analog versus digital: Extrapolating from electronics to neurobiology," *Neural Comput.*, vol. 10, no. 7, pp. 1601–1638, May 1998.

[13] J. Zhang, Z. Wang, and N. Verma, "18.4 A matrix-multiplying ADC implementing a machine-learning classifier directly with data conversion," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC' 15)*, Feb. 22–26, 2015, pp. 1–3.

[14] S.-Y. Hsu, Y. Ho, P.-Y. Chang, C. Su, and C.-Y. Lee, "A 48.6-to-105.2 μW machine learning assisted cardiac sensor SoC for mobile healthcare applications," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 801–811, Apr. 2014.

[15] B. Murmann, "Digitally assisted data converter design," in *Proc. Eur. Solid-State Circuits Conf. (ESSCIRC'13)*, Sep. 16–20, 2013, pp. 24–31.

[16] K. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "24.2 Context-aware hierarchical information-sensing in a 6 μW 90 nm CMOS voice activity detector," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC'15)*, Feb. 22–26, 2015, pp. 1–3.

[17] F. J. Fahy, "Measurement of acoustic intensity using the cross-spectral density of two microphone signals," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 1057–1059, 1977.

[18] Knowles. *Microphones*. [Online]. Available: http://www.knowles.com/eng/Products/Microphones, accessed on Apr. 30, 2015.

[19] Invensense. *Microphones*. [Online]. Available: http://www.invensense.com/mems/microphone/, accessed on Apr. 30, 2015.

[20] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an International Speech Test Signal (ISTS)," *Int. J. Audiol.*, vol. 49, no. 12, pp. 891–903, 2010.

[21] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, 1937.

[22] S. Lauwereins, K. Badami, W. Meert, and M. Verhelst, "Optimal resource usage in ultra-low-power sensor interfaces through context-, and resource-cost-aware machine learning," *Neurocomputing*, vol. 169, pp. 236–245, Dec. 2, 2015.

[23] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7–8, pp. 588–601, 2007.

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

[25] J. Kola, C. Espy-Wilson, and T. Pruthi, "Voice activity detection," MERIT BIEN, pp. 1–6, 2011.

[26] M. Yip and A. P. Chandrakasan, "A resolution-reconfigurable 5-to-10-bit 0.4-to-1 V power scalable SAR ADC for sensor applications," *IEEE J. Solid-State Circuits*, vol. 48, no. 6, pp. 1453–1464, Jun. 2013.

[27] B. Rumberg, D. W. Graham, V. Kulathumani, and R. Fernandez, "Hibernets: Energy-efficient sensor networks using analog signal processing," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 1, no. 3, pp. 321–334, Sep. 2011.

[28] S. A. House, "On vowel duration in English," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1174–1178, 1961.

Komail M. H. Badami (S'14) received the M.S. degree in electrical engineering from the Indian Institute of Technology Madras, Chennai, India, in 2011. Currently, he is pursuing the Ph.D. degree in microelectronics at MICAS Laboratories, KU Leuven, Leuven, Belgium.

From August 2011 to December 2012, he was with Intel Tech. India Pvt. Ltd, Bangalore, India, where he worked on high speed memory I/O design. His research interests include ultra low power analog/mixed-signal circuit design for context aware applications.

Steven Lauwereins (S'14) received the M.S. degree in electrical engineering from ESAT-MICAS Laboratories, KU Leuven, Leuven, Belgium, in 2013.

In 2013, he became a Research Assistant with ESAT-MICAS Laboratories, KU Leuven. His research interests include implementing machine-learning methods at circuit level to optimize power consumption in sensor interfaces.

Wannes Meert received the M.S. degree in electrical engineering, the M.S. degree in artificial intelligence, and the Ph.D. degree in machine learning from ESAT-MICAS Laboratories and CS-DTAI, KU Leuven, Leuven, Belgium, in 2005, 2006, and 2011, respectively.

He is currently a Postdoctoral Researcher with CS-DTAI, KU Leuven. His research interests include machine learning, data mining, and artificial intelligence, in general, for industrial applications.

Marian Verhelst (M'08–SM'13) received the M.S. and Ph.D. degrees in electrical engineering from ESAT-MICAS Laboratories, KU Leuven, Leuven, Belgium, in 2003 and 2008, respectively.

She was a Visiting Scholar at Berkeley Wireless Research Center, University of California at Berkeley, Berkeley, CA, USA, in 2005. From 2008 to 2011, she was with the Radio Integration Research Laboratory, Intel Corporation, Hillsboro, OR, USA, involved in research on digital-assisted analog and radio frequency. In 2012, she became a Professor with the ESAT-MICAS Laboratories, KU Leuven. Her research interests include smart, self-adaptive system architectures and circuits for ubiquitous sensing and computing.