Analog Acoustic Feature Extraction for Always-On Voice Activity Detection

Dual Degree Project Stage I Report

submitted in partial fulfillment of the requirements

for the degree of

Bachelor of Technology in Electrical Engineering and Master of Technology in Microelectronics

by

Mihir Kavishwar

(Roll No: 17D070004)

Supervisor:

Prof. Rajesh Zele

Department of Electrical Engineering

Indian Institute of Technology Bombay

Powai, Mumbai - 400076

October 27, 2021

Declaration

I declare that this written submission represents my ideas in my own words and where other ideas or words or diagrams have been included from books/ papers/electronic media, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will result in disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken.

Mihir Kavishwar

(Roll No:17D070004)

October 27, 2021

Acknowledgement

I would like to express my gratitude to my guide Prof. Rajesh Zele for giving me the opportunity to work on this project and for his constant motivation and guidance. Secondly, I would like to thank my project partner, Prashant Kurrey, for the insightful discussions. I would also like to thank all the member of Advanced Integrated Circuits and Systems Laboratory for their valuable suggestions.

Mihir Kavishwar
Electrical Engineering
IIT Bombay

Abstract

This report presents a critical analysis of various Analog Feature Extraction techniques used in State-of-the-art (SotA) ultra-low-power Voice Activity Detectors (VADs). A VAD architecture is proposed which performs acoustic feature extraction in Analog domain while Neural Network based classification in Digital domain. The Analog Frontend of the proposed architecture is based on a previous work that uses Switched Capacitor N-Path Bandpass Filters, Zero-IF Mixers and Low Pass Filters to extract energy of input signal in different frequency bands during fixed time frames. Major circuit blocks were implemented at the transistor level in UMC 65nm technology and their simulation results have also been discussed in this report.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Automatic Speech Recognition (ASR) has become increasingly popular in recent years and is widely used in smartphones, wearables and other Internet of Things (IoT) devices. Complex tasks such as Keyword Spotting, Speaker Verification and Speech-to-Text Conversion are typically performed using Machine Learning (ML) algorithms that require significant computational power. However, Edge devices have severe energy constraints since they are powered by small batteries and therefore cannot continuously run these algorithms. State-of-the-art (SotA) systems overcome this issue by using the concept of hierarchical detection - a set of tasks with increasing complexity are cascaded, allowing activation of posterior stages by previous steps in the pipeline [11]. Only the first stage in the classifier cascade is always-on, thus making the overall system much more energy efficient.

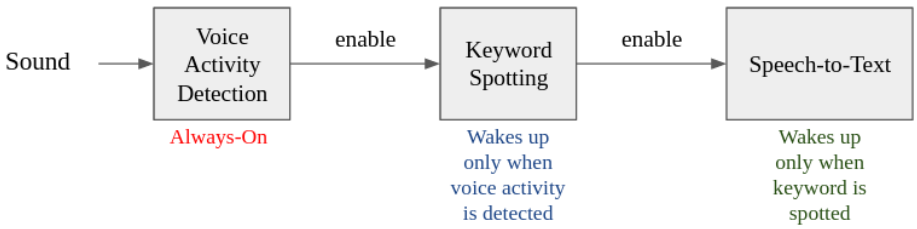


Figure 1.1: Always-on voice activity detection as a wakeup mechanism for keyword spotting and speech-to-text conversion

A Voice Activity Detector (VAD) identifies the presence or absence of human speech in an audio signal [13]. In most SotA systems, VADs are used as the first stage in the classifier cascade. They remain always-on and serve as a wake up mechanism for the DSP blocks which perform more advanced tasks. Therefore, the power consumption of a VAD is extremely critical and

can have significant impact on the battery life of a device. Moreover, energy efficiency shouldn't come at the expense of a significant accuracy degradation because if a VAD fails to detect speech, it won't wake up the subsequent stages in the classifier chain which perform advanced processing.

Several techniques for implementing energy efficient VADs have been discussed in the recent literature. Typical VADs consist of two parts [8]:

1. **Acoustic Feature Extractor -** It converts the incoming signal into low-dimensional but dense acoustic features. Previous works have extracted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) [7], [11], [14], [16]; input signal energies in different frequency bands [3], [4], [8], [12], [17]; or non-linear spiking events based on band energies [10], [18].

2. **Classifier -** It takes a feature set input and produces a binary decision: "speech" or "non-speech". Previous works have used different classifier models such as Decision Trees [4], [12]; Support Vector Machines [6], [9]; or more commonly Neural Networks [7], [8], [10], [12], [14], [16], [17], [18]. Some works have used an energy thresholding based method [3], [13].

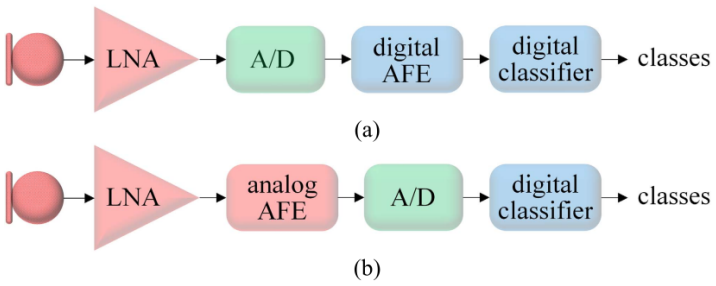


Figure 1.2: (a) Conventional digital-intensive and (b) analog-intensive signal processing chain in acoustic inference sensing systems. AFE = Acoustic Feature Extractor. *Source: Minhao Yang et al., JSSC '21 [18].*

Depending on the choice of features and the classification algorithm, either of these tasks can be more computationally expensive. The feature extractor can be made more energy efficient by performing some of the computation in the analog domain, very close to the microphone sensor node. This relaxes the specifications for the ADC as well as relaxes the complexity of our classifier. The classifier, which typically has to perform several Multiply and Accumulate (MAC) operations, can be made very energy efficient using

latest advancements in mixed-signal computing for neural network inference [15]. This work focuses understanding different analog feature extraction architectures as well as proposes a new architecture based on previous works. Our goal is to build an ultra-low-power low latency Voice Activity Detector which exploits the power of Analog signal processing for feature extraction and embedded Machine Learning techniques for classification.

The rest of this report is organised as follows. Chapter 2 gives an overview of fundamental concepts that have been used in subsequent chapters. Chapter 3 has a detailed literature survey of many SotA VAD implementations. Chapter 4 introduces a VAD architecture based on previous works. Chapter 5 presents the simulation results of circuit blocks that were implemented at the transistor level or as behavioural models. Chapter 6 concludes this reports and discusses future work. Lastly, Chapter 7 lists all the references that were cited in this report.

Chapter 2

Overview of Spectrograms and MFCCs

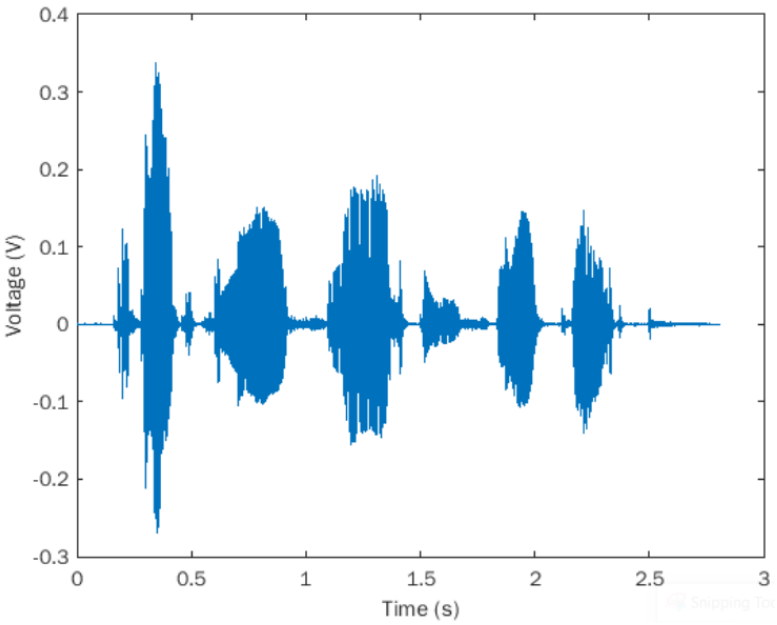
In this chapter, I present an overview of some of the key concepts that would be helpful for understanding the work presented in subsequent chapters.

Spectrogram is a visual representation of the frequency spectrum of a signal as it varies over time. Conceptually, to plot the spectrogram of a continuous time signal we need to perform the following steps:

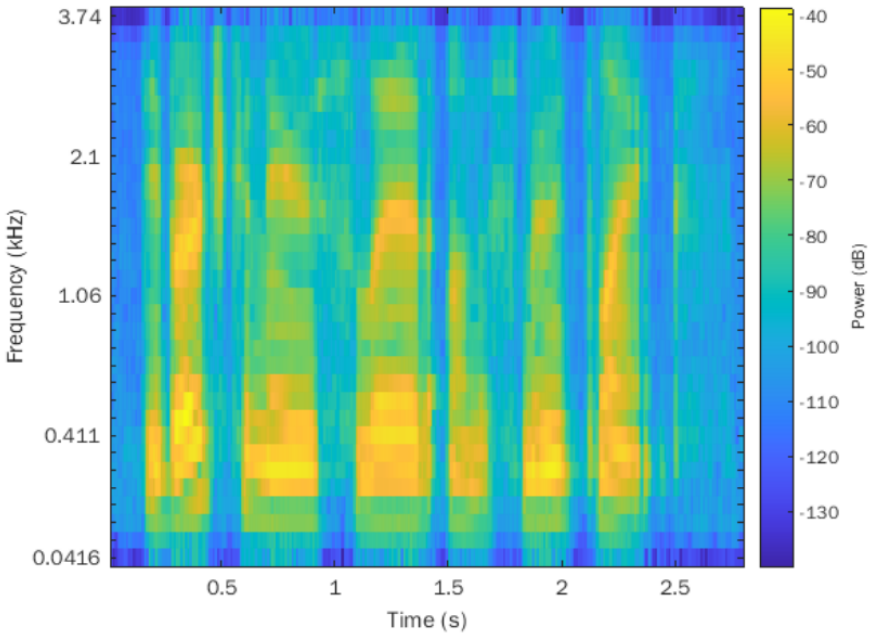
1. Section the long signal into shorter, fixed length frames
2. Compute the Power Spectral Density (PSD) of each frame and plot it on the Y-axis using different color shades to represent magnitude
3. Stack the spectral densities of successive time frames beside one another on the X-axis

The **mel scale** is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The log spectrum on a **mel frequency scale** (the mel log spectrum) is considered to be a more effective representation of the spectral envelope of speech than that on the linear frequency scale [1]. Thus, in audio speech processing we typically use **mel spectrogram**, which is a spectrogram where the frequencies are converted to the mel scale. The formula to convert from f hertz to m mels is:

$$m = 2595 \times log_{10}\left(1 + \frac{f}{700}\right)$$



(a) Time domain signal



(b) Mel Spectrogram

Figure 2.1: Different representations of the same speech signal

The **mel-frequency cepstrum** (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a

log power spectrum on a nonlinear mel scale of frequency. **Mel-frequency cepstral coefficients** (MFCCs) are coefficients that collectively make up an MFC. MFCCs are commonly derived as follows [2]:

1. Take the Fourier transform of (a windowed excerpt of) a signal
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows or alternatively, cosine overlapping windows
3. Take the logs of the powers at each of the mel frequencies
4. Take the discrete cosine transform of the list of mel log powers
5. The MFCCs are the amplitudes of the resulting spectrum

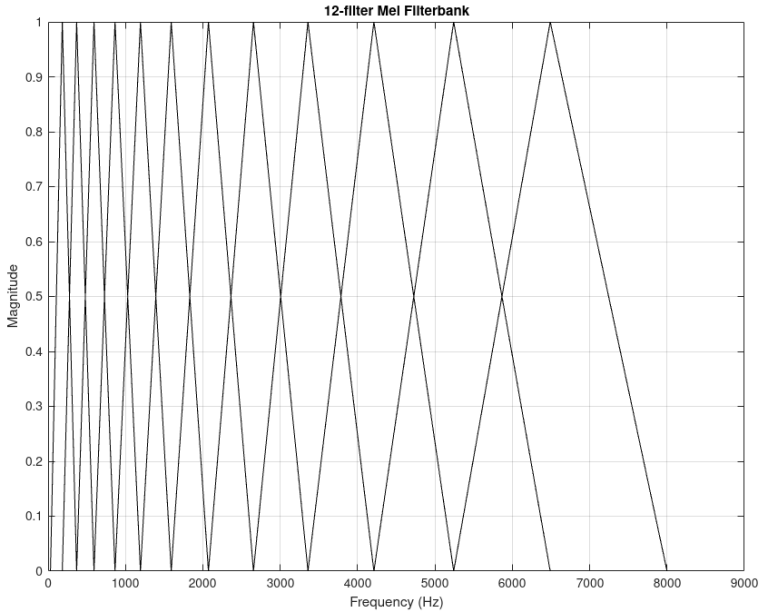


Figure 2.2: 12-filter Mel Filterbank for computing MFCCs

Chapter 3

Literature Survey

3.1 Power-Proportional Acoustic Sensing

This section is focused on the work by Badami et al.[4] in where they first introduced the idea of power-proportional sensing which has now been adopted by several other SotA speech processing systems. Power-proportional sensing paradigm aims to scale the power consumption of a system in proportion to the complexity of the sensing task. Thus, power hungry blocks in the signal processing tool chain are turned on only as the task of information extraction gets more complex.

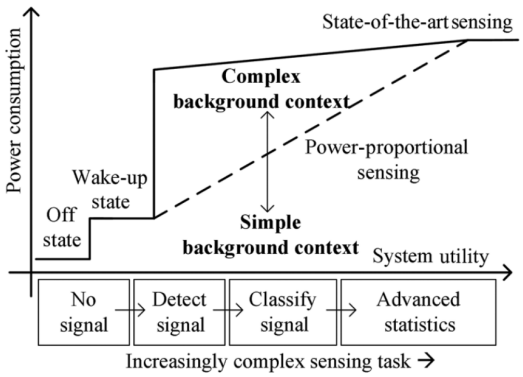


Figure 3.1: Power-proportional sensing in contrast with then SotA systems. *Source: Komail Badami et al., JSSC '16 [4].*

The work presents a proof of concept acoustic frontend for voice activity detection based on the power-proportional sensing paradigm. The system architecture is shown in figure 3.2. The operation of this system is as follows. An always-on threshold-based wakeup block keeps checking the passive

microphone for sound activity. When any signal is detected, it wakes up the analog feature-extractor that translates the input signal into a set of features. The on-chip classifier uses some of these features to classify incoming signal as speech or non-speech. If the signal is speech, the classifier wakes up the microcontroller for more advanced processing with the complete feature set.

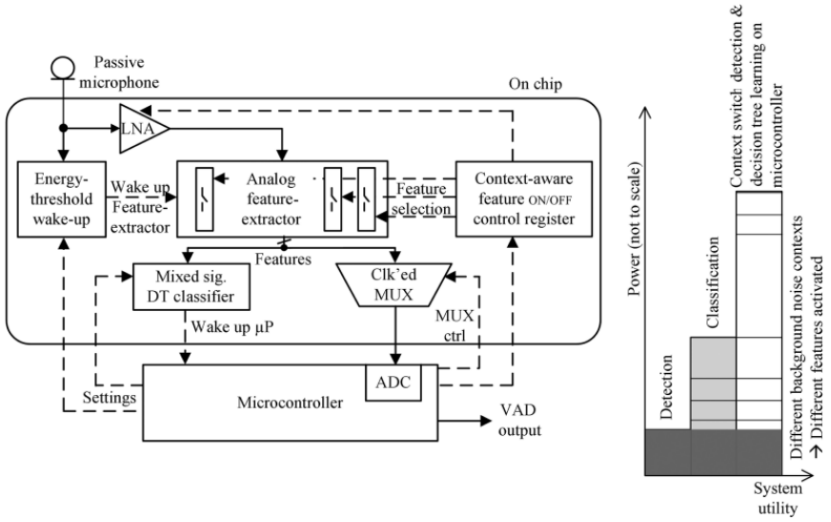


Figure 3.2: System architecture of power-proportional VAD (left) and it's power scaling with sensing complexity (right). *Source: Komail Badami et al., JSSC '16 [4].*

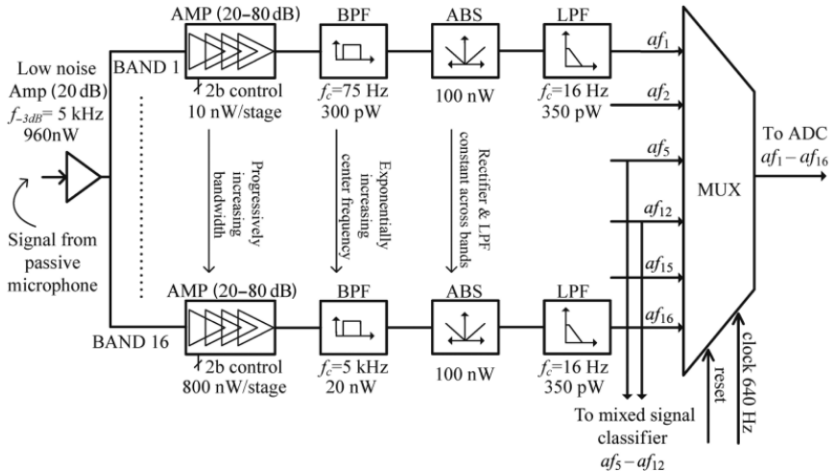


Figure 3.3: Schematic and design parameters of the analog feature extraction block. *Source: Komail Badami et al., JSSC '16 [4].*

The analog feature extractor decomposes the input signal into a set of 16 features, which are used by a trained Decision Tree classifier to evaluate if the signal is speech or non-speech. Mathematically, each analog feature af_i is defined as

$$af_i = \overline{abs[Ax(t) * h_i^{BPF}]}$$

For both band pass and low pass functions, active gm-C filters are used. The rectifier and low pass filter are implemented in current mode to form an active averaging circuit. The decision tree classifier uses a modified C4.5 machine-learning algorithm.

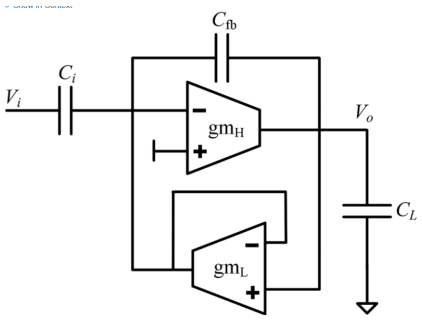


Figure 3.4: Second order gm-C Bandpass Filter. *Source: Komail Badami et al., JSSC '16 [4].*

Small signal analysis of this circuit gives us

$$\frac{V_o(s)}{V_i(s)} = \frac{sC_i(sC_{fb} - gm_H)}{s^2(C_iC_L + C_{fb}C_L + C_{fb}C_i) + s(C_Lgm_L + C_{fb}gm_H) + gm_Lgm_H}$$

While it is possible to get a good quality factor with this filter topology, it is not easy to tune the center frequencies if we have large number of bands.

3.2 Event-based Acoustic Feature Extraction

This section is focused the work presented by Minhao Yang et al. in JSSC '19 [10] which proposes an Analog intensive VAD architecture taking biomimetic inspiration. The acoustic features (band energies) are extracted in Analog domain, and then encoded by event-driven ADCs into parallel events. A digital Deep Neural Network receives parallel event streams from the ED-ADCs and classifies the signal as speech or non-speech as shown in Figure 3.5.

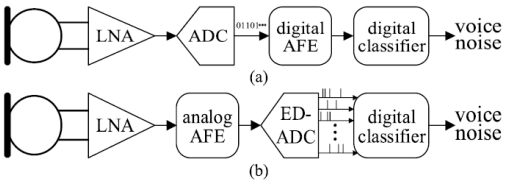


Figure 3.5: (a) Conventional digital intensive and (b) analog-intensive signal processing chain in a VAD. *Source: Minhao Yang et al., JSSC '19 [10]*

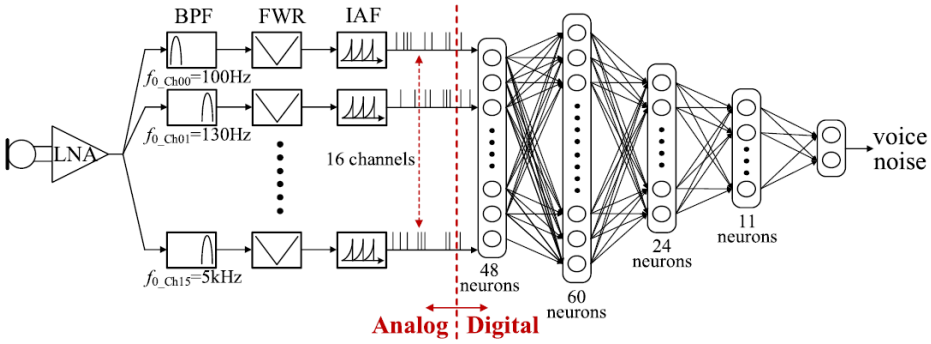


Figure 3.6: VAD system architecture using analog acoustic feature extraction and digital classification with event-driven analog-to-digital conversion. *Source: Minhao Yang et al., JSSC '19 [10]*

The system architecture is shown in Figure 3.6. The input audio signals are amplified by a LNA and then analysed by 16 parallel channels, each composed of a Bandpass Filter (BPF), a Full-wave Rectifier (FWR), and an integrate-and-fire (IAF) encoder as ED-ADC. The event sequence can be ideally described by the following equation:

$$\frac{1}{C_{\text{int}}} \int_{t_j}^{t_{j+1}} |f_{v\to i}(v_{o\text{BPF\_}k}(t))|dt = V_{\text{refdn}}$$

where t_j is the time stamp of the j^{th} event,$v_{o\text{BPF\_}k}$ is the BPF output voltage in channel k, $f_{v\to i}$ is the voltage-to-current conversion function, and C_{int} and V_{refdn} are the integration capacitance and the threshold voltage of IAF, respectively. The time interval between two adjacent events is the function of the integrated $v_{o\text{BPF\_}k}$.

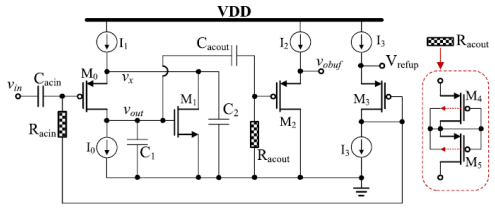


Figure 3.7: Schematic of the SSF-based BPF with output buffer and input dc bias. The fabricated circuit was a differential version. *Source: Minhao Yang et al., JSSC '19 [10]*

The work presents a new 2nd-order BPF circuit shown in Figure 3.7 which is based on super-source-follower (SSF) topology. The authors figured out that classification results are satisfying even with low quality factors and low-BPF orders and hence did not chose high order filter topologies. The transfer functions $H_{BPF}(s)$, central frequency f_0 , quality factor Q, and peak gain A_0 are derived as

$$H_{\mathrm{BPF}}(s) = \frac{v_{\mathrm{out}}}{v_{\mathrm{in}}} = -\frac{s\frac{C_2}{g_{m2}}}{s^2\frac{C_1C_2}{g_{m1}g_{m2}} + s\frac{C_1}{g_{m2}} + 1}$$

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{g_{m1}g_{m2}}{C_1C_2}}, \quad Q = \sqrt{\frac{g_{m2}C_2}{g_{m1}C_1}}, \; A_0 = \frac{C_2}{C_1}$$

The schematic of FWR and IAF is shown in Figure 3.8. The detailed analysis can be found in the original paper.

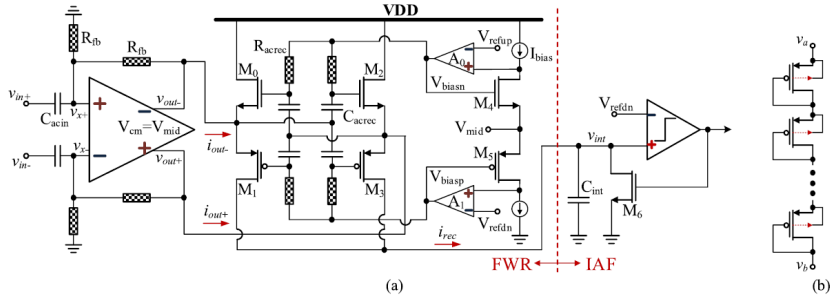


Figure 3.8: Schematic of (a) FWR and IAF and (b) pseudo-resistor R_{fb} that is composed of eight diode-connected pFETs connected in series *Source: Minhao Yang et al., JSSC '19 [10]*

The authors also recently published a new paper [18] which builds on

their previous work and exploits some non-linear properties of analog feature extractor.

3.3 Mixer-based Sequential Frequency Scanning

This section is focused the work presented by Sechang Oh et al. in JSSC '19 [8] which discusses a programmable acoustic signal processing system for both VAD and non-voice acoustic event detection based on Neural Network classifier. The authors propose using a sequential frequency scanning technique as shown in figure 3.9 instead of parallel feature extraction like in [4] and [10]. This allows them to further reduce the power consumption to sub-μW levels.

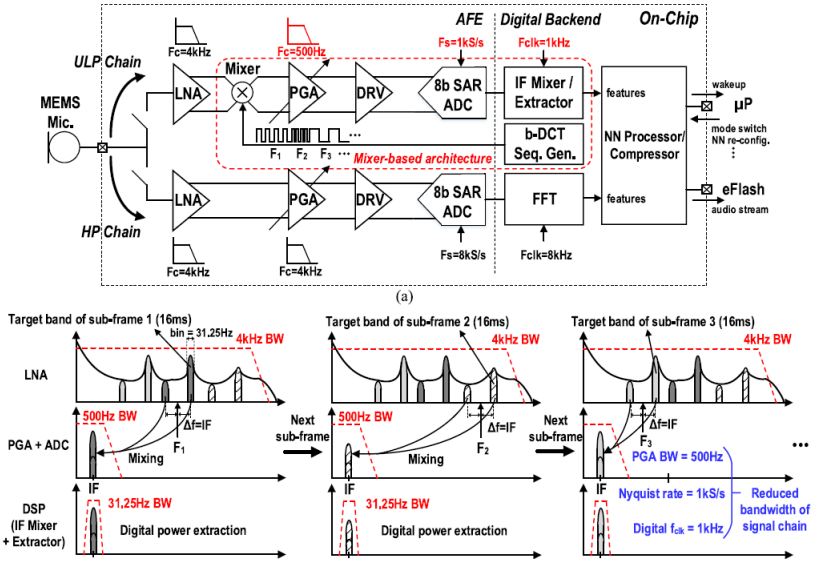


Figure 3.9: (a) VAD system architecture. (b) Operating principle of mixer-based sequential frequency scanning. *Source: Sechang Oh et al., JSSC '19 [8]*

Based on the same concept of power-proportional sensing discussed in previous section, this system has two signal chains: an always-on ultra-low-power (ULP) chain and a high performance (HP) chain that wakes upon event detection by the ULP chain. In the ULP mode, the system consumes just 142-nW while in HP mode it consumes 18-μW. The HP chain consists of 4-kHz bandwidth and 8-kS/s sampling rate with conventional AFE architectures consisting of low-noise aplifier (LNA), programmable amplifier

(PGA), ADC driver (DRV), and ADC. In contrast, the ULP chain employs a digitally controlled mixer between LNA and PGA to shift the desired signal frequency down to 500-Hz to lower the Nyquist rate to 1 kS/s after the PGA.

The amplifiers used in the AFE are based on capacitively coupled amplifier topology . The input transistors are biased using a DC common-mode feedback between the input and output. This results in the AFE having a bandpass nature, which is suitable for filtering acoustic signals that have similar bandwidths. Therefore, the signal cannot be mixed down to DC in the AFE iteself and has to be downconverted in the digital backend. Moreover, the impact of flicker noise of the PGA is also reduced when we divide the downconversion process between AFE and digital backend.

The authors have reported measurement results with actual audio signals, and not just electric analog audio signals like some previous works. While the system performs well on energy efficiency metric, it has significantly more latency than other SotA systems due to its sequential processing architecture.

3.4 N-Path Bandpass Filtering

This section is focused the work presented by Villamizar et al. in TCAS1 '21 [17] which demonstrates an excellent application of N-Path Switched Capacitor Bandpass Filters for Acoustic Feature Extraction. N-Path filters have been studied in great detail in past [5]. Figure 3.10 shows they key idea used in N-Path Filters.

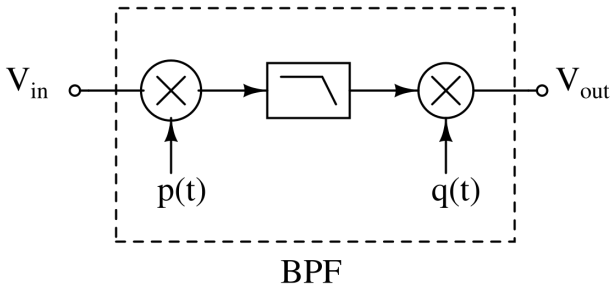


Figure 3.10: The fundamental principle used in an N-Path Bandpass Filter is Downconversion + Lowpass Filtering + Upconversion = Bandpass Filtering

In N-Path filters, the several BPF blocks as shown in Figure 3.10 are connected in parallel and get turned on one after another in a cyclic fashion.

Figure 3.11 shows how we can derive the final differential N-Path filter circuit from the original idea.

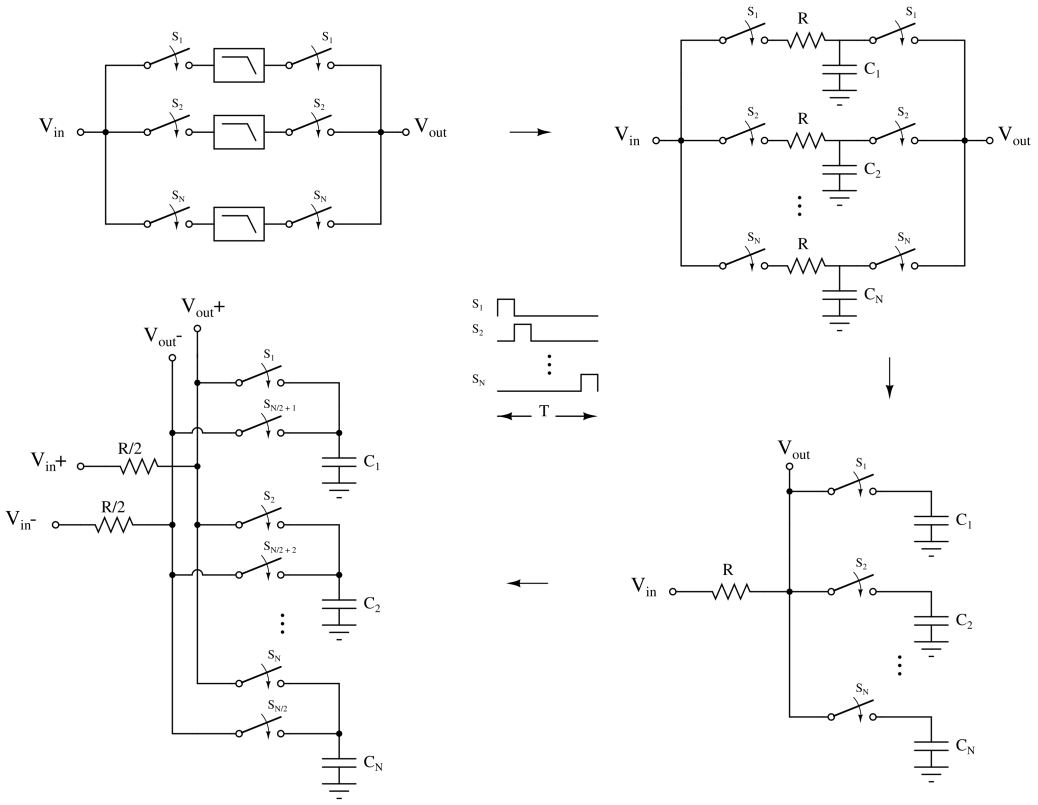


Figure 3.11: Derivation of differential N-Path Bandpass Filter from the original conceptual diagram

The center frequency is same as clock frequency (f_c) while the 3 dB bandwidth is given by ($\frac{1}{2\pi NRC}$). N-Path Filters have the several advantages [5]:

1. High quality factor is easily achievable
2. They are extremely tunable since the center frequency of the filter is determined by the clock frequency
3. Energy efficiency is high since because power is required only to drive switches

These features makes them an ideal candidate for acoustic feature extraction. The two main issues with N-Path Filters - presence of harmonic responses and folding, can be absorbed by the Machine Learning model of the classifier.

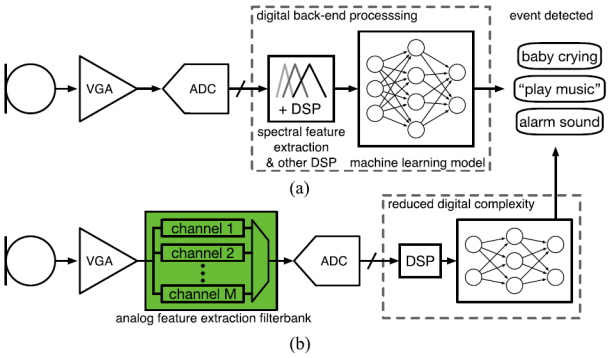


Figure 3.12: Architecture comparison. (a) Traditional digital system performing analog-to-digital conversion as close as possible to the sensor signal. (b) Analog feature extraction before the ADC to reduce digital processing and ML model complexity. *Source: Villamizar et al., TCAS1 '21 [17]*

Notice that the N-Path schematic in figure 3.13 is a specific case of the final circuit in figure 3.11 with N = 4. Switched capacitor resistors are used so that the system is sensitive to only capacitor matching and clock frequency, both of which are well within are well controlled in fabrication and sufficiently stable during operation. Changing the center frequency and quality factor of such BPF is easy. The center frequency can be changed by modifying the clock rate, The quality factor can be changed by varying the capacitor ratio between C_i and C_{SC-R}.

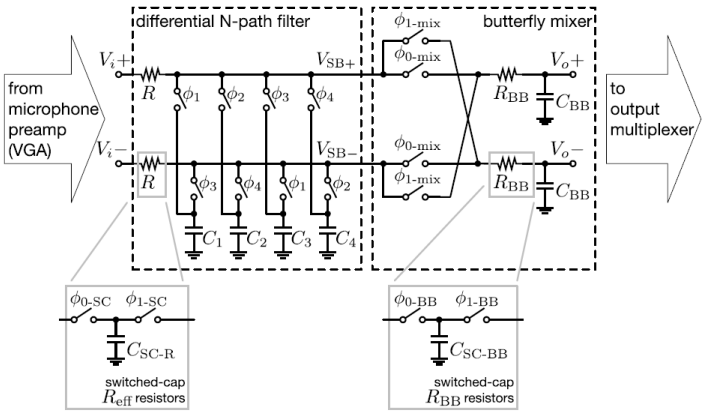


Figure 3.13: Single channel schematic. N-Path bandpass topology used for subband filtering followed by a butterfly differential mixer using a low-pass filter load for demodulation. *Source: Villamizar et al., TCAS1 '21 [17]*

The work also presents a software model of the analog circuits for Machine

Learning dataset processing. Without such a model, the simulation time for running transient simulations would be too high and we wouldn't be able to generate a sufficiently large training dataset.

3.5 Energy-Quality Scaling

This section is focused the work presented by Jinq Horng Teo et al. in TCAS1 '20 [12] which investigates system-level Energy-Quality (EQ) scaling in VAD systems. EQ scaling is a design dimension to minimize the energy under a given detection quality target. It exploits the observation that sensing/processing quality degradation is typically acceptable in noise-resilient applications, compared to the most demanding tasks, contexts, and input datasets. In this work, the authors pursue low-energy voice activity detection through end-to-end insertion of EQ knobs along the signal chain, and system-level simultaneous co-optimization of such knobs to minimize the energy under a given detection quality target. EQ knobs are inserted from the sensor interface in the form of analog bias current and data converter resolution, to machine learning-based classification in the form of decision tree node count (i.e. ML model size). Figure 3.14 shows the proposed VAD architecture with EQ knobs.

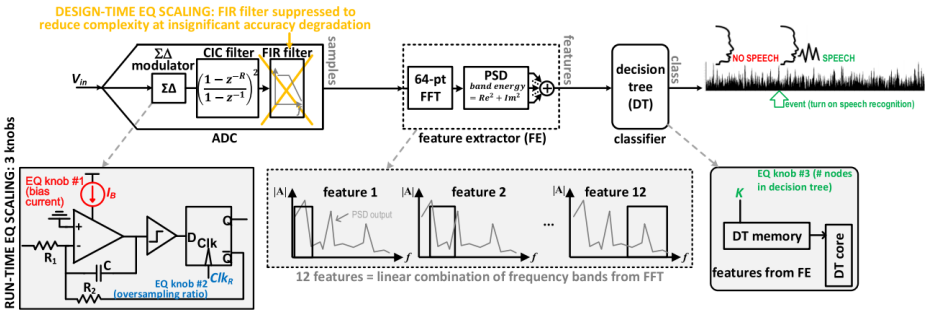


Figure 3.14: Proposed Energy-Quality scalable VAD architecture with three run-time knobs and one design-time knob. *Source: Jinq Horng Teo et al., TCAS1 '20 [12]*

This architecture has three run-time EQ knobs and one design-time EQ knob. Design-time EQ scaling was applied at architectural level by eliminating the finite impulse response (FIR) low-pass filter that is routinely found after the cascaded integrator comb (CIC) filter in ΣΔ modulators. Omitting the FIR filter at design time enabled significant energy saving (50%)

at insignificant quality degradation with or without retraining (1.15% and below). The three run-time EQ knobs are:

1. Bias current I_B in the OTA used in the integrator of ΣΔ modulator loop
2. Oversampling ratio (and thus resolution) of ΣΔ modulator
3. Number of Decision Tree nodes in the classifier

The energy-quality sensitivity is an effective metric that quantifies both the gracefulness of quality degradation and the potential for energy savings. The EQ sensitivity pertaining to a given knob X around a given operating point (E,Q) is defined as the sensitivity of quality with respect to the energy when X is adjusted, according to

$$S_E^Q\Big|_X = \frac{\partial Q}{\partial E} \cdot \frac{E}{Q}$$

Values of EQ sensitivity lower than one pertain to effective EQ knobs that have graceful degradation, as the energy quality is more pronounced than the quality reduction. Values of EQ sensitivity higher than one instead refer to EQ knobs that are less effective, as the energy benefit is smaller than the quality degradation.

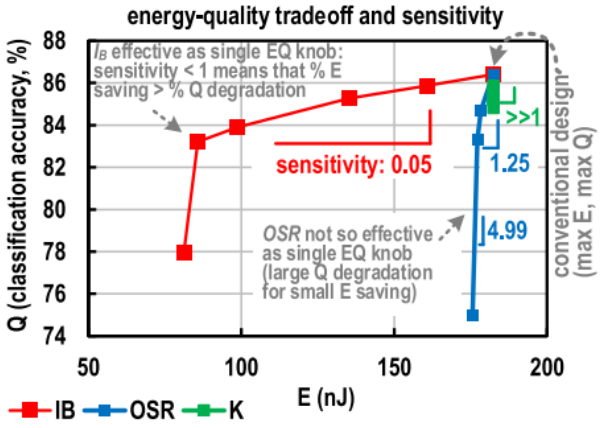


Figure 3.15: Energy-quality plots under individual EQ knob tuning (voice activity detection under a 10-dB noise environment). *Source: Jinq Horng Teo et al., TCAS1 '20 [12]*

Figure 3.15 shows the effect of EQ scaling for the different run-time knobs mentioned before. Even though K is not an effective individual knob, Figure

3.16 shows that K needs to be tuned to avoid overfitting/underfitting and related ungraceful quality degradation.

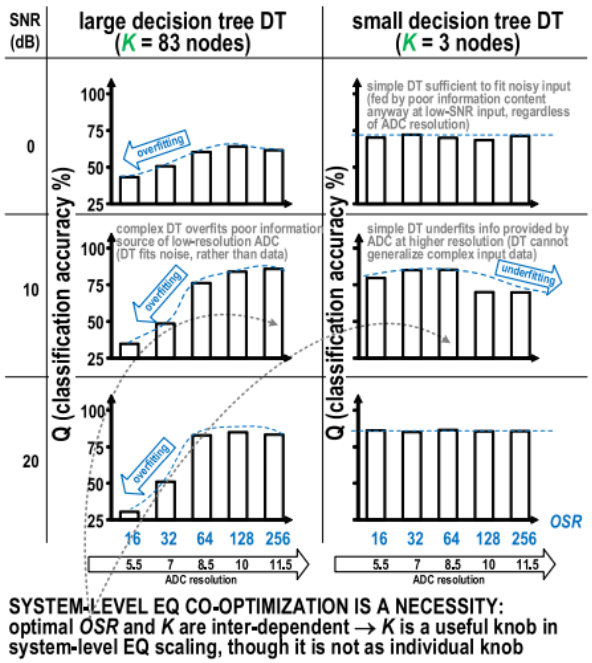


Figure 3.16: The effect of tuning OSR on classification accuracy at different K. *Source: Jinq Horng Teo et al., TCAS1 '20 [12]*

3.6 Fully Analog VAD

This section is focused on the work presented by Marco Croce et al. in JSSC '21 [13] which demonstrates an end-to-end Analog VAD based on signal-to-noise ratio. The overall chain is composed of a programmable-gain amplifier (PGA), a squarer, an integrator, an SC-based signal averaging circuit, and a periodic threshold update circuit for adaptability. The block diagram and schematic of the proposed analog VAD are shown in Figures 3.17 and 3.18 respectively.

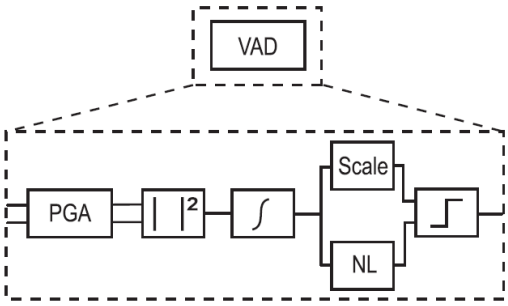


Figure 3.17: Detailed block diagram of the prosed analog VAD. NL = Noise Level. *Source: Marco Croce et al., JSSC '21 [13]*

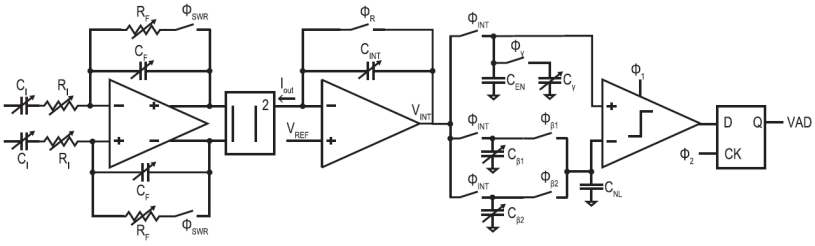


Figure 3.18: Schematic of the prosed analog VAD. *Source: Marco Croce et al., JSSC '21 [13]*

This work, although seems to perform well under most circumstances, can fail in various situations in my opinion. Since this work uses an SNR-based Decision Rule, it won't be able to distinguish speech signal from a high amplitude non-speech signal such as clapping or knocking. Moreover, since the noise level adapts to the background sound, it is possible that when a continuous speech signal is present for a very long duration, the system starts to classify it as noise.

3.7 Comparison Table

	TCAS1 '21 [17]	**JSSC '21 [13]**	**TCAS1 '20 [12]**	**JSSC '19 [8]**	**JSSC '19 [10]**	**JSSC '16 [4]**
Technology	130nm	180nm	28nm	180nm	180nm	90nm
Band (Hz)	30-8k	300-6.8k	NA	0-4k	100-5k	75-5k
Feature	Analog	Analog	Digital	Analog	Events	Analog
Feature Extraction Method	SC-BPF, SC-Mix	square, integrate	FFT	SC-Mix, LPF, DSP	gmC, FWR, IAF	gmC, FWR, LPF
Classifier	SVM/NN	SNR	DT	NN	NN	DT
Power (nW)	6200	760	6490	142	380	6000
Dataset	Proprietary	Proprietary	Proprietary	LibriSpeech + NOISEX-92	Aurora4 w/ DEMAND	NOISEUS
Function	KWS	VAD	VAD	VAD	VAD	VAD
Accuracy	92.4%	99.5%	87.3%	91.5%	85%	89%
Latency (ms)	26	32	8	512	10	< 100

Table 3.1: Comparison of SotA acoustic sensing chips. Accuracy is computed over different datasets and therefore fair comparison is difficult. Power consumption values are for entire system and not just analog frontend.

Chapter 4

Proposed VAD Architecture

System level simulations in MATLAB with digital Mel-Filtering showed that good accuracy for speech vs non-speech classification can be achieved even with just 12 filter channels and a fairly simple ML model. Using overlapping frames was necessary to prevent loss of information as well as to increase throughput of the filterbank. Incorporating multiple contextual neighboring frames also helps in improving the classification accuracy [3],[10]. Based on this assessment and after analysing previous VAD implementations, we propose the following VAD architecture.

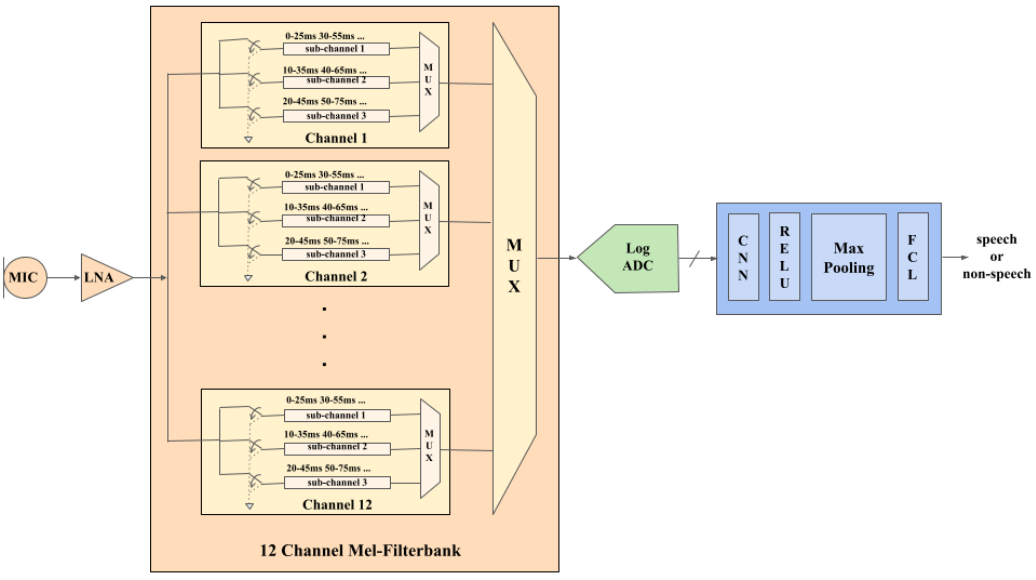


Figure 4.1: Proposed VAD architecture. The orange, green and blue blocks are implemented in analog, mixed-signal and digital domains respectively.

4.1 System Description

Figure 4.1 shows the system level block diagram of the proposed VAD architecture. The microphone converts the audio input into an electrical signal. The Low Noise Amplifier (LNA) amplifies this signal to sufficiently high amplitude for further processing. The output of LNA goes into 12 different Filterbank channels. Every channel has 3 sub-channels which have the same circuitry but are time-multiplexed as shown in Figure 4.2. Every sub-channel is made up of a Band-Pass Filter, a Mixer and a Low-Pass Filter. The center frequencies and quality factors of the band-pass filters are adjusted to match the specification of Mel-Filters. The DC output of every channel is again time-multiplexed at a rate such that all outputs are sampled within the 10ms time window. The serial outputs of the 12:1 Multiplexer are converted to fixed-point Digital values using a log ADC and stored in a buffer to construct a spectrogram. The run-time generated spectrogram is used as the input to a Neural Network based classifier. All the digital processing is done using fixed-point numeric representation. There are only 4-layers in the ML model: 1 Convolutional Neural Network (CNN) Layer, 1 Rectified Linear Unit (ReLU) layer, 1 Max-Pooling layer and finally a Fully Connected Layer (FCL). The classier produces a binary output indicating speech or non-speech every 100ms.

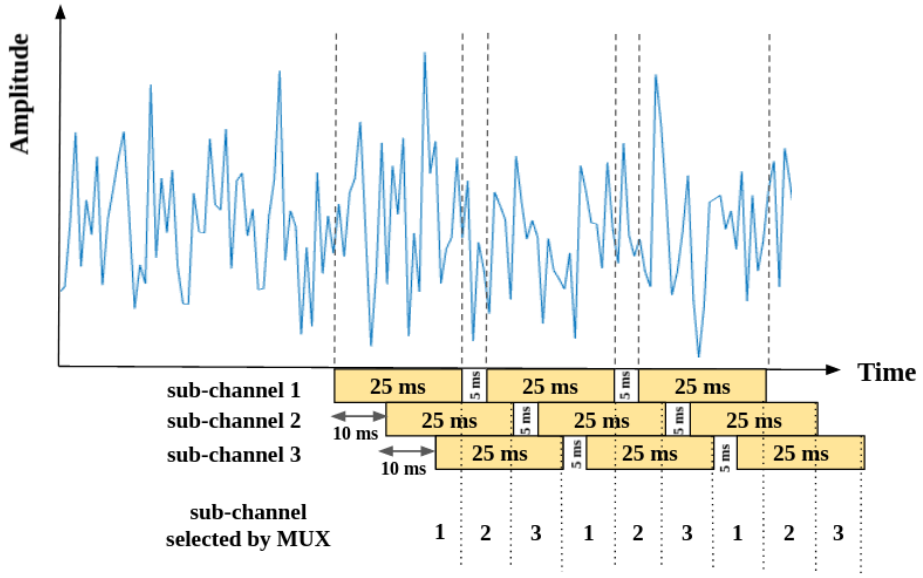


Figure 4.2: Description of how sub-channels process the input signal in different frames. Due to such arrangement we get the throughput of the Filterbank as 10ms although the frame length is 25ms.

4.2 Circuit Description

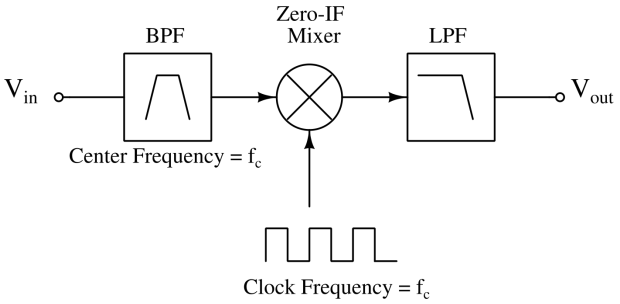


Figure 4.3: Block diagram of a sub-channel

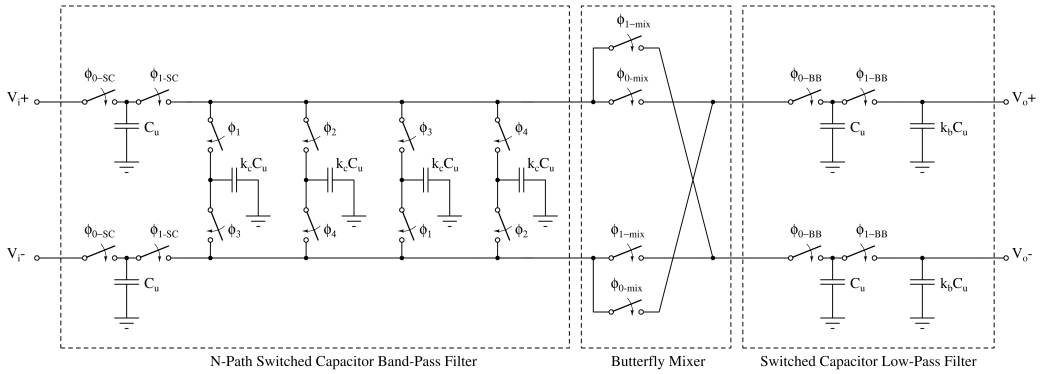


Figure 4.4: Schematic of a sub-channel. Redrawn from original source: Villamizar et al., TCAS1 '21 [17]

The exact same circuit as given in [17] was used for implementing the sub-channels of the Filterbank. The circuit has three blocks:

1. N-Path Switched Capacitor Band-Pass Filter: It's center frequency corresponds to the period of ϕ_1, ϕ_2, ϕ_3 and ϕ_4. It's quality factor can be tuned by varying the parameter k_c.

2. Butterfly Mixer: It mixes the bandpass signal with a clock of same frequency for down-conversion

3. Switched Capacitor Low-Pass Filter: It computes the average value of the filtered signal in a time frame and gives a DC output.

4.3 Specifications for Band-Pass Filters

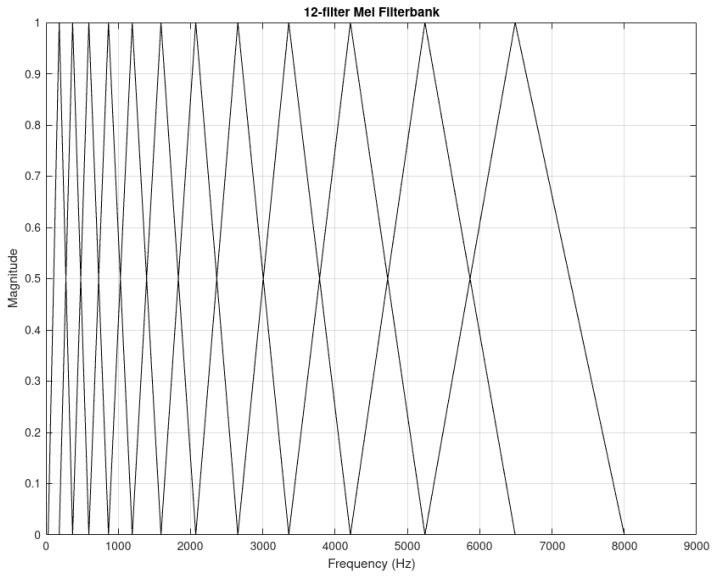


Figure 4.5: Ideal 12-channel Mel-Filterbank

A raange of 30Hz to 8KHz was chosen for the filterbank. Based on the ideal Filterbank response shown in Figure 4.5, the required approximate center frequencies and quality factors were determined and k_c was chosen accordingly. The Table 4.1 gives these specifications.

Center Frequency, f_c (Hz)	**Bandwidth, BW (Hz)**	**Quality Factor, $Q = \frac{f_c}{BW}$**
180	100	1.8
360	120	3
600	145	4.14
860	176	4.89
1200	213	5.64
1600	257	6.22
2070	311	6.65
2650	377	7.03
3360	456	7.37
4200	552	7.61
5240	668	7.85
6500	808	8.05

Table 4.1: Specifications for Band-Pass Filters in the Mel-Filterbank

Chapter 5

Circuit Implementation and Simulation Results

The Mel-Filterbank in the architecture shown in 4.1 was implemented in Cadence. The circuit shown in Figure 4.4 was implemented at the transistor level, while the and Analog MUXes were implemented in Verilog-A. Ideal clocks with appropriate rise and fall times were used. The circuit parameters chosen for different channels are given in Table 5.1.

	$\mathbf{f_{mix}}$ (**KHz**)	$\mathbf{C_{u,BPF}}$ (**pF**)	$\mathbf{k_c}$	$\mathbf{C_{u,LPF}}$ (**pF**)	$\mathbf{k_b}$	$\mathbf{f_{BB}}$ (**KHz**)
Channel 1	0.18	10	1	1	8	6
Channel 2	0.36	5	1.5	1	8	6
Channel 3	0.6	4	2.25	1	8	6
Channel 4	0.86	3.5	2.5	1	8	6
Channel 5	1.2	3	3	1	8	6
Channel 6	1.6	2.5	3.25	1	8	6
Channel 7	2.07	2	3.5	1	8	6
Channel 8	2.65	1.5	3.75	1	8	6
Channel 9	3.36	1.5	4	1	8	6
Channel 10	4.2	1.5	4	1	8	6
Channel 11	5.24	1	4.25	1	8	6
Channel 12	6.5	1	4.5	1	8	6

Table 5.1: Parameters chosen for 12-channel Mel-Filterbank. NMOS switch of same size were used everywhere with $\frac{W}{L} = \frac{400n}{65n}$.

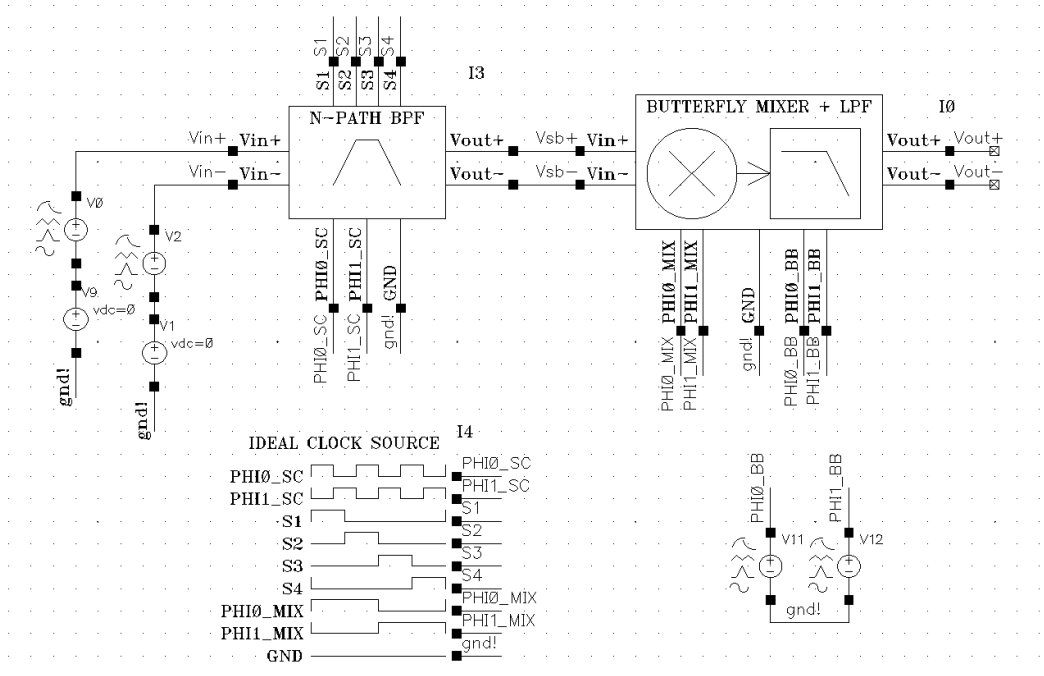


Figure 5.1: Testbench for single sub-channel of the Mel-Filterbank

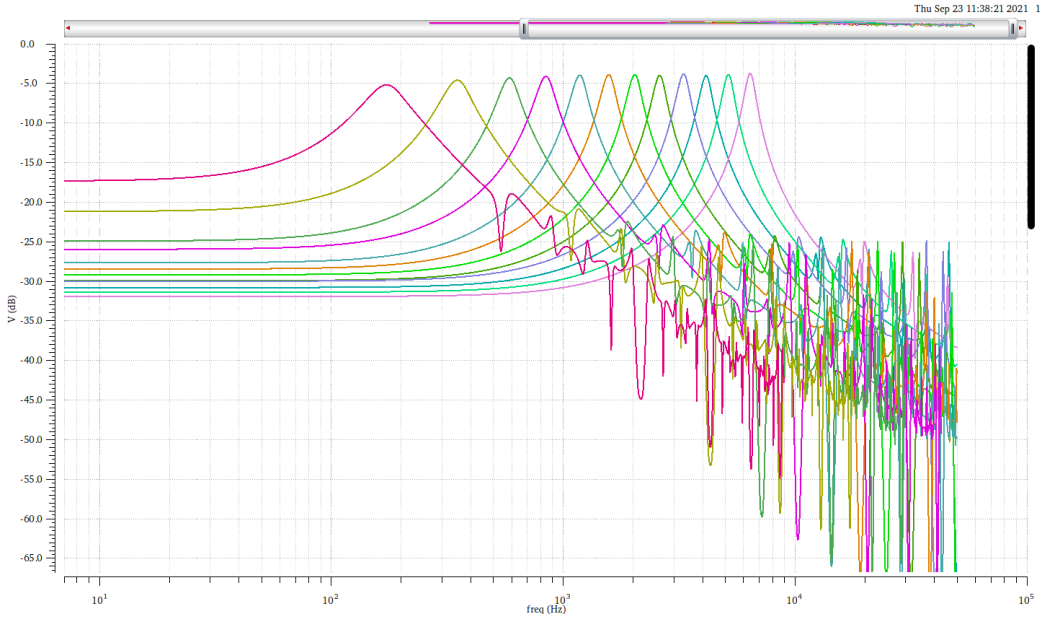


Figure 5.2: Spectre Periodic AC Analysis of the 12-channel Mel-Filterbank

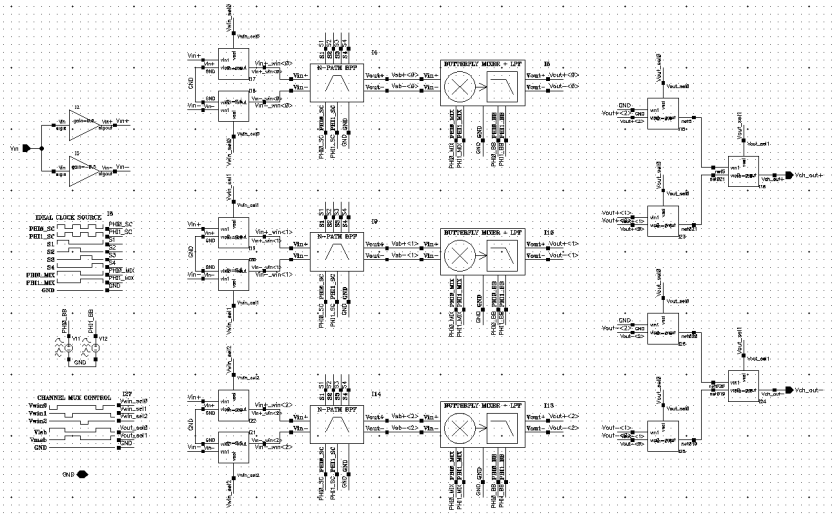


Figure 5.3: Schematic for single channel in Mel-Filterbank

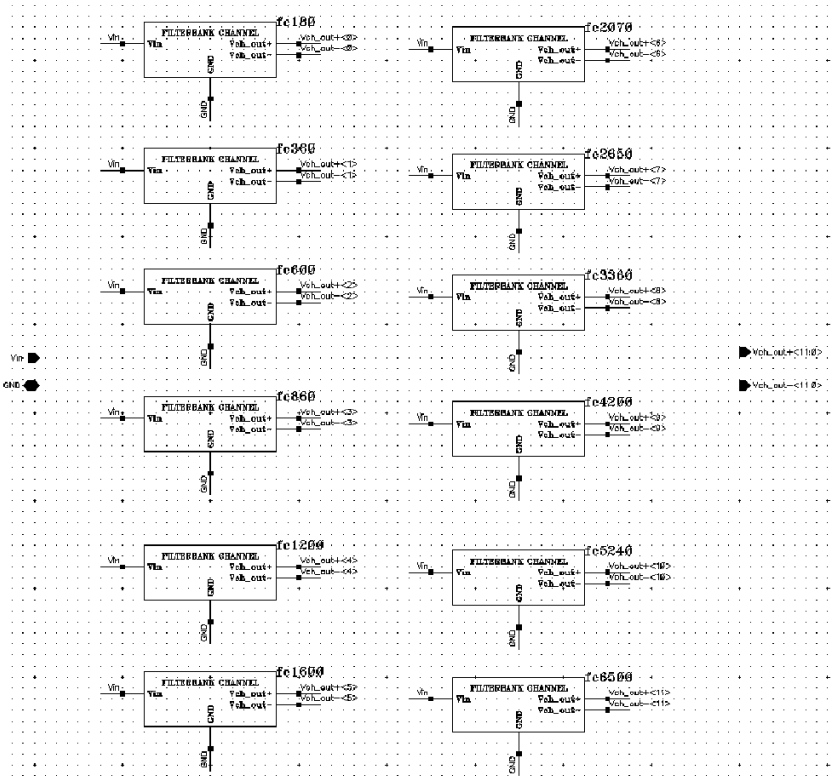
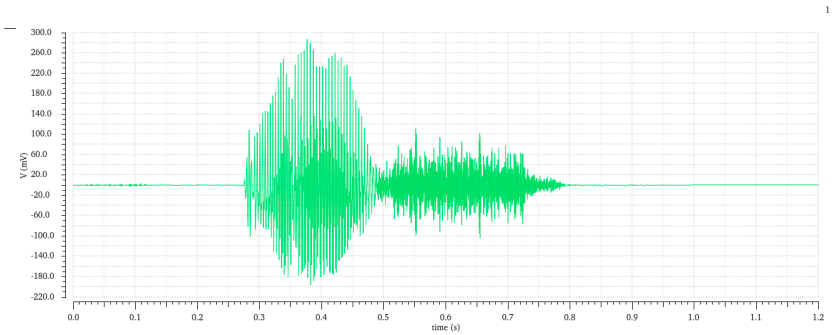
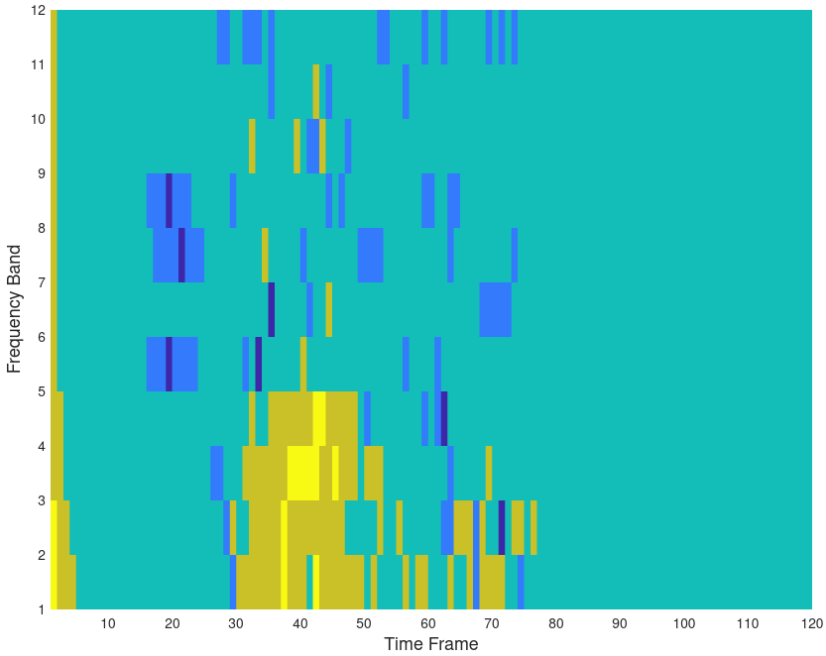


Figure 5.4: Schematic for 12-channel Mel-Filterbank



(a) Speech input imported through a WAV file



(b) Spectrogram generated using simulation data of circuit implementation

Figure 5.5: Results of transient simulation with a test input

After looking at the simulation results an issue was identified - the output of the Filterbank was dependent on the input phase. Since VADs typically have to be always-on, the feature set corresponding to a particular speech signal should not be significantly affected by when the speech signal arrives at the microphone. In other words, we would like our VAD to approximate a time invariant system. However, this property doesn't seem to hold in this implementation.

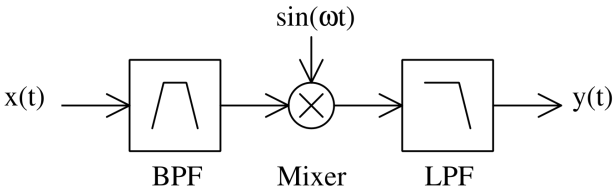


Figure 5.6: Simplified block diagram of filterbank sub-channel

For quick analysis, let's assume all the blocks are ideal and ignore the effects like harmonic mixing. Consider the following two cases :

Case 1:

ω is in passband and during a particular time frame $x(t) = A.sin(\omega t)$

$$\implies x(t) \times sin(\omega t) = \frac{A.(1 - cos(2\omega t))}{2} \implies y(t) = \frac{A}{2}$$

Case 2:

ω is in passband and during a particular time frame $x(t) = A.sin(\omega(t - t_0))$

$$\implies x(t) \times sin(\omega t) = \frac{A.(cos(\omega t_0) - cos(2\omega t))}{2} \implies y(t) = \frac{A}{2}.cos(\omega t_0)$$

Therefore, just delaying the input by t_0 changes the output of LPF by a factor of $cos(\omega t_0)$. Moreover, $cos(\omega t_0)$ can be any value between -1 and 1, and therefore can't be ignored. Here are the results of a transient simulation for my implementation of a single channel of the Filterbank.

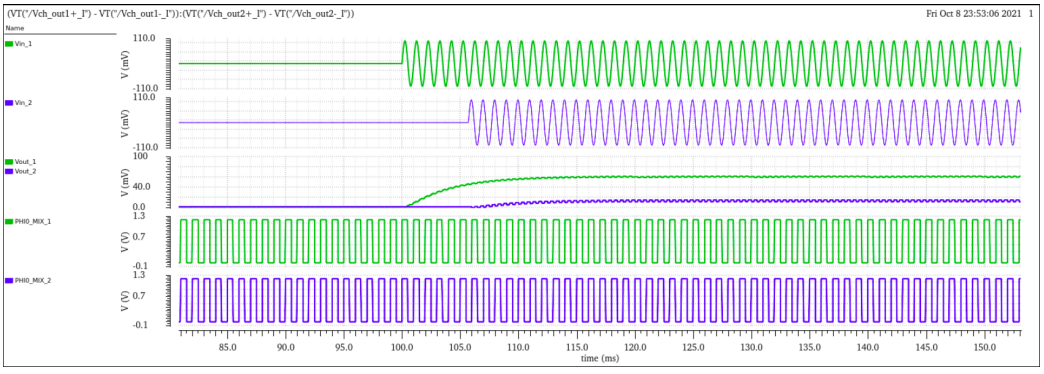


Figure 5.7: Effect of input phase on filterbank channel output

The ratio of DC outputs for the two cases also matched with cosine of the phase difference between the input signals. Moreover, negative DC outputs were also observed in other tests. While this analysis is for a single frequency tone, it can be extended to any arbitrary input signal by using Fourier transform. Thus further analysis of the current architecture is necessary.

Chapter 6

Conclusion & Future Work

A VAD architecture was proposed based on previous SotA implementations. The 12-channel Mel-Filterbank was tested in Cadence using Periodic AC and Transient analyses with transistor level implementation of Band-Pass Filters, Mixers and Low-Pass Filters. Spectrograms were derived from simulation data for custom audio inputs. An issue of input phase dependence was identified which would require further analysis.

Future work includes:

1. Optimizing the N-Path Filters to reduce the capacitor values
2. Exploring rectifier-based architectures for Filterbank and comparing the results with mixer-based implementation
3. Building a software model of the Analog Frontend for generation of data required for training the Machine Learning classifier
4. Implementing the Frontend using discrete components on a PCB and interface with the digital classifier running on an FPGA
5. Exploring efficient in-memory computing or other mixed-signal architectures for running machine learning algorithms

Chapter 7

References

[1] S. Imai. "Cepstral analysis synthesis on the mel frequency scale". In: *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 8. 1983, pp. 93–96. DOI: 10.1109/ICASSP.1983.1172250.

[2] Md. Sahidullah and Goutam Saha. "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". In: *Speech Communication* 54.4 (2012), pp. 543–565. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2011.11.004. URL: https://www.sciencedirect.com/science/article/pii/S0167639311001622.

[3] Arijit Raychowdhury et al. "A 2.3 nJ/Frame Voice Activity Detector-Based Audio Front-End for Context-Aware System-On-Chip Applications in 32-nm CMOS". In: *IEEE Journal of Solid-State Circuits* 48.8 (2013), pp. 1963–1969. DOI: 10.1109/JSSC.2013.2258827.

[4] Komail M. H. Badami et al. "A 90 nm CMOS, 6 μW Power-Proportional Acoustic Sensing Frontend for Voice Activity Detection". In: *IEEE Journal of Solid-State Circuits* 51.1 (2016), pp. 291–302. DOI: 10.1109/JSSC.2015.2487276.

[5] Bram Nauta. *ISSCC Videos: N-Path Filters*. 2017. URL: https://www.youtube.com/watch?v=MP7m5OjXWUg&t=1s.

[6] Seokhyeon Jeong et al. "Always-On 12-nW Acoustic Sensing and Object Recognition Microsystem for Unattended Ground Sensor Nodes". In: *IEEE Journal of Solid-State Circuits* 53.1 (2018), pp. 261–274. DOI: 10.1109/JSSC.2017.2728787.

[7] Michael Price, James Glass, and Anantha P. Chandrakasan. "A Low-Power Speech Recognizer and Voice Activity Detector Using Deep Neural Networks". In: *IEEE Journal of Solid-State Circuits* 53.1 (2018), pp. 66–75. DOI: 10.1109/JSSC.2017.2752838.

[8] Sechang Oh et al. "An Acoustic Signal Processing Chip With 142-nW Voice Activity Detection Using Mixer-Based Sequential Frequency Scanning and Neural Network Classification". In: *IEEE Journal of Solid-State Circuits* 54.11 (2019), pp. 3005–3016. DOI: 10.1109/JSSC.2019.2936756.

[9] Daniel Villamizar et al. "Sound Classification using Summary Statistics and N-Path Filtering". In: *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2019, pp. 1–5. DOI: 10.1109/ISCAS.2019.8702364.

[10] Minhao Yang et al. "Design of an Always-On Deep Neural Network-Based 1-μW Voice Activity Detector Aided With a Customized Software Model for Analog Feature Extraction". In: *IEEE Journal of Solid-State Circuits* 54.6 (2019), pp. 1764–1777. DOI: 10.1109/JSSC.2019.2894360.

[11] Juan Sebastian P. Giraldo et al. "Vocell: A 65-nm Speech-Triggered Wake-Up SoC for 10-μW Keyword Spotting and Speaker Verification". In: *IEEE Journal of Solid-State Circuits* 55.4 (2020), pp. 868–878. DOI: 10.1109/JSSC.2020.2968800.

[12] Jinq Horng Teo, Shuai Cheng, and Massimo Alioto. "Low-Energy Voice Activity Detection via Energy-Quality Scaling From Data Conversion to Machine Learning". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 67.4 (2020), pp. 1378–1388. DOI: 10.1109/TCSI.2019.2960843.

[13] Marco Croce et al. "A 760-nW, 180-nm CMOS Fully Analog Voice Activity Detection System for Domestic Environment". In: *IEEE Journal of Solid-State Circuits* 56.3 (2021), pp. 778–787. DOI: 10.1109/JSSC.2020.3038253.

[14] Hassan Dbouk et al. "A 0.44-μJ/dec, 39.9-μs/dec, Recurrent Attention In-Memory Processor for Keyword Spotting". In: *IEEE Journal of Solid-State Circuits* 56.7 (2021), pp. 2234–2244. DOI: 10.1109/JSSC.2020.3029586.

[15] Boris Murmann. "Mixed-Signal Computing for Deep Neural Network Inference". In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 29.1 (2021), pp. 3–13. DOI: 10.1109/TVLSI.2020.3020286.

[16] Weiwei Shan et al. "A 510-nW Wake-Up Keyword-Spotting Chip Using Serial-FFT-Based MFCC and Binarized Depthwise Separable CNN in 28-nm CMOS". In: *IEEE Journal of Solid-State Circuits* 56.1 (2021), pp. 151–164. DOI: 10.1109/JSSC.2020.3029097.

[17] Daniel Augusto Villamizar et al. "An 800 nW Switched-Capacitor Feature Extraction Filterbank for Sound Classification". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 68.4 (2021), pp. 1578–1588. DOI: 10.1109/TCSI.2020.3047035.

[18] Minhao Yang et al. "Nanowatt Acoustic Inference Sensing Exploiting Nonlinear Analog Feature Extraction". In: *IEEE Journal of Solid-State Circuits* 56.10 (2021), pp. 3123–3133. DOI: 10.1109/JSSC.2021.3076344.