Analog Acoustic Feature Extraction and Mixed-Signal / Digital Classification for Voice Activity Detection

Dual Degree Project Stage II 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

June 2022

Dissertation Approval

The dissertation entitled

Analog Acoustic Feature Extraction and Mixed-Signal / Digital Classification for Voice Activity Detection

by

Mihir Kavishwar

is approved for the degree of

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

Prof. Rajesh Zele
Department of Electrical Engineering
(Supervisor)

Prof. Sachin B. Patkar
Department of Electrical Engineering
(Chairperson and Examiner)

Prof. Laxmeesha Somappa
Department of Electrical Engineering
(Examiner)

Date: June 2022
Place: Mumbai.

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)

June 2022

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. I would also like to thank Mr. Prashant Kurrey for the several insightful discussions and all other members of Advanced Integrated Circuits and Systems Laboratory for their valuable suggestions.

Abstract

Analog and mixed-signal processing for edge AI systems has been gaining a lot of traction in recent years due to its promise of achieving higher energy efficiency and performance compared to traditional digital processing approaches. This thesis explores the above idea and presents the circuit implementation of a novel voice activity detection (VAD) system that performs acoustic feature extraction in the analog domain and neural network based classification in mixed-signal and digital domains. The feature extractor is comprised of switched capacitor N-Path bandpass filters, comparator based full wave rectifiers, switched capacitor lowpass filters, non-linear transform circuits and analog multiplexers. The classifier comprises of one convolutional neural network layer, which is implemented within a first order delta-sigma modulator; ReLU, max pooling and fully connected layers, all of which are implemented in the digital domain. All analog and mixed-signal circuit blocks were designed up to transistor level schematics in UMC 65nm technology, while the digital circuits were implemented as functional modules in Verilog-A. System and circuit level simulation results have been discussed, along with challenges encountered, proposed solutions and future work.

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 - cascading a set of tasks with increasing complexity so that the posterior stages are only activated by the previous stages in the pipeline [12]. Only the first stage in the classifier cascade is always-on, thus making the overall system much more energy efficient.

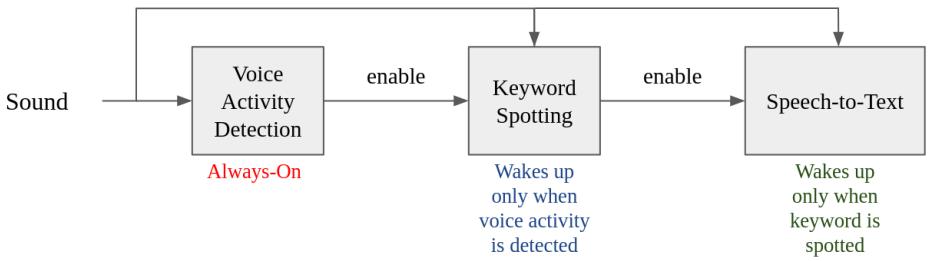


Figure 1.1: Voice activity detection application for keyword spotting and speech-to-text conversion

A voice activity detector (VAD) identifies if the input audio signal is human speech or some other sound. 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 a considerable 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 (and most edge AI systems in general) consist of two main parts [9]:

1. **Feature Extractor -** It converts the high dimensional raw input signal into low-dimensional but dense features. Previous works have extracted acoustic features such as Mel-frequency cepstral coefficients (MFCCs) [8], [12], [15], [18]; input signal energies in different frequency bands [2], [4], [9], [13], [19]; or non-linear spiking events based on band energies [11], [20].

2. **Classifier -** It infers from the feature set input if the signal is: "speech" or "non-speech". Previous works have used different classifier models such as decision trees [4], [13]; support vector machines [6], [10]; or more commonly neural networks [8], [9], [11], [13], [15], [18], [19], [20]. Some works have used an energy thresholding based method [2], [14].

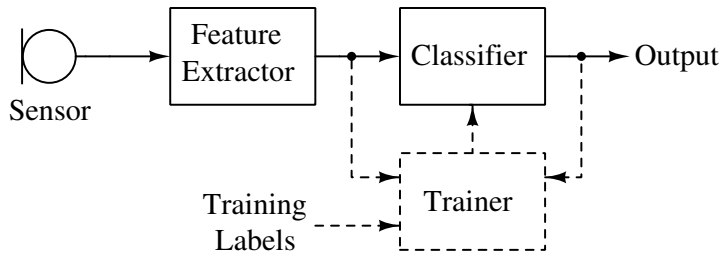
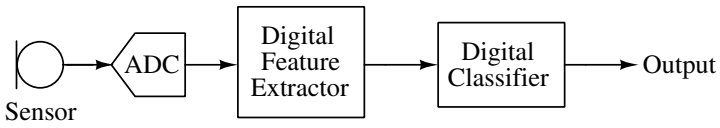


Figure 1.2: Key blocks of typical SotA edge AI systems

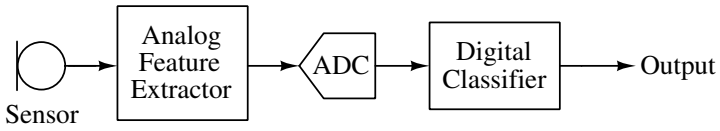
Classification algorithms employing supervised learning have to be trained using labeled datasets to determine the values of the parameters that minimize their loss functions. Since the classifier takes extracted feature data and not the raw sensor data as the input, the training dataset must be generated accordingly. The training can be done offline using a software model of the feature extractor [11], [19], [11], [20] or online once the chip is taped out.

Traditionally, both feature extraction and classification were implemented in the digital domain. This required using a high performance ADC immediately in front of the sensor as shown in 1.3a. Recent works however have

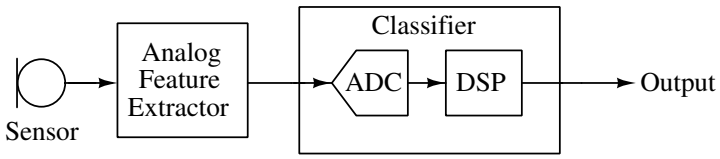
shown that implementing the feature extractor in analog, as shown in 1.3b, makes it very energy efficient, relaxes the specifications for the ADC and relaxes the complexity of our classifier. Further, some of the computation involved in the classifier can be performed within the ADC itself [3] as shown in 1.3c. The classifier typically has to perform several Multiply and Accumulate (MAC) operations and can be made very energy efficient using the latest advancements in mixed-signal computing for neural network inference [17]. The ADC can be avoided in this case as shown in 1.3a.



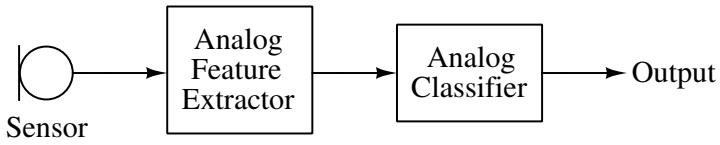
(a) Both feature extraction and classification in digital



(b) Feature extraction in analog and classification in digital



(c) Feature extraction in analog and classification within ADC and in digital



(d) Both feature extraction and classification in analog

Figure 1.3: Different ways in which edge AI systems can be implemented

This work proposes a novel VAD architecture employing analog acoustic feature extraction and multiply-accumulate (MAC) computation technique which implements the CNN operation within a delta-sigma ADC.

Chapter 2

Overview of Relevant Concepts

2.1 Mel Spectrograms

A spectrogram is a visual representation of the frequency spectrum of a signal as it varies over time [22]. Conceptually, to plot the spectrogram of a continuous time signal, we need to perform the following steps:

1. Section the long signal into shorter frames using some windowing scheme
2. Compute the power spectral density (PSD) of the windowed signal and plot it on the Y-axis using different color shades to represent the magnitude
3. Stack the PSDs of successive time frames beside one another on the X-axis

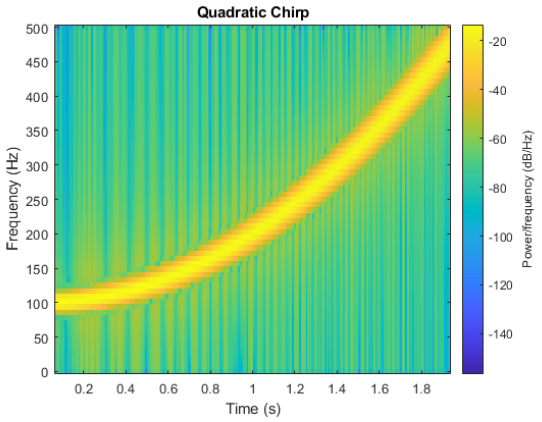


Figure 2.1: Spectrogram example. Source: *Matlab documentation*

The Mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another [21]. The log spectrum on a mel frequency scale (the mel log spectrum) is a more effective representation of the speech signal 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 represented on the mel scale. The formula to convert from f hertz to m mels is:

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

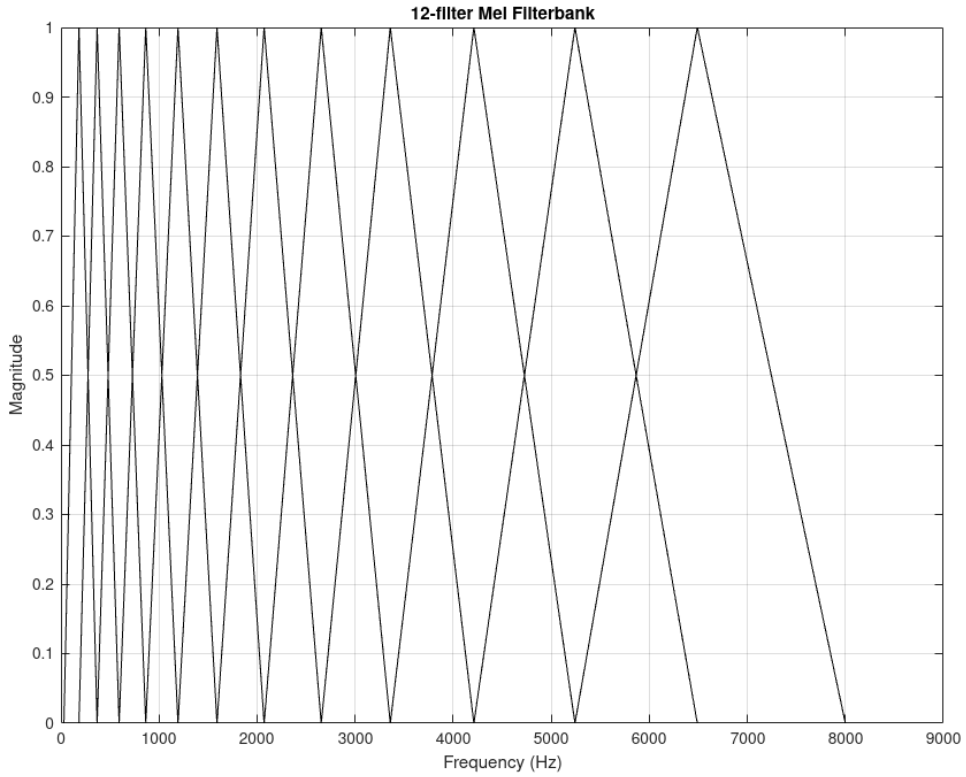
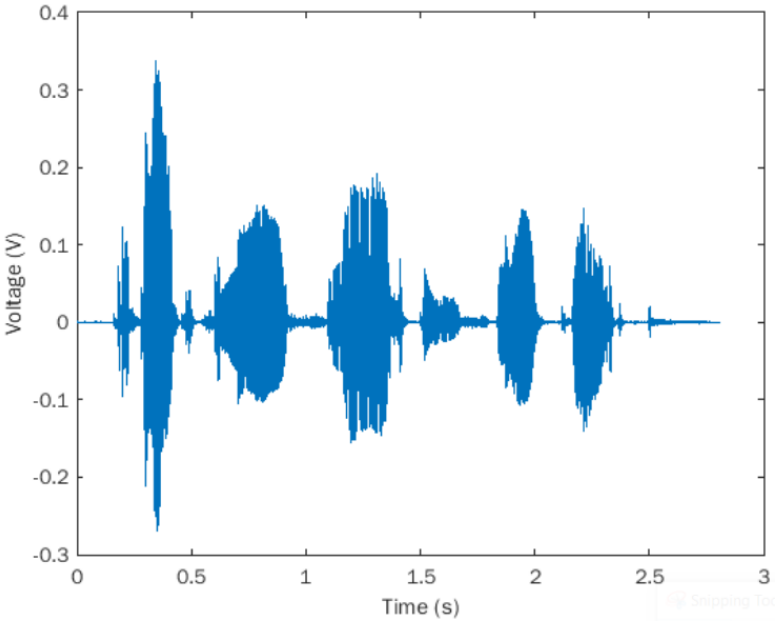
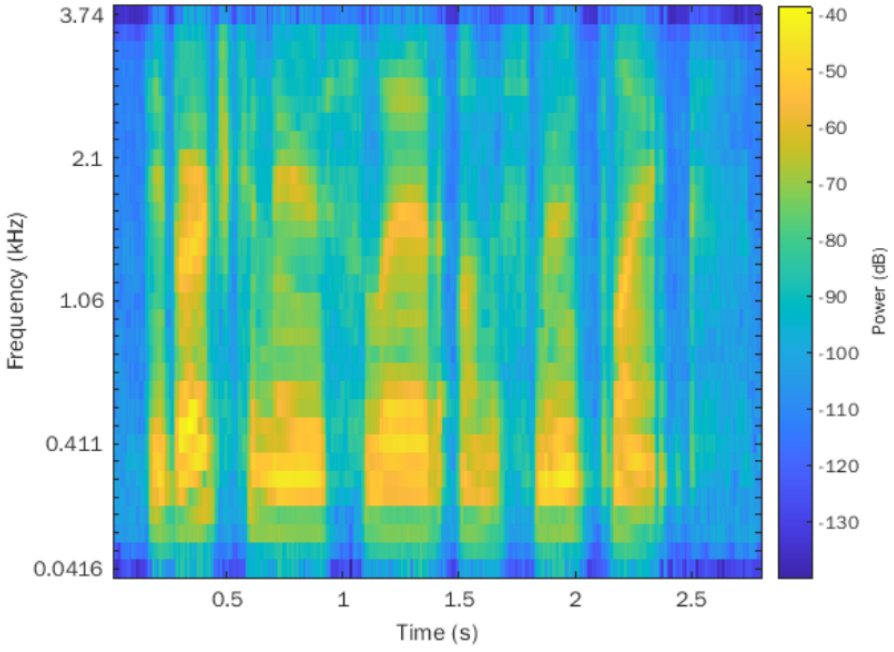


Figure 2.2: 12-filter Mel Filterbank between frequencies 30 Hz to 8 KHz

The frequency bands in a Mel Spectrogram are logarithmically spaced as shown in Figure 2.2 and the transfer functions of filters are triangular on the log scale. Figure 2.3 illustrates how the same audio signal can be represented in two different ways.



(a) Time domain signal



(b) Mel Spectrogram

Figure 2.3: Different representations of the same audio signal

2.2 Neural Network Classifiers

Neural network based classification is one of the most widely used machine learning techniques for applications such as computer vision, speech recognition, healthcare monitoring, etc. Artificial neural networks (ANNs) are made up of tiny computational units called neurons (or perceptrons). A neuron takes multiple inputs and produces a single output based on the network's weights and biases and the neuron's activation function. During the training phase, the weights and biases are updated by the training algorithm to minimize a specific loss function.

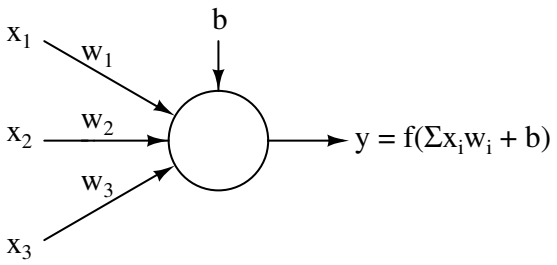


Figure 2.4: Neuron model. f represents the activation function, w_i are the weights and b is the bias.

Examples of activation functions:

1. Sigmoid : $f(x) = \frac{1}{1+e^{-x}}$
2. Rectified linear unit (ReLU) : $f(x) = max(0, x)$
3. Softmax : $f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$

Several neurons are connected together to form a neural network. Different architectures exist which are suitable for different applications -

1. Multi-layer Perceptron (MLP)
2. Convolutional Neural Network (CNN)
3. Recurrent Neural Network (RNN)

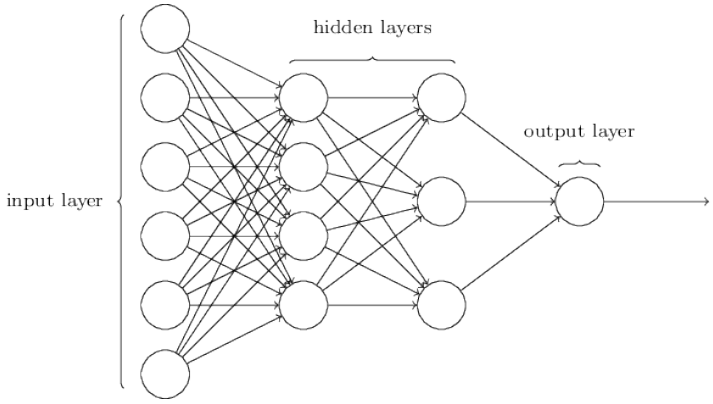


Figure 2.5: Multi-layer perceptron neural network model. Source : Dr. Michael Nielsen's online book [7]

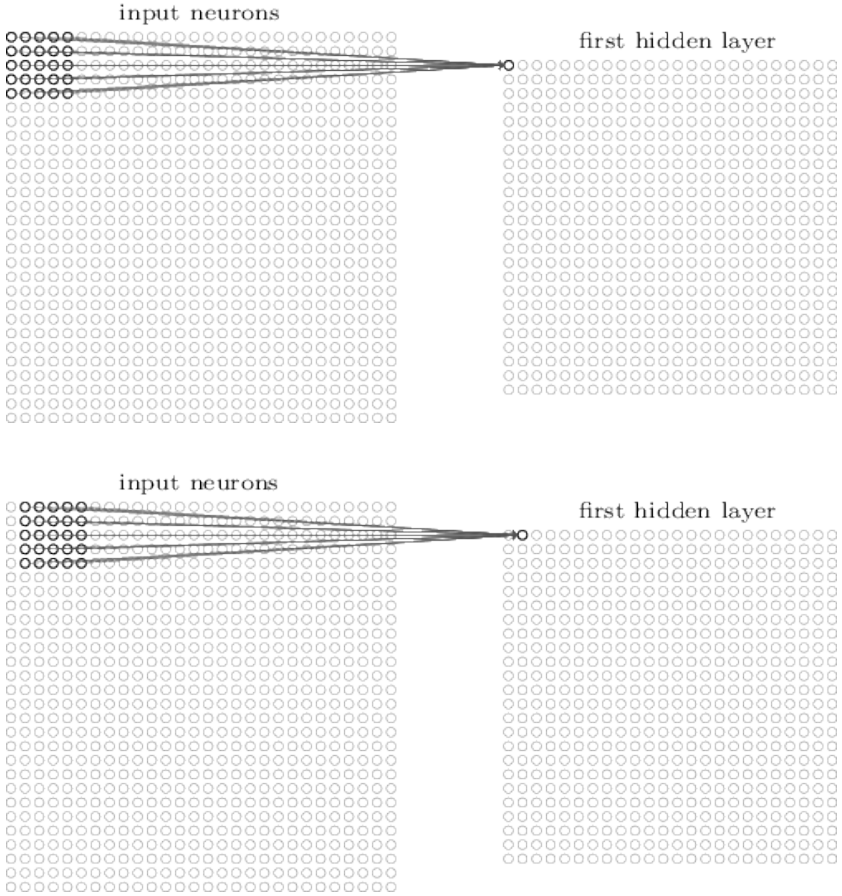


Figure 2.6: Convolutional neural network model. Source : Dr. Michael Nielsen's online book [7]

2.3 N-Path Bandpass Filtering

N-Path filters have been studied in great detail in the past [5]. Figure 2.7 shows the key idea used in N-Path filters.

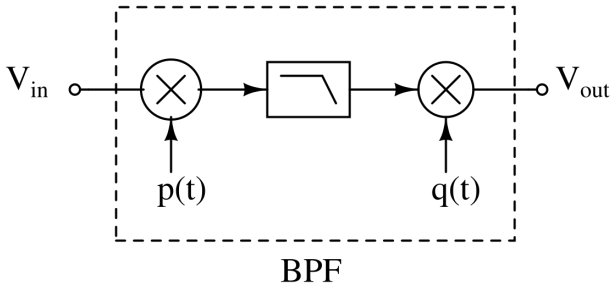


Figure 2.7: The fundamental principle used in an N-Path Bandpass Filter is Downconversion + Lowpass Filtering + Upconversion = Bandpass Filtering

In N-Path filters, the several BPF blocks, as shown in Figure 2.7 are connected in parallel and get cyclically turned on one after another. Figure 2.8 shows the block diagram of the complete N-Path filter.

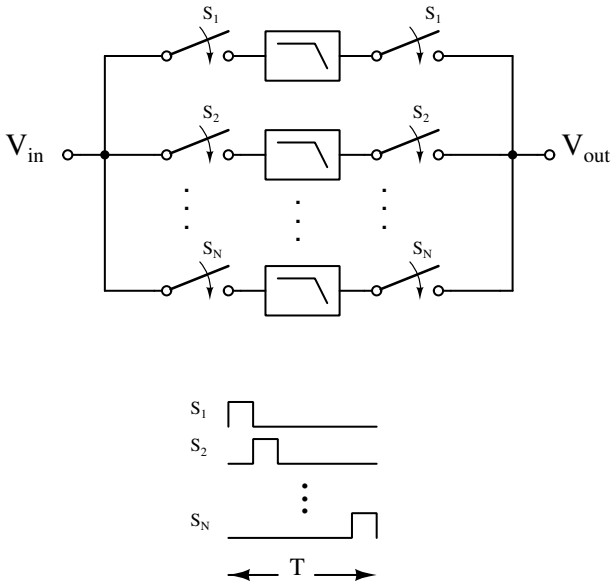


Figure 2.8: N-Path filter block digram

The main disadvantage of the N-Path filter is significant harmonic distortion. The even harmonics can be eliminated by going for a differential

topology. Further, the number of capacitors can be halved by sampling the input on the same capacitor in two different phases. Figure 2.9 shows how we can derive the final differential N-Path filter circuit from the original idea.

The center frequency is same as clock frequency (f_c) while the 3 dB bandwidth is given by $\left(\frac{1}{2\pi NRC}\right)$. 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 because power is required only to drive switches

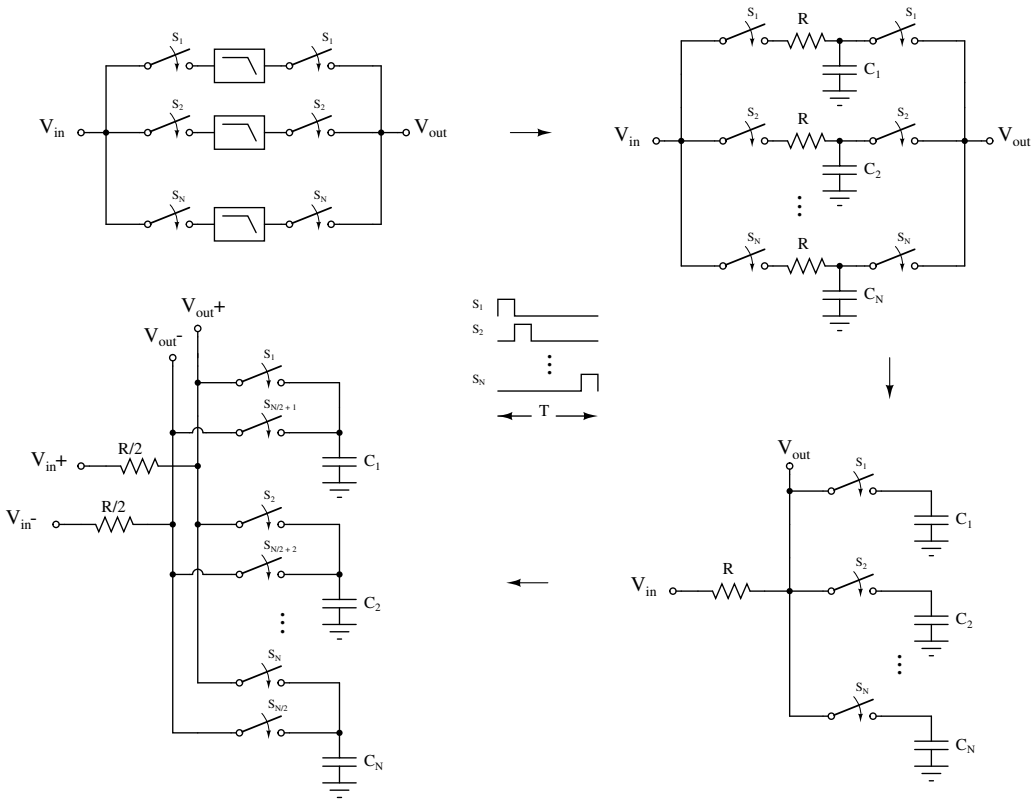


Figure 2.9: Derivation of differential N-Path Bandpass Filter from the original conceptual diagram

2.4 $\Delta\Sigma$ Modulation

$\Delta\Sigma$ modulation is a data conversion technique in which the input signal is sampled at a rate much greater than the Nyquist frequency ($f_s >> 2f_b$). Oversampling reduces the quantization noise power in the signal band. Further, the $\Delta\Sigma$ modulator shapes the quantization noise out of the signal band. $\Delta\Sigma$ analog to digital converters ($\Delta\Sigma$ ADCs) are extremely popular in applications that demand high bit resolution, low area and low power.

Figure 2.10 shows the conceptual block diagram of $\Delta\Sigma$ modulator. A key feature of this scheme is that with a single bit quantizer, we can get multi-bit resolution.

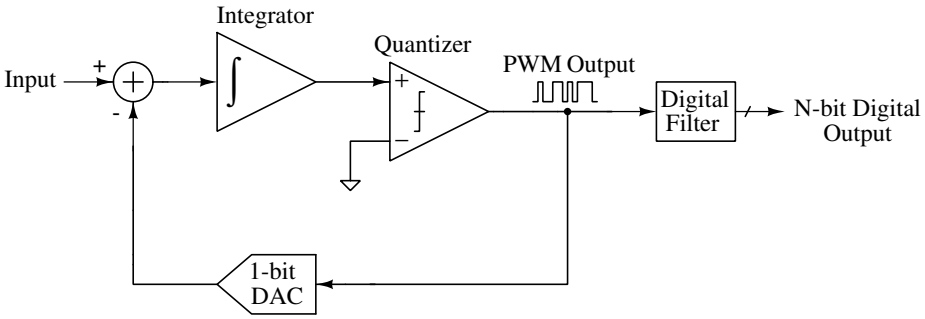


Figure 2.10: First order $\Delta\Sigma$ modulator block diagram

The delta-sigma operation can be realized in discrete time or continuous time. The Z transform as shown in Figure 2.11 is commonly used to analyze discrete time $\Delta\Sigma$ modulators.

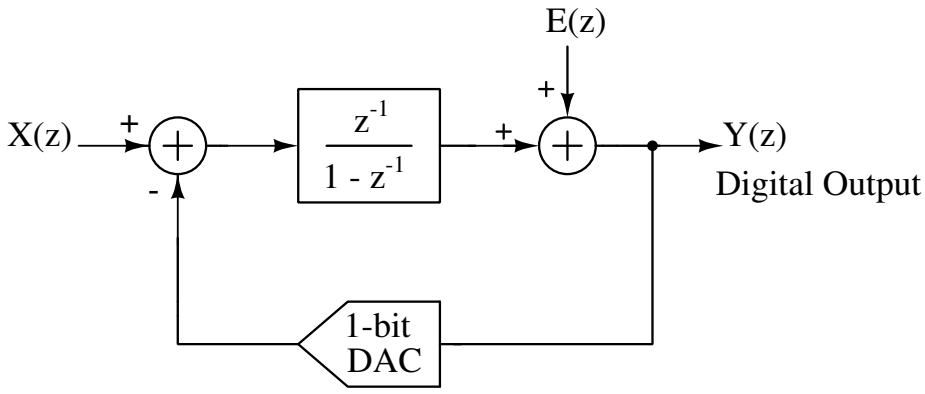


Figure 2.11: First order $\Delta\Sigma$ modulator Z-transform representation

For this architecture:

$$Y(z) = z^{-1}X(z) + (1 - z^{-1})E(z)$$

Signal Transfer Function (STF) $= z^{-1}$
Noise Transfer Function (NTF) $= (1 - z^{-1})$

The circuit realization of the above architecture is shown in Figure 2.12.

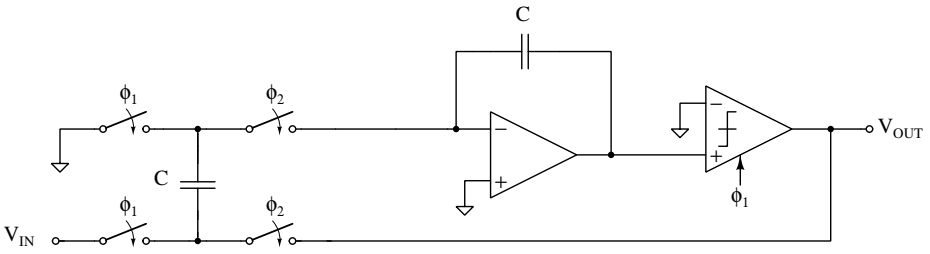


Figure 2.12: First order $\Delta\Sigma$ modulator circuit

Chapter 3

Literature Survey

3.1 Power-Proportional Acoustic Sensing

This section is focused on the work **Badami et al., JSSC '16** [4] in which 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 when more complicated tasks need to be performed .

In this work, the first stage in the signal processing tool chain is a threshold based always-on sound activity detector. On detection of a sound signal, the analog feature extractor wakes up and converts the sound into a set of acoustic features. Some of these features go in to the on-chip classifier which infers if the sound signal is human speech or non-speech. If human speech is detected, the classifers sends a signal to the microcontroller which turns on to perform more complex sensing tasks with the complete feature set.

In the analog feature extractor, the input audio signal is decomposed into a set of 16 features. Mathematically, each analog feature af_i is defined as

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

These features can be turned ON/OFF depending on their utility for a particular sensing task.

In the analog frontend, active gm-C filters are used for both band pass and low pass functions. The rectifier and low pass filters are implemented in

current mode to form an active averaging circuit. The decision tree classifier is implemented in mixed-signal domain and uses a modified C4.5 machine-learning algorithm.

Key idea of the paper : The power consumption of a system can be scaled in proportion to the complexity of the sensing task by cascading blocks in such a way that low complexity blocks wake-up high complexity blocks as per the sensing task.

3.2 Event-based Acoustic Feature Extraction

This section is focused on the work **Minhao Yang et al., JSSC '19 [11] and JSSC '21 [20]** which proposes a VAD architecture based on analog spiking events. The feature extraction happens in the analog domain. A low noise amplifier (LNA) first amplifies the input signal. The output of LNA is fed into 16 different channels, each of which computes the signal energy in some frequency band. Each channel is composed of a bandpass filter (BPF), a full-wave rectifier (FWR) and an event driven ADC (ED-ADC). The work presents a new 2nd-order BPF circuit which is based on super-source-follower (SSF) topology. The ED-ADC encodes the analog information in spiking events which can be measured by a counter. The counter output goes to a neural network classifier implemented in the digital domain. The neural network classifier produces a speech or non-speech output.

The recent paper [20] builds on their previous work and exploits some non-linear properties of analog feature extractor. An additional clipping amplifier is used in the system.

Key idea of the papers : Event-driven analog to digital conversion is useful because it combines the functions of integration and quantization. Non-linearity of analog circuits can be exploited to our benefit in certain applications.

3.3 Mixer-based Sequential Frequency Scanning

This section is focused on the work **Sechang Oh et al., JSSC '19 [9]** which discusses a programmable acoustic signal processing system based on neural network classification. The authors propose using a sequential frequency scanning technique instead of parallel feature extraction like in other works.

This allows them to further reduce the power consumption to sub-μW levels.

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 low-noise aplifier (LNA), programmable amplifier (PGA), ADC driver (DRV), and ADC. The ULP chain consists of an additional mixer between LNA and PGA for down conversion so that the Nyquist rate is reduced for subsequent blocks.

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 itself 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. While the system performs well on energy efficiency metric, it has significantly more latency than other SotA systems due to its sequential processing architecture.

Key idea of the paper : Sequential frequency scanning enables extremely high energy efficiency since it doesn't need a multi-channel filterbank, but there is a trade off with latency.

3.4 Switched Capacitor Feature Extraction Filterbank

This section is focused the work **Villamizar et al., TCAS1 '21 [19]** which demonstrates an application of N-Path switched capacitor bandpass filters for acoustic feature extraction. N-Path filters have been studied in great detail in past [5]. N-Path filters are highly tunable, can provide great quality factor and are extremely energy efficient. 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 compensated by the machine learning model of the classifier, as demonstrated in this paper.

The filterbank has 32 channels and simultaneously process the input audio signal in different frequency bands. In a channel, the output of the bandpass filter is fed into a butterfly mixer and then passed through a lowpass filter to get dc output. The extracted features are used by the classification algorithm. The paper discusses two classification tasks: baby-cry detection and keyword spotting. For baby-cry detection, support vector machine (SVM) algorithm is used. For keyword spotting, recurrent neural network (RNN) algorithm is used.

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.

Key idea of the paper : N-Path switched capacitor bandpass filters are suitable for acoustic sensing applications because of their high tunability and low power consumption. Harmonic responses and folding can be absorbed by machine learning model of the classifier.

3.5 Energy-Quality Scaling

This section is focused the work **Jinq Horng Teo et al., TCAS1 '20 [13]** which investigates the trade off between energy and quality in VAD systems. The paper discusses the use of Energy Quality (EQ) 'knobs' - which are essentially some parameters in the system that can be tuned (either during design-time or run-time) to vary the energy consumption and quality of the system. Examples of EQ knobs in this work are - analog bias current, resolution of the ADC and number of nodes in the decision tree classifier.

The paper also discusses the concept of energy-quality sensitivity. EQ sensitivity is a way to quantify the amount by which the quality of a system suffers and while energy savings increase on varying the EQ knobs. The EQ sensitivity is defined by the expression

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

Therefore, it is desirable to have values of EQ sensitivity greater than 1,

since the energy savings are greater than the quality degradation.

Key idea of the paper : Energy-Quality knobs can be inserted in a system and optimization of these knobs allows us to attain minimum energy consumption at a given accuracy target.

3.6 Fully Analog Voice Activity Detectors

This section is focused on the works **Marco Croce et al. in JSSC '21 [14]** and **Udita Mukherjee et al. in ISCAS '21 [16]** .

In [14], the authors demonstrate an end-to-end Analog VAD based on computation of signal-to-noise ratio. The input signal is first amplified by a programmable-gain amplifier (PGA). It is then squared using a circuit that exploits the square relation between current and gate to source voltage in a mosfet. The squared output is integrated and then averaged using a switched capacitor based circuit. Finally, a thresholding circuit is implemented whose reference voltage is periodically updated to adapt with the background noise.

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.

In [16], the authors demonstrate a fully analog neural network based VAD. The weights of the neural network are quantized to -1, 0 and $+1$. The signal chain consist of a bandpass filter and rectifier bank, 2-layer neural network, lowpass filter and a decision slicer. Although the entire system is implemented in analog, its power consumption is significantly greater than SotA VADs which have digital classifiers.

Key idea of the papers : Complete analog implementation of VAD is possible in energy thresholding based methods or quantized neural network based methods.

3.7 Matrix Multiplication within an ADC

This section is focused on the work **Zhuo Wang et al., in TBCAS '15 [3]** which proposes the idea of performing significant computation required for classification tasks directly within an ADC with negligible energy overhead. The work has novel contributions both on the algorithmic front and the circuit design front.

On the algorithmic front, the paper presents a way to combine linear feature extraction and classification into a single matrix transformation. It also demonstrates the benefits of using adaptive boosting (AdaBoost), which is an algorithm than combines multiple weak classifiers to create a strong classifier, over conventional classifiers like radial basis function (RBF) kernel and support vector machine (SVM).

On the circuits front, the paper presents the implementation a SAR ADC that can perform matrix multiplication with it's analog inputs and digital weights. A passive analog multiplication operation is realised by adding a programmable divider in the feedback path of the SAR ADC. To increase the multiplier range, mixed-signal floating point multiplication is implemented.

The utility of the above ideas is shown for two different applications:

1. Detection of cardiac arrhythmia from an ECG
2. Detection of gender from image pixels

Key idea of the paper : In resource constrained edge AI systems, significant classification computation can be performed in the ADC itself without much additional energy consumption. Algorithm-hardware co-design can be extremely beneficial.

3.8 Comparison Table

	TCAS1 '21 [19]	**JSSC '21 [14]**	**TCAS1 '20 [13]**	**JSSC '19 [9]**	**JSSC '19 [11]**	**JSSC '16 [4]**
Task	KWS	VAD	VAD	VAD	VAD	VAD
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
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 could be achieved with just 12 filter channels and a relatively simple neural network based machine learning model. A rectangular windowing scheme was employed to prevent the loss of information and increase the filterbank's throughput. Incorporating multiple contextual neighboring frames also helps in improving the classification accuracy [2],[11]. Based on this assessment and after studying relevant literature, the following novel VAD architecture is proposed:

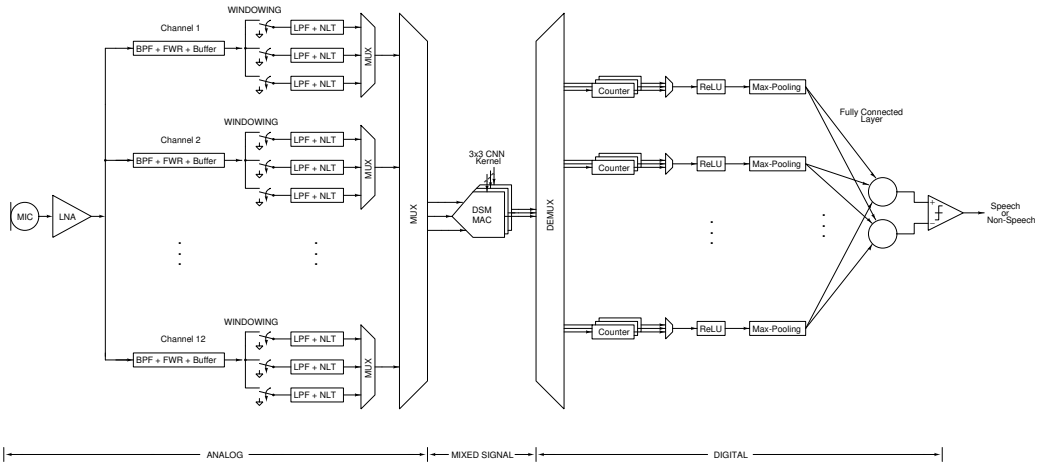


Figure 4.1: Proposed VAD architecture. LNA = low noise amplifier, BPF = bandpass filter, FWR = full wave rectifier, LPF = lowpass filter, NLT = nonlinear transform, MUX = multiplexer, DSM MAC = delta-sigma modulation based multiply and accumulate, ReLU = rectified linear unit

4.1 System Description

The system level block diagram of the proposed VAD architecture is shown in Figure 4.1. The microphone converts the audio input into an electrical signal. The LNA amplifies this signal to a sufficiently high amplitude for further processing (in my design, I have not implemented an LNA since the test input is sufficiently large). The output of LNA goes into 12 different filterbank channels. Every channel has a bandpass filter, a full wave rectifier, a buffer and 3 sub-channels, each of which consists of a low pass filter and a non-linear transform circuit time-multiplexed to implement the windowing scheme shown in Figure 4.2.

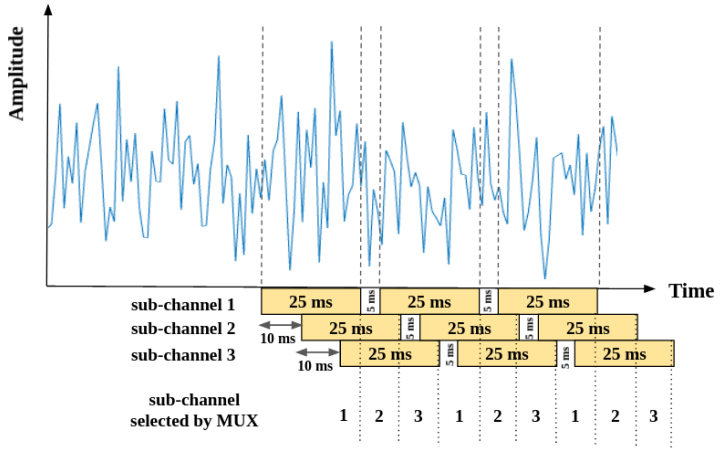


Figure 4.2: Description of how sub-channels process the input signal in different frames. Due to such arrangement, we get the throughput of the Filterbank as 10ms although the frame length is 25ms.

The center frequencies and quality factors of the bandpass filters are adjusted to match the specification of a 12 channel mel-filterbank. Therefore, every channel computes the energy of the input signal in the respective frequency band in a time window. Thus the output of the analog feature extractor produces a spectrogram, with each column vector of dimension 12x1. We use a CNN based machine learning classifier which takes this spectrogram as input and produces a speech vs non-speech decision as shown in Figure 4.3.

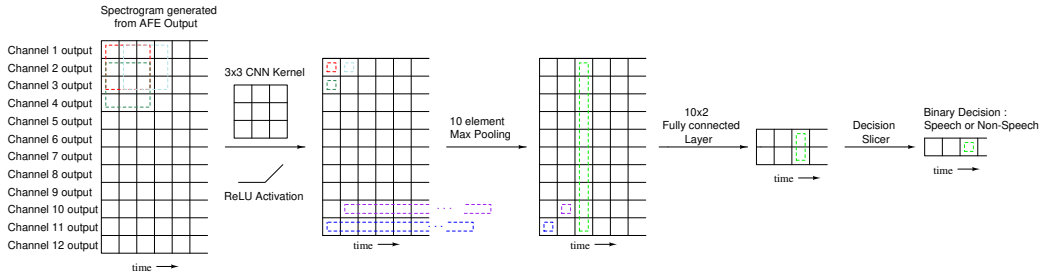


Figure 4.3: Illustration of classification algorithm used in the VAD system

The computation of the CNN layer is directly implemented within a $\Delta\Sigma$ ADC and counters as shown in Figure 4.4.

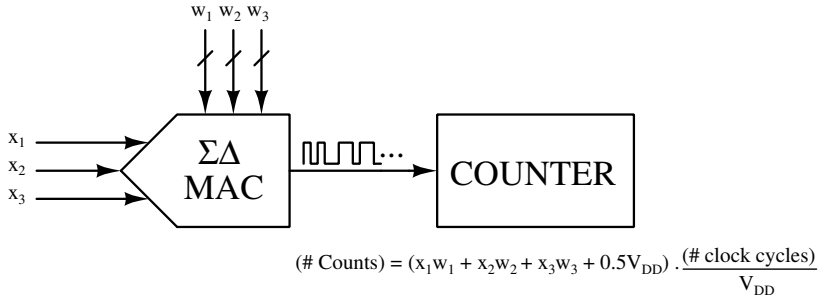


Figure 4.4: Block diagram of MAC computation using $\Delta\Sigma$ modulation

Assuming all input voltage values are with respect to analog ground $= 0.5V_{DD}$ and the logic state voltages are 0 and V_{DD}, the expression for number of counts is given by:

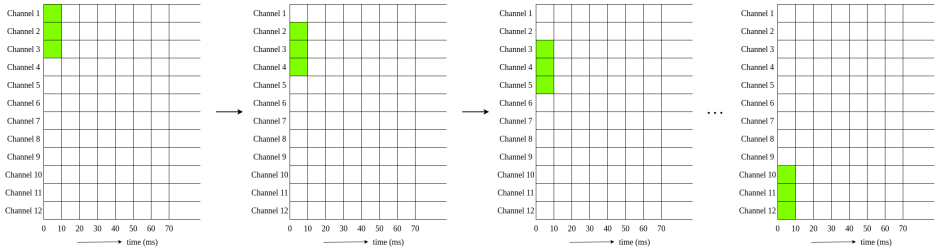
$$\boxed{(\#\text{ Counts}) = \left(\left(\sum_{i=1}^{N=3} x_i w_i\right) + 0.5V_{DD}\right) \times \frac{\#\text{ clock cycles}}{V_{DD}}}$$

$$\implies (\#\text{ Counts}) = k\left(\sum_{i=1}^{N=3} x_i w_i\right) + b$$

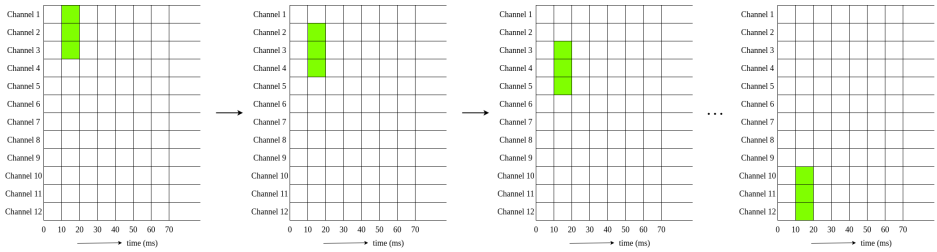
$$\text{where, } k = \frac{\#\text{ clock cycles}}{V_{DD}} \quad , \quad b = 0.5 \times (\#\text{ clock cycles})$$

The bias parameter can be incorporated by changing the reset state of the counter. The counter stores the current state unless it is reset. Thus we can use the counter as an accumulator. When the next computation output is available from the $\Delta\Sigma$ MAC, the counter adds it to the previously stored

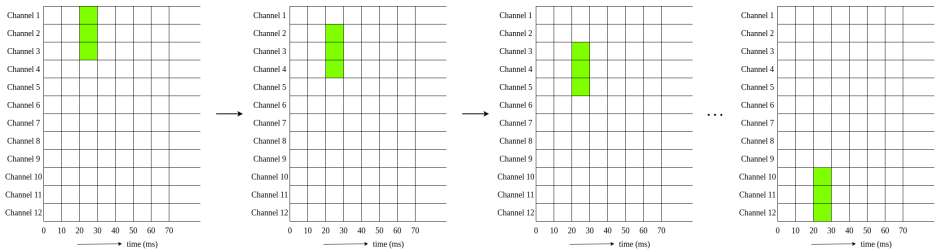
state, and thus, the final count is a MAC across time. The counter output is fed to the subsequent layers in the ML model, all of which are implemented in the digital domain.



(a) Inputs to $\Delta\Sigma$ MAC during first time window



(b) Inputs to $\Delta\Sigma$ MAC during second time window



(c) Inputs to $\Delta\Sigma$ MAC during third time window

Figure 4.5: Illustration of CNN computation of the real-time spectrogram being generated by the AFE. The shaded boxes indicate the AFE outputs selected by the MUX to be given as input to $\Delta\Sigma$ MACs.

Figure 4.5 illustrates which AFE outputs are selected as inputs to the $\Delta\Sigma$ MACs. Note that the same three AFE outputs must be multiplied by three different kernel columns depending on which CNN output is being computed. To avoid storing AFE outputs in analog buffers, we use three $\Delta\Sigma$ MACs in parallel, each of which takes a different kernel column as its weight input. The results also need to be stored in three parallel counters. The counters are then multiplexed and the result is sent to the next DSP block.

4.2 Circuit Description

4.2.1 Bandpass Filters

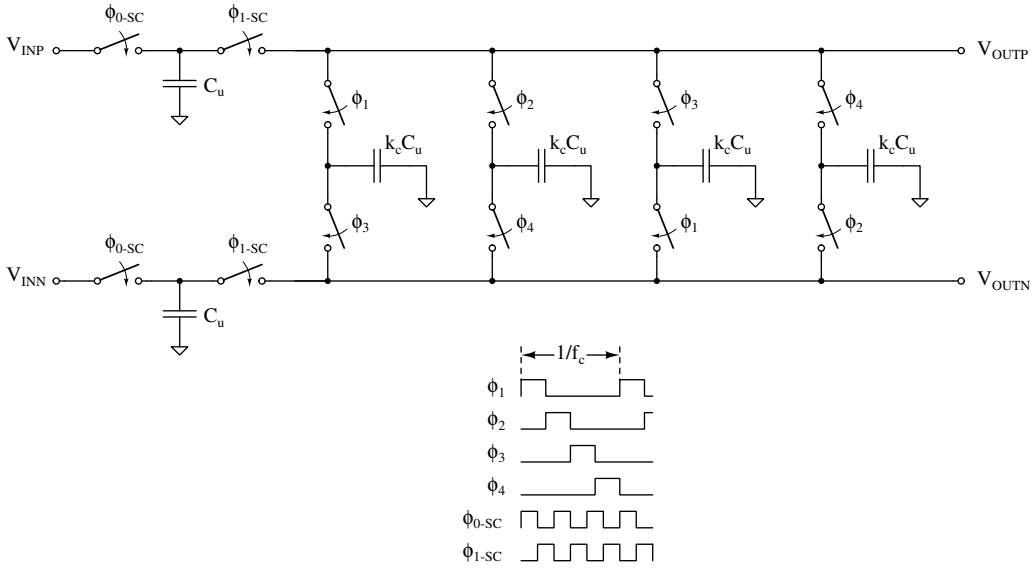


Figure 4.6: Switched capacitor differential N-Path filter

The switched capacitor differential N-Path filter topology given in [19] was used for implementing the mel-spaced bandpass filters. It's center frequency corresponds to the period of ϕ_1, ϕ_2, ϕ_3 and ϕ_4. Its quality factor can be tuned by varying the parameter k_c.

A range of 30Hz to 8KHz was chosen for the filterbank. Based on the ideal Filterbank response shown in Figure 2.2, the required approximate center frequencies and quality factors were determined and k_c was chosen accordingly. The Table 4.1 gives these specifications and chosen k_c.

Center Frequency, f_c (Hz)	**Bandwidth, BW (Hz)**	**Quality Factor, $Q = \frac{f_c}{BW}$**	**Scaling Factor, k_c**
180	100	1.8	1
360	120	3	1.5
600	145	4.14	2.25
860	176	4.89	2.5
1200	213	5.64	3
1600	257	6.22	3.25
2070	311	6.65	3.5
2650	377	7.03	3.75
3360	456	7.37	4
4200	552	7.61	4
5240	668	7.85	4.25
6500	808	8.05	4.5

Table 4.1: Specifications and chosen parameter k_c for Band-Pass Filters in the Mel-Filterbank

4.2.2 Full Wave Rectifier

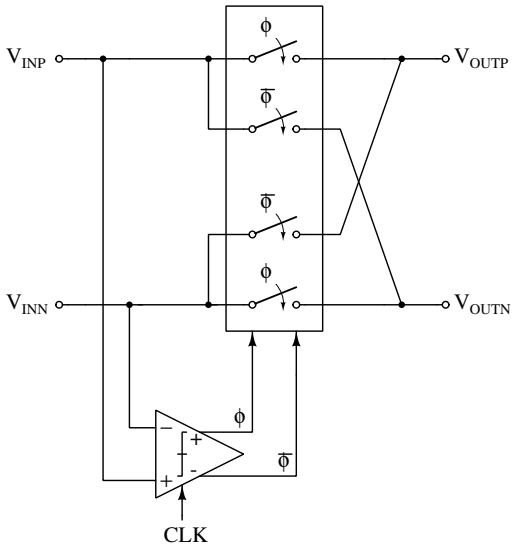


Figure 4.7: Full wave rectifier schematic

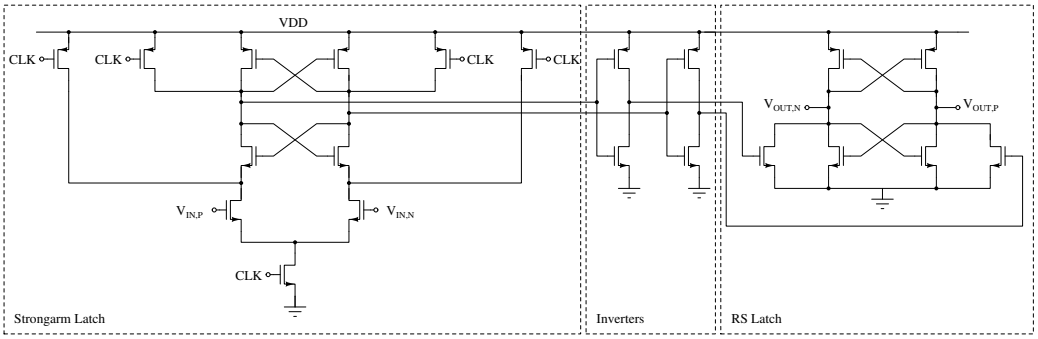


Figure 4.8: StrongARM comparator schematic

For the full wave rectifier, the circuit topology used in [16] as shown in Figure 4.7 was chosen since it only has dynamic power loss. The working principle of this rectifier is simple. The comparator constantly senses the input signal. If $V_{INP} > V_{INN}$ then the input is connected as is to the output, else the V_{INP} is connected to V_{OUTN} and V_{INN} is connected to V_{OUTP}. A StrongARM comparator was used as shown in Figure 4.8. The main advantage of StrongARM comparator is that it is a dynamic comparator, so there is negligible static power loss. However, the StrongARM comparator suffers from the issue of significant clock feedthrough. In this VAD architecture, clock feedthrough is not a big issue since it just occurs during clock edges and those high frequency spikes get filtered by the low pass filter.

4.2.3 Windowing Circuit

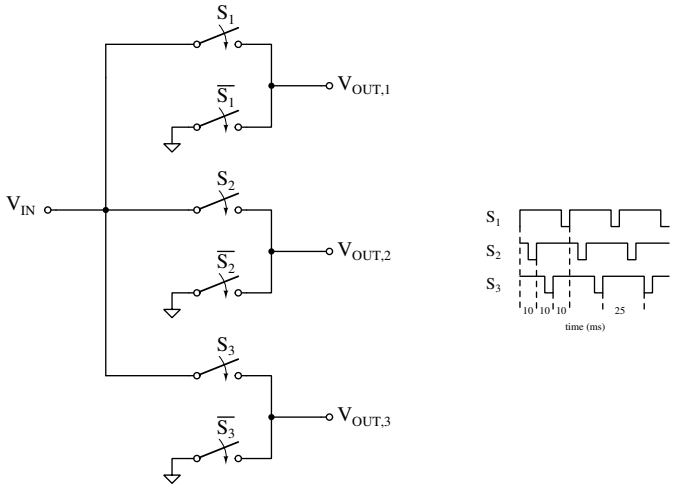


Figure 4.9: Windowing circuit schematic

Every channel has to be split into three sub-channels as per the windowing scheme described in Figure 4.2. A simple switch based implementation is used to achieve this, as shown in Figure 4.9. Since the load at the input pin is constantly changing, we need to precede this circuit with a unity gain buffer stage.

4.2.4 Lowpass Filter

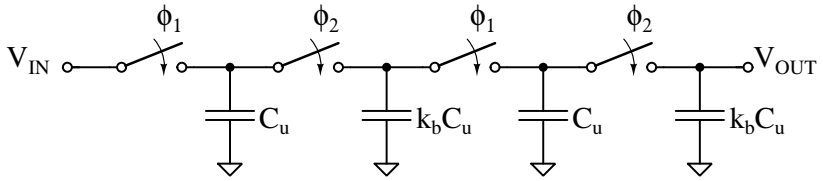


Figure 4.10: Second order switched capacitor low pass filter circuit schematic

To convert the rectified signal to a DC value, a second order lowpass filter is implemented. The cutoff frequency is chosen such that the time constant is sufficiently small so that the output settles within 15ms. The transfer of a second order RC filter is given by

$$H(s) = \frac{1}{(1 + sRC)^2}$$

In this case, $R = \frac{1}{f_s C_u}$ and $C = k_b C_u$. Therefore,

$$H(s) = \frac{1}{(1 + s\frac{k_b}{f_s})^2}$$

I have chosen $k_b = 8$ and $f_s = 6\ KHz$.

4.2.5 Non-linear Transform

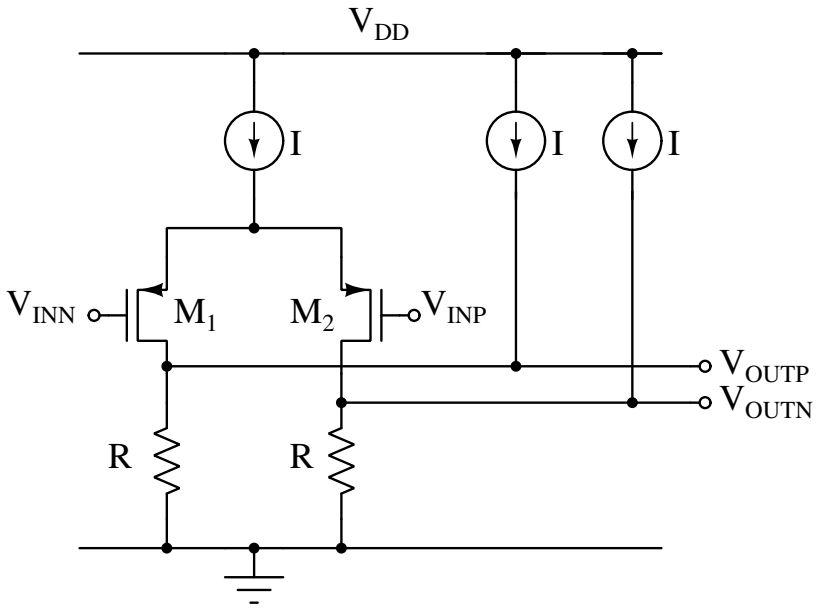


Figure 4.11: Non-linear transform circuit schematic

While analyzing the machine learning model in MATLAB, it was observed that we get better accuracy by performing a non-linear transform (like logarithm) on the extracted features. This non-linear transform should be monotonic increasing and should compress large inputs to smaller outputs. The exact shape of the transform is not very important and the training algorithm ensures that weights and biases are updated appropriately to give an accurate result. Therefore, the response of a simple differential pair was chosen, which also ensures that the AFE output is bound within $\pm 200\ mV$ voltage range.

4.2.6 $\Delta\Sigma$ Multiply Accumulate

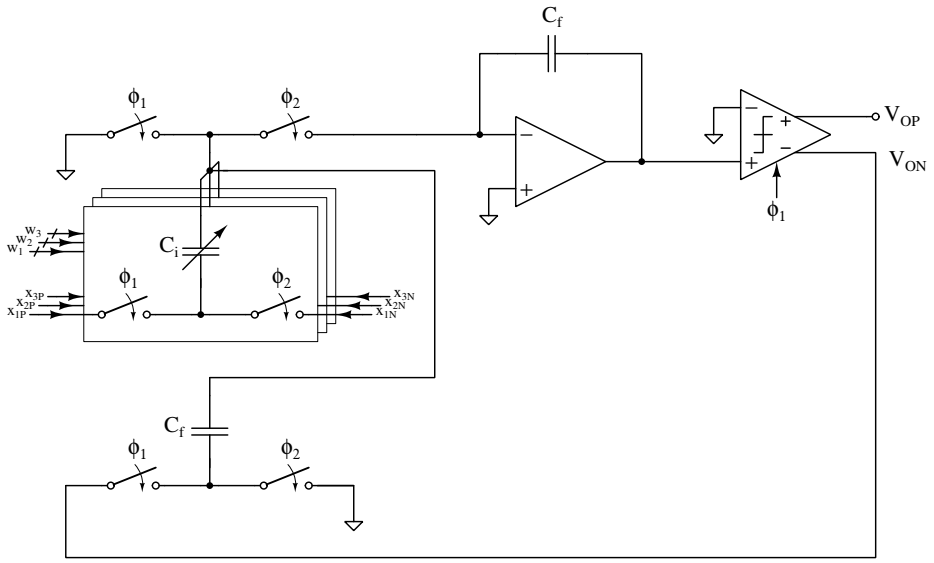
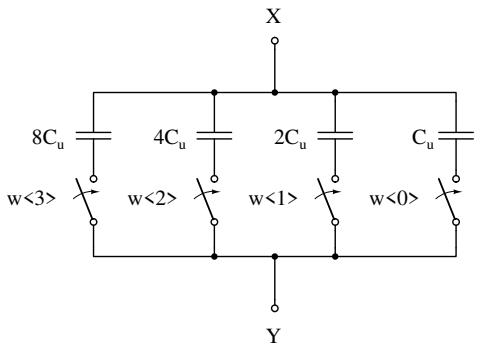
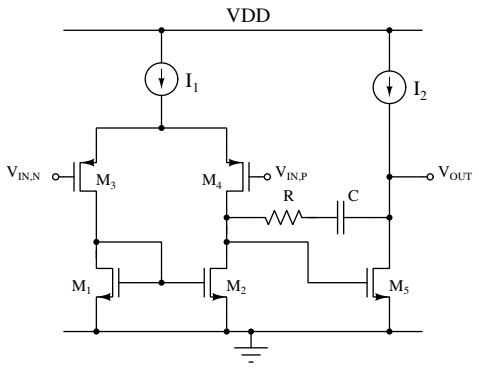


Figure 4.12: $\Delta\Sigma$ multiply accumulate circuit schematic. ϕ_1 and ϕ_2 are non-overlapping clocks.



(a) Digitally programmable capacitor schematic. $C_{eff} = wC_u$



(b) Two stage operational transconductance amplifier schematic

Figure 4.13: Sub-circuits in $\Delta\Sigma$ MAC

Figure 4.12 shows the circuit which performs multiply-accumulate operation within a $\Delta\Sigma$ ADC. The variable capacitors in Figure 4.12 are controlled by the digital weights as shown in Figure 4.13a. A two stage operational transconductance amplifier (OTA) was used with RC compensation as shown in Figure 4.13b. The same StrongARM comparator shown in Figure 4.8 was

also used here.

The working principle of the $\Delta\Sigma$ MAC circuit can be understood by applying the charge conservation principle. Assume all voltages are with respect to analog ground $= 0.5V_{DD}$. Assume that $V_O = V_{OP} = -V_{ON}$.

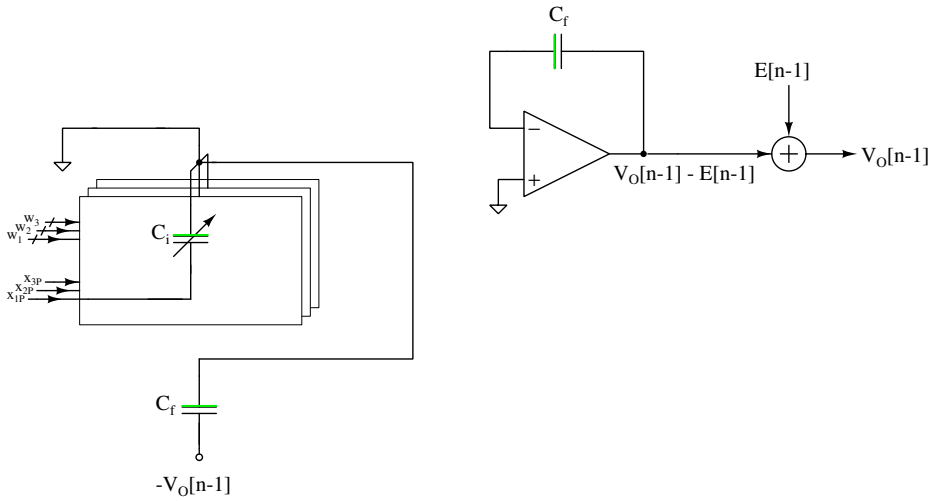


Figure 4.14: $\Delta\Sigma$ multiply accumulate circuit in ϕ_1

$$\begin{aligned} Q_{tot.\phi_1} &= -(x_{1p}[n]w_1 + x_{2p}[n]w_2 + x_{3p}[n]w_3)C_u \\ &\quad + V_O[n-1]C_f - (V_O[n-1] - E[n-1])C_f \\ &= -(x_{1p}[n]w_1 + x_{2p}[n]w_2 + x_{3p}[n]w_3)C_u + E[n-1]C_f \end{aligned}$$

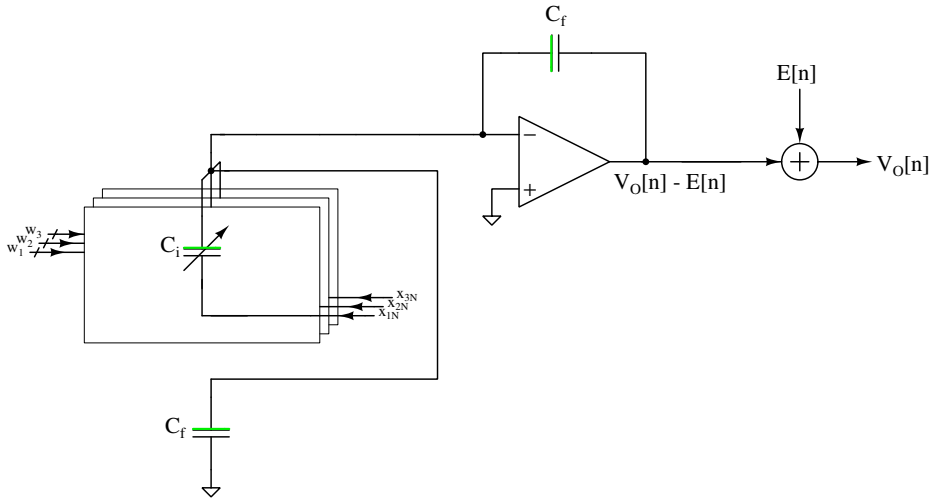


Figure 4.15: $\Delta\Sigma$ multiply accumulate circuit in ϕ_2

$$Q_{tot.\phi_2} = -(x_{1n}[n]w_1 + x_{2n}[n]w_2 + x_{3n}[n]w_3)C_u - (V_O[n] - E[n])C_f$$

Given that ϕ_1 and ϕ_2 are non-overlapping, we can apply charge conservation principle and say $Q_{tot,\phi_1} = Q_{tot,\phi_2}$.

$$\begin{aligned}
\implies & -(x_{1p}[n]w_1 + x_{2p}[n]w_2 + x_{3p}[n]w_3)C_u + E[n-1]C_f \\
& = -(x_{1n}[n]w_1 + x_{2n}[n]w_2 + x_{3n}[n]w_3)C_u - (V_O[n] - E[n])C_f \\
\implies & V_O[n]C_f = (x_1[n]w_1 + x_2[n]w_2 + x_3[n]w_3)C_u + (E[n] - E[n-1])C_f \\
\implies & V_O[n] = (x_1[n]w_1 + x_2[n]w_2 + x_3[n]w_3)\frac{C_u}{C_f} + E[n] - E[n-1]
\end{aligned}$$

Where $x_i[n] = x_{ip}[n] - x_{in}[n]$, $i \in \{1, 2, 3\}$. Taking Z-transform of the above equation, we get:

$$\boxed{V_O(z) = (X_1(z)w_1 + X_2(z)w_2 + X_3(z)w_3)\,\frac{C_u}{C_f} + \left(1 - z^{-1}\right)E(z)}$$

This expression resembles the standard equation of first order $\Delta\Sigma$ modulator with signal and noise transfer functions given by:

$$STF_{X_1} = \frac{w_1 C_u}{C_f}\ ,\ STF_{X_2} = \frac{w_2 C_u}{C_f}\ ,\ STF_{X_3} = \frac{w_3 C_u}{C_f}\ ,\ NTF = (1 - z^{-1})$$

Thus the $\Delta\Sigma$ multiply accumulate circuit implements the function shown in Figure 4.4. In my design, I have chosen $\frac{C_u}{C_f} = \frac{1}{16}$ and 4-bit integer weights. To allow for negative weights, I added an additional sign bit which interchanges x_{ip} and x_{in} when HIGH. Therefore, the domain of weights is

$$weights\ \in\ \left\{-\frac{15}{16}, -\frac{14}{16}, \ldots, \frac{14}{16}, \frac{15}{16}\right\}$$

4.2.7 Digital Backend

The following digital blocks were implemented as functional modules in Verilog-A in order to carry out simulations with Spectre simulator.

1. 10-bit up-counter
2. Register
3. Multiplexer
4. ReLU
5. Max pooling
6. Fully connected layer
7. Decision slicer

Chapter 5

Circuit Implementation and Simulation Results

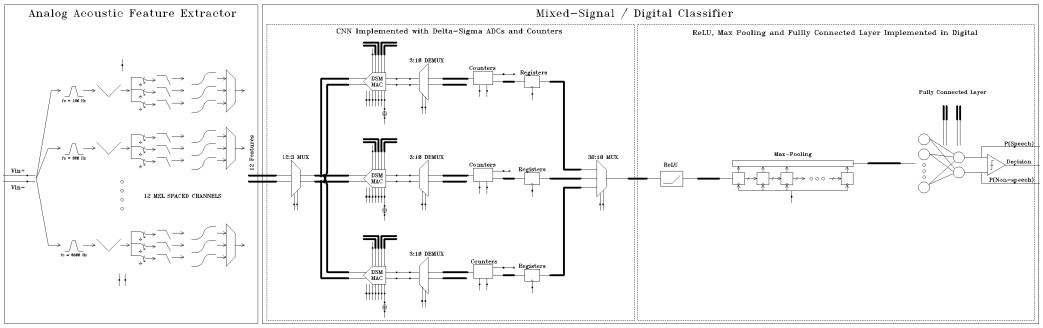


Figure 5.1: VAD System Cadence Implementation

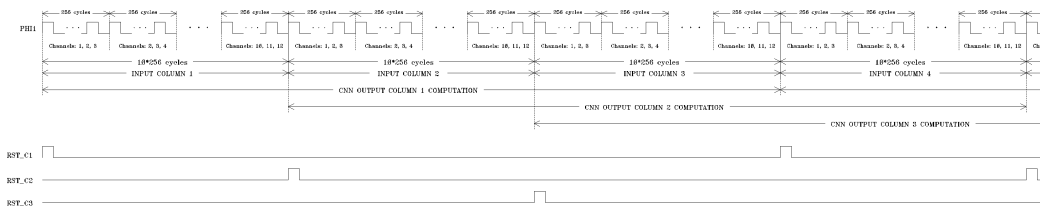


Figure 5.2: Timing diagram of the VAD system. PHI1 is the clock to $\Delta\Sigma$ MAC. RST_C1, RST_C2, RST_C3 are active high reset signals to the three parallel counters.

Figure 5.1 shows the top level view of the entire VAD system implemented in Cadence. The LNA was not implemented since its specification would depend on the microphone being used for testing. For simulation purposes,

the input audio was directly imported from a WAV file. The timing diagram is shown in Figure 5.2.

5.1 Analog Acoustic Feature Extractor

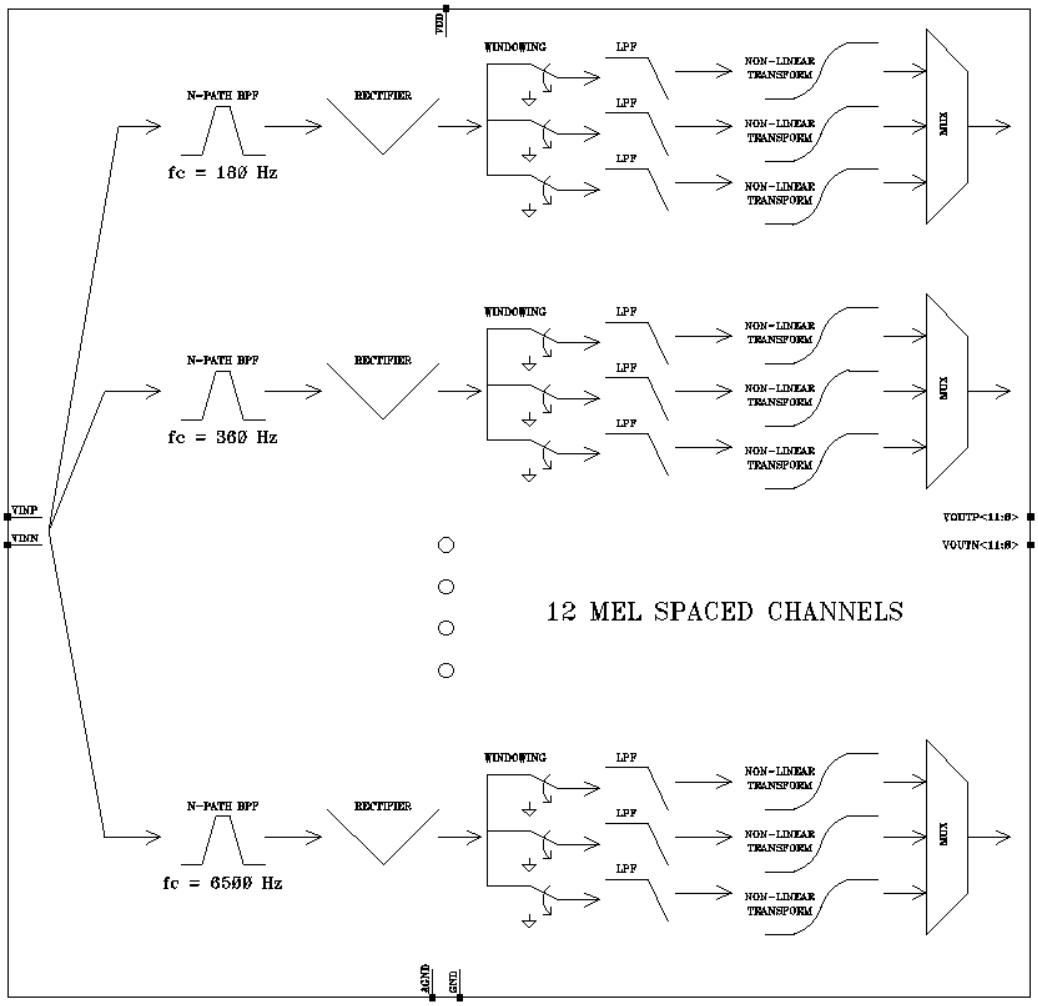


Figure 5.3: Analog acoustic feature extractor symbol

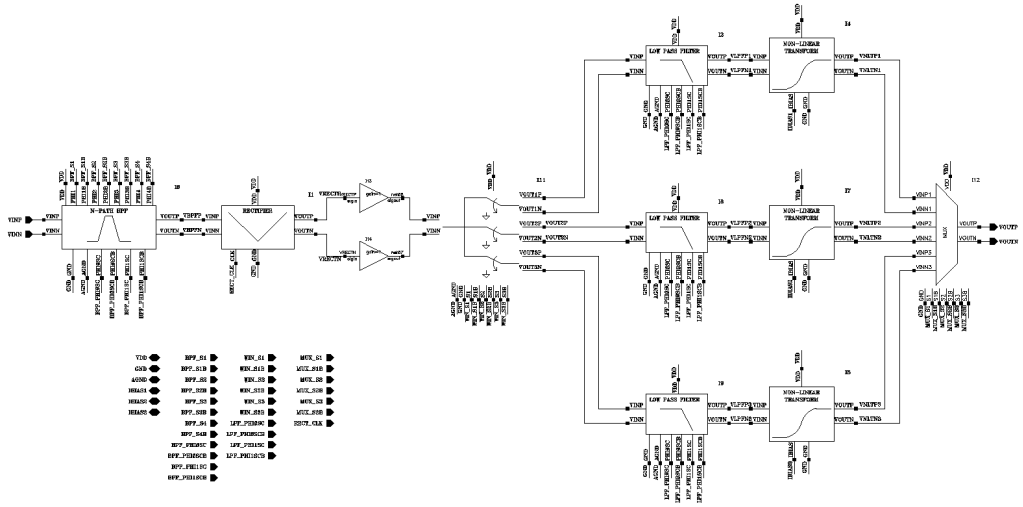


Figure 5.4: Single channel in the analog acoustic feature extractor

Figure 5.3 shows the symbol of the analog acoustic feature extractor implemented in Cadence. The feature extractor takes an audio signal as input and produces 12 output features in every time window. Figure 5.4 shows the implementation of single AFE channel.

5.1.1 Bandpass Filter

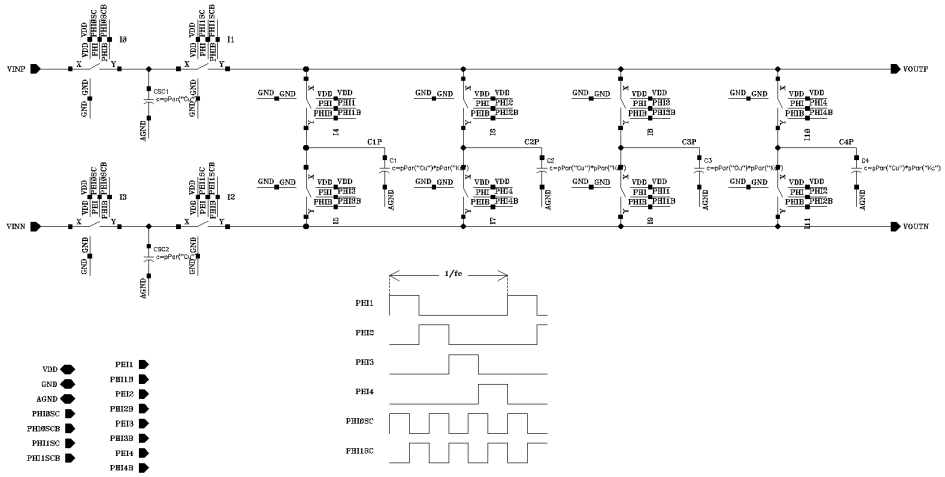


Figure 5.5: Switched capacitor differential N-Path bandpass filter

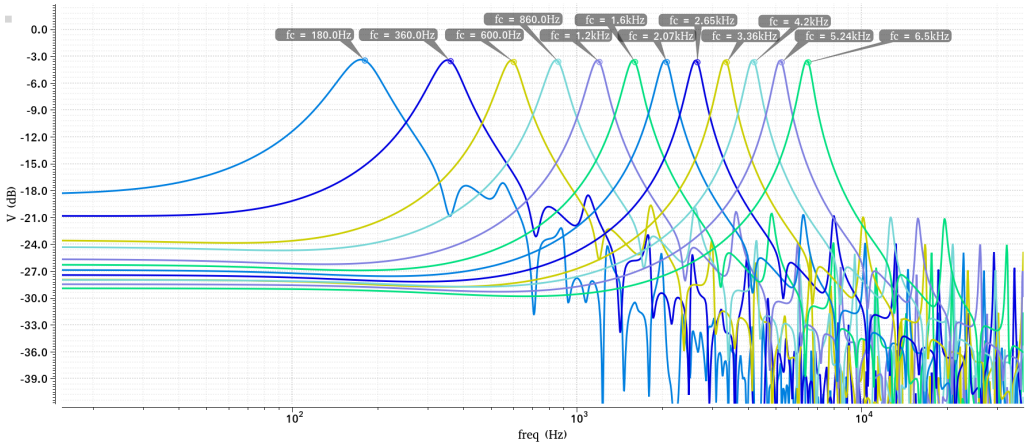


Figure 5.6: Spectre periodic AC analysis of 12 mel-spaced N-Path bandpass filters

5.1.2 Comparator

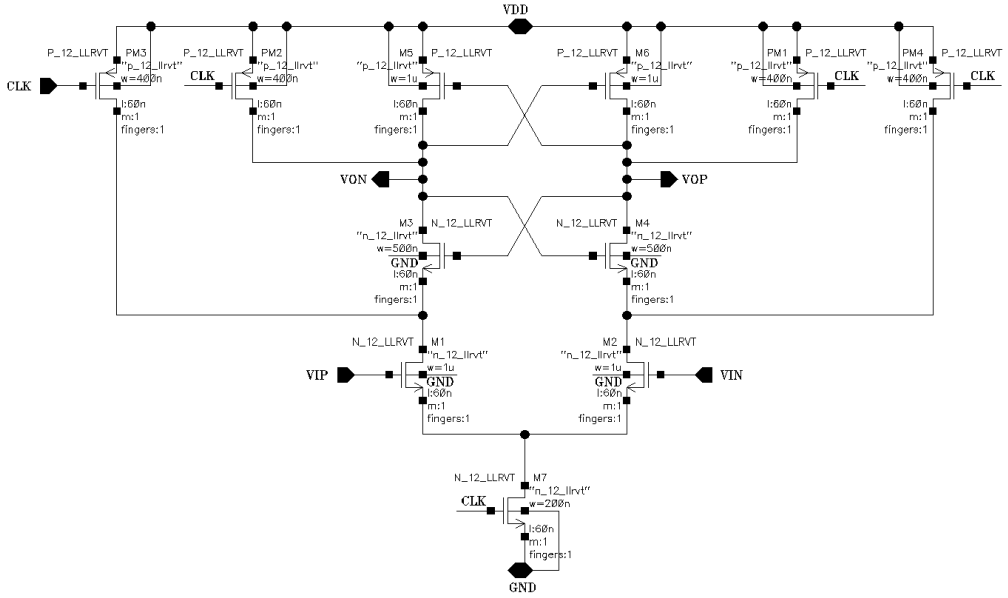


Figure 5.7: StrongARM latch

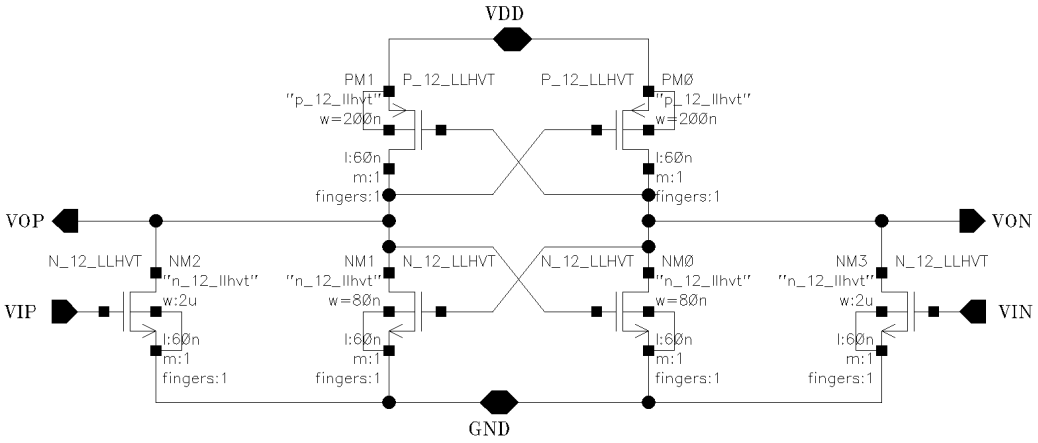


Figure 5.8: RS latch

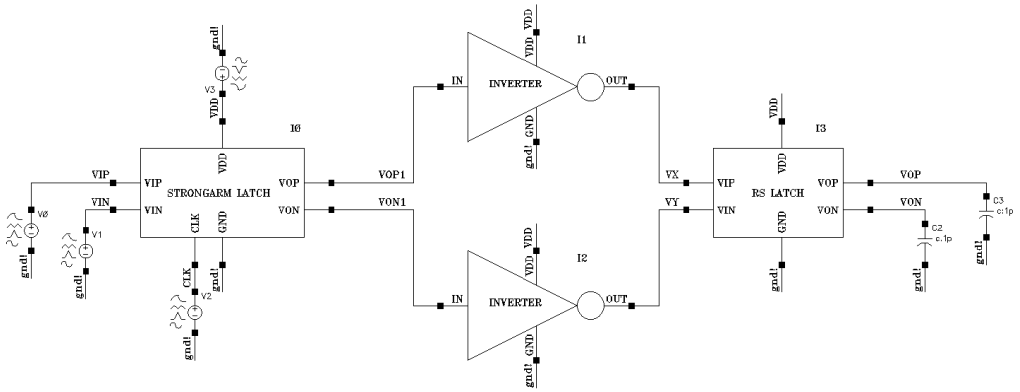


Figure 5.9: Full comparator testbench

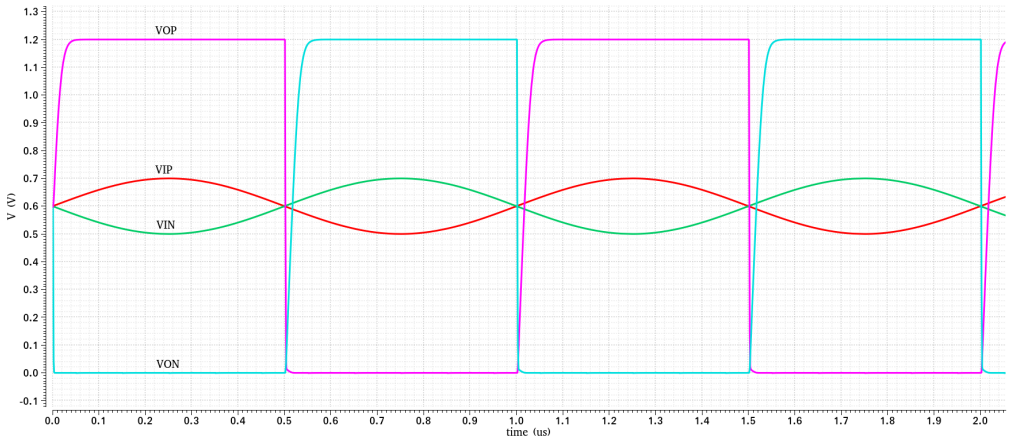


Figure 5.10: Comparator simulation results with 1 MHz sine wave input

5.1.3 Rectifier

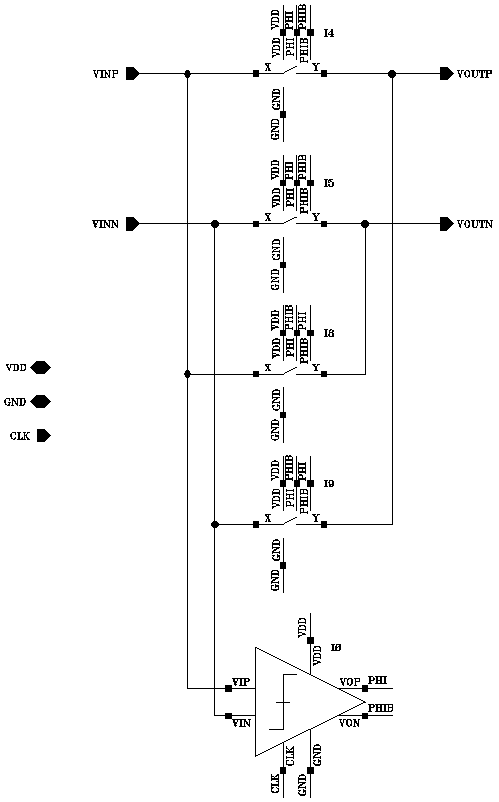


Figure 5.11: Full wave rectifier

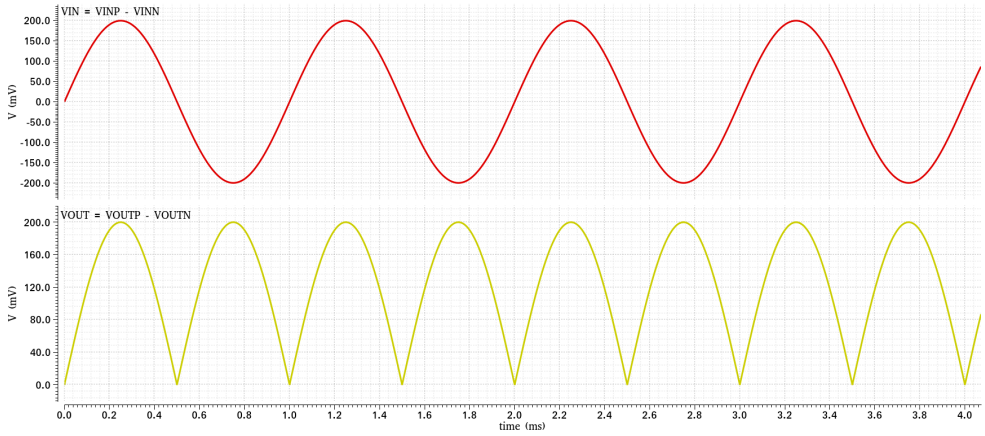


Figure 5.12: Transient simulation of the rectifier

5.1.4 Windowing

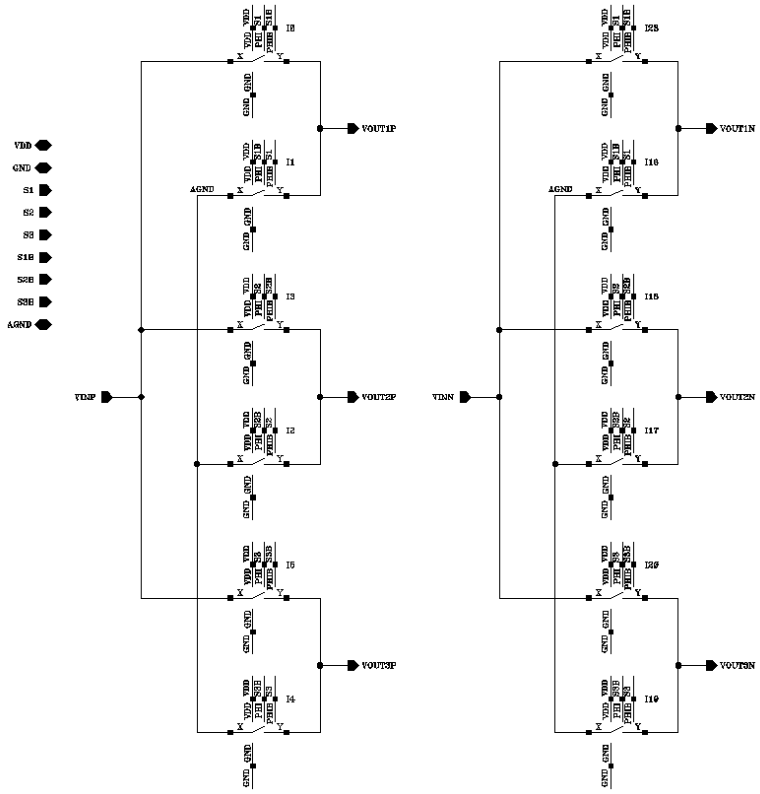


Figure 5.13: Windowing circuit

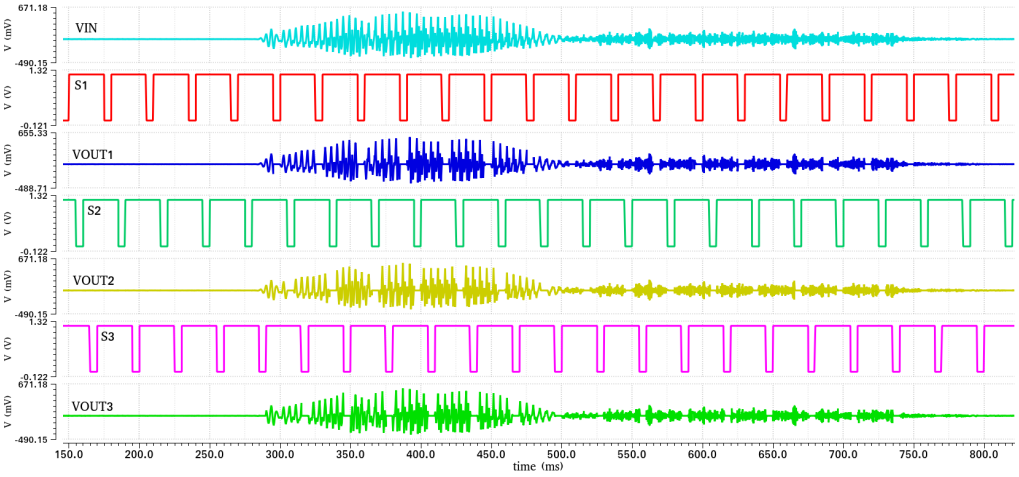


Figure 5.14: Transient simulation of the windowing circuit

5.1.5 Lowpass Filter

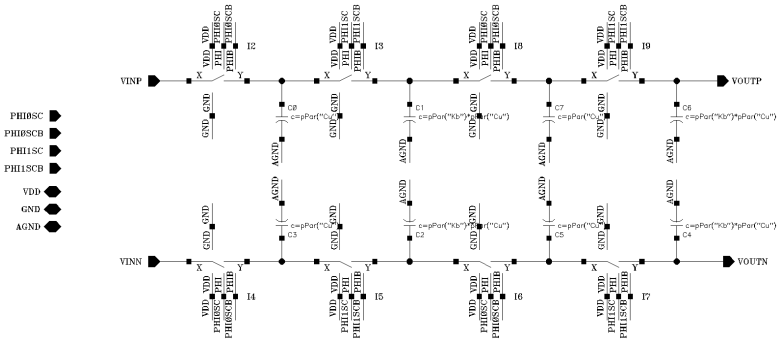


Figure 5.15: Lowpass filter

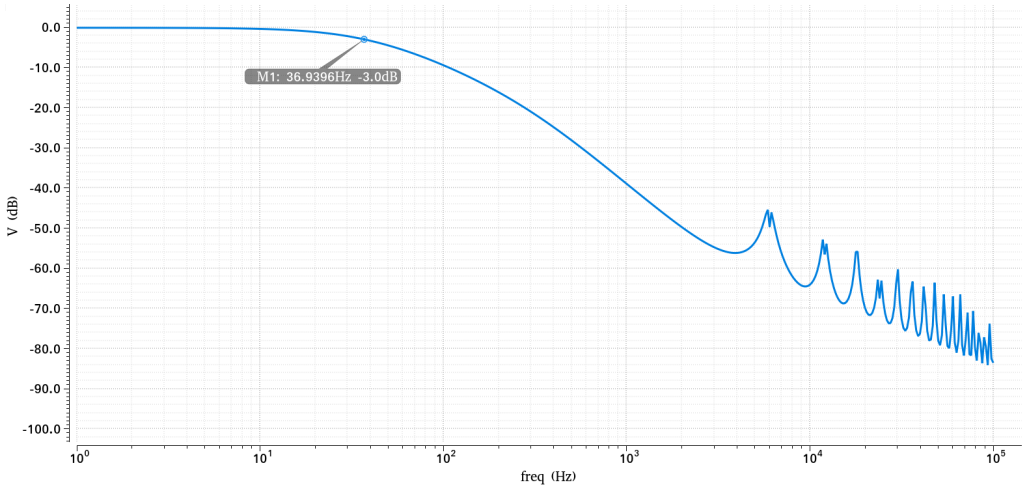


Figure 5.16: Periodic AC analysis of the lowpass filter

5.1.6 Non-linear Transform

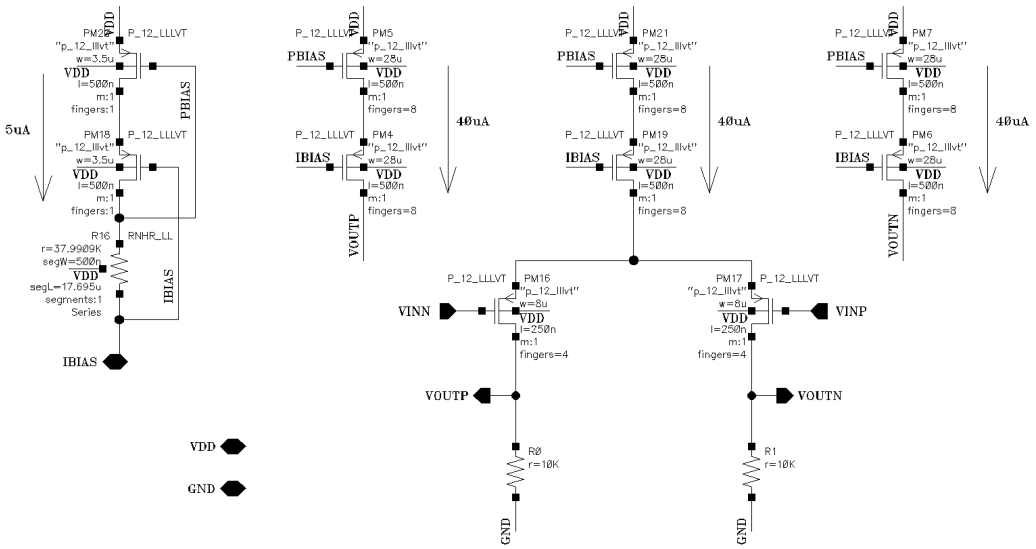


Figure 5.17: Non-linear Transform circuit

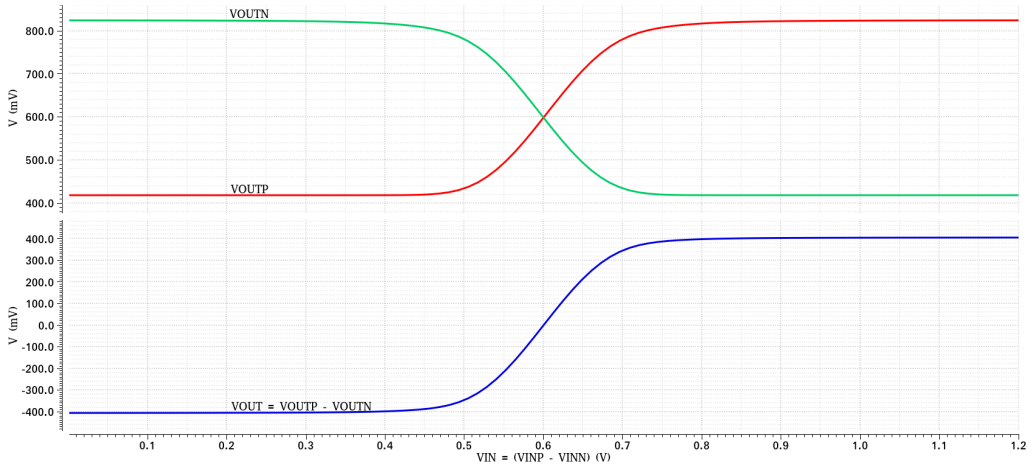


Figure 5.18: Parametric analysis of non-linear transform circuit

5.1.7 Full AFE Results

Test 1:

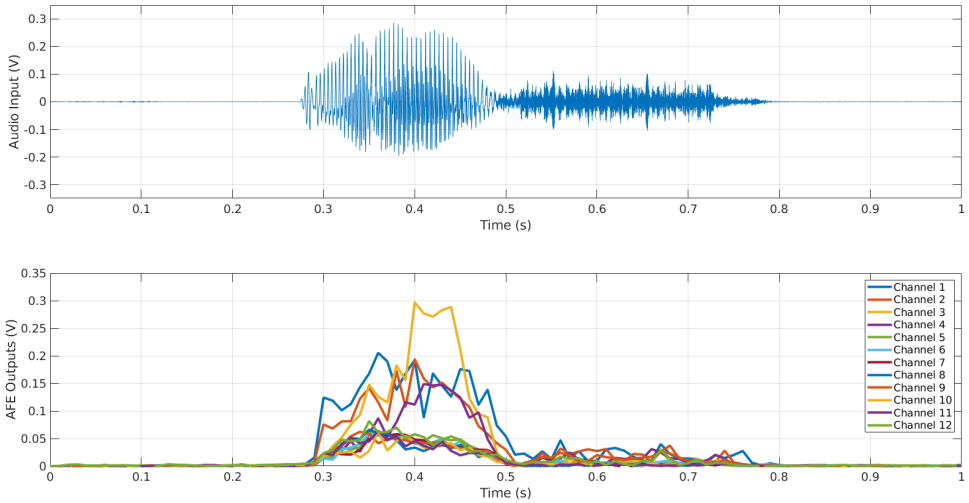


Figure 5.19: Audio input and the corresponding acoustic features extracted by the analog frontend. The audio recording is of a human saying the word 'yes'.

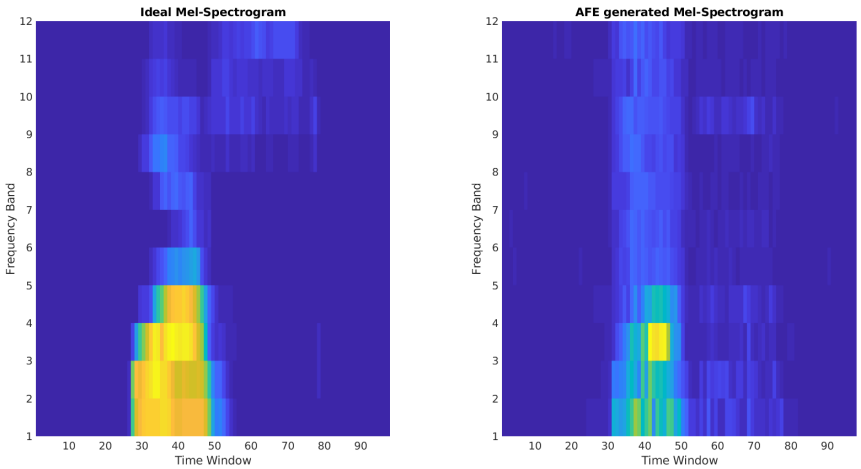


Figure 5.20: Comparison of ideal mel-spectrogram with mel-spectrogram generated from AFE outputs

Test 2:

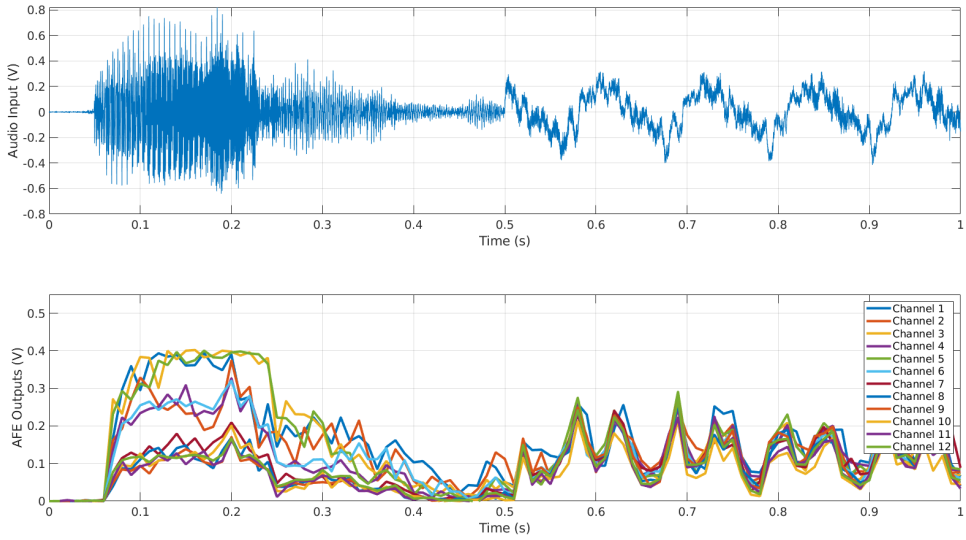


Figure 5.21: Audio input and the corresponding acoustic features extracted by the analog frontend. The audio recording is first of a human saying the word 'bird' and then some drilling noise.

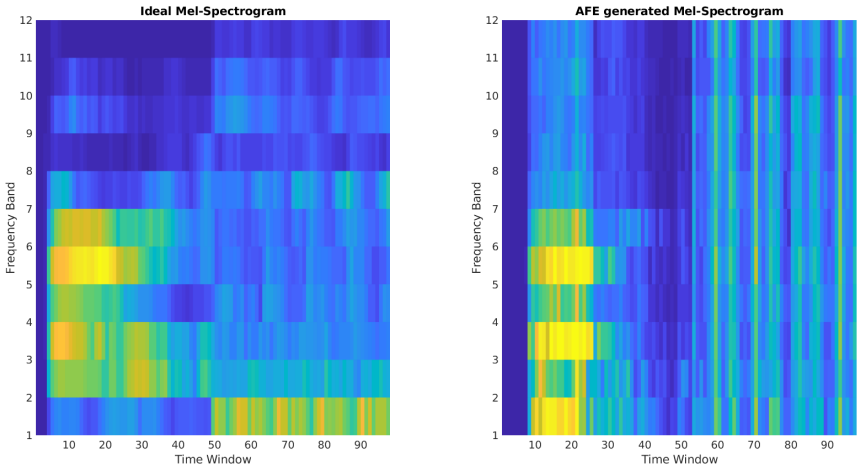


Figure 5.22: Comparison of ideal mel-spectrogram with mel-spectrogram generated from AFE outputs

5.2 $\Delta\Sigma$ Multiply Accumulate

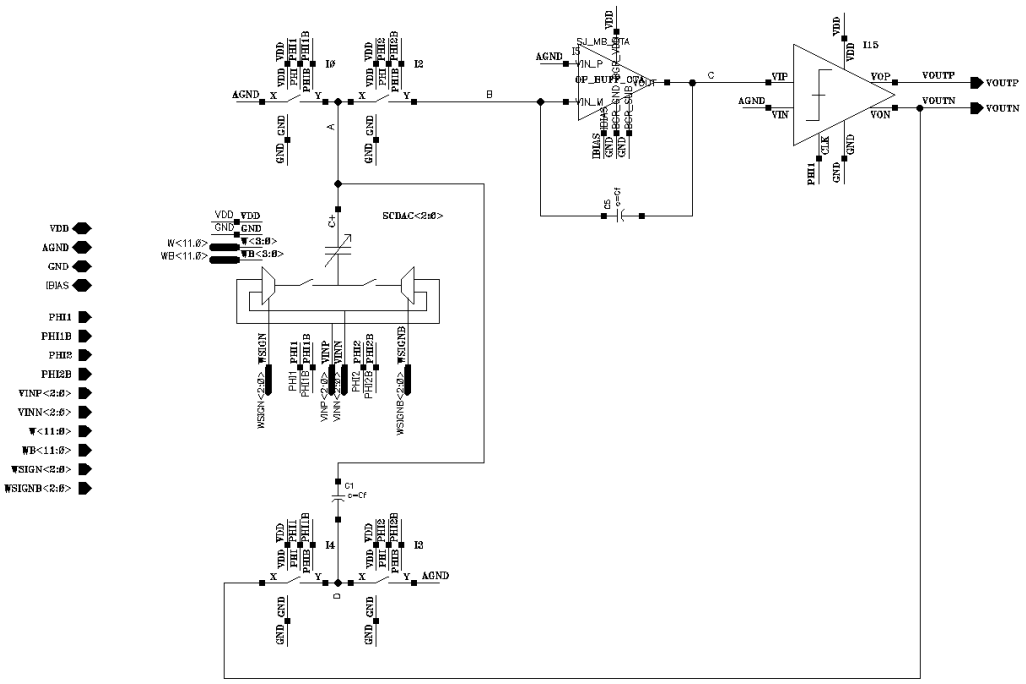


Figure 5.23: $\Delta\Sigma$ multiply accumulate circuit

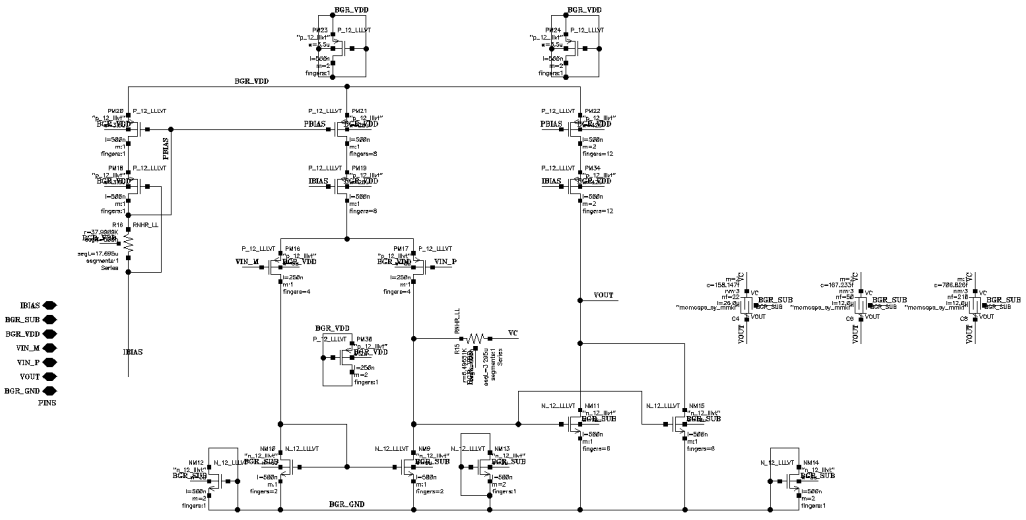


Figure 5.24: Two stage operational transconductance amplifier

Open Loop DC Gain	57.27 dB
Unity Gain Frequency	708.4 MHz
Phase Margin	55.78^o
3 dB Bandwidth	39.86 KHz

Table 5.1: Operational transconductance amplifier characterization

5.2.1 $\Delta\Sigma$ MAC and Counter Results

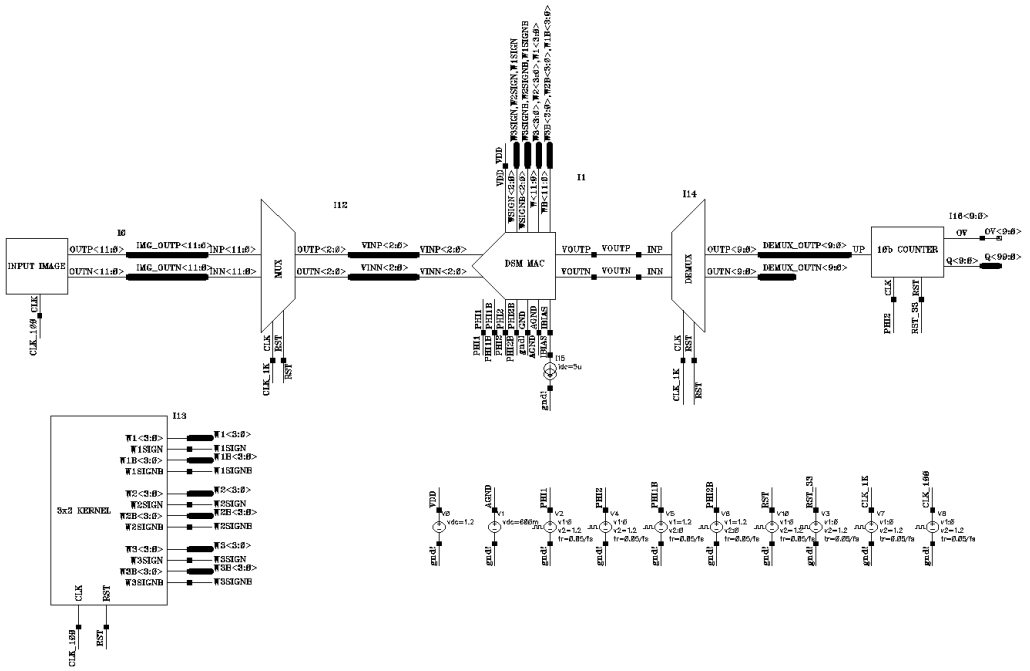


Figure 5.25: Testbench for analysing single $\Delta\Sigma$ multiply accumulate circuit

Figure 5.25 shows the testbench for analyzing the $\Delta\Sigma$ MAC circuit. The input image block continuously generates a random input image with 12 rows. The analog multiplexer selects 3 of these inputs at a time and connects them to the $\Delta\Sigma$ MAC block. The $\Delta\Sigma$ MAC output is connected to one of the counters based on which image input rows were selected.

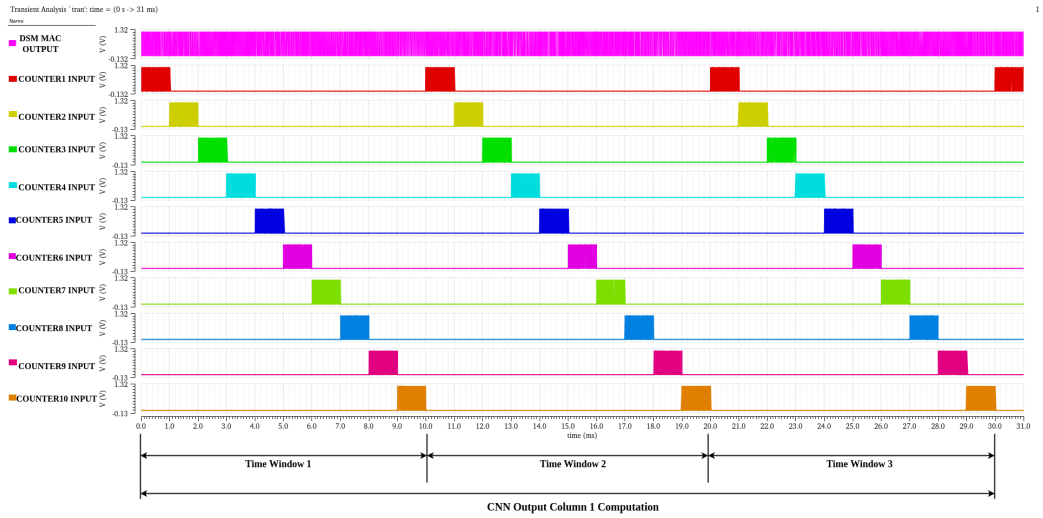


Figure 5.26: $\Delta\Sigma$ MAC output being connected as input to different counters

Figure 5.26 shows the simulation results illustrating the working of the $\Delta\Sigma$ MAC based CNN computation as described previously in Figure 4.5. The shaded regions represent the output of the $\Delta\Sigma$ MAC. Since the circuit operates at 256 KHz clock frequency, the pulses are very closely spaced on the given time scale and appear as shaded regions. Every multiply operation happens in 256 clock cycles. Using the expression for the number of counts given in section 4.1, the equation to convert from total MAC value (in mV) to final count is given by:

$$Count_1 = (MAC_1 + 600)\frac{256}{1200}$$

$$Count_2 = (MAC_2 + 600)\frac{256}{1200}$$

$$Count_3 = (MAC_3 + 600)\frac{256}{1200}$$

$$Count_{tot} = Count_1 + Count_2 + Count_3$$

$$\boxed{Count_{tot} = (MAC_{tot})\,\frac{256}{1200} + 384}$$

The above simulation was carried out with a 12x3 random input image and two different CNN kernels. Simulation results for both cases have been given below.

Input Image Column 1 (mV)	**Input Image Column 2 (mV)**	**Input Image Column 3 (mV)**
-163	35	196
59	184	53
143	-56	34
72	75	65
13	-107	-36
133	72	60
4	-138	-98
-108	127	20
-105	-1	-131
-3	-199	-103
-64	-162	188
-30	-66	193

Table 5.2: Test input data

-0.625	-0.25	-0.1875
-0.4375	-0.375	-0.25
-0.625	0	-0.9375

(a) 3x3 CNN Kernel

Expected CNN Output	**Expected Count**	**Simulated Count**	**Equivalent CNN Output**	**% Error**
-172.9	347	352	-150	13.3
-248.8	331	334	-234	5.8
-132	356	358	-121.9	7.7
-171.9	347	350	-159.4	7.3
14.6	387	387	14.1	3.4
10.8	386	387	14.1	29.3
233.4	434	431	220.3	5.6
209.5	429	425	192.2	8.3
55.9	396	397	60.9	9.1
-49.5	373	373	-51.6	4.2

(b) Comparison of expected and achieved CNN output

Table 5.3: $\Delta\Sigma$ MAC based CNN Test 1 results

0.5	−0.375	0.125
0.5625	0.875	0.75
−0.1875	0.3125	0.6875

(a) 3x3 CNN Kernel

Expected CNN Output	**Expected Count**	**Simulated Count**	**Equivalent CNN Output**	**% Error**
142.9	414	412	131.2	8.1
78.7	401	401	79.7	1.3
191.0	425	424	187.5	1.8
-58.5	372	372	-56.2	3.8
113.7	408	407	107.8	5.2
-71.3	369	370	-65.6	8
36.2	392	390	28.1	22.2
-389.7	301	305	-370.3	5
-230.9	335	341	-201.6	12.7
141.2	414	401	79.7	43.6

(b) Comparison of expected and achieved CNN output

Table 5.4: ΔΣ MAC based CNN Test 2 results

5.3 Digital Backend

All digital blocks have veen implemented as functional modules in Verilog-A.

5.3.1 ReLU

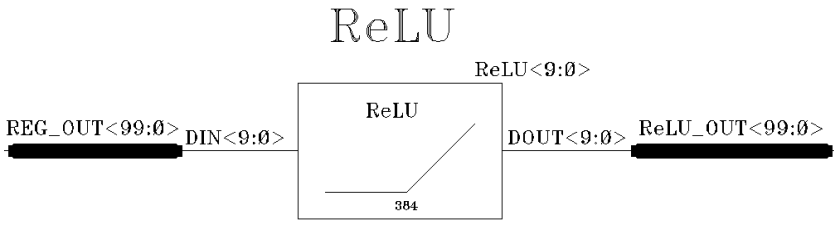


Figure 5.27: ReLU symbol

5.3.2 Max Pooling

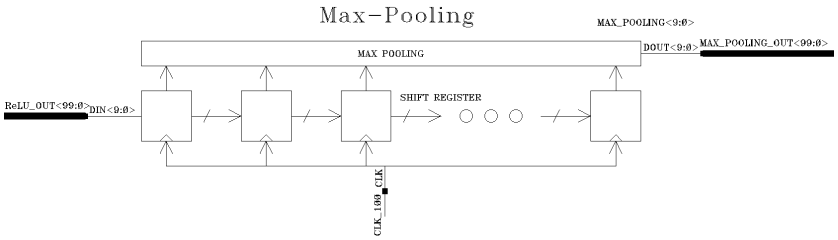


Figure 5.28: Max-Pooling symbol

5.3.3 Fully connected layer

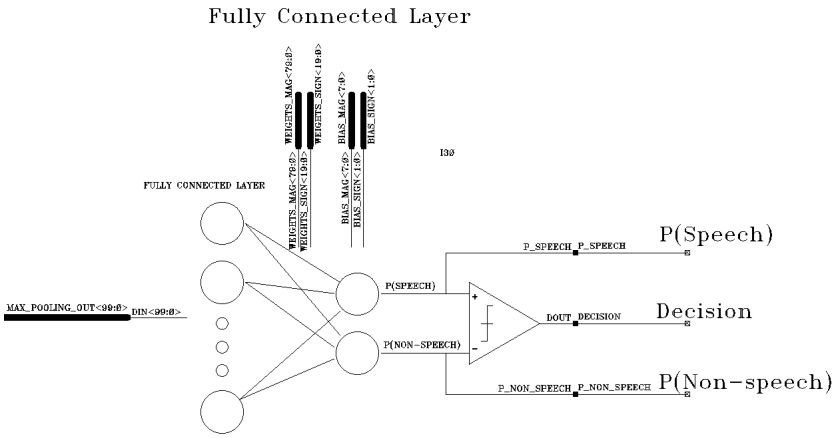


Figure 5.29: Fully connected layer symbol

Chapter 6

Conclusion & Future Work

A novel voice activity detector has been proposed which performs acoustic feature extraction in the analog domain and neural network based classification in the mixed-signal and digital domains. All the analog and mixed-signal blocks were designed up to the schematic level. The digital blocks were implemented as functional modules in Verilog-A. Transistor level simulations of individual circuit blocks were carried out to validate their behavior and ensure that they meet all the specifications imposed by the system architecture.

The entire acoustic feature extractor was tested with two different audio inputs. The features extracted by the AFE were compared with the corresponding ideal features. For both the cases, it was observed that the AFE generated mel-spectrogram shows a good resemblance with the ideal mel-spectrogram.

The $\Delta\Sigma$ MAC based CNN implementation was tested with a 12x3 random input image and two different CNN kernels. The simulation results showed a good overall match between expected and actual count values. Accuracy improvements, if necessary, can be achieved by increasing the number of clock cycles per MAC operation.

Future work includes:

1. Optimizing individual circuit blocks to minimize power consumption
2. Designing the clocking circuits
3. Improving the accuracy of $\Delta\Sigma$ MAC
4. Layout of the VAD and post-layout simulations

5. Building a software model of the AFE for training data generation
6. Training the VAD using the feature dataset generated by the software model of AFE
7. **C**haracterizing the VAD - power, speech / non-speech hit rate
8. Exploring alternate ML algorithms for VADs
9. Exploring in-memory computing architectures and their utility in VADs
10. Applying ideas presented in this work to other edge AI systems

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] 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.

[3] Zhuo Wang, Jintao Zhang, and Naveen Verma. "Realizing Low-Energy Classification Systems by Implementing Matrix Multiplication Directly Within an ADC". In: *IEEE Transactions on Biomedical Circuits and Systems* 9.6 (2015), pp. 825–837. DOI: 10.1109/TBCAS.2015.2500101.

[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 A. Nielsen. *Neural Networks and Deep Learning*. misc. 2018. URL: http://neuralnetworksanddeeplearning.com/.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Udita Mukherjee et al. "A 28.5µW All-Analog Voice-Activity Detector". In: *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2021, pp. 1–5. DOI: 10.1109/ISCAS51556.2021.9401504.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Wikipedia contributors. *Mel scale — Wikipedia, The Free Encyclopedia*. [Online; accessed 16-June-2022]. 2022. URL: https://en.wikipedia.org/w/index.php?title=Mel_scale&oldid=1085283171.

[22] Wikipedia contributors. *Spectrogram — Wikipedia, The Free Encyclopedia*. [Online; accessed 16-June-2022]. 2022. URL: https://en.wikipedia.org/w/index.php?title=Spectrogram&oldid=1091793825.