

EE593 Dual Degree Project Phase 1

Analog Acoustic Feature Extraction for Always-On Voice Activity Detection

Guide: Prof. Rajesh Zele

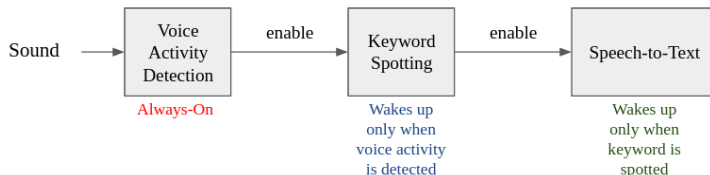
Mihir Kavishwar (17D070004)
Advanced Integrated Circuits and Systems Lab
Electrical Engineering, IIT Bombay

Outline

- 1 Introduction
- 2 Literature Survey
- 3 N-Path Filtering
- 4 Mel Spectrogram & Mel Filterbank
- 5 Proposed VAD Architecture
- 6 Testbench & Simulation Results
- 7 Conclusion
- 8 Future Work
- 9 References

Introduction

- A **Voice Activity Detector** (VAD) identifies the presence or absence of **human speech** in an audio signal
- In mobile phones, wearables and other Internet-of-Things (IoT) devices, VADs serve as a **wakeup mechanism** for the DSP blocks which perform advanced speech processing tasks



- Our goal is to build an ultra-low-power VAD for Edge applications by exploiting the **energy efficiency** of Analog signal processing and **robustness** of Machine Learning models to non-idealities

Introduction

- Typical VADs consist of two parts:
 - 1 **Acoustic Feature Extractor** - It converts the incoming signal into low-dimensional but dense acoustic features
 - 2 **Machine Learning Classifier** - It takes a feature set input and produces a binary decision: *speech* or *non-speech*
- The feature extractor can be made more energy efficient by performing some or all of the computation in the **Analog** domain

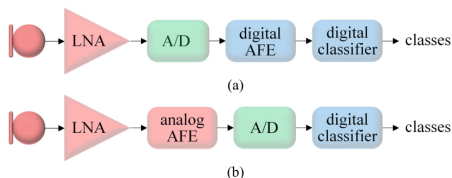


Figure: (a) Conventional digital-intensive and (b) analog-intensive signal processing chain in acoustic inference sensing systems. AFE = Acoustic Feature Extractor. *Source: Minhao Yang et al., JSSC '21 [7].*

Literature Survey 1: Power-Proportional Acoustic Sensing

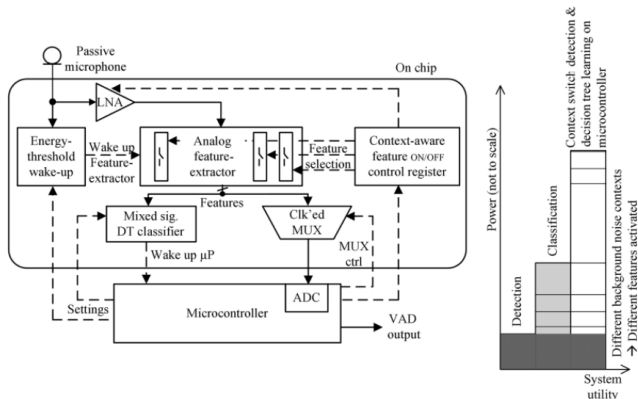


Figure: System architecture of power-proportional VAD (left) and its power scaling with sensing complexity (right). *Source: Komail Badami et al., JSSC '16 [1].*

Literature Survey 1: Power-Proportional Acoustic Sensing

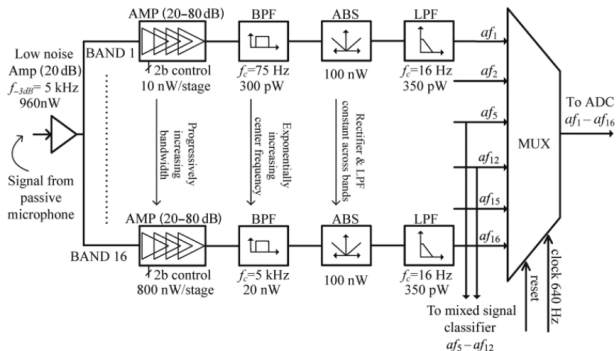


Figure: Schematic and design parameters of the analog feature extraction block. *Source: Komail Badami et al., JSSC '16 [1].*

Key Idea

The power consumption of a system can be scaled in proportion to the complexity of the sensing task.

Literature Survey 2: Event-based Acoustic Feature Extraction

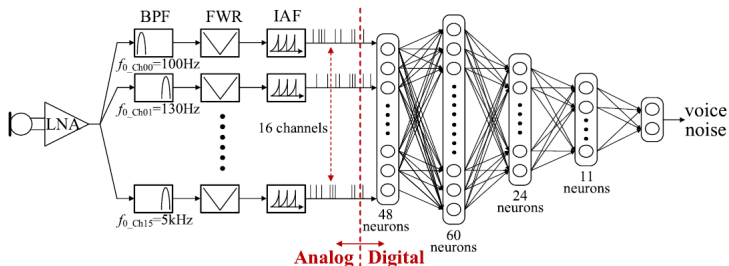


Figure: VAD system architecture using analog acoustic feature extraction and digital classification with event-driven analog-to-digital conversion. *Source: Minhao Yang et al., JSSC '19 [3]*

Literature Survey 2: Event-based Acoustic Feature Extraction

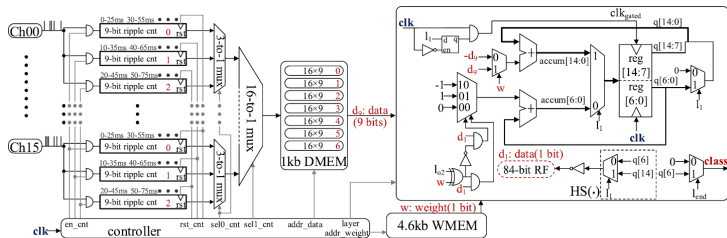
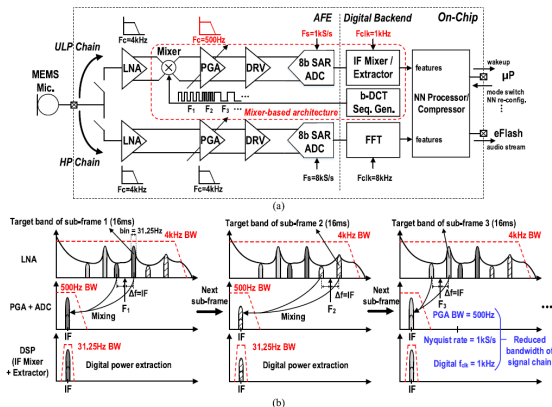


Figure: Block diagram of binarized MLP digital classifier. Source: Minhao Yang et al., JSSC '19 [3]

Key Idea

Event-driven Analog to Digital conversion is useful because it combines the functions of integration and quantization.

Literature Survey 3: Mixer-based Sequential Frequency Scanning



Key Idea

Sequential frequency scanning enables extremely high energy efficiency since it doesn't need a multi-channel filterbank, but there is a tradeoff with latency.

Figure: (a) VAD system architecture. (b) Operating principle of mixer-based sequential frequency scanning. *Source: Sechang Oh et al., JSSC '19 [2]*

Literature Survey 4: Energy-Quality Scaling

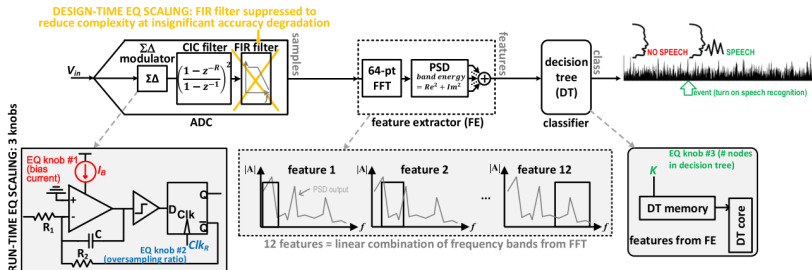


Figure: Proposed Energy-Quality scalable VAD architecture with three run-time knobs and one design-time knob. *Source: Jinq Horng Teo et al., TCAS1 '20 [4]*

Literature Survey 4: Energy-Quality Scaling

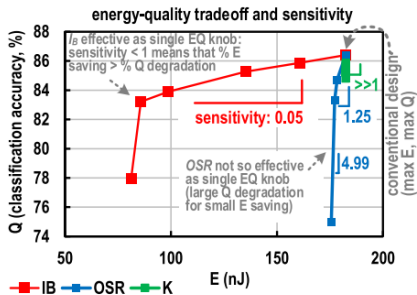
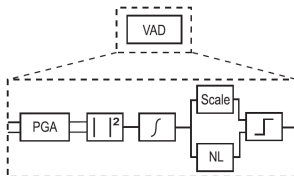


Figure: Energy-quality plots under individual EQ knob tuning (voice activity detection under a 10-dB noise environment). Source: Jinq Horng Teo et al., TCAS1 '20 [4]

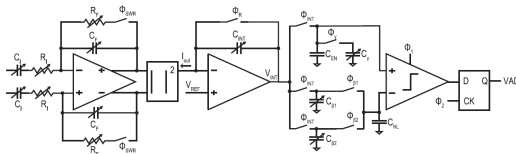
Key Idea

Energy-Quality knobs can be inserted along the signal chain, and system-level simultaneous co-optimization of such knobs can be done to minimize the energy under a given detection quality target.

Literature Survey 5: Fully Analog VAD



(a) Block diagram of the proposed analog VAD. NL = Noise Level.



(b) Schematic of the proposed analog VAD.

Figure: Source: Marco Croce et al., JSSC '21 [5]

Key Idea

Complete analog implementation of VAD is possible if we only want to detect high amplitude sounds (like human speech) in noisy environments.

Literature Survey 6: N-Path Switched Capacitor Filterbank

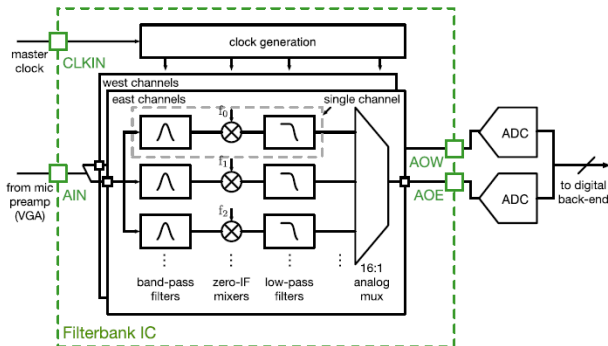


Figure: Feature extraction IC block diagram. The 32 channels are split into two 16:1 time-multiplexed outputs: AOW and AOE. Center frequencies and bandwidths are reconfigurable by setting clock divider ratios in the block labeled “clock generation.” Signal input and outputs are implemented pseudo differentially. *Source: Villamizar et al., TCAS1 '21 [6]*

Literature Survey 6: N-Path Switched Capacitor Filterbank

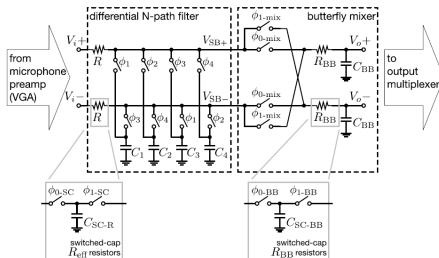


Figure: Single channel schematic. N-Path bandpass topology used for subband filtering followed by a butterfly differential mixer using a low-pass filter load for demodulation. *Source: Villamizar et al., TCAS1 '21 [6]*

Key Idea

N-Path SC BPFs are suitable for VAD because of high tunability and low power consumption. Harmonic responses and folding can be absorbed by ML model of the classifier.

Literature Survey: Comparison Table

	TCAS1 '21 [6]	JSSC '21 [5]	TCAS1 '20 [4]	JSSC '19 [2]	JSSC '19 [3]	JSSC '16 [1]
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 1h	NOISEUS
Accuracy	KWS: 92.4%	VAD: 99.5%	VAD: 87.3%	VAD: 91.5%	VAD: 85%	VAD: 89%
Latency (ms)	26	32	8	512	10	< 100

Table: Comparison of Voice Activity Detectors. Accuracy is computed over different datasets and therefore fair comparison is difficult. Power consumption values are for entire system and not just analog frontend.

N-Path Filtering

N-Path Filters have the several advantages:

- 1 High quality factor is easily achievable
- 2 Easily tunable center frequency and quality factor
- 3 High energy efficiency because power is required only to drive switches

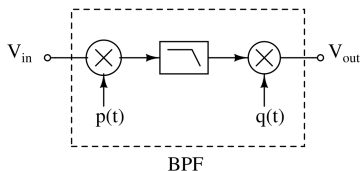


Figure: The fundamental principle used in an N-Path Bandpass Filter is
Downconversion + Lowpass Filtering + Upconversion = Bandpass Filtering

N-Path Filtering

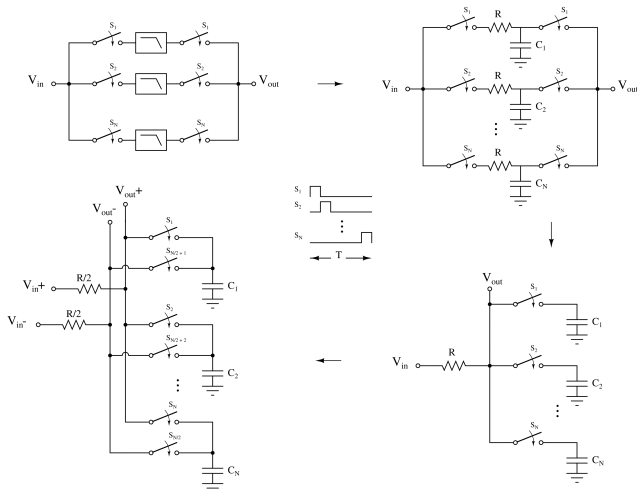


Figure: Derivation of differential N-Path Bandpass Filter from the original conceptual diagram

Mel Spectrogram & Mel Filterbank

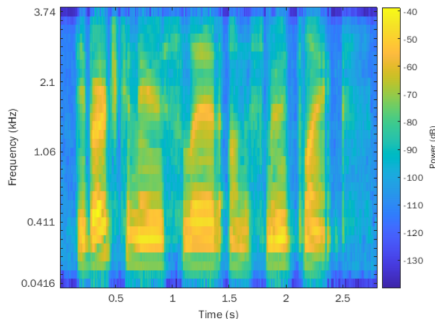
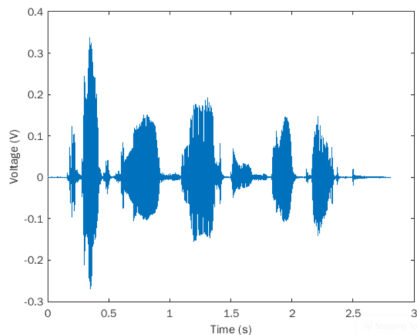


Figure: (a) Time domain (b) Spectral representation of the same speech signal

The **mel scale** is a perceptual scale of pitches judged by listeners to be equal in distance from one another. In audio speech processing we typically use **mel spectrogram**, where the frequencies are converted to the mel scale.

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

Mel Spectrogram & Mel Filterbank

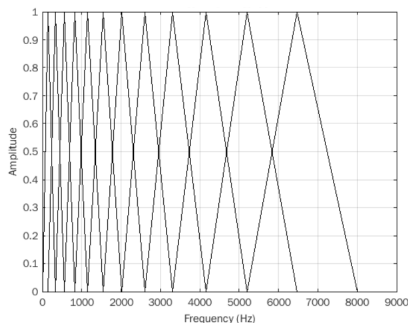


Figure: 12-filter Mel Filterbank with frequency range of 30Hz to 8KHz

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

Proposed VAD Architecture

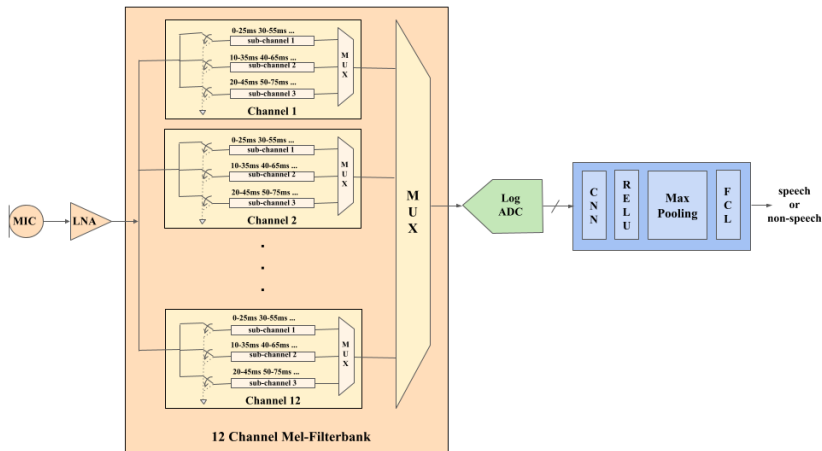


Figure: Proposed VAD architecture. The orange, green and blue blocks are implemented in analog, mixed-signal and digital domains respectively.

Proposed VAD Architecture

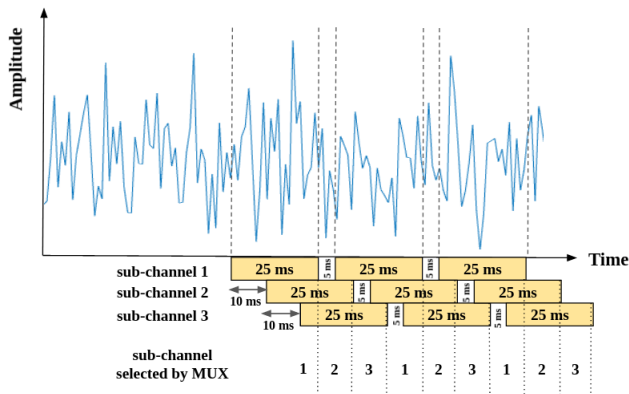


Figure: Description of how sub-channels process the input signal in different frames. Due to such arrangement we get the throughput of the Filterbank as 10ms although the frame length is 25ms.

Proposed VAD Architecture

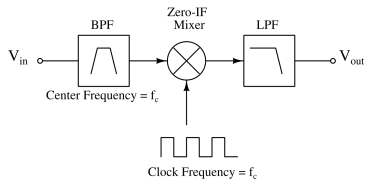


Figure: Block diagram of a sub-channel

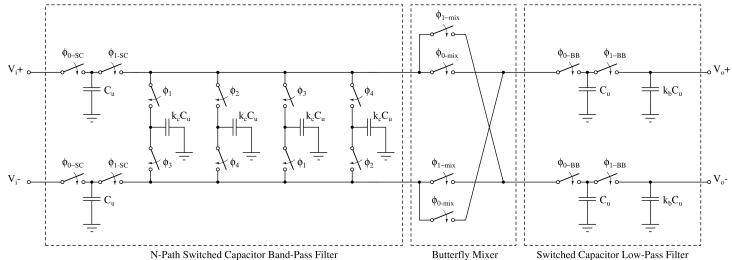


Figure: Schematic of a sub-channel. Source: Villamizar et al., TCAS1 '21 [6]

Testbench & Simulation Results

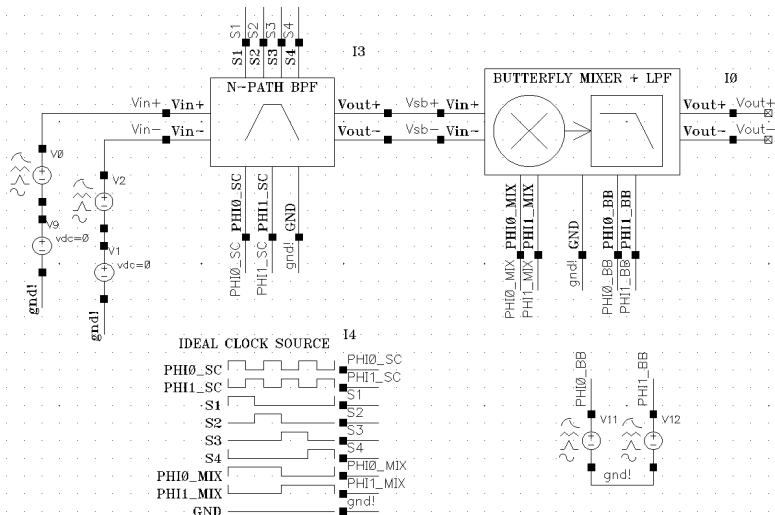


Figure: Testbench for single sub-channel of the Mel-Filterbank

Testbench & Simulation Results

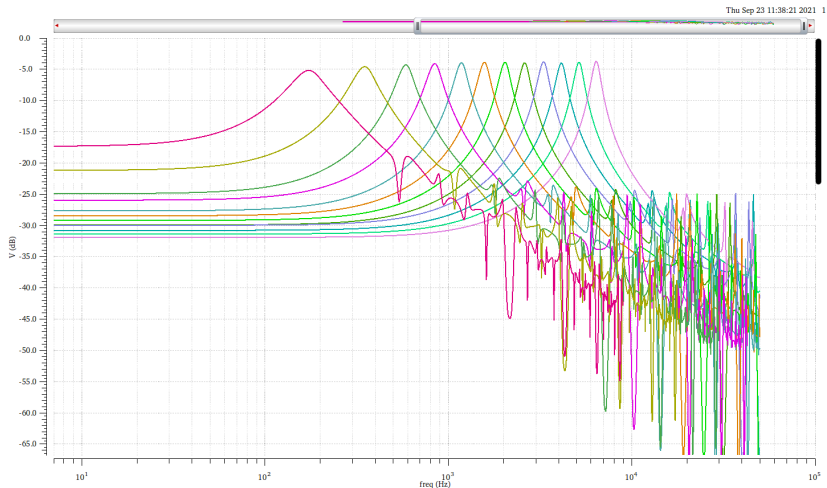


Figure: Spectre Periodic AC Analysis of the 12-channel Mel-Filterbank

Testbench & Simulation Results

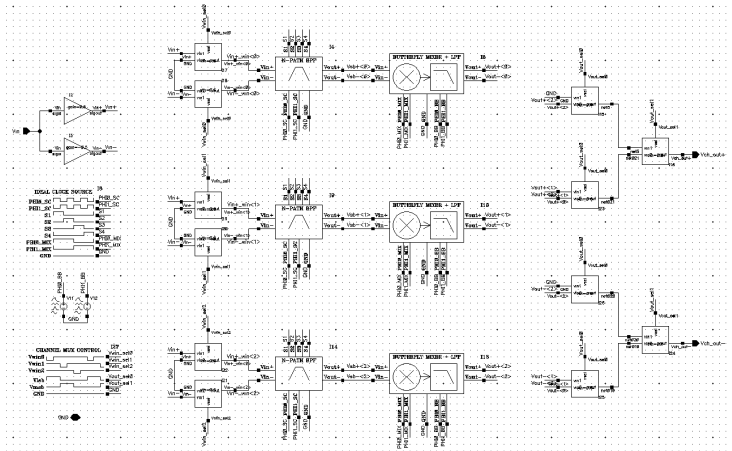
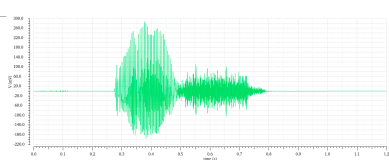
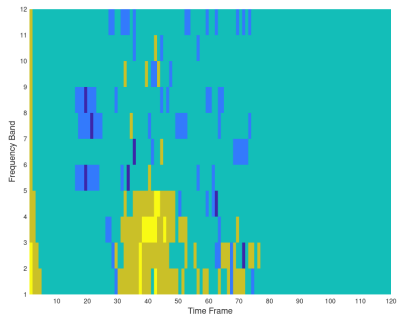


Figure: Schematic for single channel in Mel-Filterbank

Testbench & Simulation Results



(a) Speech input imported through a WAV file



(b) Spectrogram generated using simulation data of circuit implementation

Figure: Results of transient simulation with a test input. Although the results were similar to the expected output from ideal filterbank, the simulation time was very high which makes this procedure impractical for generating large training data. Another issue with this architecture is that the output is dependent on the input phase since we are not doing IQ Demodulation.








Conclusion

- 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, however, the simulation time was very large
- An issue of input phase dependence was identified which would require further analysis

Future Work

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

References

-  [Komail M. H. Badami et al.](#) "A 90 nm CMOS, 6 μ W 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.
-  [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.
-  [Minhao Yang et al.](#) "Design of an Always-On Deep Neural Network-Based 1- μ W Voice Activity Detector Aided With a Customized Software Model for Analog Feature Extraction". In: *IEEE Journal of Solid-State Circuits* 54.6 (2019), pp. 1764–1777. DOI: 10.1109/JSSC.2019.2894360.
-  [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.
-  [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.
-  [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.
-  [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.