DISS. ETH NO. 23110

Silicon Retina and Cochlea with Asynchronous Delta Modulator for Spike Encoding

A dissertation submitted to attain the degree of
DOCTOR of SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
MINHAO YANG
M.Sc., Peking University
born on 25.Nov.1985
citizen of Chongqing, P. R. China

accepted on the recommendation of
Prof. Tobias Delbrück, examiner
Prof. Christian Enz, co-examiner
Dr. Shih-Chii Liu, co-examiner

Nov.2015

Acknowledgements

This thesis would not come into existence without my supervisor Prof. Tobias Delbruck accepting me for my doctoral study in the first place. He has generously indulged me in pursuing my own research interests despite of the deviation from the original plan. The relaxed atmosphere he provided greatly helped me achieve the research goals steadily. He and my co-advisor Dr. Shih-Chii Liu have opened the door for me to explore the wonderland of bio-inspired spiking sensors including silicon retina and cochlea which are promising for future practical applications, and I shall continue the path on studying the means of power-efficient spike processing thanks to their initial inspiration.

The discussions on noise of retina pixels with Prof. Bernabé Linares-Barranco and Prof. Teresa Serrano Gotarredona at IMSE-CNM-CSIC, and the valuable inputs from Prof. Aurel Lazar at Columbia University on queueing theory and from Gaspard Hiblot at STMicroelectronics on DITS effect in silicon MOSFET are very much appreciated. I am in debt to Prof. Christian Enz at EPFL for agreeing to be my co-examiner. Prof. Yannis Tsividis at Columbia University kindly introduced me to his colleague Prof. Mingoo Seok for my postdoc stay funded by SNF, and I am grateful to them.

Fellow doctoral students amicably lent their hands along the way: Raphael Berner, Christian Brändli, Chenghan Li, and especially Chen-Han Chien who suffered from the almost impossible cochlea tape-out with me for several months. Raphael also helped me translate the abstract of the thesis into German. Colleagues Vicente Villanueva and Luca Longinotti affiliated with iniLabs GmbH helped with the PCB, firmware logic and jAER interface.

All my friends in Switzerland for dinner, movie, sports and philosophical discussions. My old friends outside of Switzerland for conversations and reminiscence.

My grandparents for their constant caring, and my parents for their eternal support and love.

Abstract

Spike encoding in bio-inspired sensory chips such as silicon retina and cochlea has been more of an empirical practice drawing inspirations from neuroscience since the dawn of neuromorphic engineering. From a system design and optimization point of view, the two natural questions to ask are how faithfully the encoded spike train represents the original analog input, and what minimum level of encoding fidelity is required according to some output performance specifications like the classification accuracy of a spiking neural network classifier. This thesis develops from the first question in the context of spiking sensors with massively parallel spike encoders, and covers both computational spike coding analysis and silicon chip design of spiking sensors.

The difficulty in rigorous mathematical analysis of spike encoding integrity especially when non-idealities in practical silicon implementations are considered particularly lies in its irregular sampling and nonlinear nature. Although simple linear spike decoding is cost-efficient in terms of computational and hardware expenses, it does not fully appreciate the timing information, i.e. the precise timestamps contained in the output spike train. Therefore it appears that to achieve the same signal-to-distortion ratio (SDR) which is used as the evaluation metric of encoding quality, linear decoding requires much more encoding quantization levels compared to nonlinear decoding based on the so-called frame theory. Nonetheless linear decoding reveals that, when feedback delay in the encoders is taken into account, the encoding mechanism of self-timed reset (STR) used in prior silicon retina results in lower SDR than the asynchronous delta modulation (ADM) commonly adopted for level-crossing ADCs. To further study the impact of jitter in spike timestamps caused by signal-dependent comparison delay during spike generation and spike transmission queueing due to limited communication bandwidth, nonlinear decoding algorithms are developed for STR and ADM. Circuit analysis and queueing theory are used to emulate the jitter generation with the examples of a specific comparator topology and two queueing models. A quantitative link is established between the system and circuit parameters and the decoding SDR metric, which is useful for future specifications-guided design of spiking sensory systems.

One type of silicon retina called dynamic vision sensor (DVS) encodes temporal contrast (TC) change of light intensity into output spike trains. Prior DVSs normally have a TC sensitivity of about 10% and all use STR spike encoders. Although DVSs have been successfully used in object tracking, for other potential applications like optical neuroimaging, the TC sensitivity and spike encoding quality need to be improved. The major advantage of adopting DVSs over prevailing APS imagers is much reduced output data rate given the sparsity of neuronal activity in neuroimaging, which could largely save power in wireless data transmission and thus facilitate continuous monitoring of free-moving animals. To improve the TC sensitivity, a low-noise transimpedance photoreceptor with pFET common-gate feedback and a programmable gain amplifier (PGA) employing a compact two-stage Opamp with pseudo-cascode compensation are used. To improve the spike encoding, a compact asynchronous switched-capacitor circuit is proposed for in-pixel ADM. The measured results show that a 1% TC sensitivity is achieved with 35% relative standard deviation across the pixel array, and using an exemplary visual stimulus input, up to 3.5× more spikes are preserved compared to a prior DVS with STR encoding. A simulated optical neuroimaging experiment is demonstrated.

To aim for ubiquitous smart audio sensing in the context of internet of everything, a 0.5-V ultra-low-power binaural silicon cochlea with ADM for spike encoding is designed. The spike output of a silicon cochlea is the natural input to event-driven DSPs and spiking neural networks for cognitive audio information processing. The cochlea has a parallel architecture with 64×2 channels. Besides bias distribution network, each binaural channel consists of a shared translinear loop for BPF *Q*-tuning, a pair of identical programmable attenuators, asymmetrical 1-zero 4-pole bandpass filters (BPFs), ADMs with adaptive self-oscillating comparators, and asynchronous logic blocks. The proposed asymmetrical BPF is composed of a low-power 4^{th}-order source-follower-based lowpass filter (LPF) and a summing PGA. The proposed ADM employs latched comparators instead of commonly-adopted continuous-time comparators based on multistage amplifiers in clockless systems for improved power efficiency. The reset signal needed to initialize the regenerative latch for each comparison is generated through a self-oscillation loop, and the oscillation frequency is adaptive to the output spike rate of a local channel. The measured power consumption of the 0.5-V core is about 55 μW at a 100k spike/s output rate and the system has a >70 dB dynamic range. Using the normalized power metric, this design is about 18× more power efficient compared to the best prior art. Moreover, the transfer functions between the corresponding binaural channels exhibit good matching and each channel has a wide *Q*-tuning range from 1 to about 40. This cochlea targets integration with ultra-low-power spike processors to form a smart audio sensing SoC.

Zusammenfassung

Die Impulskodierung (IK), welche in biologisch inspirierten Sensoren wie Silizium-Retinae und Silizium-Cochleae verwendet wird, ist seit den Anfängen der neuromorphen Ingenieurswissenschaften basiert auf empirischen Beobachtungen der Neurowissenschaften. Ausgehend von optimalem Systemdesign stellen sich zwei Fragen: Wie genau repräsentiert die Pulsefolge am Ausgang des Sensors das analoge Eingangssignal, und welche minimale Genauigkeit ist nötig um eine gewisse Leistung zum Beispiel eines Klassierers zu erreichen. Diese Doktorarbeit behandelt die erste Frage im Kontext von Sensoren mit vielen, parallel arbeitenden Impulskodierern.

IK im Detail mathematisch zu analysieren ist sehr schwierig, insbesondere wenn Nicht-Idealitäten von praktischen Schaltungen einbezogen werden sollen, da die Signalabtastung unregelmässig ist, und sich die Schaltungen in der Realität nicht-linear verhalten können. Eine einfache lineare IK ist effizient bezüglich Rechenaufwand und Hardwareanforderungen, jedoch wird damit die zeitliche Information der Impulsfolge nicht vollständig genutzt. Um debselben Signal-zu-Verzerrungs-Abstand (SDR) wie mit einer nicht-linearen, auf der sogenannten „Frame-Theorie“ basierten IK zu erreichen, müssen bei einer linearen IK deshalb mehr Quantisierungsstufen benutzt werden. Trotzdem zeigt schon lineare IK, dass der in bisherigen Silizium-Retinae verwendete Kodierungsmechanismus des „selbst-rücksetzens“ (STR) einen schlechteren SDR als der häufig in Schranken-Analog-zu-Digital-Konvertern verwendete asynchrone Delta-Modulator (ADM) zur Folge hat. Um auch den Einfluss von Ungenauigkeiten der Zeitstempel der Impulse genauer zu untersuchen, wurden nicht-lineare Dekodieralgorithmen für STR und ADM entwickelt. Die Ungenauigkeiten derZeitstempel werden verursacht durch signalabhängige Komparatorverzögerung und/oder Verzögerungen in der Kommunikation der Impulse durch die limitierte Bandbreite der Kommunikations-Schnittstelle. Eine Schaltungsanalyse der benutzen Komparatoren und zwei verschiedene Warteschlangenmodelle werden benutzt um die Zeitungenauigkeiten zu emulieren, und damit einen quantitativen Zusammenhang zwischen System- und Schaltungsparametern und dem SDR des Dekodierers herzustellen. Dieser Zusammenhang kann nützlich sein für die Entwicklung zukünftiger impulsbasierter Sensorsysteme.

Der am Institut für Neuroinformatik entwickelte dynamische Sehsensor (DVS) enkodiert die zeitliche Änderung (TC) des einfallenden Lichts in eine Impulsfolge am Ausgang. Bisherige DVS-Sensoren hatten eine TC-Empfindlichkeit von etwa 10% und benutzen ausschliesslich STR Kodierer. DVS werden benutzt für Standortverfolgung von Objekten, aber für Anwendungen zum Beispiel in neurologischen Bildgebungsverfahren müssen der TC und die Kodierqualität verbessert werden. Der Vorteil von DVS gegenüber konventionellen APS-Bildsensoren ist die reduzierte Datenmenge durch die geringe Dichte neuronaler Aktivität. Die verminderte Datenmenge würde einen tieferen Energieverbrauch von drahtlosen Sensoranbindungen und damit eine kontinuierliche Überwachung freilaufender Tiere ermöglichen.

Diese Doktorarbeit präsentiert einen neuen DVS-Sensor mit verbesserter TC-Empfindlichkeit. Der neue Sensor verwendet einen rauscharmen Transimpedanz-Fotorezeptor mit einem common-gate pFET im Signalrückführungspfad und einen programmierbarer Verstärker. Der benutzte zweistufige Operationsverstärker ist kompakt und verwendet eine Pseudo-Kaskoden-Kompensation. Um die Kodierqualität zu verbessern wird ein kompakter, asynchroner Schaltkreis mit geschalteten Kondensatoren verwendet, welcher einen Intrapixel ADM implementiert. Die Messungen zeigen eine TC-Empfindlichkeit von 1% und 35% relative Standardabweichung auf dem Pixelfeld. Mit einem spezifischen Stimulus werden im Vergleich zu bisherigen DVS-Sensoren mit STR–Kodierung bis zu 3.5-

mal mehr Impulse generiert. Die Fähigkeiten des Sensors werden auch an Hand eines simulierten Neuro-Bildgebungs-Experiments demonstriert.

In einem zweiten Teil dieser Doktorarbeit wurde im Kontext von intelligenten Audiosensoren für das „Internet of everything“ eine verbrauchsarme 0.5V Silizium-Cochlea mit ADM Impulskodierer entwickelt. Die Impulsfolge am Ausgang einer Silizium-Cochlea eignet sich sehr gut als Eingangssignal von impulsgesteuerten digitalen Signalprozessoren und neuronalen Netzwerken für die Audio-Signalverarbeitung. Die Cochlea hat eine parallele Architektur mit zwei Mal 64 Kanälen. Jeder binaurale Kanal besteht aus einem asymmetrischen Bandpassfilter (BPF) mit verteilter translinearer Schlaufe für die BPF Qualitätsabstimmung, einem Paar programmierbarer Attenuatoren, ADMs mit adaptiven, eigenschwingenden Komparatoren und asynchroner Logik. Der BPF hat eine Nullstelle und 4 Pole und besteht aus verbrauchsarmen Source-Folger-basierten Tiefpassfiltern vierter Ordnung und einem programmierbaren Summierverstärker. Um den Stromverbrauch zu senken benutzt der ADM einen verriegelten Komparator anstelle des oft in taktlosen Systemen verwendeten zeitkontinuierlichen mehrstufigen Komparators. Das Signal zur Rückstellung der Verriegelung wird generiert durch eine eigenschwingende Schaltung, die Schwingungsfrequenz adaptiert sich an die Impulsrate des lokalen Kanals. Der gemessene Stromverbrauch des 0.5V-Kerns der Cochlea ist 55 Mikrowatt bei einer Ausgangsrate von hunderttausend Impulsen pro Sekunde, und das System hat einen Dynamikumfang von mehr als 70dB. Wenn man den normalisierten Leistungsverbrauch vergleicht, ist das hier präsentierte System etwa 18 Mal effizienter als der Stand der Technik. Die Übertragungsfunktionen der binauralen Kanäle zeigen eine gute Abstimmung, und jeder Kanal hat einen grossen Qualitäts-Abstimmung-Bereich von 1 bis 40. Diese Cochlea könnte zusammen mit sehr verbrauchsarmen Impulsprozessoren integriert werden um ein intelligentes Audiosensor-System zu bilden.

Contents

Abstract ii

Zusammenfassung iv

Contents vi

List of Figures ix

List of Tables xiii

Chapter 1: Introduction .... - 1 -

1.1 Brief History of Spiking Sensors .... - 1 -

1.1.1 Spiking Silicon Retina .... - 1 -

1.1.2 Spiking Silicon Cochlea .... - 3 -

1.2 Other Related Works .... - 5 -

1.2.1 Revival of Neuro-Inspired Computing .... - 5 -

1.2.2 Computational Spike Encoding/Processing/Decoding .... - 6 -

1.2.3 Clockless Continuous-time ADC/DSP/DAC Systems .... - 7 -

1.3 Thesis Contribution & Organization .... - 9 -

Chapter 2: Analysis of Spike Encoding Mechanisms in Spiking Sensors .... - 11 -

2.1 STR and ADM with Fixed Feedback Delay T_D .... - 12 -

2.1.1 Modeling of STR and ADM .... - 12 -

2.1.2 Linear Decoding Comparison .... - 13 -

2.2 SDR Degradation due to T_D Variation .... - 15 -

2.2.1 Nonlinear Decoding .... - 15 -

2.2.2 Two Sources of T_D Variation .... - 18 -

2.2.3 SDR Degradation .... - 27 -

2.3 Conclusion and Remarks .... - 28 -

2.4 Appendix .... - 29 -

Chapter 3: Silicon Retina with Enhanced Temporal Contrast Sensitivity and Spike Encoding .... - 31 -

3.1 System Architecture .... - 31 -

3.2 Pixel Design .... - 33 -

3.2.1 Photoreceptor .... - 33 -

3.2.2 Capacitively-Coupled Programmable Gain Amplifier (CC-PGA) .... - 39 -

3.2.3 Asynchronous Delta Modulator (ADM) .... - 47 -

3.2.4 In-Pixel Asynchronous Logic (IPAsyncL) .... - 50 -

3.3 Experimental Results ..... - 52 -
3.3.1 Pixel Layout and Chip Microphotograph ..... - 52 -
3.3.2 Split-Transistor Gain ..... - 52 -
3.3.3 Sensor Array Noise ..... - 53 -
3.3.4 Temporal Contrast Sensitivity ..... - 54 -
3.3.5 Spike Encoding Comparison ..... - 56 -
3.3.6 Simulated Optical Neuroimaging using ADMDVS ..... - 58 -
3.4 Conclusion and Remarks ..... - 59 -
3.5 Appendix ..... - 61 -
3.5.a Spike Rate of ADM ..... - 61 -
3.5.b From Screen Luminance to Chip Illuminance ..... - 62 -
Chapter 4: Ultra-Low-Power Binaural Silicon Cochlea ..... - 63 -
4.1 Architecture ..... - 63 -
4.2 0.5-V Cochlea Core Design ..... - 64 -
4.2.1 64 Geometrically-Scaled Channel Bias Currents ..... - 64 -
4.2.2 Translinear Loop for *Q*-Tuning and In-Channel Bias Distribution ..... - 66 -
4.2.3 Programmable Capacitive Attenuator ..... - 67 -
4.2.4 Source-Follower-Based Asymmetrical Bandpass Filter ..... - 69 -
4.2.5 Asynchronous Delta Modulator with Self-Oscillating Comparison ..... - 86 -
4.3 System Design and Chip Microphotograph ..... - 94 -
4.4 Measurement Results ..... - 95 -
4.4.1 BPF Transfer Functions ..... - 96 -
4.4.2 BPF Noise, Distortion and Dynamic Range ..... - 97 -
4.4.3 CMRR and PSRR ..... - 99 -
4.4.4 Spike Output ..... - 100 -
4.4.5 Comparison with Prior Arts ..... - 106 -
4.5 Conclusion and Discussion ..... - 108 -
4.6 Appendix ..... - 109 -
4.6.a Bandwidth of the Summing CC-PGA ..... - 109 -
4.6.b Transient Analysis of the ADM Amplifier ..... - 109 -
4.6.c CT- and DT-Comparator Power Consumption Comparison ..... - 111 -
Chapter 5: Addressable Current Reference Array for Spiking Sensors ..... - 113 -
5.1 Coarse-Fine System Architecture ..... - 113 -

- 5.2 Circuit Implementation ... - 115 -
  - 5.2.1 PTAT Master Bias ... - 115 -
  - 5.2.2 Eight Coarse Currents ... - 115 -
  - 5.2.3 Coarse-Fine Current Interface ... - 117 -
  - 5.2.4 8-bit Current DAC ... - 119 -
  - 5.2.5 Configurable Output Buffer ... - 120 -
- 5.3 Experimental Results ... - 121 -
- 5.4 Conclusion and Remarks ... - 122 -

Chapter 6: Summary and Future Work ... - 123 -

Bibliography ... - 126 -

Curriculum Vitae ... - 136 -

List of Figures

1.1 Simplified illustration of a dynamic vision sensor (DVS) pixel 2
1.2 Cascaded second-order sections (SOSs) and the detailed circuits of one SOS 3
1.3 Zero-crossing and integrate-and-fire spike encoding 4
1.4 TEM/TDP/TDM system 6
1.5 Clockless CT-ADC/DSP/DAC system 8
2.1 Ideal spike encoding model of STR named as silicon ON-OFF TEM 12
2.2 Spike encoding model of STR and ADM with delay time and switch-hold period 13
2.3 Waveforms of STR and ADM, and the linear reconstruction error 14
2.4 Linear spike decoder comprised of an ideal integrator and a brick-wall LPF 14
2.5 SDR improvement of ADM over STR 15
2.6 Illustration of the *t*-transformation principle for STR and ADM 16
2.7 Reconstruction error of STR and ADM using nonlinear decoding 17
2.8 Simulated comparison delay versus input signal slope 19
2.9 Queueing models of 1-D and 2-D sensors 20
2.10 Mean waiting time and waiting time distribution of an M/D/1 queue 22
2.11 Mean waiting time and waiting time distribution of an $M^X/D/1$ queue 24
2.12 Waveform of the white Gaussian noise input and exemplary spike train output 25
2.13 Reconstruction SDR as the function of the comparison speed indicator y_0 and exemplary recovered signals 26
2.14 SDR of STR and ADM versus the ratio T_D/T_{Davg} 26
2.15 SDR degradation in 1-D silicon cochlea and in 2-D silicon retina 27
3.1 System architecture of the ADMDVS chip 32
3.2 Building blocks of one pixel 33
3.3 Circuit diagrams of the SFPR and the CGPR 34
3.4 Small signal equivalent circuits of the SFPR and the CGPR 34
3.5 Spectre simulation and theoretical calculation of the output noise of SFPR and CGPR 40
3.6 Previous circuit solutions to high in-pixel gain 40
3.7 CC-PGA circuit diagram with the transistor implementation of the pseudo-resistor 41
3.8 Effect of finite off-state switch resistance on CC-PGA 41
3.9 Schematic of the CC-PGA Opamp and transistor channel length splitting 43
3.10 Schematic of the ADM and timing diagram of the switch control signals 47
3.11 Representative transistor-level implementation of STR 49
3.12 Circuit diagram of the IPAsyncL and timing diagram for ON spike transmission. 50

3.13 2×2 pixel layout and chip microphotograph 51

3.14 Measured DC-gain ratio of split to non-split nFETs and measured open-loop DC-gain distribution of the Opamp used in CC-PGA 52

3.15 Setup for measuring ADMDVS noise; measured average noise spike rate of the ADMDVS array and calculated average equivalent noise voltage at CGPR output 53

3.16 Setup for measuring the average TC sensitivity 54

3.17 Measured average TC sensitivity and its relative standard deviation across array 55

3.18 Effect of different TC sensitivity settings on detecting fine palm lines 55

3.19 Testing setup and the original image for the spike encoding comparison experiment 57

3.20 Comparison of the spike-accumulated histogram images acquired by ADMDVS and DAVIS at different rotational speeds of the testing image 57

3.21 Comparison of the average number of ON/OFF spikes over the 108 brightest/darkest pixels in the histogram images 58

3.22 Measured grayscale-luminance relationship of the screen and one frame of the optical neuroimaging video used for simulated experiment 59

3.23 Temporal waveform averaged over a 5×5 pixel window within the target neuron of the video and the linear signal reconstruction using output ON and OFF spikes 59

4.1 Complete binaural silicon cochlea system 64

4.2 Schematic of the circuit that generates 64 geometrically-scaled bias currents 65

4.3 Translinear loop circuit that generates the BPF biases I_{BPF1} and I_{BPF2} and R-2R programmable current DAC that is used to adjust I_{BPF1} 66

4.4 Complete in-channel bias distribution network 66

4.5 3-bit programmable capacitive attenuator 67

4.6 RC-equivalent circuits of the attenuator considering parasitic R and C 68

4.7 Attenuator AC characteristics and its output IM3 at 1 V_{pp} input with no attenuation 69

4.8 Original SF-based LPF biquad and its alternative topology 70

4.9 RC-equivalent circuits of the LPF biquads in 4.8 70

4.10 Existing SF-based BPF biquads in literature 71

4.11 RC-equivalent circuits of the BPF biquads in 4.10 71

4.12 Proposed SF-based BPF biquad 72

4.13 The impact of current source mismatch at Q=1 and Q=10 73

4.14 Half-circuit of the proposed SF-based BPF biquad with noise sources 75

4.15 Simulated noise PSD of the proposed biquad at Q=1 and Q=10 76

4.16 Complete asymmetrical SF-based BPF 78

4.17 Simulated transfer functions of the 4^{th}-order asymmetrical BPF at different Q's 79

4.18 Simulated 3σ variation of Q, central frequency and peak gain at different mean Q's 80

4.19 Simulated maximum BPF rms output with specified distortion and output-referred noise (ORN); accordingly calculated DR, all at different mean Q's 81
4.20 Circuit diagram of the fully differential CC-PGA 82
4.21 Circuit diagram of the fully-differential two-stage Opamp used in the CC-PGA 83
4.22 Detailed circuit diagrams of the compensation capacitors in the Opamp in 4.21 84
4.23 Conventional CMFB loop and the CMFB loop used in the proposed Opamp in 4.21 85
4.24 Block diagram of the ADM with adaptive self-oscillating comparison 87
4.25 Circuit diagram and symbolic abstraction of the unity-gain inverting amplifier 88
4.26 Circuit diagram of the two-stage Opamp used in ADM 89
4.27 Detailed switched-capacitor circuits used in ADM 90
4.28 Circuit diagrams of the ON- and OFF-comparators, the asynchronous logic that generates the *CLK* and *nCLK* for the comparators, and the EN_{busy} generator 91
4.29 Impact of the threshold setting on maximum allowable comparison delay 92
4.30 In-channel asynchronous logic circuits and exemplary timing diagram of an ON spike 93
4.31 Chip microphotograph of the silicon cochlea 94
4.32 Measured transfer functions of the right ear at $Q\approx1.3$ and $Q\approx9.2$ 95
4.33 Central frequency scaling as the function of channel number 96
4.34 Mismatch of the central frequencies and Q's between two ears at $Q\approx1.3$ and $Q\approx9.2$ 96
4.35 Q-tuning examples of 4 different channels 97
4.36 Noise PSD of channel 00 and channel 54 at the output of the summing PGA 97
4.37 Measured distortion of channel 00 and 54 with the PGA gain of 18 dB 98
4.38 Measured CMRR at the output of the ADM amplifier 100
4.39 Measured PSRR at the output of the ADM amplifier 100
4.40 Spike cochleagram in response to an exponential chirp input 101
4.41 Output spike train samples from channel 18 in response to a single-tone input; the spectra and time-domain waveforms of reconstructed signals 101
4.42 Output spike train samples from the all channels and channel 18 in response to a two-tone input; the spectra and time-domain waveforms of reconstructed signals 102
4.43 Reconstruction method for speech synthesis from spike trains 103
4.44 Spectrogram and time-domain waveform of the input male digit speech; output spike cochleagram, and spectrogram and time-domain waveform of the reconstructed speech at $Q\approx1$and $Q\approx10$ 104
4.45 Spectrogram and time-domain waveform of the input female sentence speech; output spike cochleagram, and spectrogram and time-domain waveform of the reconstructed speech at $Q\approx1$and $Q\approx10$ 105
4.46 Spectrogram of 'heed' and 'had'; normalized histogram of generated spikes

in response to the 'heed' and 'had' input at Q≈1 and Q≈10 106
4.47 Alternative summation scheme and alternative circuit topologies to obtain one zero 108
4.a1 ADM amplifier models considering three different kinds of non-idealities 110
4.a2 Comparison of StrongARM latch comparator and two-stage amplifier comparator 110
5.1 Coarse-fine architecture of the addressable current reference array; timing diagram of the main digital signals for programming bias currents 114
5.2 Circuit diagram of the PTAT master bias 115
5.3 Circuit diagram of a compact current multiplier 116
5.4 Complete circuit diagram of the generator of the eight coarse currents 117
5.5 Conventional and proposed current selection schemes 118
5.6 Complete circuit of the coarse-current selection 118
5.7 Auto-configured current mirror to convey the selected coarse current to an 8-bit current DAC for fine-current selection 119
5.8 8-bit R-2R current-splitting DAC for fine-current selection 119
5.9 Fine-current buffer to generate the output bias voltage 120
5.10 Layout of the current reference array 121
5.11 Measured currents as the function of the 8-bit fine code and the 3-bit coarse code 122
6.1 Comparison of the architectures of a smart sensor node comprising sensing, signal conversion and processing stages 124

List of Tables

2.1 Parameter values for computing mean and distribution of queueing waiting time 22

3.1 Comparison of the output voltages of SFPR and CGPR 39

3.2 Sizes of the transistors in SFPR and CGPR 39

3.3 Stimulus contrast values used for measuring TC sensitivity 54

3.4 Comparison with previous DVSs 60

4.1 Simulated means and 3σ variations of some of the output currents in Figure 4.2 65

4.2 Simulated means and 3σ variations of the multiplication and division of I_{BPF1} and I_{BPF2} 67

4.3 *Q* value as the function of the bias ratio I_{BPF2}/I_{BPF1} 79

4.4 Body biasing of the transistors in the Opamp circuit in Figure 4.21 83

4.5 Body biasing of the transistors in the amplifier in Figure 4.25 88

4.6 Body biasing of the transistors in the amplifier in Figure 4.26 89

4.7 Programmable values of δ-subtraction of the ADM amplifier 90

4.8 Measured integrated rms noise at the PGA output 97

4.9 Measured PGA rms output amplitude at THD=-40dB and IM3=-26dB, and the calculated SNR 99

4.10 CMRR and PSRR at central frequencies measured at the output of the ADM amplifier 99

4.11 Cochlea chip performance summary and comparison with prior works 107

5.1 Simulated nominal coarse currents under room temperature 116

Chapter 1: Introduction

While neuroscientists are delving deep into the working mechanisms of the ultra-intricate neural networks in our central nervous system that has the number of synaptic connections a thousand times larger than the number of the stars in our Milky Way galaxy, engineers have been going to great lengths to emulate the extraordinary capabilities of our information sensing and processing organs, particularly the retina, the cochlea and the neocortex. This thesis is about another engineering endeavor at evolving artificial bio-inspired sensors, including the spiking silicon retina and cochlea.

1.1 Brief History of Spiking Sensors

1.1.1 Spiking Silicon Retina

In the late 1980s, Carver Mead brought in the IC community the concept of neuromorphic engineering trying to replicate the morphologies and functions of biological nervous systems with the fast-evolving silicon technologies, and a number of his protégées devoted great efforts in bridging the knowledge gap between biology and electronics. One such representative figure is Misha Mahowald who built the first ever silicon retina [1]. In a most simplified way, it mimicked the four layers of cells in a mammalian retina, namely the photoreceptor, the horizontal cell, the bipolar cell and the ganglion cell. Back then the design philosophy was dominated by mimicking biological details and did not emphasize possible practical applications and performance specifications. The results were demonstrations of interesting engineering feat with biological flavors other than something practically useful. Consequently, despite the pioneering creation of many circuit and system rudiments that are still academically in use today such as the adaptive logarithmic photoreceptor and the address event representation (AER) protocol for spike transmission, the industrial attention was mostly drawn to the famous active pixel sensor (APS) [2] which has a more straightforward working principle to comprehend and offers more reliable performances for commercial production with wide applications in consumer digital cameras, high-end telescopes for astrophysics, optical bioimaging, computer/machine vision etc.

Proliferated in 1990s, APS imagers capture images or record videos based on the notion of frame, the 2-D equivalence of the classic 1-D sampling with equidistance time steps. This uniform sampling has a well-established mathematical foundation [3], and is ubiquitously used in analog-to-digital converters (ADCs) to transform real-world continuous analog signals into discrete digital representation, both in amplitude and time. An APS pixel normally has a very simple circuitry composed of three or four transistors to transform photons to sampled voltages, which has substantial advantages in spatial resolution and uniformity. Because of these advantages, APS imagers have become the de facto standard in image/video acquisition. On the other hand, the original silicon retina and the subsequent improvements to morph all the five retinal layers (including the amacrine cells) [4] suffered from significant pixel mismatch due to the pixel complexity, making them barely useable. In addition, asynchronous non-uniform sampling associated with the spike generation of each individual pixel and the asynchronous (arbitrated) spike transmission in a sensor array is not as well understood by electronic engineers as its counterpart of synchronous uniform sampling.

The so-called octopus imager published in early 2000s [5] indicates a design style shift of spiking retinae. The grand idea of faithfully morphing biological retinae was displaced by keeping the functional essence for performance gain. Each pixel performs a nearly linear conversion of light intensity to spike rate, and the spike transmission follows the AER protocol. Although the 80×60 pixel array achieved large dy-

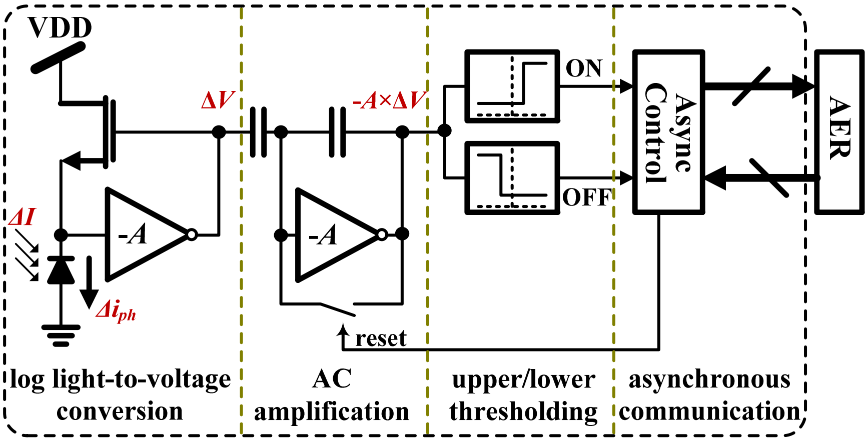


Figure 1.1. Simplified illustration of a dynamic vision sensor (DVS) pixel [6].

namic range, high effective frame rate, relatively low power, and reasonably small fixed pattern noise (FPN), the light-intensity-to-frequency encoding, i.e. the pulse frequency modulation (PFM), is not an efficient way of conveying information considering its temporal redundancy and in turn the heavy load on data transmission and post-processing.

Reduction of temporal redundancy could be done via frame-based methods by subtracting the sampled voltages in each pixel of two consecutive frames [7], [8], but the repetitive frame scan still wastes unnecessary energy especially when there is no temporal change at all, and the temporal resolution is limited to frame rate. It has been discovered that human eyes rely on constant microsaccades to prevent visual fading during fixation [9], which indicates that the ganglion cells in human retina produce almost no spike if no temporal contrast of light intensity ΔI is detected. The dynamic vision sensor (DVS) invented in 2006 by Lichtsteiner et al. [6], [10] exploited this mechanism of spike generation. As illustrated in Figure 1.1, the pixel functions as follows: light intensity temporal contrast ΔI is converted to voltage change ΔV; ΔV is amplified by A times; the amplified AC signal is compared with two comparators with upper and lower thresholds; the asynchronous control block generates a spike if an ON or OFF threshold crossing is detected; the spikes are communicated off-chip via the on-chip peripheral AER circuits; the capacitively-coupled amplifier is reset once the acknowledge in response to the request of this pixel comes back. Because of the AC coupling, a DVS pixel only generates a spike if the ΔI within some frequency passband is sufficiently large, which means no output spike is generated for a static scene or a scene with merely very high frequency content. This is similar to the experimental observations from human eyes [11], except the eye bandwidth is restricted within tens of Hz whereas the DVS can response to stimuli above kHz depending on the ambient illuminance and the circuit bias settings. The DVS is a big step for spiking retinae towards practical applications where dynamic content recording and analysis is of particular interest. The pixel-autonomous and frameless operation with in-pixel self-timed reset spike encoding largely reduces the temporal redundancy of output data and breaks the tight trade-off between data rate and power consumption in conventional APS imagers [12], [13]. The pixel uniformity is also greatly improved compared to prior spiking retinae thanks to the signal amplification before threshold crossing and the well-matched gain among pixels using capacitance ratio. Further works around DVS include high-speed application in optical line sensors [14], improving temporal contrast sensitivity [15], [16] and adding the function of absolute light intensity acquisition [17], [18].

Extracting spatial contrast on the focal plane is also of interest because it reduces spatial redundancy in

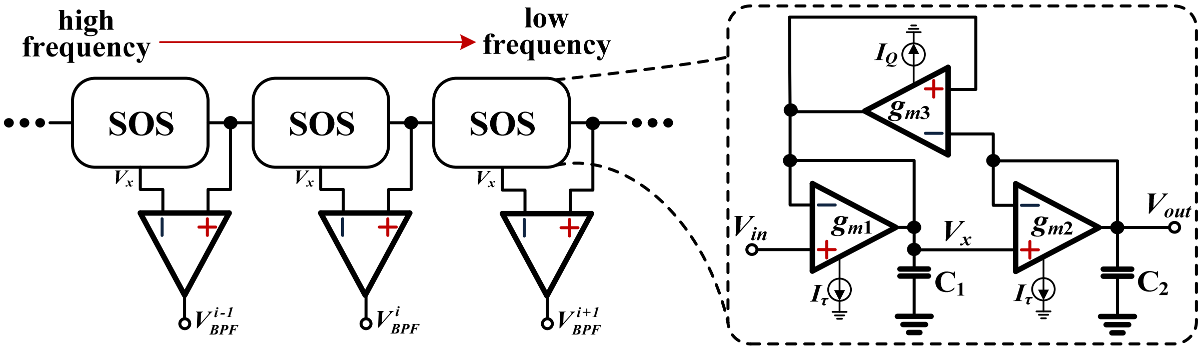


Figure 1.2. Cascaded second-order sections (SOSs) and the detailed circuits of one SOS [23].

output data while keeping relevant contour information for shape and object recognition, which is dual in functionality to temporal redundancy reduction. Frame-based solutions [19], [20] give good output uniformity and low power consumption. Spiking sensors that use diffusive grid connecting neighboring pixels [21] for spatial contrast computation still suffer from large component mismatch which needs in-pixel calibration and consequently results in large pixel size with very small fill factor (e.g. 2% in [22]).

1.1.2 Spiking Silicon Cochlea

Early in-silico cochlea modeling in Mead's lab focused on the frequency division of the basilar membrane as a function of place [23], [24]. Multiple (tens to hundreds of) second-order sections (SOSs) with each composed of two feedforward g_m-C integrators and a feedback OTA are connected in series as shown in Figure 1.2. The bandpass filtering (BPF) outputs are obtained by subtracting V_x and V_{out} in each SOS. If $g_{m1}=g_{m2}=g_{m\tau}$, $g_{m2}=g_{mQ}$, and $C_1=C_2=C$, the BPF transfer function of one single SOS can be then written as:

$$H(s)_{SOS}=\frac{V_{out}-V_x}{V_i}=-\frac{\frac{g_{m\tau}}{C}s}{s^2+\frac{2g_{m\tau}-g_{mQ}}{C}s+\frac{g_{m\tau}^2}{C^2}}=-\frac{\omega_0 s}{s^2+\frac{\omega_0}{Q}s+\omega_0^2}$$

where ω_0 is the angular central frequency and Q is the quality factor, which can be written as:

$$f_0=\frac{\omega_0}{2\pi}=\frac{g_{m\tau}}{2\pi C},\quad Q=\frac{1}{2-g_{mQ}/g_{m\tau}}$$

where f_0 is the central frequency. All f_0's of the cascaded SOSs are proportional to the bias current I_τ of each individual SOS, and are geometrically scaled. Q's are controlled by the bias current I_Q. The first SOS in the serial chain with its input normally connected to a preamplifier is a 2nd-order symmetrical BPF, and due to the accumulation of poles, the following stages have increasingly larger lowpass roll-off, theoretically 40 dB/decade more after each stage. The highpass roll-off is however always 20 dB/decade because the single zero is created by the subtraction of V_x and V_{out} in each SOS.

There are several problems with the cascaded structure: 1. I_τ was generated by using linearly-graded voltages to bias nFETs in subthreshold and thus had large mismatch; 2. Although the BPF transfer functions have progressively steeper roll-off as f_0 decreases, the frequency selectivity is poor. Partly this is because the Q of each SOS has to remain low (usually <2) to prevent the prohibitively large passband

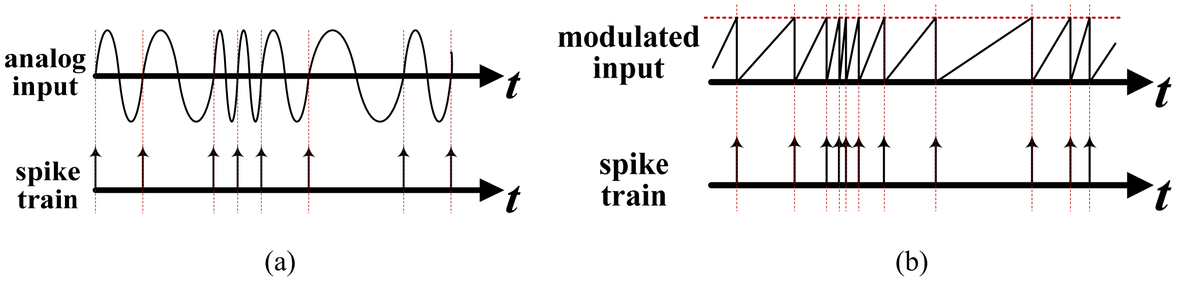


Figure 1.3. (a) Zero-crossing spike encoding; (b) Integrate-and-fire spike encoding.

gain due to gain accumulation; 3. Noise is aggregated along the cascaded chain, so the SNR of the low-frequency stages could be much lower than the input stage. The random mismatch was addressed by using the compatible lateral bipolar transistors (CLBTs) in standard CMOS instead of subthreshold nFETs [25], which is basically the parasitic bipolar transistors often used in bandgap reference as the substitute of diodes [26]. The results showed significant improvement in linearity of logarithmic scaling of f_0 over SOS stages. The low Q and noise aggregation problem can be circumvented by using a parallel architecture even though it does not quite resemble the structure of the basilar membrane that allows cascaded propagation of mechanical waves. In the analog bionic ear processor presented by Sarpeshkar et al. [27], the 16 parallel channels employing input-attenuated 4th-order symmetrical g_m-C BPFs [28] is said to be capable of having Q up to 10. Parallel filter banks nevertheless fall short in providing steep roll-off unless multiple filter biquads are used. For example, 8th-order one-zero gammatone filter transfer function with 8 poles and 1 zero was implemented using 3 lowpass and 1 bandpass pseudo-differential class-AB logarithmic current-mode biquads [29]. The area is 2.25 mm^2 for one channel in 0.35 μm CMOS and the power consumption is 3.4 μW without AGC at f_0=3.3 kHz, whereas in [27] the area for one BPF is about 2.8 mm^2 in 1.5 μm BiCMOS and the power consumption is 5.4 μW at f_0=7.1 kHz. Considering the difference in the technology nodes, area and power are clearly the penalty for higher roll-off with more biquads.

Besides the cascaded and parallel architectures, new models have been proposed to better emulate the human cochlea. One example is to use the resistive passive coupling among parallel resonators to model the cochlea fluid [30]–[32], and another example takes inspiration from the function of the outer hair cells and uses active bidirectional coupling to counteract the destructive interference among channels in passive coupling [33]. The efforts towards more bio-realistic modeling on analog VLSI were unfortunately buried in unsatisfactory performances mainly due to the obstinate mismatch problem. In [31], the f_0 variation is as large as 15% while the ideal f_0 ratio between neighboring BPFs is less than 1.05, and in [33], the expected Q_{10} (f_0 divided by the bandwidth taken at 10 dB below the peak) enhancement in software simulation is hindered by abrupt changes in basilar membrane properties in silicon.

Integration of spike encoders together with basilar membrane has rarely been implemented in silicon. About six such chips sparsely scattered spanning twenty years from 1990s to 2010s, and their encoding mechanisms mainly include zero-crossing [27], [34] and integrate-and-fire [33], [35]–[37], as illustrated in Figure 1.3. Zero-crossing rate is useful for relatively primitive sound processing, e.g. zero-crossing-based ultra-low-power voice activity detection [38]. However, zero-crossing alone usually does not constitute a complete representation of the original signal [34], [39] and thus certain information like amplitude could be lost or corrupted during encoding. On the contrary, integrate-and-fire can theoretically have perfect reconstruction even with absolute refractory period if certain conditions are satisfied [40]. For real-world-scenario applications, besides digit recognition that used the cochlea with zero-crossing encoding [34], the spike trains of the binaural cochlea with integrate-and-fire encoding called AER-EAR was readily used for sound source localization by exploiting the interaural time difference even though the implemented encoders suffered from substantial mismatch among channels in one monaural cochlea and between binaural channels [36], [37].

Conventional sound acquisition uses high-precision synchronous ADCs, and the frequency analysis is done in digital domain by FFT which can still be power-consuming even with the state-of-the-art development in 32 nm CMOS [41]. For always-on smart sensing like voice activity detection, a spike-driven processor built on spiking silicon cochlea has the potential advantage of reducing the system power consumption due to the low-power analog filtering and adaptive digital power in response to the incoming spike trains. A recently developed voice activity detector with cochlea-like 16-channel analog BPFs and a mixed-signal decision-tree machine leaning kernel has shown a ×10 reduction in power compared to prior arts [42], even though the microcontroller for configuring the decision tree is clocked and off-chip.

1.2 Other Related Works

1.2.1 Revival of Neuro-Inspired Computing

14 nm processors are in mass production from Intel and Samsung. Moore’s law is predicted to have only 5 to 10 more years before it literally hits the end according to ITRS. The technology scaling that has been successfully lasting for half a century has undoubtedly aided the seamless penetration of electronics into our daily life, and yet nothing that is sufficiently intelligent to make a decision on its own in complex situations has come to life. Of course we should count out supercomputers sitting in a cozy lab and consuming immense electricity. The ever-increasing motivation of maintaining the prospects of the commercially profitable semiconductor industry at the end of Moore’s law has spurred the diverse quests for alternative means of building smarter and more energy-efficient systems such as carbon-nanotube-based computing [43], inexact computing [44], [45], etc. Revived to be the neuro-inspired computing with more emphasis on taking functional inspiration other than trying to be bio-realistic, neuromorphic engineering has also reemerged to the surface of common technological interests after years of relatively quiet progress. In August 2014, IBM’s announcement of the TrueNorth chip with a million digital spiking neurons has attracted a lot of public attention [46]. A visual recognition task was demonstrated with 72 mW power consumption at 775 mV power supply which is ×176,000 and ×769 more energy-efficient compared to running on a modern general-purpose microprocessor and the state-of-the-art multiprocessor neuromorphic platform SpiNNaker [47], respectively. TrueNorth largely deviates from the stereotyped Von Neumann architecture of centralized memory and processing units, and instead is composed of massively parallel neurons and adopts the distributed slow asynchronous spike communication among synaptic cores other than high-speed digital communication which is inevitable in conventional digital processors. As pointed out in [46], spiking sensors like the DVS and AER-EAR could be the perfectly suitable candidates for TrueNorth in live settings for real-time applications because their spike train output is exactly what TrueNorth needs as input. It is interesting to see that the campaign of neuro-inspired computing is expanding beyond the traditional neuromorphic engineering community, and has drawn the attention of engineers who are more on the industry-oriented IC design side: Jan Rabaey at UC Berkeley is travelling around the world to give the presentation 'The Return of Neuro-Inspired Computing – Why Now', Zhengya Zhang at University of Michigan is developing digital neural network IC with on-chip learning for image feature extraction [48], Mingoo Seok at Columbia University is also developing digital spiking neural network chip for brain-machine interface applications like spike sorting, just to name a few.

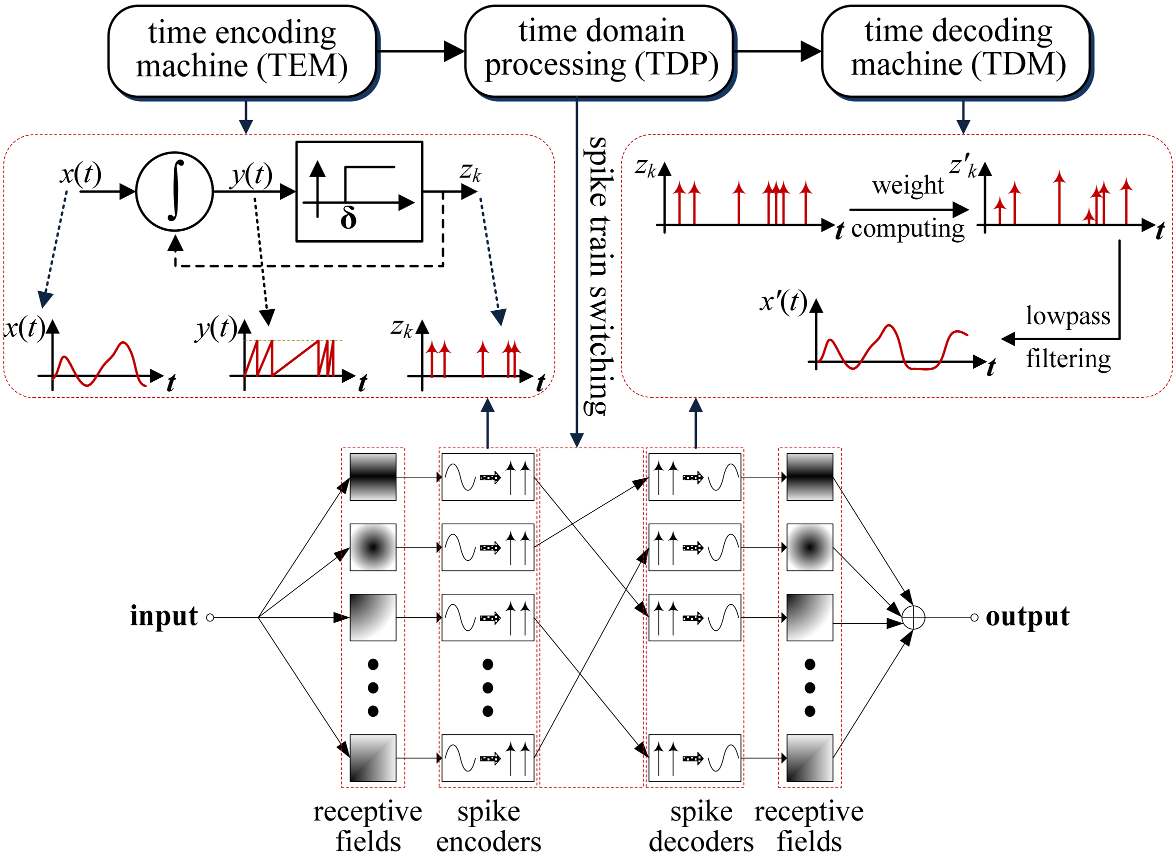


Figure 1.4. A TEM/TDP/TDM system, where information is represented and processed in time domain in the form of continuous-time and discrete-amplitude spikes [49].

1.2.2 Computational Spike Encoding/Processing/Decoding

The intense debate about the nature of neural coding within the neuroscience community – whether it is rate coding or temporal coding – has not yet settled. At least, from an engineering point of view, rate coding is far less data- and power-efficient than temporal coding, e.g. comparing the Octopus imager with the DVS. Although the Octopus is capable of acquiring absolute intensity information, temporal coding can do the same by using the information of inter-spike interval of a single spike pair [17]. Aurel Lazar shares the same belief in temporal coding. He pondered the system TEM/TDP/TDM as illustrated in Figure 1.4 [49], where TEM, TDP and TDM stand for time encoding machine, time domain processing and time decoding machine, respectively, in analogy to the conventional clocked ADC/DSP/DAC system. Signals are represented and processed all in time domain in the form of spike trains. The precise timestamps of the encoded spikes are deemed to contain all the information about the original input given that the parameters of the spike encoders are known. The importance of precise timestamps is substantiated by the observation of the cat lateral geniculate nucleus (LGN) where the temporal structure of the spiking response of an LGN neuron was found to have a finer timescale than that of the input stimuli filtered by the neuron's temporal receptive field (RF), and large timing jitter of spikes with its reciprocal comparable to the average firing rate results in much degraded signal reconstruction [50]. Lazar approached the analysis of the TEM/TDP/TDM system dealing with precise timestamps in a rigorous signal-processing way, and studied many forms of TEM and their corresponding TDM, from single-input and single-output neuron to multi-input and multi-output neuron ensemble [51].

Mathematically TEM maps input analog waveforms to output spike timestamps following the orthodox *t*-transformation. Taking the simplest integrate-and-fire encoding depicted in Figure 1.4 as an example with immediate spike-triggered feedback reset and no refractory period, its *t*-transformation can be written as:

$$\int_{t_k}^{t_{k+1}} x(t)dt = \delta$$

where δ is the encoder threshold, $x(t)$ is the analog input, and t_k is the output k^{th} timestamp with $k \epsilon Z$. It has been shown that perfect reconstruction of $x(t)$ from the output spike train z_k is possible if the average spike rate is at least as large as the Nyquist rate and the dimensionality of z_k is infinite [52]. A TDM implements the reconstruction process by giving appropriate weight to each spike in z_k and then passing the weighted z'_k through a brick-wall lowpass filter (LPF) with the same bandwidth of $x(t)$. Each spike is treated as an ideal Dirac-delta function so that its impulse response of an ideal LPF is a sinc function, and therefore the reconstructed signal $x'(t)$ can be seen as superposition of differently-weighted sinc functions that are time-shifted to t_k. The method of calculating the weights is adapted from irregular sampling based on frame theory [51]. Simply put, in contrast to a basis with linearly independent elements in a vector space V equipped with an inner product, the elements in a frame may be linearly dependent and thus could have more than one set of coefficients to represent an arbitrary element in $\boldsymbol{V}$. To form a frame $\{f_k\}_{k \epsilon Z^*}$, it needs to satisfy the inequality below:

$$A \| v \|^2 \leq \sum_{k \in Z^*} |< v, f_k >|^2 \leq B \| v \|^2 \quad \forall v \in \boldsymbol{V}$$

where A and B are the frame bounds, and have 0<A,B<∞. In TDM, a frame can be formed by the time-shifted sinc functions. TDP was first demonstrated in a video coding setup as shown in Figure 1.4 [53]. A video passes through multiple RFs with spatiotemporal filtering before being fed into identical spike encoders. The processed spike trains by the TDP are recovered back to analog waveforms which are then summed together after passing through the same RFs to obtain the recovered video. The TDP block is merely composed of switches that control the routing of spike trains among different RF channels. Identity preserving transformations such as video recovery, rotation and zooming can be achieved by this simple TDP.

The mathematical treatment of the TEM/TDP/TDM system provides a first rigorous insight into how spike timestamps play the role in real-world analog signal representation and processing. The TEM/TDM modeling is especially useful for evaluating the quality of spike encoding in spiking sensors. The change of information currency from binary digits 0/1 to spikes however increases the analysis difficulties. This whole system including the TDP could be a primitive model for future neuro-inspired silicon SoC with integrated sensing and processing modules working in time domain without any clock, but it is yet to be shown that the functional mechanisms of timestamps in advanced processing tasks like classification through learning using more complex neural networks.

1.2.3 Clockless Continuous-Time ADC/DSP/DAC Systems

A conceptually related system to the TEM/TDP/TDM that has been implemented in silicon is the clockless continuous-time (CT) ADC/DSP/DAC [54] (Figure 1.5) proposed by Yannis Tsividis. Like

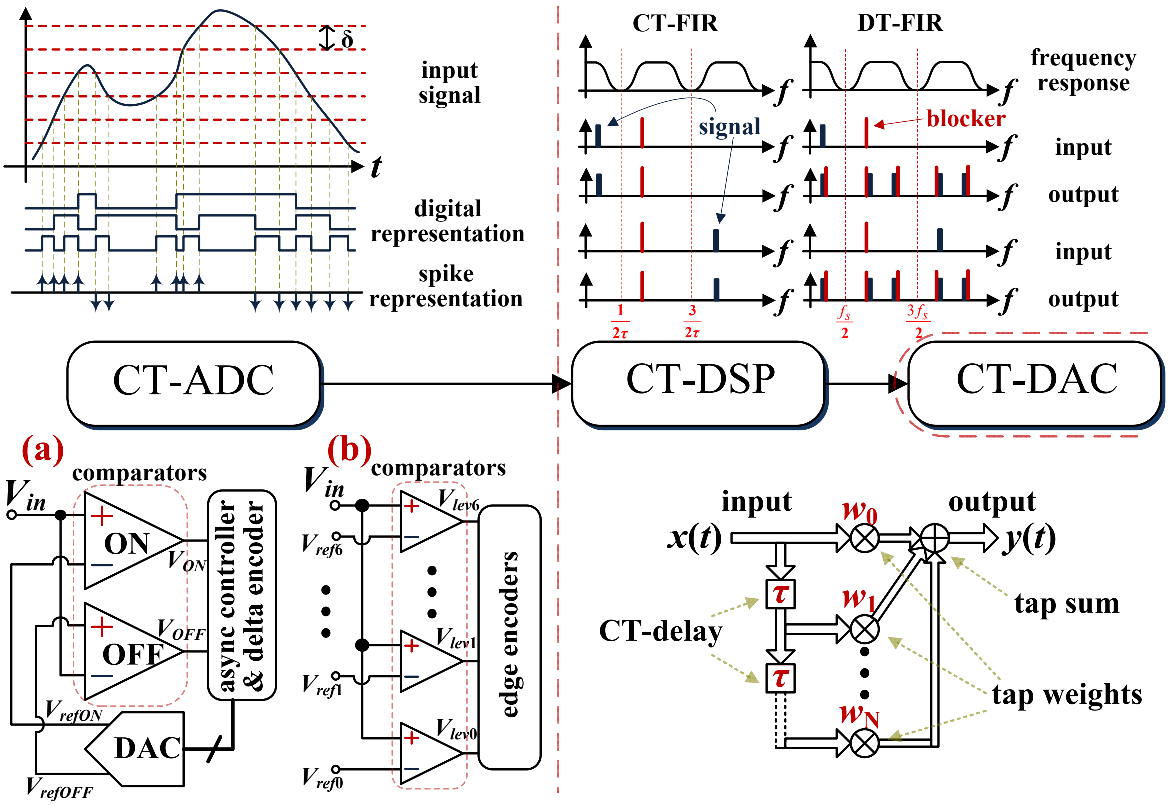


Figure 1.5. A clockless CT-ADC/DSP/DAC system, where usually the ADC uses level-crossing sampling, the DSP implements CT digital filtering, and the DAC performs zero-order-hold [54].

TEM, the CT-ADC only quantizes the input by its amplitude whereas in conventional ADC both time and amplitude are quantized. The conversion is based on the level-crossing sampling scheme [55] which is a mutation of the asynchronous delta modulation (ADM) proposed back in 1960s [56]. Every time the input passes a predefined amplitude level (normally the levels are equally spaced), an output spike (also called event) is generated and its polarity depends on the sign of the signal slope at the moment of level-crossing. Using the signal-to-quantization-noise ratio (SQNR) metric as in clocked ADCs, the performance can be quantified the same as the ADM with white Gaussian noise input [57]:

$$\mathrm{SQNR} = 18.8 + 30\log_{10}\frac{\sigma}{\delta}\ \mathrm{dB}$$

where σ is the rms amplitude and δ is the level-crossing threshold. Note that the SQNR increases 9 dB for every increased 1-bit resolution, i.e. halved δ, in contrast to 6 dB in clocked ADCs. This is also true for sinusoidal input [58]. The encoded signal can be in either digital or spike (also called delta) format, and is sent to the CT-DSP for processing, where only digital filtering functions including FIR and IIR have been demonstrated so far. Taking the digital CT-FIR as an example, it is readily derived from the analog FIR with the output $y(t)$ written as:

$$y(t) = \sum_{n=0}^{N} w_n x(t - n\tau)$$

where $x(t)$ is the input, τ is the tap delay, w_n is the weight for tap n, and (N+1) taps are used. Note that 'CT-DSP' is particularly distinguished from 'asynchronous DSP' like asynchronous microprocessors pioneered by Alain Martin et al. [59] – the former emphasize the exact timing of spikes while the later only cares to preserve the spike order. The CT-DAC usually performs zero-order-hold (ZOH) to produce the staircase-like output waveform. Recently the derivative level-crossing sampling is proposed for CT-ADC so that a much improved signal-to-distortion ratio (SDR) at the CT-DAC output can be obtained by ZOH and integration [60]. Thanks to the lack of clock and time quantization, the whole system is free of aliasing and has adaptive power consumption – if no spike is generated by the CT-ADC, the CT-DSP only dissipates leakage power.

Several circuit implementations of CT-ADC and CT-DSP have been presented in accord with different system requirements. For voice band applications, the circuit (a) at the bottom left of Figure 1.5 was used as CT-ADC [61]. The change of the crossing levels is achieved by using a feedback DAC in response to the outputs of the ON and OFF comparators. If V_{ON} gets high, V_{refON} and V_{refOFF} increase one LSB; if V_{OFF} gets high, V_{refON} and V_{refOFF} decrease one LSB. The encoded digital signal in spike/delta form is sent to the CT-DSP via asynchronous handshake. The tap delay circuit in the CT-DSP based on capacitor discharging is customized for delaying burst signals like spikes [62]. The weight multiplication and tap summation are implemented in asynchronous digital circuits. For GHz-range applications, the delay in the feedback loop of circuit (a) is too long for V_{refON} and V_{refOFF} to track the input, and instead the circuit (b) was employed which is basically a clockless flash ADC with 7 comparators for 3-bit encoding [63]. The output of each comparator is encoded into rising and falling edges which are sent to the CT-DSP composed of charge-pump-based weight-multiplier and current-mode adder.

The CT-ADC/DSP/DAC system may be regarded as the first rational physical embodiment of the computational TEM/TDP/TDM in silicon which closely resembles the prevalent clocked ADC/DSP/DAC system. It involves development of novel clockless circuits and systems for encoding and processing signals in time domain. Recent advancements include adaptive-resolution CT-ADCs [64], [65], flexible CT-DSPs that allow different digital input formats with variable data rate [66], etc. Spiking sensors like silicon cochlea and retina can in fact be regarded as 1-D and 2-D arrays respectively of CT-ADCs with analog preprocessing before spike encoding. Compact implementation is usually required considering the large number of array elements.

1.3 Thesis Contribution & Organization

This thesis analytically and numerically studies the quality of sensory information representation in spike domain with non-idealities of silicon implementation considered. ADM is found in general to deliver better encoding quality over previously adopted self-timed reset (STR) or integrate-and-fire, and is chosen for spike encoder implementations in spiking silicon retina and cochlea. In additional to compact ADM designs, the fabricated silicon retina focuses on improvement of temporal contrast sensitivity for potential fine-texture recognition and optical neuroimaging applications, and the silicon cochlea emphasizes low supply voltage and low power consumption for wireless sensor networks in the context of internet of things or ambient intelligence where harvested energy is scarcely available. The rest of the thesis is organized as follows.

Chapter 2 first compares the encoding quality between the ADM and STR encoders using the SDR metric with linear decoding of the output spike train. With comparison delay T_{DC} and queueing delay T_{DQ} considered, the ADM encoder shows an increasingly higher SDR improvement over the STR as the input signal frequency and the encoder's quantization bit number increase as long as no slope overload occurs in ADM. Second, the effect of jitter of spike timestamps caused by T_{DC} and T_{DQ} variations is quantitatively measured by the reconstruction SDR using nonlinear decoding algorithms based on frame theory. Because T_{DC} and T_{DQ} variations are related to several circuit and system parameters, this analysis can provide guidance for specifications-oriented design of spiking sensors.

Chapter 3 presents the spiking silicon retina with enhanced temporal contrast sensitivity and spike encoding quality with in-pixel ADM. One pixel is composed of a low-noise common-gate photoreceptor that logarithmically converts light intensity change ΔI to output voltage ΔV_{pr}, a capacitively-coupled programmable gain amplifier that amplifies ΔV_{pr} with four levels of adjustable gains, a switched-capacitor ADM that encodes the amplified ΔV_{pr} into spike trains, and an in-pixel asynchronous logic block that generates the switching signals for the ADM and communicates spikes to the peripheral 2-D burst-mode word-serial AER. The chip named as ADMDVS was fabricated in UMC 0.18 µm 1P6M RF/MM CMOS. The measurement results include noise, TC sensitivity, and the comparison of ADM spike encoding with a prior DVS using STR encoding. A simulated optical neuroimaging experiment is also performed to demonstrate a potential practical application.

Chapter 4 elaborates on the design of the 0.5-V 55-µW 64×2-channel binaural silicon cochlea. A power-efficient source-follower-based 1-zero 4-pole asymmetrical BPF with tunable Q is proposed which comprises a 4th-order source-follower-based LPF and a summing PGA. A self-oscillating comparison scheme is proposed to replace the CT-comparators in ADM with more energy-efficient dynamic latched comparators. The oscillation frequency is designed to be adaptive in response to the output spike activity to further save power. The biases of all the channels are geometrically scaled, covering a frequency range from 8 to 20k Hz. The chip named as CochLP was fabricated in TowerJazz 0.18 µm 1P6M IS CMOS. The measurement results mainly include the transfer functions, noise and distortion of the BPFs, and the spike output using chirp and natural sound input to the chip. The 0.5-V CochLP core is about ×18 times power-efficient compared to the best prior art, exhibits good matching within a monaural ear and between both ears, and has a wide Q tuning range.

Chapter 5 describes a wide dynamic range current reference array that provides all the biases for spiking sensors. To cover the current range from about 30 fA to 26 µA, a hierarchical coarse-fine selection architecture is employed. A subthreshold PTAT master bias sends a 400 nA current to a coarse current divider and multiplier to generate eight coarse currents with a scaling ratio of 8. Each bias branch selects one of the eight coarse currents and uses an 8-bit R-2R DAC based on MOSFET-only current-splitting technique to generate a fine-tuned current which is sent to a configurable buffer. A bias voltage is generated by the buffer via a diode-connected transistor. The measurement results are obtained from chips fabricated in UMC 0.18 µm CMOS.

Chapter 6 gives a brief summary of the thesis and some outlooks of possible future directions.

Chapter 2: Analysis of Spike Encoding Mechanisms in Spiking Sensors

This chapter presents the analysis of the spike encoding mechanisms that can be feasibly implemented in array sensors like silicon retina and cochlea, particularly the self-timed reset (STR) and asynchronous delta modulation (ADM). STR and ADM both belong to the direct-threshold-crossing encoding category without integration in the feedforward path in contrast to other encoding schemes like asynchronous sigma-delta modulation (ASDM) [52] and integrate-and-fire (IAF) [40] where an integrator is used before threshold-crossing. STR and ADM mainly differ in the feedback path as will be discussed later. STR has been used in all previous dynamic vision sensors (DVSs) [67] because of its simple circuit implementation. ADM has so far only been realized in the form of level-crossing sampling in clockless continuous-time ADCs [54]. Different encoding schemes have been shown to have largely different encoding quality [53]. With a nonlinear decoding algorithm derived based on frame theory, regardless of the encoding mechanisms, perfect reconstruction is possible provided the average spike rate is at least as large as the Nyquist rate and the time support is infinite [52]. In practice however, finite time support results in degradation of signal representation integrity, and with the same time length and similar spike number some encoding is superior compared to others measured by the signal-to-distortion ratio (SDR) metric that is used in the rest of this chapter and is defined as below:

$$\mathrm{SDR} = \frac{P_{signal}}{P_{distortion}} = \frac{\int_{t_1}^{t_2} (x(t))^2\, dt}{\int_{t_1}^{t_2} [x(t) - x_{rc}(t)]^2\, dt} \tag{2.1}$$

where P_{signal} is the power of the original input signal $x(t)$ and $P_{distortion}$ is the power of the reconstruction error which is the difference between $x(t)$ and the recovered signal $x_{rc}(t)$.

The calculated SDR depends not only on the encoding methods but also on the decoding algorithms. As pointed out in [52], linear decoding that uses simple integration and/or filtering [68] gives much lower SDR compared to nonlinear decoding, because the former does not consider the nonlinear nature of the encoding process where the threshold-crossing takes place. To obtain the same SDR, linear decoding requires much higher quantization resolution, i.e. lower threshold, and in turn leads to high average spike rate and high power consumption in silicon implementations. As shown in the simulation of level-crossing sampling [69], with 4-bit amplitude quantization, nonlinear decoding gives a larger than 100 dB SNR in frequency domain for a sinusoidal input while linear decoding could only achieve less than 40 dB [70]. This infers that the information of the input waveform is entirely contained in the encoded timestamps, and therefore to evaluate the encoding integrity, nonlinear decoding should be used. However, nonlinear decoding is much more computational intensive, and hence if hardware efficiency is paramount, linear decoding should be employed. But for spike processing, it is unclear yet how to evaluate the tradeoff between encoding resolution and algorithm complexity with some targeted performance like classification accuracy of a spike-based classifier. This is something to be explored in the future.

According to the data-processing inequality in information theory [71], any information loss caused by the encoder cannot be retrieved by further processing, and therefore in a rational design, the SDR lower bound of a spike train that is encoded by a spike encoder (asynchronous ADCs) or a population of spike encoders (spiking sensors) should be determined by the performance requirement of a spike processing algorithm (e.g. classification accuracy of a classifier). With specified encoding quality, the encoder type

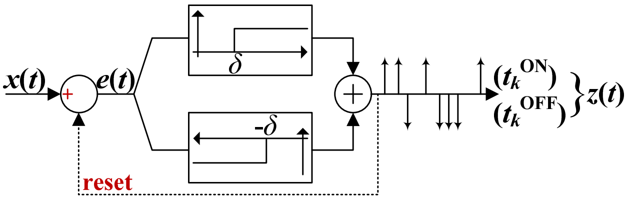


Figure 2.1. Ideal spike encoding model of STR named as silicon ON-OFF TEM in [53].

and resolution can be accordingly chosen. This chapter mainly compares two types of encoders, i.e. the STR and ADM.

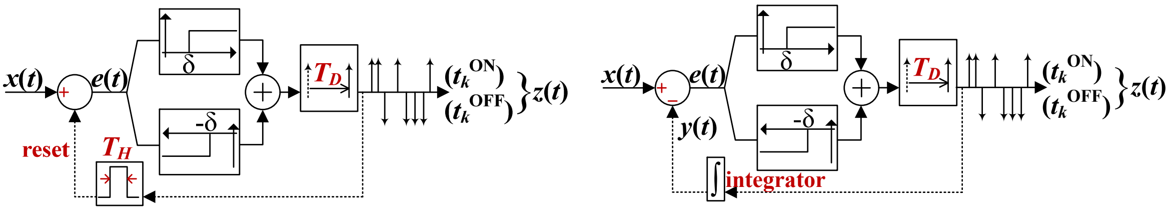
The content of this chapter is organized as follows: Section 2.1 compares the STR and ADM with fixed feedback delay, including: Section 2.1.1 describes the abstract models of STR and ADM considering fixed feedback delay; Section 2.1.2 shows the SDR metric comparison of STR and ADM using linear decoding; Section 2.2 discuss the SDR degradation problem due to imprecise timestamps caused by communication delay variation, including: Section 2.2.1 provides the nonlinear spike decoding algorithms of STR and ADM; Section 2.2.2 quantifies two sources of delay variation, namely the comparison delay variation and the queueing delay variation; Section 2.2.3 shows the encoding SDR degradation due to the two types of delay variation; Section 2.3 gives conclusion and remarks.

2.1 STR and ADM with Fixed Feedback Delay T_D

2.1.1 Modeling of STR and ADM

The STR was ideally modeled as the silicon ON-OFF time encoding machine (TEM) [53] as illustrated in Figure 2.1. Starting from the reset level, say 0, the error signal $\boldsymbol{e}(t)$ tracks the input signal $\boldsymbol{x}(t)$. When $\boldsymbol{e}(t)$ crosses the threshold δ/-δ at time $t_k^{\mathrm{ON}}/t_k^{\mathrm{OFF}}$ (k∊Z), an ON/OFF spike is generated. The generated spike immediately reset $\boldsymbol{e}(t)$ back to 0 with no delay and $\boldsymbol{e}(t)$ instantaneously starts to track $\boldsymbol{x}(t)$ again. This process repeats to produce spikes over time which forms the spike train $\boldsymbol{z}(t)$. The assumptions of no feedback delay T_D and zero reset time T_H are invalid in practical silicon implementation. Two sources contribute to T_D. One is the comparison delay. The detection of threshold-crossing is normally done by a comparator that has a finite bandwidth set by its bias current and capacitive load. At nA-bias, the delay of a 2T current-mode comparator is in the range of tens of µs. The other source is the spike transmission delay. The generated spikes from each retina pixel or cochlea channel are transmitted via the address-event-representation (AER) communication protocol. Besides the intrinsic delay of the customized on-chip asynchronous circuits and the time needed for an off-chip CPLD or FPGA to register the addresses and timestamps of the spikes, queueing caused by arbitration in a congested AER channel also contributes to transmission delay. The time for a full four-phase handshake in DVS128 is about 1.2 µs [72], and it can increase drastically without bounds when the average spike rate of a sensor array approaches the throughput capacity of an AER channel. The reset in STR is usually done by a MOS switch [6]. For $\boldsymbol{e}(t)$ to have a certain settling accuracy within e.g. a few percent of δ, T_H has to be larger than 0 and is determined by the slew rate and bandwidth of the STR amplifier. With a 10 fF capacitance and a 0.5 V threshold voltage, a 1 nA slewing current give a 5 µs T_H. In addition, the refractory period deliberately introduced to limit the spike rate of a pixel may further increase T_H.

For the reasons mentioned above, the silicon ON-OFF TEM model needs to be modified to account for the practical non-idealities. We first assume that T_D and T_H have fixed values. Figure 2.2(a) shows the



(a) (b)

Figure 2.2. Spike encoding model of (a) STR and (b) ADM with delay time T_D and switch-hold period T_H.

STR model with T_D and T_H considered. The two sources of T_D are lumped together and the timestamp of each spike is the moment of threshold-crossing plus T_D. The reset occurs after the acknowledge to each spike transmission comes back, and the T_H represents the reset time and refractory period during which $\boldsymbol{e}(t)$ is held at 0. $\boldsymbol{e}(t)$ can be expressed as:

$$\boldsymbol{e}(t)=\begin{cases}0 & (t_k^i < t \le t_k^i + T_H)\\ \boldsymbol{x}(t) & (t_k^i + T_H < t < t_{k+1}^j)\end{cases} \quad (i,j=\text{ON/OFF})$$

The ADM model is similar to STR, as shown in Figure 2.2(b). The reset is replaced by an ideal integrator with the impulse response given below:

$$\boldsymbol{y}_{k+1}(t)=\begin{cases}\boldsymbol{y}_k(t)+\delta & (t_k^i < t \le t_{k+1}^j,\ i,j=\text{ON})\\ \boldsymbol{y}_k(t)-\delta & (t_k^i < t \le t_{k+1}^j,\ i,j=\text{OFF})\\ 0 & (t \le 0)\end{cases}$$

And $\boldsymbol{e}(t)$ is now written as:

$$\boldsymbol{e}(t)=\boldsymbol{x}(t)-\boldsymbol{y}(t)$$

The ADM model used here is adapted from the model given in [57], where the incremental value in $\boldsymbol{y}(t)$ at each t_k is 2δ/-2δ instead of δ/-δ. Using a less than 2δ incremental step can help prevent false crossing of the complementary threshold (δ and -δ are complementary thresholds for each other) due to noise or digital coupling. Here δ is chosen to be consistent with the ideal STR case.

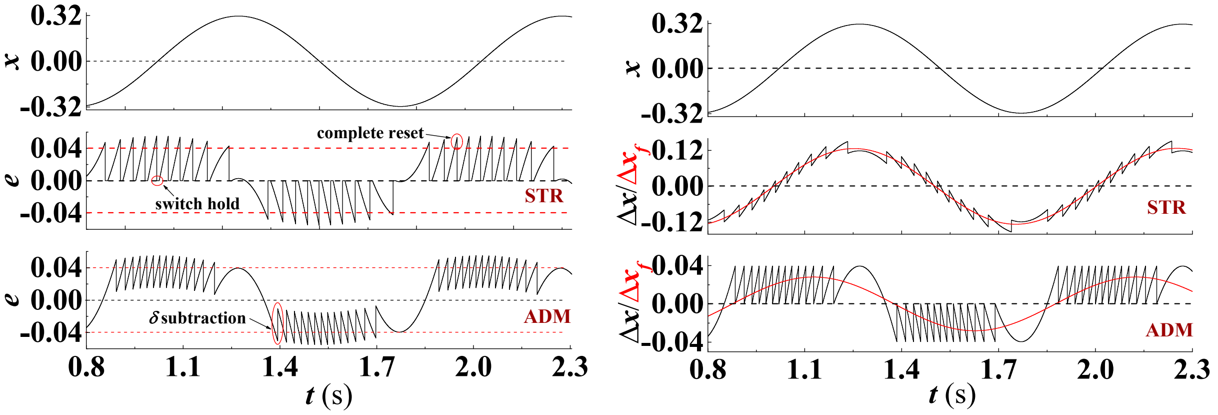
2.1.2 Linear Decoding Comparison

Figure 2.3(a) shows the simulated waveform $\boldsymbol{e}(t)$ of both STR and ADM using MATLAB within a time window of [0.78 s, 2.28 s]. The input $\boldsymbol{x}(t)$ is a 1 Hz sinusoidal signal:

$$\boldsymbol{x}(t)=0.316\times\sin(2\pi\times1\times t)$$

T_D and T_H are both about 7.8 ms, and δ is set to $0.316/2^3$=0.0395 for a 4-bit quantization (4 is called the quantization bit number (QBN)). The signal loss due to complete reset and switch hold is labeled in red circle in STR, and so is the δ subtraction in ADM. The parts of the signal that are beyond δ and during T_H are discarded in STR because in linear reconstruction, every spike is assumed to contribute a δ incremental which is not true. In ADM, the amount in each feedback subtraction is exactly δ, and therefore ideally no signal loss occurs.

To compare the encoding quality using the SDR metric, a linear decoder composed of an ideal integrator and a brick-wall LPF is used as depicted in Figure 2.4. The lowpass corner frequency of the LPF is set



(a) (b)

Figure 2.3. (a) Waveforms of $\boldsymbol{x}(t)$ and $\boldsymbol{e}(t)$ of the STR and ADM, and (b) the linear reconstruction errors $\Delta\boldsymbol{x}(t)$ and $\Delta\boldsymbol{x}_f(t)$ within a time window of [0.78 s, 2.28 s].

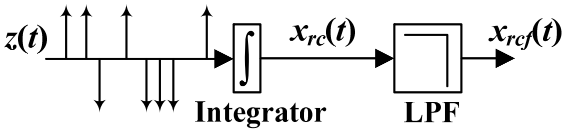


Figure 2.4. Linear spike decoder comprised of an ideal integrator and a brick-wall LPF.

to 2 Hz. Let $\boldsymbol{x}_{rc}(t)$ and $\boldsymbol{x}_{rcf}(t)$ represent the reconstructed signal before and after LPF. The reconstruction errors are written as:

$$\Delta x(t) = x(t) - x_{rc}(t + T_D)\ ,\ \Delta x_f(t) = x(t) - x_{rcf}(t + T_D)$$

Note that $\boldsymbol{x}_{rc}(t)$ and $\boldsymbol{x}_{rcf}(t)$ are shifted by T_D to account for the spike delay. Figure 2.3(b) shows both $\Delta\boldsymbol{x}(t)$ and $\Delta\boldsymbol{x}_f(t)$ of STR and ADM within a time window [0.78 s, 2.28 s]. The maximum $\Delta\boldsymbol{x}_f(t)$ of STR is about 3LSB (~0.12), and of ADM is less than 1LSB (~0.04). The calculated SDR using $\Delta\boldsymbol{x}_f(t)$ is 7.95 dB for STR and 21.0 dB for ADM, respectively. The SDR improvement ΔSDR is 13.1 dB. It may seem that the SDR improvement comes from the more frequent sampling in ADM, 28 versus 20 spikes of STR during one $\boldsymbol{x}(t)$ period. However, if the threshold of ADM δ is increased so that ADM also takes only 20 samples per period, its SDR only slightly decreases to 19.0 dB, and if the STR threshold δ is decreased so that 28 samples are taken per period, its SDR is further decreased to 4.91 dB. This is because the higher the sampling frequency of STR, the larger portion of signal is discarded due to complete reset and switch hold.

The SDR improvement ΔSDR is a function of both QBN and input frequency f_{in} for a given signal amplitude. Figure 2.5 shows the dependence of ΔSDR on QBN and f_{in}. QBN is swept from 2 to 11 in simulation, and f_{in} from 1 to 20 Hz. The 20 Hz bandwidth is reported to contain most information in natural human vision [50]. Smaller and more practical T_D and T_H are used, 61 μs and 7.6 μs respectively. When QBN is less than 4, ΔSDR is negligible within the frequency range, but it becomes increasingly large as QBN goes up above 5, and reaches the peak value of about 57 dB at f_{in}=2.5 Hz and QBN=11. ΔSDR also generally increases with f_{in}. The abrupt ΔSDR drop at large f_{in} and QBN is caused by slope overload in ADM where the feedback $\boldsymbol{y}(t)$ is not able to tract the input $\boldsymbol{x}(t)$ due to the feedback delay. In this region, $\boldsymbol{e}(t)$ can even exceed ±2δ. The critical parameter that determines the occurrence of slope overload is T_D. In the case of a sinusoidal input, the bounds on f_{in} and QBN can be linked to T_D as:

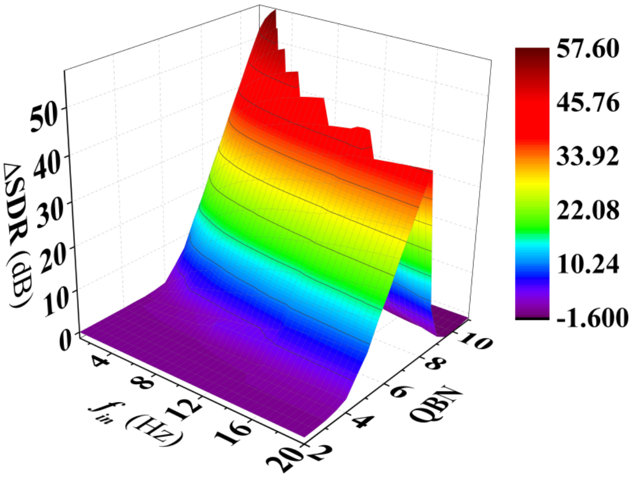


Figure 2.5. ΔSDR of ADM over STR as a function of f_{in} and QBN.

$$T_D \leq \frac{1}{\pi f_{in} \cdot 2^{QBN+1}} \tag{2.2}$$

If an ADM is designed for a large f_{in}, e.g. 1k Hz, to have an 8 QBN, T_D must be significantly reduced to about 0.62 μs to avoid slope overload. The input amplitude of a sinusoidal signal $V_{\sin}$ has been assumed to be fixed so far, but no additional analysis is needed for a varying $V_{\sin}$ because with a given threshold δ, it is directly linked to QBN by $V_{\sin}/\delta=2^{\mathrm{QBN}}$.

Other spike encoding mechanisms like IAF with reset in the feedback also have the same problems of STR in practical implementations, i.e. complete reset and switch hold, which may also lead to severe signal loss using linear reconstruction.

2.2 SDR Degradation due to T_D Variation

With the assumption of fixed T_D and T_H, the numerical analysis in Section 2.1 reveals that ADM generally favors a higher SDR than STR with linear decoding, especially at large f_{in} and QBN, which is attributed to the uncertainty of signal representation in STR due to the complete reset and switch hold. It has been shown in [40] that the uncertainty caused by absolute refractory period in IAF neurons can be accounted for and a perfect reconstruction algorithm was thereof obtained based on the methods from irregular sampling. For both STR and ADM, similar algorithms can be developed and give much higher SDR. This section develops the nonlinear decoding algorithms to study the encoding quality of ADM and STR considering the variation of T_D which is caused by several factors in spike transmission of spiking sensors.

2.2.1 Nonlinear Decoding

To develop the nonlinear decoding algorithms, the first step is to map the input signal to timestamps following the so-called t-transform [52]. Figure 2.6 illustrates the principle for STR and ADM. The waveforms are the modulated $\boldsymbol{x}(t)$, i.e. the $\boldsymbol{e}(t)$ in Figure 2.2. For simplicity, $\boldsymbol{x}(0)$ is assumed to be 0; otherwise an offset $\boldsymbol{x}_0=\boldsymbol{x}(0)$ can be added to $\boldsymbol{x}(t)$. The numbers 1 and 2 in the superscript of the timestamps represent ON and OFF spikes respectively. For STR, the timestamps have the following relationships:

$$x(t_2^1 - T_{D2}) - x(t_1^1 + T_H) = (-1)^{1-1}\delta\,,\ x(t_3^2 - T_{D3}) - x(t_2^1 + T_H) = (-1)^{2-1}\delta$$

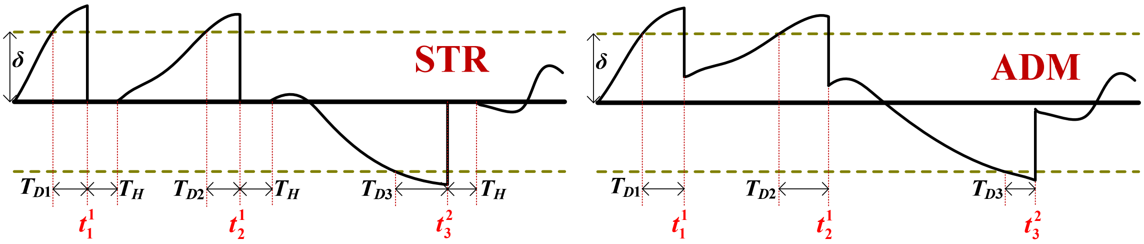


Figure 2.6. Illustration of the *t*-transformation principle for STR and ADM.

And for ADM, the equations below hold:

$$x(t_1^1 - T_{D1}) = \delta\cdot(1-2\cdot 0)\,,\ x(t_2^1 - T_{D2}) = \delta\cdot(2-2\cdot 0)\,,\ x(t_3^2 - T_{D3}) = -\delta\cdot(3-2\cdot 2)$$

T_{D1}, T_{D2} and T_{D3} represent the different delays associated with each spike. From the specific examples given above, more general formulization can be derived. For STR, assume T_H is fixed, and given $\forall k\epsilon N^+$:

$$x\left(t_k^i - T_{Dk}\right) - x\left(t_{k-1}^j + T_H\right) = (-1)^{i-1}\delta \quad (k \geq 2)\,,\ x\left(t_k^i - T_{Dk}\right) = (-1)^{i-1}\delta \quad (k=1) \tag{2.3}$$

where i,j=1,2, i and j are independent of each other, and T_{Dk} is the delay associated with the k^{th} spike. For ADM, given $\forall k$, $l\epsilon N^+$:

$$x(t_k^1 - T_{Dk}) = \delta\cdot(k - 2\sum_{l\in N^+} 1_{[t_l^2 < t_k^1]})\,,\ x(t_k^2 - T_{Dk}) = -\delta\cdot(k - 2\sum_{l\in N^+} 1_{[t_l^1 < t_k^2]}) \tag{2.4}$$

The inner product form of Eq. (2.3) can be written as:

$$< x, \chi_{D,k}^i > - < x, \chi_{H,k-1}^j >= q_{kS}^i$$

where q_{kS}^i is the right side of Eq. (2.3), and χ is the sampling function and has the form of:

$$\chi_{D,k}^i = g(t - t_k^i + T_{Dk})\,,\ \chi_{H,k-1}^j = g(t - t_{k-1}^j - T_H) \quad (k \geq 2)\,,\ \chi_{H,0}^j = 0$$

In this chapter, $g(t)$ represents the sinc function $g(t)=\sin(\Omega t)/\pi t$. It is the impulse response of a brick-wall LPF with a cutoff frequency of Ω.

The inner product form of Eq. (2.4) can be written as:

$$< x, \chi_k^i >= q_{kA}^i$$

where q_{kA}^i is the right side of Eq. (2.4), i=1,2, and the sampling function χ has the form of:

$$\chi_k^i = g(t - t_k^i + T_{Dk})$$

In light of the *Proposition* 1 and its proof in [53], in STR the recovered signal $\boldsymbol{x}_{rc}(t)$ from spike trains can be obtained by weighted summation of the representation functions $\eta_k(t) = g(t - t_k^i)$, which form a frame for the space of bandlimited functions:

$$x_{rc}(t) = \sum_{k\in N^+} c_k \eta_k(t) \tag{2.5}$$

The weights c_k, i.e. the coefficient vector $\boldsymbol{c}$ can be computed as:

$$\mathbf{c}=\mathbf{G}^+\mathbf{q}$$

where $[\mathbf{q}]_k = q_{kS}^i$, i=1,2. $\mathbf{G}^+$ is the pseudo-inverse of the matrix $\mathbf{G}$. The elements in $\mathbf{G}$ are:

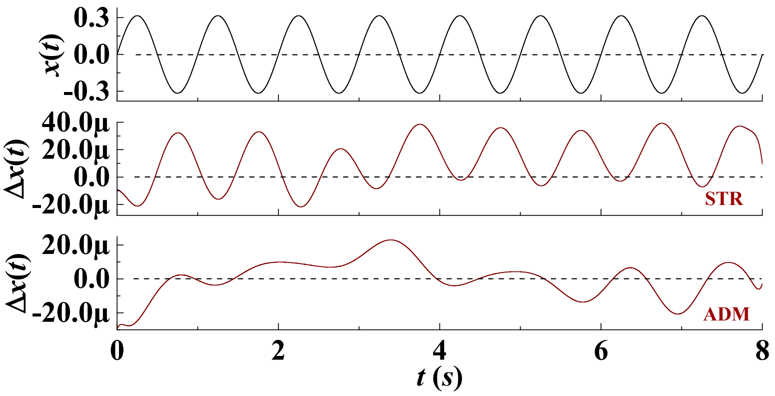


Figure 2.7. Reconstruction error of STR and ADM using nonlinear decoding.

$$[\mathbf{G}]_{kl} = <\chi^{i}_{D,k}, \eta_l> - <\chi^{j}_{H,k-1}, \eta_l>$$

for all i,j=1,2, and $k,l\epsilon N^{+}$, $k\geq 2$. When k=1,

$$[\mathbf{G}]_{kl} = <\chi^{i}_{D,k}, \eta_l>$$

Using Parseval's formula [73], $[\mathbf{G}]_{kl}$ can be computed as:

$$\begin{aligned}[\mathbf{G}]_{kl} &= \int_{-\infty}^{\infty} g(t-t_k^i+T_{Dk})g(t-t_l)dt - \int_{-\infty}^{\infty} g(t-t_k^i-T_H)g(t-t_l)d \\ &= \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{-i\omega(t_k^i-T_D-t_l)}d\omega + \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{-i\omega(t_k^j+T_H-t_l)}d\omega \\ &= \mathrm{g}(t_k^i-T_{Dk}-t_l)\text{-}\mathrm{g}(t_{k-1}^j+T_H-t_l) \quad (k\geq 2)\end{aligned}$$

$$[\mathbf{G}]_{kl} = \mathrm{g}(t_k^i-T_{Dk}-t_l) \quad (k=1)$$

Similarly, in ADM the recovered signal $\boldsymbol{x}_{rc}(t)$ from spike trains can be computed as:

$$x_{rc}(t) = \sum_{k\in N^{+}} c_k^1\eta_k^1(t) + \sum_{k\in N^{+}} c_k^2\eta_k^2(t) \tag{2.6}$$

where $\eta_k^i(t) = g(t-t_k^i)$, i=1,2. With $\mathbf{c}=[\mathbf{c}^1; \mathbf{c}^2]$ and $[\mathbf{c}^{\mathbf{i}}]_k = c_k^i$, the vector of coefficients $\mathbf{c}$ can be computed as:

$$\mathbf{c}=\mathbf{G}^{+}\mathbf{q}$$

where q=$[\mathbf{q}^{\mathbf{1}}; \mathbf{q}^{\mathbf{2}}]$ with $[\mathbf{q}^{\mathbf{i}}]_k = q_{kA}^i$, i=1,2, and the matrix $\mathbf{G}$ can be written as:

$$\mathbf{G}=\begin{bmatrix}\mathbf{G}^{11} & \mathbf{G}^{12} \\ \mathbf{G}^{21} & \mathbf{G}^{22}\end{bmatrix},\ [\mathbf{G}^{ij}]_{kl} = <\chi_k^i, \eta_l^j>$$

where k, $l\epsilon N^{+}$. $[\mathbf{G}^{\mathbf{ij}}]_{kl}$ can be computed as:

$$\begin{aligned}[\mathbf{G}^{\mathbf{ij}}]_{kl} &= \int_{-\infty}^{\infty} g(t-t_k^i+T_{Dk})g(t-t_l^j)dt \\ &= \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{-i\omega(t_k^i-T_D-t_l^j)}d\omega = \mathrm{g}(t_k^i-T_{Dk}-t_l^j)\end{aligned}$$

With the same input signal $\boldsymbol{x}(t)$ and the encoding threshold δ that are used to obtain the plots in Figure 2.3, the reconstruction errors $\Delta\boldsymbol{x}(t)=\boldsymbol{x}(t)\text{-}\boldsymbol{x}_{rc}(t)$ of STR and ADM using the nonlinear decoding algorithms described above are shown in Figure 2.7. The time support is 8 s, Ω=2π rad/s, and time step is about 3.8 μs.

Obviously the reconstruction errors $\Delta\boldsymbol{x}(t)$ are reduced by orders of magnitude in both cases compared to linear decoding. The ADM has an SDR of 87 dB and even the STR has 80 dB, only 7 dB lower. The residual $\Delta\boldsymbol{x}(t)$ is caused by two sources, the finite time support and finite simulation accuracy in terms of non-infinitesimal time step. To study a more general case, a white Gaussian noise will be used as the input $\boldsymbol{x}(t)$ instead of a sinusoid.

2.2.2 Two Sources of T_D Variation

2.2.2.1 Comparison Delay T_{DC} Variation

Comparators are used for the detection of threshold-crossing, and comparison delay T_{DC} variation is the first source of T_D variation because T_{DC} is dependent on the slope s_x of the input signal $\boldsymbol{x}(t)$ at the moment of threshold-crossing. For continuous-time comparators commonly used in clockless systems, the relevant parameters are the bias current and the DC gain with a given circuit topology. The following analysis will take the simplest 2T common-source amplifier (CSA) comparator as an example which has been used in many DVSs [6]. Assuming one dominant pole and no slewing [74], the transfer function of the CSA can be written as:

$$H(s) = \frac{A_{DC}}{s\tau_c + 1}$$

where A_{DC} is the DC gain of the CSA, and τ_c is the time constant that is associated with the dominant pole and determined by the CSA bias and output load capacitance. The input signal change at the moment of threshold-crossing is approximated by $s_x\Delta t$, where Δt is time needed for the CSA output to have a change of δ, and the Laplace transform of $s_x\Delta t$ is s_x/s^2. The CSA output can be then written as:

$$V_{out}(s) = \frac{s_x}{s^2}\cdot H(s) = \frac{A_{DC}s_x}{s^2(s\tau_c + 1)}$$

The inverse-Laplace transform of $V_{out}(s)$ is:

$$\begin{aligned} V_{out}(t) &= \mathcal{L}^{-1}\{V_{out}(s)\} = \mathcal{L}^{-1}\{\frac{A_{DC}s_x}{s^2(s\tau_c + 1)}\} \\ &= A_{DC}s_x\mathcal{L}^{-1}\{\frac{1}{s^2} - \frac{\tau_c}{s} + \frac{\tau_c^2}{s\tau_c + 1}\} = s_x A_{DC}(t - \tau_c + \tau_c e^{-t/\tau_c}) \end{aligned}$$

Using T_{DC} to replace t and making Taylor expansion on the exponential term, $V_{out}(t)$ is rewritten as:

$$\begin{aligned} V_{out}(t) &= s_x A_{DC}[T_{DC} - \tau_c + \tau_c(1 - \frac{T_{DC}}{\tau_c} + \frac{T_{DC}^2}{2\tau_c^2} - O(T_{DC}^3)] \\ &\approx s_x A_{DC}\frac{T_{DC}^2}{2\tau_c} \quad (T_{DC} >> 3\tau_c) \end{aligned}$$

T_{DC} can be written as the function of s_x in logarithmic:

$$\log_{10}T_{DC} = \frac{1}{2}\log_{10}\frac{2|V_{out}|\tau_c}{A_{DC}} - \frac{1}{2}\log_{10}|s_x| = y_0 - \frac{1}{2}\log_{10}|s_x| \tag{2.7}$$

where y_0 is the lumped indicator of comparison speed as the function of A_{DC} and τ_c with a given V_{out} range, usually from 10% to 90% rail-to-rail. The simulated T_{DC} versus s_x of a CAS comparator biased at 0.2 nA,

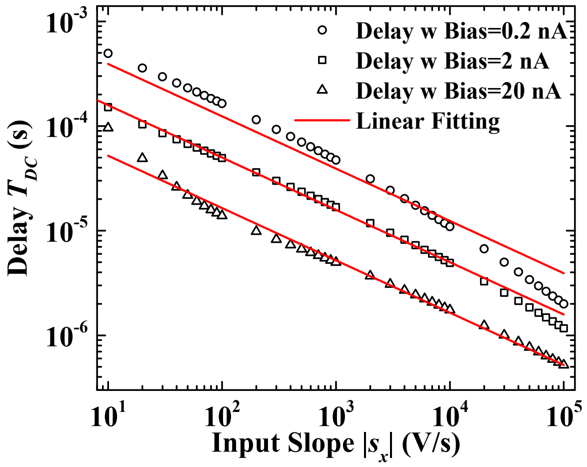


Figure 2.8. Simulated comparison delay T_{DC} versus input signal slope s_x at different biases of a CSA comparator by Spectre in 0.18 μm CMOS, and the linear fittings using Eq. (2.7).

2 nA and 20 nA using Spectre in 0.18 μm CMOS is shown in Figure 2.8 as the scattered dots. The red lines are the linear fits using Eq. (2.7) with y_0=-2.91, -3.30, and -3.78, respectively. It can be seen that Eq. (2.7) is a relatively good approximation of practical T_{DC}, and therefore in spike encoding simulation, it is used to generate the delays T_{Dk} in Eq. (2.3) and (2.4) that correspond to the signal slope s_x at the moment of each threshold-crossing. However in signal recovery, T_{Dk} is not available due to the lack of the knowledge of s_x. One simple possible way of approximate $\boldsymbol{x}(t)$ recovery is to assume a fixed average delay at all the threshold-crossing moments according to some general characteristics of $\boldsymbol{x}(t)$, such as its bandwidth and power. In this case, all T_{Dk}'s are replaced by T_{Davg} and are independent of k. T_{Davg} will be derived as follows with characteristic parameters of a given input signal.

For a more general analysis, a random noise $\boldsymbol{x}(t)$ with Gaussian amplitude distribution is used instead of a sinusoid as in Section 2.1. To obtain the T_{Davg}, the average slope at threshold-crossing moments is approximated by the mean slope of $\boldsymbol{x}(t)$. The joint probability density function of $\boldsymbol{x}(t)$'s amplitude a_x and slope s_x has been derived by S. Rice as [75]:

$$p(a_x,s_x)=\frac{(-\psi_0\psi_0'')^{-1/2}}{2\pi}\exp(-\frac{a_x^2}{2\psi_0}+\frac{s_x^2}{2\psi_0''}) \tag{2.8}$$

where ψ_0 and ψ_0'' are the correlation function $\psi(\tau)$ of $\boldsymbol{x}(t)$ and the second derivative of $\psi(\tau)$ at τ=0, respectively. The probability density function of s_x can be computed as [57]:

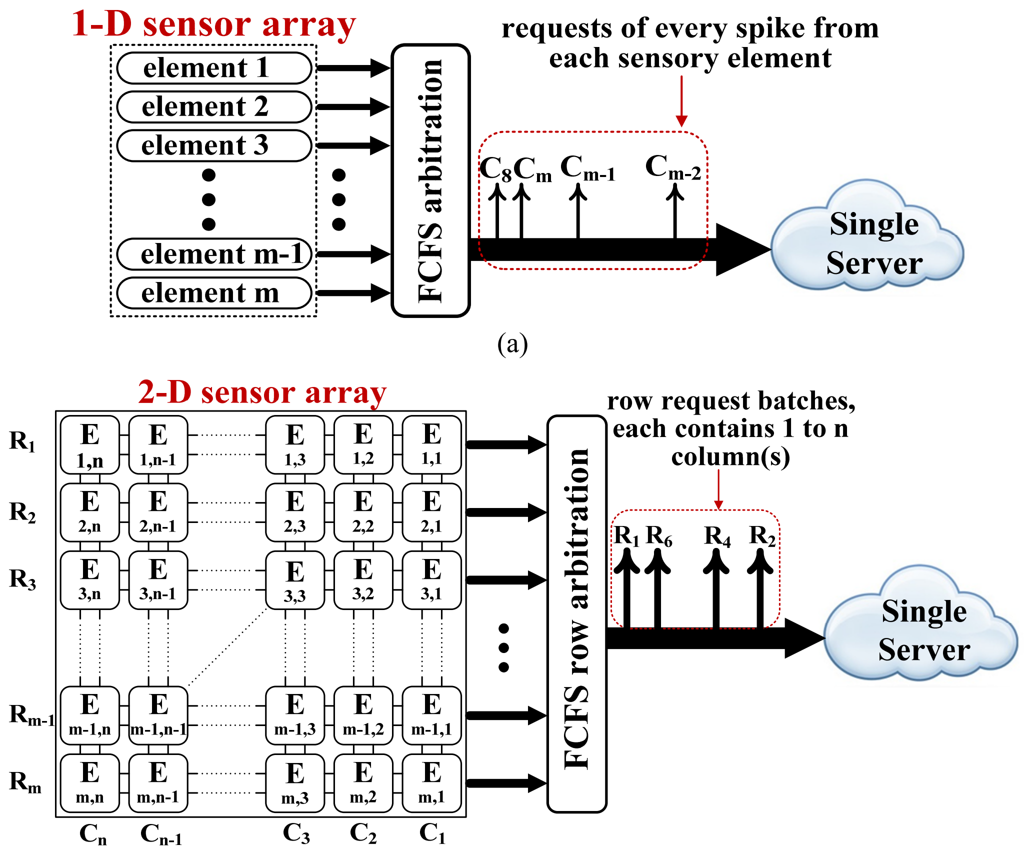
$$p(s_x)=\int_{-\infty}^{\infty}p(a_x,s_x)da_x=\frac{(-\psi_0'')^{-1/2}}{\sqrt{2\pi}}\exp(\frac{s_x^2}{2\psi_0''})$$

The average absolute slope s_{xavg} of $\boldsymbol{x}(t)$ can be calculated as:

$$s_{xavg}=2\int_0^{\infty}|s_x|\,p(|s_x|)d|s_x|=\sqrt{\frac{2}{\pi}}(-\psi_0'')^{1/2}$$

$\psi''(\tau)$ is related to $\boldsymbol{x}(t)$'s power spectrum $\omega(f)$ by [76]:

$$\psi''(\tau)=-(2\pi f)^2\int_0^{\infty}\omega(f)\cos(2\pi f\tau)df$$



(a)

(b)

Figure 2.9. The queueing models of (a) 1-D sensors like a 1-D spiking silicon cochlea and (b) 2-D sensors like a 2-D spiking silicon retina.

For bandlimited white noise with zero mean, σ^2 power and f_n bandwidth, ψ_0'' is:

$$\psi_0'' = -\frac{4\pi^2\sigma^2}{f_n}\int_0^{f_n} f^2 df = -\frac{4}{3}(\pi\sigma f_n)^2$$

s_{xavg} can hence be written as:

$$s_{xavg} = 2\sqrt{\frac{2\pi}{3}}\sigma f_n \tag{2.9}$$

Using Eq. (2.7), T_{Davg} can be written as:

$$T_{Davg} = 10^{y_0} s_{xavg}^{-1/2} \tag{2.10}$$

From Eq. (2.9) and (2.10), one can see that the required parameters for computing T_{Davg} are $\boldsymbol{x}(t)$'s bandwidth f_n, power σ^2, encoder's threshold δ, and the comparator speed indicator y_0, which is related to comparator's DC gain A_{DC} and time constant τ_c that is determined by its bias current and output load capacitance. The estimation of s_{xavg} using Eq. (2.9) may deviate from the actual value of an arbitrary $\boldsymbol{x}(t)$ as the threshold δ approaches σ. This is because in practice, s_x only takes values at quantized amplitudes in ADM that are integer-multiple of δ. In STR, the s_x sampling occurs at more randomized a_x due to complete reset and refractory period.

2.2.2.2 Queueing Delay T_{DQ} Variation

Arbitration is essential for sparse spike encoding schemes like STR and ADM. Unlike some early spiking sensors that use pulse-frequency modulation (PFM) for spike encoding [5], [77] where several missing spikes due to collision in unfettered communication channels do not have significant impact on the average spike frequency within a time window, the reconstruction SDR from spike trains encoded by STR and ADM, i.e. the representation integrity of input signals in spike domain, relies on the presence and precise timestamp of each spike. Arbitration guarantees the transmission of every spike, but it compromises the timestamp accuracy. Because the spikes are only recorded or processed on the receiving end, if the generated spikes pass through the transmission channel freely without any waiting, the timestamps are intact and otherwise skewed if they need to wait for the transmission of earlier generated spikes. This kind of system can be studied using queueing models, and the variation of queueing delay T_{DQ} can be characterized by the waiting time distribution $P(W)$.

Integrated CMOS sensors can be categorized as either 1-D sensors like the silicon cochlea and optical line sensors [14], or 2-D sensors like the silicon retina, and neural recording [78] and biochemical sensing [79] arrays. For 1-D spiking sensors as shown in Figure 2.9(a), each sensing element generates a stream of spike trains. The spike trains from all sensing elements are arbitrated according to the first-come-first-serve (FCFS) rule, which can be implemented as the 1-D version of the fair address-event-representation (AER) circuits [80]. Assuming an ideally fair arbiter which does not distinguish the identity among different sensing elements and assign priorities, all spikes enter one queueing line abiding by the time order of their generation. Each spike is served by a single server that is either a CPLD or an FPGA for off-chip recording, or a TDC for on-chip recording. The service time is normally fixed in the recording case. For a practical sensor under some input stimuli, the pattern of spike arrival forms a particular stochastic process. In general, the queueing model G/D/1 can be used, where G indicates that the inter-spike interval of the arrival spike train has an arbitrary distribution, D means the deterministic service time, and 1 means a single server. Here we consider the Poisson arrival to demonstrate the effect of spike queueing on the integrity of signal representation using the SDR metric, and hence the M/D/1 model is used where M stands for the Poisson arrival.

To obtain the $P(W)$ of an M/D/1 queue, two parameters the average rate of spike arrival λ_{q1D} and the service rate μ_q need to be determined. μ_q is the reciprocal of the service time which is the sum of the intrinsic delay of the AER circuit and the fixed time assigned for registration of one spike address and timestamp. λ_{q1D} depends on a specific sensor model. Taking silicon cochlea with m total sensing elements as the example, each sensing element or channel contains a BPF with a central frequency of f_i ($i \in N^+ \cap i \in [1, m]$) cascaded by a spike encoder with a threshold δ_i. The f_i's of all the BPFs are geometrically scaled with a ratio of r_{fi}. With the predefined frequency range from $f_1 = f_L$ to $f_m = f_H$, r_{fi} can be calculated as:

$$r_{fi} = \left(\frac{f_H}{f_L}\right)^{\frac{1}{m-1}}$$

If the spike rate of the ith ADM spike encoder is λ_i, λ_{q1D} is the sum of all λ_i:

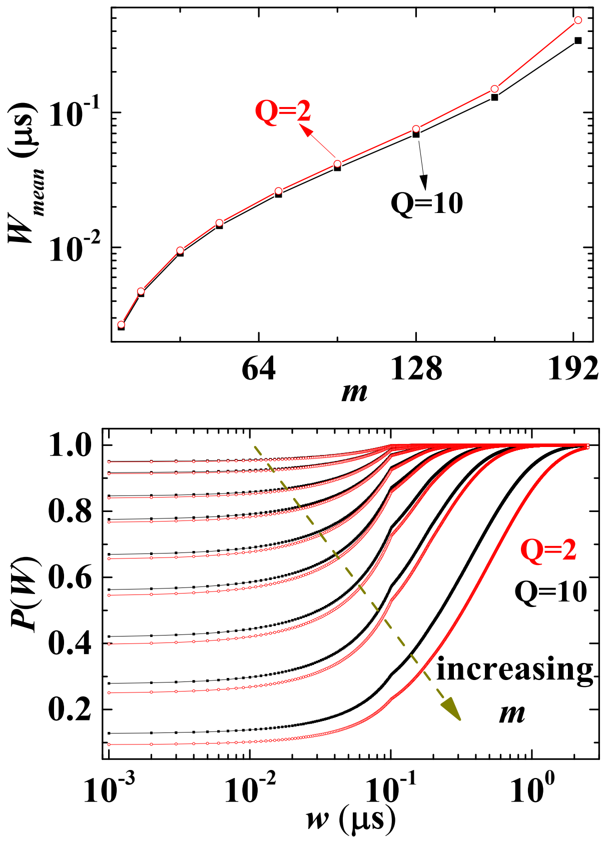
$$\lambda_{q1D} = \sum_{i=1}^{m} \lambda_i$$

To calculate λ_i of a bandpass-filtered white Gaussian noise input $\boldsymbol{x}(t)$, its ψ_0'' is first computed as:

$$\psi_0'' = -\frac{4\pi^2\sigma_i^2}{f_{Hi} - f_{Li}} \int_{f_{Li}}^{f_{Hi}} f^2 df = -\frac{4}{3}(\pi\sigma_i)^2 (f_{Hi}^2 + f_{Hi} f_{Li} + f_{Li}^2)$$

Table 2.1. Parameter values for computing W_{mean} and $P(W)$ of queueing models.

Symbol	**Description**	**Value**
σ	Signal rms amplitude	0.382
δ	Encoder threshold	0.125
f_H	Highest frequency for cochlea	20k Hz
f_L	Lowest frequency for cochlea	20 Hz
f_{vis}	Pixel bandwidth for retina	20 Hz
μ_q	Service rate or spike departure rate	10^7/s
p	Probability of a retina pixel being active	0.16



(a) (b)

Figure 2.10. (a) Mean waiting time W_{mean} and (b) waiting time distribution $P(W)$ of an M/D/1 queue with the specified parameters in Table 2.1.

where σ_i is the rms amplitude, and f_{Hi} and f_{Li} are the lowpass and highpass corner frequencies of the i^{th} BPF. f_{Hi} and f_{Li} are related to f_i by the equations below:

$$f_{Hi} - f_{Li} = f_i / Q_i\,,\ \ f_{Hi} f_{Li} = f_i^2$$

where Q_i is the quality factor of the i^{th} BPF. Using Eq. (2.8), λ_i can be calculated as:

$$\lambda_i = \frac{2\int_0^\infty s_x \int_{-\infty}^{\infty} p(a_x, s_x) da_x ds_x}{\delta_i} = \sqrt{\frac{2}{\pi}} \frac{\sqrt{-\psi_0''}}{\delta_i} = \frac{2\pi\sigma_i}{\delta_i} \sqrt{\frac{2(f_{Hi}^2 + f_{Hi} f_{Li} + f_{Li}^2)}{3\pi}} = 2\pi \frac{\sigma_i f_i}{\delta_i} [\frac{2}{3\pi}(3 + \frac{1}{Q_i^2})]^{1/2}$$

Note that an STR encoder gives a spike rate lower than the λ_i computed above, depending on the values of T_{Dk} and T_H. For simplicity, it is assumed for an STR encoder to have the same λ_i as an ADM with the consequence of underestimated reconstruction SDR. More realistic λ_i estimation of both ADM and STR should take the actual input signal statistics into account. Further assumptions of identical σ_i, δ_i and Q_i for all channels denoted as σ, δ and Q give the following simplified explicit expression of λ_{q1D}:

$$\lambda_{q1D} = \frac{2\sigma}{\delta}\left[\frac{2\pi}{3}\left(3+\frac{1}{Q^2}\right)\right]^{1/2}\frac{f_m(1-r^{-m})}{1-r^{-1}} \tag{2.11}$$

Now that we have both λ_{q1D} and μ_q, the mean waiting time W_{mean} [81] and the waiting time distribution $P(W)$ [82] of an M/D/1 queue can be computed by:

$$W_{mean} = \frac{\rho}{2\mu_q(1-\rho)} \tag{2.12}$$

$$P(W \le w) = (1-\rho)\sum_{i=0}^{N}\frac{[-\lambda_{q1D}(w-\frac{i}{\mu_q})]^i}{i!}e^{\lambda_{q1D}(w-\frac{i}{\mu_q})} \quad \left(\frac{N}{\mu_q} \le w < \frac{N+1}{\mu_q}\right) \tag{2.13}$$

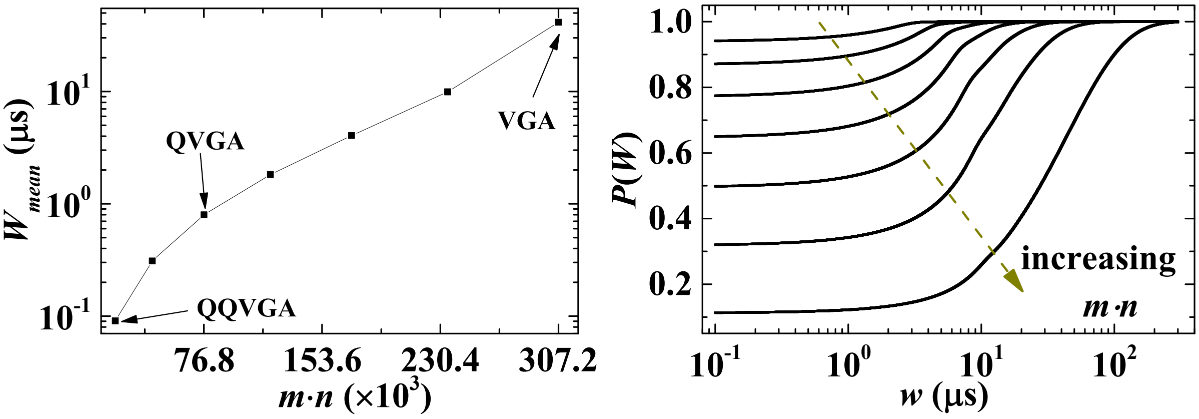
where $\rho=\lambda_{q1D}/\mu_q$ is the traffic intensity. With the parameter values given in Table 2.1, W_{mean} and $P(W)$ as a function of the number of channels m are plotted in Figure 2.10. The resolution of the waiting time w in computation is 1 ns. As m increases, W_{mean} increases exponentially, the sigmoid $P(W)$ shifts towards larger waiting time with increasing variance, and the probability of immediate service with no delay decreases. Both W_{mean} and $P(W)$ are slightly dependent on Q, and they will be later used for the computation of reconstruction SDR of the 1-D cochlea.

The first generation of 2-D AER implemented handshakes of the row and column requests of a sensing element E(i,j) in series [80], and then the row and column addresses are transmitted in parallel. The average service rate of one pixel μ_{qpixel} in this transmission scheme is:

$$\mu_{qpixel} = \frac{1}{T_{row} + T_{col}}$$

where T_{row} and T_{col} are the durations needed for a row and a column handshake, respectively. A G/D/1 model can still apply to this case, and the 1-D sensor analysis above is also valid if the spike arrival is Poisson. To enhance the service rate, the second generation of 2-D AER employs the burst-mode word-serial transmission scheme [83]. As illustrated in Figure 2.9(b), for a pixel array of size $m \cdot n$, whenever one or several pixels in row R_i ($i \in N^+ \cap i \in [1, m]$) initiate the spike transmission, R_i enters the queue of row requests waiting for its row acknowledge from the single server. The fair arbitration also follows the FCFS rule. By the time when the R_i request gets served, i.e. the row address and the row request timestamp are recorded, all the 'active' pixels in R_i ('active' means pixels having spike generated before R_i gets served) starts to transmit their column addresses in burst from C_1 to C_n and they all have the same timestamp as the R_i request. After the column address of the last active pixel in R_i is registered, the next row request waiting in the queue will be served, and the process repeats. Because one active pixel is only allowed to generate another spike after the current spike is served and acknowledged, the number of active columns in one row request cannot be larger than n. In this case, μ_{qpixel} is:

$$\mu_{qpixel} = \frac{1}{T_{row}/n_{avg} + T_{col}} \approx \frac{1}{T_{col}} \quad (\text{if } 1 << n_{avg} \le n)$$



(a) (b)

Figure 2.11. (a) Mean waiting time W_{mean} and (b) waiting time distribution $P(W)$ of an $M^X/D/1$ queue with the specified parameters in Table 2.1.

where n_{avg} is the average number of active columns in one row request. In silicon retina, n_{avg} is mostly determined by stimulus patterns which are often spatiotemporally correlated. If n_{avg} is large, μ_q is improved by a ratio of $1+T_{row}/T_{col}$. T_{row} is usually not smaller than T_{col}, and could be even as twice as large because of the fair arbitration [84]. Therefore, μ_q can be improved by ×2~×3 at the cost of precise timestamps of the spikes in the same row request.

With no assumption on the arrival pattern of the row requests, the second generation of 2-D AER that is currently in use for DVSs [18] can be described by a $G^X/D/1$ queueing model, where X denotes the bulk arrival of multiple active columns in one row request and the number of active columns n_a is drawn from another independent distribution. The model implies that the delay of a row request is the delay for all the active columns in this row, which is only true if all the active columns generate spikes at the same time. In practice, one pixel may become active at the moment just before the burst column transmission, which may be much later than its row request due to queue waiting, and consequently its delay is overestimated. In this sense, the delay distribution from the $G^X/D/1$ model gives an upper bound of T_{DQ} variation. If a Poisson process is assumed for the row requests in the following analysis, the combination of multiple active columns and a Poisson row requests lead to the compound Poisson process [85], and the corresponding queueing model is $M^X/D/1$.

Let λ_{q2D} be the mean arrival rate of row requests in a 2-D silicon retina. In light of the principle of spike number conservation of a whole 2-D array, the equality below holds:

$$\lambda_{q2D} \cdot n_{avg} = \lambda_{pix} \cdot m \cdot n \cdot p$$

where λ_{pix} is the mean spike rate of one active/stimulated pixel, and p is the probability of one pixel being active. n_{avg} was shown to be positively dependent on the traffic intensity because the spikes contained in a row request can be accumulated during the queue waiting [86]. However, it will be shown later that this dependence is negligible in a particular case of silicon retina where λ_{pix} and p are small and in turn the probability of spike accumulation is very low. With the assumption of binomial distribution of n_a, the probability mass function p_a of spike number in a row request is written as:

$$p_a = \binom{n-1}{a-3} p^{a-1} (1-p)^{n-a+2}$$

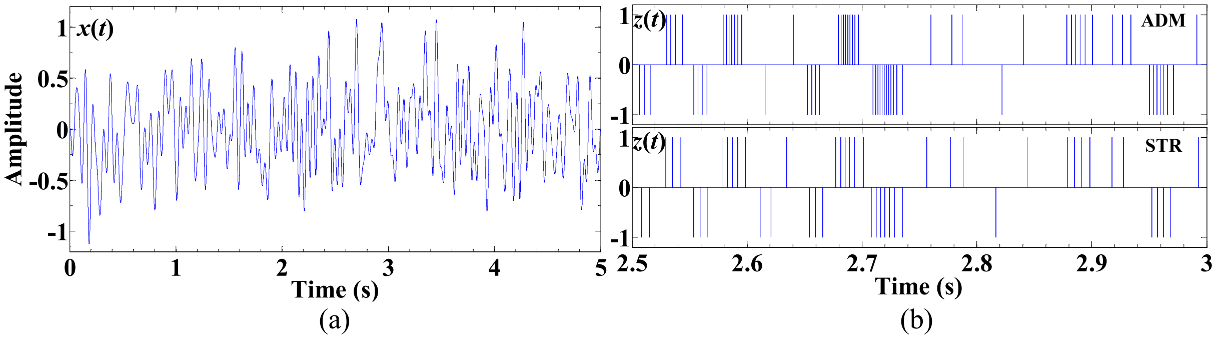


Figure 2.12. (a) Waveform of the white Gaussian noise input $\boldsymbol{x}(t)$ used for simulation of SDR degradation; (b) exemplary encoded spike train $\boldsymbol{z}(t)$ with both ADM and STR within the time window [2.5 s, 3 s].

where a has the support of $a\epsilon N^{+}\cap a\epsilon[3, n+2]$. Note that a is not from 1 to n because $T_{row}\approx 2T_{col}$ is considered. n_{avg} therefore can be written as:

$$n_{avg} = \sum_{a=3}^{n+2}(a-2)p_a$$

For a bandlimited white Gaussian noise input, λ_{pix} of an ADM is [57]:

$$\lambda_{pix} = 2\sqrt{\frac{2\pi}{3}}\frac{\sigma f_{vis}}{\delta}$$

where f_{vis} is the visual signal bandwidth. With λ_{pix} and n_{avg} obtained from the equations above, λ_{q2D} can be computed as:

$$\lambda_{q2D} = \frac{\lambda_{pix}\cdot m\cdot n\cdot p}{n_{avg}} \tag{2.14}$$

The mean waiting time W_{mean} [87] and waiting time distribution $P(W)$ [88] of row requests in an M^{X}/D/1 queue can be computed as:

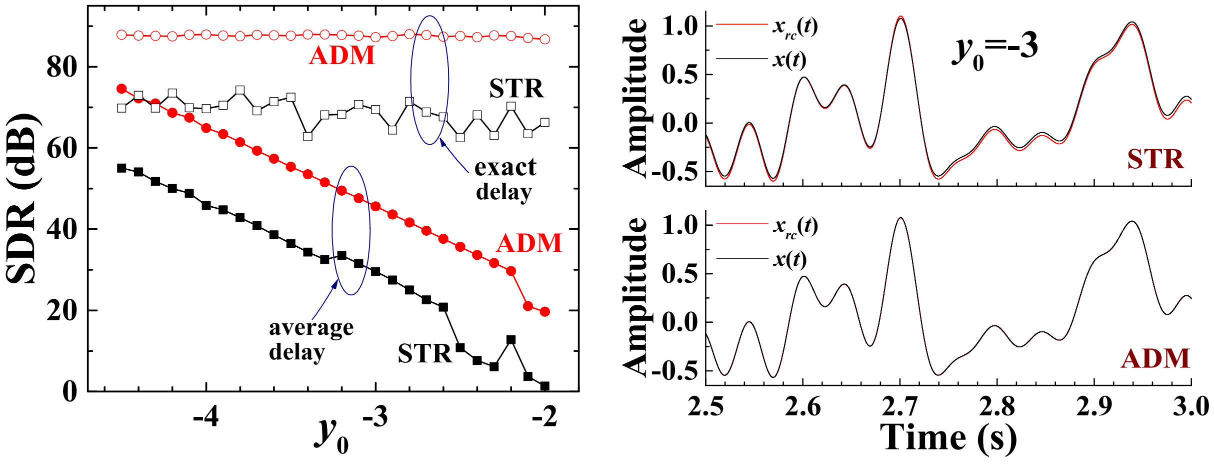
$$W_{mean} = \frac{\rho}{2\mu_q(1-\rho)}\left(\frac{\sum_{a=3}^{n+2}a(a-1)p_a}{\sum_{a=3}^{n+2}ap_a}+1\right) \tag{2.15}$$

$$P(W\le w) = \sum_{k=0}^{N}\sum_{i=0}^{N+1-k} f_i g_k\left(\frac{N+1}{\mu_q}-w\right) \quad \frac{N}{\mu_q}\le w < \frac{N+1}{\mu_q} \tag{2.16}$$

where ρ is the traffic intensity that is defined as:

$$\rho = \frac{\lambda_{q2D}}{\mu_q}\sum_{a=3}^{n+2}ap_a$$

f_i denotes the probability of the stationary distribution with i row requests being held in the system, and $g_k(t)$ denotes the probability that exact k new row requests enter the queue during an arbitrary time interval of length t (the computation of f_i and $g_k(t)$ sees Appendix). With the parameter values given in Table 2.1, W_{mean} and $P(W)$ as a function of the total pixel number $m{\cdot}n$ are plotted in Figure 2.11. An aspect ratio



(a) (b)

Figure 2.13. (a) Reconstruction SDR of ADM and STR as the function of the comparison speed indicator y_0. Hollow and solid symbols represent reconstruction using exact delay and average delay, respectively; (b) Exemplary recovered signal $\boldsymbol{x}_{rc}(t)$ at y_0=-3 within the time window of [2.5 s, 3 s].

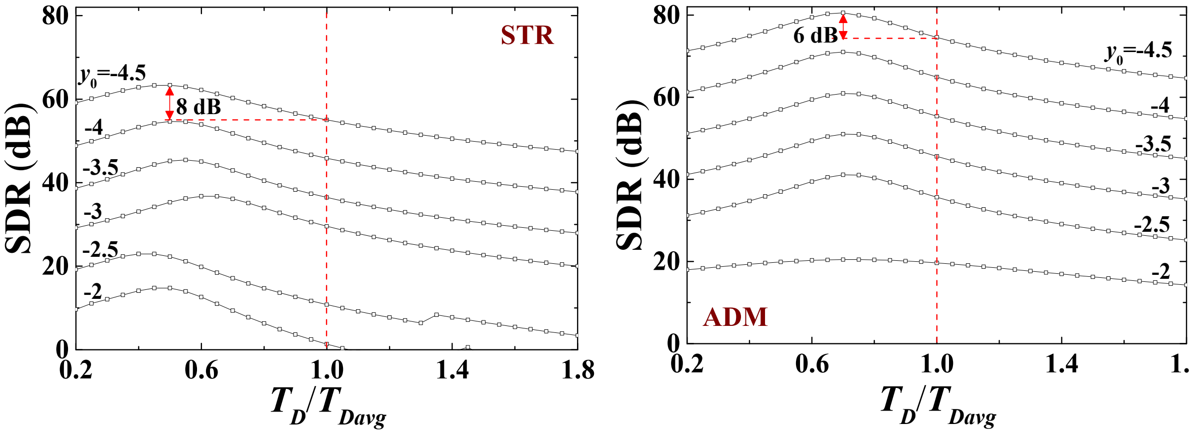


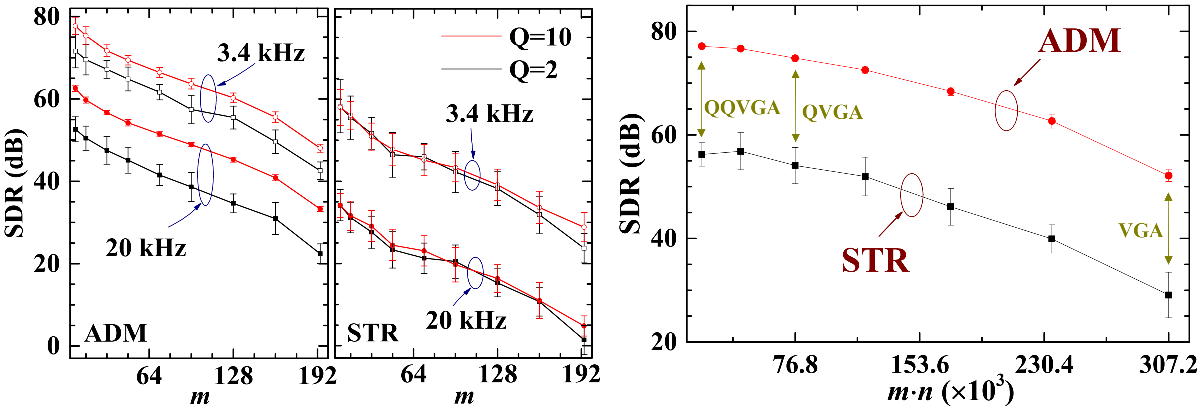
Figure 2.14. SDR versus the ratio T_D/T_{Davg} for STR and ADM. T_D is the actual average delay used for spike train decoding, and T_{Davg} is the estimated average delay calculated by Eq. (2.10).

of m:n=3:4 is used. The popular video graphics array (VGA) resolution and its scaled versions are labeled in Figure 2.11(a). Similar to the 1-D case, W_{mean} increases exponentially as $m \cdot n$ increases, and so does the variance of $P(W)$. However, the W_{mean} values are about ×100 larger than the 1-D case with comparable traffic intensities (e.g. both around 0.88 for the largest W_{mean}), which is attributed to bulk arrival. Note that a decreased m/n results in increased W_{mean} even though $m \cdot n$ is kept constant.

We now verify an earlier statement that spike accumulation during waiting can be neglected in the 2-D silicon retina. In the worst delay case that we have considered when the array has a VGA size, the average accumulated spike number N_{accu} can be calculated as:

$$N_{accu} = \lambda_{pix} n p W_D$$

where W_D is the waiting delay before the requesting row gets served. We take a large W_D of 200 μs ac-



(a) (b)

Figure 2.15. (a) SDR degradation in (a) 1-D silicon cochlea as the function of the channel number m with two different central frequencies 3.4k and 20k Hz, and two different Q's 2 and 10 evaluated, and (b) in 2-D silicon retina as the function of the array size $m \cdot n$ for both ADM and STR.

cording to Figure 2.11(b) which already has a very small chance to happen, and N_{accu} is calculated to be 3.6. In the VGA case, n_{avg} is calculated to be 103, much larger than 3.6.

2.2.3 SDR Degradation

The white Gaussian noise shown in Figure 2.12(a) is used as the input $\boldsymbol{x}(t)$. It has zero mean, variance σ of 0.382 and bandwidth f_n of 20 Hz. Figure 2.12(b) shows the example spike train $\boldsymbol{z}(t)$ encoded by ADM and STR encoders within the time window [2.5 s, 3 s] respectively, with threshold δ of 0.125, fixed delay T_D and refractory period T_H both of 1 ms. It is clear that STR gives a $\boldsymbol{z}(t)$ with a smaller sampling rate compared to ADM. With this input $\boldsymbol{x}(t)$ and using the nonlinear decoding algorithms described in Section 2.2.1 for signal reconstruction from $\boldsymbol{z}(t)$, SDR degradation due to both T_{DC} and T_{DQ} variation will be discussed as follows.

The SDR degradation due to T_{DC} variation is plotted as the function of the comparator speed indicator y_0 in Figure 2.13(a) for both ADM and STR. The time resolution in simulation is 1 µs, and T_H for STR is fixed at 1 ms. The exact delay T_{Dk} at each threshold-crossing is generated according to Eq. (2.7) with computed slope values. The average delay T_{Davg} is calculated using Eq. (2.10). The hollow symbols and the solid symbols represent the SDR computed with exact delay and average delay, respectively. The SDR only accounts for the time interval [0.2 s, 4.8 s] to exclude the relatively large distortion in the vicinity of time boundaries. Both ADM and STR maintain high SDR (87.6 dB and 68.8 dB in average) regardless of y_0 if exact delay is used for reconstruction. In practice however, T_{Dk} is not available due to the lack of instantaneous slope at threshold-crossings, therefore T_{Davg} is used to approximately recover $\boldsymbol{x}(t)$, and SDR decreases by about 20 dB for the increase of y_0 by 1. Intuitively this is because the variance of the delay error T_{Dk}-T_{Davg} increases with y_0. Figure 2.13(b) gives an example of recovered signal $\boldsymbol{x}_{rc}(t)$ at y_0=-3. With a given SDR specification, the y_0 value can be determined, and in turn the comparator parameters A_{DC} and τ_c. Larger SDR requires either larger A_{DC} that can be achieved by using long channel-length transistors, or smaller τ_c by increasing comparator biases. Increased signal bandwidth with the same A_{DC} and τ_c is expected to result in decreased SDR because of comparatively larger jitter in timestamps. Figure 2.14 examines the effectiveness of T_{Davg} estimation using Eq. (2.10) with replacing the T_{Davg} by a sweeping T_D

for decoding, which clearly shows that the calculated T_{Davg} is not optimal in terms of the highest achievable reconstruction SDR. For example, at y_0=-4.5, the SDR obtained with T_{Davg} is 8 dB lower than the best SDR at T_D=0.5T_{Davg} in STR, and 6 dB lower at T_D=0.7T_{Davg}. T_{Davg} becomes an increasingly better estimation as the threshold δ decreases.

The SDR degradation in the 1-D silicon cochlea case is plotted as the function of the channel number m in Figure 2.15(a). The error bar represents the SDR variance obtained in 20 runs. The waveform in Figure 2.12(a) is scaled to 5 ms with a bandwidth of 20k Hz to be used as the input to the BPF bank. As examples, the input signal is filtered by two BPFs with different central frequencies at 3.4k and 20k Hz and different Q's of 2 and 10 before it is encoded into spikes. The 3.4k and 20k Hz are chosen for they are the telephony voice frequency band and the upper limit of human hearing frequency range. The time step for simulation is 1 ns, and T_H is 5.9 μs for 3.4k Hz and 1 μs for 20k Hz. The exact delay T_{Dk} at each threshold-crossing is drawn from the distribution $P(W)$ in Figure 2.10(b), and the average delay T_{Davg} is from Figure 2.10(a). The decoding bandwidth is set to the 3-dB lowpass corner frequency. With m increasing from 8 to 194, the SDR approximately decreases by 30 dB for both ADM and STR in all combinations of central frequencies and Q's. The 20k-Hz case always gives a worse reconstruction SDR than the 3.4k-Hz case because signals with higher frequency component result in finer timescale of the encoded spike train which is more prone to the negative impact of spike timestamp uncertainty due to time quantization. Although theoretically the decoding algorithm for STR given in Section 2.2.1 can perform perfect reconstruction, the lower SDR compared to ADM in Figure 2.15(a) can be attributed to the finite time support, i.e. less samples in a given period. Q=10 gives a higher SDR in ADM than Q=2, which is probably due to the fact of slightly smaller T_{DQ} variation as can be seen in Figure 2.10(b). The less obvious benefit of a higher Q in STR could be the result of its much lower reconstruction SDR, masking the effect of Q dependence.

The SDR degradation in the 2-D silicon retina case is plotted as the function of the array size $m \cdot n$ in Figure 2.15(a). The waveform in Figure 2.12(a) is used as the input to spike encoders. The time resolution in simulation is 0.1 μs, and T_H is 1 ms. The exact delay T_{Dk} at each threshold-crossing is drawn from the distribution $P(W)$ in Figure 2.11(b), and the average delay T_{Davg} is from Figure 2.11(a). The decoding bandwidth is set to 20 Hz. With $m \cdot n$ increasing from 19.2k (QQVGA) to 307.2k (VGA), the SDR approximately degrades by 25 dB for both ADM and STR.

2.3 Conclusion and Remarks

This chapter presents the performance analysis of and comparison between two spike encoding mechanisms, namely the self-timed reset (STR) and the asynchronous delta modulation (AMD), both of which have been implemented in silicon spiking sensors. Linear decoding of the spike trains produced by these two types of encoders shows that, with non-idealities considered like feedback delay due to comparison and spike queueing during spike transmission, and switch-holding during complete reset, STR usually results in less reconstruction SDR than ADM. Further analysis using nonlinear decoding algorithms quantitatively reveals the effects of jitters in spike timestamps caused by comparison and queueing on SDR: larger jitters results in more SDR degradation

For practical system designs of spiking sensors, the type of spike encoder can be chosen according to the specified SDR requirements. An STR encoder or an integrate-and-fire encoder should be adopted if sufficient SDR can be obtained because of the simpler and more area-efficient circuit implementations compared to an ADM encoder even though the later usually have a higher-fidelity spike-domain representation. Circuit parameters that are related to timestamp jitters caused by comparison delay T_{DC} variation and queueing delay T_{DQ} variation can also be determined by the targeted SDR. T_{DC} variation is associated with the comparator DC gain and pole frequency, and T_{DQ} variation is related to the encoder threshold, number of encoders, spike train service rate, and system architecture that can be modeled by a corresponding queueing model. The queueing models M/D/1 and M^X/D/1 used in the analysis above can be replaced by others if necessary which can more accurately describe the patterns of spike arrival/departure and server number of the system under consideration.

In the context of internet of things, the analysis in this chapter has implications in future low-power spiking sensor design. The comparator speed and the AER communication channel bandwidth can be reduced to the minimal level where the required encoding quality specifications (e.g. measured by SDR) can still be satisfied, and therefore the bias current of the comparators and the supply voltage of the AER circuits and be lowered to a certain degree to save system power. The SDR metric used here is directly relevant to faithful recording applications like optical neuroimaging [89] and electrical neural signal acquisition [78] where a small encoding error is of paramount importance. Compared to traditional clocked Nyquist sampling, asynchronous spike encoding schemes like STR and ADM have the advantage of reduced data redundancy, which is essential in minimizing RF transmission power for wireless sensors. For emerging smart sensing systems with low-power embedded processing for within-sensor classification and recognition [42], SDR may not be the best measure and further study is needed to establish the link between signal encoding quality and system performance in terms of classification/recognition accuracy.

2.4 Appendix

This appendix gives the method of computing f_i and $g_k(t)$ in Eq. (2.16) based on the fully probabilistic analysis described in [88]. Recall that $g_k(t)$ denotes the probability that exact k new row requests enter the queue during an arbitrary time interval of length t. If j row requests arrive within t, the probability $\psi_k(j)$ of these j row requests containing k spikes can be derived by recursion (note that one row request is equivalent to two real spikes because of $T_{row}\approx 2T_{col}$):

$$\psi_k(1) = p_k \ (k \in N \cap k \in [3, n+2]) \ , \ \psi_k(j) = \sum_{a=3}^{k-3(j-1)} p_a \psi_{k-a}(j-1) \ (k \geq 3j) \ , \ \psi_k(j) = 0 \ (k < 3j)$$

where p_k is the probability mass function of spike number in a row request. Given the assumption that the number of row request arrival within time t is Poisson-distributed, $g_k(t)$ can be expressed as:

$$g_k(t) = \sum_{j=1}^{\infty} \psi_k(j) \frac{(\lambda_{q2D} t)^j}{j!} e^{-\lambda_{q2D} t} \tag{2.a1}$$

Let $f_i(t_0)$ denote the probability of the system holding i row requests at time t_0. By conditioning on the number of spikes present at t_0, the equation below holds:

$$f_i(t_0 + D) = \sum_{l=0}^{1} f_l(t_0) g_i(D) + \sum_{l=2}^{i+1} f_l(t_0) g_{i+1-l}(D) \ (i \in N_0)$$

where $D=1/\mu_q$ is the deterministic service time. The stationary distribution f_i is found by letting $t_0 \rightarrow \infty$:

$$f_i = \sum_{l=0}^{1} f_l g_i(D) + \sum_{l=2}^{i+1} f_l g_{i+1-l}(D) \ (i \in N_0) \tag{2.a2}$$

To solve f_i from Eq. (2.a2), the geometric tail approach described in [90] is used. For $i \geq M$ where M is a sufficiently large positive integer, f_i is approximated as:

$$f_i \approx f_M \varepsilon^{M-i} \ (i \geq M)$$

The scaling factor ε can be solved by setting the denominator of the probability generating function of the $M^X/D/1$ queue [91] to 0:

$$1 - \varepsilon \cdot \exp[-\frac{\lambda_{q2D}}{\mu_q}(\sum_{a=3}^{n+2} p_a \varepsilon^a - 1)] = 0$$

Eq. (2.a2) can be then written as:

$$f_i = \sum_{l=0}^{1} f_l g_i(D) + \sum_{l=2}^{i+1} f_l g_{i+1-l}(D) \ (i < M)$$

which is a linear equation system with M dimension. The normalization equation can be written as:

$$\sum_{i=0}^{M-1} f_i + f_M \sum_{i=M}^{\infty} \varepsilon^{M-i} = 1 \Rightarrow \sum_{i=0}^{M-1} f_i + \frac{f_M}{1-\varepsilon^{-1}} = 1$$

Now an $(M+1)$-dimensional linear equation system is complete for $M+1$ variables f_i, $i \in [0, M]$. In the numerical simulation to obtain Figure 2.11(b), $M=200$ is used.

Chapter 3: Silicon Retina with Enhanced Temporal Contrast Sensitivity and Spike Encoding

This chapter elaborates on a new silicon retina design [89] based on the original dynamic vision sensor (DVS) [6] that is briefly introduced in Chapter 1. The new features of this design are down to 1% temporal contrast (TC) sensitivity with less than 35% 1σ variation and improved spike encoding thanks to the employment of in-pixel asynchronous delta modulator (ADM).

Enhanced TC sensitivity can be useful for applications such as fine texture recognition, fluorometric calcium imaging and voltage-sensitive dye imaging (VSDI). For example, in VSDI, the transient fluorescence signal change is typically below 1% within tens of milliseconds [92], while in calcium imaging the signal is about 10% for a single action potential [93]. Several reported minimum TC sensitivity values of DVSs lie around 10%. Setting the thresholds of spike encoders smaller than minimum TC sensitivity results in excessive noise spikes, which makes it difficult to reliably detect any transient visual change. A recent design improved the TC sensitivity to 1.5% by incorporating a subthreshold transimpedance preamplifier stage [16]. Despite the measured 2.1%-2.5% rms equivalent input TC noise, the 1.5% TC sensitivity in [16] was obtained only by averaging the output spikes over the entire sensor array. Hence a low noise design of the pixel front-end is essential for a reasonable signal-to-noise ratio (SNR) at the sensor output enabling an even smaller TC sensitivity setting.

As already discussed in Chapter 2, with linear decoding, ADM normally results in a better encoding quality than self-timed reset (STR) that is commonly used in prior DVSs. This is also true for nonlinear decoding with finite time support. Improved spike encoding can profit applications where video reconstruction from spikes is required. For example, the continuous-time signal dynamics is of interests in optical bioimaging. The input waveform needs to be recovered from the output spike train using the linear or nonlinear decoding methods presented in Chapter 2. Another example is the video decompression from vision sensors that can acquire both intensity images and asynchronous spikes in response to temporal intensity changes [94], [95]. It was pointed out in [94] that the signal loss due to STR encoding greatly degrades the decompression quality.

The content of this chapter is organized as follows: Section 3.1 describes the system architecture and design considerations of pixel gain division; Section 3.2 gives the detailed pixel design, including: Section 3.2.1, analysis of the photoreceptor; Section 3.2.2, the capacitively-coupled programmable gain amplifier (CC-PGA) for sensitivity enhancement including a compact two-stage Opamp with pseudo-cascode compensation; Section 3.2.3, the in-pixel ADM design; Section 3.2.4, the in-pixel asynchronous logic (IPAsyncL) to generate ADM control signals and for communication of generated spikes with peripheral address-event-representation (AER) circuitry; Section 3.3 presents the experimental results; Section 3.4 concludes the chapter with several remarks.

3.1 System Architecture

The system architecture of the new DVS (called ADMDVS) is shown in Figure 3.1. It consists of a 60×30 pixel array, the X/Y address encoder that encodes the column and row addresses into 6- and 5-bits output respectively, the X/Y AER logic and arbiter for asynchronous transmission of pixel spike output, and the asynchronous state machine for the communication with off-chip complex programmable logic device (CPLD) or field programmable gate array (FPGA). The AER is a modified version [84] of the

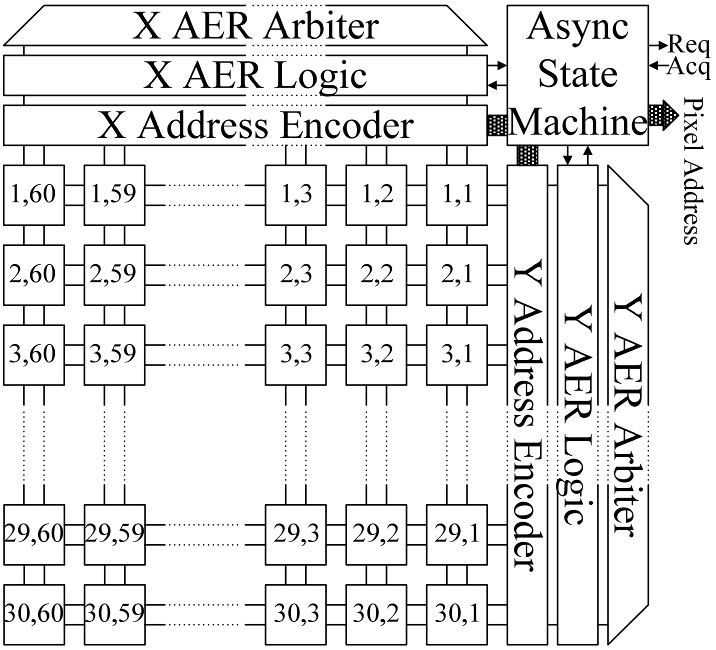


Figure 3.1. System architecture of the ADMDVS chip, including the pixel array, address encoder, address-event representation (AER) logic and arbiter, and asynchronous state machine. The bias generator and shift-register configuration chain are not shown.

word-serial burst-mode protocol as proposed in [83]. The working mechanism of this AER scheme can be understood from the queueing model described in Chapter 2.

The building blocks of one pixel are shown in Figure 3.2. The photoreceptor that is composed of a photodiode and a transimpedance amplifier (TIA) logarithmically converts a small-signal photocurrent i_{ph} into a voltage output ΔV, which is then buffered by a source follower (SF) before it is amplified by a CC-PGA. The amplified analog signal is encoded into spikes by the ADM and the spikes are communicated off-chip by in-pixel asynchronous logic (IPAsyncL) and the peripheral AER circuitry. The bias currents of the TIA, SF and CC-PGA are adjustable to control the front-end bandwidth. The gain of the CC-PGA is programmable with 2 bits, and the threshold voltages of the ADM can be continuously adjusted for different TC sensitivity settings. The illustration of the pixel communication with the periphery AER is simplified; the complete sensor array has the 2-dimensional AER to communicate X and Y addresses of active pixels in a burst-mode word-serial fashion as depicted in Figure 3.1.

The aim of the ADMDVS design is to achieve below 1% TC sensitivity using ADM for spike encoding instead of STR. Sufficient front-end gain is necessary to amplifier small TC signals. For example, 1% TC corresponds to about 0.25 mV photoreceptor output. With a 250 mV ADM threshold, a gain of 60 dB is needed. If the gain is only provided by the CC-PGA, the total capacitance would be about 5 pF given a 5 fF feedback capacitor. In a 0.18 μm CMOS process, the unit capacitance of MIM is normally 1 fF/μm^2, which leads to a 71×71 μm^2 capacitor size. Although large closed-loop gain reduces the CC-PGA's noise contribution (explained in Section 3.3), in order to have a pixel size of about 30×30 μm^2 that is comparable to previous DVS pixels, the ADM needs to take part of the gain. If the CC-PGA has a gain of 36 dB and the ADM 24 dB, the total capacitance is about 400 fF if both CC-PGA and ADM have a 5 fF feedback capacitor, and the area is about 20×20 μm^2 which is affordable by a 30×30 μm^2 pixel.

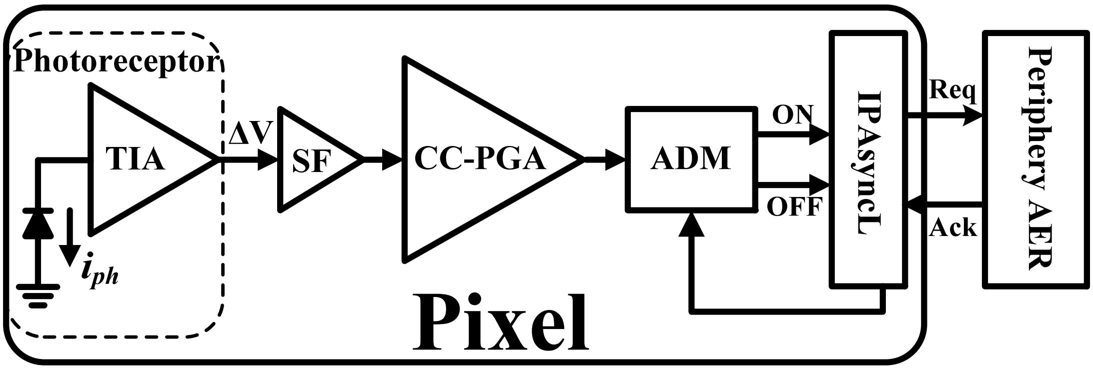


Figure 3.2. Building blocks of one pixel, including the photoreceptor, source follower (SF), capacitively-coupled programmable gain amplifier (CC-PGA), asynchronous delta modulator (ADM) and in-pixel asynchronous logic (IPAsyncL).

3.2 Pixel Design

3.2.1 Photoreceptor

Two types of photoreceptors have been used in DVSs, and they differ in the TIA feedback element as shown in Figure 3.3: the one in Figure 3.3(a) uses an nFET, and is called a source-follower photoreceptor (SFPR) [6]; the one in Figure 3.3(b) uses a pFET, and is called a common-gate photoreceptor (CGPR) [16]. The detailed analysis of SFPR and CGPR is given below.

3.2.1.1 SFPR

The output DC V_{outDC} of SFPR is at $V_{gs1}+V_{gs3}$. M_{3s} is usually in deep subthreshold because of small DC photocurrent I_{PH}, which can result in negative V_{gs3}. The headroom problem may arise when V_{outDC} is too low for M_{2s} or even M_{1s} to stay in saturation. V_{outDC} could always stay at a reasonably high level to guarantee the saturation of M_{2s} in 0.35 µm CMOS [6], but not anymore as process scales and transistor threshold voltage decreases. Recent vision sensors with SFPR in 0.18 µm CMOS used 3.3-V nFETs for M_{1s} and M_{3s} to circumvent this problem [18].

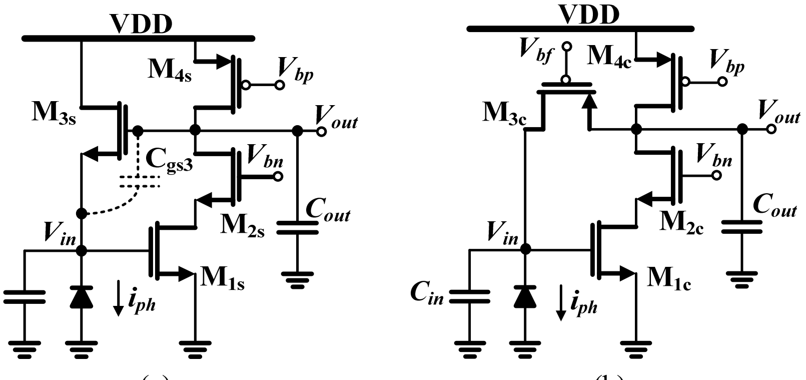
The small signal equivalent circuit of SFPR is shown in Figure 3.4(a). The transfer function is accordingly calculated as in Eq. (3.1):

$$\frac{V_{out}}{V_T\frac{i_{ph}}{I_{PH}}}=\frac{V_{out}}{\frac{i_{ph}}{g_{s3}}}=\frac{A_{DC}(1-\frac{s}{\omega_z})}{\frac{s^2}{\omega_n^2}+\frac{s}{\omega_n Q}+1}=\frac{\frac{A}{1+\kappa A}(1-s\beta\tau_{out})}{s^2\frac{A}{1+\kappa A}(1+\alpha+\beta)\tau_{in}\tau_{out}+s\frac{A}{1+\kappa A}[\tau_{in}(\alpha(1+\frac{1}{A})+\frac{1}{A})+\tau_{out}(\beta(1-\kappa)+1)]+} \tag{3.1}$$

where V_T is the thermal voltage, A_{DC} is the DC gain of the TIA, ω_n is the natural frequency, ω_z is the zero frequency, Q is the quality factor, κ is the subthreshold slope factor (assume the same for both nFET and pFET), and the other parameters are defined as below:

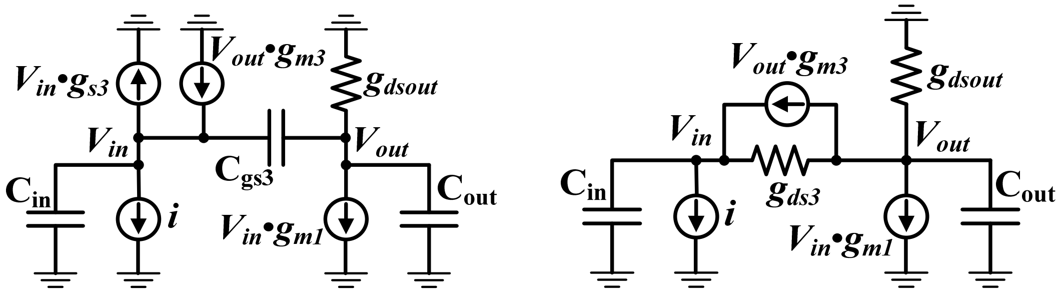
$$A=\frac{g_{m1}}{g_{dsout}},\ \tau_{in}=\frac{\mathrm{C_{in}}}{g_{s3}},\ \tau_{out}=\frac{\mathrm{C_{out}}}{g_{m1}},\ \alpha=\frac{\mathrm{C_{gs3}}}{\mathrm{C_{in}}},\ \beta=\frac{\mathrm{C_{gs3}}}{\mathrm{C_{out}}},\ R_c=\frac{\mathrm{C_{in}}}{\mathrm{C_{out}}},\ R_i=\frac{I_{amp}}{I_{PH}},\ R_\tau=\frac{\tau_{in}}{\tau_{out}}=\kappa R_c R_i,$$

$$m=\kappa(1+\alpha+\beta)\ ,\ n=1+\beta(1-\kappa)$$



(a) (b)

Figure 3.3. Circuit diagrams of (a) the source-follower photoreceptor (SFPR) and (b) the common-gate photoreceptor (CGPR).



(a) (b)

Figure 3.4. Small signal equivalent circuits of (a) the SFPR and (b) the CGPR.

where g_{dsout} and I_{amp} are the output conductance and the bias current of the TIA, respectively. If A>>1, we have:

$$\mathrm{A}_{\mathrm{DC}} = \frac{1}{\kappa},\ \omega_n = \sqrt{\frac{\kappa^2}{m\tau_{in}\tau_{out}}},\ \omega_z = \frac{1}{\beta\tau_{out}},\ Q = \frac{\sqrt{mR_\tau}}{R_\tau(\alpha + \frac{1}{A}) + n}$$

The maximum Q over all illuminance conditions is derived as:

$$Q_{max} = \frac{\sqrt{\frac{mA}{n(1+\alpha A)}}}{2},\ \text{when}\ R_\tau = \frac{nA}{1+\alpha A}$$

If αA>>1, the Q_{max} is simplified as:

$$Q_{max} = \frac{\sqrt{m/n\alpha}}{2},\ \text{when}\ R_\tau = \frac{n}{\alpha} \tag{3.2}$$

To have real poles in this 2nd-order system, R_i has to satisfy:

$$R_i < \frac{-n(\alpha+\frac{1}{A}) + 2m - 2\sqrt{m(m-n\left(\alpha+\frac{1}{A}\right))}}{\kappa R_c(\alpha+\frac{1}{A})^2}\ \text{or}\ R_i < \frac{-n(\alpha+\frac{1}{A}) + 2m - 2\sqrt{m(m-n\left(\alpha+\frac{1}{A}\right))}}{\kappa R_c(\alpha+\frac{1}{A})^2}$$

With practical design parameters such as A=100, κ=0.8, α=0.05, β=0.25, Q_{max} is about 2.03 at R_i=4.38. To have real poles, the range of R_i is calculated to be: $R_i<6.83\times10^{-2}$ or $R_i>2.80\times10^{2}$. Under low illumination, e.g. $R_i=10^4$, the 3-dB bandwidth is about:

$$\omega_{3dB} = 13.4 / \tau_{in}$$

Several comments on the calculations above are given below.

1. If the C_{gs} of M_{3s} is not considered, Q_{max} is directly related to the gain of the amplifier A:

$$Q_{max} = \frac{\sqrt{\kappa A}}{2}, \text{ when } R_\tau = A$$

which is about 4.47. With the presence of C_{gs}, if $\alpha A>>1$, Q_{max} is not dependent on A as shown in Eq. (3.2), and can be designed to be relatively small by choosing the size of M_{3s}. However, the parasitic C_{gs} limits the speed: the 3-dB bandwidth is only 13.4 times larger than that of the simple source-follower photoreceptor. For a relatively large range of photocurrent, the system has two real poles, and the speed is either determined by τ_{in} or τ_{out}. Another benefit of this configuration compared to CGPR might be that the photocurrent I_{PH} could be larger than I_{amp}, and I_{PH} can be readily reused for APS intensity readout [18].

2. The DC gain is dependent on illumination level, which implies contrast inconsistency over a wide dynamic range. (1) Under low illumination, dark current is prominent, and g_{s3} is $(I_{PH}+I_{dark})/V_T$ instead of I_{PH}/V_T. Therefore, the effective DC gain decreases. (2) The DC point of V_{out} changes with I_{PH}, and so does the κ of M_{3s} [96]. Short channel devices have less κ dependence due to the effect of the capacitance between the channel and the drain [97]. This is still valid in UMC 0.18 µm CMOS in Spectre simulation. (3) Under high illumination, M_{3s} may enter moderate inversion, and the effective DC gain may become larger. So overall, M_{3s} with a small L and large W is good for less κ variation over a large illumination range, but it aggravates the headroom problem under low illumination due to more negative V_{gs3}.

3. The C_{gs} of M_{3s} is not constant but dependent on illumination level. It increases with the V_{gs} of M_{3s} [98], which is logarithmically dependent on the photocurrent. This may indicate larger photoreceptor bandwidth under low illumination than expected and smaller Q under high illumination. It is shown that short-channel transistors have larger normalized C_{gs} in deep sub-threshold region than long-channel transistors [99]. Therefore, smaller L does not necessarily guarantee smaller C_{gs} or higher bandwidth, and simulation is needed for optimized Q and speed.

The output noise power of SFPR can be analytically calculated using simple noise models [26]. The transfer function of the noise current from the photodiode and M_{3s} are the same as that of the input small signal i_{ph}. Their contribution to output noise power spectral density (psd) can thus be written as:

$$\overline{V}^2_{nout,ph}(\omega) = \overline{i}^2_{n,ph}\frac{V_T^2}{I_{PH}^2}|H(\omega)|^2 = \overline{i}^2_{n,ph}\frac{V_T^2}{I_{PH}^2}\frac{\frac{1}{\kappa^2}(1+\frac{\omega^2}{\omega_z^2})}{(\frac{\omega^2}{\omega_n^2})^2+\left(\frac{1}{Q^2}-2\right)\frac{\omega^2}{\omega_n^2}+1}, \quad \overline{V}^2_{nout,M_3}(\omega) = \overline{i}^2_{n,M_3}\frac{V_T^2}{I_{PH}^2}\frac{\frac{1}{\kappa^2}(1+\frac{\omega^2}{\omega_z^2})}{(\frac{\omega^2}{\omega_n^2})^2+\left(\frac{1}{Q^2}-2\right)\frac{\omega^2}{\omega_n^2}+1}$$

The output noise psd from the forward amplifier is calculated as:

$$\overline{V}^2_{nout,amp}(\omega) = \overline{i}^2_{n,amp}\frac{V_T^2}{I_{PH}^2}\frac{\frac{1}{\kappa^2}(\frac{1}{(\kappa R_i)^2}+\frac{\omega^2}{\omega_{z^*}^2})}{\left(\frac{\omega^2}{\omega_n^2}\right)^2+\left(\frac{1}{Q^2}-2\right)\frac{\omega^2}{\omega_n^2}+1}, \text{ where } \omega_{z^*} = \frac{1}{\beta\tau_{out}+\frac{\tau_{in}}{\kappa R_i}}$$

It is often stated that the main noise sources in photoreceptors are shot noise from the photodiode and thermal noise from the feedback transistor M_{3s}, and the noise of the amplifier only becomes non-negligible under high illumination [100], [101]. For a complete analysis, flicker noise (1/*f*) is also included. For small-sized transistors, 1/*f* noise is actually believed to become random telegraph signal (RTS) noise with a Lorentzian-shaped psd [98]. Nevertheless, the simple 1/*f* noise model is used here for approximate calculations to gain some insights. The noise current psd from the photodiode, M_{3s}, and the amplifier are written as:

$$\overline{i}_{n,ph}^{2} = 2qI_{PH}\,,\ \overline{i}_{n,M_3}^{2} = 2qI_{PH} + \frac{K_n}{(WL)_3 C_{ox}}\frac{1}{f}g_{m3}^2\,,\ \overline{i}_{n,amp}^{2} = 4qI_{amp} + \frac{K_p}{(WL)_4 C_{ox}}\frac{1}{f}g_{m4}^2 + \frac{K_n}{(WL)_1 C_{ox}}\frac{1}{f}g_{m1}^2$$

Noise from the cascode nFET M_{2s} can be normally neglected, if the output resistance of M_{1s} is large enough and signal frequency is sufficiently low [26]. We define the following transistor size ratios:

$$\frac{(WL)_3}{(WL)_4} = R_{34}\,,\ \frac{(WL)_3}{(WL)_1} = R_{31}$$

The total output noise can be calculated as in Eq. (3.3):

$$\begin{aligned}\overline{V}_{nout,tot}^{2}(\omega) &\approx \frac{1}{\kappa^2}\frac{4qV_T^2}{I_{PH}}\int_0^{BW}\frac{T_1 + T_2\dfrac{\omega^2}{\omega_n^2}}{\left(\dfrac{\omega^2}{\omega_n^2}\right)^2 + \left(\dfrac{1}{Q^2} - 2\right)\dfrac{\omega^2}{\omega_n^2} + 1}d\omega + \frac{2\pi K_n}{(WL)_3 C_{ox}}\int_0^{BW}\frac{1}{\omega}\frac{F_1 + F_2\dfrac{\omega^2}{\omega_n^2}}{\left(\dfrac{\omega^2}{\omega_n^2}\right)^2 + \left(\dfrac{1}{Q^2} - 2\right)\dfrac{\omega^2}{\omega_n^2} + 1}d\omega \\ &\overset{x=\frac{\omega}{\omega_n}}{\Rightarrow} \frac{4kT}{C_{in}}\sqrt{\frac{R_i R_c}{\kappa m}}\int_0^{\frac{BW}{\omega_n}}\frac{T_1 + T_2 x^2}{x^4 + \left(\dfrac{1}{Q^2} - 2\right)x^2 + 1}dx + \frac{2\pi K_n}{(WL)_3 C_{ox}}\int_0^{\frac{BW}{\omega_n}}\frac{1}{x}\frac{F_1 + F_2 x^2}{x^4 + \left(\dfrac{1}{Q^2} - 2\right)x^2 + 1}dx\end{aligned} \tag{3.3}$$

where BW is the pixel bandwidth, and the coefficients T_1, T_2, F_1 and F_2 are defined as:

$$T_1 = 1 + \frac{1}{\kappa^2 R_i}\,,\ T_2 = \frac{\kappa R_c}{m}\,,\ F_1 = 1 + \frac{1}{\kappa^2}\frac{K_p}{K_n}R_{34} + \frac{1}{\kappa^2}R_{31}\,,\ F_2 = (1 + R_i^2\frac{K_p}{K_n}R_{34} + R_i^2\frac{R_c^2}{\beta^2}R_{31})\frac{\kappa\beta^2}{mR_c R_i}$$

3.2.1.2 CGPR

The output DC V_{outDC} of CGPR is $V_{bf}+|V_{gs3}|$. V_{bf} can be set at a reasonably high voltage so that unlike SFPR both M_{1c} and M_{2c} can stay in saturation even under low illumination. The TIA bias current I_{amp} set by pFET M_{4c} should always be larger than the photocurrent I_{ph} with some margin to supply sufficient current to M_{1c} and M_{2c}. The C_{ds} of M_{3c} in CGPR is much smaller than that of M_{3s} in SFPR, and hence no Miller effect needs to be considered which lends to speed advantage albeit with more stability concern. The body of M_{3c} is tied to VDD, sharing the same n-well with M_{4c} to save area.

The small signal equivalent circuit of the CGPR is shown in Figure 3.4(b). The transfer function is calculated as in Eq. (3.4):

$$\frac{V_{out}}{V_T\dfrac{i_{ph}}{I_{PH}}} = \frac{V_{out}}{\dfrac{i_{ph}}{g_{s3}}} = \frac{A_{DC}}{\dfrac{s^2}{\omega_n^2} + \dfrac{s}{\omega_n Q} + 1} = \frac{A(1 - \dfrac{1}{A^{'}R_i})/A + \dfrac{\kappa}{A^{'}}(1 + A)}{s^2\dfrac{A}{A + \dfrac{\kappa}{A^{'}}(1 + A)}\tau_{in}\tau_{out} + s\dfrac{A}{A + \dfrac{\kappa}{A^{'}}(1 + A)}(\dfrac{\tau_{in}}{A}(1 + \dfrac{A}{\kappa R_i} + \dfrac{A}{A^{'}R_i}) + \tau_{out}\dfrac{\kappa}{A^{'}}) + 1} \tag{3.4}$$

where A' is defined as:

$$A' = \frac{\kappa g_{s3}}{g_{ds3}}$$

If A and A' are much larger than 1, A_{DC}, ω_n and Q can be simplified as:

$$A_{DC} = 1,\ \omega_n = \sqrt{\frac{1}{\tau_{in}\tau_{out}}},\ Q = \frac{\sqrt{R_\tau}}{R_\tau\left(\frac{1}{A} + \frac{1}{\kappa R_i}\right) + \frac{\kappa}{A'}} = \frac{\sqrt{R_\tau}}{\frac{R_\tau}{A} + R_c + \frac{\kappa}{A'}}$$

The maximum Q over all illumination conditions is derived as:

$$Q_{max} = \frac{\sqrt{\frac{AA'}{\kappa + A'R_c}}}{2},\ \text{when}\ R_\tau = \frac{A}{A'}(\kappa + A'R_c)$$

If $A'R_c >> \kappa$, the Q_{max} is simplified as Eq. (3.5):

$$Q_{max} = \frac{\sqrt{A/R_c}}{2},\ \text{when}\ R_\tau = AR_c \tag{3.5}$$

To have real poles in this 2nd-order system, R_i has to satisfy:

$$R_i < \frac{2A^2}{\kappa R_c}\left(1 - \sqrt{1 - \frac{R_c}{A}}\right) - \frac{A}{\kappa}\ \text{or}\ R_i < \frac{2A^2}{\kappa R_c}\left(1 + \sqrt{1 - \frac{R_c}{A}}\right) - \frac{A}{\kappa}$$

With practical design parameters such as A=100, κ=0.8, R_c=5, Q_{max} is about 2.24 at R_i=125. To have real poles, the range of R_i is calculated as: $R_i<1.60$ or $R_i>9.75\times10^3$. Under low illumination, e.g. $R_i=10^4$, the 3-dB bandwidth is about:

$$\omega_{3dB} = 126/\tau_{in}$$

Comments on the calculations above are given below.

1. Note that Q_{max} is not only dependent on A but also R_c, as shown in Eq. (3.5). To keep a small Q_{max}, large R_c and small A are needed. Therefore, if a CC-PGA is used, as depicted in Figure 3.2, an SF connected to the output of the photoreceptor is indispensable to avoid large capacitive loading. By tying the gate of M_{2c} to VDD, smaller A could be obtained to keep Q_{max} low, especially under high illumination. Without the Miller effect in contrast to the case of SFPR where C_{gs3} plays the role of Miller capacitance, the 3-dB bandwidth of CGPR is over 9 times larger than that of SFPR at $R_i=10^4$. Under high illumination, the bandwidth is limited by τ_{out} in both cases, and the bandwidth difference becomes negligible.

The noise analysis follows the method used in the SFPR case. The output noise psd contributed by the photodiode is written as:

$$\overline{V}^2_{nout,ph}(\omega) = \overline{i}^2_{n,ph}\frac{V_T^2}{I_{PH}^2}|H(\omega)|^2 = \overline{i}^2_{n,ph}\frac{V_T^2}{I_{ph}^2}\frac{1}{\left(\frac{\omega^2}{\omega_n^2}\right)^2 + \left(\frac{1}{Q^2} - 2\right)\frac{\omega^2}{\omega_n^2} + 1}$$

The output noise psd contributed by M_{3c} and the amplifier can be respectively calculated as:

$$\overline{V}^2_{nout,M_3}(\omega) = \overline{i}^2_{n,M_3} \frac{V_T^2}{I_{PH}^2} \frac{1+\frac{\omega^2}{\omega_z^2}}{(\frac{\omega^2}{\omega_n^2})^2 + \left(\frac{1}{Q^2}-2\right)\frac{\omega^2}{\omega_n^2}+1}, \text{ where } \omega_z = A\frac{g_{dsout}}{C_{in}}$$

$$\overline{V}^2_{nout,amp}(\omega) = \overline{i}^2_{n,amp} \frac{V_T^2}{I_{ph}^2} \frac{\frac{1}{\left(A^{'}R_i\right)^2}+\frac{\omega^2}{\omega_z^2}}{\left(\frac{\omega^2}{\omega_n^2}\right)^2 + \left(\frac{1}{Q^2}-2\right)\frac{\omega^2}{\omega_n^2}+1}$$

The noise current psd from the photodiode, M$_{3c}$, and the amplifier can be written as:

$$\overline{i}^2_{n,ph} = 2qI_{PH}\,,\ \overline{i}^2_{n,M_3} = 2qI_{PH} + \frac{K_p}{(WL)_3 C_{ox}}\frac{1}{f}g_{m3}^2\,,\ \overline{i}^2_{n,amp} = 2q\left(2I_{amp}-I_{ph}\right) + \frac{K_p}{(WL)_4 C_{ox}}\frac{1}{f}g_{m4}^2 + \frac{K_n}{(WL)_1 C_{ox}}\frac{1}{f}g_{m1}^2$$

The total output noise can be calculated as in Eq. (3.6):

$$\begin{aligned}\overline{V}^2_{nout,tot}(\omega) &\approx \frac{4qV_T^2}{I_{ph}}\int_0^{BW}\frac{T_1+T_2\frac{\omega^2}{\omega_n^2}}{\left(\frac{\omega^2}{\omega_n^2}\right)^2+\left(\frac{1}{Q^2}-2\right)\frac{\omega^2}{\omega_n^2}+1}d\omega + \frac{2\pi\kappa^2K_p}{(WL)_3C_{ox}}\int_0^{BW}\frac{1}{\omega}\frac{F_1+F_2\frac{\omega^2}{\omega_n^2}}{\left(\frac{\omega^2}{\omega_n^2}\right)^2+\left(\frac{1}{Q^2}-2\right)\frac{\omega^2}{\omega_n^2}+1}d\omega \\ &\overset{x=\frac{\omega}{\omega_n}}{\Rightarrow} \frac{4kT}{C_{in}}\sqrt{\kappa R_i R_c}\int_0^{\frac{BW}{\omega_n}}\frac{T_1+T_2x^2}{x^4+\left(\frac{1}{Q^2}-2\right)x^2+1}dx + \frac{2\pi\kappa^2K_p}{(WL)_3C_{ox}}\int_0^{\frac{BW}{\omega_n}}\frac{1}{x}\frac{F_1+F_2x^2}{x^4+\left(\frac{1}{Q^2}-2\right)x^2+1}dx\end{aligned} \tag{3.6}$$

where BW is the bandwidth of the pixel analog front end, and the coefficients T_1, T_2, F_1 and F_2 are defined as:

$$T_1 = 1\,,\ T_2 = \frac{R_c}{\kappa}\,,\ F_1 = 1\,,\ F_2 = (1+R_i^2R_{34}+(R_i-1)^2\frac{K_n}{K_p}R_{31})\frac{R_c}{\kappa R_i}$$

3.2.1.3 Noise Comparison

The bandwidth improvement of CGPR due to the lack of C$_{gs}$ Miller capacitance has been exploited to improve the minimum latency of DVSs [15], even though this is at the cost of increased Q_{max}. In the context of designing a DVS with enhanced TC sensitivity, noise is what matters more because high noise level at the photoreceptor output can mask any transient visual signal with amplitude below the noise rms value. This is true for a single pixel. For an array of pixels, however, the visual detection of an object also depends on the percentage of its 2D geometrical size in the whole visual field. As pointed out in [102], statistically the minimum number of photons required to detect a spatial contrast in an image is inversely proportional to the object size, which also applies to the DVS case. In this regard, even though the TC signal may be below the noise rms, an object whose size is sufficiently large could still be detected. Because object size is unpredictable, in the following analysis, we still apply the most stringent criterion that the photoreceptor output noise should be below the aimed TC sensitivity.

The only difference between the two photoreceptors is the feedback transistor, an nFET in SFPR and a pFET in CGPR respectively. CGPR is expected to have a lower output noise because pFETs are often observed to have a lower $1/f$ noise compared to nFETs. As shown by the theoretically calculated results

Table 3.1. Comparison of the output voltages of SFPR and CGPR with different i_{ph} change.

	SFPR	CGPR	Notes
normalized DC gain $\boldsymbol{V_{out}/(i_{ph}/I_{PH})}$	V_T / κ	V_T	V_T: thermal voltage, 25.9 mV; i_{ph}: small signal photocurrent; I_{PH}: DC background photocurrent; κ: subthreshold slope factor, 0.8.
$\boldsymbol{\Delta V_{out}}$ (mV) at 10% $\boldsymbol{i_{ph}}$ change	3.05	2.59	
$\boldsymbol{\Delta V_{out}}$ (μV) at 1% $\boldsymbol{i_{ph}}$ change	305	259	

Table 3.2. Sizes of the transistors in Fig. 2.

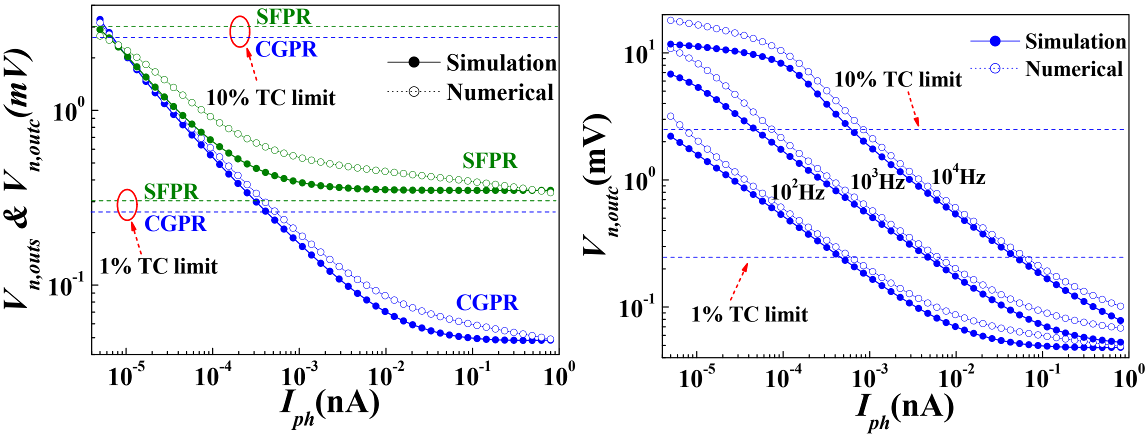
Device	Width(μm)/Length(μm)
M_{1s}/M_{1c}	0.24/2.00
M_{2s}/M_{2c}	0.24/0.18
M_{3s}/M_{3c}	0.60/0.20
M_{4s}/M_{4c}	0.24/1.00

listed in Table 3.1, for 1% TC detection by a single pixel, the rms integrated output noise should be less than 305 μV and 259 μV for SFPR and CGPR, respectively. For Spectre simulation, resistors are used to emulate the shot noise of a photodiode [103]. With the transistor sizes given in Table 3.2, the simulated output noise for both SFPR $V_{n,outs}$ and CGPR $V_{n,outc}$ within 100 Hz bandwidth are plotted in Figure 3.5(a) along with the numerically calculated results according to Eq. (3.3) and Eq. (3.6) with the parameters provided by UMC 0.18 μm RF/MM CMOS. M_{1s}/M_{1c} is a 3.3-V transistor in simulation to avoid the situation where V_{out} is too low for M_{2s} in SFPR to stay in saturation at low I_{PH}. For small I_{PH}<100 fA, $V_{n,outs}$ and $V_{n,outc}$ are approximately the same because the photocurrent shot noise dominates. As I_{PH} increases, the shot noise contribution decreases, and transistor $1/f$ noise becomes prominent. $V_{n,outs}$ saturates to 350 μV at $I_{PH}\approx 1$ pA, still above its 1% TC level. In comparison, $V_{n,outc}$ continues to decreases until $I_{PH}\approx 50$ pA, and saturates at about 50 μV, 7 times lower compared to the SFPR. Figure 3.5(b) shows the simulated $V_{n,outc}$ as a function of both I_{PH} and pixel front-end bandwidth using a 1.8-V M_{1c}. Larger bandwidth results in increased $V_{n,outc}$ under the same illumination. The plateau of the 10^4-Hz curve is due to the fact that the CGPR bandwidth under low illumination is already less than 10^4 Hz, and thus the integrated noise remains relatively constant. The pixel front-end bandwidth can be controlled via the CC-PGA as will be discussed in the next section.

From the comparison of the output noise between SFPR and CGPR, the latter is chosen in the design for its better noise performance under moderate to high illumination which ensures the goal of 1% TC sensitivity. Note that a 1.8-V M_{1c} is used in fabrication for CGPR.

3.2.2 Capacitively-Coupled Programmable Gain Amplifier (CC-PGA)

CC-PGA is chosen for the additional in-pixel gain to amplify small TC signals. The previous solutions include using above-threshold common-source amplifiers (Figure 3.6(a) [15]) and subthreshold transimpedance amplifiers (Figure 3.6(b) [16]). The common drawback in these two approaches is that they both need a global feedback mechanism to control the bias of the preamplifiers I_{preamp} in order to keep the transistors in saturation because I_{preamp} is dependent on the local photocurrent. In Figure 3.6(a), it is done by tuning the source voltage of the nFET, and in Figure 3.6(b), it is done by tuning the gate voltage of the pFET. This global feedback limits the intrascene dynamic range (to about 60 dB with a 3.3-V

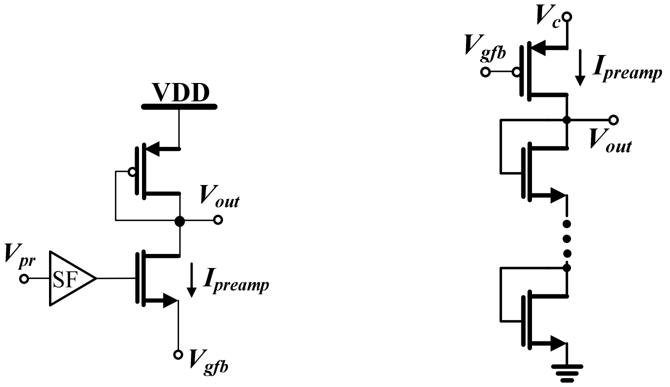


(a) (b)

Figure 3.5. Spectre simulation and theoretical calculation plots of (a) the output noise of both SFPR and CGPR bandlimited within 100 Hz, and (b) the output noise of CGPR at different bandwidths.

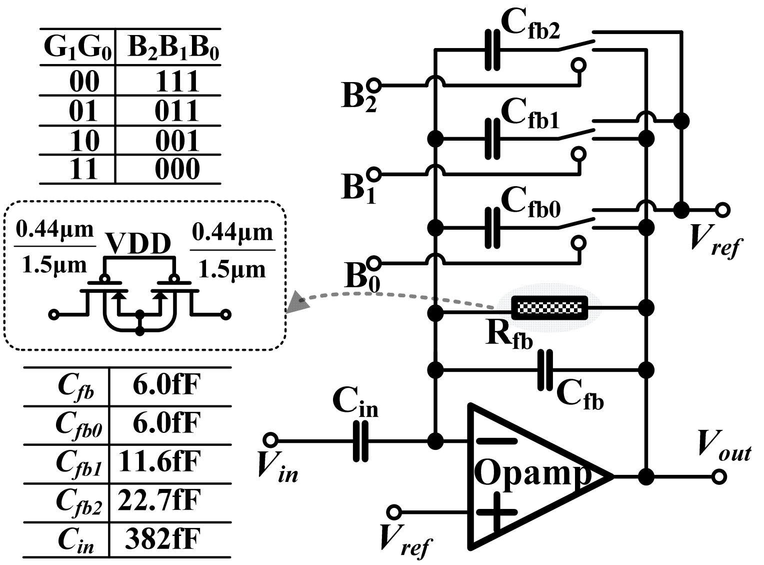
supply in 0.35 µm CMOS) because of possible large I_{preamp} span across the whole pixel array. In addition, the obvious penalty of using the circuit in Figure 3.6(a) is large power consumption for above-threshold operation to obtain proper gain via transistor sizing, and in Figure 3.6(b) is the compromised output swing because of the stacking of diode-connected nFETs. To overcome the drawbacks in the previous designs, a compact in-pixel CC-PGA is used, and the design requirements are low power, large output swing and independence of bias current on local photocurrent. The proposed circuit is shown in Figure 3.7.

The closed-loop gain is defined by the capacitance ratio $A_{CC\text{-}PGA}$=-C_{in}/C_{fbtot}, where C_{fbtot} is the total feedback capacitance. 2 digits G_1G_0 are used for 4-level programmable $A_{CC\text{-}PGA}$ control via combinational logic, from 18 dB (G_1G_0=00) to 36 dB (G_1G_0=11) with a 6 dB step. The capacitance values are given in Figure 3.7. Note that when the capacitors C_{fbi} are not connected in the feedback loop, their right plates are connected to V_{ref}. This is adapted from [104] to prevent frequency response distortion at low frequencies where the off-state switch resistance can no longer be assumed to be infinitely large, especially in deep submicron technology. A simplified circuit shown in Figure 3.8(a) is used to explain the effect. Let R_x denote the finite off-state switch resistance, and assume infinite open-loop gain of the Opamp. The



(a) (b)

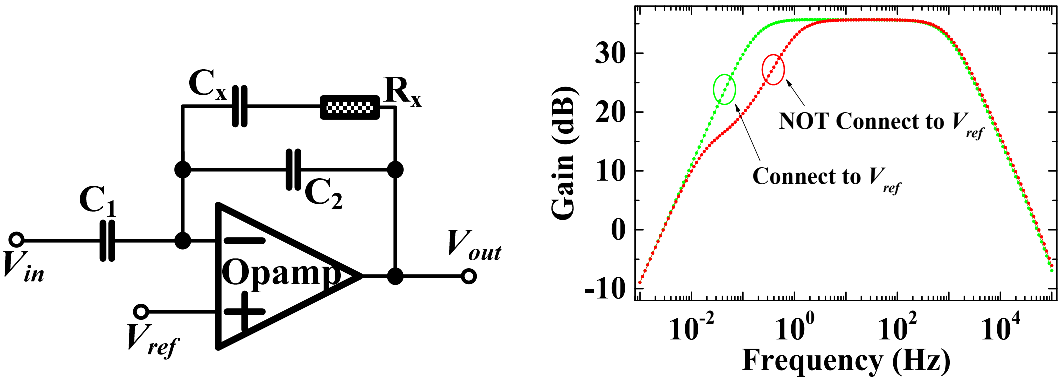
Figure 3.6. Previous circuit solutions to high in-pixel gain (a) above-threshold common-source amplifier [15] and (b) subthreshold transimpedance amplifier [16].



G_1G_0	$B_2B_1B_0$
00	111
01	011
10	001
11	000

C_{fb}	6.0fF
C_{fb0}	6.0fF
C_{fb1}	11.6fF
C_{fb2}	22.7fF
C_{in}	382fF

Figure 3.7. CC-PGA circuit diagram with the pseudo-resistor transistor implementation. The digital control bits and capacitor values are also listed.



(a) (b)

Figure 3.8. (a) Simplified circuit for explaining the effect of finite off-state switch resistance; (b) Simulated results using Spectre at G_1G_0=11 with the right plates of C_{fbi} connected (green line) or not connected (red line) to V_{ref} when not used in the feedback.

closed-loop gain is calculated to be:

$$\frac{V_{out}}{V_{in}} = -\frac{C_1(s+\frac{1}{R_xC_x})}{C_2(s+\frac{1}{R_x}(\frac{1}{C_x}+\frac{1}{C_2}))}$$

The finite R_x creates an additional pair of pole and zero. If the pole frequency is larger than the highpass corner frequency f_{hc} determined by C_{fbtot} and R_{fb} in Figure 3.7, $A_{CC\text{-}PGA}$ starts to degrade at a frequency higher than f_{hc}. By connecting the right plates of C_{fbi} to V_{ref}, C_{fbi} is completely avoided in the feedback loop, and therefore the parasitic pole/zero pair is avoided. Figure 3.8 shows the simulated frequency response in both cases: the red curve represents C_{fbi} not connected to V_{ref}, and the green curve represents C_{fbi}

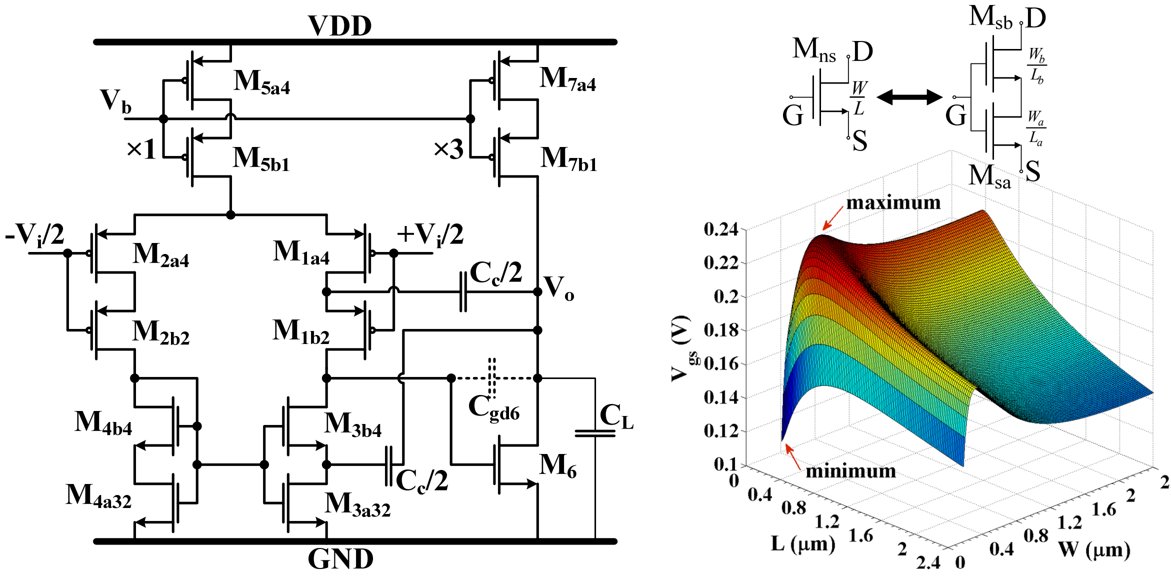
connected to V_{ref}. The highpass corner frequency is shifted from 942 mHz to 165 mHz, about 6 times improved. The sufficiently low f_{hc}=165 mHz for not filtering out low-frequency visual signals is ensured by using the pseudo-resistor R_{fb} as depicted in Figure 3.7. It is composed of two off-state pFETs with the bulk connected to source [105]. The only conducting channel is formed by the back-to-back drain-bulk PN junctions whose leakage current provides the DC feedback.

The main challenge lies in the Opamp design. For the CC-PGA, the output swing needs to be maximized for wide input dynamic range, because unlike ADM it does not have any reset mechanism to pull its output towards the resetting level VDD/2 and therefore can be saturated by large input transient. This precludes the use of simple 5T one-stage amplifiers or the telescopic topology. Folded cascode with moderate output swing is not considered either for its four bias voltages needed which complicates the pixel array wiring. Another reason for not using one-stage amplifiers is the insufficient open-loop gain. Assume the open-loop gain of the Opamp in Figure 3.8(a) is A_{open}, and ignore C_x and R_x. The actual closed-loop gain A'_{CC-PGA} is calculated as:

$$A'_{CC-PGA} = \frac{V_{out}}{V_{in}} = -\frac{C_1}{C_2}\frac{1}{\dfrac{\dfrac{C_1}{C_2}+1}{A_{open}}+1} = -A_{CC-PGA}\frac{1}{\dfrac{A_{CC-PGA}+1}{A_{open}}+1}$$

At G_1G_0=36 dB, i.e. the desired A_{CC-PGA}=64, with $A_{open}=10^2$, A'_{CC-PGA} is only 38.8; with $A_{open}=10^3$, A'_{CC-PGA} is about 60.1; with $A_{open}=10^4$, A'_{CC-PGA} is about 63.6. Large A_{open} also helps desensitize the variation of A'_{CC-PGA} to the variation of A_{open}. For example, given a ±10% variation at $A_{open}=10^3$, A'_{CC-PGA} varies about 1.2%; give a ±40% variation at $A_{open}=10^4$, A'_{CC-PGA} varies about 0.6%. Even though the variation at $A_{open}=10^4$ is about 4 times larger than that at $A_{open}=10^3$, the variation of A'_{CC-PGA} is only half. It is difficult to achieve a design goal of $A_{open}=10^4$ in a single-stage amplifier in 0.18 μm CMOS. The next reasonable option is the symmetrical amplifier which is virtually a single-stage amplifier but with additional gain that can be obtained from its inherent current-mirror structure. Although it offers a large output swing, this topology has twice more noise than a single stage amplifier [106], not mention that it is area- and power-consuming if large gain is to be obtained from the transistor ratio of the current mirrors. Now the best candidate seems to be the Miller-compensated two-stage amplifier. It has a large output swing, almost the same noise as a single-stage amplifier thanks to the large gain of the first stage, and needs only 2 more transistors. However, as suggested by its name, it needs a compensation capacitor connected between the outputs of the first and second stages, which is usually fine for an Opamp with a given bandwidth. In DVS, on the other hand, the Opamp bandwidth is required to be widely tunable by changing the bias to define the pixel front-end bandwidth, which can result in insufficient phase margin of the Opamp because the right-half plane zero caused by Miller compensation can only be well cancelled at a particular bias current using a nulling resistor. The nulling resistor would have unacceptably large size given small bias current of the Opamp. It is possible to use a pFET operating in linear region as the nulling resistor whose resistance tracks the bias of the Opamp [26]; the cost is obviously the extra area for the resistance-tracking circuits with at least 4 additional transistors.

The schematic of the Opamp used in the design is shown in Figure 3.9(a). The "Miller-like" cascode compensation is based on the technique proposed in [107] back in 1984. The compensation capacitor is connected between the sources of the cascode transistors (M_{1b2} and M_{3b4}) in the first stage and the output of the second stage. This configuration helps push the right-half plane zero to a much higher frequency compared to the original Miller compensation, and thus no nulling resistor is needed. Nevertheless, the



(a) (b)

Figure 3.9. (a) Schematic of the Opamp used in the CC-PGA with pseudo-cascode compensation; (b) transistor channel length splitting, and the sweep of the V_{gs} of an nFET as the function of its L and W.

value of the compensation capacitance C_c should be carefully chosen to avoid insufficient gain margin caused by gain peaking beyond the GBW as pointed out in [107]. Note that the cascode transistors M_{xby} (x is the transistor index, y is the number of unit transistors connected in parallel) is not an ordinary cascode transistor, because its gate is not connected to an external fixed bias voltage but the gate of the common-source transistors M_{xaz} (z is the number of unit transistors connected in series). The compensation in this Opamp topology is called pseudo-cascode compensation [108]. Compared to the traditional cascode compensation, besides the benefits such as lower VDD, and higher unit-gain frequency [108], pseudo-cascode compensation can save two extra biases and thus simplify the wiring in a pixel array. Usually M_{xaz} is in deep triode region due to its small V_{ds}, and yet still relax the value of C_c to a large extent as demonstrated in [109]. To increase the open-loop gain, body-biasing of M_{xby} was used [108] to lower its V_{th} so that M_{xaz} can move towards the edge of saturation. Not only is this impossible for using area-consuming deep n-well in pixels, but also two extra wires for biasing voltages are needed. To circumvent the aforementioned problems, two small-size effects of MOSFETs are exploited in a split-transistor technique in subthreshold [110].

The schematic in Figure 3.9(b) shows the concept of the split-transistor technique. nFET M_{ns} is divided into two transistors M_{sa} and M_{sb} with equivalent channel length $L=L_a+L_b$. M_{sa} and M_{sb} form the pseudo-cascode structure. In subthreshold, as long as the V_{ds} of M_{sa} is larger than 100 mV, it is considered to be in saturation and the output impedance is large. In contrast, to stay in saturation in above threshold, V_{ds} usually needs to be larger than the overdrive voltage V_{ov}>200 mV. Sizing M_{sa} with a very small W/L ratio and M_{sb} with a very large W/L ratio can also push M_{sa} towards saturation, but it is as impractical as the body-biasing means with a restricted pixel area. Note that the transistor threshold voltage V_{th} varies with L and W. The well-known short-channel effect (SCE) [111] indicates that in sub-micrometer processes V_{th} decreases with L due to effects like the drain-induced barrier lowering (DIBL). To reduce the resulting increased leakage current at small L, the so-called halo implantation was devised to mitigate channel depletion, and hence V_{th} may decrease slower or even increases within some range of L. This phenomenon is termed as reverse SCE (RSCE) [111]. Modern processes normally use shallow-trench-isolation (STI) instead of local oxidation of silicon (LOCOS) for device isolation to allow a denser integration. In STI, V_{th} decreases with decreasing W due to an increased electrostatic potential near the channel edge. The phenomenon is termed as inverse narrow-width effect (INWE), in contrast to NWE observed in LOCOS [112]. To verify these effects by simulation in UMC 0.18 μm CMOS, the V_{gs} of an nFET is plotted over its L and W at 1 nA bias current as shown in Figure 3.9(b). V_{ds} is fixed at 0.9 V because V_{th} is also dependent on V_{ds} [111]. It can be seen that V_{gs} does not monotonically decrease as the W/L ratio increases. At small L and W, the small size effects play an important role: With fixed W (L), V_{gs} first increases with decreasing L (W) and then drops down rapidly after a certain point, indicating RSCE (INWE). The V_{gs} sweep of PMOS shows the RSCE and the INWE as well, but they are less evident. If the sizes of M_{sa} and M_{sb} are chosen according to Figure 3.9(b) such that the V_{gs} of M_{sa} is at least 100 mV larger than that of M_{sb}, then M_{sa} has its V_{ds}>100 mV and is considered to be in saturation. The size of M_{sa} and M_{sb} in this design are $(W/L)_{sa}$=0.48/0.36 and $(W/L)_{sb}$=0.24/0.18 in μm corresponding to the maximum (116 mV) and minimum (235 mV) V_{gs} in Figure 3.9(b). The size of M_{ns} is $(W/L)_{ns}$=0.40/0.54 in μm, which has the same area as the sum of M_{sa} and M_{sb}. Theoretical analysis of the gain ratio of the split transistor A_s over the non-split transistor A_{ns} is given as follows.

The ratio A_s/A_{ns} can be written as:

$$\frac{A_s}{A_{ns}} = \frac{g_{msa} r_{osa} g_{msb} r_{osb}}{g_{mns} r_{ons}}$$

where g_{mi}'s and r_{oi}'s are the transconductance and output resistance of transistor M_i (i=sa, sb, ns) respectively. In subthreshold, the transconductance is $\kappa I/V_T$, but the output conductance can have different formulizations depending on the physical mechanisms considered. Let us first derive the output resistance with DIBL considered r_{oDIBL}. The I_{ds} in subthreshold can be written as:

$$I_{ds} = \frac{W}{L} I_0 e^{\kappa(V_{gs} - V_{th0} + \Delta V_{thDIBL})/V_T} (1 - e^{-V_{ds}/V_T}) \tag{3.7}$$

where I_0 is a lumped parameter determined by unit gate capacitance, carrier mobility and subthreshold slope factor, V_{th0} is the intrinsic transistor threshold voltage, and ΔV_{thDIBL} is the threshold voltage shift due to DIBL. If in saturation, I_{ds} can be simplified as:

$$I_{ds} = \frac{W}{L} I_0 e^{\kappa(V_{gs} - V_{th0} + \Delta V_{thDIBL})/V_T} \tag{3.8}$$

A typical expression of ΔV_{thDIBL} is [98]:

$$\Delta V_{thDIBL} = [3(\phi_{bi} - \phi_0) + V_{ds}] e^{-L/l_0}$$

where ϕ_{bi} is the drain/source-bulk pn junction built-in potential, ϕ_0 is the surface potential of two-terminal MOS structure in strong inversion, and l_0 is the characteristic length or scale length [98], [113]. r_{oDIBL} is calculated as:

$$r_{oDIBL} = 1 / \frac{\partial I_{ds}}{\partial V_{ds}} = \frac{V_T}{\kappa I_{ds}} e^{\frac{L}{l_0}} \tag{3.9}$$

The scale length l_0 can be written as:

$$l_0 = 2\sqrt{\frac{\varepsilon_{\mathrm{si}} t_{\mathrm{ox}}}{\varepsilon_{\mathrm{ox}}} W_{dm}}$$

where W_{dm} is the maximum depletion depth under gate, ε_{si} and ε_{ox} are the permittivity of silicon and silicon dioxide respectively, and t_{ox} is the thickness of the gate dioxide. For a uniformly doped channel, W_{dm} can be calculated as:

$$W_{dm} = \sqrt{\frac{2\varepsilon_{si}(2V_T \ln\frac{N_a}{n_i} - V_{bs})}{qN_a}}$$

where N_a is the channel doping concentration and n_i is the intrinsic carrier concentration of silicon. For V_{bs}=-0.1 V, W_{dm} is about 40.63 nm, and l_0 is about 45.22 nm. If the r_{oi} of all the nFETs follow the exponential dependence on L as in Eq. (3.9), A_s/A_{ns} is calculated as:

$$\frac{A_s}{A_{ns}} = \frac{\kappa I_{ds}}{V_T} \frac{\frac{V_T}{I_{ds}} e^{\frac{L_{sa}}{l_0}} \frac{V_T}{\kappa I_{ds}} e^{\frac{L_{sb}}{l_0}}}{\frac{V_T}{I_{ds}} e^{\frac{L_{ns}}{l_0}}} = 1 = 0 \text{ dB}$$

As we will learn later, the resulting ratio 0 dB is not consistent with the simulation and testing results where A_s/A_{ns} often shows to be much larger than 1. This is probably due to the fact that the halo implantation suppresses the DIBL and modifies the dependence of r_o on L in light of the drain-induced-threshold-shift (DITS) effect [114]. Although the DITS is lately attributed to the difference between the source and drain barriers [115], we still use the widely accepted phenomenological long-channel DIBL model provided in BSIM where the V_{th} shift is written as [116]:

$$\Delta V_{thDITS} = \frac{V_T}{\kappa} \ln[\frac{L}{L + DVTP0 \cdot (1 + e^{-DVTP1 \cdot V_{ds}})}]$$

where $DVTP0$ and $DVTP1$ are the DITS coefficients. $DVTP0$ is only related to process parameters and $DVTP1$ depends on L via the equation below [114]:

$$DVTP1 = \frac{e^{-L_p/l_0} + 2e^{-2L_p/l_0}}{V_T}$$

where L_p is the pocket length of halo doping that can be estimated with L and the channel and halo doping concentration N_{ch} and N_p by [116]:

$$L_p = \frac{N_{ch}}{N_p - N_{ch}} \frac{Nlx + lNlx / L}{2}$$

where Nlx is the lateral non-uniform doping parameter, and $lNlx$ is its corresponding binning parameter.

Similar to Eq. (3.9), the output resistance considering DITS can be derived as:

$$r_{oDITS} = 1 / \frac{\partial I_{ds}}{\partial V_{ds}} = \frac{\kappa L}{I_{ds}} \frac{e^{DVTP1 \cdot V_{ds}}}{DVTP0 \cdot DVTP1} = \frac{V_E}{I_{ds}} \tag{3.10}$$

where V_E can be seen as the Early voltage and is proportional to L. Because M_{sb} is clearly affected by the DIBL with its V_{gs} at the nadir of the sweep in Figure 3.9(b), its r_o still uses Eq. (3.9), and Eq. (3.10) applies to M_{sa} and M_{ns}. Note that the L in the equations above should be in fact the effective length L_{eff} which can be calculated from the nominal drawn length L_{drawn} as:

$$L_{eff} = L_{drawn} - 2dL$$

where dL accounts for the depletion length in channel due to heavily-doped drain/source. Herein, A_s/A_{ns} can be again calculated as:

$$\frac{A_s}{A_{ns}} = \frac{\kappa I_{ds}}{V_T} \frac{\frac{\kappa L_{effsa}}{I_{ds}} \frac{e^{DVTP1_{sa} \cdot V_{dssa}}}{DVTP0_{sa} \cdot DVTP1_{sa}} \frac{V_T}{\kappa I_{ds}} e^{\frac{L_{effsb}}{l_0}}}{\frac{\kappa L_{effns}}{I_{ds}} \frac{e^{DVTP1_{ns} \cdot V_{dsns}}}{DVTP0_{ns} \cdot DVTP1_{ns}}} = \frac{L_{effsa} DVTP1_{ns}}{L_{effns} DVTP1_{sa}} e^{DVTP1_{sa} \cdot V_{dssa} - DVTP1_{ns} \cdot V_{dsns}} e^{\frac{L_{effsb}}{l_0}} = 3.96 = 12\mathrm{dB} \tag{3.11}$$

M_{sa} could be in linear region with its V_{ds}<100 mV due to variation of V_{th} in both M_{sa} and M_{sb}. In this case, using Eq. (3.7), the output resistance of M_{sa} is derived as:

$$r_{osa} = \frac{V_T}{I_{ds}} e^{\frac{V_{tha} - V_{thb}}{V_T}}$$

With r_{osb} using Eq. (3.9) and r_{ons} using Eq. (3.10), A_s/A_{ns} is calculated as:

$$\frac{A_s}{A_{ns}} = \frac{\kappa I_{ds}}{V_T} \frac{\frac{V_T}{I_{ds}} e^{\frac{V_{tha} - V_{thb}}{V_T}} \frac{V_T}{\kappa I_{ds}} e^{\frac{L_{effsb}}{l_0}}}{\frac{V_E}{I_{ds}}} = \frac{V_T}{V_E} e^{\frac{V_{tha} - V_{thb}}{V_T}} e^{\frac{L_{effsb}}{l_0}} \tag{3.12}$$

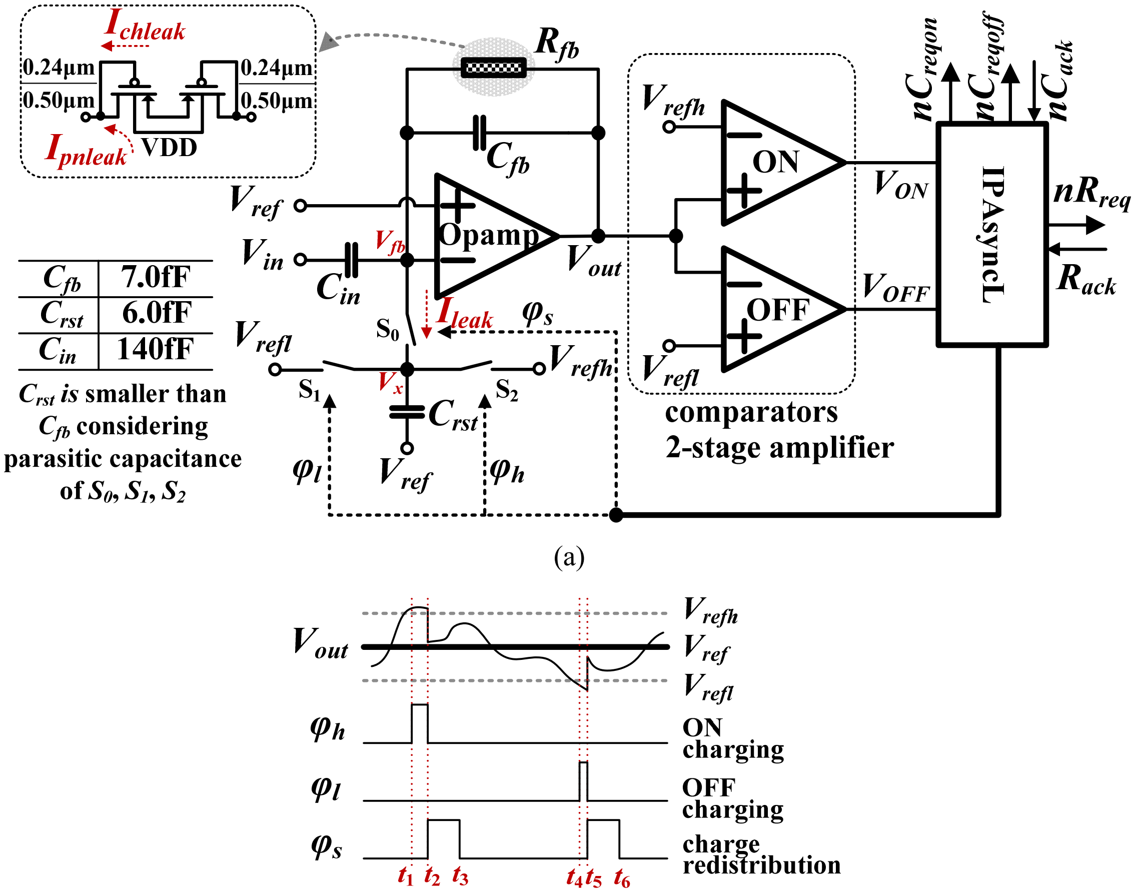
V_E is simulated to be about 1.76 V for M_{ns} biased at 10 nA. From the equation above, it is possible to have a gain ratio less than 1 if V_{tha}-V_{thb}<31.9 mV, which can happen due to process variation. For more robust gain increase, more M_{sa}′s are connected in series and more M_{sb}′s are connected in parallel, as in the case of the Opamp design in Figure 3.9(a).

The noise of the CC-PGA has a direct tradeoff with its closed-loop gain A_c. Considering only thermal noise, the input-referred noise can be approximated by [117]:

$$V_{n,ir} = \sqrt{\frac{8kT}{3C_c A_c}}$$

With C_c=32.8 fF, $V_{n,ir}$=73 μV and 202 μV when A_c=36 dB and 18 dB respectively. If a 50 μV noise is contributed by the CGPR, a 6 dB SNR can still be obtained for a single pixel at A_c=36 dB with a 1% TC signal. Limiting the bandwidth by either lowering the bias current of the CC-PGA or using larger C_c could further improve SNR, but they are limited by voltage headroom and pixel area, respectively. To minimize the 1/*f* noise contribution from the Opamp, the nFETs in the first stage are designed to occupy the most area of the CC-PGA.

Instantaneous signal dynamic range is limited by the output range of the CC-PGA. Let us assume that the output range is 0.2 V-1.6 V without severe output signal distortion or clipping. At A_c=18 dB, the input range of the CC-PGA is about 170 mV which allows 57 dB instantaneous i_{ph} change. At A_c=36 dB, the input range is about 22 mV, which allows only 7.4 dB instantaneous i_{ph} change. In terms of signal integrity, higher gain is at the cost of smaller allowable dynamic range of a dynamic visual scene. However, in



(a)

(b)

Figure 3.10. (a) Schematic of the ADM with illustration of the pseudo-resistor; (b) Timing diagram of the switch control signals φ_s, φ_l and φ_h, and the V_{out} waveform.

the long run, the CC-PGA output will eventually adapt to its DC level of 0.9 V through R_{fb}, and is independent of I_{PH}, therefore the intrascene dynamic range is only limited by the CGPR, in contrast to [16] where it is limited by the transimpedance preamplifier to about only 60 dB with a 3.3 V supply.

3.2.3 Asynchronous Delta Modulator (ADM)

Area constraint is the main challenge for the design of an in-pixel ADM. Previous implementations of ADMs as level-crossing ADCs often require a large area of resistive or capacitive feedback DACs [61], [65] which determines the resolution of the ADCs. The 1-bit capacitive DAC used in [118] can potentially reduce the area, but the switching of the control signals is complicated and because of the capacitive division, the input signal is attenuated after the DAC. With 36 dB gain from the CC-PGA, the 1% TC signal is amplified to about 16 mV which is still too small for comparison considering the output DC variation of the ADM amplifier and the input offset of the comparators. Additional gain from the ADM is hence necessary.

The idea of realizing the δ subtraction feedback in ADM was first suggested in [70], but instead of using a full DAC for feedback, a novel asynchronous δ-subtraction switched-capacitor circuit is proposed here for a more area-efficient implementation. The circuit diagram is shown in Figure 3.10(a). The closed-loop gain of the ADM is C_{in}/C_{fb}=24 dB, and the Opamp that has a similar topology as in Figure 3.9(a) is optimized for slew rate instead of noise. The amplified 1% TC after the ADM is about 260 mV, and is detectable by comparators as long as V_{refh}<1.16 V and V_{refl}>0.64 V with V_{ref}=0.9 V. The control signals φ_x (x=s,l,h) of switches S_i (i=0,1,2) are generated by the in-pixel asynchronous logic (IPAsyncL). The switching sequence for ON spikes is described as follows: (1) When V_{out} exceeds V_{refh}, φ_h becomes high, and S_2 is connected. The top plate of C_{rst} is charged to V_{refh}. The row request nR_{req} (active low) is communicated to the periphery AER; (2) after a four-phase AER handshake to transmit the row and column addresses of this pixel, the column acknowledge nC_{ack} (active low) is activated leading to φ_h switching low and φ_s switching high, and therefore disconnecting S_2 and connecting S_0. V_{out} is reset towards V_{ref}; (3) after a certain reset time controlled by the IPAsyncL, φ_s switches low and S_0 is disconnected. The ADM is ready for the next communication cycle. For OFF spikes, the sequence is similar except that the top plate of C_{rst} is charged to V_{refl} first by connecting S_1 via a high φ_l. The switching sequence along with the V_{out} waveform is illustrated in Figure 3.10(b). Comparing with the ADM model given in Chapter 2, the integration is done by ON/OFF charging on C_{rst} via S_2/S_1 connected to V_{refh}/V_{refl}, and the subtraction is done by connecting S_0 so that the charge on C_{rst} is redistributed. Mathematically, the subtraction process can be explained as follows. For an ON spike, at time t_2, the voltage at the top plate of C_{rst} V_x is charged to V_{refh}, and before the end of t_3, V_x is shorted to V_{fb} which is approximately the same as V_{ref} due to virtual ground. In light of charge conservation, the total charge on C_{in}, C_{fb} and C_{rst} at t_2 equals to that at t_3:

$$\begin{aligned}&(V_{in}(t_2)-V_{fb}(t_2))\times C_{in}+(V_{out}(t_2)-V_{fb}(t_2))\times C_{fb}+(V_{ref}(t_2)-V_x(t_2))\times C_{rst}\\&=(V_{in}(t_3)-V_{fb}(t_3))\times C_{in}+(V_{out}(t_3)-V_{fb}(t_3))\times C_{fb}+(V_{ref}(t_3)-V_x(t_3))\times C_{rst}\end{aligned}$$

The equalities below hold:

$$V_{fb}(t_2)=V_{fb}(t_3),\quad V_{ref}(t_2)=V_{ref}(t_3),\quad V_x(t_2)=V_x(t_3)+\delta$$

where $\delta=V_{refh}-V_{ref}=V_{ref}-V_{refl}$. The charge conservation equation can then be simplified as:

$$V_{in}(t_2)C_{in}+V_{out}(t_2)C_{fb}-\delta C_{rst}=V_{in}(t_3)C_{in}+V_{out}(t_3)C_{fb}$$

The output change $\Delta V_{out}=V_{out}(t_3)-V_{out}(t_2)$ can be written as:

$$\Delta V_{out}=-\Delta V_{in}\frac{C_{in}}{C_{fb}}-\delta\frac{C_{rst}}{C_{fb}} \tag{3.13}$$

where $\Delta V_{in}=V_{in}(t_3)-V_{in}(t_2)$. It is clear from Eq. (3.13) that the subtraction does not interfere with the input change amplification, in contrast to STR where the input is blocked from the output during reset.

S_1 and S_2 are connected to low impedance voltage sources V_{refl} and V_{refh} respectively, and hence clock feedthrough usually has a more detrimental effect than charge injection on the precision of the sampled voltage at V_x. If both S_1 and S_2 use nFETs or pFETs, after being disconnected, V_x would be $V_{refl}+\Delta V_x$ for OFF spikes and $V_{refh}+\Delta V_x$ for ON spikes. ΔV_x has a negative/positive sign if nFETs/pFETs are used. The nominal subtracted voltages for OFF and ON spikes now become $-|\delta-\Delta V_x|$ and $\delta+\Delta V_x$, respectively. Because the reset capacitance C_{rst} is quite small compared to the parasitic C_{gs}/C_{gd}, ΔV_x could be as large as tens of mV. To reduce the subtraction imbalance, a pFET is used for S_1, and an nFET for S_2. The nFET and pFET are sized so that the ΔV_x's are approximately the same. A complementary transmission gate is used for S_0 to minimize the switching effects. Note that the reset capacitance is in effect C_{rst} plus the parasitic C_{db}/C_{sb} of all the switches connected to C_{rst}. Therefore even though the clock feedthrough makes the

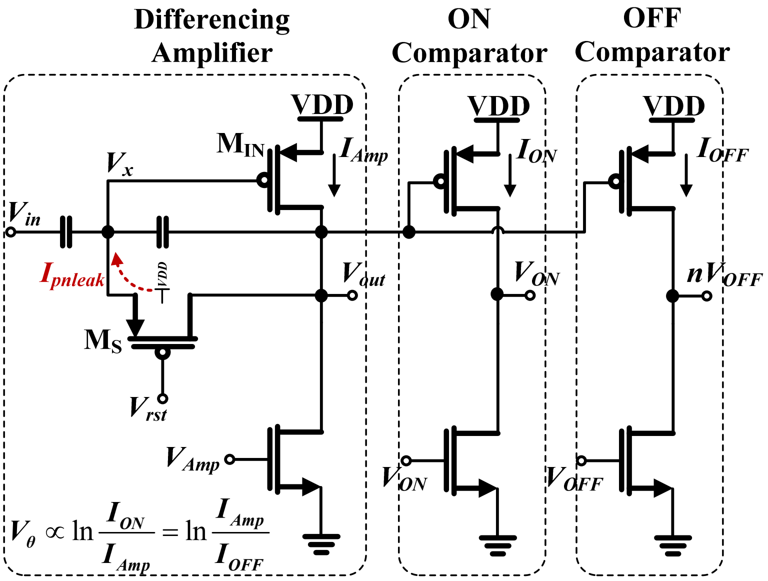
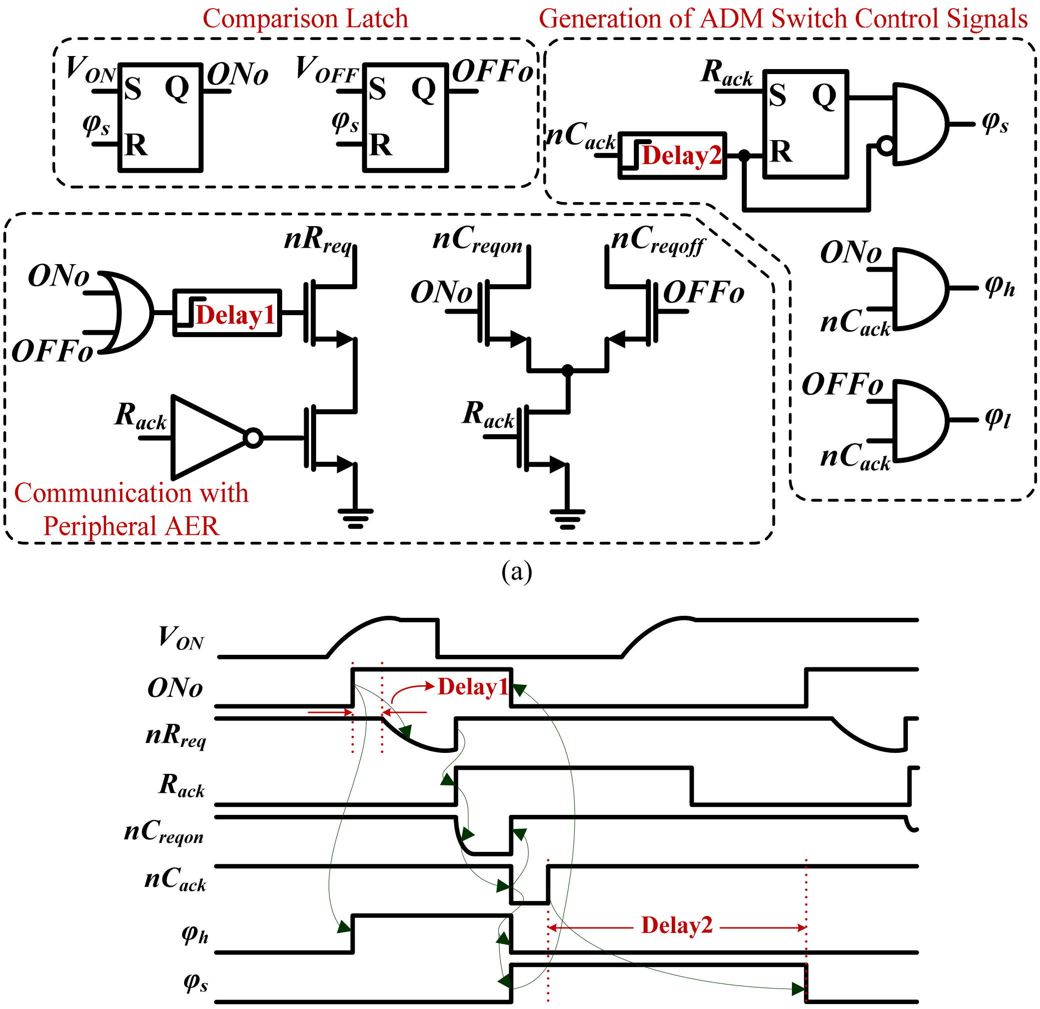


Figure 3.11. Representative transistor-level implementation of STR [6]. It comprises a differencing amplifier and two 2T current-mode comparators.

nominal subtracted voltage smaller than δ and C_{rst} is 1 fF smaller than the feedback capacitance C_{fb}, the overall actual subtracted voltage at the amplifier output V_{out} is about δ in post-layout simulation. In retrospect, for signal reconstruction from spike trains, the subtraction imbalance does not necessarily have to be dealt with in encoding as long as it is taken into account in decoding by using different incremental values for ON and OFF spikes. For spike processing, the imbalance could also be absorbed in hardware design, especially if machine learning algorithms are adopted for classification tasks.

The pseudo-resistor R_{fb} is used to establish the DC operating point of the Opamp, however it cannot employ the off-state pFETs in series as used in the CC-PGA because the switch S_0 implemented by a complementary transmission gate has considerable leakage compared to the drain-bulk PN junction leakage of off-state pFETs, which can cause DC feedback to fail. On the other hand, the traditional pseudo-resistor of two diode-connected pFETs in series [119] cannot provide sufficiently large resistance for a sub-Hz highpass corner frequency in 0.18 μm CMOS with C_{fb} only about 7 fF. The proposed pseudo-resistor shown in Figure 3.10(a) is composed of two diode-connected pFETs with bulk connected to VDD. This topology supplies both PN junction leakage and channel leakage currents. The connection of the bulk to VDD instead of the source not only prevents forward conduction of PN junction at large output swing and in turn distortion, but also increases the effective resistance. The simulated highpass corner frequency is <0.25 Hz under room temperature. One benefit of continuous-time feedback via R_{fb} instead of using reset switch to refresh the output DC level as is done in STR is the avoidance of background ON spikes that are not correlated to input stimulus. A representative transistor implementation of STR is shown in Figure 3.11. The reset switch M_S connects the gate and drain of the input transistor M_{IN} of the differencing amplifier every time after either the ON or OFF threshold set by the ON or OFF comparators is crossed. When M_S is off, even with no input signal present, ON spikes can still occur due to the charging of V_x towards VDD via the bulk-source junction leakage.

The equally spaced thresholds of the two 2T current-mode ON and OFF comparators in Figure 3.11 are derived as:



(a)

(b)

Figure 3.12. (a) Circuit diagram of the IPAsyncL and (b) timing diagram of the important signals involved in an ON spike transmission.

$$V_{\theta ON} = \frac{V_T}{\kappa}\ln\frac{I_{ON}}{I_{Amp}}, \quad V_{\theta OFF} = \frac{V_T}{\kappa}\ln\frac{I_{Amp}}{I_{OFF}}$$

To have an empirically minimum 50 mV threshold, the current ratio I_{ON}/I_{OFF} is about 22. The largely different biases of the ON and OFF comparators results in the speed gap, i.e. the OFF comparator has a much larger delay than the ON comparator, and hence severely limits the SDR of the encoded spike train, as analyzed in Chapter 2. To circumvent this problem, a two-stage uncompensated amplifier is used as the comparator whose threshold is directly set by a reference voltage and is independent of its bias current.

3.2.4 In-Pixel Asynchronous Logic (IPAsyncL)

The IPAsyncL communicates with the peripheral AER and generates the switch control signals φ_h, φ_l, and φ_s for ADM. The circuit diagram is shown in Figure 3.12(a), and the timing diagram of the main signals for an ON spike is depicted in Figure 3.12(b). The IPAsyncL can be divided into three parts:

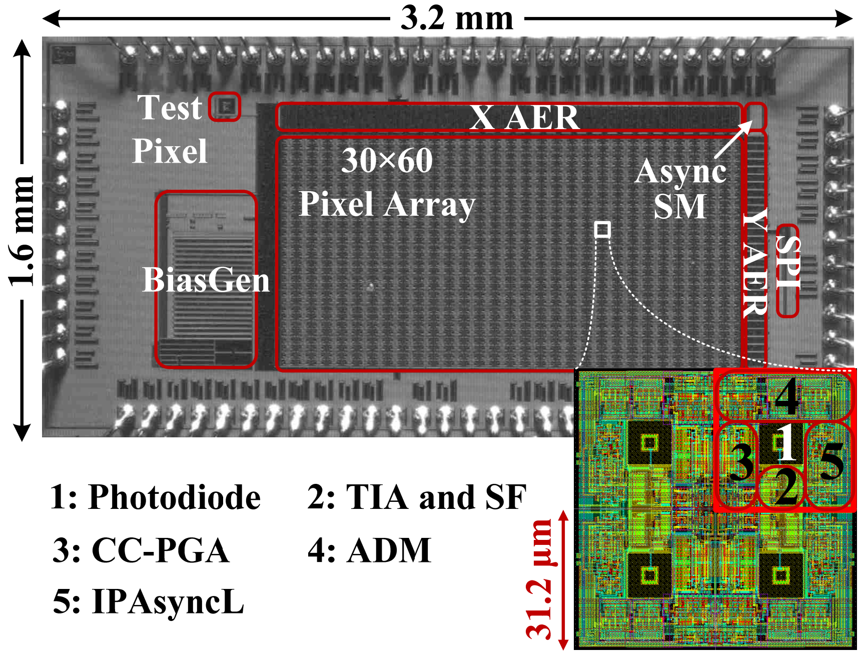
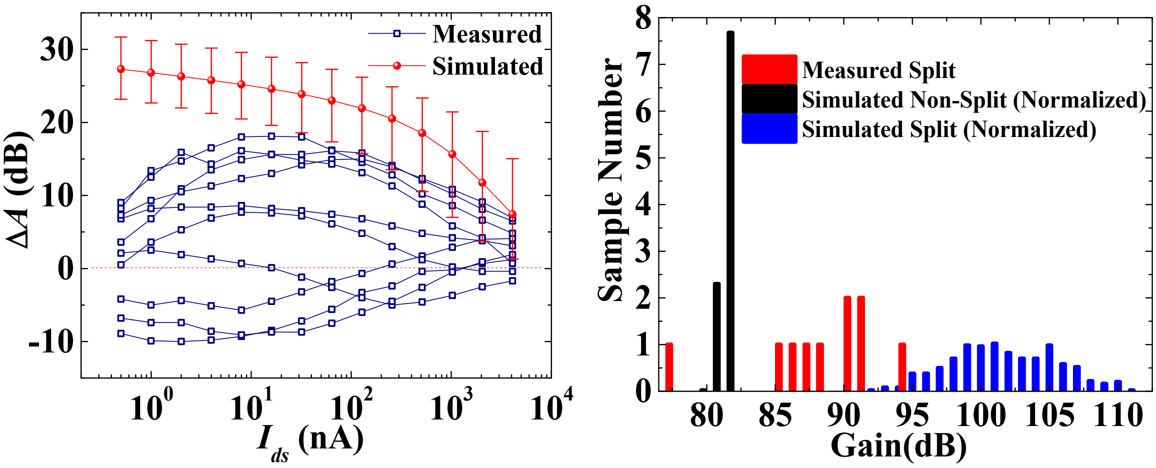


Figure 3.13. 2×2 pixel layout arranged in common centroid and the chip microphotograph.

(1). The comparison latch part is used to latch the threshold-crossing events detected by the two comparators in the ADM. In previous DVS designs, the active output of the comparators was directly used to trigger the row request nR_{req}. This is problematic because the comparator output can become inactive due to noise or digital coupling before the column request is sent for off-chip registration of the pixel address. Row-only spikes were often observed in prior DVSs which indicates failure of spike transmission due to missing column addresses. In the present design, two SR latches lock the positive state of threshold-crossing for ON and OFF spikes respectively. The latch output ONo or $OFFo$ stays active until φ_s gets high which means the charge redistribution starts in the ADM; ONo and $OFFo$ cannot get high until φ_s gets low which means the charge redistribution is finished. As depicted in Figure 3.12(b), V_{ON} goes low before nC_{reqon} is sent, but ONo still stays high to finish the communication cycle.

(2). The communication logic is designed to be compatible with the peripheral AER circuits that implement the burst-mode word-serial spike transmission [83], [84]. Once either ONo or $OFFo$ is active, after a small delay controlled by the rising edge delay element Delay1, the row request nR_{req} gets low. After the row acknowledge R_{ack} becomes high in response to nR_{req}, the column request either nC_{reqon} or nC_{reqoff} gets low. The column acknowledge nC_{ack} becomes low in response to active nC_{reqon}/nC_{reqoff}, and it becomes high after a certain time window (about ~30 ns minimum in current design in 0.18 µm CMOS at 1.8 V) controlled by the off-chip CPLD/FPGA. R_{ack} only becomes low again after all the active pixels in this row have transmitted their column addresses. Delay1 is to ensure sufficient pulse width of φ_h or φ_l for charging the C_{rst} in ADM.

(3). The three switch control signals φ_h, φ_l, and φ_s for ADM are generated according to the timing of the handshake communication described above. φ_h or φ_l gets high in response to high ONo or $OFFo$, and gets low once nC_{ack} becomes active low. φ_s only goes high if both R_{ack} and nC_{ack} are active. Its pulse width is controlled by the rising edge delay element Delay2. The SR latch associated with φ_s is to prevent φ_s goes low prematurely in response to R_{ack} going low. However, the delay from Delay2 cannot be too long,



(a) (b)

Figure 3.14. (a) Measured DC-gain ratio A_s/A_{ns} in dB of split to non-split nFETs over 10 samples at different bias currents, along with Monte Carlo simulation results; (b) Measured open-loop DC-gain distribution of the Opamp in Figure 3.9(a) over 10 samples, along with Monte Carlo simulation results.

otherwise nC_{ack} from the same column but another row could block φ_s from going low and in turn block this pixel from requesting again until φ_s could eventually get low. A better design is to disassociate the pulse width of φ_s from the rising edge of nC_{ack} by only using simultaneously active R_{ack} and nC_{ack} as the trigger to a simple timer which independently determines the φ_s pulse width.

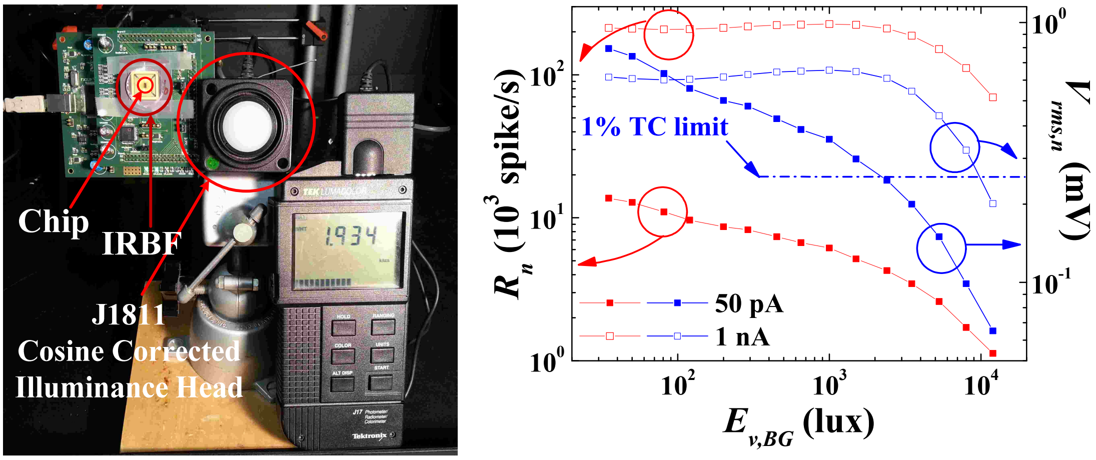
3.3 Experimental Results

3.3.1 Pixel Layout and Chip Microphotograph

A 60×30 vision sensor prototype called ADMDVS was designed and fabricated in UMC 0.18 μm RF/MM CMOS. The pixel layout and the chip microphotograph are shown in Figure 3.13. The 2×2 pixel layout shows a common centroid arrangement. Each pixel has a pitch size of 31.2 μm and a fill factor of about 10.3%. The 10×10 μm^2 photodiode is formed by the n-well/p-substrate junction. A metal ring made of M1 to M6 that is connected to ground shields the photodiode from any signal coupling from the other building blocks. To minimize the digital coupling, all request and acknowledge signals are wired using mostly M1 at the outer edge of block 4 and 5. The analog biases using M3 and M4 are carefully shielded from digital signals using ground or VDD intermediate metal layers. The whole chip including the pads occupies 3.2×1.6 mm^2. Besides the X/Y AER and the asynchronous state machine (SM) for spike transmission, the chip includes a digitally programmable proportional-to-absolute-temperature (PTAT) bias generator array (see Chapter 5) [120] to provide all the current biases, a serial-to-parallel interface (SPI) to configure digital bits for pixel gain control, chip power down and debug, a test pixel and test structures for split transistors described in Section 3.2.2. The USB interface, firmware logic, and host side codes in jAER [121] are based on existing designs.

3.3.2 Split-Transistor Gain

To validate the split-transistor technique exploiting small-size effects, a compound split nFET composed of two M_{sa} in series and two M_{sb} in parallel (M_{sa}/M_{sb} has the same size as the one in Figure 3.9(b)) was fabricated along with a normal nFET with $L_{ns}=2*L_{sa}+L_{sb}=0.9$ μm and $W_{ns}=2(W_{sa}L_{sa}+W_{sb}L_{sb})/L_{ns}=0.48$



(a) (b)

Figure 3.15. (a) Experiment setup for measuring the noise of the ADMDVS array; (b) Measured average noise spike rate R_n of the ADMDVS array and calculated average equivalent noise voltage $V_{rms,n}$ at the output of one CGPR with two different bias current settings of the CC-PGA, 50 pA and 1 nA.

μm. The I_{ds}-V_{ds} characteristic curves of the split and non-split nFETs were measured using Keithley 236 over 10 samples from different dies and the DC gains were extrapolated from the slope of the I_{ds}-V_{ds} curves in the saturation region. The calculated gain ratio $\Delta A=20\log_{10}(A_s/A_{ns})$ in dB at different bias currents are plotted in Figure 3.14(a), along with the 500 Monte Carlo simulation trial results using Cadence Spectre. The error bar shows the simulated variance. The measured results give a maximum 16 dB DC-gain increases of the split transistor within the 8-64 nA range. The measured samples exhibit a large variation that can be attributed to the small device area. Some samples even have decreased DC gain which was implied by Eq. (3.12) if the M_{sa} or compound M_{sa} is not in saturation. The simulation generally overestimates the DC gain enhancement, especially at very low biases. Such large discrepancy could be the consequence of unreliable device modeling and inaccurate device parameter extraction. Particularly it is known that accurate g_{ds} modeling of MOSFETs is difficult in deep subthreshold region.

With a 4 nA total current consumption sufficient for 500 Hz bandwidth in-pixel gain amplification at $G_1G_0=11$, the open-loop DC gain of the Opamp in Figure 3.9(a) was measured over 10 samples by an SR780 network signal analyzer. The distribution of the measured and simulated DC gain is plotted in Figure 3.14(b). The 500 Monte Carlo simulation trials are scaled to 10 so that it is evident to be compared with 10 tested samples. Statistically the measured DC gain of the Opamp is less than the simulated one using split transistor while mostly larger than the simulated one using non-split transistors. In all, 9 out of the 10 samples have DC gains larger than 85.4 dB with a maximum up to 94.0 dB. Only one sample has a gain of about 77.1, slightly less than the simulated mean 81.7 dB of the Opamp using non-split transistors.

3.3.3 Sensor Array Noise

Noise performance of the ADMDVS is important because it determines its minimum TC sensitivity. To estimate the average equivalent noise $V_{rms,n}$ at the output of a CGPR, the average noise spike rate R_n of the sensor array under DC illumination was measured with the setup shown in Figure 3.15(a). The chip is covered with an infrared blocking filter (IRBF) so that the illuminance value shown on the Tektronix J17 photometer approximately reflects the actual illuminance on the pixel array. The tunable light source was

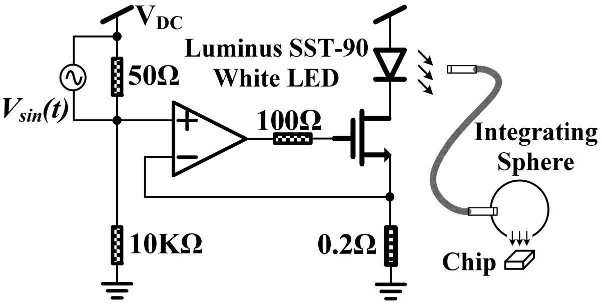


Figure 3.16. Experiment setup for measuring the average TC sensitivity θ of the ADMDVS array

Table 3.3. Stimulus contrast C_{sti} used for measuring θ at different gain codes G_1G_0 of the CC-PGA.

G_1G_0	C_{sti}
00	1.17
01	0.628
10	0.340
11	0.174

a QT-DE12R7s floodlight lamp with 500-W maximum power. A white Gaussian noise input is assumed for the ADMs in all pixels. Using Eq. (3.a1) in the Appendix, $V_{rms,n}$ can be calculated as:

$$V_{rms,n} = \frac{R_n \cdot \delta}{M \cdot N \cdot 2\sqrt{\frac{2\pi}{3} f_{BW} A_{FE}}} \tag{3.14}$$

where $M{\cdot}N$=1.8×10^3 is the total pixel number, δ is the ADM threshold, f_{BW} is the CC-PGA bandwidth, and A_{FE} is the front-end gain including the gain of CC-PGA and ADM.

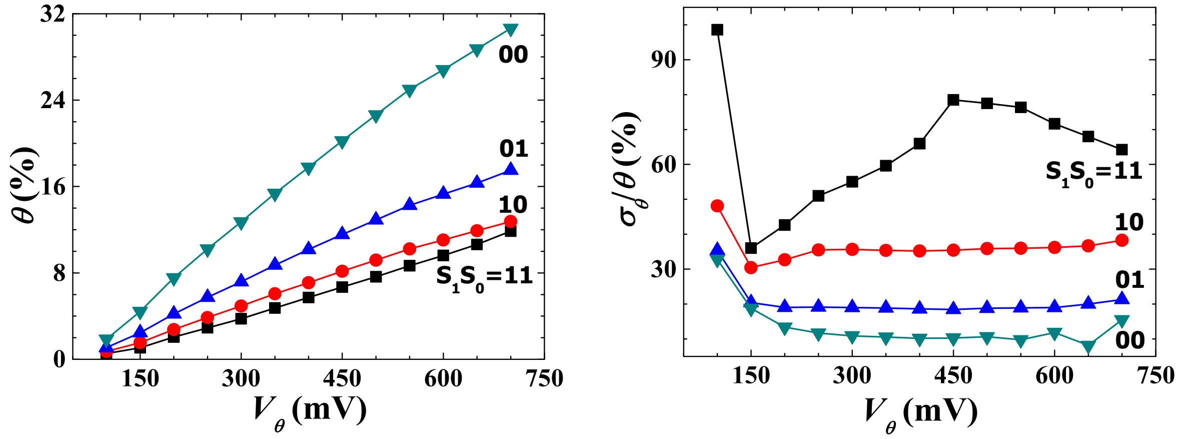
Figure 3.15(b) shows the plot of the measured R_n and calculated $V_{rms,n}$ versus background illuminance $E_{v,BG}$ with the TC threshold set to about 2.4% under two different nominal biases of the CC-PGA (50 pA and 1 nA). At 1 nA bias, the maximum R_n of about 220k spike/s occupies a considerable portion of the array's total bandwidth of 10M spike/s. R_n would increase linearly as the array size increases. Limiting f_{BW} by setting the bias to 50 pA lowers R_n, and also lowers $V_{rms,n}$, because the noise contributed by CGPRs is further filtered. The $V_{rms,n}$ curves resemble the simulated CGPR output noise in Figure 3.5(b). To keep $V_{rms,n}$ less than the 1% TC limit, $E_{v,BG}$>2.5k lux is needed at 50 pA bias, and $E_{v,BG}$ >10k lux at 1 nA bias. Therefore, to achieve a <1% TC sensitivity in a single pixel, low pixel bandwidth and sufficiently high illumination are the two essential factors. In fact, it is observed that removal of IRBF from the chip coverage induces significant R_n drop (up to 50× at about 8k lux) which is believed to be the result of significantly increased photocurrent thanks to the exposure to a large amount of infrared. The flattened tail of the photoreceptor output noise at high illuminance limited by $1/f$ noise in Figure 3.5 also becomes visible in measurement when IRBF is removed even though it is not shown in Figure 3.15(b).

3.3.4 Temporal Contrast Sensitivity

The experimental setup for the sensitivity test is shown in Figure 3.16. An SST-90 white LED from Luminus Devices with a maximum luminous flux of 2500 lm modulated by a sinusoidal signal $V_{sin}(t)$ is used to provide the input stimulus for the chip through an integrating sphere. The peak-to-peak voltage of $V_{sin}(t)$ and DC bias current of the LED determines the stimulus contrast C_{sti}. The different C_{sti} values used for sensitivity testing (listed in TABLE 3.3) at different CC-PGA gain codes G_1G_0 are used to keep a relatively constant output spike rate. C_{sti} is calculated as

$$C_{sti} = \ln \frac{I_{\max}}{I_{\min}}$$

where I_{max} and I_{min} are the measured maximum and minimum illuminance from the LED. Using the signal



(a) (b)

Figure 3.17. (a) Measured average TC sensitivity θ of the ADMDVS and (b) measured relative standard deviation of the TC sensitivity σ_θ/θ at different threshold voltages V_θ of the ADM and different gain codes G_1G_0 of the CC-PGA.

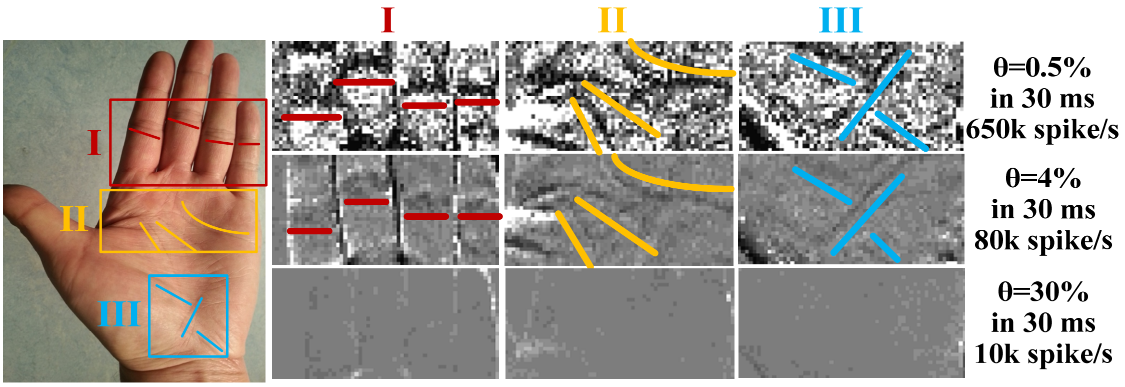


Figure 3.18. Effect of different TC sensitivity settings on detecting fine palm lines. The vertically moving hand is divided into three different regions marked as I, II, and III. The ADMDVS is used to detect the lines of each region separately.

spike rate R_{sin} of the array in response to the stimulus $V_{sin}(t)$, the average detectable TC threshold, i.e. the TC sensitivity θ can be calculated as

$$\theta = \frac{2M \cdot N \cdot f_{sin} \cdot C_{sti}}{R_{sin}} \tag{3.15}$$

where f_{sin} is the frequency of $V_{sin}(t)$. To evaluate the standard deviation of the sensitivity σ_θ among all pixels, the signal spike rate of each pixel $R_{sin,pixel}$ is used to calculate the sensitivity for each pixel θ_{pixel}, and σ_θ is obtained by applying a Gaussian fit to all θ_{pixel}.

As pointed out in Section 3.3.3, the noise spike rate R_n is quite considerable when the ADM threshold is set small even with a high illuminance, so the actual measured spike rate R_{sin+n} under stimulus is due to both the sinusoidal signal stimulus and the noise. Therefore, directly using R_{sin+n} to calculate θ results in overestimated sensitivity. However, with measured R_{sin+n} and R_n, the R_{sin} can be calculated using Eq. (3.16) below (see Appendix) assuming the noise is white and Gaussian

$$R_{sin+n} = \sqrt{\frac{2}{\pi}} R_n e^{-c^2/4} \left[\left(1.25+0.63c^2\right) I_0\left(\frac{c^2}{4}\right) + 0.63b^2 I_1\left(\frac{c^2}{4}\right) \right] \quad (3.16)$$

where the parameter c is given as:

$$c = \sqrt{\frac{\pi}{2}} \frac{R_{sin}}{R_n}$$

and $I_v(x)$ (v=0, 1) is the modified Bessel function of the first kind. Note that this noise de-embedding is only necessary for low gain settings at G_1G_0=00 and 01 where R_n is substantial when V_θ is small, and even though the sinusoidal input stimuli is logarithmically transformed to voltage signals by the photoreceptor, the ADM input can still be approximately regarded as a sinusoid because $\ln(1+\Delta) \approx \Delta$ when Δ is much less than 1.

The TC sensitivity θ versus V_θ at different gain codes G_1G_0 of the CC-PGA and the corresponding relative standard deviation σ_θ/θ versus V_θ are plotted in Figure 3.17. Figure 3.17(a) shows that θ increases by less than a factor of two at the same V_θ when G_1G_0 decreases one step, especially from 11 to 10 where θ increases by only ×1.25 on average. This non-ideal gain step is because the underestimated parasitic capacitance in the fabricated chips is comparable to the feedback capacitance C_{fb} in Figure 3.7. θ increases relatively linearly with V_θ for all G_1G_0. The minimum measured θ is about 0.54% at G_1G_0=11 and V_θ=100 mV; below this V_θ, R_n is too large for the array to detect any visual signal. The obvious penalty of small θ at high gain and low V_θ is the large θ variation as can be seen in Figure 3.17(b). For a reasonable $\sigma_\theta/\theta \approx 35\%$, θ is about 1%. At high gain, the feedback capacitance in CC-PGA is small, which causes significant capacitance mismatch and in turn large σ_θ/θ. The mismatch decreases as the capacitance increases, i.e. the gain decreases. At low V_θ, the variation of the ADM amplifier's DC output and the input offset of the two comparators contribute substantially to σ_θ/θ, which is mitigated as V_θ increases because the DC variation and offset become a smaller portion of V_θ.

The effect of different TC sensitivity settings is demonstrated by detecting the fine palm lines of a moving hand under office lighting (~500 lux) as shown in Figure 3.18. The hand moves at a speed of about 15 cm/s and it is about 6 cm away from the lens. The experiment was repeated for three different sensitivity settings. The palm lines in different parts of the hand are marked in the original hand image, and also in the accumulated-spike histogram images (over a time window of 30 ms for θ=0.5%, 4% and 30%) wherever they are visually detectable. It is clear that a small θ setting helps detection of low contrast objects, although with a larger fixed-pattern noise. On the other hand, a large θ setting can be used to detect the contour of high contrast objects with minimal noise.

3.3.5 Spike Encoding Comparison

To verify the improved spike encoding of the proposed in-pixel ADM over STR, the ADMDVS chip is compared with a DAVIS chip with STR that is previously developed in our group [18], using a moving visual pattern as the input stimulus and a simple histogram reconstruction as output. For a fair comparison between the two chips, the following three factors were paid attention to in the designed experiment: 1. To ensure that the front-end bandwidth is not limited by the photoreceptor, the 500 W floodlight lamp is used to provide additional lighting so that the illuminance at the position of the rotating image is about 4k lux; 2. The front-end bandwidth is set to about 300 Hz by the source follower in STR pixel and the CC-PGA in ADM pixel respectively so that the signals fed into the two encoders have approximately the same bandwidth; 3. The cutoff frequency of the amplifiers for spike encoding is set to about 500 Hz, and

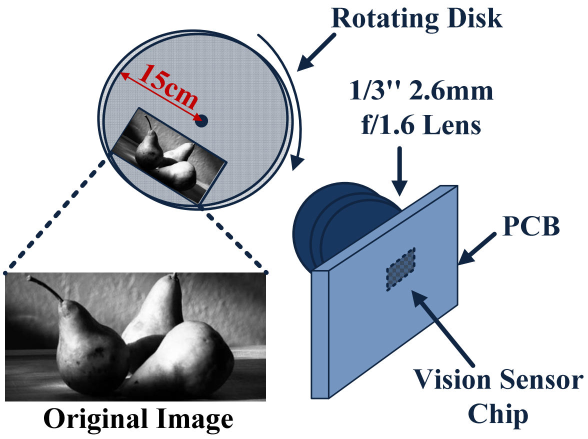


Figure 3.19. Testing setup and the original image for the spike encoding comparison experiment.

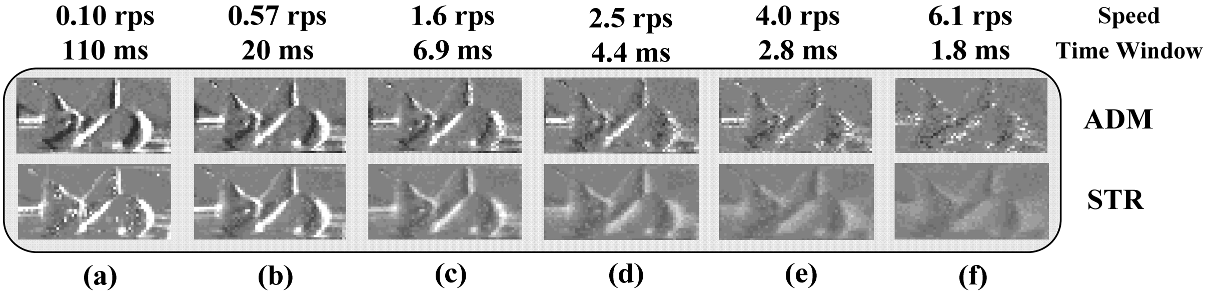


Figure 3.20. Comparison of the spike-accumulated histogram images acquired by the chip with ADM encoding (ADMDVS) and the chip with STR encoding (DAVIS) [18] at different rotational speeds within time windows which are inversely proportional to the rotational speeds.

the sensitivity to about 15%. As illustrated in Figure 3.19, the sensors are mounted with a 1/3″ 2.6 mm f/1.6 lens, and a pear image is attached to a disk driven by a motor with an adjustable rotational speed. Because the DAVIS has a 240×180 resolution, a 60×30 region was selected to match with the ADMDVS. The spike-accumulated histogram images in Figure 3.20 are acquired by the two sensors at different rotational speeds from 0.10 to 6.1 rps within time windows inversely proportional to the rotational speeds. The dark edges in the ADM images are more clearly defined and the bright edges become obscure more slowly as rotational speed increases compared to those in the STR images. Note that in Figure 3.20(a), the STR image shows clearly visible scattered white dots, which probably indicates background ON spikes caused by leakage charging of the reset switch M_S in Figure 3.11.

To quantify the spike encoding improvement, the following method is proposed for comparing the number of produced spikes from the two sensors. Figure 3.21 shows the comparison of the average number of ON/OFF spikes (SN=spike number) from the 108 most active pixels of the images in Figure 3.19. The number of pixels chosen for SN averaging depends on the percentage of active pixels in the array. Ideally the average ON/OFF SN should stay constant with stimulus speed given infinite pixel bandwidth and instantaneous spike feedback. However because of the limited 300 Hz analog front-end bandwidth used in this experiment, the average SN in both STR and ADM decreases. Although the average ON SN in STR is approximately the same as that in ADM at 0.1 rps, implying the nearly identical sensitivity set-

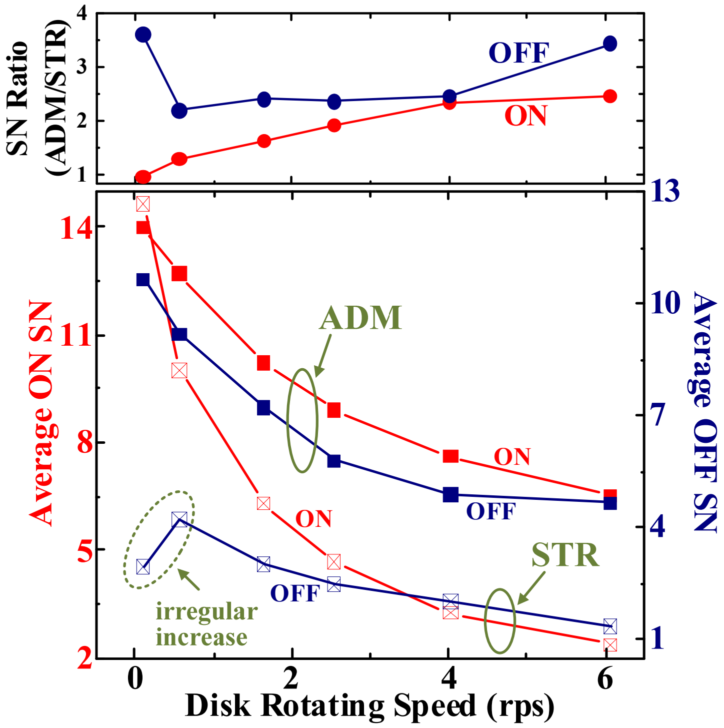


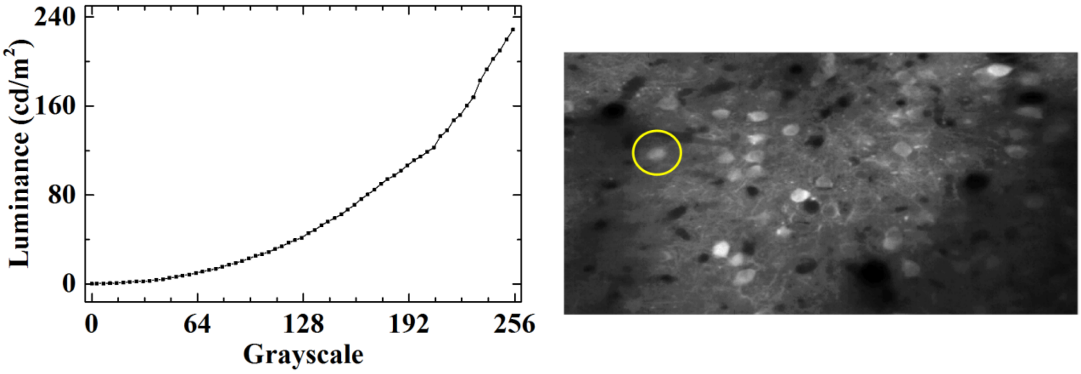
Figure 3.21. Comparison of the average number of ON/OFF spikes (SN=spike number) over the 108 brightest/darkest pixels and the SN ratio of ADMDVS over DAVIS.

ting, it decreases faster with increasing rotational speeds as indicated by the increasing ON SN ratio up to 3 because of signal loss in STR during reset and refractory period. This ON SN ratio increase supports the SDR improvement of ADM against STR in model simulation as shown in Chapter 2. The much lower average OFF SN in STR even at low rotational speeds is due to severe signal loss caused by large feedback delay with the threshold-determined low bias of the OFF comparator, and the maximum SN ratio is up to >3.5. The irregular increase of average OFF SN in STR from 0.1 to 0.57 rps is due to the junction leakage of the reset switch M_S in Fig. 11 in accordance with the comparison of the STR images between Figure 3.20(a) and 3.20(b). This is eliminated in ADM thanks to the continuous-time DC feedback using the pseudo-resistor.

3.3.6 Simulated Optical Neuroimaging using ADMDVS

The prototype ADMDVS is applied to a simulated optical neuroimaging experiment. A fluorescence imaging video recorded from a region in mouse cortex is displayed on a screen and the ADMDVS mounted with the same lens as in Section 3.3.5 is placed in front of the screen. The measured grayscale-luminance relationship using photometer is given in Figure 3.22(a). Assuming a lens transmissibility of 0.9, the average illuminance on the chip is calculated to be about 6.9 lux with the measured screen luminance of about 25 cd/m^2 (using Eq. (3.b1) in the Appendix). The TC sensitivity of the ADMDVS is set to 2.7%. Figure 3.22(b) shows one frame of the optical neuroimaging recording. The target neuron circled in yellow is used to demonstrate the effectiveness of temporal signal reconstruction from the ADMDVS output spikes.

The upper waveform in Figure 3.23 is the grayscale value over time averaged from 5×5 pixels within



(a) (b)

Figure 3.22. (a) Measured grayscale-luminance relationship of the screen; (b) one frame of the optical neuroimaging video with the target neuron circled in yellow.

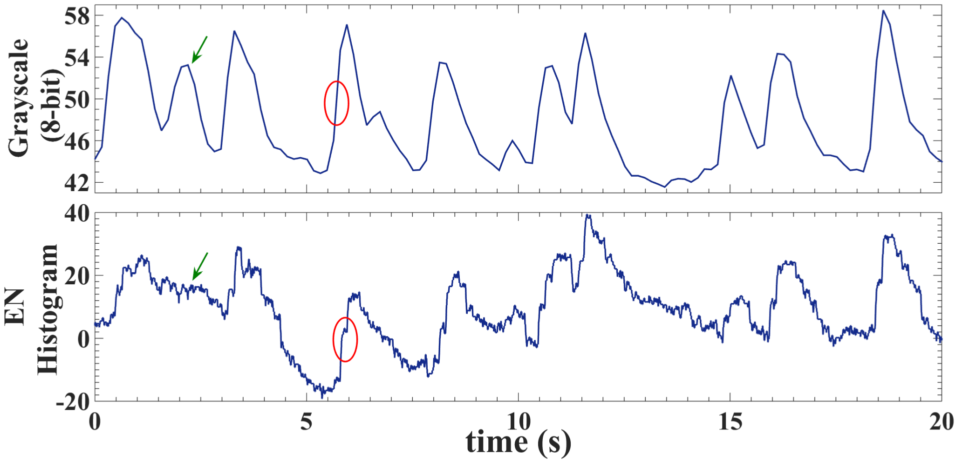


Figure 3.23. The temporal waveform averaged over a 5×5 pixel window within the target neuron of the video in (a) (upper plot); The linear signal reconstruction using the output ON and OFF spikes from one ADMDVS pixel whose visual field covers the region around the target neuron (lower plot).

the target neuron in the video. The lower waveform in Figure 3.23 is the simple linear reconstruction from the ON and OFF spikes recorded by one ADMDVS pixel that has the visual field around the target neuron. The corresponding peaks in the two waveforms are evident. The missing peak pointed by the green arrow is likely due to the fact that the ADMDVS pixel may not have the exact visual field of the 5×5 pixels. On the other hand, the rising edge circled in red in the upper waveform has a contrast of 61% calculated according to the measured grayscale-luminance relationship in Figure 3.22(a), and the one in the lower waveform is composed of 26 ON events corresponding to a contrast of $\ln[(1+2.7\%)^{26}]=69\%$, well close to 61%. The long-term DC level fluctuation of the peaks in the reconstructed waveform is caused by the unbalanced ON and OFF thresholds of the fabricated ADM circuit.

3.4 Conclusion and Remarks

This chapter described a specific type of silicon retina, i.e. the dynamic vision sensor (DVS) with enhanced TC sensitivity and spike encoding. The performance is summarized and compared with prior DVSs in Table 3.4. Although the circuit area of this proposed pixel is approximately 3 times larger than

Table 3.4. Comparison with previous DVSs.

	This work	2014 [18]	2013 [16]	2011 [15]	2011 [17]	2008 [6]
Technology	0.18 µm MM/RF	0.18 µm IS	0.35 µm IS	0.35 µm MM/RF	0.18 µm MM/RF	0.35 µm MM/RF
Resolution	60×30	240×180	128×128	128×128	304×240	128×128
Chip Area (mm^2)	3.2×1.6	5×5	4.9×4.9	5.6×5.5	9.9×8.2	6.3×6
Pixel Area (μm^2)	31.2×31.2	18.5×18.5[a]	31×30	35×35	30×30[b]	40×40
Fill Factor (%)	10.3	22	10.5	8.7	10[c]	8.1
Supply Voltage (V)	1.8	1.8/3.3	3.3	3.3	1.8/3.3	3.3
Power (mW)	0.72[d]	14[e]	4[d]	145[d]	175[e]	24[f]
Power/Pixel (µW)	0.40	0.32	0.24	8.8	2.4	1.5
Min. TC Sensitivity (%)	**1**	11	1.5	10	13	17
DR (dB)	130[g]	130	120	100	N.A.[h]	120
Intra-Scene DR (dB)	**130**	130	60	56	N.A.	120
Event Encoding	**ADM**	STR	STR	STR	STR	STR

a. Including 4T APS; b. Including PWM imaging circuits; c. Only DVS photodiode; d. At 10^5 event/s; e. High activity, including the imaging functionality; f. Non-optimized power-consuming biasgen; g. About 0.03 lux to >100k lux; h. Only DR of PWM imaging given.

[18], it achieves comparable power consumption per pixel, and a 1% TC sensitivity with a 35% relative standard deviation without sacrificing the intra-scene DR by using an in-pixel CC-PGA. The TC sensitivity record is 0.3% [122], but it was only demonstrated in a single pixel with a very limited pixel bandwidth. The prototype ADMDVS also employs an in-pixel ADM for spike encoding which has been in MATLAB simulation (Chapter 2) [123] and here experimentally verified to have a better encoding quality compared to STR (Note that a similar switched-cap circuit used in relaxation oscillator for constant charge subtraction like the one used for the in-pixel ADM here is almost simultaneously published on JSSC by a group at Michigan University [124]). These improvements together with the intrinsic low-latency sparse-output features of DVSs pave the way for applications like wireless in-vivo optical neuroimaging on free-moving animals, where the energy spent on RF data transmission can be reduced.

For VSDI with signal temporal contrast often less than 1% [92], further improved SNR at a high sensitivity setting is still necessary. One obvious means is to increase the photodiode size. Optimized photodiodes in a dedicated image sensor process with higher quantum efficiency and micro-lenses can be used to obtain high photocurrent so that shot noise is reduced. Transistors with large L and W help lower the contribution of $1/f$ noise. Large compensation capacitance and large closed-loop gain can be used to reduce the CC-PGA noise at the cost of pixel area and power consumption.

Although an ADM improves the spike encoding integrity compared to STR, for real-time high-fidelity signal reconstruction [94], [95], the problems of sensitivity variation among pixels and unbalanced ON and OFF threshold remain. They could be addressed at the circuit level by using larger transistors and capacitors with increased pixel size. A novel encoding mechanism might give an area-efficient solution. For example, a threshold-variation-insensitive decoding algorithm was developed for the asynchronous sigma-delta modulation (ASDM) [52], although the ASDM generates idle output without any input signal change and thus results in a much more limited pixel spike-output bandwidth.

3.5 Appendix

3.5.a Spike Rate of ADM

Let us assume a white Gaussian noise input $x_n(t)$ that has zero mean, rms amplitude $V_{rms,n}$, and bandwidth $f_{bw,n}$, then the output spike rate R_n of an ADM can be derived as [57]:

$$R_n = 2\sqrt{\frac{2\pi}{3}}\frac{f_{bw,n}V_{rms,n}}{\delta} \tag{3.a1}$$

where δ is the ADM threshold. For a sinusoidal input $x_{sin}(t)$, the output spike rate R_{sin} is [57]:

$$R_{sin} = 4\sqrt{2}\frac{f_{sin}V_{rms,sin}}{\delta} \tag{3.a2}$$

where f_{sin} is the signal frequency, and $V_{rms,sin}$ is the rms amplitude of $x_{sin}(t)$.

Let $x_{sin+n}(t)=x_{sin}(t)+x_n(t)$ represent the sum of a sinusoidal signal and a white Gaussian noise. To derive its corresponding spike rate R_{sin+n}, the mean of the absolute slope needs to be obtained. The joint probability density function $p(a_x,s_x,t)$ of the amplitude a_x and the slope s_x of x_{sin+n} at time t is given as [125]:

$$p(a_x,s_x,t) = \frac{\pi N_0}{-\psi_0''}\varphi(y-c_1\cos(2\pi f_{sin}t))\varphi(x+c_2\sin(2\pi f_{sin}t))$$

where the parameters N_0, x, y, c_1, c_2, and the function $\varphi(z)$ are given below:

$$N_0 = \frac{1}{\pi}\sqrt{\frac{-\psi_0''}{\psi_0}}\,,\ x = \frac{s_x}{\sqrt{-\psi_0''}}\,,\ y = \frac{a_x}{\sqrt{\psi_0}}\,,\ c_1 = \sqrt{\frac{2}{\psi_0}}V_{rms,sin}\,,\ c_2 = \sqrt{\frac{2}{-\psi_0''}}2\pi V_{rms,sin}f_{sin}\,,\ \varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

ψ_0 is the correlation function $\psi(\tau)$ of $x_n(t)$ at τ=0, and ψ_0'' is the second derivative of $\psi(\tau)$ at τ=0. For bandlimited white noise that has a power of $V^2_{rms,n}$ within $f_{bw,n}$, ψ_0 and ψ_0'' can be calculated as:

$$\psi_0 = \int_0^{f_{bw,n}}\frac{V^2_{rms,n}}{f_{bw,n}}\mathrm{d}f = V^2_{rms,n}\,,\ \psi_0'' = -4\pi^2\int_0^{f_{bw,n}}\frac{V^2_{rms,n}}{f_{bw,n}}f^2\mathrm{d}f = -\frac{4}{3}\pi^2V^2_{rms,n}f^2_{bw.n}$$

The probability density function $p(s_x,t)$ of the slope s_x can be obtained by integrating the amplitude over its range, namely from -∞ to +∞:

$$p(s_x,t) = \frac{\pi N_0}{-\psi_0''}\varphi(x+c_2\sin(2\pi f_{sin}t))\int_{-\infty}^{+\infty}\varphi(y-c_1\cos(2\pi f_{sin}t))\mathrm{d}a_x = \frac{\pi N_0}{-\psi_0''}V_{rms,n}\varphi(x+c_2\sin(2\pi f_{sin}t))$$

The mean of $|s_x(t)|$ can be calculated as:

$$<|s_x(t)|> = 2\int_0^{\infty}s_xp(s_xt)\mathrm{d}s_x = \frac{2\pi N_0}{-\psi_0''}V_{rms,n}\int_0^{\infty}s_x\varphi(x+c_2\sin(2\pi f_{sin}t))\mathrm{d}s_x$$

$$\xrightarrow{v=c_2\sin(2\pi f_{sin}t)} 2\pi N_0V_{rms,n}\int_0^{\infty}x\varphi(x+v)\mathrm{d}x = 2\pi N_0V_{rms,n}[-\frac{v}{2}+\varphi(v)+v\int_0^{v}\varphi(z)\mathrm{d}z]$$

The mean of $|s_x|$ can be calculated by taking the limit below:

$$\begin{aligned}
<|s_x|> &= \lim_{T\to\infty}\frac{1}{T}\int_0^T <|s_x(t)|> \mathrm{d}t = 2\pi N_0 V_{rms,n}\lim_{T\to\infty}\frac{1}{T}\int_0^T[-\frac{v}{2}+\varphi(v)+v\int_0^v\varphi(z)\mathrm{d}z]\mathrm{d}t \\
&= 2\pi N_0 V_{rms,n}\lim_{T\to\infty}\frac{1}{T}\int_0^T[-\frac{c_2\sin(2\pi f_{sin}t)}{2}+\varphi(c_2\sin(2\pi f_{sin}t))+c_2\sin(2\pi f_{sin}t)\int_0^{c_2\sin(2\pi f_{sin}t)}\varphi(z)\mathrm{d}z]\mathrm{d}t \\
&\xrightarrow{\theta=2\pi f_{sin}t} 2\pi N_0 V_{rms,n}\lim_{T\to\infty}\frac{1}{2\pi f_{sin}T}\int_0^{2\pi f_{sin}T}[-\frac{c_2\sin\theta}{2}+\varphi(c_2\sin\theta)+c_2\sin\theta\int_0^{c_2\sin\theta}\varphi(z)\mathrm{d}z]\mathrm{d}t \\
&= 2N_0 V_{rms,n}\int_0^{\pi}[\varphi(c_2\sin\theta)+c_2\sin\theta\int_0^{c_2\sin\theta}\varphi(z)\mathrm{d}z]\mathrm{d}\theta
\end{aligned}$$

The last step is simply obtained based on the parity of each term in the square brackets. Therefore, together with Eq. (3.a1) and Eq. (3.a2), the output spike rate R_{sin+n} of an ADM with $x_{sin+n}(t)$ as the input can be written as:

$$\begin{aligned}
R_{sin+n} &= \frac{<|s_x|>}{\delta} = \frac{4f_{bw,n}V_{rms,n}}{\sqrt{3}\delta}\int_0^{\pi}[\varphi(c_2\sin\theta)+c_2\sin\theta\int_0^{c_2\sin\theta}\varphi(z)\mathrm{d}z]\mathrm{d}\theta \\
&= \sqrt{\frac{2}{\pi}}R_n\int_0^{\pi}[\varphi(c_2\sin\theta)+c_2\sin\theta\int_0^{c_2\sin\theta}\varphi(z)\mathrm{d}z]\mathrm{d}\theta
\end{aligned} \tag{3.a3}$$

Note that the equation below holds:

$$c_2 = \sqrt{\frac{2}{-\psi_0''}}2\pi V_{rms,sin}f_{sin} = \sqrt{\frac{3}{2}}\frac{2V_{rms,sin}f_{sin}}{V_{rms,n}f_{bw,n}} = \sqrt{\frac{\pi}{2}}\frac{R_{sin}}{R_n}$$

Eq. (3.16) is the numerically integrated version of Eq. (3.a3) for faster computation.

3.5.b From Screen Luminance to Chip Illuminance

With the assumption of lumen conservation, the equation below holds:

$$L_{screen}\cdot A_{screen}\cdot\frac{\pi\frac{d_{lens}^2}{4}}{d_{scr-lens}^2}\cdot T = I_{sensor}\cdot A_{sensor}$$

where L_{screen} (cd/m^2) is the screen luminance, I_{sensor} (lux) is the illuminance at the focal plane of the ADMDVS chip, A_{screen} and A_{sensor} are the areas of the screen and the sensor respectively, d_{lens} is the lens diameter, $d_{scr\text{-}lens}$ is the distance between the screen and the lens, and T is the lens transmissibility. A_{screen} and A_{sensor} have the geometrical relationship below:

$$\frac{A_{screen}}{A_{sensor}} = \frac{d_{scr-lens}^2}{d_{sen-lens}^2}$$

where $d_{sen\text{-}lens}$ is the distance between the sensor and the lens. Therefore, I_{sensor} can be expressed as:

$$I_{sensor} = L_{screen}\cdot\frac{d_{scr-lens}^2}{d_{sen-lens}^2}\cdot\frac{d_{lens}^2}{d_{scr-lens}^2}\cdot\frac{\pi T}{4} = L_{screen}\frac{\pi T}{4f^2} \tag{3.b1}$$

where f is the lens aperture ratio.

Chapter 4: Ultra-Low-Power Binaural Silicon Cochlea

This chapter presents an ultra-low-power binaural spiking silicon cochlea with 0.5 V power supply in 0.18 μm CMOS aiming for energy-scarce applications like voice activity detection and speaker identification in scenarios of wireless sensor networks (WSNs), internet of things (IoTs), etc. As projected by major semiconductor companies including Intel and Qualcomm, deployment of trillions of sensors is envisioned to interweave the physical and cyber worlds and facilitate human-environment, human-human interactions [126]. One of the key enabling technologies is smart ultra-low-power sensor nodes that are able to make simple local decisions for information transmission instead of power-consuming raw data transmission. Hence processing units should be tightly integrated with sensing front end. Conventional audio signal acquisition and processing are dominated by clocked ADCs and DSPs. The obvious drawback of this approach is the waste of energy, because sound signals are bursty and constant sampling in clocked systems produces redundant data for processing. As one essential functional block in audio DSPs, FFT has been optimized for very high energy efficiency [41], but it still falls short compared to analog frequency division at least in some classification tasks like voice activity detection where signal SNR requirement is low-to-medium [42].

The newly-developed silicon cochlea [127] features 64×2 channels with biomimetic asymmetrical analog BPFs covering frequency range from 8 to 20k Hz and an asynchronous delta modulator (ADM) for spike encoding in each channel to facilitate event-driven clockless processing. To improve power efficiency, source-follower-based BPFs and ADMs with adaptive self-oscillating comparison are proposed. The SF-based BPF is composed of a 4th-order source-follower-based LPF and a summing PGA, and the self-oscillation is realized by employing dynamic latched comparators, simple logic gates and a delay element. The circuit details will be described in the subsections. The content of this chapter is organized as follows: Section 4.1 describes the system architecture and the composition of one cochlea channel; Section 4.2 gives the detailed 0.5-V cochlea core design, including: Section 4.2.1, the geometrically-scaled channel bias generation; Section 4.2.2, the translinear loop for Q-tuning and the in-channel bias distribution circuits; Section 4.2.3, the programmable capacitive attenuator; Section 4.2.4, the source-follower-based asymmetrical BPF; Section 4.2.5, the ADM with adaptive self-oscillating comparison; Section 4.3 describes the system design considerations; Section 4.4 presents the measurement results.

4.1 Architecture

The whole system of the silicon cochlea is illustrated in Figure 4.1. Binaural audio signals are acquired and amplified by off-chip microphones and preamplifiers, respectively. The on-chip part consists of the 0.5-V core and the 1.8-V AER. The AER could be designed with a 0.5-V supply as well, but we did not pursue this because of limited time budget during tape-out. The chip request C_{req}, chip acknowledge C_{ack}, as well as the channel addresses are communicated between the AER and the off-chip FPGA. The 1-D AER contains an address encoder, an arbiter logic block and a binary-tree arbiter [84]. The address bus has 8 bits: the LSB $\mathbf{b}_{\mathbf{AER0}}$ for the spike polarity (whether ON or OFF), $\mathbf{b}_{\mathbf{AER1}}$ for the distinction of the left or right cochlea, and $\mathbf{b}_{\mathbf{AER2}}$-$\mathbf{b}_{\mathbf{AER7}}$ for the 64 channels. The biases of the 64 channels I_{ch0}~I_{ch63} are geometrically scaled with a designed ratio of 1.108 covering the frequency range from about 32 to 20k Hz. In each channel, the input signal passes through a programmable capacitive attenuator before being filtered by a 4th-order source-follower-based LPF. The subsequent programmable-gain amplifier (PGA) sums the output of the LPF with its internal nodes to create the zero for a BPF transfer function (TF). The central frequency of the BPF f_{ci} is proportional to the channel bias I_{chi} ($i \in \mathcal{N}_0 \cup i \in [0, 63]$). The quality factor Q of

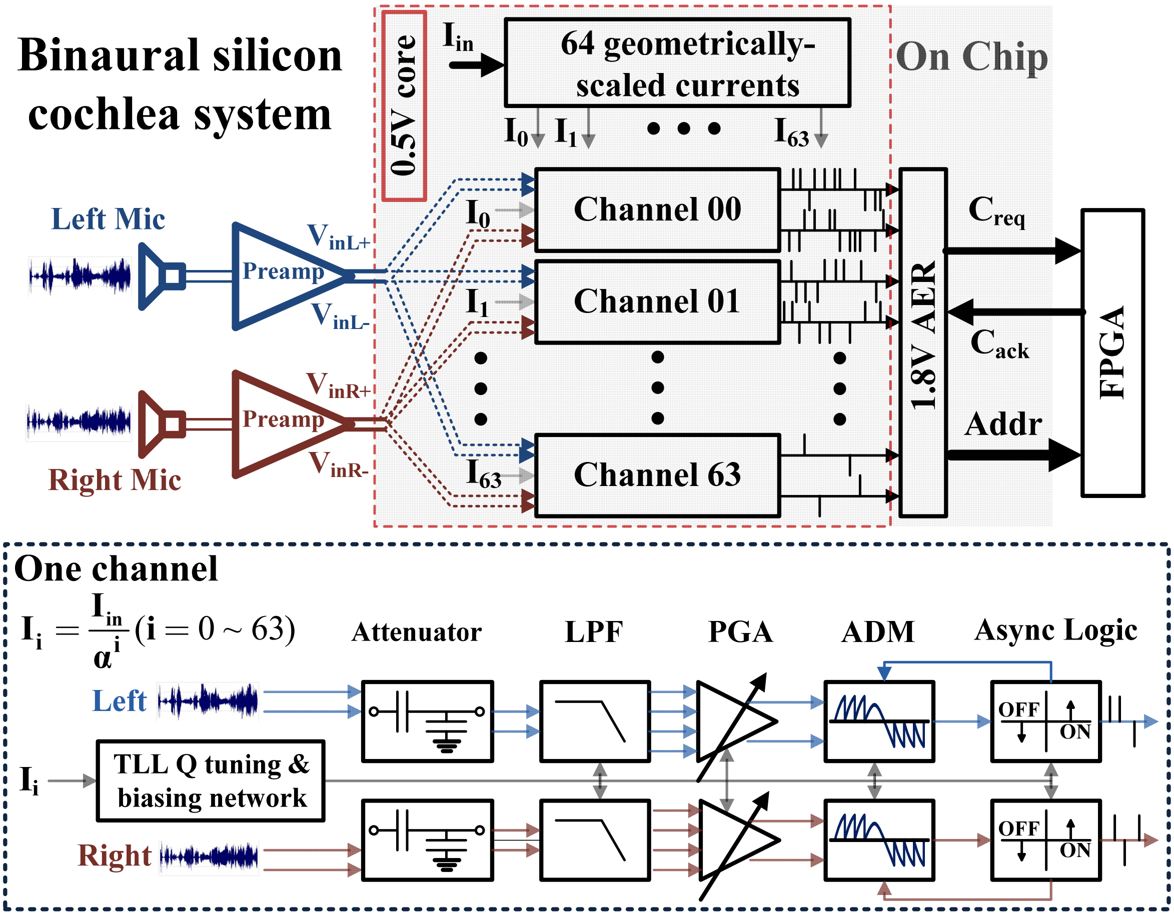


Figure 4.1. The complete binaural silicon cochlea system. The microphones, preamplifiers and FPGA are off-chip, and the on-chip part consists of the 0.5-V cochlea core with 64×2 binaural channels and the 1.8-V AER. Each cochlea channel contains the left and right monaural branches, and the building blocks are the programmable capacitive attenuator, 4th-order LPF, summing programmable gain amplifier (PGA), asynchronous delta modulator (ADM), asynchronous logic, shared translinear loop (TLL) for *Q*-tuning, and in-channel bias distribution circuitry.

the BPF is tunable via a translinear loop. The filtered signal is modulated by an ADM with self-oscillating comparison whose oscillating frequency is adaptive in accordance with the spike output activity. The asynchronous logic generates the control signals for the ADM and communicates with the 1.8-V AER abiding by the fourth-phase handshake protocol.

4.2 0.5-V Cochlea Core Design

4.2.1 64 Geometrically-Scaled Channel Bias Currents

The original cochlea design used subthreshold MOS transistors to produce the biases for the second-order section (SOS) BPFs [23], [24]. The transistor gate voltages are linearly scaled by a resistive divider so that the drain currents are geometrically scaled because of the exponential V_{gs}-I_{ds} relationship. With specified frequency range and scaling ratio, the voltages on the two terminals of the resistive divider can be determined. Besides the problem of large I_{ds} deviation resulted from the mismatch of MOS threshold

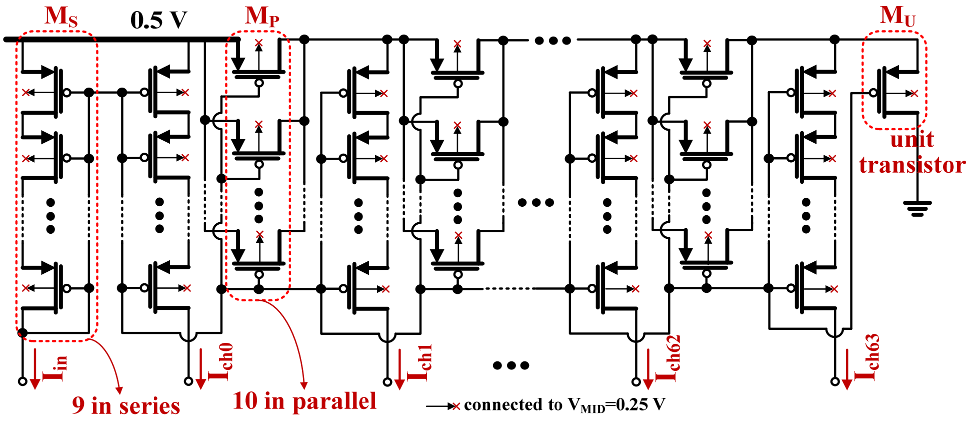


Figure 4.2. Schematic of the circuit that generate4 64 geometrically-scaled bias currents.

Table 4.1. Mean μ and 3σ variation of some of the output currents in Figure 4.2 in a 250-run Monte Carlo simulation.

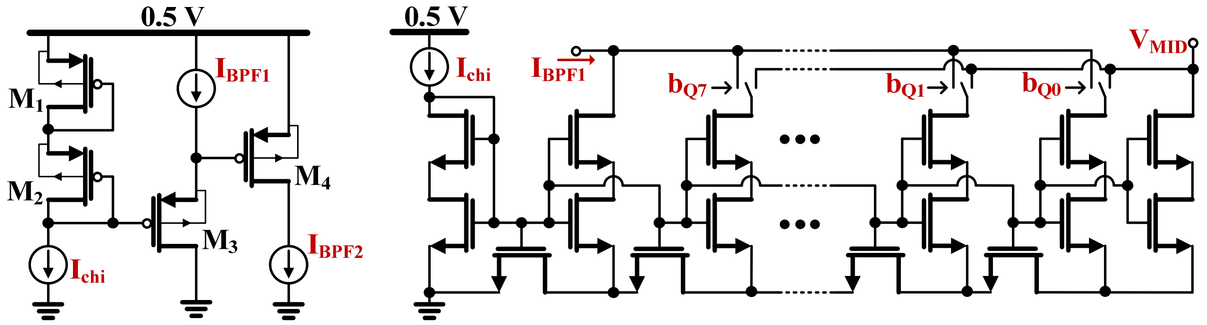
Current	I_{ch0}	I_{ch8}	I_{ch16}	I_{ch24}	I_{ch32}	I_{ch40}	I_{ch48}	I_{ch56}	I_{ch63}
μ (A)	50.1n	22.0n	9.63n	4.21n	1.83n	798p	346p	152p	76.6p
3σ (A)	273p	122p	53.6p	23.2p	11.5p	4.54p	2.27p	966f	684f
3σ/μ (%)	0.54	0.55	0.56	0.55	0.63	0.57	0.66	0.64	0.89

voltages, a large area of resistors is needed to minimize the power consumption. The required two additional reference voltages can be defined by two diode-connected transistors with two defined currents flowing through each of them to counteract the global corner die-to-die threshold voltage variation; however, two buffers are needed to supply the current flowing on the resistive divider. To mitigate the problem of large I_{ds} deviation using subthreshold MOS transistors, CMOS compatible lateral BJTs were used as the alternative owing to the well-matched device parameters of BJTs [25], and consequently the monotonicity of the central frequency scaling was greatly improved. Nonetheless, the base-emitter junction voltage V_{BE} of a BJT is usually around 0.6~0.7 V that is already larger than the targeted supply voltage of 0.5 V, not to mention that the base current causes deviation of linear V_{BE} scaling with resistive divider providing the base voltage.

Another approach that uses only MOS transistors is to employ the MOS-based current-splitting technique as described in [128], [129] for geometrical current scaling. The complete circuit is illustrated in Figure 4.2. All the unit pFETs have the same width and length. The numbers of vertical pFETs in series (lumped as M_S) and horizontal pFETs in parallel (lumped as M_P) are designed to be 9 and 10, respectively, and the chain is terminated with a single unit pFET M_U. The choice of 9 and 10 is justified as follows. As explained in [129], for a scaling ratio of r, the size ratio R_{SP} of M_S to M_P should be:

$$R_{SP} = \frac{(r-1)^2}{r}$$

With a frequency range of 32 to 20k Hz for 64 channels, $r \approx 1.108$, and in turn $R_{SP} \approx 0.011$. The arrangement of 9 vertical in series and 10 horizontal in parallel gives a size ratio of $1/90 \approx 0.011$, the same as the calculated R_{SP}. For correct operation of the current divider, the V_{ds}'s of M_S and M_P need to be larger than 100 mV so that they can stay in saturation. The body voltages of the pFETs are all biased at half VDD, i.e.



(a) (b)

Figure 4.3. (a) Translinear loop (TLL) that generates the BPF biases I_{BPF1} and I_{BPF2} whose multiplication equals to the square of I_{chi}; (b) R-2R current DAC that adjusts I_{BPF1} with 8 programming digital bits.

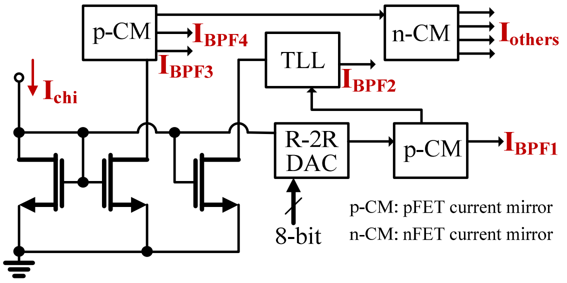


Figure 4.4. Complete in-channel bias distribution network.

250 mV to relax the *W*/*L* ratio with a gate voltage of larger than 200 mV. Given the theoretical *r* of 1.108, the ideal incremental of neighboring central frequencies is 10.8%, which limits the variation of I_{chi} to be at least less than ±5.4% to avoid crossover of central frequencies. Considering the variation accumulatively contributed by later stages in current copying, the budget at this stage should be much less. With a *W*/*L* of 200/2 in μm for a unit pFET and I_{in}=50 nA, the mean μ and 3σ variation of some of the output currents obtained from a 250-run Monte Carlo simulation are listed in Table 4.1. As I_{chi} decreases, 3σ/μ generally increases, and the maximum is within 0.89%, about 1/6 of 5.4%.

4.2.2 Translinear Loop for *Q*-Tuning and In-Channel Bias Distribution

As will be shown in Section 4.2.4, the *Q*-tuning of the BPFs needs two bias currents I_{BPF1} and I_{BPF2} whose ratio is tunable and multiplication stays constant. A translinear loop (TLL) as shown in Figure 4.3(a) can fulfill the requirement. M_1~M_4 are in subthreshold and saturation, and their bulks are connected to their own sources. The currents should ideally satisfy the following TLL equation:

$$I_{chi}^2 = I_{BPF1} \cdot I_{BPF2}$$

I_{BPF1} is generated by the circuit in Figure 4.3(b), a 9-bit R-2R current DAC. I_{BPF1} can be expressed as:

$$I_{BPF1} = I_{chi}\left(2^{-1} + \sum_{i=0}^{7} b_{Q(7-i)} 2^{-(i+2)}\right)$$

where $\mathbf{b}_{\mathbf{Qi}}$ ($i \in \mathbb{N}_0 \cup i \in [0, 7]$) are the 8 digital bits for programming I_{BPF1}. I_{BPF2} can be accordingly generated from the TLL circuit. Note that even though the DAC has a 9-bit resolution, the number of digital pro-

Table 4.2. Mean μ and 3σ variation of the multiplication and division of I_{BPF1} and I_{BPF2} in some of the channels in a 250-run Monte Carlo simulation.

	channel	0	8	16	24	32	40	48	56	63
$\sqrt{I_{BPF1}\times I_{BPF2}}$ $\mathbf{B_Q}$=11111111	μ (A)	50.2n	22.0n	9.65n	4.22n	1.84n	800p	347p	152p	77.0p
	3σ (A)	560p	236p	103p	46.6p	22.0p	9.18p	3.95p	1.80p	904f
	3σ/μ (%)	1.1	1.1	1.1	1.1	1.2	1.1	1.1	1.2	1.2
I_{BPF2}/I_{BPF1} $\mathbf{B_Q}$=11111111	μ	0.982	0.986	0.989	0.991	1.01	1.01	1.01	1.00	1.00
	3σ	26.6m	27.7m	27.1m	28.6m	24.7m	23.4m	23.6m	23.7m	25.4m
	3σ/μ (%)	2.7	2.8	2.7	2.9	2.4	2.3	2.3	2.4	2.5
$\sqrt{I_{BPF1}\times I_{BPF2}}$ $\mathbf{B_Q}$=00101111	μ (A)	50.1n	22.0n	9.63n	4.21n	1.83n	798p	346p	152p	76.8p
	3σ (A)	556p	234p	102p	46.1p	21.9p	9.13p	3.92p	1.79p	899f
	3σ/μ (%)	1.1	1.1	1.1	1.1	1.2	1.1	1.1	1.2	1.2
I_{BPF2}/I_{BPF1} $\mathbf{B_Q}$=00101111	μ	2.79	2.80	2.81	2.81	2.85	2.85	2.84	2.84	2.83
	3σ	78.8m	85.3m	84.2m	88.1m	73.0m	70.7m	70.9m	71.9m	75.2m
	3σ/μ (%)	2.8	3.0	3.0	3.1	2.6	2.5	2.5	2.5	2.7

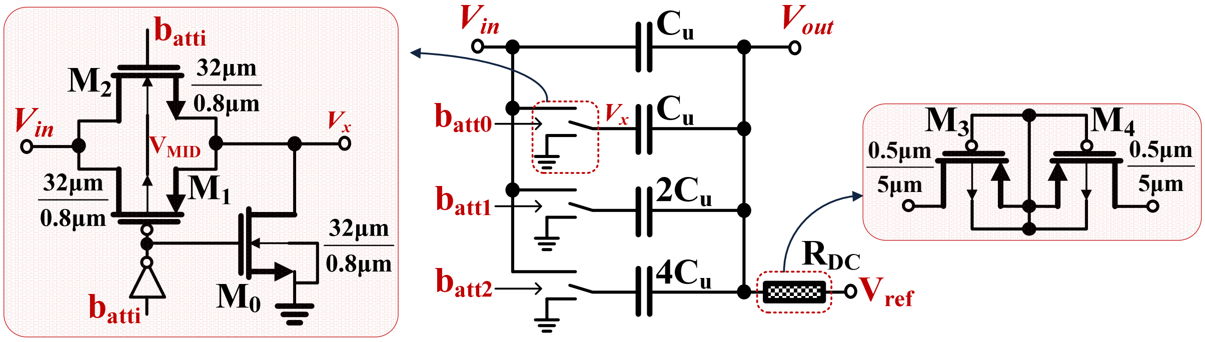


Figure 4.5. 3-bit programmable capacitive attenuator.

gramming bits is 8, and in turn I_{BPF1} is always larger than $0.5I_{chi}$. The design rationale is that I_{BPF1} less than $0.5I_{chi}$ is not needed for *Q* larger than 1 as will be clear in Section 4.2.4. The simplified complete in-channel bias distribution network including the TLL and the R-2R DAC is shown in Figure 4.4. The I_{chi} is directly the current generated by the current-splitting circuit in Figure 4.2. I_{BPF3} and I_{BPF4} are for the second filter biquad, and I_{others} are distributed to the other building blocks including the PGA and ADM.

Monte Carlo simulation is performed to evaluate the variation of the multiplication of I_{BPF1} and I_{BPF2}, and the division of I_{BPF2} over I_{BPF1}. The results are listed in Table 4.2. Comparing with the channel biases in Table 4.1, the largest deviation of the square root of $I_{BPF1}\times I_{BPF2}$ is about 0.55% at channel 32 with $\mathbf{B_Q}$=11111111. The 3σ variation is within 1.2%, and together with the variation of the channel bias I_{chi} itself, the total 3σ variation is within 1.5% at the worst case channel 63, which is still below 1/3 of 5.4%. The ratio I_{BPF2}/I_{BPF1} is related to the *Q* of the BPFs, and its variation is more important for high *Q* settings because the central frequency gain and *Q* value are more sensitive to I_{BPF2}/I_{BPF1} variation as will be discussed in Section 4.2.4. $\mathbf{B_Q}$=00101111 usually sets a high *Q*, and the worst 3σ variation is about 3.1%.

4.2.3 Programmable Capacitive Attenuator

A fixed capacitive attenuator has been used in a prior bionic ear chip to increase the maximum allowable input amplitude with a 5% output THD [28], and in wireless transceivers programmable resistive at-

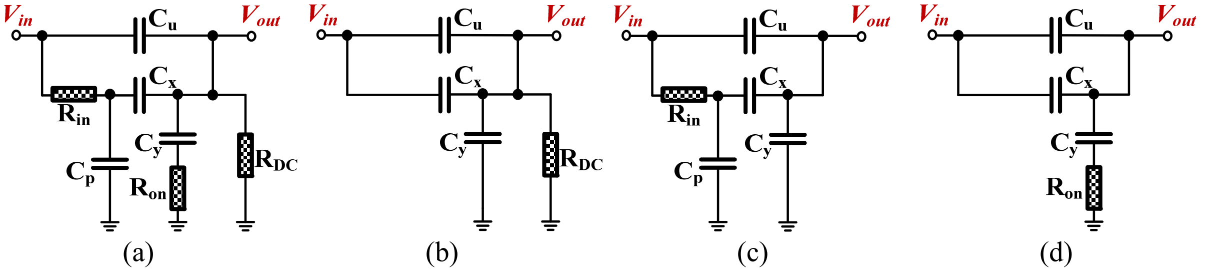


Figure 4.6. (a) The complete RC-equivalent circuit of the attenuator including all important parasitics; (b) consider only the pseudo-resistor R_{DC}; (c) consider only the ON-resistance and parasitic capacitance when switches are connected to V_{in}; (d) consider only the ON-resistance when switches are connected to ground.

tenuators are used to limit signal power in both the receiving and transmitting paths [130]. In this work, a programmable capacitive attenuator is used before the main BPF to allow flexible linearity control in accordance with the input audio power. The circuit schematic is shown in Figure 4.5. 8 attenuation levels from 0 dB to -18 dB are selected by 3 digital bits $\mathbf{b}_{\mathbf{atti}}$ (i=[1,2,3]). Because the output V_{out} is directly connected to the gate of the BPF input transistor, a pseudo-resistor R_{DC} is used to set the output DC level to V_{ref} whose generation will be discussed in Section 4.2.4. R_{DC} is composed of two pFETs M_3 and M_4 with the gates and bulks connected to their own sources. The switches are connected to the left plates of the capacitors because their leakage currents can pull V_{out} to largely deviate from V_{ref} if they are at the same side of R_{DC}.

The RC-equivalent circuit considering the parasitic resistance (R_{in} and R_{on}) and capacitance (C_p) of the switches is shown in Figure 4.6(a). The capacitors and R_{DC} form a highpass filter whose corner frequency needs to be at least lower than the lowest central frequency in cochlea channels, i.e. 32 Hz. The resistance of R_{DC} is positively correlated with the W/L of the pFETs. A conservative design with W/L=0.5/5 in µm gives a resistance that is larger than 56G ohm at room temperature. With the simplified circuit in Figure 4.6(b), the highpass corner frequency f_{c1} is computed as:

$$f_{c1} = \frac{1}{2\pi \times 8 C_u R_{DC}}$$

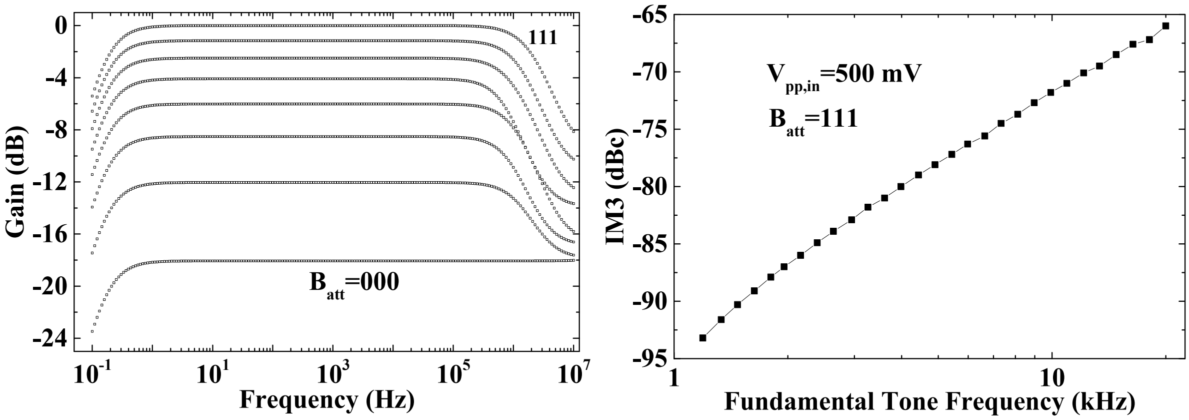
With a relaxed f_{c1}=0.4 Hz«32 Hz, C_u is calculated to have the capacitance of about 910 fF which occupies an area of 30×30 μm^2. Larger L of the pFET in R_{DC} results in larger resistance, and consequently the capacitance and area of C_u can be reduced.

When switches are connected to V_{in}, the pole/zero pair caused by the parasitic resistance R_{in} and capacitance C_p limits the attenuator bandwidth. The simplified circuit in Figure 4.6(c) assumes that the individual switch resistances and the capacitors between V_{in} and V_{out} are lumped together. Let $C_x = a \cdot C_u$, the pole and zero are located at:

$$\omega_p = \frac{1}{(C_p + aC_u - a^2 C_u / 8) R_{in}}, \quad \omega_z = \frac{a+1}{(C_p + aC_u) R_{in}}$$

With $C_p \ll C_u$, the lowest ω_p is obtained when a=4, and the value of ω_p and ω_z are:

$$\omega_p = \frac{1}{2 C_u R_{in}}, \quad \omega_z = \frac{1}{0.8 C_u R_{in}}$$



(a) (b)

Figure 4.7. (a) AC characteristics of the attenuator. Eight gain levels have flat response within the cochlea passband; (b) the attenuator's output IM3 at V_{in}=1 V_{pp} and $\mathbf{B}_{att}$=111.

ω_p should be larger than the highest cochlea channel frequency 20k Hz, which gives the lower limit of the W/L ratio of the complimentary pFET M_1 and nFET M_2 whose bulk voltages are biased at V_{mid}=0.25 V to further reduce the equivalent resistance. Another factor that constrains the ON-resistance of the switch is distortion. The odd-order harmonics mainly the 3rd-order are of concern because the even-orders are largely suppressed thanks to the differential input. The theoretical analysis in [130] shows that with a given input amplitude, the 3rd-order intermodulation resulted from a single transistor switch is proportional to its cubic ON-resistance. Usually a 5% output THD is targeted in applications like cochlea implant [28], [29], and it corresponds to only a -16 dBc IM3 if HD3 is the main distortion component. In biological basilar membrane, the IM3 can decrease from about -17 dBc at 30 SPL fundamental tones spaced with a 1.1 ratio to less than -50 dBc at 90 SPL [131]. In this design, an IM3<-60 dBc at a 20k Hz differential input of V_{in}=1 V_{pp} and $\mathbf{B}_{att}$=111 (no attenuation) is specified not to limit the system performance, and with the sizes of M_1 and M_2 shown in Figure 4.5, an IM3=-66 dBc is obtained in simulation at room temperature and slow process corner. IM3 decreases as input frequency decreases because it is inversely proportional to the cubic capacitor impedance [130].

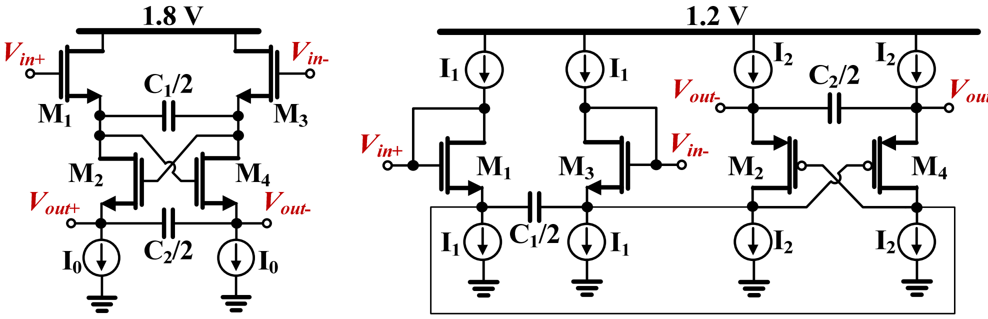
When switches are connected to ground via nFET M_0, the ON-resistance R_{on} in series with the capacitor C_y as shown in Figure 4.6(d) also creates a pole/zero pair which are located at:

$$\omega_p = \frac{1}{(\frac{1}{8} + \frac{C_x}{8C_u})C_y R_{on}}, \ \omega_z = \frac{1}{C_y R_{on}}$$

It is obvious that $\omega_z \le \omega_p$, and hence R_{on} should be sufficiently low so that $\omega_z/2\pi$ is at least larger than 20 kHz. M_0 is chosen to have the same size of M_2 which can guarantee a sufficiently small R_{on}. To summarize the attenuator design, the simulation results of the transfer function and IM3 as the function of input signal frequency with a V_{in}=1 V_{pp} and $\mathbf{B}_{att}$=111 are plotted in Figure 4.7.

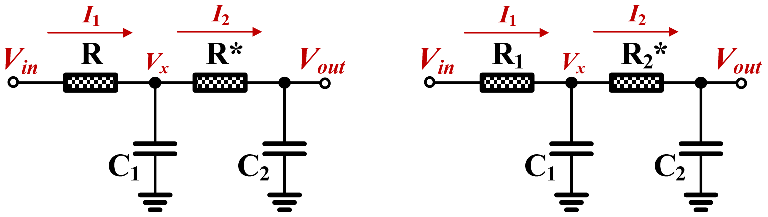
4.2.4 Source-Follower-Based Asymmetrical Bandpass Filter

Many types of active filters have been proposed for various applications ranging from wireless and wireline communication to biomedical signal acquisition. Opamp-RC, g_m-C and switched-cap are among the most widely used types, and more recently developed ones including the ring-oscillator-based [132]



(a) (b)

Figure 4.8. (a) The original SF-based LPF biquad with local positive feedback via nFETs M_2 and M_4, the biases for the nFET pairs M_1/M_3 and M_2/M_4 are the same, i.e. I_0 [134], [135]; (b) the alternative SF-based LPF biquad with separate biases I_1 and I_2 for nFETs M_1/M_3 and pFETs M_2/M_4 respectively [136], [137].



(a) (b)

Figure 4.9. The RC-equivalent circuits for (a) the SF-based LPF biquad in Figure 4.8(a) and (b) the SF-based LPF biquad in Figure 4.8(b).

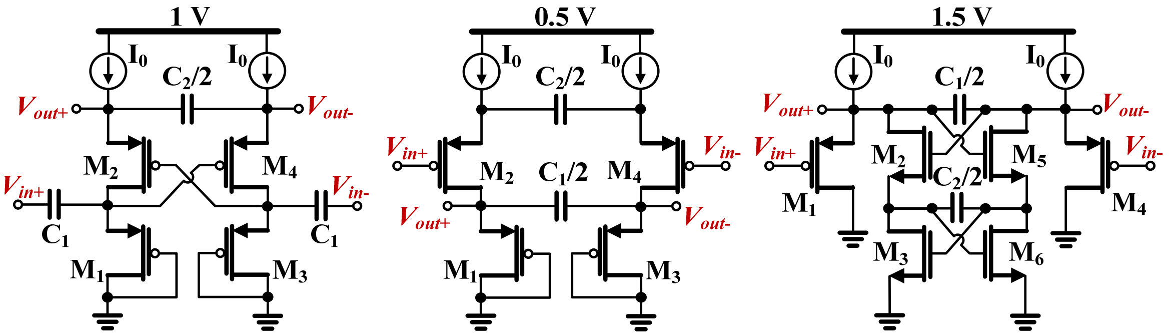
and switched-Opamp-based [133] specifically aim for systems with low supply voltages down to 0.55 V. If power efficiency is the most paramount requirement in system design, the source-follower-based (SF-based) filter is the best choice because of its simple topological configuration using minimal number of transistors. The patented SF-based LPF biquads are depicted in Figure 4.8. The first version of a 2nd-order biquad proposed in 2006 is shown in Figure 4.8(a) [134], [135]. The nFET pairs M_1/M_3 and M_2/M_4 form the composite source follower structure. To avoid gain loss, the bulks of the nFETs should be connected to their sources, which requires deep-Nwell in a p-substrate process. The bias current I_0 and the capacitors C_1 and C_2 determine the LPF cutoff frequency. The nFETs can be equivalently seen as resistors which have the same small-signal resistance as in Eq. (4.1) if the nFETs are in weak inversion:

$$R = 1/g_m = \frac{V_T}{\kappa I_0} \tag{4.1}$$

where g_m is the transconductance, V_T is the thermal voltage and κ is the subthreshold slope factor. Note that all the transconductance symbols below like g_m or g_{mi} $(i \in \mathcal{N}_0)$ represent the reciprocal of the corresponding resistance symbols like R or R_i. The RC-equivalent circuit is illustrated in Figure 4.9(a). The equations related to the small-signal currents I_1 and I_2 are:

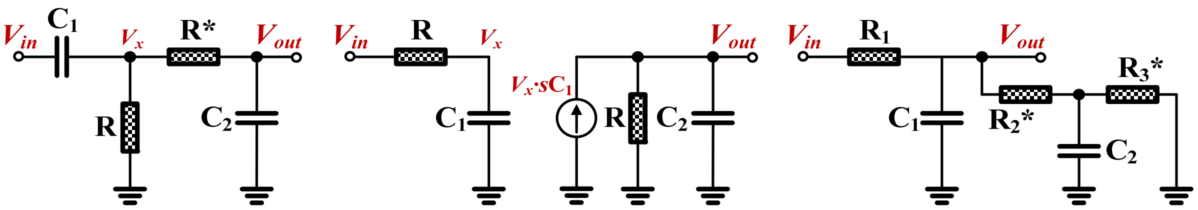
$$I_1 = \frac{V_{in} - V_x}{R} = I_2 + V_x \cdot sC_1,\ I_2 = -\frac{V_x + V_{out}}{R*} = V_{out} \cdot sC_2$$

Note that R* has the same value as R, and the unusual expression of I_2 is due to the local positive feedback in the cross-coupled transistors. The lowpass transfer function can be then derived as:



(a) (b) (c)

Figure 4.10. The SF-based BPF biquads in (a) [141]; (b) [142]; (c) [143].



(a) (b) (c)

Figure 4.11. The RC-equivalent circuits of the BPF biquads in Figure 4.10.

$$H_{out}(s) = \frac{V_{out}(s)}{V_{in}(s)} = -\frac{1}{s^2 \cdot \frac{C_1 C_2}{g_m^2} + s \cdot \frac{C_1}{g_m} + 1} \tag{4.2}$$

Thanks to the local positive feedback, synthesis of complex poles becomes possible, and the quality factor can exceed 0.5 which allows steeper roll-off. This composite SF-based biquad was used to build a 4th-order Bessel LPF which was fabricated in 0.18 μm CMOS and had a 10 MHz cutoff frequency and 79-dB dynamic range with a 4.1 mW power consumption.

An alternative way of constructing the biquad is shown in Figure 4.8(b), which was published in 2008 [136], [137]. Instead of stacking the normal nFET pair on top of the cross-coupled nFET pair, the two pairs are separated in two branches, and hence their bias currents do not have to be the same, although in [136] the same bias currents were used. Even though this topology does not have the current-reuse advantage as the one in Figure 4.8(a), the flexibility of choosing different biasing currents for M_1/M_3 and M_2/M_4 makes Q tuning easy as will be evident later. With the aid of the RC-equivalent circuit in Figure 4.9(b), the lowpass transfer function can be derived as:

$$H_{out}(s) = -\frac{1}{s^2 \cdot \frac{C_1 C_2}{g_{m1} g_{m2}} + s \cdot (\frac{C_1}{g_{m1}} + \frac{C_2}{g_{m2}} - \frac{C_2}{g_{m1}}) + 1} \tag{4.3}$$

Three of this biquad were cascaded to build a 6th-order LPF for UWB applications, which was fabricated in 0.13 μm CMOS and had a 280 MHz cutoff frequency with a 120 μW power consumption. Compared to the state-of-the-art g_m-C LPFs with large bandwidth [138]–[140], the SF-based solution retains the lowest power/pole/Hz with only slightly inferior linearity.

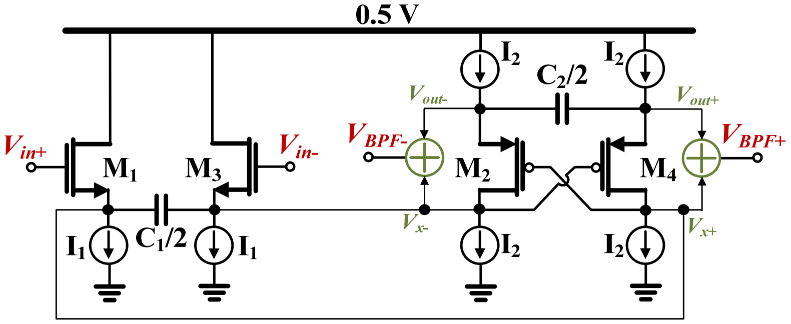


Figure 4.12. Proposed SF-based BPF biquad.

Bandpass filtering is needed in silicon cochlea for channel frequency division, and the previous works on building SF-based BPFs are first briefly summarized here [141]–[143]. The transistor implementations of the biquads and the corresponding RC-equivalent circuits are depicted in Figure 4.10 and Figure 4.11, respectively. The biquads (a) and (b) were for low-frequency biomedical applications, and (c) was for MHz-range IF baseband in heterodyne receivers. Their transfer functions can be derived as follows:

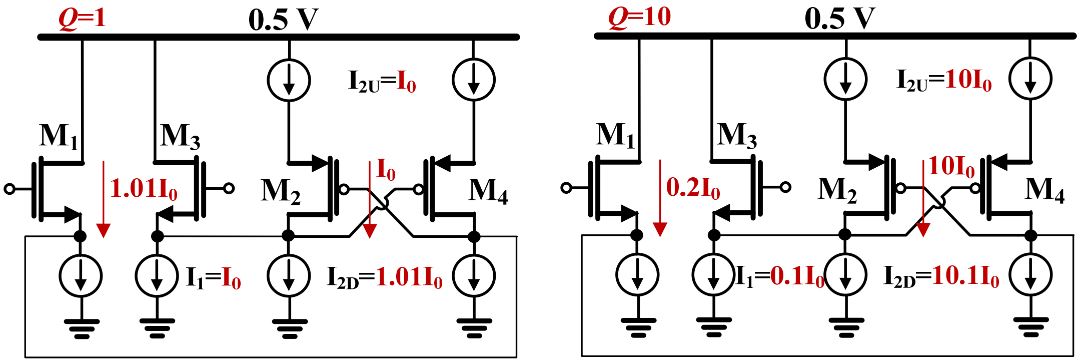
$$H_a(s) = -\frac{s\cdot\frac{C_1}{g_m}}{s^2\cdot\frac{C_1C_2}{g_m^2} + s\cdot\frac{C_1}{g_m} + 1},\quad H_b(s) = -\frac{s\cdot\frac{C_1}{g_m}}{s^2\cdot\frac{C_1C_2}{g_m^2} + s\cdot(\frac{C_1}{g_m} + \frac{C_2}{g_m}) + 1},$$

$$H_c(s) = \frac{\frac{g_{m1}}{C_1}(s + \frac{g_{m2} - g_{m3}}{C_2})}{s^2 + s\cdot(\frac{g_{m1} - g_{m2}}{C_1} + \frac{g_{m2} - g_{m3}}{C_2}) + \frac{g_{m1}g_{m2} - g_{m1}g_{m3} + g_{m2}g_{m3}}{C_1C_2}}$$

The BPF biquad in Figure 4.10(a) has a straightforward RC-prototype as shown in Figure 4.11(a): a 1^{st}-order HPF cascaded by a 1^{st}-ordre LPF. The cross-coupled pFETs M_2 and M_4 allow the synthesis of complex poles, and in turn a Q larger than 0.5. Q is equal to the square root of the capacitance ratio C_2/C_1, and for a Q=10, the ratio needs to be as large as 100, which is not area-efficient. Also as will be discussed later, large Q by large capacitance ratio results in its impractically high relative sensitivity to the variation of the g_m ratio of M_2/M_4 over M_1/M_3. Higher-order BPFs can be built by repeatedly cascading the same biquads; however each cascading causes about 6-dB gain loss in the passband due to the finite input impedance which is frequency dependent. The problem could be solved by inserting extra 2T source followers between biquads. The BPF biquad in Figure 4.10(b) does not have the gain loss problem in cascading owing to the gate input, but the lack of negative g_m prevents the possibility of a Q larger than 0.5. The biquad in Figure 4.10(c) can both have Q>0.5 and avoid gain loss in cascading, nevertheless besides the complicated form of the transfer function, the major problem is that the values of g_{m1}, g_{m2} and g_{m3} are not well-defined by bias currents and are dependent on the common-mode input voltage. This makes its transfer function very sensitive to any common mode fluctuation from input and power supply.

4.2.4.1 Proposed Source-Follower-Based BPF Biquad

The proposed SF-based BPF biquad used in cochlea channels are based on the LPF biquad in Figure 4.8(b). As depicted in Figure 4.12, this biquad has the same small-signal RC-equivalent circuit as in Figure 4.9(b), except that the final output is not V_{out} but the summation of V_x and V_{out}. The transistor pairs M_1/M_3 and M_2/M_4 all have the bodies connected to their own sources to prevent gain loss, which requires



(a) (b)

Figure 4.13. The impact of current source mismatch at (a) Q=1 and (b) Q=10.

deep-nWell for nFETs. The transfer function of V_x can be derived as:

$$H_x(s)=\frac{s\cdot\frac{C_2}{g_{m2}}+1}{s^2\cdot\frac{C_1C_2}{g_{m1}g_{m2}}+s\cdot\left(\frac{C_1}{g_{m1}}+\frac{C_2}{g_{m2}}-\frac{C_2}{g_{m1}}\right)+1} \tag{4.4}$$

It is easy to obtain the bandpass transfer function by summing Eq. (4.3) and Eq. (4.4):

$$H_{BPF}(s)=H_{out}(s)+H_x(s)=\frac{s\cdot\frac{C_2}{g_{m2}}}{s^2\cdot\frac{C_1C_2}{g_{m1}g_{m2}}+s\cdot\left(\frac{C_1}{g_{m1}}+\frac{C_2}{g_{m2}}-\frac{C_2}{g_{m1}}\right)+1} \tag{4.5}$$

The central frequency f_0, quality factor Q and passband gain K are written as:

$$f_0=\frac{1}{2\pi}\sqrt{\frac{g_{m1}g_{m2}}{C_1C_2}},\ Q=\frac{\sqrt{\frac{g_{m1}g_{m2}}{C_1C_2}}}{\frac{g_{m1}}{C_1}+\frac{g_{m2}}{C_2}-\frac{g_{m2}}{C_1}},\ K=\frac{\frac{g_{m1}}{C_1}}{\frac{g_{m1}}{C_1}+\frac{g_{m2}}{C_2}-\frac{g_{m2}}{C_1}} \tag{4.6}$$

In the simplest case, the same capacitance C_0 is used for both C_1 and C_2, as in the case of the SOS-BPF in the original cochlea design [23]. The transfer function and the characteristic parameters are simplified as:

$$H_{BPF}(s)=\frac{s\cdot\frac{C_0}{g_{m2}}}{s^2\cdot\frac{C_0^2}{g_{m1}g_{m2}}+s\cdot\frac{C_0}{g_{m2}}+1},\ f_0=\frac{1}{2\pi}\frac{\sqrt{g_{m1}g_{m2}}}{C_0},\ Q=\sqrt{\frac{g_{m2}}{g_{m1}}},\ K=1$$

Or the same bias current I_0 is used for both I_1 and I_2, as in the case of the original SF-based LPF in Figure 4.8(a). The transfer function and the characteristic parameters are simplified as:

$$H_{BPF}(s)=\frac{s\cdot\frac{C_2}{g_{m0}}}{s^2\cdot\frac{C_1C_2}{g_{m0}^2}+s\cdot\frac{C_1}{g_{m0}}+1},\ f_0=\frac{1}{2\pi}\frac{g_{m0}}{\sqrt{C_1C_2}},\ Q=\sqrt{\frac{C_2}{C_1}},\ K=\frac{C_2}{C_1}$$

Although these two cases provide simple expressions of parameters which can be easily calculated, the following analysis will show that they cannot be practically employed in designs where relatively high Q, e.g. Q=10, is required. A design methodology is proposed to circumvent the problems.

First, the case of equal capacitance is considered. The main problem is identified as the mismatch of the upper and lower current sources (I_{2U} and I_{2D}) in the cross-coupled pFET branch, as illustrated in Figure 4.13. For simplicity, the capacitors and summation symbols are neglected. Assume the g_m values of nFETs and pFETs are only functions of the bias currents. When Q=1, ideally $I_{2U}=I_{2D}=I_1=I_0$. If a 1% mismatch exists between I_{2U} and I_{2D}, e.g. $I_{2D}=1.01I_{2U}$ which is already not easy to achieve in MOS process, the currents flowing through the nFET and pFET pairs are $1.01I_0$ and I_0, respectively. The central frequency considering the mismatch f_{0mis} is calculated as:

$$f_{0mis}=\frac{1}{2\pi}\frac{\sqrt{1.01g_{m0}\cdot g_{m0}}}{C_0}=1.005f_0$$

which means the actual central frequency is only 0.5% off the targeted value without mismatch. When Q=10, ideally $I_{2U}=I_{2D}=10I_0$ and $I_1=0.1I_0$. Still assume $I_{2D}=1.01I_{2U}$. The currents flowing through the nFET and pFET pairs are $0.2I_0$ and $10I_0$, respectively. The central frequency considering the mismatch f_{0mis} is calculated as:

$$f_{0mis}=\frac{1}{2\pi}\frac{\sqrt{0.2g_{m0}\cdot 10g_{m0}}}{C_0}=1.414f_0$$

which means the actual central frequency is 41.4% off the targeted value without mismatch. This is unacceptable because as mentioned in Section 4.2.1, the variation of central frequency should be limited within ±5.4%.

To circumvent the I_{2U}/I_{2D} mismatch problem, let us consider a more general case and write the capacitance and transconductance in the form below:

$$C_1=\frac{C_0}{k},\ C_2=kC_0,\ g_{m1}=\frac{g_{m0}}{n},\ g_{m2}=ng_{m0},\ (k,n\in R\bigcap\geq 1) \tag{4.7}$$

The characteristic parameters in Eq. (4.6) can be then rewritten as:

$$f_0=\frac{g_{m0}}{2\pi C_0},\ Q=\frac{1}{\frac{k}{n}+\frac{n}{k}-nk},\ K=\frac{1}{1+\frac{n^2}{k^2}-n^2} \tag{4.8}$$

Let m_{is} be the mismatch of the current sources and t_{ol} the limit of tolerable relative central frequency variation. Assume $I_{2D}=(1+m_{is})\cdot I_{2U}=(1+m_{is})\cdot n\cdot I_0$, and $I_1=I_0/n$. The current flowing through M_2/M_4 is $n\cdot I_0$, and the current flowing through M_1/M_3 can be calculated as:

$$I_{2D}+I_1-n\cdot I_0=\frac{I_0}{n}+m_{is}\cdot n\cdot I_0$$

We have the following inequality:

$$\sqrt{(\frac{1}{n}+m_{is}\cdot n)\cdot n}-1\leq t_{ol}$$

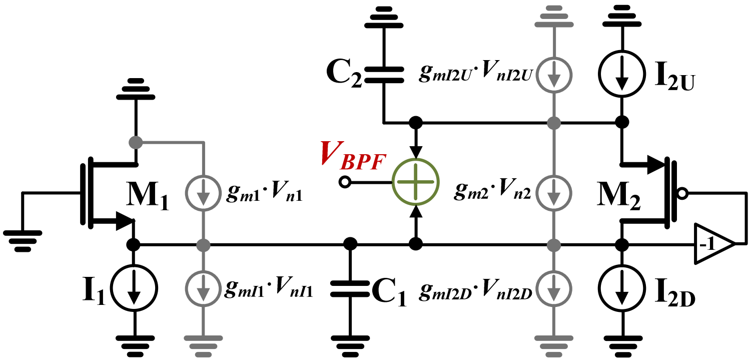


Figure 4.14. Half-circuit of the proposed SF-based BPF biquad with noise sources shown in gray.

The scale parameter n can be calculated to have the upper bound as:

$$n \le \sqrt{\frac{t_{ol}^2 + 2t_{ol}}{m_{is}}} \tag{4.9}$$

If n is determined, the scale parameter k can be accordingly calculated using the Q expression in Eq. (4.8):

$$k = \frac{n - \sqrt{n^2 - 4Q^2 n^2 (1-n^2)}}{2(1-n^2)Q} \tag{4.10}$$

However, Eq. (4.9) only gives the upper limit of n; the determination of its lower limit is related to the impractical case of equal g_m. Define r_c and r_{gm} as follows:

$$r_c = \frac{C_2}{C_1} = k^2, \; r_{gm} = \frac{g_{m2}}{g_{m1}} = n^2$$

Using the Q expression in Eq. (4.6), the relative sensitivity of Q over r_{gm} can be calculated as:

$$S = \frac{\partial Q}{\partial r_{gm}} / \frac{Q}{r_{gm}} = \frac{1}{2}\frac{r_c - r_{gm} + r_c r_{gm}}{r_c + r_{gm} - r_c r_{gm}}$$

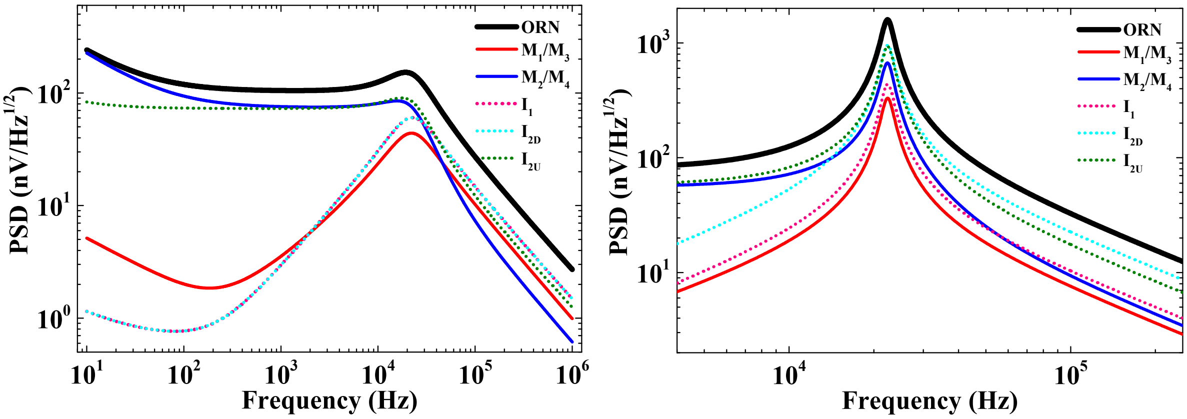
If r_c=1, i.e. the equal capacitance case, S is calculated to be 0.5 and independent of r_{gm}. With r_{gm}=100 for Q=10 and a 1% variation of r_{gm}, the variation of Q is 0.5%. If r_{gm}=1, i.e. the equal transconductance case, with r_c=100 for Q=10, S is calculated to be 99.5. This means a 1% variation of r_{gm} results in 99.5% variation of Q. The r_{gm} variation comes from both the variation of I_1 and I_2, and the parameter variation of the nFET and pFET pairs like the subthreshold slope factor. Now we can see why the equal transconductance case is impractical as well in terms of a well-controlled Q value that should be relatively insensitive to r_{gm} variation. If the upper limit of S is specified as S_0, the inequality below can be obtained:

$$(2S_0 + 1)\cdot n^2 - 2Q\cdot n + 1 - 2S_0 \ge 0$$

With $S_0 \ge 0.5$, the lower limit of n is calculated as:

$$n \ge \frac{Q + \sqrt{Q^2 + 4S_0^2 - 1}}{2S_0 + 1} \tag{4.11}$$

Therefore the value of the scale parameter n can be found between the upper and lower bounds calculated



(a) (b)

Figure 4.15. Simulated noise power spectrum density (PSD) of the proposed biquad at (a) Q=1 and (b) Q=10.

by Eq. (4.9) and Eq. (4.11), respectively. With m_{is}=0.02, t_{ol}=0.025 and S=20, the range of n at Q=10 is calculated to be $1.249 \leq n \leq 1.591$. In the fabricated design, r_c is set to be 26/15, and using Eq. (4.10) n is calculated to be 1.450 at Q=10, which is within the targeted range. This design methodology is verified by the results of Monte Carlo simulation of the complete asymmetrical BPF in Section 4.2.4.3.

4.2.4.2 Noise and Linearity

The half-circuit of the proposed SF-based BPF biquad with noise sources is shown in Figure 4.14. V_{n1} and V_{n2} are the equivalent gate noise voltages of transistors M$_1$/M$_3$ and M$_2$/M$_4$, respectively, and V_{nI1}, V_{nI2D} and V_{nI2U} are of the current sources I$_1$, I$_{2D}$ and I$_{2U}$, respectively. The noise transfer function of V_{n1}, $H_{n1}(s)$, is the same as the BPF transfer function $H_{BPF}(s)$ given in Eq. (4.5). The noise transfer functions of V_{n2}, V_{nI1}, V_{nI2D} and V_{nI2U} are denoted as $H_{n2}(s)$, $H_{nI1}(s)$, $H_{nI2D}(s)$ and $|H_{nI2U}(s)|$, respectively, and have the expressions shown below:

$$|H_{n2}(s)| = \left| \frac{s \cdot \frac{C_2 - C_1}{g_{m1}} - 1}{s^2 \cdot \frac{C_1 C_2}{g_{m1} g_{m2}} + s \cdot (\frac{C_1}{g_{m1}} + \frac{C_2}{g_{m2}} - \frac{C_2}{g_{m1}}) + 1} \right|, \quad |H_{nI1}(s)| = \left| \frac{s \cdot \frac{g_{mI1} C_2}{g_{m1} g_{m2}}}{s^2 \cdot \frac{C_1 C_2}{g_{m1} g_{m2}} + s \cdot (\frac{C_1}{g_{m1}} + \frac{C_2}{g_{m2}} - \frac{C_2}{g_{m1}}) + 1} \right|$$

$$|H_{nI2D}(s)| = \left| \frac{s \cdot \frac{g_{mI2D} C_2}{g_{m1} g_{m2}}}{s^2 \cdot \frac{C_1 C_2}{g_{m1} g_{m2}} + s \cdot (\frac{C_1}{g_{m1}} + \frac{C_2}{g_{m2}} - \frac{C_2}{g_{m1}}) + 1} \right|, \quad |H_{nI2U}(s)| = \left| \frac{s \cdot \frac{g_{mI2U} C_1}{g_{m1} g_{m2}} + \frac{g_{mI2U}}{g_{m2}}}{s^2 \cdot \frac{C_1 C_2}{g_{m1} g_{m2}} + s \cdot (\frac{C_1}{g_{m1}} + \frac{C_2}{g_{m2}} - \frac{C_2}{g_{m1}}) + 1} \right|$$

Note that in subthreshold, $g_{mI1} \approx g_{m1}$ and $g_{mI2D} \approx g_{mI2U} \approx g_{m2}$. The zeros of $|H_{n2}(s)|$ and $|H_{nI2U}(s)|$ are located at:

$$\omega_{zHn2} = \frac{g_{m1}}{C_2 - C_1}, \quad \omega_{zHn3} = \frac{g_{m1}}{C_1}$$

These zeros are either larger or comparable to the angular central frequency $\omega_0 = 2\pi f_0$ of $H_{BPF}(s)$, and therefore $|H_{n2}(s)|$ and $|H_{nI2U}(s)|$ are almost flat below ω_0, and behave similarly to a lowpass transfer function. On the other hand, $|H_{nI1}(s)|$ and $|H_{nI2D}(s)|$ has a bandpass shape that is the same as $H_{BPF}(s)$. Consequently at low Q, the major noise contributions to the output V_{BPF} are from V_{n2} and V_{nI2U}, i.e. the pFET pair M_2/M_4 and the current source pair I_{2U}. This is validated by the power spectrum density (PSD) simulation of the output referred noise (ORN) at Q=1 and f_0=20k Hz and the individual transistor noise contribution as shown in Figure 4.15(a). Note that the noise curves of I_1 and I_{2D} almost overlap. Because of the lowpass-alike behavior of $H_{n2}(s)$, the use of pFETs for M_2/M_4 is beneficial in terms of lower flick noise compared to nFETs with the same area. Be aware that all the noise simulation results in this chapter are obtained by using the process parameters from UMC even though the chip is fabricated in TowerJazz who admitted that their noise parameter extraction in subthreshold is extremely unreliable. As Q increases, the low frequency noise PSD decreases because g_{m2} increases, but the in-band noise PSD increases significantly because of gain peaking as shown in Figure 4.15(b) where Q=10. The most dominant in-band noise contributions are from the current sources I_{2U} and I_{2D}. This can be attributed to their large in-band gain. For comparison, the ratios of the central frequency gain of the noise transfer functions to the BPF's central frequency gain K given in Eq. (4.6) are listed below:

$$\frac{K_{n1}}{K}=1,\ \frac{K_{n2}}{K}=\sqrt{\frac{(C_2-C_1)^2 g_{m2}^2}{C_2^2 g_{m1}^2}+\frac{C_1 g_{m2}}{C_2 g_{m1}}},\ \frac{K_{nI1}}{K}\approx 1,\ \frac{K_{nI2D}}{K}\approx\frac{g_{m2}}{g_{m1}},\ \frac{K_{nI2U}}{K}\approx\sqrt{\frac{C_1^2 g_{m2}^2}{C_2^2 g_{m1}^2}+\frac{C_1 g_{m2}}{C_2 g_{m1}}}$$

Using the capacitance and transconductance ratios calculated in Section 4.2.4.1, the gain ratios are:

$$\frac{K_{n1}}{K}=1,\ \frac{K_{n2}}{K}=1.44,\ \frac{K_{nI1}}{K}\approx 1,\ \frac{K_{nI2D}}{K}\approx 2.16,\ \frac{K_{nI2U}}{K}\approx 1.68$$

These calculated values are approximately consistent with the simulation results. In both low and high Q cases, the dominant output noise contributions are related to the noise sources of the cross-coupled branch. The noise PSD can only be reduced by larger bias currents and in turn larger capacitor area.

It is well-known that MOSFET in above-threshold is superior compared to BJT in terms of smaller distortion because the latter have an exponential I-V characteristics while the former generates no 3rd-order distortion if the ideal square law is assumed [106]. However, with a 0.5-V power supply, the transistor pairs M_1/M_3 and M_2/M_4 work in deep-subthreshold, and the I-V curve becomes exponential, resembling BTJ's. While the SF-based LPFs with above-MHz cutoff frequencies for transceiver applications can easily have less than -40 dBc IM3 with hundreds of mV input amplitude especially when the signal frequency is far below the cutoff frequency [134], [136], [144], the kHz-range SF-based BPFs usually exhibit merely around -30 dBc IM3 with tens of mV input signals at low Q [141], [142], and the input is more limited at high Q because of increased passband gain. The inferior linearity performance in those low-frequency BPFs is attributed to both subthreshold operation of MOSFETs which is the same reason for poor linearity of low-frequency biomedical g_m-C filters [145], and the restricted effective operating frequency range where the nonlinear V-I conversion occurs more than the case of SFB-LPF at frequencies well below its cutoff frequency. In other words, at low frequencies, the intrinsic negative feedback due to the current source in a source follower allows the input transistor to have negligible voltage-to-current conversion which is the source of nonlinearity, and the input and output signals are in phase; as input frequency increases, the loading capacitance at the output reduces the effective impedance of the current source, so the voltage-to-current conversion of the input transistor becomes significant, and the phase difference between input and output signals increases. Nevertheless, our biological basilar membrane in cochlea also has only around -30 dBc IM3 at 60 SPL input which is the loudness of our normal speech [131], and we still perform recognition task well. This of course relies on the functions of the intricate neural networks in our auditory cortex.

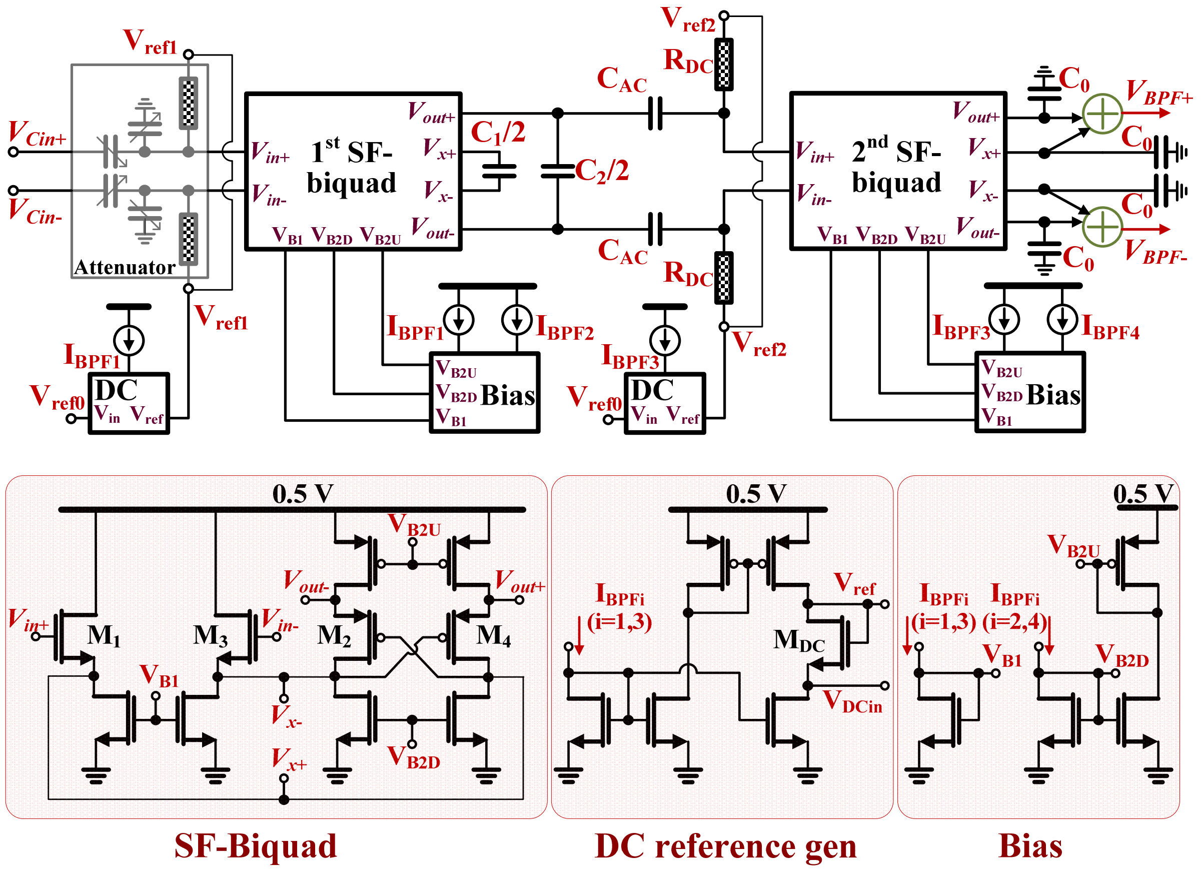


Figure 4.16. Complete asymmetrical SF-based BPF with 4 poles and 1 zero. The building blocks, including the SF-biquad, the input DC reference generator, and the bias are respectively illustrated.

The 0.5-V supply enforces boundaries on signal swing and transistor voltage overhead. The branch with cross-coupled pFETs is determinant because three transistors are stacked between ground and VDD. The expression below holds:

$$V_{sig,swing} + 2V_{ds,sat} + V_{gs} = 0.5$$

A minimum 100-mV $V_{ds,sat}$ needs to be ensured for the saturation of the upper and lower current sources because the overall linearity can benefit from their larger output resistance [144]. If a 50-mV single-ended signal swing $V_{sig,swing}$ is assumed, the pFET pair M_2/M_4 can have a maximum V_{gs} of 250 mV, and the W/L ratio can be determined accordingly with a specific bias current. The bias current, as mentioned in the noise part, is related to the required noise PSD and the BPF central frequency.

4.2.4.3 4-Pole, 1-Zero 4th-Order Asymmetrical BPF

The BPFs in silicon cochlea are usually designed to be asymmetrical to mimic the biological cochlea transfer function. The original SOSs were cascaded to obtain the accumulatively steeper lowpass roll-off while the highpass roll-off is kept to be 1st-order [23], [24]. For parallel architectures, one example is the one-zero gammatone filter transfer function which was implemented by cascading four log-domain current-mode biquads where three are LPFs and one is BPF, and therefore it has 8 poles and 1 zero [29]. Another recent example is constructed by cascading one 2nd-order BPF, one 2nd-order tunable LPF and one 5th-order elliptic LPF, which makes it a 9-pole 1-zero filter [146]. Symmetrical 4th-order BPFs have also

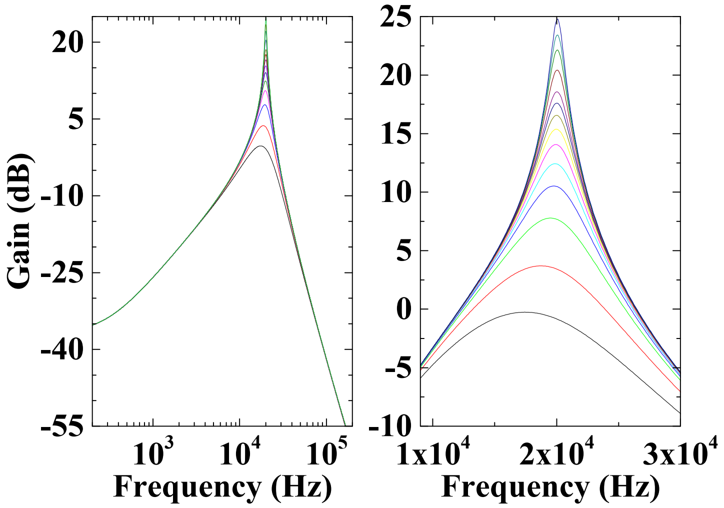


Figure 4.17. Simulated transfer functions of the 4th-order asymmetrical BPF at different Q's. The right graph is the zoomed-in around the central frequency of the left graph.

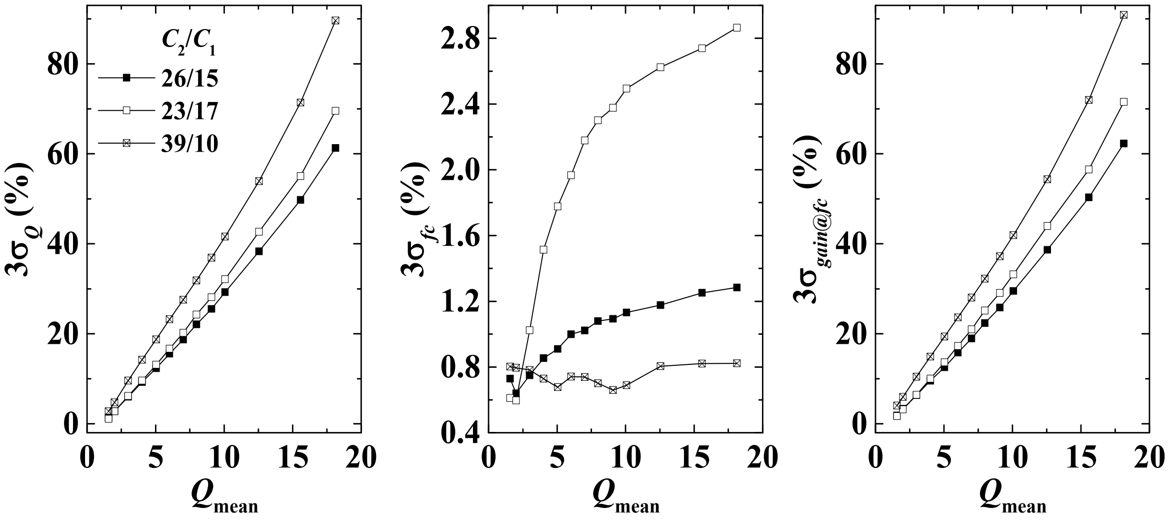
Table 4.3. Q value as the function of the bias ratio $\mathrm{I_{BPF2}/I_{BPF1}}$.

$\mathrm{I_{BPF2}/I_{BPF1}}$	Q
1	1.43
1.5876	2.00
2.1492	3.01
2.4649	4.04
2.6504	5.02
2.7889	6.03
2.8866	7.02
2.9653	8.03
3.0276	9.04
3.0800	10.09
3.1684	12.46
3.2364	15.23
3.2797	17.65
3.3197	20.63

been used [28], and it is not clear yet how the BPF order can have impact on some post-processing tasks. For example, simple 2nd-order BPFs that are used in a voice activity detector with a 16-channel cochlea-like analog frontend and a subsequent mixed-signal processing stage implementing a decision-tree machine learning algorithm seems to provide reasonable classification accuracy [42]. However, good frequency selectivity that is associated with relatively high Q and steep roll-off is speculated to be of importance in more advanced tasks like speech recognition considering the ability of fine-pitch discrimination in human auditory systems. Steeper roll-off normally comes from higher filter orders at the cost of more power consumption in parallel architectures. One exception could be a previous work from CSEM Neuchâtel, i.e. the bank of a hundred 2nd-order resonators coupled via resistive network which creates several hundreds of dB/decade lowpass roll-off because of destructive interferences of properly weighted frequency components in adjacent channels [31]. Besides the poor monotonicity of BPF central frequency scaling, the penalty is largely reduced dynamic range from 50 dB to 25 dB when the coupling is turned on.

A reasonable tradeoff between power consumption and cutoff steepness is to choose a moderate filter order, like the symmetrical 4th-order BPF in [28]. In this design, an asymmetrical 4th-order BPF is implemented with 4 poles and 1 zero, as illustrated in Figure 4.16. Two cascaded SF-based LPF biquads give rise to 4 poles, and the summation of V_x and V_{out} to obtain the final output V_{BPF} at the second biquad produces 1 zero. Several design considerations of the proposed asymmetrical BPF are listed below:

1) As will be discussed later in Section 4.2.4.4, the summation at the 2nd biquad is implemented via a capacitively-coupled PGA (CC-PGA). The input capacitors $\mathrm{C_0}$ of the PGA are in effect the capacitive loading of the 2nd SF-biquad and determine the pole positions. The four $\mathrm{C_0}$'s need to have the same capacitance in order to make the same gain for both V_x and V_{out} during summation, as required by the principle of the BPF transfer function generation shown in Eq. (4.5). We already know that with the same capacitance at V_x and V_{out}, obtaining high Q by tuning the ratio of $\mathrm{I_{BPF4}}$ over $\mathrm{I_{BPF3}}$ is unacceptable because of significant central frequency variation. Therefore, the task of Q-tuning is borne by the 1st biquad with C_2/C_1=26/15. To superimpose the resonance frequencies of the two biquads, the capacitances should satisfy $C_1 \times C_2 = C_0 \times C_0$, and the bias currents should satisfy $\mathrm{I_{BPF1} \times I_{BPF2} = I_{BPF3} \times I_{BPF4}}$, where



(a) (b) (c)

Figure 4.18. 3σ variation of (a) Q, (b) central frequency and (c) peak gain at different mean Q obtained from 250-run Monte Carlo simulations of the 20k-Hz channel.

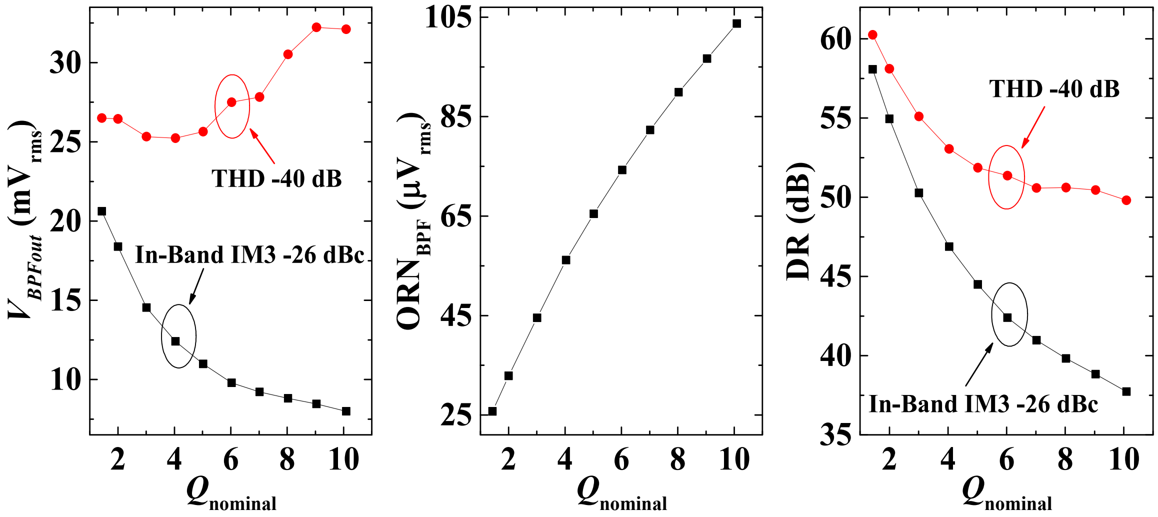
I_{BPF1} and I_{BPF2} are tunable via the DAC and TTL described in Section 4.2.2, and I_{BPF3} and I_{BPF4} are the same as the channel bias I_{chi} generated by the circuit in Figure 4.2.

2) The Q-tuning DAC illustrated in Figure 4.3(b) has a 9-bit resolution even though the number of digital controlling bits is 8. The DAC′s output is the BPF bias current I_{BPF1}, and I_{BPF2} is generated via the TLL. The resolution of the DAC is determined by the granularity of the Q-tuning steps. The simulated transfer functions of the asymmetrical BPF at different Qs for the 20k-Hz channel are shown in Figure 4.17, and the corresponding values of I_{BPF2}/I_{BPF1} are given in Table 4.3. It is easy to see that as Q increases, the interval between the bias ratios becomes smaller. If a Q-tuning step of 1 is aimed when Q<10, the necessary number of tuning bits is calculated to be:

$$-\log_2\left(\frac{1}{\sqrt{3.0276}}-\frac{1}{\sqrt{3.0800}}\right)=7.7$$

At least 8 bits are needed for the aimed Q-tuning resolution. One extra bit, i.e. 9-bit resolution in total, is assigned to the DAC for some design margin. The lowest Q value 1.43 is obtained when $I_{BPF2}/I_{BPF1}=1$. Lower Q is not considered to be of much use because of its poor frequency selectivity, and for Q within 20, the bias ratio is always less than 4. These two factors help limit the tuning range of I_{BPF1} within $[0.5I_{chi}, I_{chi}]$, and hence 8 controlling bits are sufficient for the DAC.

3) To verify the design methodology of controlling the variations of BPF′s central frequency and Q described in Section 4.2.4.1, 250-run Monte Carlo simulations are performed on the complete 4th-order BPF of the 20k-Hz channel and the results are shown in Figure 4.18. The x-axis is the mean Q value, and the y-axes are the 3σ variation of Q, central frequency and peak gain. Different C_2/C_1 ratios are used. C_2/C_1=26/15 is the one used in the fabricated chip. For the same Q, C_2/C_1=23/17 has more balanced capacitance C and requires larger I_{BPF2}/I_{BPF1}, whereas larger C_2/C_1=39/10 results in smaller I_{BPF2}/I_{BPF1}, i.e. more balanced transconductance g_m. More balanced g_m favors smaller central frequency variation, especially at high Q. This is validated by the curves in Figure 4.18(b). More balanced C, on the other hand, favors less Q sensitivity on g_m ratio. Figure 4.18(a) indeed indicates that the Q varia-



(a) (b) (c)

Figure 4.19. Simulated results at different Q values of (a) the maximum BPF rms output with in-band IM3=-26 dBc and THD=-40 dB, respectively, (b) the output-referred noise (ORN) and (c) the calculated DR. All the results are from the 4th-order BPF of the 20k-Hz channel.

tion of C_2/C_1=26/15 is smaller than that of C_2/C_1=39/10. The reason why the case of C_2/C_1=23/17 has a larger Q variation than that of C_2/C_1=26/15 might be increased g_m ratio variation due to increased I_{BPF2}/I_{BPF1}. The peak gain variation in Figure 4.18(c) resembles the Q variation. Except the tradeoff between the variations of central frequency and Q, practical circuit considerations also put constraints on the allowable C_2/C_1 ratio. If C_2/C_1 is too small, I_{BPF2}/I_{BPF1} needs to be very large for high Q values, which significantly limits signal swing in the cross-coupled pFET branch especially under low supply voltage because of the increase V_{gs} of M_2/M_4. If C_2/C_1 is too large, not only is it not area-efficient, but also a DAC with an impractically high resolution much higher than 9 bits is needed for fine Q-tuning. Therefore C_2/C_1=26/15 is an optimal choice, in terms of both BPF parameter variations and feasibility of circuit implementations.

4) The input DC level of each SF-biquad should be properly set so that the signal swing headroom at V_x and V_{out} can be maximized. V_{ref1} and V_{ref2} set the input DC level of the 1st and the 2nd SF-biquads separately via the pseudo-resistors, one in the attenuator as described in Section 4.2.3 and one as the R_{DC} in Figure 4.16. The transistor implementation of R_{DC} is the same as the one in the attenuator. With the very low 0.5-V power supply, the two biquads are AC-coupled to avoid the impact of the output DC level variation of the 1st biquad due to Q-tuning and process variation on the 2nd biquad, in contrast to the three SF-based LPF biquads directly DC-cascaded in [136] with a 1.2-V supply. The DC reference generator is used to stabilize the DC level of V_x and set it to the same voltage as V_{DCin}, if the nFET M_{DC} size is proportionally scaled with respect to that of M_1/M_3 as their bias currents are. The downscaling of the channel bias current from high to low frequency channels reduces the V_{gs} of M_2/M_4 to a value where the input signal swing can cause the pFETs to operate at the brink of triode region. To avoid this, the W/L of M_2/M_4 as well as M_1/M_3 is scaled down every four other channels.
5) The linearity of the 4th-order BPF of the 20k-Hz channel is evaluated by the in-band IM3 and THD. In the simulation of IM3, one of the two test tones is at the central frequency f_c and the other is at f_c+1k. The maximum output rms amplitudes V_{BPFout} with IM3=-26 dBc (5%) and THD=-40 dB (1%) are plotted in Figure 4.19(a) at different Q values. Even though the THD requirement is more stringent, its

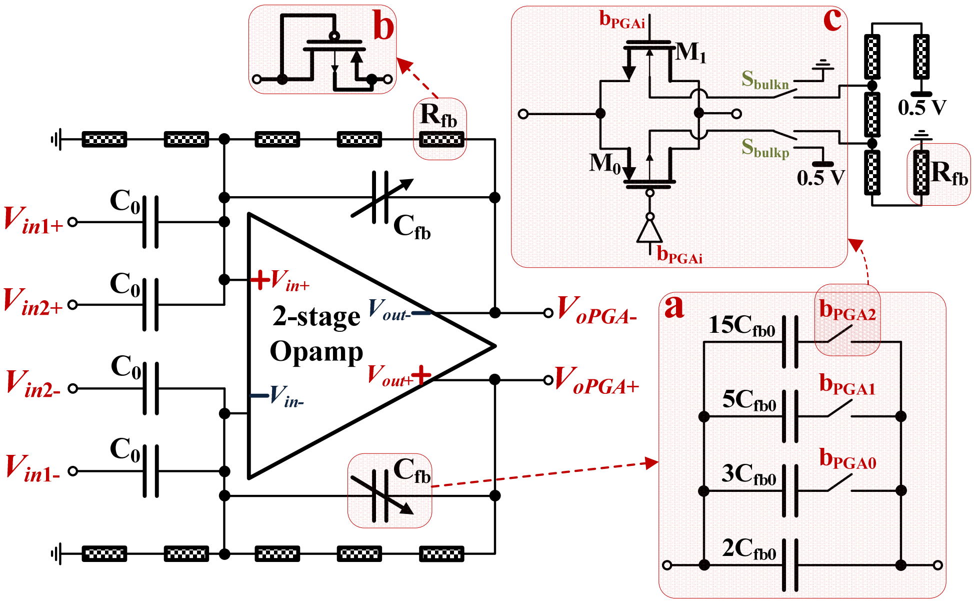


Figure 4.20. Circuit diagram of the fully differential capacitively-coupled PGA. The input signals V_{in1+}/V_{in1-} and V_{in2+}/V_{in2-} are from V_{out+}/V_{out-} and V_{x+}/V_{x-} of the 2nd SF-biquad in Figure 4.16.

maximum V_{BPFout} values are still always higher than the ones bounded by the relaxed IM3 requirement, which is attributed to the filtering effect of out-of-band harmonics. To target for a 60-dB DR at the lowest Q using the THD metric, an output-referred noise (ORN) of less than 26.5 μV should be achieved. As discussed in Section 4.2.4.2, the output noise is dominated by the cross-coupled pFETs M_2/M_4 and the current sources I_{2U} at low Q. Because the output noise of the 1st SF-biquad is filtered by the 2nd one, the total ORN of the 4th-order BPF can be approximated by considering the noise contribution only from the 2nd SF-biquad. If only thermal noise is considered, the ORN is analytical written as:

$$\text{ORN} = \sqrt{\frac{2kT}{\kappa C_0}}$$

A $C_0 \geq 14.7$ pF is calculated with $\kappa=0.8$ and $T=300$. In simulation, a $C_0=10.13$ pF is found to be sufficient, and according to the ratio of $C_2/C_1=26/15$, $C_1=3.75$ pF and $C_2=6.51$ pF are used. The simulated ORN is shown in Figure 4.19(b). It is clear that the ORN increases as Q increases due to gain peaking at the central frequency. The calculated DR is given in Figure 4.19(c). With the THD metric, the BPF DR is above 50 dB at all Q values, and has a peak value of 60.3 dB; with the IM3 metric, the highest DR is 58.1 dB and lowest 37.7 dB. The programmable attenuator described in Section 4.2.3 can add another 18-dB DR and makes the total maximum DR 78.3 dB at $Q=1.43$ with THD=-40 dB, although it has no effect on improving SNR. Note that the maximum channel bias current 50 nA given in Section 4.2.1 is determined by the calculated capacitance values and the central frequency.

4.2.4.4 Summing Programmable Gain Amplifier

The fully-differential CC-PGA for summing the V_x and V_{out} of the second SF-biquad in Figure 4.16 to ultimately obtain the 4th-order asymmetrical BPF transfer function while simultaneously provides programmable gain is illustrated in Figure 4.20. The input signals V_{in1+}/V_{in1-} and V_{in2+}/V_{in2-} are connected to V_{out+}/V_{out-} and V_{x+}/V_{x-}, respectively. Owing to the virtual ground at the input of the Opamp with large

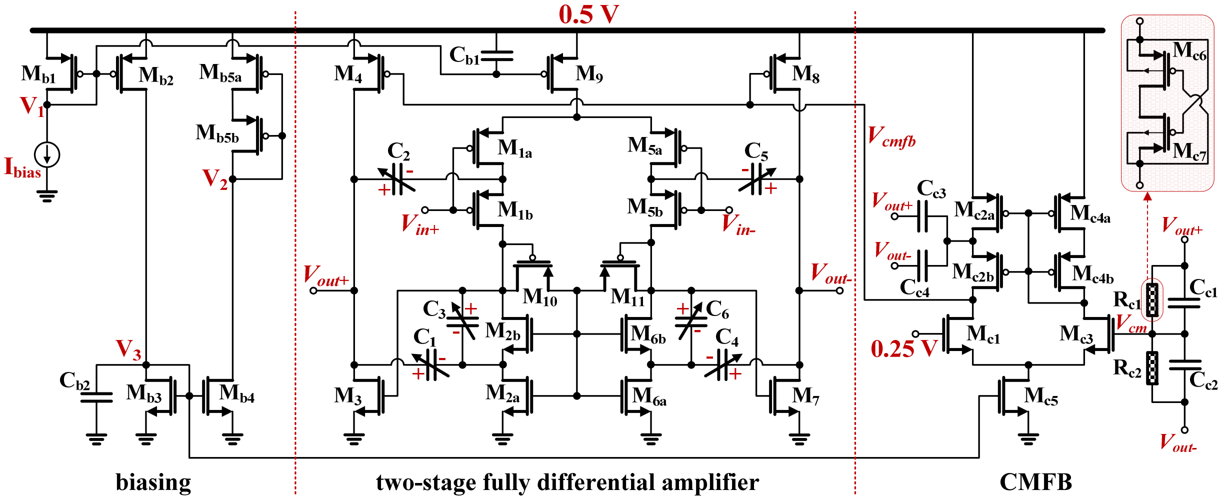


Figure 4.21. Circuit diagram of the fully-differential two-stage Opamp used in the CC-PGA in Figure 4.20 including the biasing and the common-mode feedback (CMFB) circuits.

Table 4.4. Body biasing of the transistors in the Opamp circuit in Figure 4.21.

Transistors	First 32-channel	Last 32-channel
M_{1a}, M_{1b}, M_{5a}, M_{5b}	V_2	0.5 V
M_{2a}, M_{2b}, M_{6a}, M_{6b}, M_3, M_7	V_3	0 V
M_4, M_8, M_9, M_{b1}, M_{b2}, M_{b5a}, M_{b5b}, M_{c2a}, M_{c2b}, M_{c4a}, M_{c4b}	V_1	0.5 V
M_{c1}, M_{c3}	0.4 V	0 V

open-loop gain, the four input capacitors are actually the same capacitors C_0's in Figure 4.16, and provide the loading capacitance for the second SF-biquad. The PGA is designed to have four gain levels with thermometer code controlling bits $\mathbf{b}_{\mathbf{PGAi}}$ (i=0,1,2). The unit feedback capacitance C_{fb0} is 50.6 fF, and C_0 is 10.13 pF as calculated earlier according to noise requirement. Hence the four gain levels are 18 dB, 26 dB, 32 dB and 40 dB. The same as the CC-PGA in Chapter 3, the DC feedback is to establish the input DC level and simultaneously needs to give a sufficiently low highpass corner frequency. Because of the low supply voltage, the input DC level is set to 100 mV for pFET input transistors. With the output common-mode DC level set to 250 mV for maximum output swing, the DC feedback needs to provide 2/5 voltage division. Given all the requirements, unit diode-connected pFET pseudo-resistor R_{fb} is used and the connection is shown in the insert **b**. Small *W*/*L* is used for a highpass corner frequency at least smaller than the lowest channel frequency, and the unit pFET area is kept reasonably large to minimize its contribution to output DC offset. The switch for feedback capacitance selection is depicted in the insert **c**, a complimentary transmission gate composed of pFET M_0 and nFET M_1. To reduce the ON-resistance, the bulks of M_0 and M_1 are connected to 0.1 V and 0.4 V respectively controlled by S_{bulkn} and S_{bulkp} when $\mathbf{b}_{\mathbf{PGAi}}$=1. The voltage divider to provide the 0.1 V and 0.4 V voltages is composed of five diode-connected pFETs in series between ground and VDD. When $\mathbf{b}_{\mathbf{PGAi}}$=0, the bulks of M_0 and M_1 are connected to VDD and ground respectively for large OFF-resistance.

A two-stage Opamp is chosen for the PGA core, considering sufficient open-loop gain, low noise and wide output range. The circuit diagram is shown in Figure 4.21. The biasing current I_{bias} of the Opamp in each channel is proportional to the corresponding channel bias currents I_{chi} as described in Section 4.2.1.

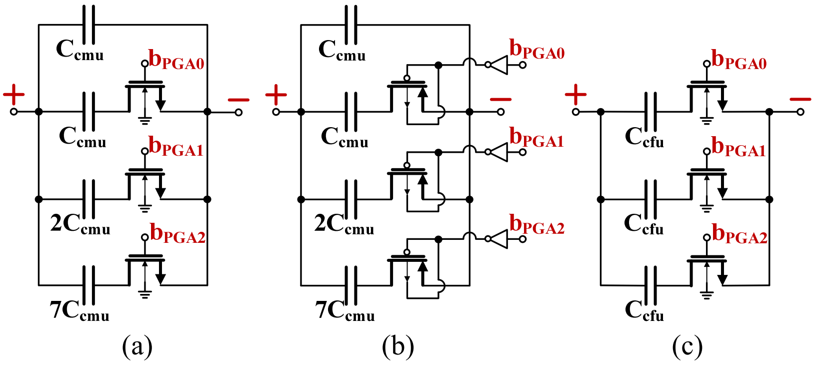


Figure 4.22. Detailed circuit diagrams of the compensation capacitors in Figure 4.12: (a) C_1 and C_4; (b) C_2 and C_5; (c) C_3 and C_6.

Because of the large current span, some of the transistor's body-biasing voltages are different in the Opamps of the first 32 and last 32 channels, which are listed in Table 4.4. The transistors not mentioned in the list have bulk connected to VDD in pFETs and ground in nFETs. Even though complimentary nFET/pFET input stage can improve power efficiency, and has been used in many neuron-activity recording biomedical ICs [147]–[150], the input transistors here are simple pFETs M_{1a}, M_{1b}, M_{5a} and M_{5b} because it is much easier to design considering current scalability. In some low-supply-voltage designs, the tail current of the first stage is avoided to save voltage headroom, and a feedforward cancellation technique is employed to suppress the common-mode gain [151]. Here however, the tail current transistor M_9 is still kept for current scaling. The voltage headroom is of less concern because of the relatively low GBW requirement for audio applications. The same pseudo-cascode compensation technique as in the CC-PGA Opamp of the retina pixel is again used here for sufficient phase margin over a wide biasing range and avoiding the use of area-consuming zero-nulling resistors. The compensation capacitors C_1, C_2, C_4 and C_5 are connected from the middle nodes of the split-transistors M_1, M_5, M_2 and M_6 instead of the output of the first stage to the output of the second stage. The problem of potentially insufficient gain margin of cascode Miller-compensation caused by gain peaking beyond GBW [107] is circumvented by the feedforward capacitors C_3 and C_6 [152].

For a one-stage Opamp, the input-referred noise (IRN) of a CC-PGA composed of this Opamp is only dependent on its closed-loop gain and loading capacitance [117]; for a two-stage Opamp like the one in Figure 4.21, the IRN is dependent on the compensation capacitance:

$$\mathrm{IRN} = \sqrt{\frac{2kT}{\kappa C_{ctot} A_c}} \tag{4.12}$$

where C_{ctot} is the total compensation capacitance and A_c is the CC-PGA closed-loop gain. To keep a relatively constant IRN while programming A_c, C_{ctot} should be accordingly changed with ideally $C_{ctot} \times A_c$ being constant. Meanwhile, the CC-PGA bandwidth should be constant and not less than the BPF central frequency with the presence of variations of bias current and capacitance. The bandwidth can be written as (see Appendix 4.6.a):

$$\mathrm{BW}_{\text{CC-PGA}} = \frac{g_{m1}}{4\pi C_{ctot} A_c} \tag{4.13}$$

where g_{m1} is the transconductance of the input transistors that is determined by the bias current of the first

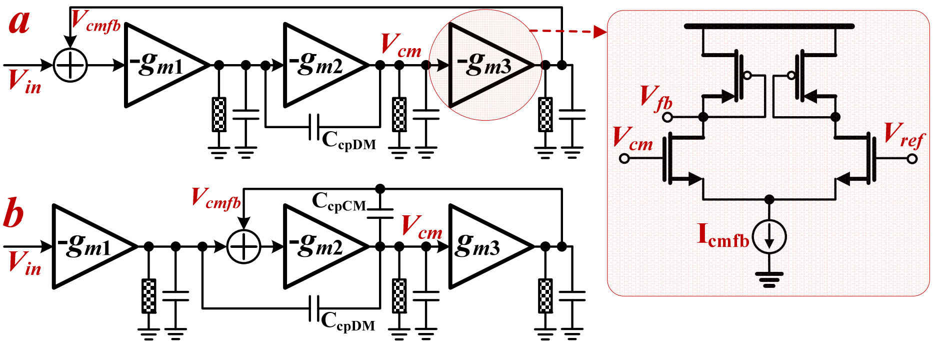


Figure 4.23. (a) Conventional CMFB loop that is composed of $-g_{m1}$, $-g_{m2}$ and $-g_{m3}$ blocks. The 5T amplifier is a representative implementation of $-g_{m3}$; (b) CMFB loop in the proposed Opamp in Figure 4.21.

stage of the Opamp. The requirement of constant $C_{ctot}\times A_c$ for BW is consistent with that for noise. For A_c=40 dB, if an IRN≤12 μV noise is aimed, C_{ctot} is calculated to be larger than 720 fF. About 700 fF is used in the design, and is split between C_1 and C_2 (or C_4 and C_5), i.e. $C_1=C_2$=350.4 fF at A_c=40 dB. With C_{cmu}=350.4 fF and C_{cfu}=987.4 fF, the complete circuits of C_1, C_2 and C_3 (the same for C_4, C_5 and C_6) are shown in Figure 4.22. Simple nFET switches are used in C_1 and C_3. Even though the bulks are connected to ground, the nFETs can have sufficiently low ON-resistance because their sources are connected to the middle node of the split-transistor M_2 (or M_6) which has a nominal voltage of less than 50 mV. pFET switches are used in C_2 because the middle node of the split-transistor M_1 (or M_5) where the pFETs' sources are connected to has a nominal voltage larger than 350 mV. To further reduce the high ON-resistance of pFET switches due to its lower mobility and higher threshold compared to nFETs, their bulks are connected to the switch control signals. Simulated results show worse-case less than 10-pA source-to-bulk leakage current when the switch is turned on.

A common-mode feedback (CMFB) circuit is an indispensable integral part for a fully differential Opamp with balanced output. Its main function is to set the Opamp output DC level at half VDD for maximum swing, i.e. 250 mV with a 0.5-V supply. A secondary function is to suppress the common-mode output swing in response to input signals and supply noise. One representative CMFB circuit for a two-stage Opamp [153]–[156] is illustrated within the red box in Figure 4.23(a) together with its symbolic small-signal abstraction $-g_{m3}$ as part of a two-stage Opamp which consists of two transconductance blocks $-g_{m1}$ and $-g_{m2}$ and a compensation capacitor C_{cpDM}. The complete CMFB loop can be seen as a three-stage amplifier in unity-gain feedback configuration. This conventional design follows the rule of thumb described in [157], particularly concerning the sharing of the main differential amplifier part with the CMFB circuit so that high common-mode open-loop gain and bandwidth can be naturally obtained. For stability consideration, the CMFB circuit is only allowed to produce a non-dominant pole whereas the dominant pole is still at the output of the $-g_{m1}$ block. This imposes a relatively high lower boundry on the bias current of the CMFB I_{cmfb}, which is usually 20% of the main amplifier's according to previously reported results [154], [158] and our own simulations. The 20% power overhead may not seem to be significant but it is already comparable to the total power consumption of the two SF-biquads in the 4th-order BPF. We argue that, in this specific system, the CMFB loop does not have to have the same high GBW as the main differential amplifier for the two reasons below:

1) The SF-biquads behave like LPFs with Q always less than 0.5 for common-mode signals, and therefore any common-mode input with frequencies higher than the BPF central frequency is filtered.

2) When the output of the summing CC-PGA is processed by the capacitively-coupled amplifier in the asynchronous delta modulator (ADM) as will be discussed in Section 4.2.5, common-mode signals will be suppressed to some extent.

Therefore, the GBW of the CMFB loop is designed to be only several times higher than the CC-PGA bandwidth to save power, and the symbolic diagram is shown in Figure 4.23(b). Only two g_m blocks $-g_{m2}$ and g_{m3} are in the loop. The first stage g_{m1} achieves its output DC balance via the diode-connected-pFET pseudo-resistors M_{10} and M_{11} in Figure 4.21. The dominant pole in the CMFB loop now is at the output of the g_{m3} block. Additional compensation by the capacitor C_{cpCM} is needed to ensure stability. C_{cpCM} is implemented as C_{c3} and C_{c4} in Figure 4.21. The power consumption of the CMFB circuit is now less than 1% of the main differential amplifier, and the CMRR of the summing CC-PGA together with the ADM amplifier as will be described in the next section is found to be larger than 50 dB around the central frequency in the worst case in Monte Carlo simulations.

One more criterion of the CMFB circuit design is that the common-mode signal detector should have a linear characteristic [157]. A commonly used detector comprises two resistors (R_{c1} and R_{c2}) and two capacitors (C_{c1} and C_{c2}) with the connections shown in Figure 4.21. The resistors and capacitors sense the low- and high-frequency common-mode signals, respectively. For good linearity, ideally the resistor and capacitors can be implemented by poly-silicon and MIM structure. However, to avoid resistive loading at the Opamp output, very large resistance is needed which is impractical in terms of area. Pseudo-resistor has to be used instead. For example in [149], diode-connected pFETs are used as R_{c1} and R_{c2}, like M_{10} and M_{11} at the Opamp's first stage. If the pseudo-resistance is sufficiently large, mainly the capacitors C_{c1} and C_{c2} play the role in detecting common-mode signals, which should be reasonably linear. The problem of this solution is that for large output swing, the asymmetrical characteristic of diode-connected pFETs causes DC drift of the generated common-mode voltage V_{cm}, and in turn drift of the DC level of the Opamp's output, which can quickly lead to signal clipping especially under a very low supply voltage. The DC drift is solved by using a symmetrical pseudo-resistor that is composed of two pFETs M_{c6} and M_{c7} as shown in the red box insert in Figure 4.21. It is adapted from the one in [104] where the working principle is detailed. The difference is that no tunable level-shifting is needed here between M_{c6}'s (M_{c7}'s) bulk and Mc_7's (M_{c6}'s) gate for adjustable resistance.

4.2.5 Asynchronous Delta Modulator with Self-Oscillating Comparison

The block diagram of the asynchronous delta modulator (ADM) is illustrated in Figure 4.24. The input signals V_{inadm+} and V_{inadm-} are from V_{oPGA+} and V_{oPGA-} of the summing CC-PGA in Figure 4.20, respectively. V_{inadm+} and V_{inadm-} are not differentially connected to the input capacitors of the ADM Opamp unlike the VGA block in [149], because in that case the Opamp's input voltages V_{ref} and V_{rst} would have large common-mode fluctuation due to the lack of feedback between the Opamp's output and its positive input, which is not allowed according to the charge conservation requirement of an ADM as explained in Section 3.2.3 in Chapter 3. To convert the differential input to single-ended output, either V_{inadm+} or V_{inadm-} needs to be inverted before summed together. Here V_{inadm+} is inverted by the unity-gain inverting amplifier, i.e. the '**-1**' block, and then combined with V_{inadm-} through the capacitively-coupled amplifier employing a differential-input single-ended-output two-stage Opamp. The Opamp's input DC level V_{ref} is as well set to 100 mV through the diode-connected-pFET pseudo-resistor R_{fb} to give sufficient headroom for the design of the Opamp's input stage. The equilibrium voltage of V_{rst} is thus also at 100 mV. The passband gain of the capacitively-coupled amplifier is 2 with C_{adm}=101.3 fF, and the working principle of the asynchro-

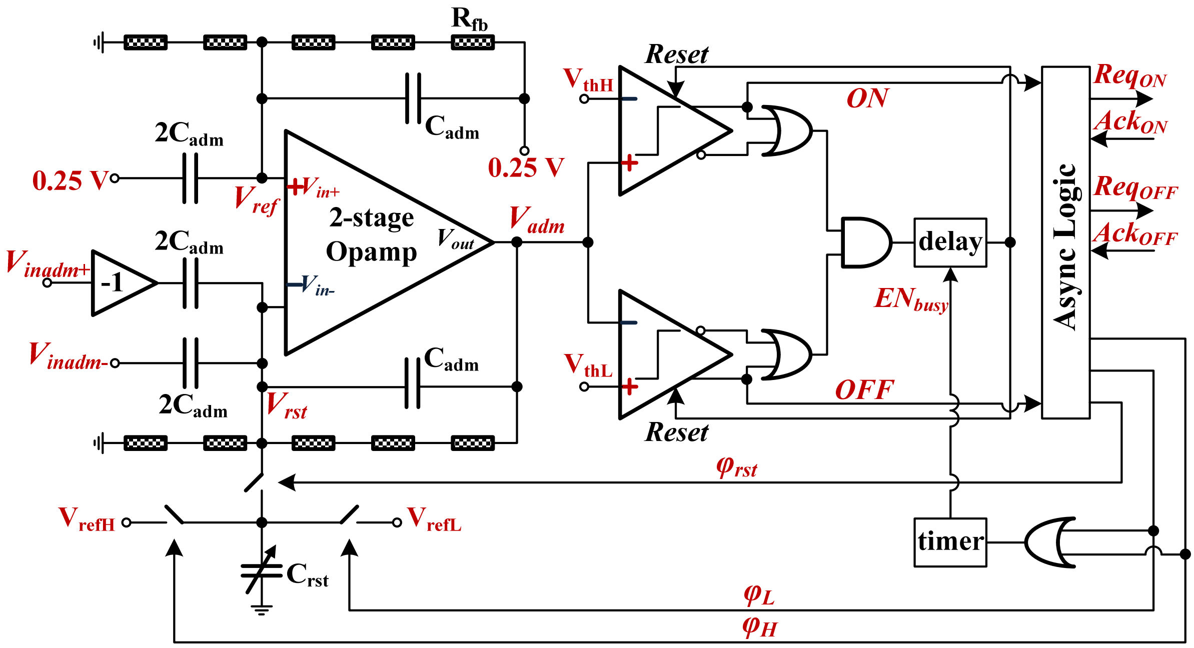
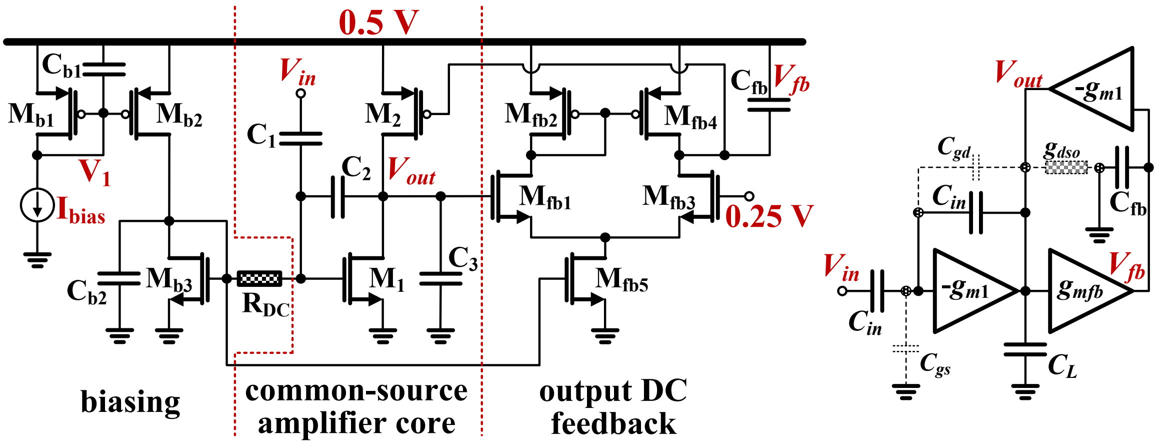


Figure 4.24. Block diagram of the ADM with adaptive self-oscillating comparison.

nous switched-capacitor circuit for delta modulation is the same as the ADM circuit in the retina pixel except that here C_{rst} is programmable. Adaptive self-oscillating comparison is proposed to improve energy efficiency compared to using continuous-time comparators. Dynamic latched comparators are used, and their psudo-clock *Reset* signal is generated by detecting the completion of each comparison. The flag signals *ON* and *OFF* indicating spike generation only become valid when the amplifier output V_{adm} is above or below the comparison thresholds V_{thH} and V_{thL}, respectively. Once the asynchronous logic is triggered by the *ON* or *OFF* signal, it communicates with the peripheral AER to transmit the generated spike and generates the switching signals φ_{rst}, φ_H and φ_L for ADM. High φ_H and φ_L are also responsible for adaptively increasing the frequency of self-oscillation which is kept at a lower value when there is no spike generation and communication. The following subsections will detail the designs of the ADM building blocks in Figure 4.24.

4.2.5.1 Unity-Gain Inverting Amplifier

Ideally the unity-gain inverting amplifier should have a gain of 1 and cause no additional phase shift at the output with respect to its input signal. Any mismatch of gain and phase between the input and output can degrade the CMRR, which is of course inevitable in practical circuits and can only be minimized. Power overhead is another main concern, and for this reason, a simple common-source amplifier is used as the core of the inverting amplifier. The complete circuit and its symbolic abstraction are illustrated in Figure 4.25. The body biasing voltages of the relevant transistors are listed in Table 4.5, and the other nFETs and pFETs by default have their bulks connected to ground and VDD, respectively. As shown in Figure 4.25(a), the 2T common-source amplifier (M_1 and M_2) is configured in a unity-gain closed-loop feedback by capacitors C_1 and C_2 with equal capacitance. The biasing current is copied via M_{b3}. A pseudo-resistor R_{DC} that has the same topology as the one used in the attenuator in Figure 4.5 blocks the low impedance path of the diode-connected M_{b3} for the input signal. The associated highpass corner frequency is sufficiently low due to R_{DC}'s large resistance. To set the DC level of V_{out} at 250 mV for maximum



(a) (b)

Figure 4.25. (a) Circuit diagram and (b) symbolic abstraction of the unity-gain inverting amplifier.

Table 4.5. Body biasing of the transistors in the amplifier in Figure 4.25.

Transistors	First 32-channel	Last 32-channel
M_2, M_{fb2}, M_{fb4}	V_1	0.5 V
M_{fb1}, M_{fb3}	0.4 V	0 V

swing, a feedback loop is formed by the 5T amplifier (M_{fb1}-M_{fb5}) and M_2, which however creates a zero that should be placed at a much lower frequency than the channel central frequency.

To analyze the frequency response in detail and determine the design parameters, the complete transfer function is derived according to the small-signal symbolic abstraction in Figure 4.25(b) where C_{in} is the capacitance of C_1 and C_2, C_L is the capacitance of C_3, C_{gs} and C_{gd} are the gate-source and gate-drain capacitances of M_1, g_{m1} is the transconductance of M_1 and M_2, g_{mfb} is the transconductance of the 5T feedback amplifier, and g_{ds0} is the output conductance of the common-source amplifier:

$$H_{UGIA}(s)=\frac{C_{in}\cdot[s^2\cdot(C_{in}+C_{gd})-s\cdot g_{m1}]}{s^2\cdot[C_{in}C_L+(C_{in}+C_L)(C_{in}+C_{gd}+C_{gs})+C_{gd}C_{gs}]+s\cdot[(C_{in}+C_{gd})g_{m1}+(2C_{in}+C_{gs}+C_{gd})g_{ds0}]+\dfrac{(2C_{in}+C_{gd}+C_{gs})g_{m1}g_{mfb}}{C_{fb}}}$$

With approximation, the locations of zeros and poles can be obtained as:

$$z_1=0,\ z_2=\frac{g_{m1}}{C_{in}},\ p_1=\frac{2g_{mfb}}{C_{fb}},\ p_2=\frac{g_{m1}}{C_{in}+2C_L}$$

The highpass corner frequency is at p_1. If $2C_L \ll C_{in}$, z_2 overlaps with p_2 and $H_{UGIA}(s)$ has a highpass characteristic. Otherwise, the lowpass corner is at p_2. To minimize the phase shift around the channel central frequency f_{ci}, p_1 and p_2 should be at least 10× away from $2\pi f_{ci}$. The passband gain A_{0UGIA} can be derived as:

$$A_{0UGIA}=\frac{1}{1+\dfrac{2g_{ds0}}{g_{m1}}+\dfrac{C_{gd}}{C_{in}}+\dfrac{g_{ds0}(C_{gs}+C_{gd})}{g_{m1}C_{in}}}$$

Increasing the open-loop gain of the common-source amplifier and reducing the C_{gs} and C_{gd} by proper sizing can help A_{0UGIA} approach 1. Connecting C_{fb} between V_{fb} and VDD instead of ground significantly

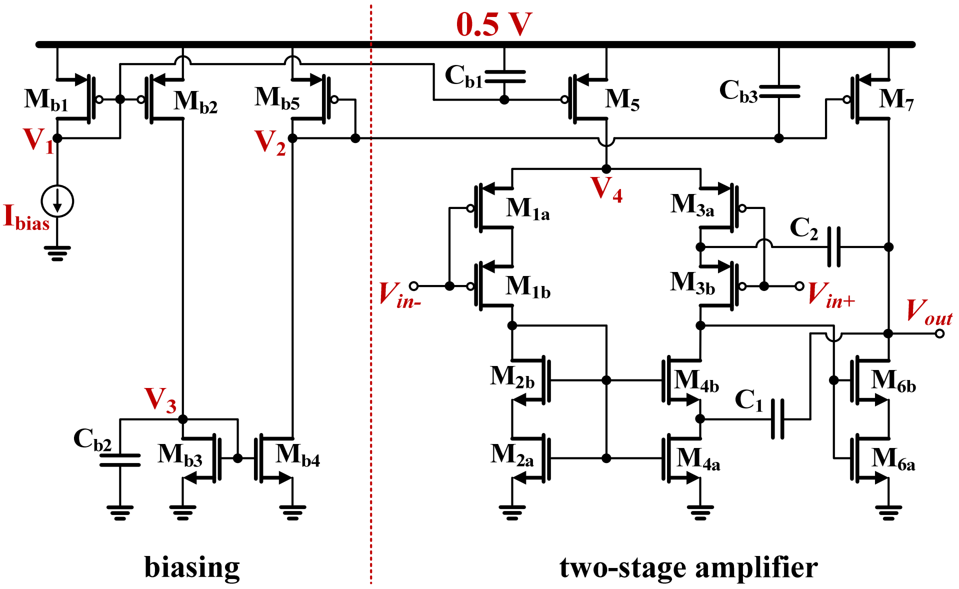


Figure 4.26. Circuit diagram of the differential-input single-ended-output two-stage Opamp used in ADM.

Table 4.6. Body biasing of the transistors in the amplifier in Figure 4.26.

Transistors	First 32-channel	Last 32-channel
M_{1a}, M_{1b}, M_{3a}, M_{3b}	V_1	V_4
M_{b5}, M_7	V_2	0.5 V
M_{2a}, M_{2b}, M_{4a}, M_{4b}, M_{6a}, M_{6b}	V_3	0 V

improves PSRR because the AC coupling stabilizes the V_{gs} of M_2; in fact, if C_{fb} is connected to ground, the PSRR has a gain of 2 at the square root of $p_1 \times p_2$, which is in vicinity of f_{ci}. The PSRR can be further optimized by reducing the C_{gd} of M_2. Another constraint on C_{fb} is the stability of the DC feedback loop. If the non-dominant pole at V_{out} is to be 3× the GBW of the loop for sufficient phase margin, the value of C_{fb} has to satisfy:

$$C_{fb} \geq \frac{3 g_{m1} g_{mfb} C_L}{g_{ds0}^2}$$

4.2.5.2 Differential-Input Single-Ended-Output Two-Stage Opamp

The schematic of the differential-input single-ended-output Opamp is given in Figure 4.26. It has the same topology as the Opamp used in the retina pixel described in Chapter 3. Its body-biasing voltages are listed in Table 4.6. The noise of this Opamp is of less concern thanks to the gain of the previous circuit stages. The speed is more important because of the fast settling requirement of the output during the charge redistribution phase in ADM. The GBW of the Opamp can be estimated as follows. According to Eq. (4.a1) in Appendix 4.6.b, to achieve a η% settling accuracy within time t, the equation below should be satisfied:

$$p_1 \cdot \frac{A_0 C_2}{C_1 + C_2 + C_{rst}} t = \ln \frac{100}{\eta}$$

where p_1 and A_0 are the dominant pole frequency and DC gain of the Opamp. Note that $p_1 \times A_0 = 2\pi \times \text{GBW}$. Assuming a sinusoidal input for a channel with a central frequency of f_{ci}, if each period produces N spikes,

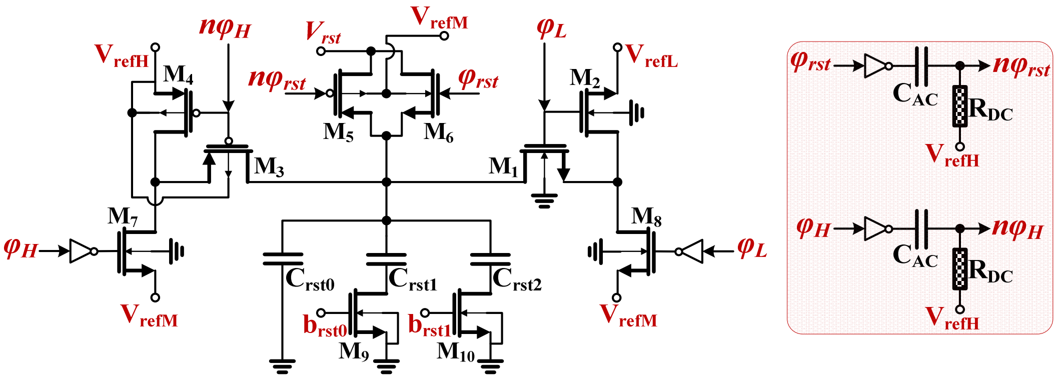


Figure 4.27. Detailed switched-capacitor circuits used in ADM.

Table 4.7. Programmable δ-subtraction of the ADM amplifier in one charge redistribution operation.

$\mathbf{b_{rst1}b_{rst0}}$	00	01	10	11
ΔV_{adm}	δ	1.5δ	2δ	2.5δ

then the settling has to be at least within the time of $1/(f_{ci}\times N)$. The GBW can be finally expressed as:

$$\mathrm{GBW}\geq\frac{1}{2\pi}\ln\frac{100}{\eta}\cdot f_{ci}\cdot N\cdot\frac{C_1+C_2+C_{rst}}{C_2}$$

For example, for the 20k-Hz channel, if N=16, η=2 and C_1=$2C_{rst}$=$4C_2$, GBW needs to be at least 1.4M rad/s. This is a tradeoff between the encoding quality and power consumption, because smaller η and larger N results in more faithful encoding at the cost of large GBW As discussed in Chapter 2, small N (3- to 4-bit encoding, i.e. N takes value from 8 to 16) can already give high encoding quality in light of the decoding SDR using nonlinear reconstruction algorithms based on frame theory. Therefore the power consumption of the sensing frontend can be reduced by using a small N at the cost of possibly more post-processing power, which might be beneficial in wireless sensor applications due to scarcely available energy at sensor nodes.

4.2.5.3 ADM Switched-Capacitor Circuits

The detailed circuit of the ADM switched-capacitor network is shown in Figure 4.27. The reset capacitor C_{rst} in Figure 4.24 is programmable by two digital bits $\mathbf{b_{rst0}}$ and $\mathbf{b_{rst1}}$. It consists of three capacitors with C_{rst0}=C_{rst2}=C_{adm} and C_{rst1}=$0.5C_{adm}$. Hence the ultimate change at the output of the ADM amplifier ΔV_{adm} can have four different values, and are listed in Table 4.7 where δ=V_{refH}-V_{refM}=V_{refM}-V_{refL}. Because in idle state, V_{rst} is 100 mV set by the negative feedback in the ADM amplifier, the middle level reference voltage V_{refM} is set to 100 mV as well. V_{refL} can be adjusted between ground and any voltage less than 100 mV, and in practice usually set to ground. For symmetry V_{refH} is set to 200 mV, which gives a δ of 100 mV. When φ_L is high, the nFET switches M_1 and M_2 have a V_{gs} of 500 mV that is enough to give a sufficiently small ON-resistance determined by the requirement of charging speed. When φ_H is high, however, the pFET switches M_3 and M_4 only have a $|V_{gs}|$ of 200 mV, leading to unacceptably large ON-resistance. To circumvent this problem, the circuit in the red box in Figure 4.27 is used to generate an $n\varphi_H$ swinging between -300 and 200 mV, and thus a 500-mV $|V_{gs}|$ can also be obtained for M_3 and M_4 when C_{rst} needs to be charged to V_{refH}. The same circuit is used to generate $n\varphi_{rst}$ which controls the pFET switch M_5. R_{DC} is the same pseudo-resistor used in the attenuator as depicted in Figure 4.5. The coupling capacitor C_{AC}

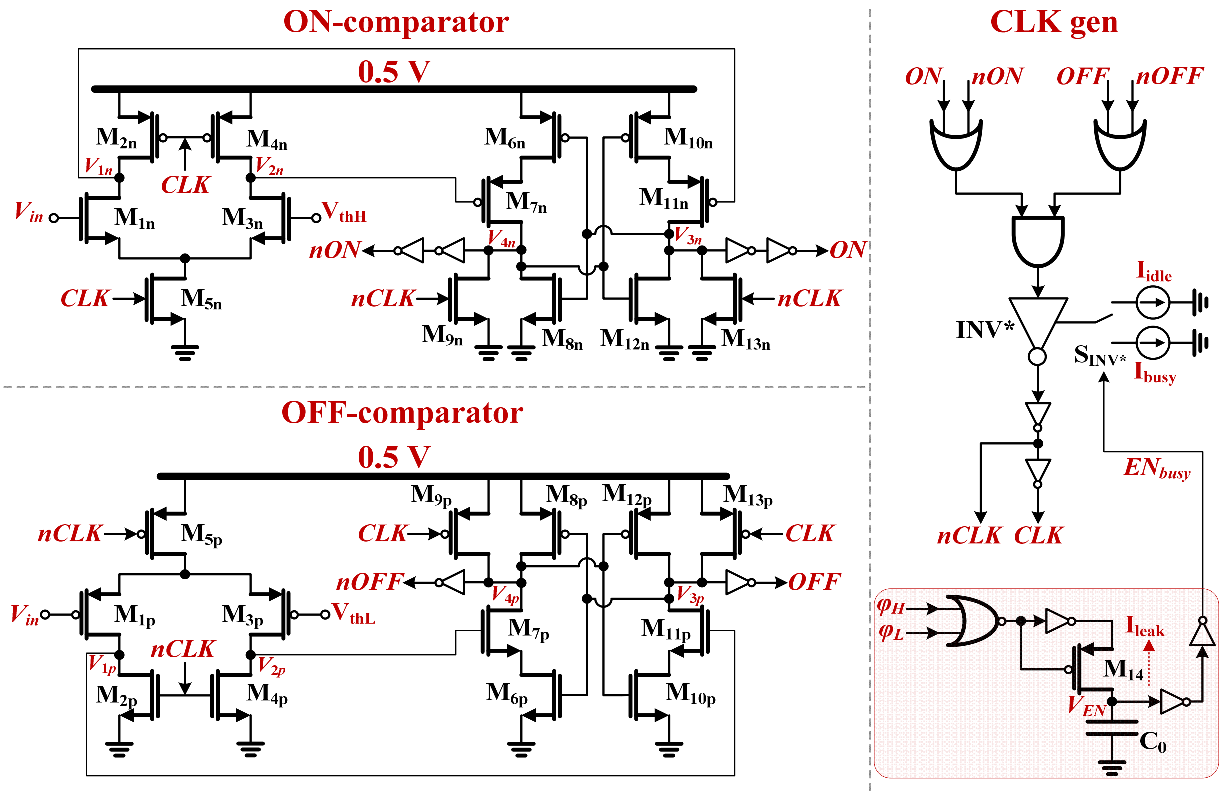


Figure 4.28. Circuit diagrams of the ON- and OFF-comparators and the asynchronous logic that generates the *CLK* and *nCLK* for the comparators. The circuit that generates the enabling signal EN_{busy} for fast self-oscillation is in the red box.

should be chosen such that $n\varphi_{rst}$ and $n\varphi_H$ stay approximately at -300 mV during a large pulse-width of φ_{rst}, particularly in low frequency channels where the ADM amplifier takes a long time to settle in each charge redistribution phase because of the scaled bias current. The pole associated with the ON-resistance R_{on} of the transmission gate switch composed of M_5 and M_6 can be approximated as $1/(R_{on}C_{rst})$ according to Eq. (4.a2) given high open-loop gain of the Opamp. This pole should be located at a higher frequency than the pole p_2 in Eq. (4.a1) determined by the GBW of the Opamp, which sets the upper limit on R_{on} as:

$$R_{on} \le \frac{1}{\ln\dfrac{100}{\eta}\cdot f_{ci}\cdot N\cdot C_{rst}}$$

4.2.5.4 Latched Comparators and Self-Oscillation Logic Circuits

Discrete-time (DT) comparators that employ latch with positive feedback generally have higher comparison speed and power efficiency compared to continuous-time (CT) comparators based on multi-stage amplifiers, especially at low supply voltages (see comparison analysis in Appendix 4.6.c). However, DT-comparators usually need clock to reset the regenerative latch to its initial state after each comparison is completed. In asynchronous systems, clock is normally not available. In order to internally generate the clock, i.e. the *Reset* signal in Figure 4.24, a self-oscillation loop can be formed in which *Reset* is derived from the completion of one comparison. This method has actually been used in SAR ADCs to automati-

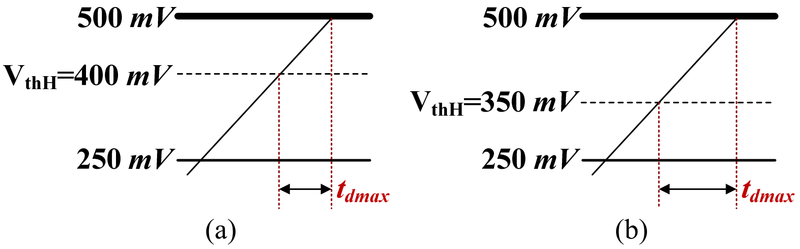
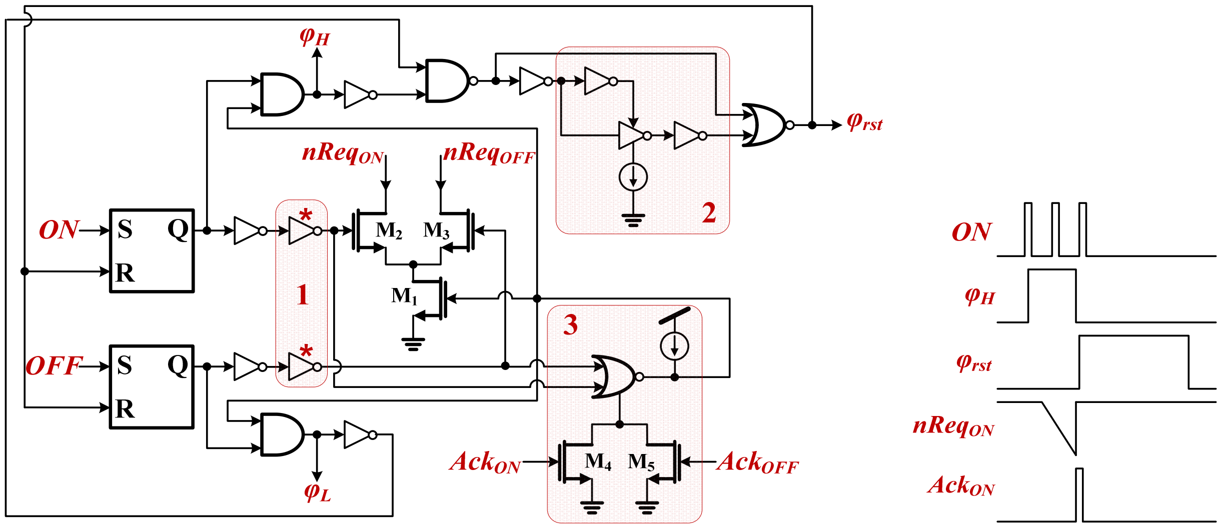


Figure 4.29. Impact of the threshold setting on maximum allowable comparison delay t_{dmax}.

cally determine the optimal time needed for each comparison [159], [160]; nonetheless, to the author's best knowledge, no such scheme has been employed in any asynchronous systems for signal modulation.

The two-stage DT-comparators used in the design are depicted in Figure 4.28. They are based on the topology proposed in [161] which has less stacking between ground and VDD compared to the conventional StrongARM latch [162]. The ON- and OFF-comparators use nFET and pFET input transistors respectively, because the nominal upper and lower threshold voltages $\mathrm{V_{thH}}$ and $\mathrm{V_{thL}}$ are 400 and 100 mV respectively. The working principle is briefly described here taking the ON-comparator as the example. When CLK=0 and $nCLK$=1, the comparator is reset. The tail nFET $\mathrm{M_{5n}}$ is off and the pFETs $\mathrm{M_{2n}}$ and $\mathrm{M_{4n}}$ are on charging the output of the first stage V_{1n}/V_{2n} to VDD. Consequently, the input pFETs of the second stage $\mathrm{M_{7n}}$ and $\mathrm{M_{11n}}$ are off, cutting off the current path from VDD to ground while the output of the second stage V_{3n}/V_{4n} are set to ground by $\mathrm{M_{9n}}$ and $\mathrm{M_{13n}}$. The final output ON and nON are both low. When CLK=1 and $nCLK$=0, the comparison starts. If $V_{in}>V_{thH}$, V_{1n} drops faster than V_{2n}, and the difference between V_{1n} and V_{2n} increases over time as long as the input nFETs $\mathrm{M_{1n}}$ and $\mathrm{M_{3n}}$ are still in saturation. As the common-mode voltage of V_{1n} and V_{2n} decreases, $\mathrm{M_{7n}}$ and $\mathrm{M_{11n}}$ start to turn on, and V_{3n} is charged faster than V_{4n}. Finally the positive feedback loop that consists of $\mathrm{M_{6n}}$, $\mathrm{M_{8n}}$, $\mathrm{M_{10n}}$, and $\mathrm{M_{12n}}$ takes over, and V_{3n} goes up to VDD and V_{4n} goes down to ground. The comparison is completed with ON=1 and nON=0. On the other hand, if $V_{in}<V_{thH}$, the final output is ON=0 and nON=1. The comparator is ready for the next comparison cycle. The OFF-comparator is summarized as follows: when CLK=0 and $nCLK$=1, the comparator is reset, and OFF=0 and $nOFF$=0; when CLK=1 and $nCLK$=0, the comparison starts, and the result is either OFF=1 and $nOFF$=0 if $V_{in}<V_{thL}$ or OFF=0 and $nOFF$=1 if $V_{in}>V_{thL}$.

A self-oscillating loop can be formed based on the fact that when the comparator is in reset, both outputs are low, and after the completion of a comparison, one output is low and the other is high. Therefore, the reset and comparison completion can be detected by an OR gate. The simplified self-oscillating logic is shown on the right side of Figure 4.28. After reset in response to CLK=0 and $nCLK$=1, all four inputs to the two OR gates are low, which leads to CLK=1 and $nCLK$=0. After comparison completion in response to CLK=1 and $nCLK$=0, each OR gate receives a high input which leads to CLK=1 and $nCLK$=0 after some delay. The delay is controlled by the current-starved inverter INV*, and it determines the frequency of the self-oscillation. For further power saving, the oscillating frequency is set to a low value by connecting the switch $\mathrm{S_{INV*}}$ to a smaller starving current $\mathrm{I_{idle}}$ when no threshold crossing occurs, i.e. no spike output is detected. Once φ_H or φ_L is triggered to high by output spikes, EN_{busy} is set to high and a larger starving current $\mathrm{I_{busy}}$ is connected to INV*, resulting in a higher oscillation frequency. Both $\mathrm{I_{idle}}$ and $\mathrm{I_{busy}}$ are derived proportionally from the corresponding channel bias $\mathrm{I_{chi}}$. The circuit in the red box in Figure 4.28 acts like an analog timer. It keeps EN_{busy} high as long as the time interval between two consecutive φ_H or φ_L pulses is less than a certain time threshold t_{th}. The t_{th} is determined by the leakage current of the pFET $\mathrm{M_{14}}$ and the capacitance of $\mathrm{C_0}$. EN_{busy} becomes low when V_{EN} is discharged low. t_{th} can be designed to be



(a) (b)

Figure 4.30. (a) In-channel asynchronous logic circuits that generate the control signals φ_H, φ_L and φ_{rst} for ADM, and communicate with the 1.8-V peripheral AER circuit via request signals Req_{ON}, Req_{OFF} and acknowledge signals Ack_{ON} and Ack_{OFF}. The inverters with star marks have long channel-length pFETs; (b) Exemplary timing diagram of an ON spike.

several seconds in accordance with the time interval of normal human speech utterance, or other values depending on the application scenarios.

The frequency of the self-oscillation sets the worst-case comparison delay. If a threshold-crossing occurs right after the comparator enters the regenerative phase, the comparator output will not become valid until this oscillation cycle is completed. Hence we approximately have:

$$t_{dmax} \approx \frac{1}{f_{osc}}$$

where t_{dmax} is the maximum delay and f_{osc} is the oscillation frequency. Note that f_{osc} is not constant because the actual comparison time t_{comp} between the moment when *CLK* becomes valid and the end of the latch regeneration is dependent on the slope of the input signal. However, f_{osc} is mostly determined by the time constant of INV* when t_{comp} is comparatively small. The low supply voltage put hard limit on t_{dmax}, which is explained by the illustration in Figure 4.29. In Figure 4.29(a), the upper threshold $\mathrm{V_{thH}}$ is set to 400 mV. If the extra delays caused by in-channel asynchronous logic and AER communication are neglected, within t_{dmax}, the increase of the ADM amplifier output ΔV_{adm} has to be less than 100 mV to avoid clipping before V_{adm} is pulled back towards 250 mV in response to delta subtraction. In Figure 4.29(b), the ΔV_{adm} is relaxed to 150 mV with $\mathrm{V_{thH}}$ set to 350 mV. It is clear that a lower threshold allows a larger t_{dmax} given the same V_{adm} slope, at the cost of more severe threshold mismatch [89]. The following calculation considers the case of $\mathrm{V_{thH}}$=400 mV. In the worst case, the two inputs of the ADM amplifier V_{inadm+} and V_{inadm-} in Figure 4.24 have maximum amplitude of 250 mV assuming sinusoidal signals, and V_{adm} in turn has amplitude of 1V. The maximum slope is therefore $2\pi\times20\times10^3$ for the 20k-Hz channel, and t_{dmax} is calculated to be $0.1/(2\pi\times20\times10^3)$=796 ns, which means f_{osc} has to be at least 1.26M Hz. Corner simulation results show that, with V_{adm} set to 250 mV, f_{osc} has minimum value of 1.92M Hz in busy mode and 0.81M Hz in idle mode. When V_{adm} is set to 401 mV, 1 mV above $\mathrm{V_{thH}}$, the minimum f_{osc} in busy mode degrades

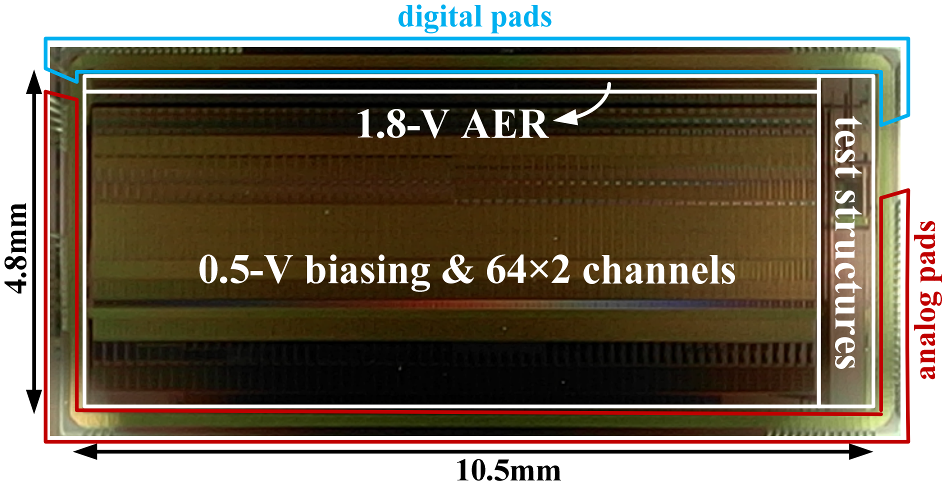


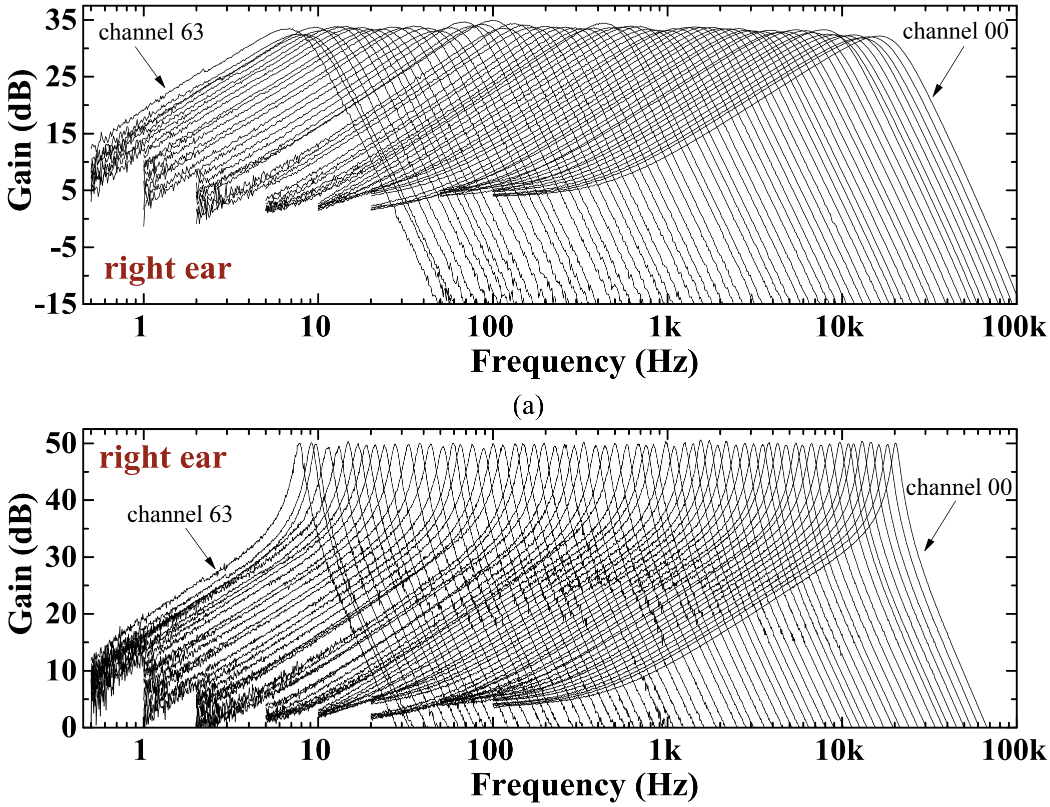
Figure 4.31. Chip microphotograph.

to 1.48M Hz. The idle-mode f_{osc} is 36% less than the minimum required 1.26M Hz, which suggests that any V_{inadm+} and V_{inadm-} with amplitude larger than 160 mV might get V_{adm} clipped at the first spike. However, keep in mind that the 1.26M-Hz requirement is overestimated by using the maximum slope of a sinusoid. The simulated worst-case power consumption of the complete circuits in Figure 4.28 in busy mode is 290 nW, and 164 nW in idle mode which is 43% less.

4.2.5.5 In-Channel Asynchronous Logic

The in-channel asynchronous logic is shown in Figure 4.30(a) and the exemplary ON-spike timing diagram is shown in Figure 4.30(b). The main functions of the logic are to generate the control signals φ_H, φ_L and φ_{rst} for ADM, and serve as the communication interface between the 0.5-V cochlea channels and the 1.8-V peripheral AER. Because the output of a DT-comparator is valid only at the end of the comparison phase and become invalid during the reset phase, an SR latch is indispensable to store the generated spike. Taking ON-spike as the example, an ON spike immediately sets φ_H high for charging the C_{rst} to V_{refH} in the ADM. The inverters in the red box 1 in Figure 4.30(a) have pFETs with long L, and thus the request $nReq_{ON}$ becomes valid after a certain delay which sets the pulse width of φ_H. Once $nReq_{ON}$ is pulled low, a valid-high Ack_{ON} is sent back from the 1.8-V peripheral AER, and brings $nReq_{ON}$ back high as long as either nFET M_1 or M_2 stays off after Ack_{ON} goes back low; otherwise $nReq_{ON}$ will be pulled low again, generating false ON spikes. The circuit in the red box 3 is used to guarantee the overlapped off-state of M_1 and M_2 after Ack_{ON} goes low. The weak pull-up current source charges the gate of M_1 slowly so that the time needed for M_2 to be turned off by valid φ_{rst} setting the SR latch output low is sufficient. Once the gate of M_2 goes low the NOR gate in the red box 3 turn on M_1 immediately, ready for the next spike transmission. The weak pull-up is necessary because otherwise the initial state of the whole logic could be in deadlock and unable to send off any spike after power up if M_1 is initially off. The circuit in the red box 2 is to generate the φ_{rst} pulse after the completion of charging C_{rst} with φ_H or φ_L turning low. The current-starved inverter controls the pulse width of φ_{rst} for sufficient settling time of the ADM amplifier, and the current source is derived proportionally from the channel bias I_{chi}. The signal flow of an OFF spike is similar, so the details will not be repeated here.

4.3 System Design and Chip Microphotograph



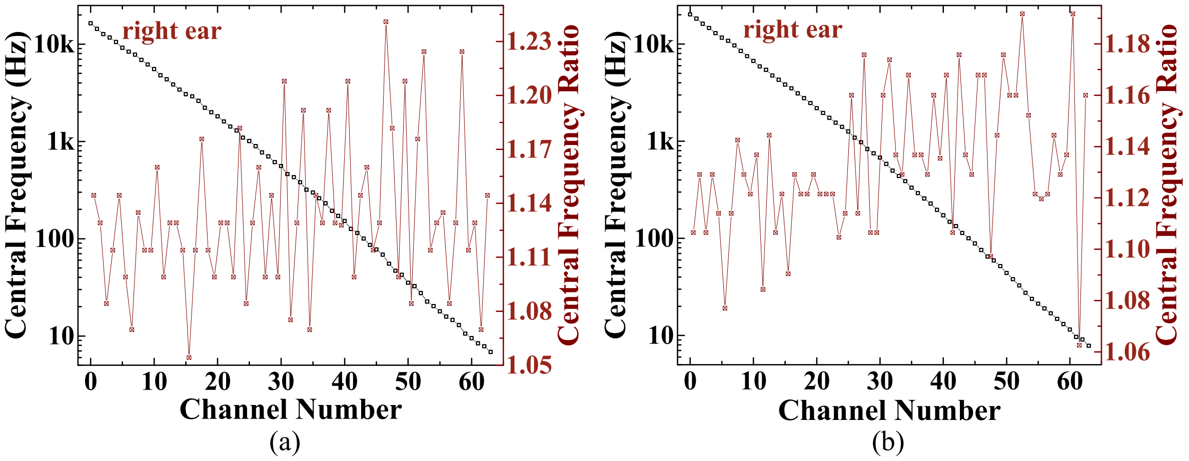
(a)

(b)

Figure 4.32. Measured transfer functions of the right ear at (a) Q≈1.3 and (b) Q≈9.2.

The chip was designed and fabricated in TowerJazz 0.18 µm IS CMOS with triple well, except the noise was simulated using UMC models. Figure 4.31 shows the chip microphotograph. It occupies an area of 10.5×4.8 mm^2 including the pads. The 128 binaural channels are placed in an interleaved manner, i.e. the left and right branches of the same channel are mirrored and put in neighbor. The 1-D AER has 256 channels (ON and OFF spike communications use separate AER channels) with fair arbitration. To obtain each individual transfer function of all the channels, the differential output of the summing PGAs are buffered by on-chip pFET source followers with a 1.8-V supply before connected to a common bus, and a shift register chain is used to select the channel connected to pads for measurements. The source follower buffer is designed to have sufficiently high bandwidth and low noise. The latches in each channel for storing configuration bits have the 0.5-V supply, and the timing diagram for channel selection and bit writing is similar to the one in Chapter 5 for programming the bias generator array. The digital nFETs in the 0.5-V core are built within deep N-well to mitigate noise injection into substrate. The digital and analog pads are arranged such that their power supplies and grounds can be easily separated. A programmable bias generator that will be described in Chapter 5 [120] is integrated but only for biasing AER, source follower buffer and test structures. A diagnostic shift register is used to configure the digital bits for test structure, which contains a testing summing PGA and a testing ADM. The USB interface, firmware logic, and host side codes in jAER [121] are based on existing designs.

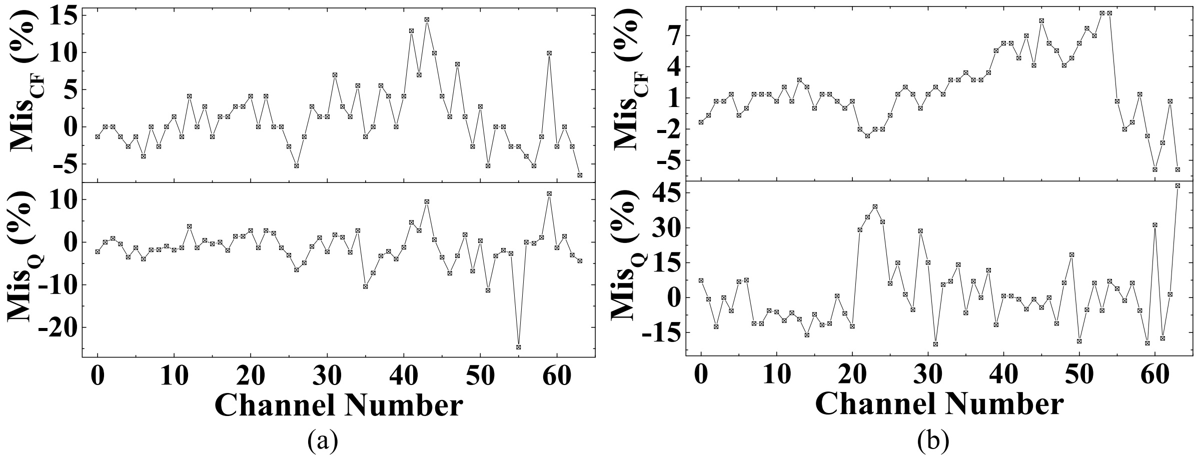
4.4 Measurement Results



(a)

(b)

Figure 4.33. Central frequency scaling as the function of channel number and the ratios of neighboring central frequencies of the transfer functions in (a) Figure 4.32(a) and (b) Figure 4.32(b).



(a)

(b)

Figure 4.34. Mismatch of the central frequencies $\mathrm{Mis}_{\mathrm{CF}}$ and Q's $\mathrm{Mis}_{\mathrm{Q}}$ between two ears at (a) $Q\approx1.3$ and (b) $Q\approx9.2$.

4.4.1 BPF Transfer Functions

All the frequency-related measurements are done with the SR780 network signal analyzer. The transfer functions of the right cochlea at low Q ($Q\approx1.3$) and high Q ($Q\approx9.2$) are plotted in Figure 4.32. The 8 bits for Q-tuning are set to 255 for all 64 channels at low Q, and are individually adjusted for each channel to have a 50-dB peak gain at high Q. Because peak gain and Q are positively correlated, the channels have approximately the same Q as well. The high Q has a 17-dB more peak gain compared to the low Q. Both of them have monotonically scaled central frequencies f_c's, as can be seen in Figure 4.33 where the ratios of neighboring f_c's are also plotted. The mean scaling ratio is about 1.13 in both cases, 2% larger than the designed value 1.108. However, accumulated over 64 stages, this seemingly small discrepancy which might be attributed to underestimated pFET leakage in the current divider in Figure 4.2 pushes the intended 32 Hz lowest frequency to around 8 Hz. The mismatches of f_c's and Q's between two ears are plotted in Figure 4.34. The higher f_c mismatch at low Q can be mostly ascribed to the difficulty in determining the peak gain frequency due to the flatter passband and the higher Q mismatch at high Q is the

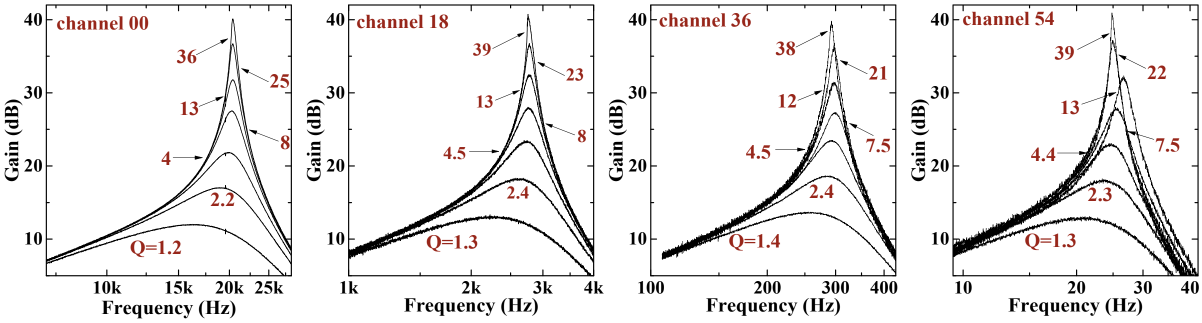


Figure 4.35. Q-tuning examples of 4 different channels.

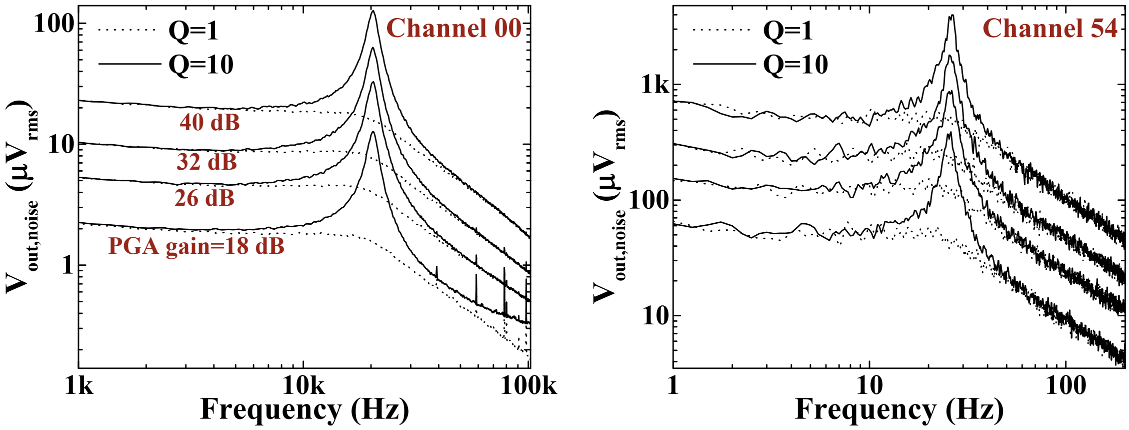


Figure 4.36. Noise PSD of channel 00 and channel 54 at the output of the summing PGA.

Table 4.8. Measured integrated rms noise at the PGA output.

Channel	00								18							
Q	1				10				1				10			
PGA gain (dB)	18	26	32	40	18	26	32	40	18	26	32	40	18	26	32	40
$V_{noise,rms}$ (mV)	0.30	0.77	1.5	3.1	0.76	1.9	3.7	7.6	0.33	0.82	1.6	3.5	0.78	2.0	3.8	8.1
Channel	36								54							
Q	1				10				1				10			
PGA gain (dB)	18	26	32	40	18	26	32	40	18	26	32	40	18	26	32	40
$V_{noise,rms}$ (mV)	0.32	0.74	1.6	3.3	0.73	1.9	3.7	7.4	0.32	0.82	1.6	3.4	0.79	2.0	3.9	8.2

direct tradeoff for well-controlled f_c matching, as analyzed in Section 4.2.4.1. Q-tunings of BPFs in four channels are demonstrated in Figure 4.35. The lowest Q's are around 1.3 and the highest can get close to 40. Note that the larger f_c variation during Q-tuning in low-frequency channels especially in channel 54 might be the indication of larger mismatch between I_{2U} and I_{2D} in Figure 4.13 due to the much smaller currents compared to high-frequency channels or nonidealities in the TLL for Q tuning.

4.4.2 BPF Noise, Distortion and Dynamic Range

The noise PSD at the output of the PGA with different PGA gains is measured for the same four channels and Figure 4.36 shows the plot of channel 00 and channel 54. The noise floor increases as the PGA gain increases and has gain peaking at the central frequency when Q=10. The integrated output rms noise

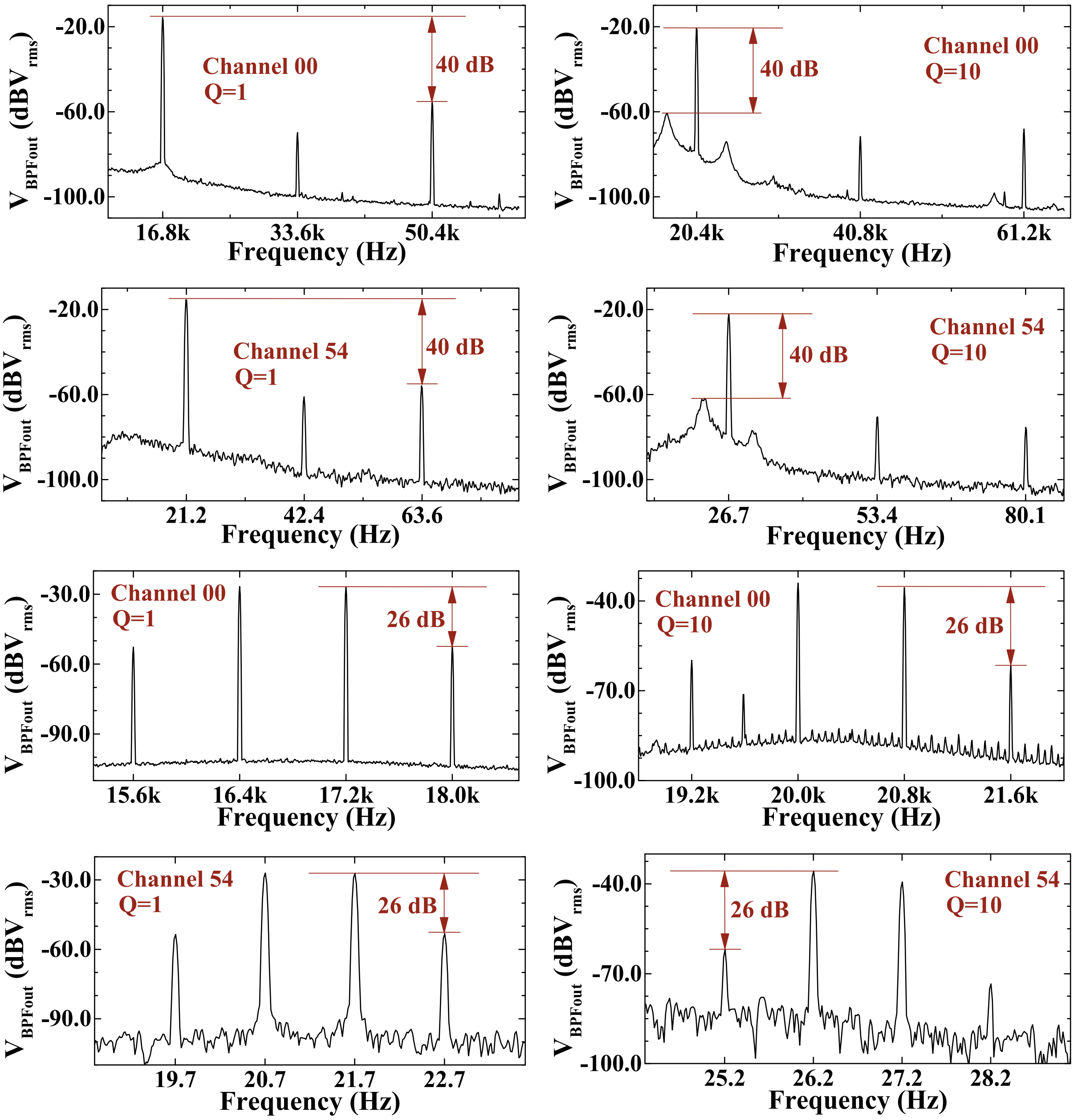


Figure 4.37. Measured distortion of THD=-40 dB and IM3=-26 dB of channel 00 and channel 54 with a PGA gain of 18 dB.

values for all the four channels are listed in Table 4.8. At the same PGA gain and *Q*, the four channels have approximately the same output noise, which is expected because the filter noise is determined by the same loading capacitance used in all channels and the PGA noise is designed to be constant over four different gains.

The distortion of one-tone and two-tone tests at *Q*=1 and *Q*=10 are plotted in Figure 4.37 for channel 00 and channel 54 at an 18-dB PGA gain. In the THD plots, the 3rd-order harmonics generally dominates over the 2nd-order thanks to the differential signal path, and it is worth noting that when *Q*=10, the dominant nonlinearity is the non-harmonic components rising beside the fundamental tone. The maximum rms

Table 4.9. Measured PGA rms output amplitude at THD=-40dB and IM3=-26dB, and the calculated SNR.

Channel	Q	PGA gain (dB)	V_{BPFout} (mV$_{rms}$)	SNR (dB)	Channel	Q	PGA gain (dB)	V_{BPFout} (mV$_{rms}$)	SNR (dB)
			THD/IM3	THD/IM3				THD/IM3	THD/IM3
00	1	18	165/91	55/50	18	1	18	181/98	55/49
		26	272/232	51/50			26	312/250	52/50
		32	261/324	45/47			32	303/370	46/47
		40	246/303	38/40			40	285/347	38/40
	10	18	92/37	42/34		10	18	98/43	42/35
		26	233/94	42/34			26	248/109	42/35
		32	251/181	37/34			32	298/199	38/34
		40	236/297	30/32			40	284/349	31/33
36	1	18	157/79	54/48	54	1	18	170/89	55/49
		26	299/203	52/49			26	257/233	50/49
		32	268/338	44/46			32	224/331	43/46
		40	232/301	37/39			40	190/293	35/39
	10	18	70/34	40/33		10	18	76/29	40/31
		26	181/87	40/33			26	194/82	40/32
		32	263/191	37/34			32	252/162	36/32
		40	216/297	29/32			40	220/305	29/31

Table 4.10. CMRR and PSRR at central frequencies measured at the output of the ADM amplifier.

Channel	00								18							
Q	1				10				1				10			
PGA gain (dB)	18	26	32	40	18	26	32	40	18	26	32	40	18	26	32	40
CMRR (dB)	51	49	50	49	40	40	40	40	53	50	51	50	37	37	37	36
PSRR (dB)	46	51	48	47	49	49	48	48	46	52	47	45	42	42	41	40
Channel	36								54							
Q	1				10				1				10			
PGA gain (dB)	18	26	32	40	18	26	32	40	18	26	32	40	18	26	32	40
CMRR (dB)	52	52	52	51	43	43	44	44	51	49	49	51	47	48	47	49
PSRR (dB)	36	44	45	44	49	48	48	47	28	37	44	46	49	46	44	44

output values at THD=-40 dB and IM3=-26 dB for all the four channels are summarized in Table 4.9, as well as the corresponding SNR. The maximum SNR for both Q=1 and Q=10 occurs at the lowest PGA gain of 18 dB. For small input signals, high PGA gains need to be used to obtain sufficiently large output swing for subsequent ADM. Ideally the gain adjustment should be achieved by automatic gain control which can be integrated in the system in the future. The system dynamic range can be extended by another 18 dB from the maximum SNR owing to the input attenuator.

4.4.3 CMRR and PSRR

The CMRR and PSRR are measured at the output of the ADM amplifier, i.e. the voltage V_{adm} in Figure 4.24. The example CMRR curves for channel 00 and channel 54 are shown in Figure 4.38, and the example PSRR curves for channel 00 are shown in Figure 4.39. The CMRR and PSRR values at central frequencies for channel 00, 18, 36 and 54 are given in Table 4.10. The significant PSRR degradation below the channel central frequencies is deemed to be aggravated by the '**-1**' analog inverter block used in the ADM. Around central frequencies, the C_{fb} in Fig. 4.25 can keep the gate voltage of M_2 tracking VDD,

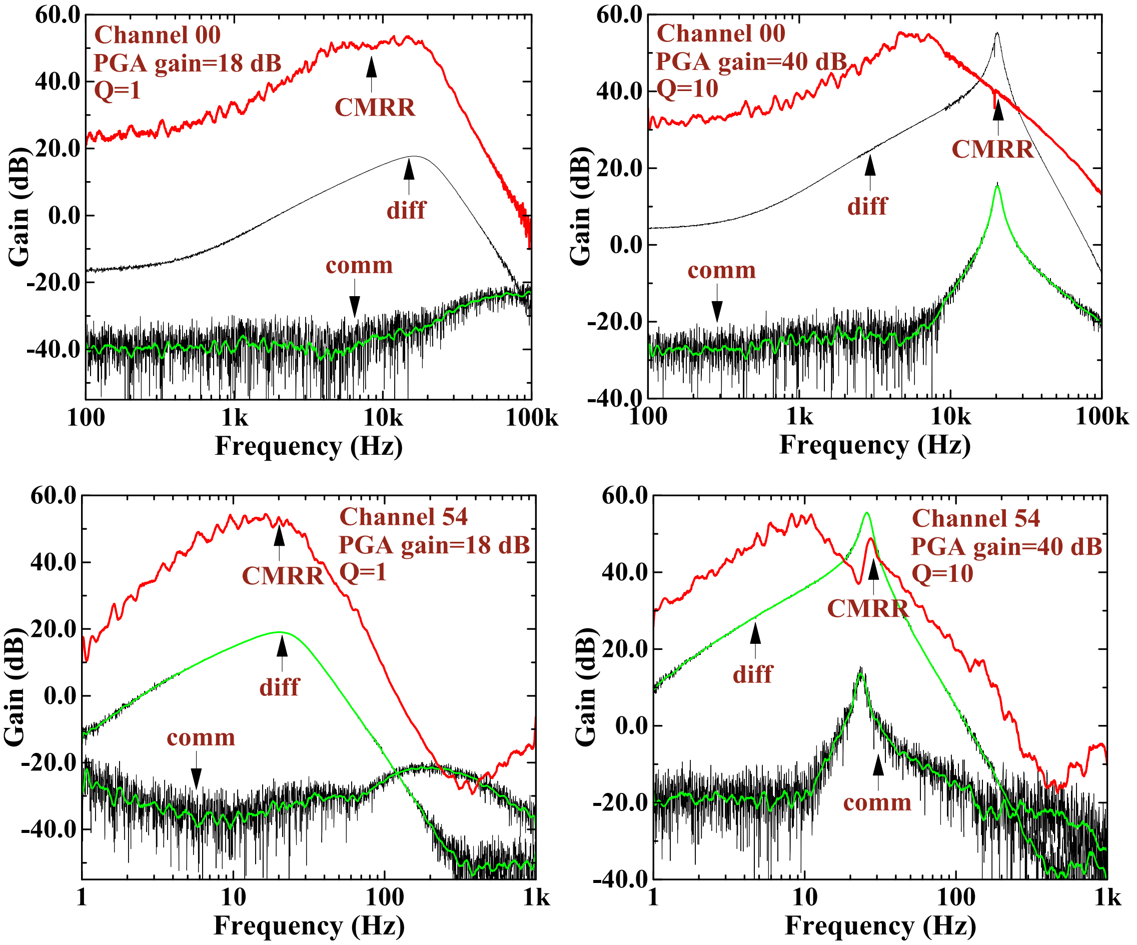


Figure 4.38. Measured CMRR at the output of the ADM amplifier.

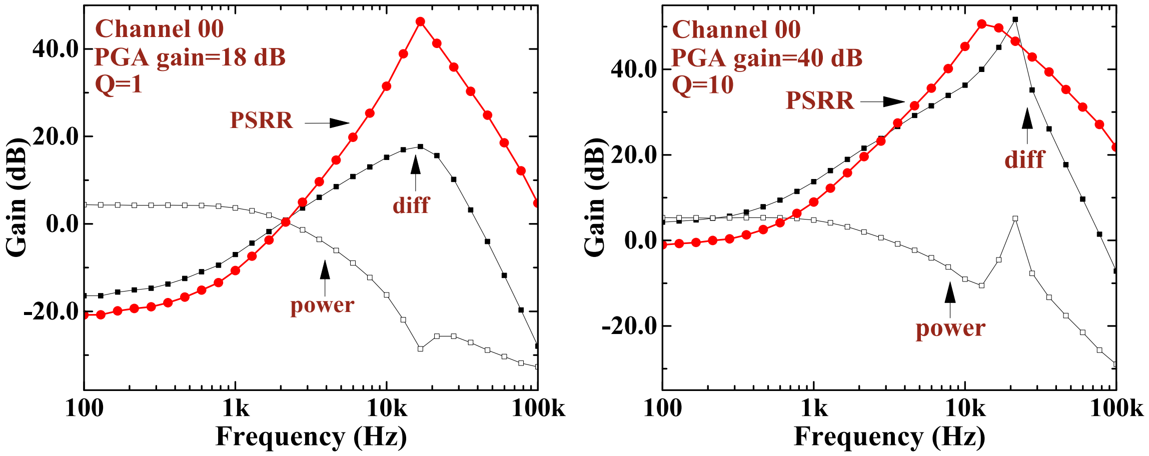
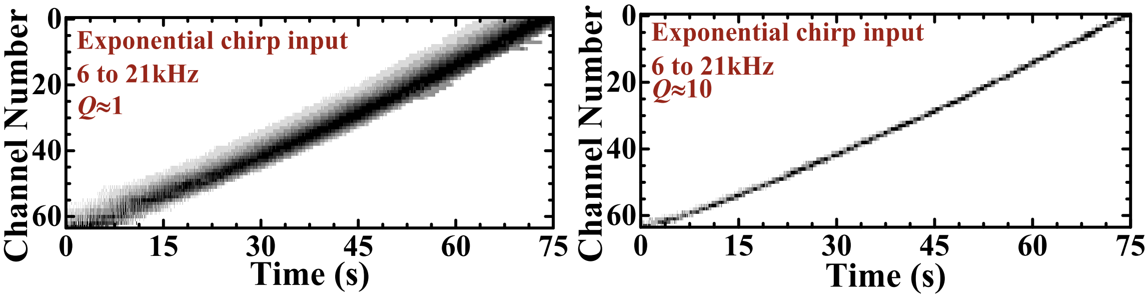


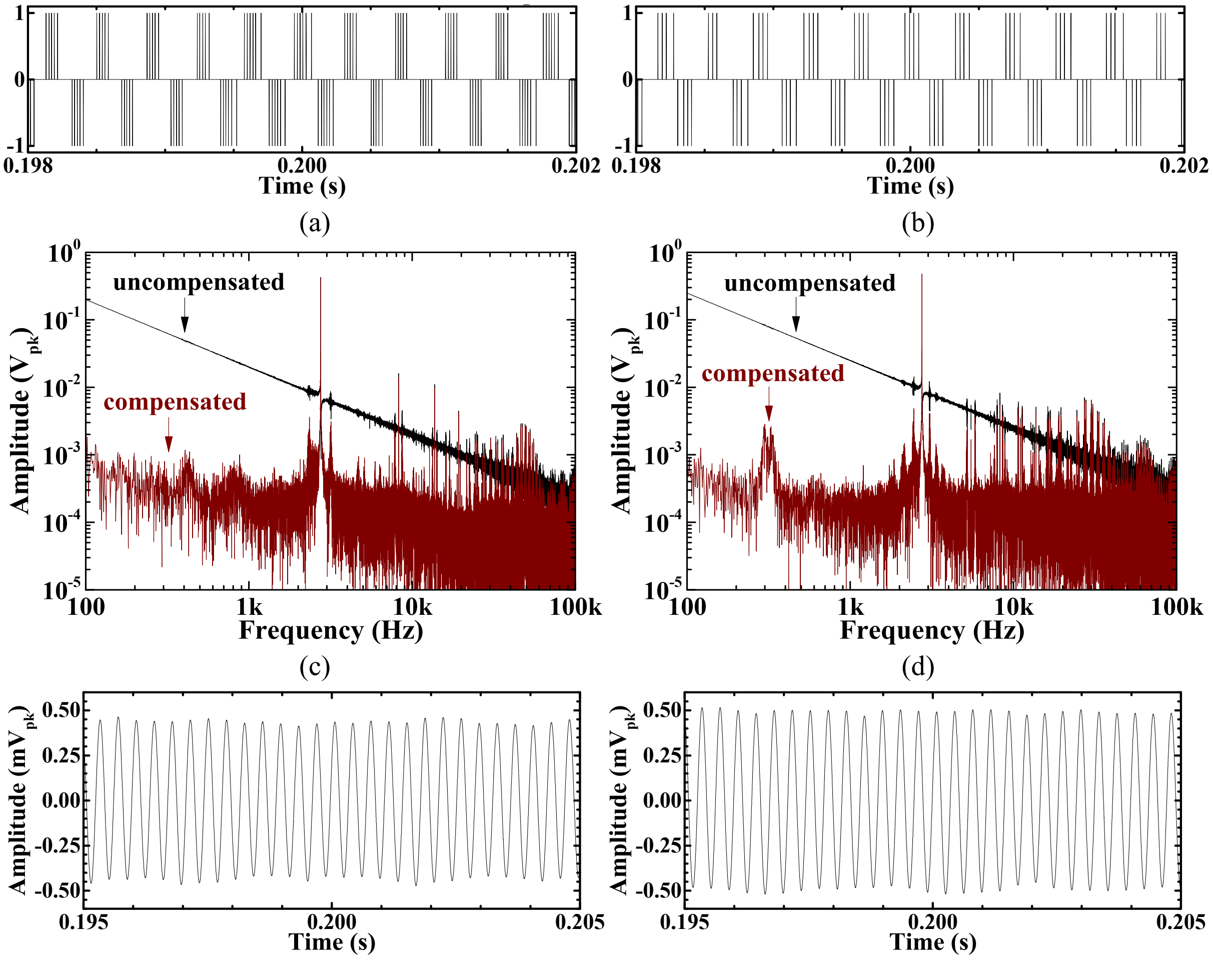
Figure 4.39. Measured PSRR at the output of the ADM amplifier.

namely the source voltage of M_2, and hence the output voltage is relatively insensitive to VDD. At low frequencies, the I_{ds} of M_2 is modulated by VDD, which directly affects the output voltage by the open-loop gain of the common-source amplifier.



(a) (b)

Figure 4.40. Spike cochleagram in response to an exponential chirp input with BPF transfer functions set as in Figure 4.32(a) for $Q \approx 1$ and Figure 4.32(b) for $Q \approx 10$.



(a) (b)

(c) (d)

(e) (f)

Figure 4.41. Output spike train samples of channel 18 in response to a 2.8k Hz sinusoid at (a) $\mathbf{b}_{rst1}\mathbf{b}_{rst0}$=01 and (b) $\mathbf{b}_{rst1}\mathbf{b}_{rst0}$=11; (c), (d) the corresponding amplitude spectra of the unfiltered reconstructed waveform with compensated and uncompensated ON and OFF thresholds; (e), (f) the uncompensated reconstruction waveform filtered by a 6th-order Butterworth HPF with a 300 Hz corner frequency and an ideal LPF with a 2.94k Hz corner frequency.

4.4.4 Spike Output

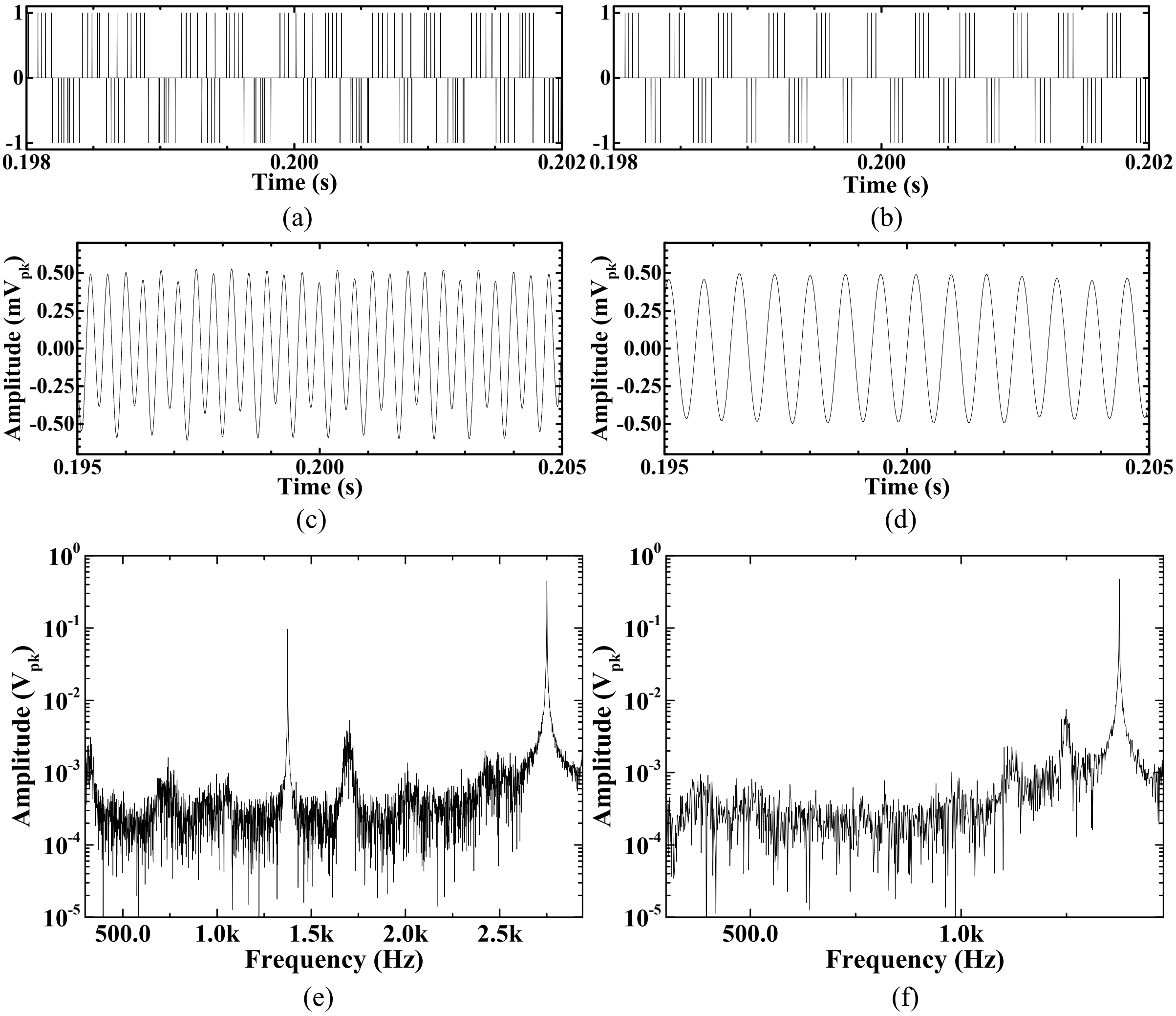


Figure 4.42. (a) Total spike train output of the cochlea chip with a two-tone input at 2.8k Hz and 1.4k Hz; (b) spike train of channel 18; reconstructed waveforms from the spike trains of (c) channel 18 and (d) channel 24, and their corresponding amplitude spectra (e) and (f).

The spike cochleagram of the chip with BPF transfer functions set as in Figure 4.32 in response to an exponential chirp input sweeping from 6 Hz to 21k Hz are shown in Figure 4.40. To plot the cochleagram, for each channel the number of spikes is binned every 0.1 s, and the maximum spike number is normalized to 1. The spike response over channels is linear with time because of the geometrically scaled central frequencies of the BPFs. It is obvious that the high *Q* cochleagram has a smaller channel response overlap in time than the low *Q* cochleagram because of the better frequency selectivity at high *Q*.

To quantify the encoding performance of a single channel, the output spike train is used to reconstruct the analog input signal employing the linear decoding methods described in Chapter 2. Spikes are recorded by the off-chip FPGA with 1-μs time resolution. Taking channel 18 as the example, the input signal is a single-tone 2.8-kHz 5-mVpk sinusoid, and the channel is configured to have a *Q*=10. The spike train samples with the configuration bits $\mathbf{b_{rst1}b_{rst0}}$ of C_{rst} set to 01 and 11 are shown in Figure 4.41(a) and 4.41(b) respectively within a 4-ms time window. Using linear decoding without filtering, the 10^6-pt FFT amplitude spectra of the reconstructed signals are shown in Figure 4.41(c) and 4.41(d). The black curve is

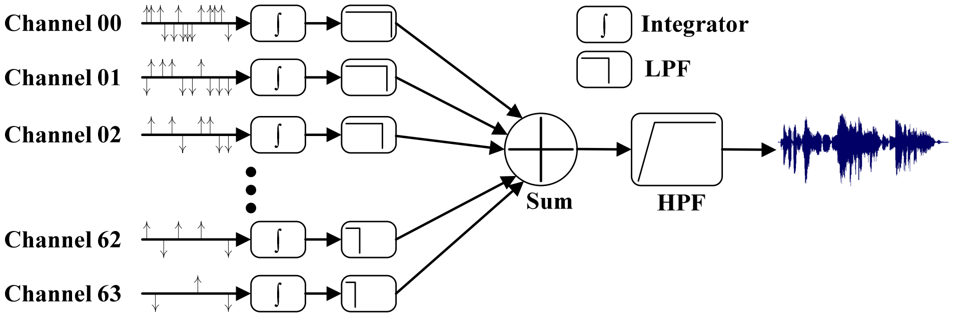


Figure 4.43. Illustration of the reconstruction method for human speech synthesis from spike trains produced by the cochlea chip.

obtained by using equal ON and OFF thresholds in reconstruction, and the dark red curve using a slightly higher ON threshold to compensate the unbalanced ON and OFF thresholds in silicon and keep a relatively stable DC level in reconstruction. The uncompensated unbalanced threshold obviously causes significantly raised noise floor at low frequencies, even though in the compensated case the ON threshold is only about 3% larger than the OFF threshold. The known channel identity can be exploited to improve the quality of reconstruction by bandpass filtering. The DC drift caused by unbalance thresholds is circumvented by passing the integrated spike train through a 6th-order Butterworth HPF with a 300-Hz corner frequency which is the lower end of the frequency range used in telephony. The high frequency harmonics are filtered by an ideal LPF with a corner frequency that equals to the mean of the central frequencies of channel 18 and 19, i.e. 2.94k Hz. After filtering, the example reconstructed time-domain waveforms with uncompensated thresholds are shown in Figure 4.41(e) and 4.41(f). Assuming the signal energy is concentrated within 20 Hz around the central frequency and using the truncated waveform for FFT (0.1 s after and before the start and end of the 1.05-s full-length waveform), the SNDR values of (e) and (f) are calculated to be 26.5 dB and 26.4 dB, respectively. As for the case of compensated thresholds, the SNDR values of the reconstructed waveform filtered by the same filters as described above at $\mathbf{b}_{\mathbf{rst1}}\mathbf{b}_{\mathbf{rst0}}$=01 and 11 are 26.8 dB and 26.2 dB, respectively.

A two-tone test is performed on channel 18 and 24 with Q set to about 10. The input signal is the combination of a 2.8k Hz and a 1.4k Hz sinusoid, both with 5-mVpk amplitude. The configuration bits $\mathbf{b}_{\mathbf{rst1}}\mathbf{b}_{\mathbf{rst0}}$ are set to 11. The total output spike train without distinguishing the spike addresses is shown in Figure 4.42(a), and the spike train from channel 18 alone is shown in Figure 4.42(b). The reconstructed waveform of channel 18 and 24 obtained by passing the integrated spike trains with uncompensated thresholds through a 6th-order Butterworth HPF with a 300 Hz corner frequency and an ideal LPF with corner frequencies at 2.94k Hz and 1.48k Hz respectively are shown in Figure 4.42(c) and 4.42(d). The corresponding amplitude spectra are shown in Figure 4.42(e) and 4.42(f).

For real-world scenario test, speech is used as the chip input. Signals are directly generated from a computer's soundcard by playing the .WAV files. The reconstruction method is depicted in Figure 4.43. The output spike train of channel i is integrated and filtered by an ideal LPF with the corner frequency at the mean of the central frequencies of channel i and i+1.The filtered signals from all 64 channels are summed and then further filtered by a 6th-order Butterworth HPF to obtain the final reconstructed waveform. Figure 4.44(a) and 4.44(b) are the spectrogram and time-domain waveform of a male speech saying '18174', respectively. Figure 4.44(c), 4.44(e) and 4.44(g) are the spike cochleagram, the spectrogram and time-domain waveform of the reconstructed signal at $Q \approx 1$. Figure 4.44(d), 4.44(f) and 4.44(h) are at

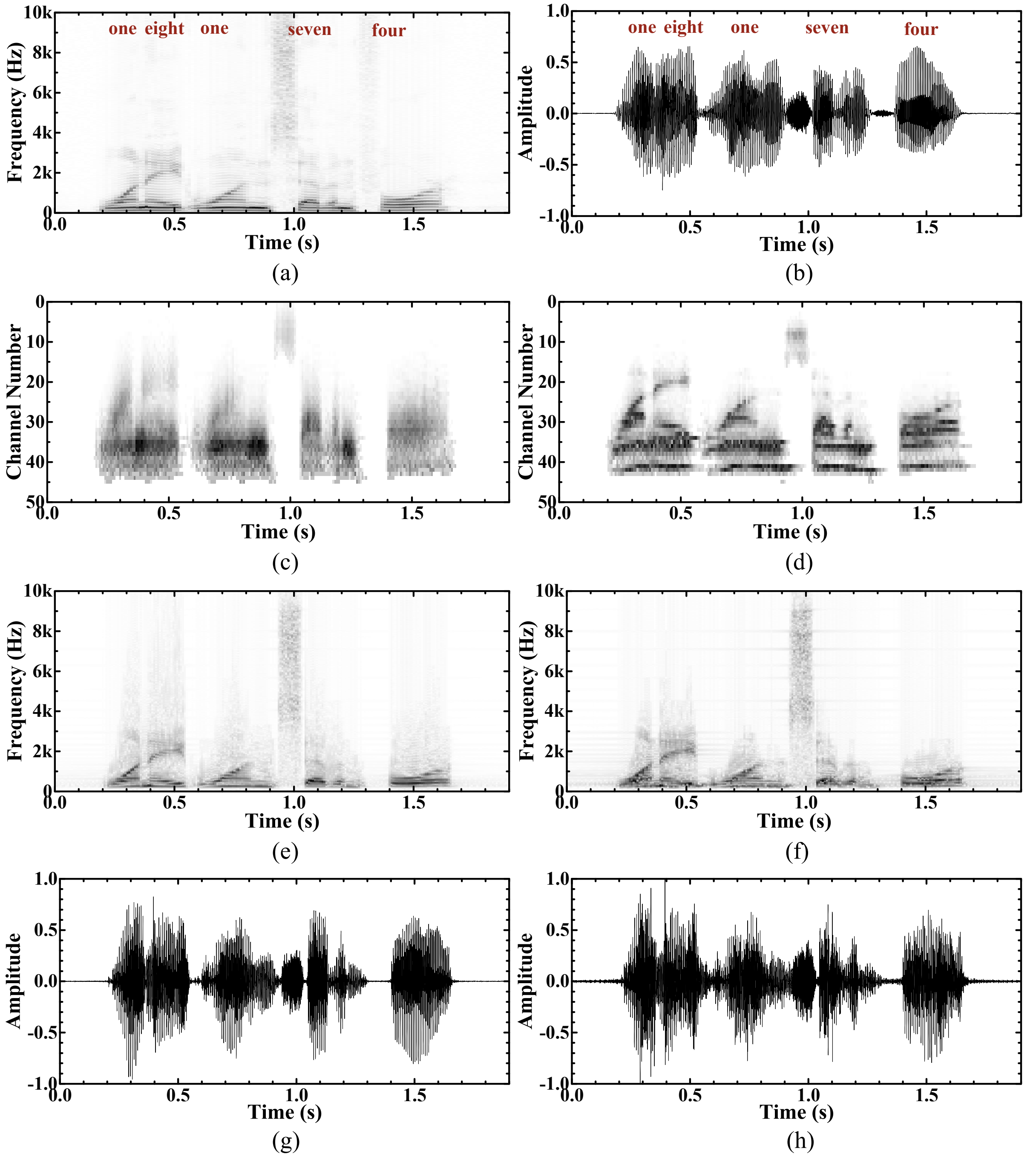
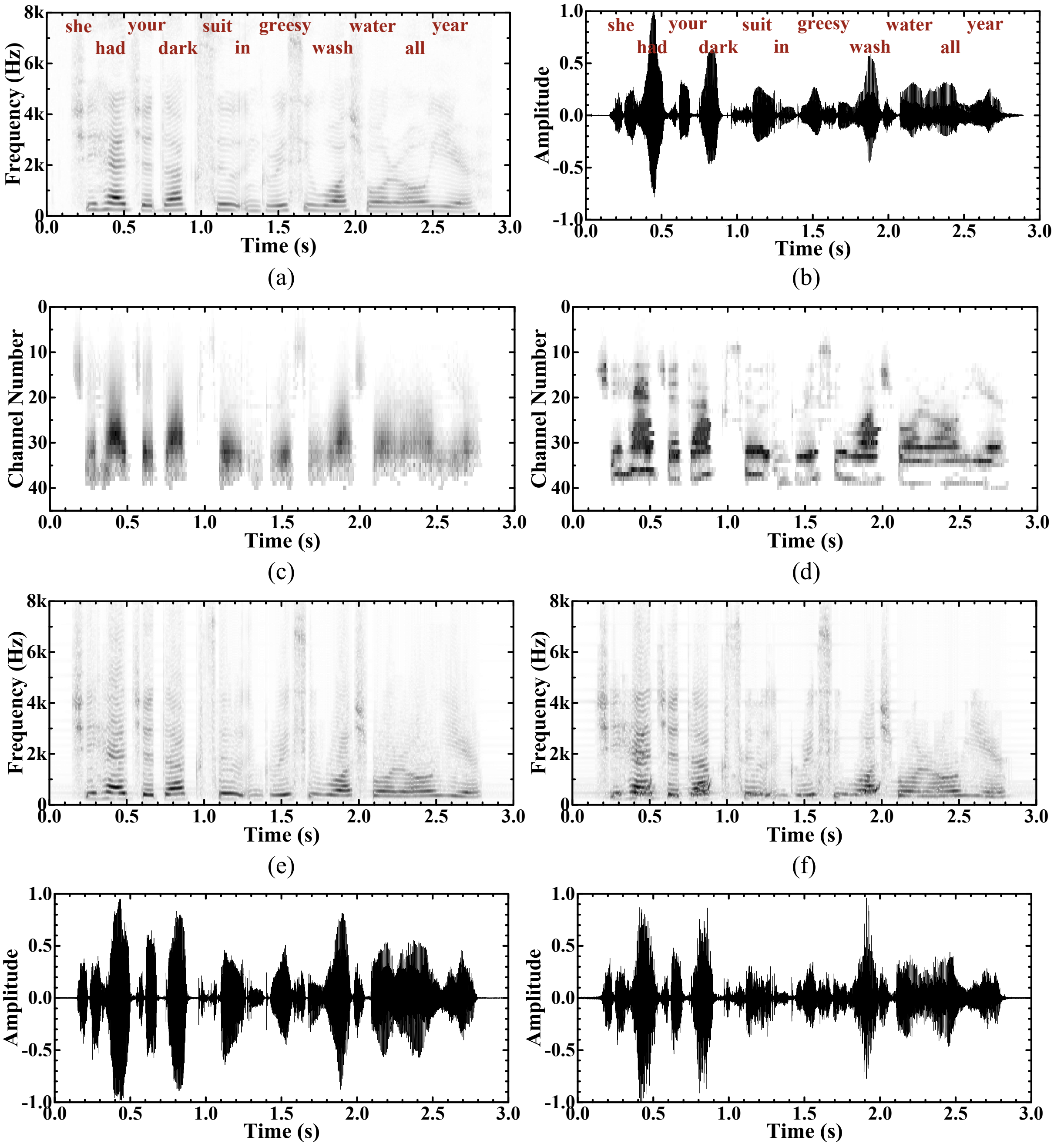


Figure 4.44. (a) Spectrogram and (b) time-domain waveform of the input male speech '18174'; (c) output spike cochleagram, and (e) spectrogram and (g) time-domain waveform of the reconstructed speech at Q≈1; (d), (f) and (h) at Q≈10.

Q≈10. The .WAV files of the original and the reconstructed audio can be downloaded from the links below:

'https://dl.dropboxusercontent.com/u/26642184/15CochLP/18174mOrigin.WAV', original '18174';

'https://dl.dropboxusercontent.com/u/26642184/15CochLP/18174mQ1.WAV' recovered at Q≈1;

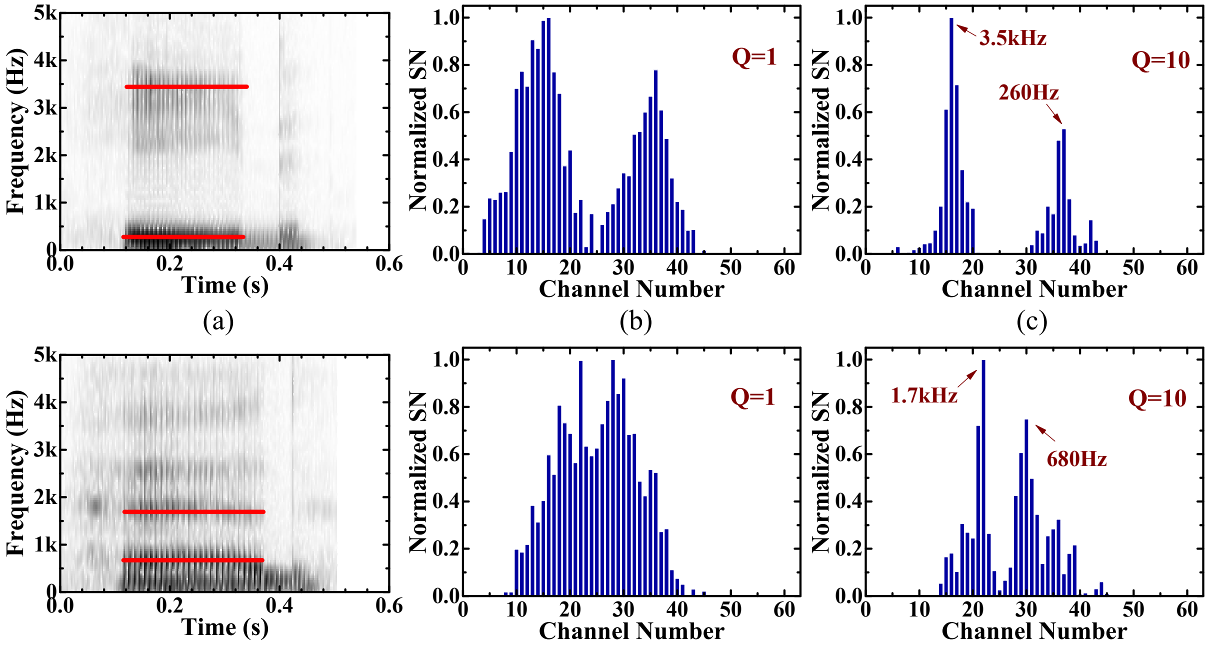


(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4.45. (a) Spectrogram and (b) time-domain waveform of the input female speech "she had your dark suit in greasy wash water all year '; (c) output spike cochleagram, and (e) spectrogram and (g) time-domain waveform of the reconstructed speech at $Q\approx1$; (d), (f) and (h) at $Q\approx10$.

'https://dl.dropboxusercontent.com/u/26642184/15CochLP/18174mQ10.WAV' recovered at $Q\approx10$.

Figure 4.45(a) and 4.45(b) are the spectrogram and time-domain waveform of a female speech 'she had your dark suit in greasy wash water all year', respectively. Figure 4.45(c), 4.45(e) and 4.45(g) are the spike cochleagram, the spectrogram and time-domain waveform of the reconstructed signal at $Q\approx1$. Fig-



(a) (b) (c)

(d) (e) (f)

Figure 4.46. Spectrogram of (a) 'heed' and (d) 'had'; normalized histogram of spike number (SN) of 'heed' at (b) Q=1 and (c) Q=10; normalized histogram of SN of 'had' at (e) Q=1 and (f) Q=10.

ure 4.45(d), 4.45(f) and 4.45(h) are at $Q \approx 10$. The .WAV files of the original and the reconstructed audio can be downloaded from the links below:

'https://dl.dropboxusercontent.com/u/26642184/15CochLP/shefOrigin.WAV', original sentence;

'https://dl.dropboxusercontent.com/u/26642184/15CochLP/shefQ1.WAV' recovered at $Q \approx 1$;

'https://dl.dropboxusercontent.com/u/26642184/15CochLP/shefQ10.WAV' recovered at $Q \approx 10$.

Subjective evaluation of the reconstruction quality indicates that the decoding methods as depicted in Figure 4.43 can synthesize well-recognizable speech from the output spike trains produced by the cochlea chip. This implies that essential speech information is preserved during the transformation from analog signals to parallel multi-frequency-channel spike trains. The high Q case is perceived to have larger distortion, which could be caused partly by the prolonged ringing of high Q BPFs, especially at large input amplitudes, or the larger distortion of the BPFs. However a high Q BPF bank may still better facilitate subsequent event-driven classification processing like voice activity detection in terms of more distinct acoustic feature extraction [163]. Figure 4.46 shows a simple example. The most two energy-significant F1 and F4 formants of the vowel 'i:' are marked in the spectrogram of the word 'heed', and the most two energy-significant F1 and F2 formants of the vowel 'æ' are marked in the spectrogram of the word 'had'. The histogram plots of normalized spike number (SN) of the two words are obtained by accumulating the spikes produced from each cochlea channel when the sound signals are sent to the chip input. As shown in Figure 4.46(b), (c) for 'heed', and 4.46(e), (f) for 'had', the formant peaks are better separated at Q=10 than at Q=1, and therefore it is easier to classify words with different vowels with a higher Q.

4.4.5 Comparison with Prior Arts

Table 4.11. Cochlea chip performance summary and comparison with prior works.

	Liu, et al. TBCAS 2014 [37]	Wen, et al. TBCAS 2009 [33]	Fragnière ISSCC 2005 [31]	Sarpeshkar, et al. ISSCC 2005 [164]	This work 2015
Architecture	Cascade	Active coupling	Passive coupling	Parallel	Parallel
Technology (µm)	0.35 CMOS	0.25 CMOS	0.5 CMOS	1.5 BiCMOS	0.18 CMOS
Power supply (V)	3.3	2.5	3.3	2.8	0.5
Power (µW)	14000[a]	35900[b]	1700[b]	60[c]	55
Channel number	64×2	360	100	16	64×2
Frequency range (Hz)	50-50k	210-14k	200-20k	100-5k	8-20k
Q range	1.5±0.4	1.16±0.92[d]	0.25-12	<10	1.3-39[e]
Normalized power[f] (µW)	291	605	78	56	3.2
Area per channel (mm^2)	0.107	0.0304	0.174	5.53	0.262
Dynamic range (dB)	52	52	50	75	73[e,g]@Q=1.3 60[e,g]@Q=10
Spike encoding	PFM	PFM	Threshold-crossing	Zero-crossing	ADM
Spike readout	AER	AER	Scanning	Scanning	AER
Binaural	Yes	No	No	No	Yes

a. preamplifier and AER readout included;

b. analog part only;

c. consist of BPFs, envelope detector and bias;

d. -10dB bandwidth Q;

e. data from channel 18, other channels are slightly different;

f. assuming geometrical scaling of currents along channels, the power consumption of the highest-frequency channel normalized to 20kHz is:

$$P_{norm} = \frac{P_{total}(1-r)}{1-r^n} \cdot \frac{20\text{k}}{f_H}, \; r = (\frac{f_L}{f_H})^{1/(n-1)}$$

where P_{total} is the total power consumption, f_H and f_L are the highest and lowest channel frequencies, and n is the channel number;

g. at 1% THD, including 18dB of the attenuator.

The performance of this cochlea chip is summarized and compared with prior works in Table 4.11. The power consumption of the 0.5-V core is 55 µW at 100k spike/s output rate. Using the normalized power metric as defined below Table 4.11, this cochlea is about 17.5× more power efficient than the best prior art, namely the bionic ear developed by Sarpeshkar et al. at MIT [164]. Besides, this cochlea has presented one of the best matching among channels not only at the BPF output but also at the output of the spike encoders. The only cochlea chip that has been used to demonstrate the information integrity from the output spike trains is the AEREAR2 [37], and the reconstruction methods described in a recent paper [165] is adapted from a linear mapping method that was used to reconstruct the auditory stimulus of a ferret from its spike responses in auditory cortex. The results show recognizable reconstructed digit speech as well, but the method requires learning procedure of new mapping if the input audio is changed from digits to some other arbitrary speech sequence. The reason that no straightforward decoding method like the one depicted in Figure 4.43 could be developed is mostly likely because of the not well-controlled circuit and system parameters in AEREAR2.

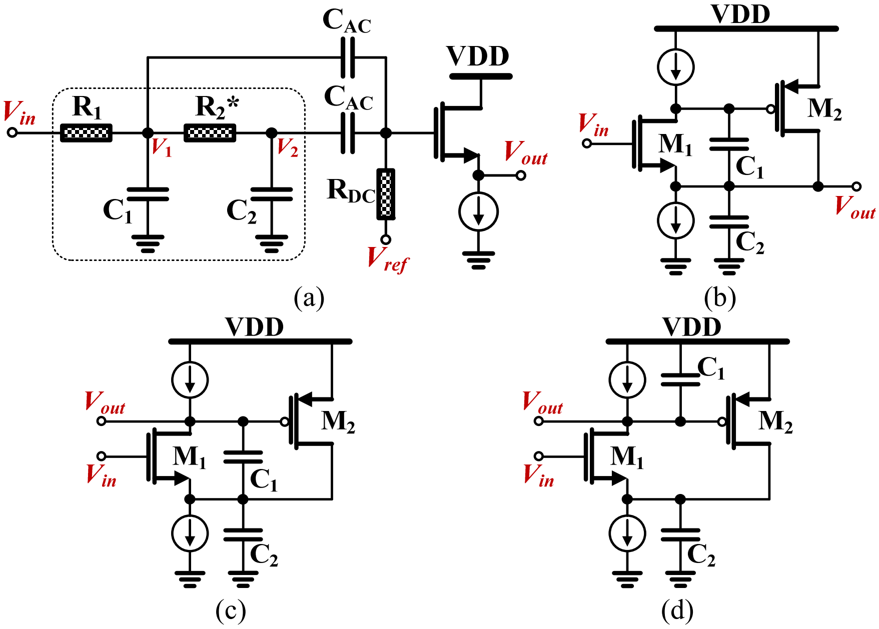


Figure 4.47. (a) Alternative summation scheme, (b) super-source-follower-based LPF [144], and (c) and (d) alternative topologies to obtain one zero.

4.5 Conclusion and Discussion

An ultra-low-voltage and ultra-low-power 64×2-channel binaural silicon cochlea is presented in this chapter. Two main techniques are proposed to achieve the 17.5× improvement in power efficiency compared to the best prior art [164]. One is the source-follower-based BPF, and the other is the ADM with adaptive self-oscillating comparison. The source-follower-based BPF is composed of a source-follower-based LPF and a summation block which in this design is particularly implemented as a summing PGA. In general, the summation block can have any form. One much more power-efficient solution is illustrated in Figure 4.47(a), where the summation is achieved simply by two AC-coupling capacitors. An alternative way of obtaining one zero is illustrated in Figure 4.47(c). The circuit in Figure 4.47(b) is the so-called super-source-follower-based LPF [144], where the input and output are the gate and source of M_1, respectively. By simply changing the output from the source to the drain of M_1, a BPF transfer function can be obtained:

$$H(s) = -\frac{s\cdot\frac{C_2}{g_{m2}}}{s^2\cdot\frac{C_1C_2}{g_{m1}g_{m2}} + s\cdot\frac{C_1}{g_{m1}} + 1},\ f_0 = \frac{1}{2\pi}\sqrt{\frac{g_{m1}g_{m2}}{C_1C_2}},\ Q = \sqrt{\frac{g_{m1}C_2}{g_{m2}C_1}},\ K = \frac{g_{m1}C_2}{g_{m2}C_1}$$

where g_{m1} and g_{m2} are the transconductance of M_1 and M_2, respectively. One obvious shortcoming of this topology is that its peak gain is the square of its Q, and hence its input dynamic range is largely limited with a high Q setting. However, this topology can be used as a low-Q BPF cascaded by the SF-based LPF so that a high Q BPF can be achieved by tuning the Q of the LPF using the same method presented in Section 4.2.4.1. A topological mutation similar to the one in [166] for RF applications is shown in Figure 4.47(d). The only difference is that the capacitor C_1 is now connected between the gate and source of M_2 instead of its gate and drain. The transfer function and the parameters are given below:

$$H(s) = -\frac{s \cdot \frac{C_1}{g_{m2}}}{s^2 \cdot \frac{C_1 C_2}{g_{m1} g_{m2}} + s \cdot \frac{C_2}{g_{m2}} + 1},\ f_0 = \frac{1}{2\pi}\sqrt{\frac{g_{m1} g_{m2}}{C_1 C_2}},\ Q = \sqrt{\frac{g_{m2} C_2}{g_{m1} C_1}},\ K = \frac{C_2}{C_1}$$

This topology does not have the drawback of quadratic dependence of K on Q, but large transconductance ratio and capacitance ratio are still needed to obtain a high Q.

Several open questions remain to be answered to further advance silicon cochlea and its related systems. How an automatic gain control block can be integrated to dynamically control the gain and Q of the analog front-end so that a larger audio input dynamic range can be self-adaptively accommodated? Existing examples can be found in cochlea papers [29] or even in wireless receivers [167]. From an SoC system point of view, if this cochlea is to be integrated with a spike-processing stage for functions like voice activity detection, how the linearity of the analog front-end and the spike encoding quality can affect the ultimate performance, e.g. classification accuracy? What kind of sensing and processing co-optimization can be done to achieve the best possible performance with constraints like available power budget? A recent interesting overview article [163] written by Marian Verhelst et al. shared some useful points for future research in the era of internet of everything (IoE). She specifically stated that unlike in traditional multimedia applications where faithful signal reconstruction is often necessary, many IoE applications do not have such requirement and therefore during the analog-to-digital conversion or analog-to-spike in our case, the sampling rate could be much lower than the Nyquist rate. Only essential information or feature should be preserved in the conversion, and thus saving both conversion and processing energy. This statement in my opinion coincides with the biological reality. The spiral ganglion cells in the cochlea can hardly achieve a firing rate larger than 1k spike/s, but we are still able to effortlessly understand human speech which undoubtedly contains frequency components above 500 Hz. Collaboration with information theorists is indispensable to better understand the working principle of smart spiking sensory systems and implement them in silicon with ever-decreasing power and increasing performance.

4.6 Appendix

4.6.a Bandwidth of the Summing CC-PGA

For a summing CC-PGA, if the Opamp has a finite DC gain of A_0 and a dominant pole at p_1, the closed-loop transfer function and the cutoff frequency f_{-3dB} can be written as:

$$H_{CC-PGA}(s) = -\frac{A_0 p_1 C_{in}}{s \cdot (2C_{in} + C_{fb}) + p_1(2C_{in} + C_{fb} + A_0 C_{fb})} \approx \frac{C_{in} / C_{fb}}{s \cdot 2C_{in} / (A_0 p_1 C_{fb}) + 1},\ f_{-3dB} = \frac{1}{2\pi}\frac{A_0 p_1 C_{fb}}{2C_{in}} = \frac{1}{2\pi}\frac{\text{GBW}}{2A_c}$$

where C_{in} and C_{fb} are the input and feedback capacitances, respectively, and GBW and A_c are the Opamp gain-bandwidth product and CC-PGA closed-loop gain, respectively. Therefore, given the requirement of f_{-3dB} larger than the BPF central frequency for minimum peak gain loss, the GBW should be $2\pi \cdot 2A_c$ times larger instead of intuitively $2\pi \cdot A_c$ times larger. Unfortunately, this was not realized during the design until approaching the tape-out deadline. The penalty of $\text{GBW} < 2\pi \cdot 2A_c f_{-3dB}$ is gain loss at the BPF central frequency and increased peak gain variation. In the fabricated design, $\text{GBW} \approx 2\pi \cdot 1.3 A_c f_{-3dB}$, and the peak gain loss is around 6 dB.

4.6.b Transient Analysis of the ADM Amplifier

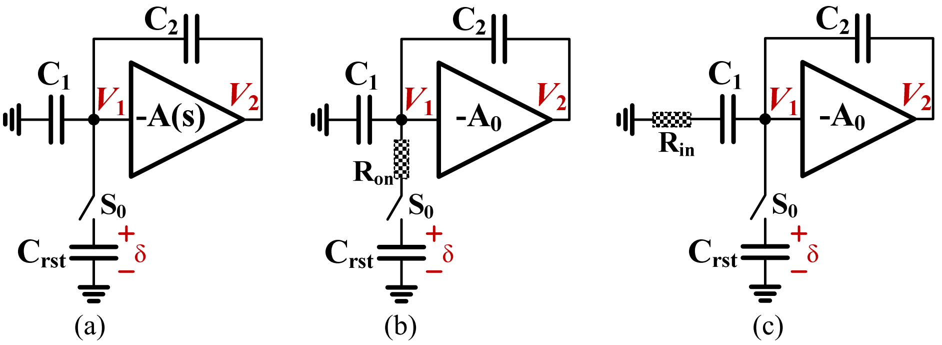


Figure 4.a1. ADM amplifier models considering (a) the limited Opamp bandwidth, (b) the switch resistance and (c) the source resistance.

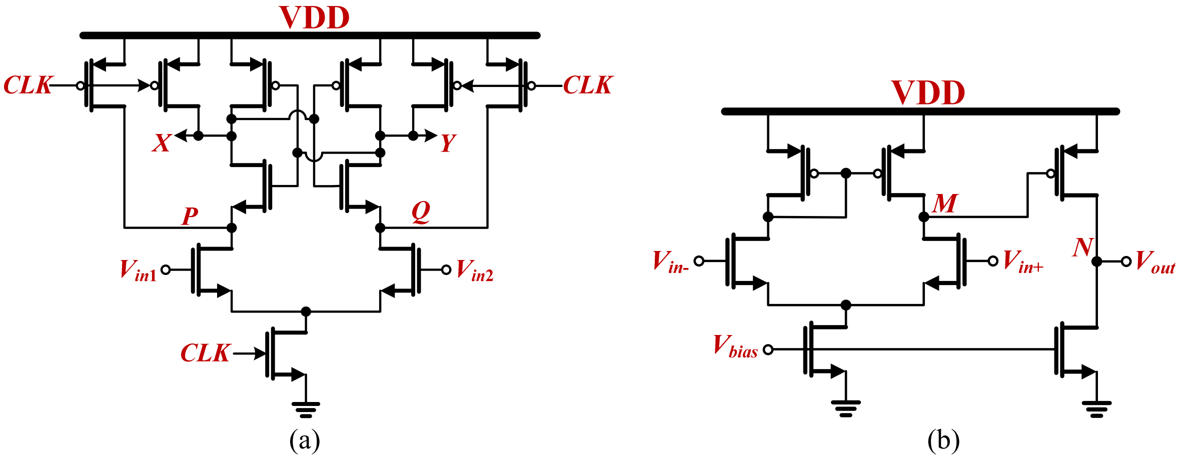


Figure 4.a2. (a) StrongARM latch DT-comparator [162] and (b) a commonly-used CT-comparator based on a two-stage amplifier without compensation.

The abstract model of the ADM amplifier is illustrated in Figure 4.a1. First let us consider the effect of limited Opamp bandwidth. We use the model in Figure 4.a1(a) where S$_0$ is an ideal switch with zero resistance. The initial voltages across C$_1$, C$_2$ and C$_{rst}$ are assumed to be 0, 0 and δ, respectively. According to KCL and the Opamp's transfer function, we have the following equations:

$$(s\cdot V_1-\delta)\cdot C_{rst}+s\cdot V_1\cdot C_1+s\cdot(V_1-V_2)\cdot C_2=0\,,\ V_1\cdot(-A(s))=V_2\,,\ A(s)=\frac{A_0}{s/p_1+1}$$

where A_0 is the DC gain of the Opamp, and p_1 is the dominant pole. The output V_2 is solved as:

$$V_2(s)=-\frac{\dfrac{A_0C_{rst}\delta}{C_1+C_2+C_{rst}}}{s(\dfrac{s}{p_1}+1+\dfrac{A_0C_2}{C_1+C_2+C_{rst}})}$$

With inverse Laplace transform, the V_2 in time domain can be found as:

$$V_2(t)=-\frac{A_0C_{rst}\delta}{C_1+C_2+C_{rst}+A_0C_2}(1-e^{-p_2t})\,,\ p_2=p_1\cdot(1+\frac{A_0C_2}{C_1+C_2+C_{rst}}) \qquad (4.a1)$$

Figure 4.a1(b) and Figure 4.a1(c) are the abstract models considering the switch resistance R$_{on}$ and the source resistance R$_{in}$ (the output resistance of the last circuit stage), respectively. Infinite Opamp bandwidth is assumed in both cases. Using the same methods above in derivation, the amplifier output V_2 in Figure 4.a1(b) can be obtained as:

$$V_2(t) = -\frac{A_0 C_{rst}\delta}{C_1 + C_2 + C_{rst} + A_0 C_2}(1 - e^{-p_3 t}),\ p_3 = \frac{C_1 + C_2 + C_{rst} + A_0 C_2}{(C_1 + C_2 + A_0 C_2) C_{rst} R_{on}} \tag{4.a2}$$

And the V_2 in Figure 4.a1(c) can be obtained as:

$$V_2(t) = -\frac{A_0 C_{rst}\delta}{C_1 + C_2 + C_{rst} + A_0 C_2}(1 + \frac{C_1}{C_2 + C_{rst} + A_0 C_2} e^{-p_4 t}),\ p_4 = \frac{C_1 + C_2 + C_{rst} + A_0 C_2}{(C_2 + C_{rst} + A_0 C_2) C_1 R_{in}} \tag{4.a3}$$

4.6.c CT- and DT-Comparator Power Consumption Comparison

The average power consumption of a representative DT-comparator, the StrongARM latch as shown in Figure 4.a2(a) can be estimated as [162]:

$$P_{avg} = (2C_{P,Q} + C_{X,Y}) f_{CLK} VDD^2$$

where C_P and C_Q are the parasitic capacitances at the nodes P and Q, and C_X and C_Y are at the nodes X and Y. This power estimation holds also for the two-stage DT-comparator used in the ADM design. For the 20k-Hz cochlea channel, an oscillation frequency that is 114× the central frequency is obtained in simulation under typical corner when the input voltage difference is 1 mV, i.e. f_{CLK}=2.28M Hz. For simplicity, let us assume that $C_{P,Q}=0.5C_{X,Y}=C_{par0}$. $C_{X,Y}$ is made larger considering capacitive loading at the output. With VDD=0.5 V, the average power becomes:

$$P_{DT} = 2.28\times10^6 C_{par0}$$

The largest comparison delay in the oscillation loop has the upper limit of $1/f_{CLK}$. Now let us compute the power consumption of a CT-comparator with a step response delay of $1/f_{CLK}$. The CT-comparator is a two-stage open-loop amplifier without compensation as shown in Figure 4.a2(b). The open-loop DC gain is usually above 5000, and thus the sensitivity is at least 0.1 mV given a 500-mV output swing. Two dominant poles are located at nodes M and N, and they are assumed to have the same capacitance as $C_{P,Q}$ and $C_{X,Y}$, respectively. Setting the two pole frequencies the same is a good tradeoff between comparison speed and power consumption, and if the input step is 1 mV, the pole locations can be calculated as [74]:

$$p_M = p_N = \frac{f_{CLK}}{\sqrt{1mV/0.1mV}} = 7.21\times10^5\ rad/s$$

where p_M and p_N are the pole frequencies of nodes M and N respectively. If transistors with L=0.5 µm is used, the same as those in the DT-comparator, the Early voltages of the nFET and pFET are simulated to be V_{En}=7.91 and V_{Ep}=7.66 V, respectively. The power consumption of the CT-comparator can be then calculated as:

$$P_{CT} = \frac{p_M \cdot C_{par0} \cdot 2 + p_N \cdot 2C_{par0}}{\frac{1}{V_{En}} + \frac{1}{V_{Ep}}} \cdot VDD = 5.61\times10^6 C_{par0}$$

This is more than twice the P_{DT}. In fact, much more power consumption is needed for CT-comparators in high-frequency channels under a low supply voltage because large W/L ratios are necessary to keep transistors in saturation which results in large extra parasitic capacitances. In DT-comparators, only the initial saturation of the input pair transistors are required to provide certain voltage gain before the regenerative latch takes effect, and this is guaranteed by charging the drain nodes to VDD in the reset phase (discharging to ground in the case of pFET input); therefore the transistor sizes in DT-comparators can be much smaller than those in CT-comparators. The two-stage DT-comparator [161] is used in the fabricated design instead of the conventional StrongARM latch because of its speed advantage with less stacking between ground and VDD [168].

Chapter 5: Addressable Current Reference Array for Spiking Sensors

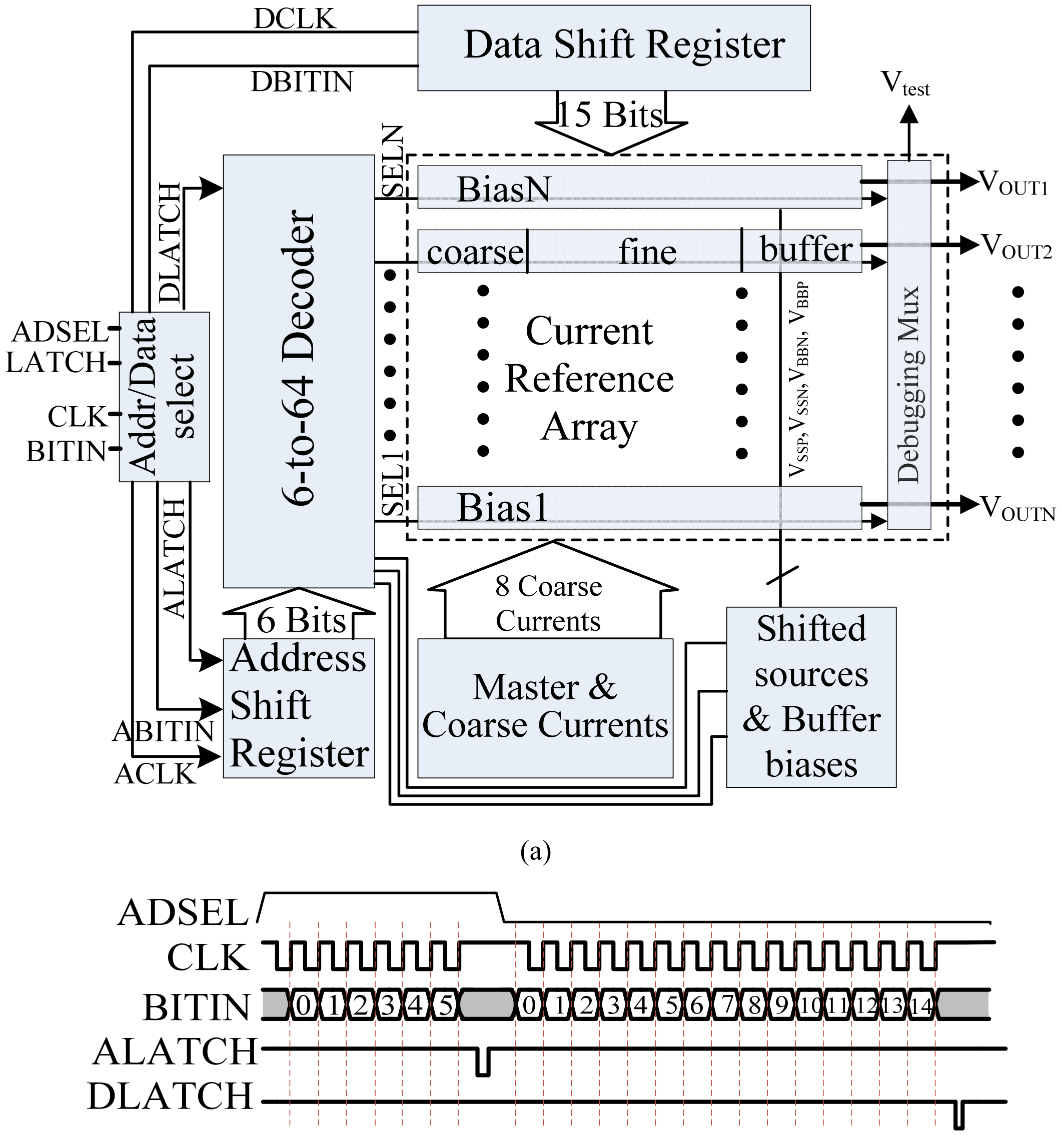
This chapter describes the addressable current reference array with wide dynamic range [120] that has been used in many of the spiking sensors developed in Sensors Group at Institute of Neuroinformatics, including the silicon retina and the silicon cochlea presented in Chapter 3 and in Chapter 4, respectively. Coarse-fine architecture is employed to cover a 174 dB DR (from 50 fA to 25 μA) of digitally programmable bias currents with extendible number of bias branches.

Configurable bias currents are desirable in many analog/mixed-mode circuits and systems, especially in array sensors [37], [89] and neuron array [169], where the circuit parameters require wide tuning. The tuning range can span from strong inversion of transistors to below leakage currents. A previous design implemented a digitally programmable current array with a nominal 22-bit precision [170]. It used shifted-source current mirror (SSCM) to enable sub-leakage current copying [129], and in turn increased the effective current DR. The capability will become increasingly important as the channel leakage continues increasing with technology scaling. However, the current array was not addressable because all the shift registers (SRs) for digital configuration were connected in series, i.e. to reprogram one bias, the whole SR chain needs to be rewritten even though other biases are in no need of changing. Each bias achieved 110 dB DR with the constant 22-bit precision; however, for large current values, such high precision adjustment is not necessary. A more reasonable design would be to adjust a nominal bias with certain precision and the nominal bias can be roughly tuned over a wide DR. This coarse-fine tuning strategy is exploited in the design of the new addressable current reference array.

The content of this chapter is organized as follows: Section 5.1 describes the coarse-fine system architecture and the timing of the digital signals for bias programming; Section 5.2 gives the detailed circuit implementation, including: Section 5.2.1, the proportional-to-absolute-temperature (PTAT) master current; Section 5.2.2, the eight coarse currents with a scaling ratio of 8 generated by using compact current divider and multiplier; Section 5.2.3, the coarse-fine current interface; Section 5.2.4, the 8-bit compact current DAC; Section 5.2.5, the configurable current buffer to generate the output bias voltage; Section 5.3 presents the experimental results; Section 5.4 concludes the chapter with several remarks.

5.1 Coarse-Fine System Architecture

The overall architecture of the addressable current reference array and the timing of the main digital configuration signals are given in Figure 5.1. To select one bias branch in the array for programming its current, the address SRs receive the 6-bit address which is then decoded by the one-hot 6-to-64 decoder. After the bias branch selected by the decoder, the data SRs receive the 15-bit configuration data which are loaded into the selected branch, and the final output bias voltage V_{outi} ($i \in [1,N], N \leq 61$) generated from the diode-connected half of a current mirror (CM) in the output buffer settles accordingly after some time. New configuration bits can be loaded into the data SRs without changing the address. At a clock frequency of 100k Hz, new data can be written to program one branch in <200 μs which is much shorter than the several ms in the previous long SR chain design [170]. Because both the address and data SRs share the same clock, latch and bit input to save pads, the one bit ADSEL is used to steer the flow of these inputs to either address (ADSEL=1) or data (ADSEL=0) SRs. The 15 configuration bits are composed of: 3 bits for coarse current selection, 8 bits for fine current selection, and 4 bits for output buffer configuration. Note that the number of bias branches is 61 instead of 64, because 3 addresses are reserved for 2 branches that



(a)

(b)

Figure 5.1. (a) Coarse-fine architecture of the addressable current reference array; (b) timing diagram of the main digital signals for programming bias currents.

generate regulated voltages V_{SSN} and V_{SSP} to support SSCM configuration, and 1 branch that generates the buffer biases V_{BBN} and V_{BBP} needed in the configurable buffer in each regular bias branch.

The 8 coarse currents are derived from a classic PTAT master bias current, ranging from about 13 pA to 25 µA, and are broadcasted in voltages to all the bias branches through a half CM. A selected copied coarse current in each bias branch is fed into a compact DAC with 8-bit resolution to generate a fine-tuned current which is then buffered to generate the output bias voltage V_{OUTi}. For debugging and measuring the generated bias current, each one of the V_{outi}'s can be selected by an analog mux to V_{test} as the gate voltage to an nFET or pFET with a large W/L ratio whose drain current can be directly measured by a Keithley 236.

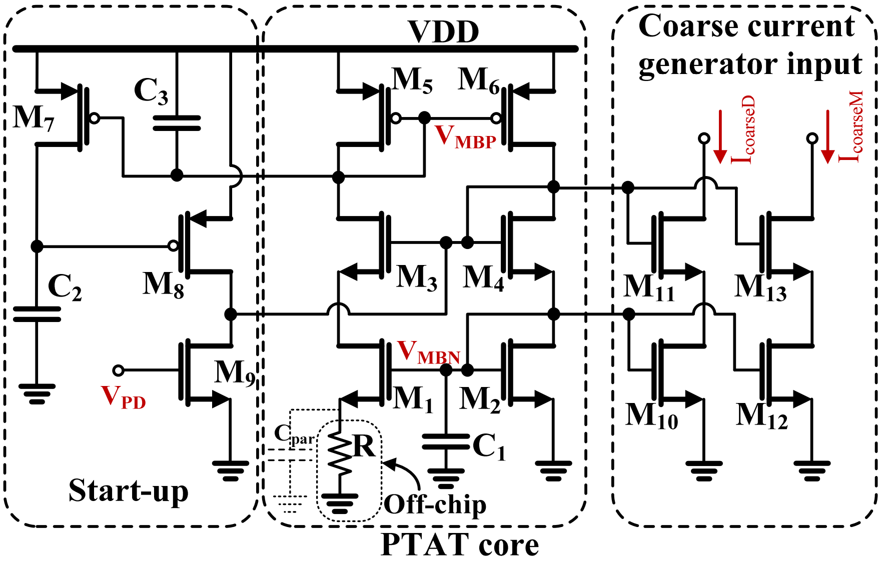


Figure 5.2. Circuit diagram of the PTAT master bias.

5.2 Circuit Implementation

5.2.1 PTAT Master Bias

The core of the PTAT master bias uses the classic positive-feedback CM topology in subthreshold [171] as circled in 'PTAT core' in Figure 5.2. The sizes of M_3 and M_5 are identical with that of M_4 and M_6, respectively. M_1 and M_2 are composed of multiple unit transistors M_u and the number of M_u in M_1 is 4 times that in M_2. The current in the off-chip resistor R, i.e. the master bias current, can be derived as:

$$I_R = \frac{V_T}{\kappa R} \ln \frac{(W/L)_1}{(W/L)_2} = \frac{V_T}{\kappa R} \ln 4 \tag{5.1}$$

where V_T is the thermal voltage, κ is the subthreshold slope factor, and R is the resistance of the off-chip resistor. V_T is proportional to absolute temperature, and so is I_R. C_1 is added to prevent the limit-cycle oscillation induced by the parasitic capacitance C_{par}. With R=100k Ω, the simulated I_R under room temperature is about 389 nA. The start-up circuit is to force the PTAT core not to enter the zero-current mode when the circuit is powered up. The digital control signal V_{PD} at the gate voltage of M_9 is to power down the master bias and in turn the whole current reference array when it is high. Two copies of I_R, $I_{coarseD}$ and $I_{coarseM}$ are sent to the coarse divider and multiplier respectively, which will be described in the next section. Cascode CMs are used to mitigate the copying inaccuracy caused by channel length modulation.

5.2.2 Eight Coarse Currents

The generated eight coarse currents are listed in Table 5.1 with the PTAT master bias current denoted as $I_{coarse3}$. The scaling ratio is approximately 8. Hence, to obtain $I_{coarse2}$ and $I_{coarse1}$, $I_{coarse3}$ needs to be multiplied by 8^1 and 8^2, respectively, and to obtain $I_{coarse4}$~$I_{coarse8}$, $I_{coarse3}$ needs to be divided by 8^1~8^5, respectively. Traditional CMs can be used for the circuit implementation which, however, is not area-efficient. The current division thus utilizes the same compact current-splitting technique [128] that has been employed to generate the geometrically scaling currents in silicon cochlea in Chapter 4. The proposed current multiplication as illustrated in Figure 5.3 can be seen as a reverse application of the technique. The

Table 5.1. Simulated nominal coarse current under room temperature.

Coarse Current	$I_{coarse1}$	$I_{coarse2}$	$I_{coarse3}$	$I_{coarse4}$	$I_{coarse5}$	$I_{coarse6}$	$I_{coarse7}$	$I_{coarse8}$
Value (A)	24.8μ	3.16μ	388n	49.7n	6.39n	821p	105p	13.4p
LSB (A)	97.01n	13.34n	1.647n	202.5p	25.21p	3.198p	409.9f	55.72f
Current Ratio	$\frac{I_{coarse1}}{I_{coarse2}}$	$\frac{I_{coarse2}}{I_{coarse3}}$	$\frac{I_{coarse3}}{I_{coarse4}}$	$\frac{I_{coarse4}}{I_{coarse5}}$	$\frac{I_{coarse5}}{I_{coarse6}}$	$\frac{I_{coarse6}}{I_{coarse7}}$	$\frac{I_{coarse7}}{I_{coarse8}}$	
Value	8.05	8.02	7.81	7.78	7.81	7.83	7.91	

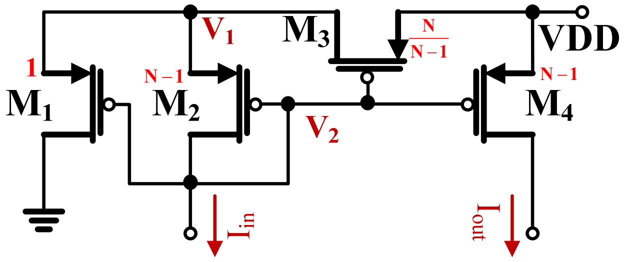


Figure 5.3. Circuit diagram of a compact current multiplier.

validity of its function can be checked by the following calculation. All the transistors are composed of multiple unit transistors M_u with the red number labeled close to their sources as the M_u number. Assume all the transistors are in above-threshold for the relatively large currents from $I_{coarse1}$ to $I_{coarse3}$ in Table 5.1. With the simplest square law [26], I_{in} and I_{out} in Figure 5.3 can be written as:

$$I_{in} = \frac{1}{2}\mu_p C_{ox}(\frac{W}{L})_u (N-1)(V_2 - V_1 - V_{th})^2$$

$$I_{out} = \frac{1}{2}\mu_p C_{ox}(\frac{W}{L})_u (N-1)(V_2 - V_{DD} - V_{th})^2$$

M_3 is in linear region, and its I_{ds} is the sum of I_{in} and the I_{ds} of M_1:

$$\frac{N}{N-1} I_{in} = I_{dsM3} = \mu_p C_{ox}(\frac{W}{L})_u \frac{N}{N-1}[(V_2 - V_{DD} - V_{th})(V_1 - V_{DD}) - \frac{1}{2}(V_1 - V_{DD})^2]$$

The equation above gives rise to the equality below:

$$\frac{1}{2}(V_2 - V_{DD} - V_{th})^2 = \frac{N}{N-1}[(V_2 - V_{DD} - V_{th})(V_1 - V_{DD}) - \frac{1}{2}(V_1 - V_{DD})^2]$$

Therefore, it is easy to draw the conclusion:

$$I_{out} = (N-1) I_{dsM3} = N I_{in} \tag{5.2}$$

Eq. (5.2) still holds if subthreshold current equation is applied. If N=8, M_2 is composed of N-1=7 M_u, and M_3 is composed of N/(N-1)=8/7 M_u. In practical implementation, M_2 should have 49 M_u, and M_3 8 M_u. In order to keep the number of M_u in M_2 low for the sake of reduced area, N/(N-1) could be replaced by 1 [129]. Eq. (5.2) needs to be modified as:

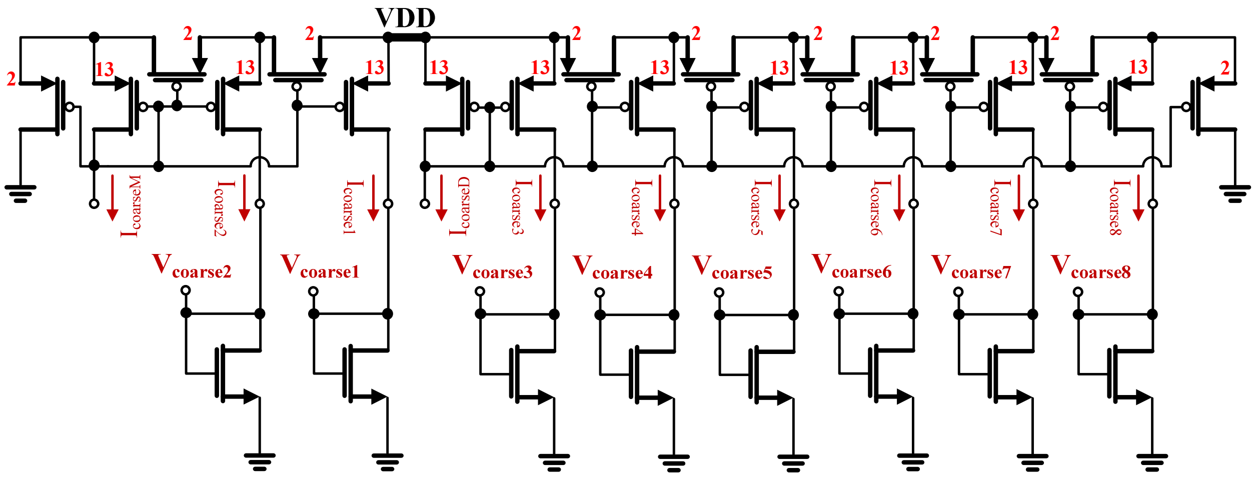


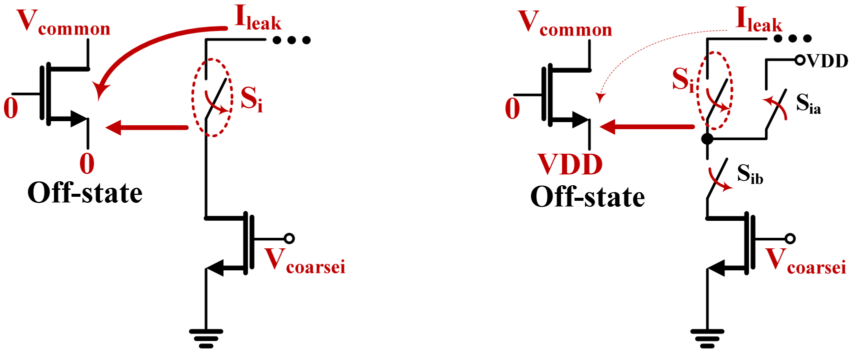
Figure 5.4. Complete circuit diagram of the current multiplier and divider for generation of eight coarse currents, and their diode-connected nFET loads.

$$I_{out} = (N + 1 - \Delta)I_{in} \quad (0 \le \Delta < 1) \tag{5.3}$$

Approximate octal weight among $I_{coarse1}$~$I_{coarse8}$ can be achieved by making N=7 or 7.5. N=7.5 is chosen for larger scaling and is implemented as a 13:2 M_u number ratio (e.g. the ratio of M_2 to M_3). The complete circuit of generating the eight coarse currents together with the diode-connected nFET loads is illustrated in Figure 5.4. Compared to conventional CM implementations [172], the adopted current multiplier and divider save considerable chip areas, especially with such wide current span. The sizes of the diode-connecteed nFET loads are largely different to accommodate the large span of the coarse currents. For example, the size of the nFET for $I_{coarse1}$ is eight (W/L)=2μm/5μm in parallel with its nominal output gate voltage $V_{coarse1}$=583 mV, and the size of the nFET for $I_{coarse8}$ is eight (W/L)=0.5μm/20μm in series with nominal $V_{coarse8}$=179 mV. Note that the nFET sizes for $I_{coarse5}$~$I_{coarse8}$ are identical to save areas in individual bias branches as will be explained in the next section. The generated voltages $V_{coarse1}$~$V_{coarse8}$ are broadcasted to all the bias branches, and one of the coarse currents is copied in each bias branch via the other half of the CM.

5.2.3 Coarse-Fine Current Interface

One of the eight coarse currents is selected in each bias branch by a one-hot 3-to-8 decoder which has the output of S_i (i=1~8) to control the selection switch. Because the smallest coarse current $I_{coarse8}$ is about 13 pA, close to the channel leakage of a normal nFET in off state in 0.18 μm CMOS, the selection circuits need careful design to minimize the impact of the leakage from the selection switches on the accuracy of current copying, especially on $I_{coarse8}$. Figure 5.5(a) shows the conventional current selection scheme. If $I_{coarsei}$ is not selected, S_i is disconnected; if $I_{coarsei}$ is selected, S_i is connected. The drain terminals of all the nFET switches are connected together. Ideally, S_i would contribute no current to the common line when it is disconnected; however, with its V_{gs}=0 and V_{ds}=V_{common} that is close to VDD, the channel leakage from the switch nFET is in the range of pA, and has large variation depending on process and temperature. In the case of selecting $I_{coarse8}$, it is completely overwhelmed by all the summed leakage currents, and in turn $I_{coarse8}$ cannot be accurately copied for the subsequent fine current selection. The leakage may be reduced to an acceptable level by using very long channel length, but the consequently large ON-resistance results



(a) (b)

Figure 5.5. (a) Conventional current selection scheme using a single nFET as the switch; (b) current selection with switch leakage current suppression using two nFETs (S_i and S_{ib}) and one pFET (S_{ia}).

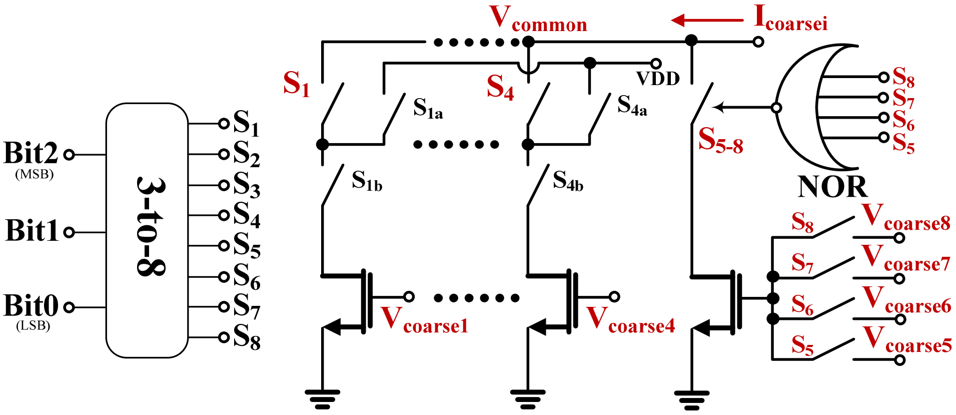


Figure 5.6. Complete circuit of the coarse-current selection.

in too large voltage drop for large coarse currents like $I_{coarse1}$. To circumvent the switch leakage problem, a leakage suppression scheme is proposed as illustrated in Figure 5.5(b). Two auxiliary switches S_{ia} and S_{ib} are added S_i. When S_i is disconnected, S_{ia} is connected and S_{ib} is disconnected so that the source voltage of S_i is elevated to VDD. This way the S_i nFET has V_{gs}=-VDD and $V_{ds}=V_{common}$-VDD that is close to 0. Therefore the leakage current from S_i is largely suppressed under various process and temperature corners to the range of fA.

The complete circuit of the coarse current selection is shown in Figure 5.6. A 3-to-8 decoder generates the 8 control signals for switches S_1-S_8. $I_{coarse1}$ to $I_{coarse4}$ are copied via separate nFETs, and the selection is done by using the leakage suppression scheme in Figure 5.5(b). $I_{coarse5}$ to $I_{coarse8}$ share the same nFET by respectively connecting its gate to $V_{coarse5}$ to $V_{coarse8}$. Compared to using separate nFETs for $I_{coarse5}$ to $I_{coarse8}$, nFET sharing saves about 38% area for each bias branch. The current selection switch $S_{5\text{-}8}$ that is controlled by a four-input NOR gate is a simple nFET as in Figure 5.5(a). No leakage suppression is needed for two reasons: the leakage from $S_{5\text{-}8}$ alone is negligible to coarse currents $I_{coarse1}$ to $I_{coarse4}$; the leakage from S_{ia} in its off state may affect the accuracy of small currents like $I_{couarse8}$.

The selected $I_{coarsei}$ is copied via an auto-configured pFET CM before being fed into the 8-bit current

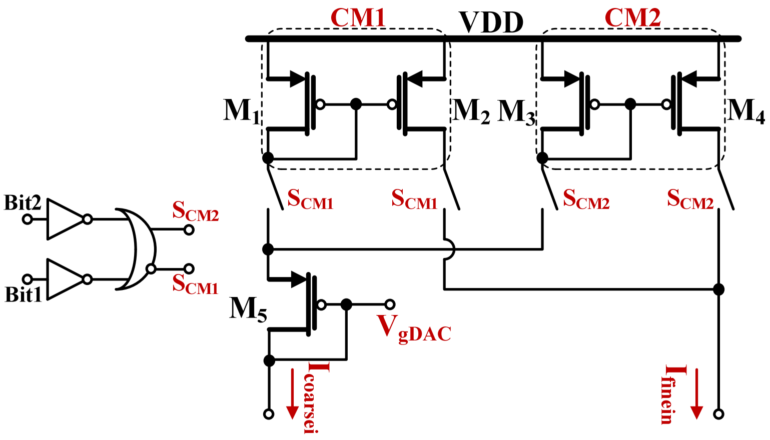


Figure 5.7. Auto-configured CM to convey the selected coarse current to an 8-bit current DAC for fine-current selection.

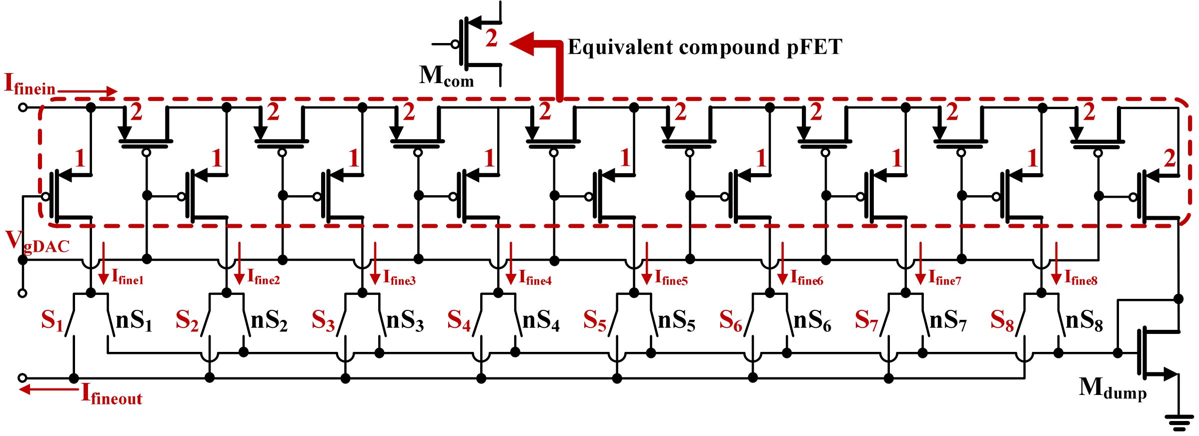


Figure 5.8. 8-bit R-2R current-splitting DAC for fine-current selection.

DAC. The auto-configured CM as shown in Figure 5.7 is composed of two different CMs. The CM1 has large *W*/*L* for coarse currents I_{coarse1} and I_{coarse2}, and the CM2 has small *W*/*L* for coarse currents from I_{coarse3} to I_{coarse8}. The CM selection logic which is a simple OR gate uses the Bit2 and Bit1 in Figure 5.6. Diode-connected pFET M_5 provides the gate voltage to the 8-bit current DAC as will be explained in the next section.

5.2.4 8-bit Current DAC

Illustrated in Figure 5.8, the compact 8-bit current DAC has the R-2R topology using pFETs. It is based on the same current-splitting principle used in Section 5.2.2 except here with binary scaling instead of octave. The number of unit transistors is marked at the source of each pFET. To the first-order approximation, the compound pFET M_{com} comprising all the DAC pFETs in Figure 5.8 is equivalent to a pFET comprising two unit pFETs in parallel with the gate voltage V_{gDAC} from M_5 in Figure 5.7. M_5 is also composed of two unit pFETs in parallel, and hence M_{com} and M_5 form the cascode stage of the CM in Figure 5.7, enabling a more accurate current copying. In previous designs [170], the gate voltage of M_{com} is connected to V_{MBN} in Figure 5.2, which not only loses the benefit of cascading but also compromises the saturation of the vertical pFETs. The eight fine currents I_{fine1}~I_{fine8} are either collected at the output node to

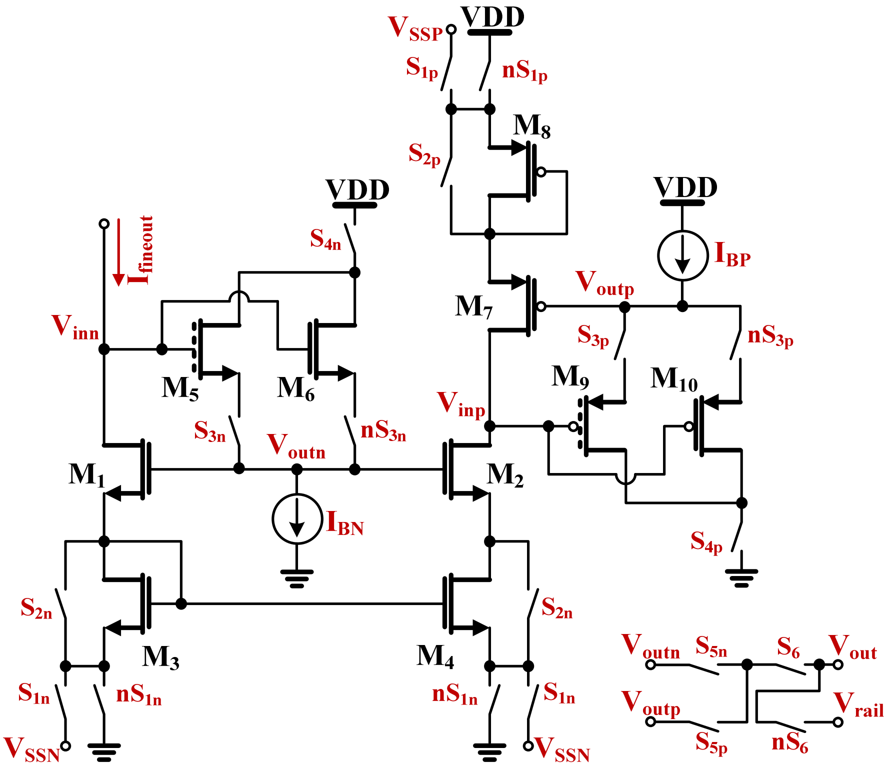


Figure 5.9. Fine-current buffer to generator the output bias voltage.

the next buffer stage as $I_{fineout}$, or at the drain of the diode-connected nFET M_{dump}. $I_{fineout}$ can be written as:

$$I_{fineout} = I_{finein}\sum_{i=1}^{8} 2^{-i} B_i \tag{5.4}$$

where B_i is the value of the i-th bit. The LSB of each coarse current is given in Table 5.1.

5.2.5 Configurable Output Buffer

The reconfigurable output buffer receives the output current of the 8-bit DAC $I_{fineout}$ described in the previous section and generates a bias voltage V_{out} to bias other circuit modules, such as retina pixels and cochlea channels, in the form of CM even though the pair transistors are distant from each other. The simplified buffer circuit is illustrated in Figure 5.9. The features of this buffer are described as follows.

(1). If $I_{course1}$ is selected in a bias branch, the cascode mode of the buffer will be automatically invalid, i.e. the switches S_{2n} and S_{2p} are always connected so that M_3 and M_8 are disabled. This is to prevent voltage headroom problems, i.e. V_{inn}/V_{inp} from being too high/low for the transistors in the previous circuit stage, namely the 8-bit DAC, to stay in saturation. If one of $I_{coarse2}$~$I_{coarse8}$ is selected, the cascode mode can be enabled by disconnecting S_{2n} and S_{2p}.

(2). The source-follower structures formed by M_5/M_6 and I_{BN}, and M_9/M_{10} and I_{BP} are used to stabilize V_{outn} and V_{outp} respectively in case of digital coupling disturbance. For $I_{corase1}$~$I_{coarse4}$, the transistor with low threshold M_5/M_9 is used to keep V_{inn}/V_{inp} not too high/low. For $I_{coarse5}$~$I_{coarse8}$, the transistor with normal threshold M_6/M_{10} is used to maintain the saturation of M_1/M_7. The control logic is derived from the

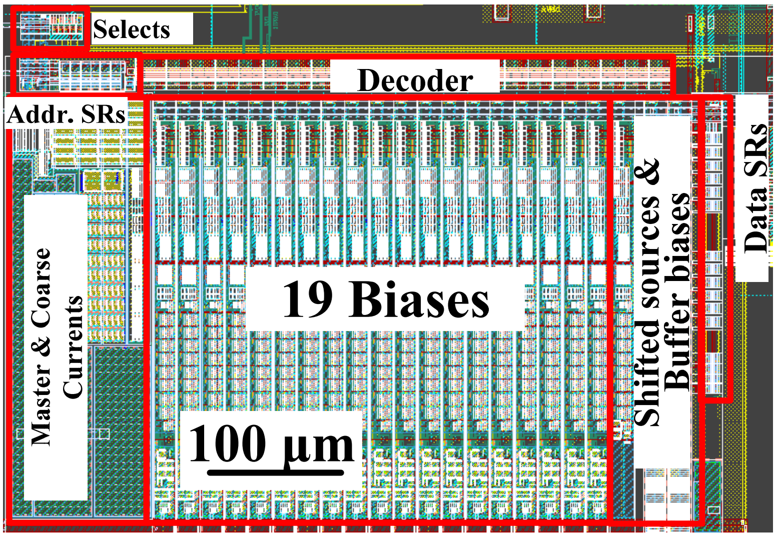


Figure 5.10. Layout of the current reference array.

Bit2 in Figure 5.6. The nominal values of I_{BN} and I_{BP} are both 50 nA, which are tunable by changing the V_{BBN} and V_{BBP} in Figure 5.1. However, they cannot be too high to avoid voltage headroom problems.

(3). Normally the sources of M_1 and M_7 are connected to ground and VDD respectively. In the non-cascode mode, to enable sub-off current generation, the source of M_1 needs to be lifted up above ground by connecting to V_{SSN} via S_{1n}, and M_7 be shifted down from VDD by connecting to V_{SSP} via S_{1p}. Together with the source-follower structure mentioned in (2), the shifted-source mode is enabled [129]. V_{SSN} and VDD-V_{SSP} are usually around 200~300 mV.

(4). If the bias branch is configured to send V_{outn} to V_{out} by connecting S_{5n} and S_6, S_{2n} can be connected or disconnected for the selection of the cascode mode. If instead V_{outp} is sent to V_{out} by connecting S_{5p} and S_6, S_{2n} is always disconnected to enable more accurate current copying via the cascode CM formed by M_1~M_4 except for the case of $I_{corase1}$.

(5). The output voltage V_{out} can be selected as one of V_{outn}, V_{outp}, and V_{rail}. V_{rail} is hard-wired to either ground or VDD in layout, and is supplied in reset to ensure minimum current consumption during start-up before the configuration data is loaded.

5.3 Experimental Results

The addressable current reference array has been fabricated and verified to be functional in both UMC 0.18 µm RF/MM CMOS and TowerJazz 0.18 µm CIS processes. The quantitative measurements are from the UMC run along with the first version of DAVIS array [18]. Figure 5.10 shows the layout of the array with 19 bias branches. Each branch occupies an area of 360×22 μm^2 that is only 25% of the previous design [170] thanks to the removal of the long SR chain in each bias branch.

The measured master bias I_R is about 390 nA under room temperature, which is consistent with the simulated $I_{couarse3}$ given in Table 5.1. Figure 5.11 gives the measured current tuning curves. The x-axis is the 8-bit fine current selection codes with the value from 1 to 256 in logarithm. The 3-bit coarse current selection codes are label along each curve. The lowest 3 curves are measured with shifted-source enabled. The specific and leakage current levels extracted from a 2×2 μm^2 transistor are also marked. The largest

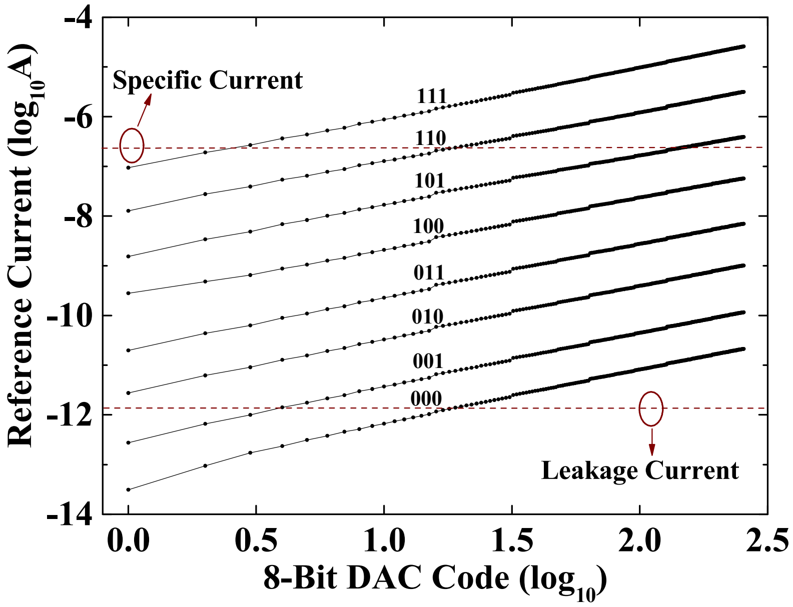


Figure 5.11. Measured currents as the function of the 8-bit fine code and 3-bit coarse code.

and smallest currents are about 25.83 μA and 31.13 fA, respectively. This gives a current dynamic range of 178 dB, close to the simulated 173 dB. The current range overlap between neighboring coarse codes allows reasonably fine resolution at all current scales, which is not the case in a non-hierarchical coarse-fine architecture. All 8 tuning curves are approximately linear with the fine code value. Note that the sub-leakage current capability is compromised at higher temperatures. According to simulation, the junction leakage currents contributed by the switches in the 8-bit R-2R DAC can reach to several pA at 80°C, which degrade the current dynamic range to about 130 dB. Within one die, the worst-case 1σ mismatch among all bias branches over the whole current range is within ±10%. The worst-case DNL is found to be about 5LSB at $I_{coarse1}$, which happens at the transition of the MSB of the 8-bit fine codes.

5.4 Conclusion and Remarks

This chapter detailed the design and measurement of an addressable current reference array with coarse-fine architecture. Compared to previous designs [170], the main benefits are: area-saving in each bias branch and faster current programming thanks to the avoidance of long SR chain; extended current dynamic range with less resolution bits owing to the hierarchical coarse-fine structure; adaptive power consumption in each bias branch in accordance to the selected coarse current, instead of a fixed up-scaling of the master bias.

To improve the linearity of the 8-bit current-splitting DAC, besides using larger unit transistors, transistors working in linear region with regulated identical voltages at the outputs of the selected current and the dumped current could be adopted at the cost of increased area. This method has been proven valid in an untrimmed 10-bit SAR ADC with -79 dB THD using MOS-only DAC [173].

Chapter 6: Summary and Future Work

This thesis starts with addressing the impact of the non-idealities of the massively-parallel asynchronous spike encoding in artificial bio-inspired sensors on encoding performances by mathematical modeling and numerical simulation. Specifically, two spike encoding mechanisms are studied, namely the self-timed reset (STR) that was implemented in prior dynamic vision sensors (DVSs), i.e. the spiking silicon retina, and the asynchronous delta modulation (ADM) that was used to implement continuous-time ADCs in the form of level-crossing sampling. The effects of spike transmission delay and switch holding (refractory period) of STR on the encoding quality in terms of signal-to-distortion ratio (SDR) of reconstructed input from spike trains are compared between STR and ADM using linear decoding. The STR is susceptible to signal loss due to the aforementioned non-idealities, whereas the ADM usually gives higher decoding SDR especially at high input frequencies and large quantization bit number because unlike STR its feedback does not disruptively interfere with the input. Nonlinear decoding based on frame theory results in much higher SDR than linear decoding, which facilitates the study of the impact of delay variation during spike transmission on encoding quality degradation. Delay variation comes from the input-slope dependent comparison delay and spike queueing owing to the AER arbitration. Using circuit analysis and queueing theory, both of the delay variation sources are quantitatively related to circuit and system design parameters like comparator biasing and AER bandwidth, and in turn the encoding quality measured by the SDR of reconstructed signals using nonlinear decoding. The results can be useful for future specifications-guided design of spiking sensors.

Two spiking sensors using ADM are developed. The new silicon retina focuses on improving previous designs in two aspects: the temporal contrast sensitivity and the encoding quality. The former is achieved by employing a low-noise photoreceptor with pFET common-gate feedback and a PGA for sufficient frontend gain, and the latter is achieved by using a proposed compact asynchronous switched-capacitor circuit for in-pixel ADM. The new 0.5-V 55-μW silicon cochlea features a proposed low-power programmable-gain source-follower-based BPF composed of a 4th-order LPF and a summing PGA, and an ADM using latched comparators with adaptive self-oscillation loop to provide the pseudo-clock for improved energy efficiency. Both chips have been fabricated in 0.18 μm CMOS and experimentally validated.

The improved temporal contrast sensitivity and encoding quality of silicon retina can facilitate its application in areas like in-vivo optical neuroimaging where the fluorescence dynamics is of particular interest. The intrinsic sparse output characteristic of event encoding can help significantly reduce the output data redundancy and thus save power in RF transmission which is beneficial to long-term monitoring of free-moving animals. As already discussed in Chapter 3, several technical hurdles including the ADM threshold mismatch still need to be overcome before any practical use. Alternative encoding mechanisms like asynchronous sigma-delta modulation (ASDM) may solve the threshold mismatch problem at the cost of much limited output bandwidth because ASDM produces idle spikes without any transient input change. However, ASDM could be used in the design of ultra-low-voltage spiking image or vision sensors by directly encoding photocurrent into spike timing instead of A/D conversion in voltage domain. One example is the Octopus imager [5] where the integrate-and-fire encoding can be seen as a special case of ASDM [174]; another example is the free-running-oscillator-pixel imager with synchronous frame sampling [175]. Neither of them exploited time-domain interpolation of the sensor output nor explored ultra-low-voltage design like in some 0.5-V imagers with pulse-width-modulation readout [176], [177],

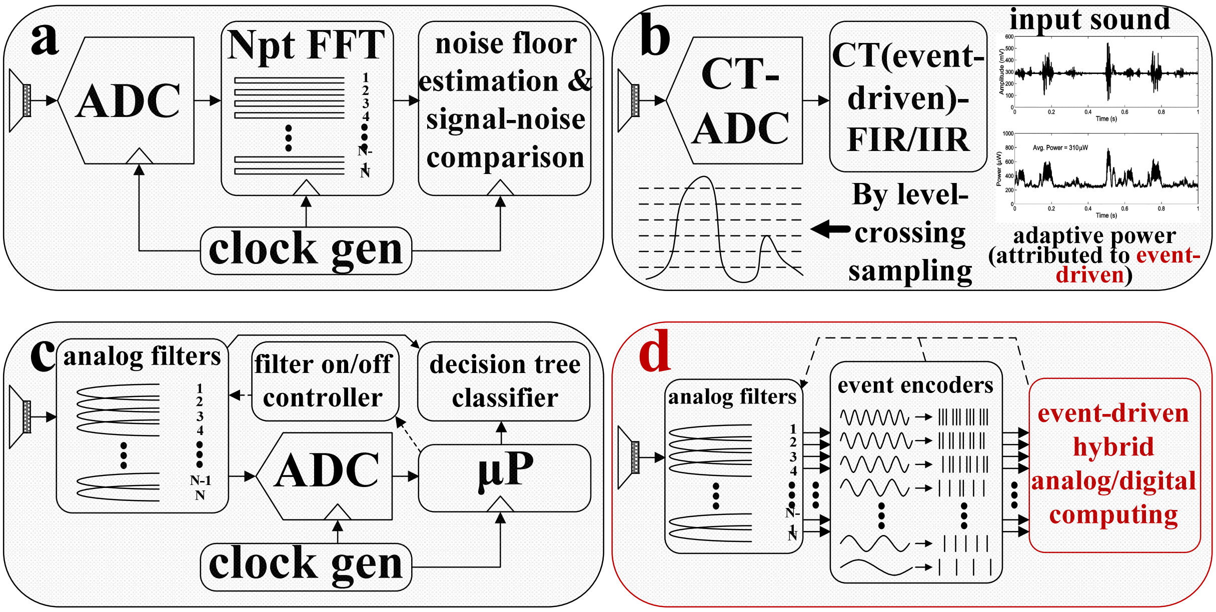


Figure 6.1. Four possible architectures for sound acquisition and processing: (a) The traditional architecture implementing an energy-based voice activity detection (VAD) algorithm with clocked ADC and DSP including FFT [41]; (b) A clockless system with CT-ADC/DSP. So far the CT-ADCs mainly implement level-crossing sampling and existing CT-DSPs only have FIR/IIR functions. The insert is the testing results from [61] showing adaptive system power consumption in response to a segment of sound signal; (c) Feature extraction of the energy in each frequency band is moved to the analog frontend. By exploiting the decision tree machine learning algorithm for VAD, unused analog filters are dynamically disabled to save power [42]; (d) Proposed architecture with cochlea-like signal acquisition and event-driven hybrid analog/digital processing.

and in both of the prior cases the problem of threshold mismatch among pixels still needs to be circumvented if no calibration or post-processing is desired. One interesting encoding scheme published recently [178] resembles ASDM except it does not produce any idle event output when the input DC is at 0. Preliminary modeling and simulation using a nonlinear decoding algorithm have shown its big advantage over ADM, i.e. the reconstructed signal does not suffer from DC drift with unbalanced ON and OFF thresholds, although the reconstruction error is still sensitive to threshold variations. One possible means of workaround is to adopt the so-called time-domain correlated double sampling [179] so that the voltage crossing represented by two spike duplets has minimized variation across pixels.

Ultra-low-voltage and ultra-low-power designs are deemed as one of the key enabling technology for ubiquitous embodiment of the concept internet of everything, which is the theme of ISSCC 2016. The cochlea design in this thesis echoes with this trend. Not only does the sensor core itself operate under a 0.5-V supply and dissipate merely tens of µW with 64×2 channels, but also the spike output is the natural input to event-driven signal processing systems with activity-dependent power consumption like the event-driven DSPs [61] and spiking neural networks [46]. Further improvement on the power efficiency of the sensor core can be achieved on both circuits and sub-system levels. For example, the summing method of creating one zero proposed in this thesis requires the bandwidth of the summing PGA to be as at least twice the BPF central frequency for non-degraded passband gain. One method of reducing the PGA bandwidth requirement is to create the zero by a source-follower-based BPF [141], [142], [166] so that the summation can be avoided. A more radical method is to replace the closed-loop PGA with a more power-efficient open-loop OTA using the spike encoding scheme proposed in [178]. Functionally it is beneficial to add automatic gain control to the silicon cochlea in order to fully harness the system dynamic range which is especially important under a low voltage supply.

In the future, the cochlea is expected to be directly integrated with spike-processing ASICs to form ultra-low-power smart sensor nodes. The first step is to implement some simple functions like voice activity detection, sound source localization and emergency word detection, which is my proposed postdoc project to work with Prof. Mingoo Seok and Prof. Yannis Tsividis at Columbia University. Although IBM′s pure digital neural network approach represented by the TrueNorth chip is becoming popular among academic communities, analog computing may still have a place in further improving computational energy efficiency [180]–[182]. The complete SoC is envisioned to contain analog preprocessing (bandpass filtering in the case of silicon cochlea), spike encoding, and digital or analog/digital hybrid spike processing as depicted in Figure 6.1(d). In terms of system optimization, one fundamental question to ask is that how the requirement of spike processing performance measured by metrics like classification or recognition accuracy enforces constraints on the analog frontend design. In principle, this type of SoC architecture can be extended to vision systems. Despite the advantage of reduced number of spike encoders and thus spike transmission bandwidth [183], an analog feature extraction pre-processing stage added before spike encoding is usually considered too area-consuming to be accomplished in focal plane of vision sensors. The area difficulty might be overcome by the maturing 3D integration technology [184]–[186]; a more challenging problem is how to minimize the performance variation in feature extraction and how much the post-processing can compensate for the front-end imperfections with a targeted system performance goal. It is also worth mentioning that the research in biomedical IC systems also seems to be following this trend [187].

Bibliography

[1] M. Mahowald, “VLSI analogs of neuronal visual processing: a synthesis of form and function,” Ph.D. Thesis, California Institute of Technology, Pasadena, CA, USA, 1992.

[2] S. K. Mendis, S. E. Kemeny, R. C. Gee, B. Pain, C. O. Staller, Q. Kim, and E. R. Fossum, “CMOS active pixel image sensors for highly integrated imaging systems,” *IEEE J. Solid-State Circuits*, vol. 32, no. 2, pp. 187–197, Feb. 1997.

[3] C. E. Shannon, “Communication in the presence of noise,” *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[4] K. Boahen, “Neuromorphic microchips,” *Sci. Am.*, vol. 292, no. 5, pp. 56–63, May 2005.

[5] E. Culurciello, R. Etienne-Cummings, and K. A. Boahen, “A biomorphic digital image sensor,” *IEEE J. Solid-State Circuits*, vol. 38, no. 2, pp. 281–294, Feb. 2003.

[6] P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor,” *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.

[7] V. Gruev and R. Etienne-Cummings, “A pipelined temporal difference imager,” *IEEE J. Solid-State Circuits*, vol. 39, no. 3, pp. 538–543, Mar. 2004.

[8] Y. M. Chi, U. Mallik, M. A. Clapp, E. Choi, G. Cauwenberghs, and R. Etienne-Cummings, “CMOS Camera With In-Pixel Temporal Change Detection and ADC,” *IEEE J. Solid-State Circuits*, vol. 42, no. 10, pp. 2187–2196, Oct. 2007.

[9] S. Martinez-Conde, S. L. Macknik, X. G. Troncoso, and T. A. Dyar, “Microsaccades counteract visual fading during fixation,” *Neuron*, vol. 49, no. 2, pp. 297–305, Jan. 2006.

[10] P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128×128 120dB 30mW asynchronous vision sensor that responds to relative intensity change,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2006, pp. 508–509.

[11] D. H. Kelly, “Visual Responses to Time-Dependent Stimuli I Amplitude Sensitivity Measurements,” *J. Opt. Soc. Am.*, vol. 51, no. 4, p. 422, Apr. 1961.

[12] Y. Tochigi, K. Hanzawa, Y. Kato, R. Kuroda, H. Mutoh, R. Hirose, H. Tominaga, K. Takubo, Y. Kondo, and S. Sugawa, “A Global-Shutter CMOS Image Sensor With Readout Speed of 1-Tpixel/s Burst and 780-Mpixel/s Continuous,” *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 329–338, Jan. 2013.

[13] R. Xu, W. C. Ng, J. Yuan, S. Yin, and S. Wei, “A 1/2.5 inch VGA 400 fps CMOS image sensor with high sensitivity for machine vision,” *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2342–2351, Oct. 2014.

[14] C. Posch, M. Hofstatter, D. Matolin, G. Vanstraelen, P. Schön, N. Donath, and M. Litzenberger, “A dual-line optical transient sensor with on-chip precision time-stamp generation,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2007, pp. 500–501.

[15] J. A. Leñero-Bardallo, T. Serrano-Gotarredona, and B. Linares-Barranco, “A 3.6 μs latency asynchronous frame-free event-driven dynamic-vision-sensor,” *IEEE J. Solid-State Circuits*, vol. 46, no. 6, pp. 1443–1455, Jun. 2011.

[16] T. Serrano-Gotarredona and B. Linares-Barranco, “A 128×128 1.5% contrast sensitivity 0.9% FPN 3 μs latency 4 mW asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers,” *IEEE J. Solid-State Circuits*, vol. 48, no. 3, pp. 827–838, Mar. 2013.

[17] C. Posch, D. Matolin, and R. Wohlgenannt, “A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS,” *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, Jan. 2011.

[18] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, “A 240×180 130 dB 3 μs latency global shutter spatiotemporal vision sensor,” *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, Oct. 2014.

[19] P.-F. Ruedi, P. Heim, F. Kaess, E. Grenet, F. Heitger, P.-Y. Burgi, S. Gyger, and P. Nussbaum,

"A 128×128 pixel 120-dB dynamic-range vision-sensor chip for image contrast and orientation extraction," *IEEE J. Solid-State Circuits*, vol. 38, no. 12, pp. 2325–2333, Dec. 2003.

[20] M. Gottardi, N. Massari, and S. A. Jawed, "A 100 µW 128×64 Pixels Contrast-Based Asynchronous Binary Vision Sensor for Sensor Networks Applications," *IEEE J. Solid-State Circuits*, vol. 44, no. 5, pp. 1582–1592, May 2009.

[21] K. A. Zaghloul and K. Boahen, "Optic nerve signals in a neuromorphic chip I: Outer and inner retina models," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 657–666, Apr. 2004.

[22] J. A. Leñero-Bardallo, T. Serrano-Gotarredona, and B. Linares-Barranco, "A five-decade dynamic-range ambient-light-independent calibrated signed-spatial-contrast AER retina with 0.1-ms latency and optional time-to-first-spike mode," *IEEE Trans. Circuits Syst. Regul. Pap.*, vol. 57, no. 10, pp. 2632–2643, Oct. 2010.

[23] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 7, pp. 1119–1134, Jul. 1988.

[24] L. Watts, D. A. Kerns, R. F. Lyon, and C. A. Mead, "Improved implementation of the silicon cochlea," *IEEE J. Solid-State Circuits*, vol. 27, no. 5, pp. 692–700, May 1992.

[25] A. van Schaik, E. Fragnière, and E. Vittoz, "Improved silicon cochlea using compatible lateral bipolar transistors," in *Advances in Neural Information Process. Syst. 8*, 1995, pp. 671–677.

[26] B. Razavi, *Design of Analog CMOS Integrated Circuits*, 1st edition. Boston, MA: McGraw-Hill Science/Engineering/Math, 2000.

[27] R. Sarpeshkar, C. Salthouse, J.-J. Sit, M. W. Baker, S. M. Zhak, T. K.-T. Lu, L. Turicchia, and S. Balster, "An ultra-low-power programmable analog bionic ear processor," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 4, pp. 711–727, Apr. 2005.

[28] C. D. Salthouse and R. Sarpeshkar, "A practical micropower programmable bandpass filter for use in bionic ears," *IEEE J. Solid-State Circuits*, vol. 38, no. 1, pp. 63–70, Jan. 2003.

[29] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon, "A biomimetic, 4.5 µW, 120+ dB, log-domain cochlea channel with AGC," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 1006–1022, Mar. 2009.

[30] L. Watts, "Cochlear mechanics : analysis and analog VLSI," Ph.D. Thesis, California Institute of Technology, Pasadena, CA, USA, 1993.

[31] E. Fragniere, "A 100-channel analog CMOS auditory filter bank for speech recognition," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2005, pp. 140–141.

[32] T. J. Hamilton, C. Jin, A. van Schaik, and J. Tapson, "An active 2-D silicon cochlea," *IEEE Trans. Biomed. Circuits Syst.*, vol. 2, no. 1, pp. 30–43, Mar. 2008.

[33] B. Wen and K. Boahen, "A silicon cochlea with active coupling," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 6, pp. 444–455, Dec. 2009.

[34] N. Kumar, W. Himmelbauer, G. Cauwenberghs, and A. G. Andreou, "An analog VLSI chip with asynchronous interface for auditory feature extraction," *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.*, vol. 45, no. 5, pp. 600–606, May 1998.

[35] H. Abdalla and T. K. Horiuchi, "An ultrasonic filterbank with spiking neurons," in *2005 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2005, pp. 4201–4204.

[36] V. Chan, S.-C. Liu, and A. van Schaik, "AER EAR: a matched silicon cochlea pair with address event representation interface," *IEEE Trans. Circuits Syst. Regul. Pap.*, vol. 54, no. 1, pp. 48–59, Jan. 2007.

[37] S.-C. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous binaural spatial audition sensor with 2×64×4 channel output," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 4, pp. 453–464, Aug. 2014.

[38] H. Noguchi, T. Takagi, M. Yoshimoto, and H. Kawaguchi, "An ultra-low-power VAD hardware implementation for intelligent ubiquitous sensor networks," in *2009 IEEE Workshop on Signal Process. Syst. (SiPS),* 2009, pp. 214–219.

[39] B. F. Logan, "Information in the zero crossings of bandpass signals," *Bell Syst. Tech. J.*, vol. 56, no. 4, pp. 487–510, Apr. 1977.

[40] A. A. Lazar, “Time encoding with an integrate-and-fire neuron with a refractory period,” *Neurocomputing*, vol. 58–60, pp. 53–58, Jun. 2004.

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

[42] K. Badami, S. Lauwereins, W. Meert, and M. Verhelst, “Context-aware hierarchical information-sensing in a 6μW 90nm CMOS voice activity detector,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2015, pp. 430–431.

[43] M. M. Shulaker, G. Hills, N. Patil, H. Wei, H.-Y. Chen, H.-S. P. Wong, and S. Mitra, “Carbon nanotube computer,” *Nature*, vol. 501, no. 7468, pp. 526–530, Sep. 2013.

[44] J. Kim and S. Tiwari, “Inexact computing using probabilistic circuits: ultra low-power digital processing,” *J Emerg Technol Comput Syst*, vol. 10, no. 2, pp. 16:1–16:23, Mar. 2014.

[45] D. S. Nikolopoulos, H. Vandierendonck, N. Bellas, C. D. Antonopoulos, S. Lalis, G. Karakonstantis, A. Burg, and U. Naumann, “Energy Efficiency through Significance-Based Computing,” *Computer*, vol. 47, no. 7, pp. 82–85, Jul. 2014.

[46] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, “A million spiking-neuron integrated circuit with a scalable communication network and interface,” *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[47] E. Painkras, L. A. Plana, J. Garside, S. Temple, F. Galluppi, C. Patterson, D. R. Lester, A. D. Brown, and S. B. Furber, “SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation,” *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, Aug. 2013.

[48] P. Knag, J. K. Kim, T. Chen, and Z. Zhang, “A Sparse Coding Neural Network ASIC With On-Chip Learning for Feature Extraction and Encoding,” *IEEE J. Solid-State Circuits*, vol. 50, no. 4, pp. 1070–1079, Apr. 2015.

[49] A. A. Lazar, E. A. Pnevmatikakis, and Y. Zhou, “The power of connectivity: Identity preserving transformations on visual streams in the spike domain,” *Neural Netw.*, vol. 44, pp. 22–35, Aug. 2013.

[50] D. A. Butts, C. Weng, J. Jin, C.-I. Yeh, N. A. Lesica, J.-M. Alonso, and G. B. Stanley, “Temporal precision in the neural code and the timescales of natural vision,” *Nature*, vol. 449, no. 7158, pp. 92–95, Sep. 2007.

[51] E. A. Pnevmatikakis, “Spikes as projections: Representation and processing of sensory stimuli in the time domain,” Ph.D. Thesis, Columbia University, New York, NY, USA, 2010.

[52] A. A. Lazar and L. T. Toth, “Perfect recovery and sensitivity analysis of time encoded bandlimited signals,” *IEEE Trans. Circuits Syst. Regul. Pap.*, vol. 51, no. 10, pp. 2060–2073, Oct. 2004.

[53] A. A. Lazar and E. A. Pnevmatikakis, “Video time encoding machines,” *IEEE Trans. Neural Netw.*, vol. 22, no. 3, pp. 461–473, Mar. 2011.

[54] Y. Tsividis, “Event-Driven Data Acquisition and Digital Signal Processing - A Tutorial,” *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 57, no. 8, pp. 577–581, Aug. 2010.

[55] N. Sayiner, H. V. Sorensen, and T. R. Viswanathan, “A level-crossing sampling scheme for A/D conversion,” *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.*, vol. 43, no. 4, pp. 335–339, Apr. 1996.

[56] H. Inose, T. Aoki, and K. Watanabe, “Asynchronous delta-modulation system,” *Electron. Lett.*, vol. 2, no. 3, pp. 95–96, Mar. 1966.

[57] R. Steele, *Delta Modulation Systems*. Pentech Press, 1975.

[58] B. Schell and Y. Tsividis, “Analysis of continuous-time digital signal processors,” in *2007 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2007, pp. 2232–2235.

[59] A. J. Martin and M. Nystrom, “Asynchronous Techniques for System-on-Chip Design,” *Proc. IEEE*, vol. 94, no. 6, pp. 1089–1120, Jun. 2006.

[60] P. Martinez-Nuevo, S. Patil, and Y. Tsividis, "Derivative level-crossing sampling," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 62, no. 1, pp. 11–15, Jan. 2015.

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

[62] B. Schell and Y. Tsividis, "A low power tunable delay element suitable for asynchronous delays of burst information," *IEEE J. Solid-State Circuits*, vol. 43, no. 5, pp. 1227–1234, May 2008.

[63] M. Kurchuk, C. Weltin-Wu, D. Morche, and Y. Tsividis, "Event-driven GHz-range continuous-time digital signal processor with activity-dependent power dissipation," *IEEE J. Solid-State Circuits*, vol. 47, no. 9, pp. 2164–2173, Sep. 2012.

[64] M. Trakimas and S. R. Sonkusale, "An adaptive resolution asynchronous ADC architecture for data compression in energy constrained sensing applications," *IEEE Trans. Circuits Syst. Regul. Pap.*, vol. 58, no. 5, pp. 921–934, May 2011.

[65] C. Weltin-Wu and Y. Tsividis, "An event-driven clockless level-crossing ADC with signal-dependent adaptive resolution," *IEEE J. Solid-State Circuits*, vol. 48, no. 9, pp. 2180–2190, Sep. 2013.

[66] C. Vezyrtzis, W. Jiang, S. M. Nowick, and Y. Tsividis, "A flexible, event-driven digital filter with frequency response independent of input sample rate," *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2292–2304, Oct. 2014.

[67] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, and T. Delbruck, "Retinomorphic event-based vision sensors: bioinspired cameras with spiking output," *Proc. IEEE*, vol. 102, no. 10, pp. 1470–1484, Oct. 2014.

[68] E. Roza, "Analog-to-digital conversion via duty-cycle modulation," *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.*, vol. 44, no. 11, pp. 907–914, Nov. 1997.

[69] C. Vezyrtzis and Y. Tsividis, "Processing of signals using level-crossing sampling," in *2009 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2009, pp. 2293–2296.

[70] B. Schell, "Continuous-time digital signal processors: analysis and implementation," Ph.D. Thesis, Columbia University, New York, NY, USA, 2008.

[71] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2 edition. Hoboken, N.J: Wiley-Interscience, 2006.

[72] P. Lichtsteiner, "An AER temporal contrast vision sensor," Ph.D. Thesis, ETH Zurich, Zurich, Switzerland, 2006.

[73] A. Papoulis, *Signal Analysis*. New York: Mcgraw-Hill College, 1977.

[74] P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design*, 2nd edition. New York: Oxford University Press, 2002.

[75] S. O. Rice, "Mathematical analysis of random noise," *Bell Syst. Tech. J.*, vol. 24, no. 1, pp. 46–156, Jan. 1945.

[76] S. O. Rice, "Mathematical analysis of random noise," *Bell Syst. Tech. J.*, vol. 23, no. 3, pp. 282–332, Jul. 1944.

[77] A. Mortara and E. A. Vittoz, "A communication architecture tailored for analog VLSI artificial neural networks: intrinsic performance and limitations," *IEEE Trans. Neural Netw.*, vol. 5, no. 3, pp. 459–466, May 1994.

[78] J. N. Y. Aziz, K. Abdelhalim, R. Shulyzki, R. Genov, B. L. Bardakjian, M. Derchansky, D. Serletis, and P. L. Carlen, "256-channel neural recording and delta compression microsystem with 3D electrodes," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 995–1005, Mar. 2009.

[79] M. Schienle, C. Paulus, A. Frey, F. Hofmann, B. Holzapfl, P. Schindler-Bauer, and R. Thewes, "A fully electronic DNA sensor with 128 positions and in-pixel A/D conversion," *IEEE J. Solid-State Circuits*, vol. 39, no. 12, pp. 2438–2445, Dec. 2004.

[80] K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.*, vol. 47, no. 5, pp. 416–434, May 2000.

[81] R. Cahn, *Wide Area Network Design: Concepts and Tools for Optimization*, 1st edition. San Francisco, Calif.: Morgan Kaufmann, 1998.

[82] G. J. Franx, "A simple solution for the M/D/c waiting time distribution," *Oper. Res. Lett.*, vol. 29, no. 5, pp. 221–229, Dec. 2001.

[83] K. A. Boahen, "A burst-mode word-serial address-event link-I: transmitter design," *IEEE Trans. Circuits Syst. Regul. Pap.*, vol. 51, no. 7, pp. 1269–1280, Jul. 2004.

[84] R. Berner, "Building blocks for event-based sensors," Ph.D. Thesis, ETH Zurich, Zurich, Switzerland, 2011.

[85] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th edition. Boston: McGraw-Hill, 2002.

[86] K. A. Boahen, "A burst-mode word-serial address-event link-III: analysis and test results," *IEEE Trans. Circuits Syst. Regul. Pap.*, vol. 51, no. 7, pp. 1292–1300, Jul. 2004.

[87] M. L. Chaudhry and U. C. Gupta, "Exact computational analysis of waiting-time distributions of single-server bulk-arrival queues: $M^X/G/1$," *Eur. J. Oper. Res.*, vol. 63, no. 3, pp. 445–462, Dec. 1992.

[88] G. J. Franx, "The $M^X/D/c$ Batch Arrival Queue," *Probab. Eng. Informational Sci.*, vol. 19, no. 03, pp. 345–349, Jul. 2005.

[89] M. Yang, S.-C. Liu, and T. Delbruck, "A dynamic vision sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding," *IEEE J. Solid-State Circuits*, vol. 50, no. 9, pp. 2149–2160, Sep. 2015.

[90] H. C. Tijms, *Stochastic Models: An Algorithmic Approach*. West Sussex, UK: John Wiley & Sons, 1994.

[91] M. L. Chaudhry and J. G. C. Templeton, *First Course in Bulk Queues*. New York: John Wiley & Sons Inc, 1983.

[92] A. Grinvald and R. Hildesheim, "VSDI: a new era in functional imaging of cortical dynamics," *Nat. Rev. Neurosci.*, vol. 5, no. 11, pp. 874–885, Nov. 2004.

[93] J. N. D. Kerr and W. Denk, "Imaging in vivo: watching the brain in action," *Nat. Rev. Neurosci.*, vol. 9, no. 3, pp. 195–205, Mar. 2008.

[94] C. Brandli, L. Muller, and T. Delbruck, "Real-time, high-speed video decompression using a frame- and event-based DAVIS sensor," in *2014 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2014, pp. 686–689.

[95] G. Orchard, D. Matolin, X. Lagorce, R. Benosman, and C. Posch, "Accelerated frame-free time-encoded multi-step imaging," in *2014 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2014, pp. 2644–2647.

[96] S.-C. Liu, J. Kramer, G. Indiveri, T. Delbruck, and R. Douglas, *Analog VLSI: Circuits and Principles*, 1st edition. Cambridge, Mass: A Bradford Book, 2002.

[97] S.-W. Kang, K.-S. Min, and K. Lee, "Parametric expression of subthreshold slope using threshold voltage parameters for MOSFET statistical modeling," *IEEE Trans. Electron Devices*, vol. 43, no. 9, pp. 1382–1386, Sep. 1996.

[98] Y. Tsividis and C. McAndrew, *Operation and Modeling of the MOS Transistor*, 3rd edition. New York: Oxford University Press, 2010.

[99] Y.-T. Yeow, "Measurement and numerical modeling of short-channel MOSFET gate capacitances," *IEEE Trans. Electron Devices*, vol. 34, no. 12, pp. 2510–2520, Dec. 1987.

[100] T. Delbruck and C. A. Mead, "Adaptive photoreceptor with wide dynamic range," in *1994 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 1994, pp. 339–342.

[101] R. Sarpeshkar, *Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications, and Bio-Inspired Systems*, 1st edition. Cambridge, UK: Cambridge University Press, 2010.

[102] A. Rose, *Vision: Human and Electronic*. Plenum Press, New York, 1973.

[103] W. N. Cheung and P. J. Edwards, "Simulation of noise characteristics in optical devices using PSpice," *Int. J. Electron.*, vol. 76, no. 4, pp. 627–632, Apr. 1994.

[104] X. Zou, X. Xu, L. Yao, and Y. Lian, "A 1-V 450-nW fully integrated programmable biomedical

sensor interface chip," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, pp. 1067–1077, Apr. 2009.

[105] R. F. Yazicioglu, P. Merken, R. Puers, and C. Van Hoof, "A 200 µW eight-channel EEG acquisition ASIC for ambulatory EEG systems," *IEEE J. Solid-State Circuits*, vol. 43, no. 12, pp. 3025–3038, Dec. 2008.

[106] W. M. C. Sansen, *Analog Design Essentials*, 1st edition. Springer, 2007.

[107] D. B. Ribner and M. A. Copeland, "Design techniques for cascoded CMOS Op Amps with improved PSRR and common-mode input range," *IEEE J. Solid-State Circuits*, vol. 19, no. 6, pp. 919–925, Dec. 1984.

[108] M. Taherzadeh-Sani and A. A. Hamoui, "A 1-V process-insensitive current-scalable two-stage Opamp with enhanced DC gain and settling behavior in 65-nm digital CMOS," *IEEE J. Solid-State Circuits*, vol. 46, no. 3, pp. 660–668, Mar. 2011.

[109] V. Saxena and R. Jacob Baker, "Compensation of CMOS op-amps using split-length transistors," in *2008 Midwest Symp. on Circuits and Syst. (MWSCAS)*, 2008, pp. 109–112.

[110] M. Yang, S.-C. Liu, and T. Delbruck, "Subthreshold DC-gain enhancement by exploiting small size effects of MOSFETs," *Electron. Lett.*, vol. 50, no. 11, pp. 835–837, May 2014.

[111] B. Yu, C. H. J. Wann, E. D. Nowak, K. Noda, and C. Hu, "Short-channel effect improved by lateral channel-engineering in deep-submicronmeter MOSFET's," *IEEE Trans. Electron Devices*, vol. 44, no. 4, pp. 627–634, Apr. 1997.

[112] K. K.-L. Hsueh, J. J. Sanchez, T. A. DeMassa, and L. A. Akers, "Inverse-narrow-width effects and small-geometry MOSFET threshold voltage model," *IEEE Trans. Electron Devices*, vol. 35, no. 3, pp. 325–338, Mar. 1988.

[113] Q. Xie, J. Xu, and Y. Taur, "Review and critique of analytic models of MOSFET short-channel effects in subthreshold," *IEEE Trans. Electron Devices*, vol. 59, no. 6, pp. 1569–1579, Jun. 2012.

[114] K. M. Cao, W. Liu, X. Jin, K. Vashanth, K. Green, J. Krick, T. Vrotsos, and C. Hu, "Modeling of pocket implanted MOSFETs for anomalous analog behavior," in *IEEE Int. Electron Devices Meeting (IEDM)*, 1999, pp. 171–174.

[115] A. S. Roy, S. P. Mudanai, and M. Stettler, "Mechanism of long-channel drain-induced barrier lowering in halo MOSFETs," *IEEE Trans. Electron Devices*, vol. 58, no. 4, pp. 979–984, Apr. 2011.

[116] T. H. Morshed, W. Yang, M. V. Dunga, X. Xi, J. He, W. Liu, M. Cao, X. Jin, J. J. Ou, M. Chan, A. M. Niknejad, and C. Hu, "BSIM4.6.4 MOSFET model user's manual," EECS Dept., Univ. California, Berkeley, CA, 2009.

[117] R. R. Harrison, "The design of integrated circuits to observe brain activity," *Proc. IEEE*, vol. 96, no. 7, pp. 1203–1216, Jul. 2008.

[118] Y. Li, D. Zhao, and W. A. Serdijn, "A sub-microwatt asynchronous level-crossing ADC for biomedical applications," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 2, pp. 149–157, Apr. 2013.

[119] R. R. Harrison and C. Charles, "A low-power low-noise CMOS amplifier for neural recording applications," *IEEE J. Solid-State Circuits*, vol. 38, no. 6, pp. 958–965, Jun. 2003.

[120] M. Yang, S.-C. Liu, C. Li, and T. Delbruck, "Addressable current reference array with 170dB dynamic range," in *2012 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2012, pp. 3110–3113.

[121] "jAER Open Source Project," *jAER Open Source Project*. [Online]. Available: http://jaerproject.org. [Accessed: 21-Apr-2015].

[122] T. Delbruck and R. Berner, "Temporal contrast AER pixel with 0.3%-contrast event threshold," in *2010 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2010, pp. 2442–2445.

[123] M. Yang, S.-C. Liu, and T. Delbruck, "Comparison of spike encoding schemes in asynchronous vision sensors: Modeling and design," in *2014 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2014, pp. 2632–2635.

[124] S. Jeong, I. Lee, D. Blaauw, and D. Sylvester, "A 5.8 nW CMOS wake-up timer for ultra-low-power wireless applications," *IEEE J. Solid-State Circuits*, vol. 50, no. 8, pp. 1754–1763, Aug. 2015.

[125] S. O. Rice, "Statistical properties of a sine wave plus random noise," *Bell Syst. Tech. J.*, vol. 27,

no. 1, pp. 109–157, 1948.

[126] J. M. Rabaey, "The Human Intranet–Where Swarms and Humans Meet," *IEEE Pervasive Comput.*, vol. 14, no. 1, pp. 78–83, Jan. 2015.

[127] M. Yang, C.-H. Chien, T. Delbruck, and S.-C. Liu, "A 0.5V 55µW 64×2-channel binaural silicon cochlea for event-driven stereo-audio sensing," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2016, p. accepted.

[128] K. Bult and G. J. G. M. Geelen, "An inherently linear and compact MOST-only current division technique," *IEEE J. Solid-State Circuits*, vol. 27, no. 12, pp. 1730–1735, Dec. 1992.

[129] B. Linares-Barranco and T. Serrano-Gotarredona, "On the design and characterization of femtoampere current-mode circuits," *IEEE J. Solid-State Circuits*, vol. 38, no. 8, pp. 1353–1363, Aug. 2003.

[130] W. Cheng, M. S. Oude Alink, A. J. Annema, G. J. M. Wienk, and B. Nauta, "A wideband IM3 cancellation technique for CMOS π- and T-attenuators," *IEEE J. Solid-State Circuits*, vol. 48, no. 2, pp. 358–368, Feb. 2013.

[131] L. Robles, M. A. Ruggero, and N. C. Rich, "Two-tone distortion in the basilar membrane of the cochlea," *Nature*, vol. 349, no. 6308, pp. 413–414, Jan. 1991.

[132] B. Drost, M. Talegaonkar, and P. K. Hanumolu, "Analog filter design using ring oscillator integrators," *IEEE J. Solid-State Circuits*, vol. 47, no. 12, pp. 3120–3129, Dec. 2012.

[133] B. Vigraham, J. Kuppambatti, and P. R. Kinget, "Switched-mode operational amplifiers and their application to continuous-time filters in nanoscale CMOS," *IEEE J. Solid-State Circuits*, vol. 49, no. 12, pp. 2758–2772, Dec. 2014.

[134] S. D'Amico, M. Conta, and A. Baschirotto, "A 4.1-mW 10-MHz fourth-order source-follower-based continuous-time filter with 79-dB DR," *IEEE J. Solid-State Circuits*, vol. 41, no. 12, pp. 2713–2719, Dec. 2006.

[135] M. Conta, A. Baschirotto, and S. D'Amico, "Filter circuit," US7659774 B2, 09-Feb-2010.

[136] S. D'Amico, M. De Matteis, and A. Baschirotto, "A 6th-order 100µA 280MHz source-follower-based single-loop continuous-time filter," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2008, pp. 72–73.

[137] M. Conta, A. Baschirotto, and S. D'Amico, "High order continuous time filter," US8710921 B2, 29-Apr-2014.

[138] V. Saari, M. Kaltiokallio, S. Lindfors, J. Ryynanen, and K. Halonen, "A 1.2V 240MHz CMOS continuous-time low-pass filter for a UWB radio receiver," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2007, pp. 122–591.

[139] F. Houfaf, M. Egot, A. Kaiser, A. Cathelin, and B. Nauta, "A 65nm CMOS 1-to-10GHz tunable continuous-time low-pass filter for high-data-rate communications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2012, pp. 362–364.

[140] J. Lechevallier, R. Struiksma, H. Sherry, A. Cathelin, E. Klumperink, and B. Nauta, "A forward-body-bias tuned 450MHz Gm-C 3rd-order low-pass filter in 28nm UTBB FD-SOI with >1dBVp IIP3 over a 0.7-to-1V supply," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2015, pp. 1–3.

[141] M. Yang, J. Liu, Y. Xiao, and H. Liao, "14.4 nW fourth-order bandpass filter for biomedical applications," *Electron. Lett.*, vol. 46, no. 14, pp. 973–974, Jul. 2010.

[142] C. Sawigun, W. Ngamkham, and W. A. Serdijn, "A 0.5-V, 2-nW, 55-dB DR, fourth-order bandpass filter using single branch biquads: An efficient design for FoM enhancement," *Microelectron. J.*, vol. 45, no. 4, pp. 367–374, Apr. 2014.

[143] Y. Chen, P.-I. Mak, L. Zhang, and Y. Wang, "0.07 mm^2, 2 mW, 75 MHz-IF, fourth-order BPF using source-follower-based resonator in 90 nm CMOS," *Electron. Lett.*, vol. 48, no. 10, pp. 552–554, May 2012.

[144] M. De Matteis, A. Pezzotta, S. D'Amico, and A. Baschirotto, "A 33 MHz 70 dB-SNR super-source-follower-based low-pass analog filter," *IEEE J. Solid-State Circuits*, vol. 50, no. 7, pp. 1516–1524, Jul. 2015.

[145] A. J. Casson and E. Rodriguez-Villegas, "A 60 pW g_m-C continuous wavelet transform circuit for

portable EEG systems," *IEEE J. Solid-State Circuits*, vol. 46, no. 6, pp. 1406–1415, Jun. 2011.

[146] S. Wang, T. J. Koickal, A. Hamilton, R. Cheung, and L. S. Smith, "A bio-realistic analog CMOS cochlea filter with high tunability and ultra-steep roll-off," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 3, pp. 297–311, Jun. 2015.

[147] S. Rai, J. Holleman, J. N. Pandey, F. Zhang, and B. Otis, "A 500μW neural tag with 2μVrms AFE and frequency-multiplying MICS/ISM FSK transmitter," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2009, pp. 212–213.

[148] X. Zou, W.-S. Liew, L. Yao, and Y. Lian, "A 1V 22μW 32-channel implantable EEG recording IC," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2010, pp. 126–127.

[149] D. Han, Y. Zheng, R. Rajkumar, G. Dawe, and M. Je, "A 0.45V 100-channel neural-recording IC with sub-μW/channel consumption in 0.18μm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2013, pp. 290–291.

[150] K. A. Ng and Y. P. Xu, "A multi-channel neural-recording amplifier system with 90dB CMRR employing CMOS-inverter-based OTAs with CMFB through supply rails in 65nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2015, pp. 206–207.

[151] S. Chatterjee, Y. Tsividis, and P. Kinget, "0.5-V analog circuit techniques and their application in OTA and filter design," *IEEE J. Solid-State Circuits*, vol. 40, no. 12, pp. 2373–2387, Dec. 2005.

[152] A. Yoshizawa and Y. Tsividis, "A channel-select filter with agile blocker detection and adaptive power dissipation," *IEEE J. Solid-State Circuits*, vol. 42, no. 5, pp. 1090–1099, May 2007.

[153] S. D'Amico, V. Giannini, and A. Baschirotto, "A 4th-order active-G_m-RC reconfigurable (UMTS /WLAN) filter," *IEEE J. Solid-State Circuits*, vol. 41, no. 7, pp. 1630–1637, Jul. 2006.

[154] A. Vasilopoulos, G. Vitzilaios, G. Theodoratos, and Y. Papananos, "A low-power wideband reconfigurable integrated active-RC filter with 73 dB SFDR," *IEEE J. Solid-State Circuits*, vol. 41, no. 9, pp. 1997–2008, Sep. 2006.

[155] M. De Matteis, S. D'Amico, and A. Baschirotto, "A 0.55 V 60 dB-DR fourth-order analog baseband filter," *IEEE J. Solid-State Circuits*, vol. 44, no. 9, pp. 2525–2534, Sep. 2009.

[156] F. Lin, X. Yu, S. Ranganathan, and T. Kwan, "A 70 dB MTPR integrated programmable gain/bandwidth fourth-order Chebyshev high-pass filter for ADSL/VDSL receivers in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, pp. 1290–1297, Apr. 2009.

[157] M. Banu, J. M. Khoury, and Y. Tsividis, "Fully differential operational amplifiers with accurate output balancing," *IEEE J. Solid-State Circuits*, vol. 23, no. 6, pp. 1410–1414, Dec. 1988.

[158] M. Abdulaziz, M. Tormanen, and H. Sjoland, "A compensation technique for two-stage differential OTAs," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 61, no. 8, pp. 594–598, Aug. 2014.

[159] S.-W. M. Chen and R. W. Brodersen, "A 6-bit 600-MS/s 5.3-mW asynchronous ADC in 0.13-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 41, no. 12, pp. 2669–2680, Dec. 2006.

[160] P. Harpe, E. Cantatore, and A. van Roermund, "A 10b/12b 40 kS/s SAR ADC with data-driven noise reduction achieving up to 10.1b ENOB at 2.2 fJ/conversion-step," *IEEE J. Solid-State Circuits*, vol. 48, no. 12, pp. 3011–3018, Dec. 2013.

[161] M. van Elzakker, E. van Tuijl, P. Geraedts, D. Schinkel, E. A. M. Klumperink, and B. Nauta, "A 10-bit charge-redistribution ADC consuming 1.9 μW at 1 MS/s," *IEEE J. Solid-State Circuits*, vol. 45, no. 5, pp. 1007–1015, May 2010.

[162] B. Razavi, "The StrongARM latch [A circuit for all seasons]," *IEEE Solid-State Circuits Mag.*, vol. 7, no. 2, pp. 12–17, Spring 2015.

[163] M. Verhelst and A. Bahai, "Where analog meets digital: Analog-to-information conversion and beyond," *IEEE Solid-State Circuits Mag.*, vol. 7, no. 3, pp. 67–80, 2015.

[164] R. Sarpeshkar, M. W. Baker, C. D. Salthouse, J.-J. Sit, L. Turicchia, and S. M. Zhak, "An analog bionic ear processor with zero-crossing detection," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2005, pp. 78–79.

[165] A. Zai, S. Bhargava, N. Mesgarani, and S.-C. Liu, "Reconstruction of audio waveforms from spike trains of artificial cochlea models," *Front. Neurosci.*, vol. 9, p. 347, 2015.

[166] Z. Gao, J. Ma, M. Yu, and Y. Ye, "A fully integrated CMOS active bandpass filter for multiband

RF front-ends," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 55, no. 8, pp. 718–722, Aug. 2008.

[167] O. Jeon, R. M. Fox, and B. A. Myers, "Analog AGC circuitry for a CMOS WLAN receiver," *IEEE J. Solid-State Circuits*, vol. 41, no. 10, pp. 2291–2300, Oct. 2006.

[168] D. Schinkel, E. Mensink, E. Klumperink, E. van Tuijl, and B. Nauta, "A double-tail latch-typevoltage sense amplifier with 18ps setup+hold time," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, 2007, pp. 314–605.

[169] G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 211–221, Jan. 2006.

[170] T. Delbruck, R. Berner, P. Lichtsteiner, and C. Dualibe, "32-bit configurable bias current generator with sub-off-current capability," in *2010 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2010, pp. 1647–1650.

[171] E. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operations," *IEEE J. Solid-State Circuits*, vol. 12, no. 3, pp. 224–231, Jun. 1977.

[172] S. Guo, H. Lee, and P. Loizou, "A 9-Bit configurable current source with enhanced output resistance for cochlear stimulators," in *2008 IEEE Custom Integrated Circuits Conf. (CICC)*, 2008, pp. 511–514.

[173] C. M. Hammerschmied and Q. Huang, "Design and implementation of an untrimmed MOSFET-only 10-bit A/D converter with -79-dB THD," *IEEE J. Solid-State Circuits*, vol. 33, no. 8, pp. 1148–1157, Aug. 1998.

[174] A. S. Alvarado, M. Rastogi, J. G. Harris, and J. C. Principe, "The integrate-and-fire sampler: A special type of asynchronous Σ-Δ modulator," in *2011 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2011, pp. 2031–2034.

[175] L. G. McIlrath, "A low-power low-noise ultrawide-dynamic-range CMOS imager with pixel-parallel A/D conversion," *IEEE J. Solid-State Circuits*, vol. 36, no. 5, pp. 846–853, May 2001.

[176] S. Hanson, Z. Foo, D. Blaauw, and D. Sylvester, "A 0.5 V sub-microwatt CMOS image sensor with pulse-width modulation read-out," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 759–767, Apr. 2010.

[177] M.-T. Chung, C.-L. Lee, C. Yin, and C.-C. Hsieh, "A 0.5 V PWM CMOS imager with 82 dB dynamic range and 0.055% fixed-pattern-noise," *IEEE J. Solid-State Circuits*, vol. 48, no. 10, pp. 2522–2530, Oct. 2013.

[178] S. Patil, A. Ratiu, D. Morche, and Y. Tsividis, "A 3-10fJ/conv-step 0.0032mm^2 error-shaping alias-free asynchronous ADC," in *2015 Symposium on VLSI Circuits (VLSI Circuits)*, 2015, pp. 160–161.

[179] D. Matolin, C. Posch, and R. Wohlgenannt, "True correlated double sampling and comparator design for time-based image sensors," in *2009 IEEE Int. Symp. Circuits and Syst. (ISCAS)*, 2009, pp. 1269–1272.

[180] G. E. R. Cowan, R. C. Melville, and Y. Tsividis, "A VLSI analog computer/digital computer accelerator," *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 42–53, Jan. 2006.

[181] J. Lu, S. Young, I. Arel, and J. Holleman, "A 1 TOPS/W analog deep machine-learning engine with floating-gate storage in 0.13 µm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, Jan. 2015.

[182] N. Guo, Y. Huang, T. Mai, S. Patil, C. Cao, M. Seok, S. Sethumadhavan, and Y. Tsividis, "Continuous-Time Hybrid Computation with Programmable Nonlinearities," in *IEEE European Solid State Circuits Conf. (ESSCIRC)*, 2015, pp. 279–282.

[183] A. A. Lazar and Y. Zhou, "Reconstructing natural visual scenes from spike times," *Proc. IEEE*, vol. 102, no. 10, pp. 1500–1519, Oct. 2014.

[184] M. Koyanagi, Y. Nakagawa, K.-W. Lee, T. Nakamura, Y. Yamada, K. Inamura, K.-T. Park, and H. Kurino, "Neuromorphic vision chip fabricated using three-dimensional integration technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2001, pp. 270–271.

[185] J. Burns, L. McIlrath, C. Keast, C. Lewis, A. Loomis, K. Warner, and P. Wyatt, “Three-dimensional integrated circuits for low-power, high-bandwidth systems on a chip,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2001, pp. 268–269.

[186] J. Aoki, Y. Takemoto, K. Kobayashi, N. Sakaguchi, M. Tsukimura, N. Takazawa, H. Kato, T. Kondo, H. Saito, Y. Gomi, and Y. Tadaki, “A rolling-shutter distortion-free 3D stacked image sensor with -160dB parasitic light sensitivity in-pixel storage node,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2013, pp. 482–483.

[187] J. Zhang, L. Huang, Z. Wang, and N. Verma, “A seizure-detection IC employing machine learning to overcome data-conversion and analog-processing non-idealities,” in *2015 IEEE Custom Integrated Circuits Conference (CICC)*, 2015, pp. 1–4.