A Comparison of a Hardware and a Software Integrate and Fire Neural Network for Clustering Onsets in Cochlear Filtered Sound

Leslie S. Smith*
Department of Computing Science and Mathematics
University of Stirling
Stirling FK9 4LA, Scotland, UK
email: lss@cs.stir.ac.uk
Mark Glover, Alister Hamilton
Department of Electrical Engineering
University of Edinburgh
Edinburgh EH9 3JL, Scotland, UK
email {mag, Alister.Hamilton}@ee.ed.ac.uk

Abstract

Onset clustering (which we use as part of a system for sound segmentation) uses integrate-and-fire neurons to perform across spectrum and across time clustering of increases in sound intensity in different parts of the spectrum. We show that a network of recently developed analogue VLSI integrate-and-fire neurons can perform this task in real-time, and compare its performance with a simulated network.

1 Background

A sound wave is a pressure wave, that is a variation in air pressure over time. Virtually all attempts at interpreting sound start by separating the sound out into its constituent frequencies. The mammalian ear [4] is no exception to this, and the cochlea in the inner ear of mammals performs a mechanical filtering of the sound, resulting in a pattern of vibrations on the basilar membrane, a membrane which runs the length of the cochlea. Vibrations at high frequencies are much stronger at the basal end of the cochlea, and low frequencies at the apical end. Transduction of these vibrations into signals on the auditory nerve is performed by the inner hair cells of organ of Corti (which stretches along the length of the cochlea) and the neurons of the

*To whom correspondence should be sent

spiral ganglion. This results in a pattern of neural spikes on the auditory nerve (AN).

The auditory nerve contains a large number of nerve fibres in mammals - about 30000 in man. Because of the nature of transduction, spikes are associated with movement of the basilar membrane in one particular direction. Further, the pattern of firing on an AN fiber in response to a pure tone of constant amplitude is not constant, but starts off high then falls off over a period of time. The auditory nerve innervates the cochlear nucleus, and the response types of many different classes of cells there have been characterised [4].

In [7] we showed (i) that simulated integrate-and-fire neurons could be used to behave like certain of the globular bushy cells in the cochlear nucleus, and provide an onset response (that is, to spike when the sound intensity in some part of the spectrum increased rapidly), (ii) that by using a simple excitatory network, these responses could be clustered across channels and time, producing volleys of spikes which correlated well with broadband bursts of energy in speech, even when the speech was in a considerable amount of (non-speech) noise, and (iii) that this clustering could be used to provide a useful segmentation of a speech signal.

One of the problems with computer simulations of cochlea-based approaches to sound interpretation is speed: cochlear filtering results in multiple channels of sound, and serial processing of multiple channels is slow. We are therefore interested in direct hardware implementation of parts of the system, and have produced a hardware implementation of the network of integrate-and-fire neurons [1].

2 Techniques used

For the work reported here, we used the basilar membrane module of the Gammatone cochlear filterbank [3]. We used parameters which resulted in 29 or 31 channels of data being generated, with centre frequencies (approximately) logarithmically distributed between 60 and 6000Hz. The bandwidth of these channels was set to a value which results in the channel response being similar to that shown by the ear for moderate levels [2]. The output from each channel was then rectified, roughly corresponding the the transduction action of the inner hair cells of the oran of Corti. This results in 29 or 31 channels of positive-going data, each with the same sampling rate, 22050 (claps data) or 16000 (TIMIT data) samples/second.

The number of channels is far fewer than the number of fibers in the AN. The run-time and size of the simulation depend strongly on the number of channels, limiting the number usable to about 120 with the computing facilities available; in addition, we wanted to compare the simulation results with those from the hardware integrate-and fire neural network, and this network consisted of 4 chips, each containing 8 neurons, limiting the number of channels to 32.

The output from each channel was convolved with a difference of Gaussians operator to accentuate onsets, and to model the onset response of real AN fibers. A balanced filter (convolution function) was used, so that the output would eventually fall to zero for a signal of fixed intensity. The filter used was causal, that is, the output depended only on the current and previous values of the input. Where the output was negative, it was replaced by 0. Details are in [7]. The filter output was then downsampled to 4000 samples/second, for use as input to the simulated neurons, and to 1000 samples/second for use as input to the hardware neurons. This lower update rate was necessitated by the requirement to update each channel independently because input to the hardware neurons was multiplexed.

The neural network has one integrate-and-fire neuron per channel. Between spikes, the activity (voltage) of a leaky integrate-and-fire neuron is governed by

$$\frac{dV(t)}{dt} = -\frac{V(t)}{RC} + I$$

where the leakiness is measured by the dissipation ($= \frac{1}{RC}$). The neuron fires when the activity reaches some predefined threshold. The leakiness of each neuron can be varied in both the simulated and hardware neurons (although the range of leakiness is smaller in the hardware neurons). Each neuron has an excitatory connection to its five neighbours in either direction. In the simulation, the strength of this connection can be varied, but it is of fixed strength in the hardware neurons. In the simulation, the spike output arrives at its target at the next simulated instant - that is, 0.25ms later. In the hardware network, the spike output arrives at its target almost instantanously. In both cases, the strength of the connection is such that one excitatory spike results in the post-synaptic neuron increasing its potential immediately by 10% of the total potential required to make the neuron reach threshold. In both simulated and hardware neurons, the refractory period was set to 50ms. Details of the implementation of the silicon neurons are in [1].

3 Results

We report the results of testing the systems with two different sound signals, namely a (locally generated) series of handclaps (sampled at 22050 samples/second) and an utterance from the TIMIT database (dr1/fsjk1/sa1) (sampled at 16000 samples/second). Both the simulated and the hardware integrate-and-fire neural networks give similar results, as can be seen in figure 1. The primary difference in the two onset volleys generated from the claps data occurs at the first clap: the hardware appears to have been affected by some noise. In the onset volleys generated from the utterance, the low frequency channels are less in evidence in the hardware implementation. However, increasing the strength of the input so that more of these appear results in too many pulses appearing. This may have been due to variations between the chips.

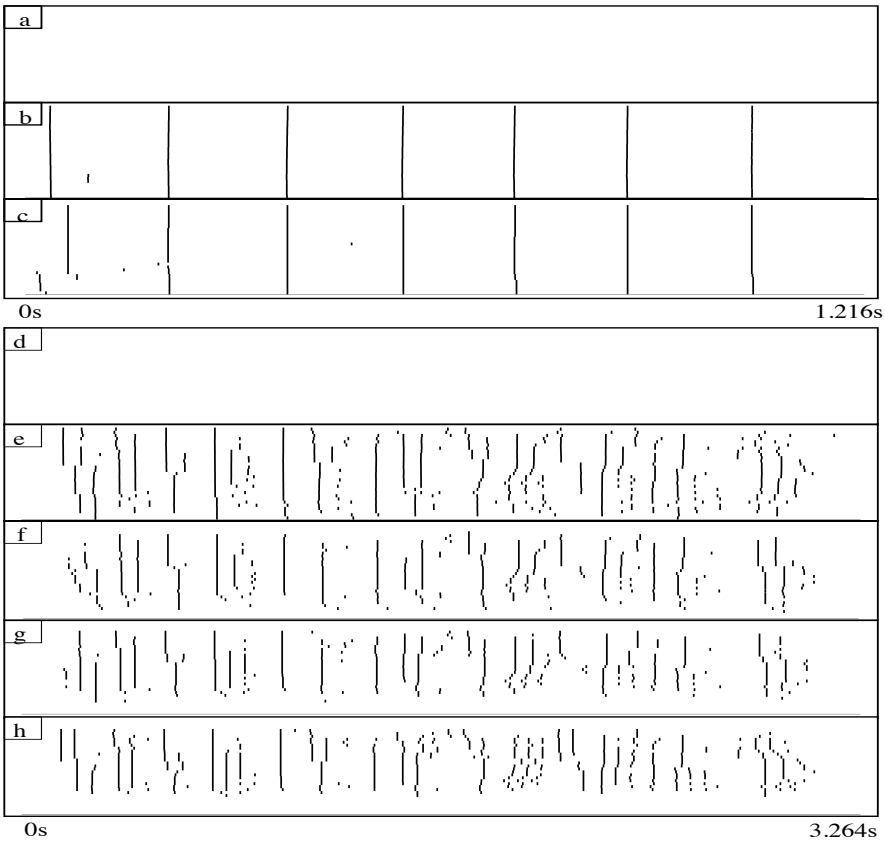

Figure 1: (a) envelope of the claps sound. (b) onset volleys found by the simulated system. Low frequency channels are at the bottom, and high frequency channels at the top. (dissipation = 50) (c) onset volleys found by the hardware system. (dissipation = 40). (d) envelope of the TIMIT utterance. (e) onset volleys found by the simulated system (dissipation=50). (f-h) onset volleys found by the hardware system (f) input at quarter of full strength, dissipation=15 (g) input at half full strength, dissipation = 50 (h) input at full strength, dissipation = 120.

The way in which the integrate-and-fire neural network clusters the data can be seen if we examine a 50ms section of the claps data: see figure 2. Looking at the 50ms section starting just before the second clap, we can turn off the excitatory weight in the simulation, and alter the dissipation of the integrate-and-fire neuron whilst adjusting the input strength to ensure that the neurons still fire. With the current chip, we cannot adjust the excitatory weights; however, we can alter the strength of the input to the chip, adjusting the dissipation to ensure roughly the same number of spikes are generated.

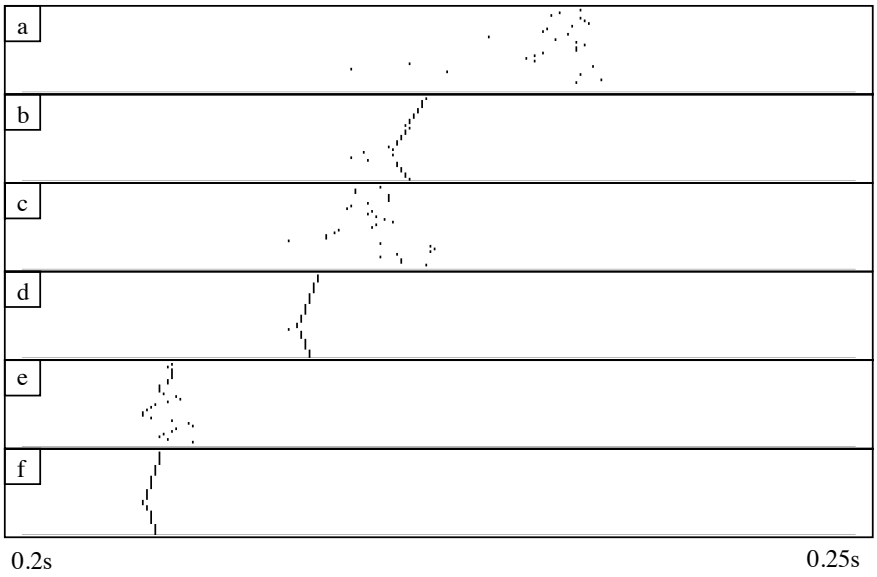

Figure 2: Results for a 50 ms section of the claps sound using the simulator. (a) Integrate-and-fire output with dissipation = 0 (i.e. no leakage), high input attenuation (weight = 0.001), and no excitatory weights between neurons. (b) as (a), except that excitatory weights are on (c) integrate-and-fire output with dissipation = 50, lower input attenuation (weight = 0.0025), and no excitatory weights between neurons. (d) as (c), except that excitatory weights are on. (e) integrate-and-fire output with dissipation = 500, low input attenuation (weight = 0.0175), and no excitatory weights. (f) as (e) except that excitatory weights are on.

The results of this are shown in 3.

Applying the same techniques to the TIMIT utterance, we see some changes occurring as the strength of the input increases and the dissipation decreases: this can be seen in figure 1f, g, and h. In figure 4 we show an enlargement of a 50ms section of the TIMIT spikes.

4 Discussion

It is clear from figure 1 that the hardware integrate-and-fire neural network can perform the same spatiotemporal clustering as the simulated network. From figure 2 one can see that the excitatory connections in the network are responsible for the across-channel clustering.

The input to each neuron is level-based, rather than pulsatile. Firing occurs only when the neuron's activity has built up to the threshold level. Decreasing the dissipation means that less of this activity leaks away over

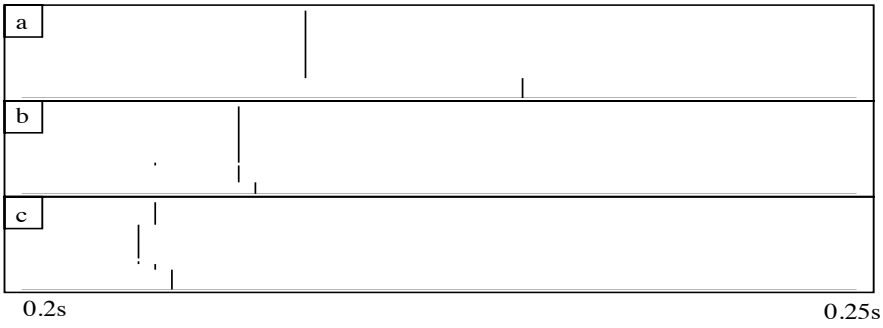

Figure 3: Results for a 50 ms section of the claps sound using the silicon neurons. (a) input attenuated to quarter original strength, dissipation = 10 (b) input attenuated to half original strength, dissipation = 20 (c) full strength input, dissipation = 40.

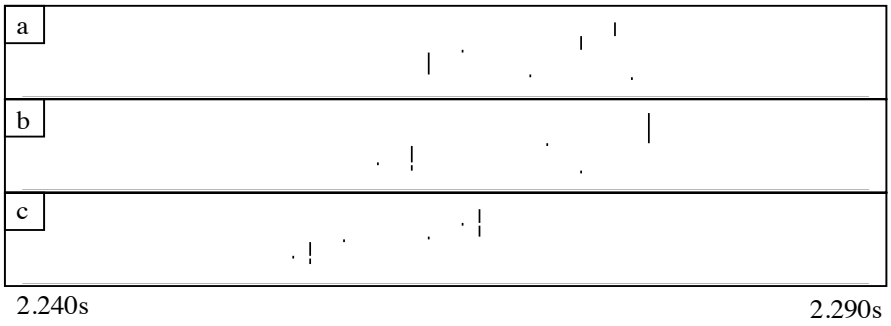

Figure 4: Results for a 50 ms section of the TIMIT sound using the silicon neurons. (a) input attenuated to quarter original strength, dissipation = 15 (b) input attenuated to half original strength, dissipation = 50 (c) full strength input, dissipation = 120.

time: however, one result of this is that small (e.g. noise) inputs can build up and cause extra unwanted spikes. This is why we reduce the strength of the input as we decrease the dissipation: we have attempted to keep the total number of spikes approximately constant. From figure 2 the latency of the first spike depends on the connection strength and dissipation: the stronger the connection, the earlier the spike occurs. Close examination of this figure shows that the timing of the first spike is independent of the inter-neuron excitatory connections, but that these excitatory connections cause the rest of the spikes to occur much more rapidly after the first spike. With the silicon neurons, we are unable to turn off the excitatory inter-connections, but the decrease in latency, and the increase in clustering as the input strength and dissipation increases is clear from figure 3.

Although the decrease in latency with increase in input strength is also visible for the TIMIT speech data in figure 4, the improvement in clustering is not visible. Indeed, looking at the spikes produced from the whole utterance in figure 1f-h, there is little visible difference in the clustering produced. Similar results have been found in simulation, varying the dissipation even up to 1000. We believe this is due to the actual distribution of the increases of energy in cochlear filtered speech. These tend to be of longer duration and skewed across channels, rather than of very short duration and nearly co-incident across channels as was the case for the percussive hand-clap. The expected improvement in clustering as the dissipation increases is offset by the rapid leakage of subthreshold excitation. Whatever the mechanism, the clustering defined by the spike volleys is not sensitive to dissipation.

Even with such such a simple model and network, there are two clustering processes running simultaneously. At the single neuron level, there is the temporal clustering of the input. This is sensitive to the dissipation of the neuron: a considerable amount of input must occur within a short time-period to make the neuron fire when the dissipation is high. At the network level, the excitatory connections result in one neuron's spiking making its neighbour neurons more likely to fire almost immediately. This results in across channel clustering.

5 Conclusions and Further Work

Both the hardware and software neural network perform similar clustering of the sound onsets. We have shown that the silicon integrate-and-fire neuron can cope with such relatively slow-varying data. A new design is under way in which the inter-neuron weight will be variable, allowing experimentation with more complex networks, for example networks which support lateral inhibition, allowing the spectral location of onset clusters to be found.

The integrate-and-fire neuron remains a very simple neural model. One result of using a simple model is the variation in latency of the onset spike depending on the input strength. Real globular bushy neurons use large (and fast) synapses, and low and high threshold potassium ionic channels to achieve rapid onset responses [6, 5], whose latency is very low, and (relatively) independent of input strength. This is of importance for estimates of sound direction based on inter-aural time differences. We intend to continue experiments with more sophisticated (and biologically realistic) neurons, while maintining the real-time response of the network.

Acknowledgements

The authors acknowledge the assistance of Adrian O'Lenskie and Frank Kelly in the design and construction of the circuitry for testing the analogue hardware. Mark Glover is supported by a Ph.D. studentship from the UK EPSRC.

References

[1] M. A. Glover, A. Hamilton, and L. S. Smith. Analogue VLSI integrate and fire neural network for clustering onset and offset signals in a sound segmentation system. In L.S. Smith and A. Hamilton, editors, *Neuromorphic Systems: Engineering Silicon from Neurobiology*. World Scientific, 1998.

[2] B.C.J. Moore and B.R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3):750–753, 1983.

[3] R.D. Patterson, M.H. Allerhand, and C. Giguere. Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform. *Journal of the Acoustical Society of America*, 98:1890–1894, 1995.

[4] J. O. Pickles. *An Introduction to the Physiology of Hearing*. Academic Press, 2nd edition, 1988.

[5] J.S. Rothman and E.D. Young. Enhancement of neural synchronization in computational models of ventral cochlear nucleus bushy cells. *Auditory Neuroscience*, 2:47–62, 1996.

[6] J.S. Rothman, E.D. Young, and P. D. Manis. Convergence of auditory nerve fibers onto bushy cells in the ventral cochlear nucleus: Implications of a computational model. *Journal of Neurophysiology*, 70(6):2562–2583, 1993.

[7] L.S. Smith. Onset-based sound segmentation. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 729–735. MIT Press, 1996.