Parallel distribution of an inner hair cell and auditory nerve model for real-time application

DOI:
10.1109/TBCAS.2018.2847562

Document Version
Final published version

Parallel Distribution of an Inner Hair Cell and Auditory Nerve Model for Real-Time Application

Robert James, *Student Member, IEEE*, Jim Garside, Luis A. Plana, Andrew Rowley, and Steve B. Furber, *Fellow, IEEE*

***Abstract*—This paper summarizes recent efforts in implementing a model of the ear's inner hair cell and auditory nerve on a neuromorphic hardware platform, the SpiNNaker machine. This exploits the massive parallelism of the target architecture to obtain real-time modeling of a biologically realistic number of human auditory nerve fibres. We show how this model can be integrated with additional modules that simulate previous stages of the early auditory pathway running on the same hardware architecture, thus producing a full-scale spiking auditory nerve output from a single sound stimulus. The results of the SpiNNaker implementation are shown to be comparable with a MATLAB version of the same model algorithms, while removing the inherent performance limitations associated with an increase in auditory model scale that are seen in the conventional computer simulations. Finally, we outline the potential for using this system as part of a full-scale, real-time digital model of the complete human auditory pathway on the SpiNNaker platform.**

***Index Terms*—Neuromorphic computing, auditory pathway modeling, inner hair cell, auditory nerve, SpiNNaker, real-time simulation.**

I. Introduction

THE mammalian ear extracts many useful features from a sound stimulus that have been essential in evolutionary survival. Such attempts to replicate these capabilities using various methods of frequency analysis and computer algorithms currently cannot match human performance in a range of hearing tasks [1]–[3]. It is believed that the brain's unique encoding of sound to spiking neuron action potentials, and the non-linear adaptive response of the inner ear, are contributing factors to what enables us to perform fast and efficient sound processing [4]. On a fundamental level the brain's capabilities of performing cognitive tasks on a low power budget are made possible by the interactions between spiking neurons arranged in complex massively parallel networks. This parallelism also features in stimulus representations on the brain's sensory pathways. We postulate that this *full scale* of parallelism should be replicated when modelling such sensory systems. This model can be used to gain a better understanding of the mechanisms in the auditory sensing brain that explains human level hearing performance. To achieve a biologically realistic model of the complete mammalian 'auditory pathway' is the aim of this (and further) research.

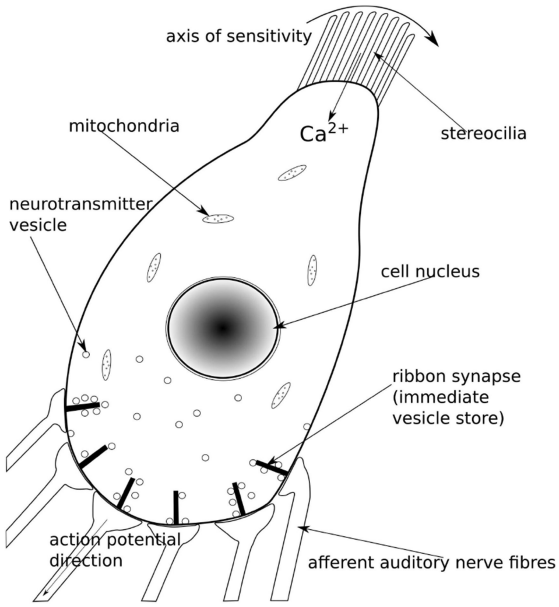

Fig. 1. A simplified inner hair cell with afferent auditory nerve fibres innervating the base of the cell. It performs mechanotransduction in the inner ear, converting physical hair displacement to auditory nerve action potentials.

A. The Auditory Pathway

The auditory pathway begins with a sound pressure wave travelling into the outer ear and eventually displacing the Tympanic Membrane (TM) which separates the outer and middle ear. Inside the middle ear the TM connects to the cochlea via three ossicle bones to continue (and amplify) this displacement into the inner ear cochlea. The cochlea is a coiled, liquid-filled organ that converts the TM displacement into a series of travelling waves along its distance, from base to apex. The frequency components of the sound stimulus dictate the location along

Manuscript received January 30, 2018; revised March 28, 2018 and May 14, 2018; accepted May 26, 2018. Date of publication July 16, 2018; date of current version October 19, 2018. This work was supported in parts by the Engineering and Physical Sciences Research Council EP/D07908X/1 and EP/G015740/1, by the European Research Council 320689, and by the Human Brain Project FP7-604102 funding. The work of R. James was supported by the Engineering and Physical Sciences Research Council Centres for Doctoral Training funding. This paper was presented in part at the 2017 IEEE BioCAS conference proceedings [21]. This paper was recommended by Associate Editor Prof. S. Carrara. *(Corresponding author: Robert James.)*

The authors are with the SpiNNaker Project, The Advance Processor Technologies group, School of Computer Science, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: robert.james@manchester.ac.uk; james.garside@manchester.ac.uk; luis.plana@manchester.ac.uk; Andrew.Rowley@manchester.ac.uk; steve.furber@manchester.ac.uk).

Digital Object Identifier 10.1109/TBCAS.2018.2847562

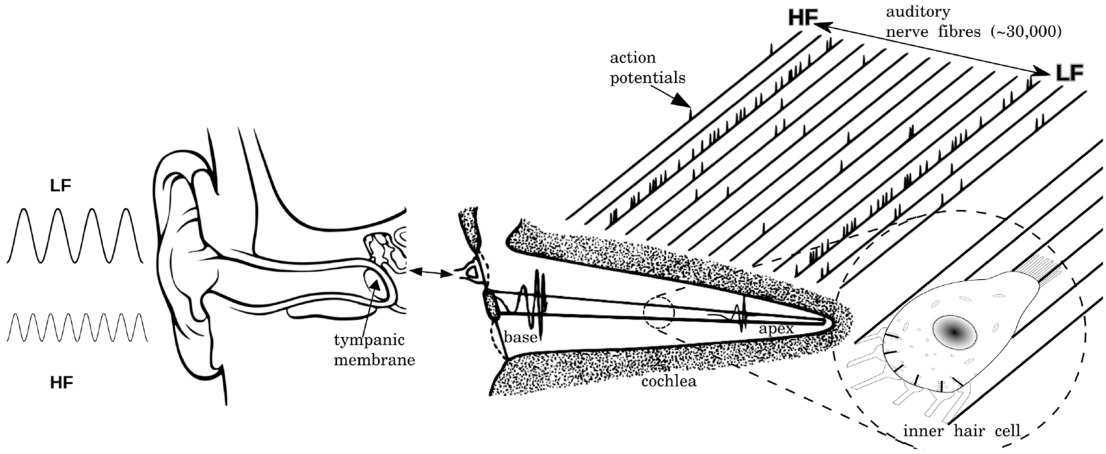

Fig. 2. An uncoiled cochlea (right) with parallel auditory nerve fibres innervating single inner hair cells along the cochlea. The spiking activity due to two stimulus frequency components: High Frequency (HF) and Low Frequency (LF) can be seen in the corresponding auditory nerve fibres.

the cochlea that will experience the most displacement along its Basilar Membrane (BM). High frequencies are absorbed at the basal regions and progressively lower frequencies reach the apical regions of the cochlea. The cochlea is lined with many motion sensitive cells, known as Inner Hair Cells (IHCs), that detect the localised displacements of the BM.

Fig. 1 shows the structure of an IHC and the afferent Auditory Nerve (AN) fibres that innervate each IHC. The stereocilia 'hairs' at the top of the cell move according to the motion of nearby regions of the BM. When the stereocilia move in the axis of sensitivity they allow an influx of calcium ions (Ca^{2+}) to enter the cell body. This influx (of Ca^{2+}) increases the likelihood of neurotransmitter release from the containing vesicles at the 'active regions' around the bottom of the IHC where AN fibres innervate the cell membrane. A release of vesicle neurotransmitter causes an action potential (spike) to occur in the corresponding AN fibre.

Fig. 2 shows a diagram of an uncoiled cochlea; it illustrates how specific characteristics of a sound stimulus are represented by spiking activity in AN fibres. The implementation featured in this work is based on a section of a model of the auditory pathway: the Inner Hair Cell and Auditory Nerve (IHC/AN). The IHCs act as the 'biological transducers' in the ear, converting physical sound-produced displacements into a corresponding spike code signal on the auditory nerve. Our method to model this process digitally is based on the algorithm described by Sumner *et al.* [5]. An overview of this algorithm and how this corresponds to believed biological processes is outlined in Section II-A. To model the functionality of all the IHCs in a cochlea on conventional computer hardware the output for each IHC would have to be generated serially. However such a system can be described as being 'embarrassingly parallel', where the processing of each individual node (an IHC model) does not depend on any other node. Therefore we can model all IHCs in a concurrent fashion. This research effort uses a massively parallel neuromorphic hardware platform to perform this simulation in a biologically inspired fashion. A single frequency-band BM 'channel' will act as an input to an instance of the IHC/AN model, where many parallel instances are distributed across the same machine thus modelling all IHCs at once.

The human auditory brainstem is fed by approximately 30,000 AN fibres from each cochlea. There are around ten AN fibres that innervate each inner hair cell, each with slightly different characteristic responses defined by their 'spontaneous rates' [6]. We argue that the number and variety of AN fibres that make up the cochlea output is vital in allowing the brain to represent sound features to the optimum spectrotemporal resolution and generalised representation in the auditory cortex. This is based on results obtained by Bronkhorst and Plomp [4] & Festen and Plomp [7] in studies that show subjects with sensorineural hearing loss (a reduced number of functioning AN fibres) fail to perform as well as normal hearing subjects across a range of speech reception experiments.

B. Motivations for a Neuromorphic Approach

A conventional computer simulation can be carried out for large scale auditory models, albeit with a compromise in processing time. Therefore we take this opportunity to outline our main motivations for using a neuromorphic hardware platform for these kind of experiments.

Firstly, we highlight the importance of 'in-the-loop' experimentation on complex systems such as large scale Spiking Neural Networks (SNNs). If the user is given rapid feedback from their experimental setup then they stand a much higher chance of understanding *what* it is that they are observing. Specifically, in the case of SNNs, the complex dynamics associated with populations of neurons can be increasingly difficult to understand when the scale of a network is increased. By experimenting with a Real-Time (RT) system the relevant model parameters can be updated to converge on a stable, realistic simulation. This can ultimately be used to answer fundamental questions on how spike-based cognition is performed in biological organisms.

Secondly, we are interested not only in modelling the biological process of converting sound into a spiking signal in the early stages of the mammalian auditory pathway, but also how such a spiking signal is further interpreted by the brain. We propose an

approach to model later stages of the auditory pathway and cortical regions of the brain on the same simulation platform, thus providing an efficient, highly parallel interface between sensory models and the SNNs that infer information from them.

Finally, we postulate the importance of the descending projections that feature between stages of the auditory pathway, even as low down as the efferent fibres that innervate the cochlea Outer Hair Cells (OHCs). It has been shown that descending projections may be providing useful feedback modulation to the incoming sound representation, 'tuning' the representations of learnt salient stimuli [8] and producing stimulus-specific adaptation in sensory neurons [9]. Therefore, if we are to gain a full understanding of the auditory system we must model it complete with multiple feedback projections. Implementing such connectivity across a large complex system in a computer simulation becomes an increasing burden on system communication resources. On our chosen hardware architecture, by using the novel one-to-many 'multi-cast' message routing mechanism, additional descending projections can be integrated into simulations without incurring large overheads and with the ability to simulate RT system feedback.

II. Modelling the IHC/AN

A. Model Algorithm

The IHC/AN model can be split into four main sections. These correspond to a sequential approach to modelling the IHC, from the stereocilia to the auditory nerve synapse.

1) **Receptor potential:** Here the displacement of the stereocilia is converted into a measure of how many ion channels open at the top of the IHC. A large number of open ion channels will cause an increase in the cell's apical conductance. This conductance is used to calculate an accumulated membrane Receptor Potential (RP).
2) **Calcium influx:** The calcium ion influx is modelled as multiple RP sensitive channels that transport calcium ions to the cell's active regions. It is the variation in the conductance parameters of these individual channels that dictates the spontaneous firing rate of the corresponding AN fibre. The spontaneous rate dictates a fibre's response to the same stimuli. It is believed that this variation in response is useful to how sounds are interpreted [10]. Fig. 3 shows different fibre type responses to the same increasing intensity stimuli. The channel Ca^{2+} current is used to calculate the accumulated Ca^{2+} concentration at each active region.
3) **Vesicle release:** An increase in Ca^{2+} concentration at the active regions of the cells increases the probability of vesicle release from the active region's immediate vesicle store (ribbon synapse). The ultimate release of a vesicle is based on this probability determining a random process. A vesicle release will cause an 'ejection' of neurotransmitter into the synaptic cleft between the AN fibre and the IHC active region. The manufacture and replenishment of new vesicles to the cell active regions and cleft neurotransmitter re-uptake are included in the model as similar stochastic processes.

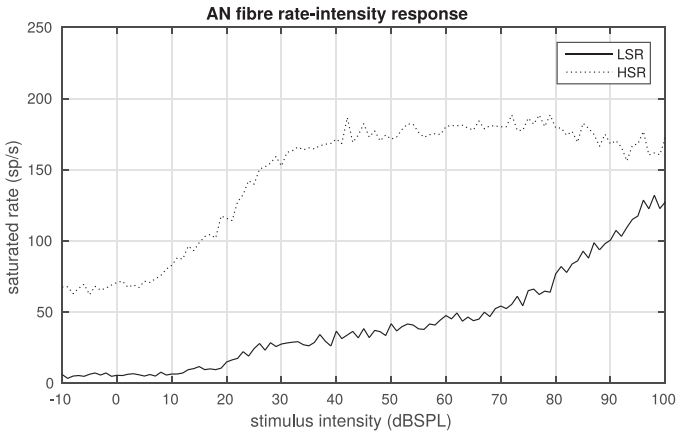

Fig. 3. Saturated AN fibre firing rates of low (LSR) and high (HSR) spontaneous rate types. Results are measured from Matlab Auditory Periphery model [13] simulations to a 1 kHz stimulus at a logarithmic range of intensities in sound pressure level (dBSPL) where 0 dBSPL is the absolute threshold of hearing (28 μ Pa).

4) **Auditory nerve action potential:** This model assumes the neurotransmitter in one released vesicle will trigger a spike in the AN fibre providing it is not already in its refractory (post-firing) period.

Additional detailed descriptions of how this model algorithm is derived are presented in the literature [5], [11]–[13].

B. Hardware Implementation

The chosen hardware platform for this model is a SpiNNaker machine [14]. The SpiNNaker architecture allows for RT execution of large scale SNNs. Its hardware is made up of multiple ARM9 microprocessor 'cores' which can run an individual application asynchronously with respect to their neighbouring cores. These processors can run any interrupt driven software model – not just spiking neuron models. SpiNNaker was designed as a custom architecture supporting SNN models by including a novel 'multi-cast' network routing method. This allows model outputs (neural spikes) to be multi-cast to many target processors efficiently without compromising scalability caused by network saturation inherent with core to core broadcasts. This facility can be exploited in other modelling applications running on the SpiNNaker architecture. The immediate 'embarrassingly parallel' nature of the IHC/AN model's separate frequency band inputs allows a large amount of parallel processing. The implementation presented here exploits this by running multiple instances of IHC/AN models as individual applications across parallel SpiNNaker cores. However, an efficient network communications protocol is necessary to provide each model instance with a continuous input within a RT constraint. Here we use the multi-cast facility to broadcast a single model input efficiently across a network of early stage auditory pathway models. This provides the scaling capacity to produce an IHC/AN model output to a biologically realistic number of fibres with capabilities of RT performance. Such performance would not be possible on conventional serial computation hardware. The low hardware clock rate of the SpiNNaker cores (200 MHz) achieves much lower energy consumption than

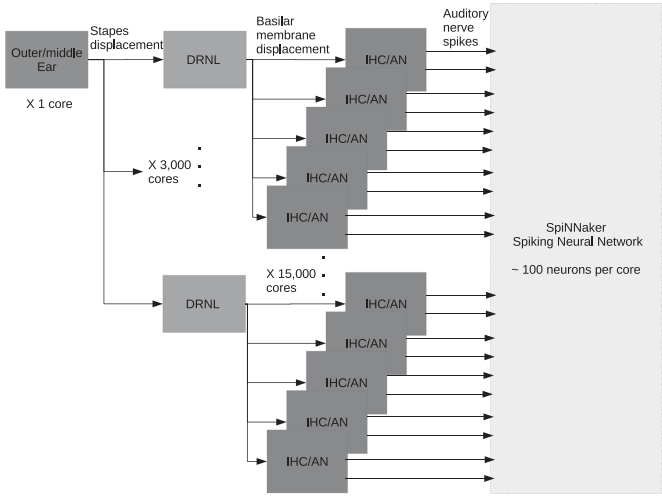

Fig. 4. A schematic for the human full scale early auditory path model distribution on SpiNNaker. The total number of cores for this simulation is 18,001 spanning across 1,500 SpiNNaker chips.

conventional processors whilst maintaining a high throughput due to the parallelisation of the computation.

One drawback of using the SpiNNaker architecture for replicating existing algorithms is that it was built without floating point hardware logic. This decision was based on the assumption that modelling spiking neurons requires fairly low precision arithmetic due to the inherent biological variation. An algorithm implementation on SpiNNaker that uses floating point variable representations uses somewhat inefficient software floating point arithmetic libraries.

C. Software Conversion

The main efforts of this work have been converting the model algorithm to run on the SpiNNaker platform. This involved converting the Matlab Auditory Periphery (MAP) [13] implementation of the model algorithm into an ARM9 executable written as an event-driven C application. The SpiNNaker implementation uses software floating point representation of variables and uniquely seeded pseudo-random number generators. To maintain RT performance each instance of the IHC/AN model processes its input data in small 96×64-bit word segments producing two AN fibre outputs, one of low (LSR) and the other a high (HSR) spontaneous rate. This is fewer than the ten observed fibres that innervate a single IHC however the SpiNNaker simulation generates multiple instances of the model to use the same input, thus allowing for a re-configurable number of fibres of any type to be allocated. An example of how the modelling of a single IHC, with ten output fibres, can be modelled by grouping five IHC/AN instances distributed to receive the same input is illustrated in Fig. 4. This arrangement enables a re-configurable specification to be set before simulation, giving the user the ability to experiment with different model parameters across simulations, e.g. investigating how the number (or type) of AN fibres that synapse a single IHC affects the stimulus representation at higher regions in the auditory pathway.

As each IHC/AN instance is running an algorithm on an ARM9 microprocessor without hardware dedicated for floating point arithmetic all such arithmetic is computed by using a software floating point library (ARM `fplib`). On this hardware, performing mathematical functions such as exponential and logarithm using the default C maths library are computationally costly and consequently jeopardise RT performance. To avoid this inefficiency we use the available fixed point arithmetic SpiNNaker library for computing exponential and logarithm functions in this implementation. This requires conversion of floating point variables to a fixed point `s16.15 accum` data type [15] before performing the operation. Such a conversion can lead to a reduction in numerical precision, this is evaluated in Section II-E.

D. Preprocessing

The input data to the IHC/AN model is generated by previous auditory pathway stage models also running on the SpiNNaker platform. These model the Outer and Middle Ear (OME) and the cochlea from a single sound stimulus input. The OME model consists of a number of digital filters to replicate the acoustic characteristics of the outer ear and middle ear bones. The cochlea model is performed using a Dual Resonance Non-Linear filterbank (DRNL) [16]. This models the cochlea as a parallel, nonlinear filterbank where each filter 'channel' output represents the displacement of a segment of the BM along the cochlea. The distribution of channel centre frequencies is a logarithmic scale from 30 Hz to 18 kHz. To ensure maximum accuracy in the OME and DRNL models all arithmetic operations in the Infinite Impulse Response (IIR) digital filter equations are performed using double (64-bit) precision floating point representation; to maintain RT processing performance in this implementation the OME output is converted to a single (32-bit) precision value before passing data to the DRNL modelling stages.

Much like the IHC/AN models the DRNL instances can be run concurrently, receiving input from the same OME model and producing the parallel BM displacement values as inputs to the IHC/AN models. The complete early auditory pathway model distribution is outlined in Fig. 4. The data transfer between the OME model and connected DRNL models is performed using the SpiNNaker multi-cast-with-payload messaging method. This efficient routing mechanism allows for the output of the OME model to be sent, a 32-bit sample at a time, as a multicast packet payload to all DRNL models located anywhere on the SpiNNaker machine. These incoming samples are stored in a local-to-core memory buffer and are batch processed when the designated processing segment size (96 samples) has been received. Following DRNL processing the output 96×64-bit word segments are stored in a shared on-chip SDRAM memory circular buffer. This allows an efficient block data transfer between a 'parent' DRNL model and its 'child' IHC/AN models (always located on the same chip) necessary for RT performance. The shared memory communication link that triggers a 'read from shared buffer' event in a child IHC/AN model is achieved using a multi-cast packet transmission from the parent DRNL model once it has processed a segment. Fig. 5(a)

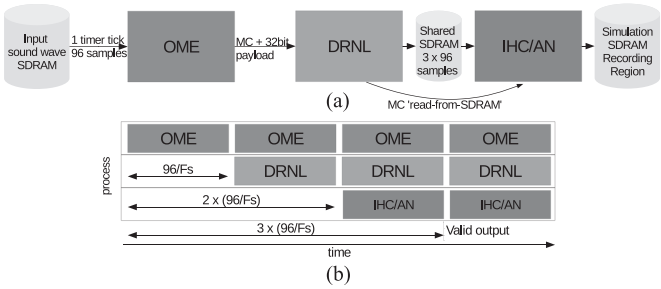

Fig. 5. (a) The data passing method from input sound wave to the output of a single IHC/AN instance using Multi-Cast (MC) and Multi-Cast with payload message routing schemes. (b) The pipeline processing structure used to achieve RT performance.

illustrates these two data communications methods used in the full model system.

In the full system the OME model application is triggered by the RT input stimulus, after which the subsequent DRNL and IHC/AN models in the software pipeline are free to run asynchronously (event driven) until the AN output stage. In a given simulation, to confirm that all model instances have initialised or have finished processing, we use 'core-ready' or 'simulation-complete' acknowledgement signals fed back through the network of all connected model instances to the parent OME model instance to ensure all cores are ready to process and data have been successfully recorded within the given time limits.

Finally we note that, due to using a pipelined model, the system does incur some initial latency between an incoming sound and the arrival of a spike at the AN. We illustrate this in Fig. 5(b) where the total latency occurring until a valid spiking output is recorded from the IHC/AN model will be $3 \times \frac{96}{Fs}$. The chosen model of the IHC/AN used in this study has been shown to replicate physiologically accurate first-spike latency with stimulus intensity relationships [12]. The processing pipeline system used in this work introduces additional latency, however this will be uniform across all stimulus intensities therefore the output fibre latency intensity curves follow the same trend as model implementations on other hardware implementations.

E. Results

The occurrences of spikes along the output AN fibres from a 'Yes' command input stimulus are shown as a raster plot in Fig. 6. This was obtained from a 30,000 AN fibre version of the early stage auditory pathway model (OME → DRNL → IHC/AN) on SpiNNaker. The stimulus audio file used in this simulation was taken from the Google Speech Command Data Set, released under the Creative Commons BY 4.0 license. The complete v0.01 dataset is available for download[1].

The execution times for each stage of the early auditory pathway model stages (OME, DRNL and IHC/AN) for both MAP and SpiNNaker implementations across a range of simulated number of channels were measured. All simulations used a 50 dBSPL sweep tone input from 30 Hz to 8 kHz of 0.5 s duration with a sampling rate of 22.05 kHz. MAP profile results

[1]http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz

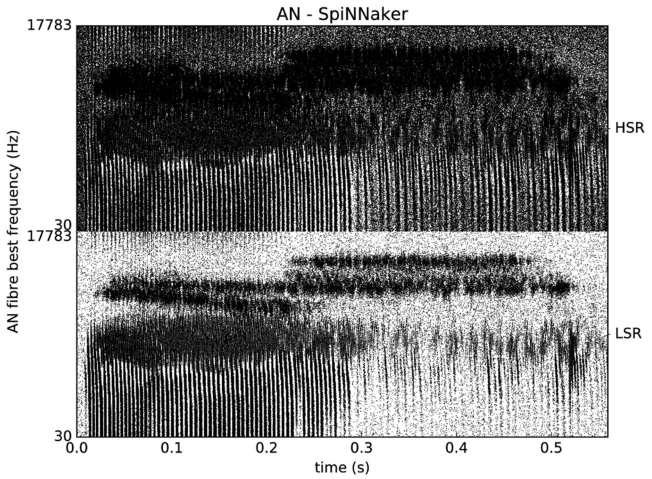

Fig. 6. Simulation results from SpiNNaker implementation responding to a 'Yes' command sound stimulus at 40 dBSPL. The AN fibre responses in the raster plot are separated by fibre type (HSR & LSR) and are tonotopically organised with ascending frequency from bottom to top.

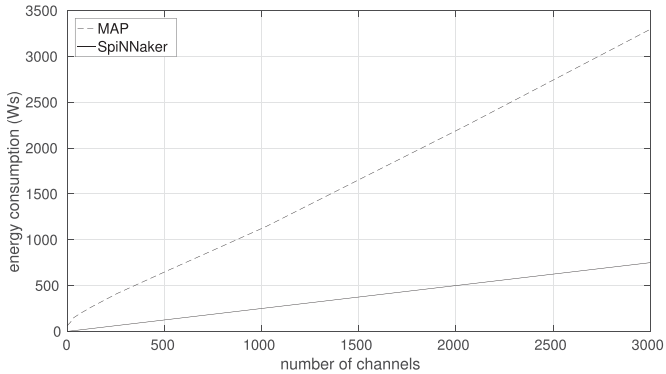

Fig. 7. Average energy consumption of processing a 0.5 s sound sample from 2 to 3,000 channels on both MAP and SpiNNaker implementations. The MAP model is executed on a desktop computer (Intel® Core™ i5-4590 CPU @ 3.3 GHz 22 nm technology) and SpiNNaker on a range of different sized SpiNNaker machines ranging from 1 to 1500 chips (130 nm technology) scaled by the number of channels in a simulation.

showed an increase in execution time with number of channels in all modules, the most significant increase was from the IHC/AN module with a processing time of 35 s at the maximum target number of channels (3,000). The performance profile results of the SpiNNaker implementation showed that with an increase in number of channels, the execution times of the three models did not increase. The average execution time measurements of the OME, DRNL and IHC/AN models were 0.202 s, 0.313 s and 0.477 s respectively. To achieve a RT system in a pipeline configuration, the processing time of each of the three modules needs to be below the input signal duration threshold (0.5 s); the results presented here confirm this.

Fig. 7 illustrates the energy consumed by both implementations across the full range of model channels tested. Energy consumption has been calculated by multiplying the complete processing time by the total power rating of the hardware used (CPU at 84 W, single SpiNNaker chip at 1 W). Here we show that both implementations incur an increase in total energy consumed – but for different reasons. The MAP implementation

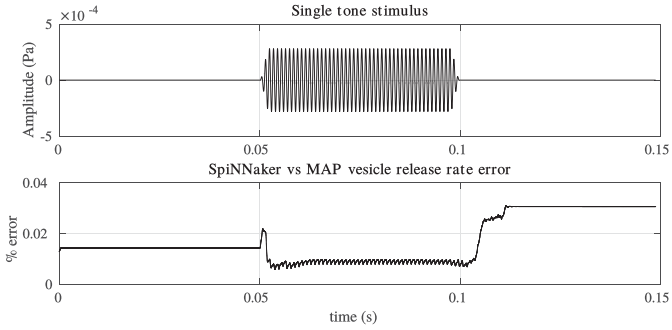

Fig. 8. Comparison of Vesicle release rate values for both MAP and SpiNNaker implementations to the same 20dBSPL 1kHz single tone stimulus.

running on a single, fixed power CPU uses more energy when the number of channels is increased due to the increase in serialised processing time. The neuromorphic hardware experiences an increase in energy consumed due to the increasing size of the machine used (number of chips) with an increase in channels. The rate of increase in energy consumed due to number of channels on neuromorphic hardware is lower than the conventional serial CPU approach. This effect illustrates the basic philosophy that underlies the functionality of SpiNNaker (and biological) processing systems: complex computation on a modest energy budget, performed by dividing overall task workload across a parallel network of simple and power efficient processing nodes.

Due to the stochastic nature of the AN model outputs, any comparisons of an individual fibre's firing patterns across runs will show variation. We first perform a comparison between implementations at the pre-stochastic stage of the IHC/AN algorithm where a vesicle release rate quantity is generated per AN fibre. Fig. 8 shows the maximum difference in the SpiNNaker implementation when compared with the MAP result to be $<0.04\%$. This small discrepancy is caused by three main differences between the implementations compared. Firstly, some precision is lost at the output of the OME in the SpiNNaker implementation due to the conversion from double to single precision values. Secondly, there are differences between the arithmetic hardware that the algorithm is performed on; Matlab simulation hardware uses the Intel x86 architecture's floating point unit (FPU) for high precision arithmetic. The FPU contains output registers of extended precision (80-bit) to prevent any loss of precision during a series of arithmetic operations. The ARM architecture does not support this extended precision operation so we find in some cases very small differences ($\sim 1e^{-24}$) occur between double precision outputs. It should be noted that the IEEE 754 floating point standard guarantees a double precision machine epsilon of $\simeq 2.22\,e^{-16}$ which corresponds to the minimum difference in two floating point numbers that cannot be due to rounding errors. Our tests have shown that the measured error in the double precision calculations from the SpiNNaker implementation fall within this limit. Finally, as mentioned in Section II-C, a loss in precision occurs in the SpiNNaker implementation due to the use of the efficient exponential and logarithm functions that require a fixed point `s16.15 accum` data type. We argue that this total error is insignificant,

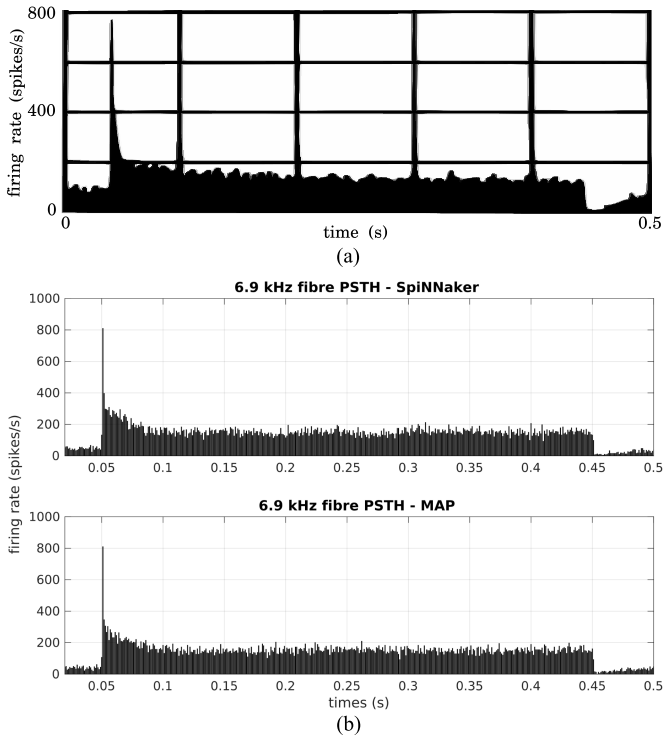

Fig. 9. (a) PSTH responses to 352 repetitions of a 400 ms 6.9 kHz 68 dBSPL stimulus from experimental data obtained by Westerman and Smith [17] of an HSR AN fibre in a gerbil and (b) the same experiment repeated for MAP and SpiNNaker implementations.

especially when it comes to replicating biological processes that are inherently noisy and therefore small-error tolerant.

Finally, to perform a credible comparison between both model implementations at the stochastic AN output Post Stimulus Time Histogram (PSTH), plots of the model outputs have been generated. The results, shown in Fig. 9, show the time varying AN spike rates across 1 ms windows to a 6.9 kHz sinusoidal 68 dBSPL stimulus, first in Fig. 9(a) from physiological data gathered by Westerman and Smith [17] and then from both model implementations in Fig. 9(b). These results show both implementations produce a biologically similar response consisting of pre-stimulus firings of approximately 50 sp/s, followed by a peak response at stimulus onset at around 800 sp/s, decaying to an adapted rate in the region of 170 sp/s. Finally at stimulus removal rates significantly drop during an offset period before returning to spontaneous firing of approximately 50 sp/s.

The largest simulation featured in this work converted a single sound stimulus into a spiking representation on a human number of AN fibres (30,000) using a total of 18,001 cores on a SpiNNaker machine. Such an implementation is easily accommodated on the available machine of 500,000 cores and uses an average power consumption $\simeq$ 1.5 kW.

III. Discussion and Future Work

A. RT Performance

In this work we presented a model of the early stages of the auditory pathway which will not incur a performance overhead

with an increase in scale. It is capable of RT processing of 64-bit input samples at a maximum sampling rate of 24 kHz with an output latency of $\frac{3\times 96}{Fs}$ seconds. Despite the RT processing constraint this solution remains flexible; for example, the addition of AN output fibres (of either HSR or LSR type) in an IHC simulation can be achieved by adding more instances of 2 fibre output IHC/AN models on the same SpiNNaker chip as the parent DRNL instance. These additional instances run concurrently on the parallel architecture thereby not incurring any additional processing time overhead. The multi-cast communications between additional instances is also parallel at no extra cost.

The primary limitation of this implementation is the maximum sampling rate possible for RT simulations. This restriction on model performance is due to the relatively low 200 MHz clock speed and lack of hardware floating point arithmetic support on the SpiNNaker platform.

We can address this limitation a number of different ways:

1) If RT performance is not essential for high audio sampling rate simulations then the SpiNNaker simulation can simply be run at a slower rate, thus increasing the maximum possible sampling rate by a factor of the simulation slowdown, e.g. a 0.5 $\times$ RT simulation will have a maximum sampling rate of 48 kHz.
2) SpiNNaker profiling results revealed the model in the processing pipeline with the largest performance overhead is the IHC/AN model, using 95.4% of signal duration. We investigated an optimisation technique in which the auditory nerve action potential section of the model (see Section II-A) is processed less frequently than the rest of the algorithm, thus reducing the average model processing overhead. The logic behind this is that individual vesicle release and re-uptake models can operate at a much lower rate than the digital filters which model the processes that determine their inputs. By reducing the vesicle processing rate by a factor of four the execution time for the IHC/AN model is reduced to that of the DRNL model. This reduction enables the pipelined system to achieve RT performance at sampling rates up to 34 kHz. It is however currently unclear whether this 'time-binning' approach could lead to a loss in precise spike time information needed for higher level processing in the auditory pathway.
3) The SpiNNaker project is currently developing a second generation SpiNNaker chip which will have hardware support for floating point algorithms, providing faster floating point arithmetic operations thus reducing the overhead of the models described in this work.

We acknowledge that this limitation could be presented as a significant disadvantage to using the proposed system, especially for RT applications. However we believe the introduction of model scalability without performance overhead increase provides a user with a greater advantage in working towards simulating the entire auditory pathway. We emphasise that a traditional computer architecture approach to tackling the problem of modelling such a vast system will inevitably face limitations when additional biological component models are added.

B. Full Auditory Pathway Model

In Section II-E we compared the processing performance of the full-scale SpiNNaker implementation of the model using a power consumption $\simeq$ 1.5 kW with a single thread Matlab implementation using $\sim$ 84 W of power. It is therefore a sound argument to suggest that alternative implementations, perhaps using clusters of CPU (or GPU) processors, could achieve the desired parallel computation in RT. However, the major strength of the approach presented in this work emerges when one considers how to interface the large parallel IHC/AN model output with subsequent auditory brainstem, midbrain and cortical neuron models that comprise the full auditory pathway. In this implementation the IHC/AN model output is already within the SpiNNaker fabric, thus allowing for massively parallel RT interfacing with SNNs running on the same machine. The multi-cast routing protocol used on SpiNNaker provides a scalable solution with the ability to model the complex network of ascending and descending projections in the auditory pathway and gives the potential for real-time, interactive interfaces. It is likely that simulations of the full auditory pathway using alternative hardware architectures, or indeed a highly parallel cluster simulation output to interface with a SpiNNaker machine, would incur performance overheads due to the inherent widespread core-to-core communications or hardware interfacing bottlenecks.

The auditory pathway contains many descending projections; we postulate the great importance of these in providing our active hearing system with useful feedback/modulation. An example of the lowest descending projection is the Medial Olivo-Cochlear (MOC) efferent connections to the cochlea's OHCs. The current SpiNNaker implementation does not currently have model of the Ventral Nucleus of the Trapezoid Body (VNTB) where MOC neurons are located [18]. MOC neurons provide fast feedback to OHCs, producing a non-linear cochlea response responsible for providing protection from loud incoming sounds, increasing the ratio of salient stimuli in a noisy background of sounds [19] and, potentially, a sensory attention mechanism [20]. In future work we will implement a model of the VNTB and the MOC efferent pathways feeding back to the DRNL cochlea model. Such a feature does exist in the complete MAP model; however, for the purpose of this work, we disabled this functionality in the implementation for a comparable algorithm performance assessment. We reiterate the suitability of the SpiNNaker implementation for simulating this biological network, specifically due to the hardware capabilities in routing multiple feedback signals between cores in RT. These fast and complex communication routing requirements would cause performance bottlenecks on the aforementioned alternative hardware implementations when it comes to simulating large scale auditory pathway networks.

IV. Conclusion

This work has shown the feasibility of implementing a human IHC/AN model to a biological scale without compromising processing performance. We exploit the existing SpiNNaker multicast communications mechanism for a novel purpose of scalable

data transmission in a pipelined processing system modelling the early stages of auditory pathway. The results presented in this work show that on our chosen hardware architecture real-time simulation of a biologically realistic sensory periphery can be achieved. For a full scale (single ear) simulation of 30,000 AN fibres the SpiNNaker implementation is distributed across 18,001 cores ($1\times$ OME $+ 3{,}000 \times$ DNRL $+ 15{,}000 \times$ IHC/AN models). This digital algorithm implementation has the flexibility to be refined according to future developments in the field of auditory research. The output of the auditory pathway model will interface directly with other spiking neural processors across the SpiNNaker fabric on an available machine of 500,000 cores.

The drawbacks of using the SpiNNaker architecture for the modelling of the early auditory pathway is evident in the lack of floating point arithmetic logic, this additional hardware will be present in the next generation SpiNNaker chip to support future similar applications. The limitations of the multi-cast data transfer method used here is in the 32-bit precision payload for a multi cast packet which can lead to a loss in model output precision if higher bit representations are required. For the purpose of the algorithms presented in this work we show the model precision losses due to the hardware drawbacks are negligible.

Despite simulating a biological process with digital neuromorphic hardware the total power necessary to achieve this is approximately two orders of magnitude greater than the biological solution, this suggests there is room for improvement in developing biologically inspired computational systems. However, we believe that using a neuromorphic platform for the models presented in this work and later processing of the auditory pathway human perception may be better understood.

ACKNOWLEDGMENT

The authors would like to thank Prof. R. Meddis and C. Sumner for their support and advice during this study.

REFERENCES

[1] M. H. Davis and O. Scharenborg, "Speech perception by humans and machines," in *Speech perception and spoken word recognition*, M.G. Gaskell and J. Mirkovic, Eds., pp. 181–203, 2017

[2] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, no. 1, pp. 1–15, 1997.

[3] R. M. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 34–43, Nov. 2012.

[4] A. Bronkhorst and R. Plomp, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *J. Acoust. Soc. Amer.*, vol. 92, no. 6, pp. 3132–3139, 1992.

[5] C. J. Sumner, E. A. Lopez-Poveda, L. P. OMard, and R. Meddis, "A revised model of the inner-hair cell and auditory-nerve complex," *J. Acoust. Soc. Amer.*, vol. 111, no. 5, pp. 2178–2188, 2002.

[6] C. D. Geisler, *From Sound to Synapse: Physiology of the Mammalian Ear*. New York, NY, USA: Oxford Univ. Press, 1998.

[7] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Amer.*, vol. 88, no. 4, pp. 1725–1736, 1990.

[8] G. Terreros and P. H. Delano, "Corticofugal modulation of peripheral auditory responses," *Front. Syst. Neurosci.*, vol. 9, p. 134, 2015.

[9] M. S. Malmierca, L. A. Anderson, and F. M. Antunes, "The cortical modulation of stimulus-specific adaptation in the auditory midbrain and thalamus: A potential neuronal correlate for predictive coding," *Frontiers Syst. Neurosci.*, vol. 9, p. 19, 2015.

[10] P. A. Fuchs, *Oxford Handbook of Auditory Science the Ear*. New York, NY, USA: Oxford Univ. Press, 2012.

[11] R. Meddis, L. P. OMard, and E. A. Lopez-Poveda, "A computational algorithm for computing nonlinear auditory frequency selectivity," *J. Acoust. Soc. Amer.*, vol. 109, no. 6, pp. 2852–2861, 2001.

[12] R. Meddis, "Auditory-nerve first-spike latency and auditory absolute threshold: A computer model," *J. Acoust. Soc. Amer.*, vol. 119, no. 1, pp. 406–417, 2006.

[13] R. Meddis *et al.*, "A computer model of the auditory periphery and its application to the study of hearing," in *Basic Aspects of Hearing*. New York, NY, USA: Springer, 2013, pp. 11–20.

[14] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, May 2014.

[15] *Programming Languages–C–Extensions to Support Embedded Processors*, International Organization for Standardization, Geneva, Switzerland, Tech. Rep. ISO/IEC TR 18037:2008, 2008. [Online]. Available: https://www.iso.org/standard/51126.html

[16] E. A. Lopez-Poveda and R. Meddis, "A human nonlinear cochlear filterbank," *J. Acoust. Soc. Amer.*, vol. 110, no. 6, pp. 3107–3118, 2001.

[17] L. A. Westerman and R. L. Smith, "Conservation of adapting components in auditory-nerve responses," *J. Acoust. Soc. Amer.*, vol. 81, no. 3, pp. 680–691, 1987.

[18] M. C. Brown, "Anatomy of olivocochlear neurons," in *Auditory and Vestibular Efferents*. New York, NY, USA: Springer, 2011, pp. 17–37.

[19] R. R. Ciuman, "The efferent system or olivocochlear function bundle–fine regulator and protector of hearing perception," *Int. J. Biomed. Sci.*, vol. 6, no. 4, pp. 276–288, 2010.

[20] G. Terreros, P. Jorratt, C. Aedo, A. B. Elgoyhen, and P. H. Delano, "Selective attention to visual stimuli using auditory distractors is altered in alpha-9 nicotinic receptor subunit knock-out mice," *J. Neurosci.*, vol. 36, no. 27, pp. 7198–7209, 2016.

[21] R. James *et al.*, "Parallel distribution of an inner hair cell and auditory nerve model for real-time application," in *Proc. IEEE Biomed. Circuits Syst. Conf*, 2018, pp. 436–440.

Robert James (S'17) received the M.Eng. degree in electronic engineering from the University of Leeds, Leeds, U.K., in 2013. He is currently working toward the Ph.D. degree with the Advanced Processor Technologies group, School of Computer Science, University of Manchester, Manchester, U.K. Before returning to academia, he worked as a Digital Electronics Engineer for a wireless communications solutions company. His research interests include biologically inspired auditory systems, spiking neural networks, and digital signal processing.

Jim Garside received the B.Sc. degree in physics and the Ph.D. degree in computer science from the University of Manchester, Manchester, U.K., in 1983 and 1987, respectively. His doctoral work looked at digital signal processing architectures. He spent some time working on parallel architectures with Transputers. After a brief sojourn in the software industry, he returned to the University of Manchester as a Lecturer in 1991. Later research work has primarily been concerned with VLSI technology, particularly with the Amulet asynchronous ARM processors, and more recently, with SpiNNaker. His current interests include power-efficient processing especially using hardware reconfiguration.

Luis A. Plana received the Ingeniero Electrónico degree from Universidad Simón Bolívar, Caracas, Venezuela, the M.Sc. degree in electrical engineering from Stanford University, Stanford, CA, USA, and the Ph.D. degree in computer science from Columbia University, New York, NY, USA. He worked with the Universidad Politécnica, Caracas, for more than 20 years, where he was a Professor of electronic engineering. He is currently a Research Fellow with the School of Computer Science, University of Manchester, Manchester, U.K. His research interests include neuromorphic systems engineering and energy-efficient, many-core systems architecture.

Andrew Rowley received the B.Sc. degree in computer science and physics and the Ph.D. degree in computer science from the University of St. Andrews, St. Andrews, U.K., he started working with the University of Manchester, Manchester, U.K., as a Research Software Engineer in 2004 and a Senior Research Software Engineer from 2007. He is currently a Member of the team responsible for the development of the software for SpiNNaker and for the interaction of the SpiNNaker machine with the Human Brain Project Collaboratory.

Steve B. Furber (M'98–SM'02–F'05) received the B.A. degree in mathematics and the Ph.D. degree in aerodynamics from the University of Cambridge, Cambridge, U.K., in 1974 and 1980, respectively. He worked with the R&D Department, Acorn Computer Ltd., from 1981 to 1990, and was a Principal Designer with the BBC Micro and the ARM 32-bit RISC microprocessor, and moved to his current position as ICL Professor of computer engineering with the University of Manchester, Manchester, U.K., in 1990. His research interests include energy-efficient many-core architectures and neural systems engineering. He is a Fellow of the Royal Society, the Royal Academy of Engineering, the BCS, and the IET.