Instrument Classification Using Spiking Neural Networks

Jainesh Doshi, Vishrant Tripathi, Onkar Desai, Shreyas Mangalgi *

***Abstract*— This report describes the design of a spiking neural network that can classify different musical instruments based on their timbre and pitch, while also describing a model inspired by the structure of the human ear that converts audio inputs to spike domain for classification by a Spiking Neural Network.**

Keywords - Spiking Neural Networks, temporal encoding. timbre

I. INTRODUCTION

The human ear is a small physical device with disproportionately large properties. When we include the information processing done in the brain and the physiological responses that it elicits, one can see why the Human Auditory System has been giving researchers a hard time since the turn of the twentieth century. Here, we tackle a small problem - musical instrument classification; which our ears and the brain are very adept at solving, but that conventionally requires sufficiently complex signal processing and algorithms to solve[5][8]. We propose a simple model based for this classification problem implemented in spiking neural networks which can be trained accordingly.

II. BIOLOGICAL BASES OF MUSICAL PERCEPTION

A. Timbre

The ANSI definition of timbre describes it as that attribute which allows us to distinguish between sounds having the same perceptual duration, loudness, and pitch, such as two different musical instruments playing exactly the same note[10][11]. In other words, it is neither duration, nor loudness, nor pitch; but is likely 'everything else'. It can be understood naively as the information containing relative amplitudes of harmonics present in a signal. Its usefulness lies in the fact that changing the person playing the instrument or the way it is played doesn't affect the timbre of the piece of sound generated by the instrument i.e. timbre is an inherent property of the instrument/sound. Studies based on psycho-physical judgments of musical timbre, ecological analyses of sounds physical characteristics as well as machine learning approaches have all suggested that timbre is a multifaceted attribute that invokes both spectral and temporal sound features and provides a rich-enough representation for recognition of musical instruments.[1]

B. The Human Ear

The cochlea, or inner ear, constitutes the hydrodynamic part of the ear. It is a small, hollow, snail shaped member formed from bone and filled with colorless liquid. The basilar membrane is a flexible gelatinous membrane that divides the cochlea longitudinally, and it contains about 25,000 nerve endings attached to numerous haircells arranged on the surface of the membrane. The basilar membrane vibrates with different frequencies at different locations causing the fluid in the ear to move and creating shear forces on the hair cells, which then excite the auditory neurons connected to these cells. Patches of hair cells are spatially located according to the propagation of frequency along the Cochlear fluid and thus respond to different frequency ranges, essentially converting the signal's information into frequency domain.[3]

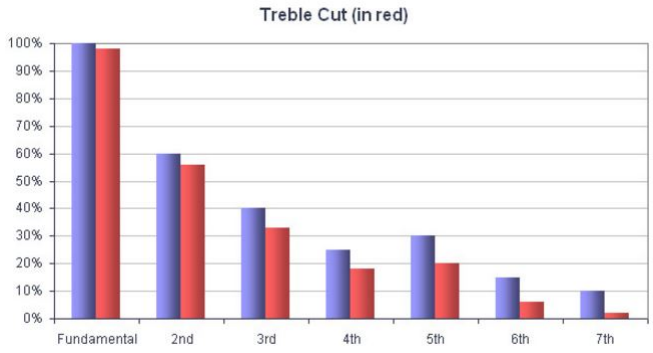

Fig. 1. The timbre of the Guitar remains the same even if it is played in a different way

III. PROPOSED MODEL

The input model that we have used is inspired by the biological structure of the inner ear while the neural network takes the inspiration from a similar network developed for composer classification [4]. The main components are as follows:

A. Modelling the input

The audio signal of an instrument can be divided into small fixed time intervals of length T (sampling). Computing the Fast Fourier Transform of this short duration signal, to gives us the frequency components present in the signal. Each neuron in the first layer (the input layer) excites to certain range of frequencies, i.e. if the played note contains frequencies in a particular band (with amplitudes above a certain threshold), then the corresponding neuron fires, with instantaneous number of spikes proportional to the average amplitude of its excitation frequency in the signal. The excitation frequencies for the input neurons cover five complete 12 note octaves. However the precision of

*Authors are listed in decreasing order of contribution to the project

distinction of frequencies decreases logarithmically as we go to higher frequencies($\geq$1kHz). Thus the outputs of those auditory neurons have to taken into consideration(weighted) accordingly as they may not be a very good indicator of the actual stimulus that we desire. Hence we identify only notes clearly upto the 5^{th}octave. We are using temporal encoding in spikes generated by the auditory neurons that is very similar to the mechanism in the human ear as discussed in section II. Studies have shown that auditory neurons use both spatio-temporal encoding to transduce signals. [2][7][9] Thus our implementation has the capability to adapt itself to nullify the errors subjected to by the input model itself.

B. The Neural Network

we use the Leaky-Integrate-and-Fire (LIF) model of a neurons in our proposition. The first layer of neurons (the input layer) is as follows, the number of neurons is 9 for each note (since we can only hear upto 9 octaves, hence 8 overtones of the least fundamental frequency) thereby one for each upper harmonic, and thus 108 overall(12 notes each having its harmonics in 9 octaves). Thus, we use the amplitudes of at least the fifth harmonic for every note in the 5 octaves.

1) Neurons in the first layer have self-inhibitory connections in the first layer (similar to [4]) to inhibit excessive spiking.
2) The second layer of neurons computes relative firing rates in consecutive harmonics(upto 5th) corresponding neurons of the first layer.
3) The third layer detects specific patterns of amplitudes from the relative firing rates to detect a timbre for a particular note whose connections have been spatially arranged accordingly. At this layer we have the detected the note and its octave by the corresponding firing neurons.
4) The neuron in the fourth layer spikes only if a note is played across any octave.
5) The last layer collects the outputs for the notes played on an instrument and spikes accordingly

The whole network from layer 3 has to replicated alongwith different timbre recognition schemes to identify other instruments at the last layer. Thus we achieve instrument classification (alongwith note detection). The topology of the network can be understood clearly from Fig 2 and Fig 10.

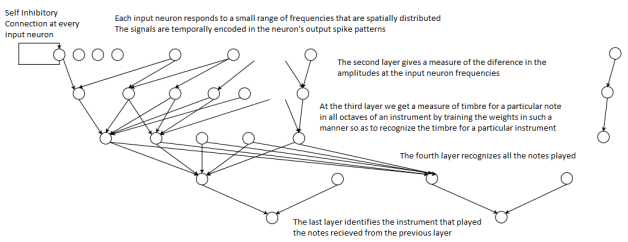

Fig. 2. The final implementation of the network that has been proposed

C. Tritonia Central Pattern Generator

To implement temporal encoding of input signals, we had initially thought that we need to time the arrival spikes from different neurons. For this, we needed some sort of a ”clock” to time events.

We used the idea of Central Pattern Generators (CPG) to create a simple 3 neuron network that generates variable controllable frequency signals of spikes followed by an off period, thus simulating a clock. The inspiration for this network was a well studied and simple CPG network in Tritonia (a mollusc) that controls its swim reflex[12]. The three neurons in the network are connected to each other by excitatory and inhibitory connections, with different delays. Suitable adjustment of weights and delays results in different behaviors being produced as shown in the figures. This CPG has not been used in our final network but can have applications of its own like acting as a global timing circuit for even based neural networks and for generating rhythmic spike trains of variable frequency and phase shifts.

D. Weight Training Rule

Since we have used temporal encoding in the network, hence every different pattern of spikes at the input represent different information. Thus we need to update the weights of the synapse after every spike arrival in a synapse. This is especially to tackle the bombarding of the network with spikes from the auditory neurons in cases of high amplitudes at excitation frequencies or any other scenario as well that might lead to excessive charge fed to the network thereby decreasing its selectivity. The graph in Fig 3 describes the training rule used to update weights between the layers. Thus for closely spaced spiking patterns this weight update rule prevents more charge from being dumped into the circuit to cause more number of spikes and lead to inconsistent results. While spikes that are spaced out in time too aren't affected by the weight update as it doesn't affect the overall pattern recognition ratio of weights. We initialize the weights to different starting values correspondingly for lower and higher frequency components of the network, and update them separately based on the training rule with different parameters. Different layers of synapses required tuning of these parameters to obtain the desired result.

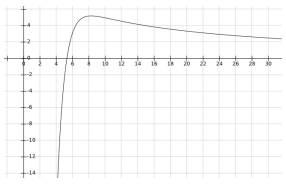

Fig. 3. This is the function used to update the weights downstream and upstream for different layers in our network. x=spike timing difference between the current and previous spike. Thus thus function helps the network to handle very frequent spiking patterns that may be due to aliasing or louder volume of the sound

E. The Complete System

The input encoding scheme described above processes an audio signal and feeds it to a layer of neurons that calculate

the difference between amplitudes of consecutive harmonics. The outputs of these neurons are added in the third layer to get spikes for specific timbre patterns (using the training algorithm) which then identify the note being played as well as the instrument on which it is being played.

IV. OBSERVATIONS

For simplicity, we have reduced the demonstration of our system to a problem of classifying a square wave and a triangular wave of the same frequency. Here, they represent two tones of the same frequency but from different instruments (since they have different timbre patterns). We first train the weights for each pattern and then observe the results for different test cases. This shows clear pattern recognition. The second case a much more practical version than the first performs octave and timbre recognition on a particular note that we have played as our stimulus. The circuit performs as per our expectations and recognizes the note only if it is of the given timbre characteristic and also the octave on which the note is being played as shown in the figure.

A. Resilience to Noise

One of the major benefits of using neural network based approaches in solving various problems is their resilience to noisy data. Here, we have simulated noisy data-sets by adding random noise at the input layer of neurons both in high frequency and low frequency ranges. The network's is resilient to noise, but more precise analysis is needed to be performed to obtain the accurate results.

B. Figures

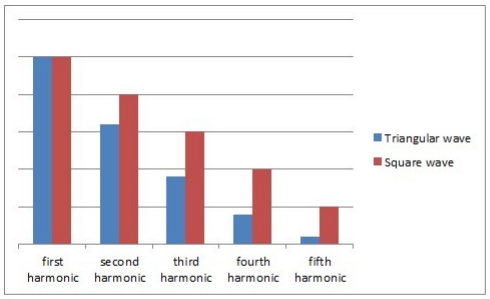

Fig. 4. Difference in Timbre for a square and a triangular wave of same tone (fundamental frequency)

V. CONCLUSION

In this report, we have described a naive spiking neural network based approach for Instrument Classification. Although based upon our model, the datasets became quite simple, We believe that the same approach can be extrapolated for classifying relatively complex timbre patterns. The most important aspect here is that there is no static computations at the inputs of our model similarly as to observed in real biological systems. Temporal encoding has led to the network classify the notes correctly although there

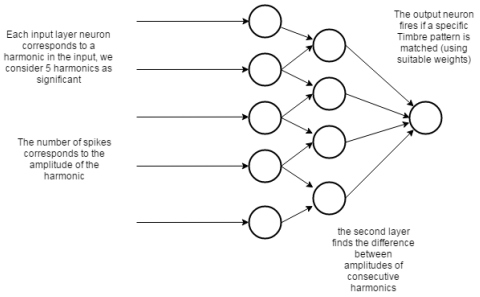

Fig. 5. Simple network for classifying triangular and square waves (Proof of Concept)

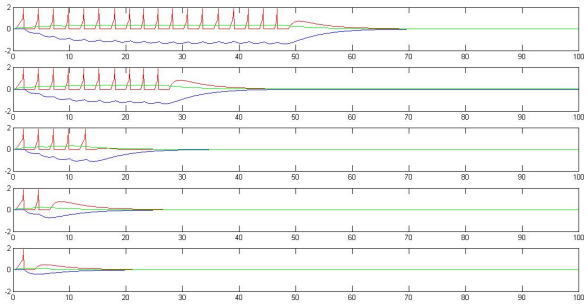

Fig. 6. Response of input layer neurons for a triangular wave (we see harmonics decay as square of the harmonic number

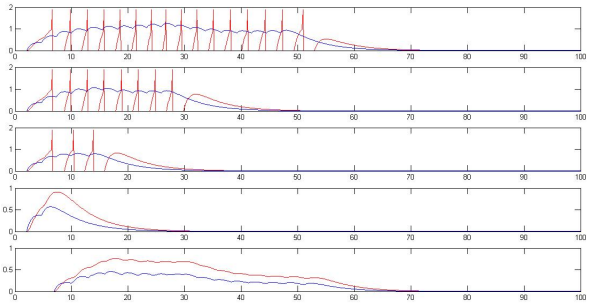

Fig. 7. Response of second layer neurons for a triangular wave (we see difference of amplitudes of harmonics) the output neuron (last in figure) does not spike

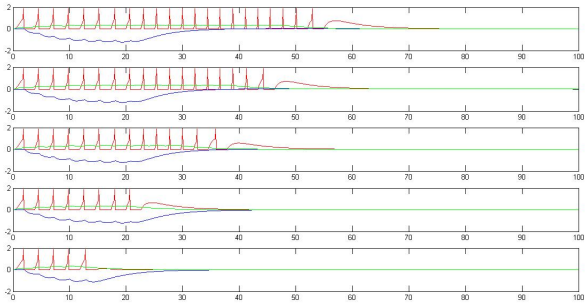

Fig. 8. Response of input layer neurons for a square wave (we see that harmonics decay linearly with the harmonic number)

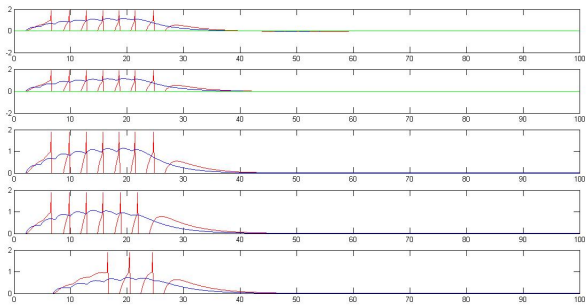

Fig. 9. Response of second layer neurons for a square wave (we see difference of amplitudes of harmonics) the output neuron (last in figure) spikes for the square wave thus achieving classification

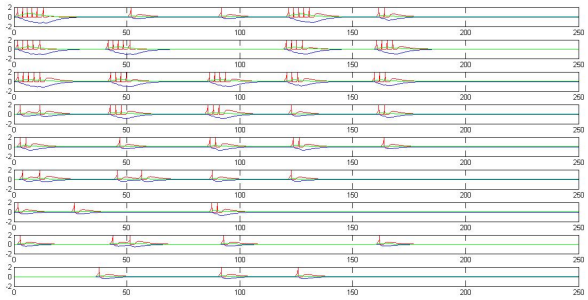

Fig. 10. This is the output of the neurons of the auditory input layer(input layer). Each input neuron has an exciting frequency that are harmonics of the fundamental tone. We have taken the stimuli as follows at t=0 we have the given note played on the guitar in octave I, at t=40 it is played again in the next octave(II), at t=80 it is played on the next octave(III), at t=120 again the same notes is played on octave I, and finally at t=160, the same note in octave I in a wind instrument

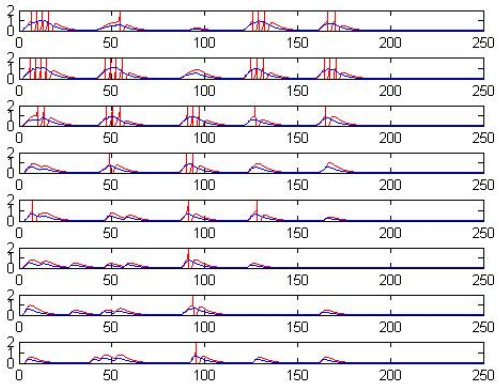

Fig. 11. This is the output of the neurons of the second layer(relative difference layer) for the above explained stimuli

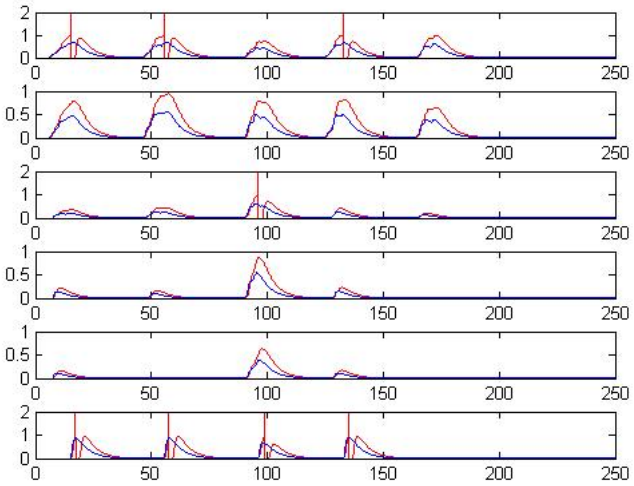

Fig. 12. This is the output of the neurons of the third layer(timbre recognition at the particular note of all 5 octaves) for the above explained stimuli. And the last plot is of the output neuron of the fourth layer that spikes when we have a the particular note played on the guitar whose timbre pattern we know. The expected results can be clearly seen right fr,m the 3rd layer spiking itself. Corresponding neurons for octaves spike after t=0, 40(neuron 18 only, octave I), 80(neuron 19 only, octave II), 120(neuron 20 only, octave III), 160(neuron 18 only, octave I) stimuli while it should not spike after t=200 stimulus that is indeed the output of the network

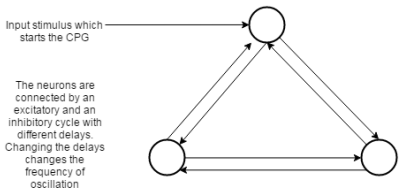

Fig. 13. The neuron network used for generating a ”clock”

are delays in the stimuli that arrive from the auditory neurons. Thus we have to update the synapse weights after every spike is encountered in this type of network as measurement of timbre is a relative pattern about the frequencies, thereby this network is also independent of amplitude changes by smaller factors as it correspondingly updates the weights by the derived function based upon the timing difference between its spikes. Introducing level based parameters made it difficult to arrive at the right parameters for this network. Thus now we have a model that can distinguish instruments that are quite different in timbre to say the least in real time.

More work needs to be done to exploit the advantages of temporal encoding of data in such networks, to be able to apply it to more applications. The tritonia inspired neural network (pattern generator) also needs to be studied further for characterising behaviour and possible applications in other domains, for example motor control.

ACKNOWLEDGMENT

We thank Prof. Bipin Rajendran, Krishnakant Saboo and Chaitanya Prasad for fruitful discussions and help throughout

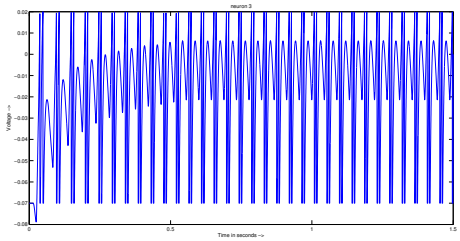

Fig. 14. Output of the Tritonia inspired three neuron pattern generator (we see stable oscillations whose frequency can be controlled)

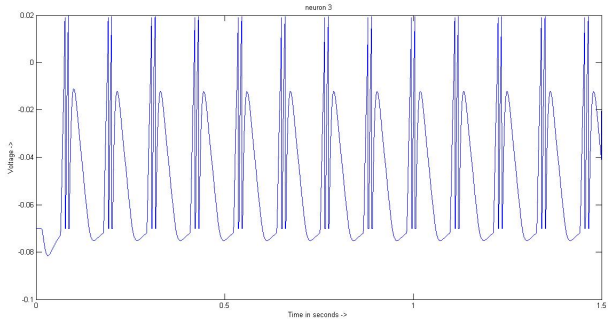

Fig. 15. Output of the Tritonia inspired three neuron pattern generator (low frequency oscillations)

the project.

REFERENCES

[1] *K. Patil, D. Pressnitzer, S. Shamma, M. Elhilali, "Music in Our Ears: The Biological Bases of Musical Timbre Perception" in PLOS Computational Biology, Volume 8, Issue 11, Nov. 2012*

[2] *R. Brette, "Computing with Neural Synchrony" in PLOS Computational Biology, Jun. 2012, DOI: 10.1371/journal.pcbi.1002561*

[3] *Engineering Acoustics/The Human Ear and Sound Perception, in Wikibooks*

[4] *C. Prasad N., K. Saboo, B. Rajendran, Composer Classification based on Temporal Coding in Adaptive Spiking Neural Networks, IJCNN, 2015*

[5] *P. Donnelly, Bayesian Approaches To Musical Instrument Classification Using Timbre Segmentation (Ph.D. Thesis), May, 2012*

[6] *www.freesound.org, F. Font, G. Roma, and X. Serra, "Freesound technical demo", Proceedings of the 21st ACM international conference on Multimedia, ACM, 2013*

[7] *Neural Coding, from Wikipedia*

[8] *X. Zhang and Z. W. Ras, Analysis of Sound Features for Music Timbre Recognition*

[9] *T. Voegtlin, Temporal Coding using the Response Properties of Spiking Neurons*

[10] *American National Standards Institute (1973), Psycho-acoustic terminology S3:20, New York: American National Standards Institute*

[11] *Definitions of Timbre (compiled by G. Sandell), 1997*

[12] *P. A. Getting, Neuronal Organization of Escape Swimming in Tritonia, Journal of Comparative Physiology by Springer-Verlag, 1977*