Neural-Symbolic Learning and Reasoning
AAAI Technical Report WS-12-11

Vowel Recognition in Simulated Neurons

Christian Huyck
Middlesex University, UK
c.huyck@mdx.ac.uk

Abstract

The neural basis of speech recognition and, more generally, sound processing is not well understood. A simple subset of the task of speech recognition, learning to categorise vowel sounds, provides some insights into the more general problems. A simulated neural system that performs this task is described. The system is based on relatively accurate fatiguing leaky integrate and fire neurons, and learns to categorise three categories of vowel sounds. The input to the system is in the form of neural stimulation that relatively accurately reflects the response of biological neurons in the ear to auditory input. The system correctly categorises 91.71% of the vowel sounds using a five-fold test. The system is a sound model of the neuropsychological task of phoneme categorisation, all be it a far from perfect model. As such, it provides an entry into a better understanding of the neuro-psychological mechanisms behind sound processing.

Introduction

Speech recognition is a very complex human skill. While aspects of the mechanisms of speech recognition are understood, the understanding of the full mechanism is far from complete. In particular, there is no simulation that approaches human level speech recognition skill. Consequently, the neural mechanisms of human speech recognition are an excellent domain to explore and simulate because this exploration will improve the scientific community's understanding of the neural mechanism of speech processing, and may provide an improved speech recognition system.

This paper describes a neural simulation of a relatively simple speech recognition task, vowel recognition. To some degree, the sound of a phoneme including a vowel phoneme is the simplest form of an auditory symbol. The system uses a relatively simple neural model with good biological fidelity. The system is trained on instances of three vowel sounds (a, i, and u), and is then used to recognise other instances of those sounds.

The system takes advantage of the Cell Assembly (CA) hypothesis (Hebb 1949). This states that the neural basis of a concept is a CA, a reverberating circuit of neurons. This circuit of neurons can remain active after stimulus has ceased. A CA is a categoriser, categorising an instance of an input as an instance of the concept that the CA represents. In this paper, there are three CAs, one for each of the three sounds. These CAs are learned, and then used to categorise inputs.

Background

While understanding of human hearing and speech recognition is far from complete, a great deal is known. Similarly, while the mechanisms of mammalian neural processing are not entirely understood, a great deal is also known.

Hearing

The ear including the cochlea is a very complex system that performs a wide range of actions crucial for hearing (Robles and Ruggero 2001). For the purposes of the simulation described in this paper, a simple description is that the ear acts as a frequency detector. Hair cells, a type of neuron in the cochlea, respond in a frequency specific manner to sound (Fettiplace and Hackney 2006). All auditory information that enters the brain comes from the cochlea via the hair cells.

There are then a series of steps between the cochlea and the auditory cortex. For example there are connections in this path from the inferior colliculus, in the midbrain, to the medial geniculate body, in the thalamus. This path branches, and there are backward connections; e.g. there are also connections from the medial geniculate body to the inferior colliculus.

Additionally, a wide range of mammals can respond to specific vowels. For examples, gerbils encode vowel sounds (Sinnott and Mosteller 2001). This enables exploration of the brain via invasive techniques, and one such study has shown that gerbils represent vowels on a two dimensional tonotopic map in the auditory cortex (Ohl and Scheich 1997).

Vowels

There are a range of vowel sounds. The first and second formants are both necessary and sufficient to recognise vowels (Peterson and Barney 1952). The first formant of a signal is the frequency with the most power. If a note was played on a piano or a tuning fork, there would be only one formant. The second formant is the one with the second most power; if there were two different tuning forks, the signal

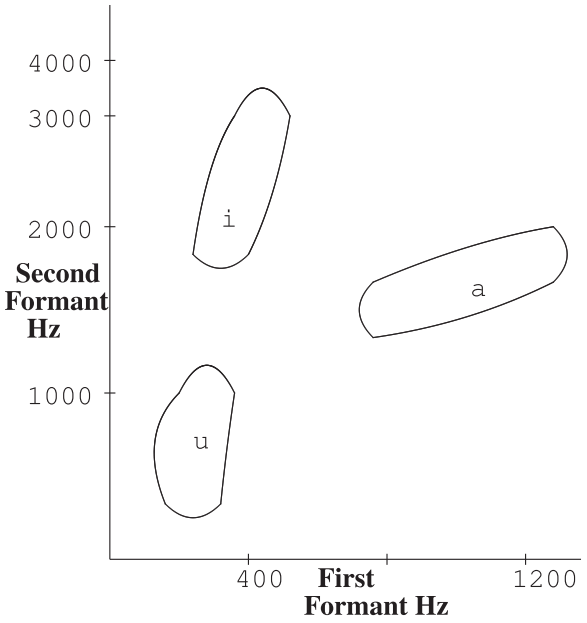

Figure 1: Vowel Matrix (Derived from Peterson and Barney (1952))

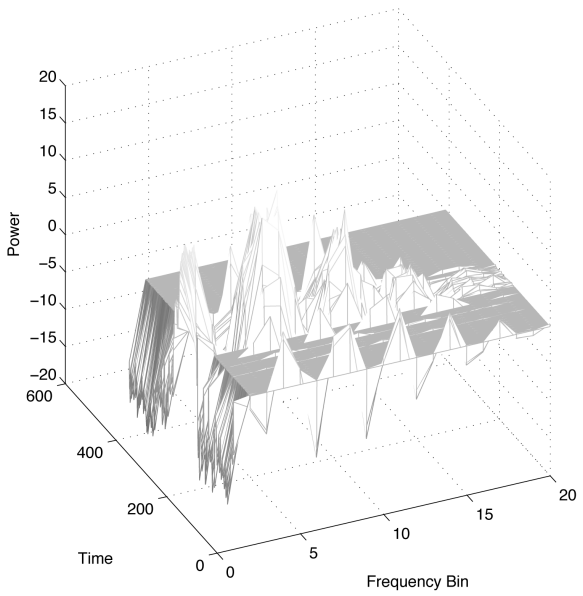

Figure 2: The input of an i sound. It has been binned into 20 frequency bins, and the duration is over roughly 600 time steps.

would have just two formants. In a typical speech signal there are several formants, and indeed in the data used in this paper there were several. Vowels can be recognised by their position in the fist formant (F1) by second formant (F2) matrix. One chart for English speakers is described in figure 1. Depending on accents, this matrix will vary, and different languages have different matrices because languages vary in the vowel sounds they use.

Neural Simulations

This paper is inspired by (Hoshino et al. 2002), who used a Cell Assembly model to learn to categorise vowel sounds. Their system learned to categorise five Japanese vowel sounds. The system was trained on these sounds from five different speakers, and could categorise novel sounds. The input was the first and second formant with input coming to neurons from zero, one, or two external inputs. Connections within the categorising net allowed a Cell Assembly to become active (ignite). This ignition categorised the input sound. Unfortunately, machine learning like results are not reported, so it is not entirely clear how well the system performs; it is entirely possible that it categorised all test sounds correctly.

In this paper, a system that categorises vowel sounds is presented. This system moves beyond the system described by (Hoshino et al. 2002). Instead of extracting the formants directly from the signal, the input to the system activates neurons as they are in the ear (see the Hearing section). Also, categorising neurons receive inputs from both formants and from other frequencies; that is the formant matrix (figure 1) is not part of the topology. This input is then used to drive a set of neurons that learn to categorise.

Data and Preprocessing

Ten instances of each of the three vowel sounds were recorded using a Galaxy AP-830 head set and head phones on five consecutive days by one individual (the author). These were recorded at 44KHz and saved in .wav format. These files and other code required to run the simulations are available from http://www.cwa.mdx.ac.uk/chris/hebb/speech1.html.

Using matlab, a fourier transfer was done to translate the signal from the time domain to the frequency domain. This was then translated to mel-frequency cepestral coefficients (MFCCs). This is thought to reflect the actual frequency mapping in hair cells (Holmberg and Gelbart 2006). This was later translated to input neural firing in the system (see Input and Training below). Before this was done the frequency matrix was normalized so that all files produced roughly the same number of input neuron spikes per sound file.

Note that, very roughly speaking, the input is what the brain receives. The MFCC sorted frequencies are an approximation of the firing of cochlear hair cells. Also note that the normalization does not reflect correct hair cell behaviour. Hair cells fire more rapidly at higher volumes, which is not the case in this simulation.

System Description

The system described below is a neural model. It has many similarities to the brain, but in many ways is an extreme simplification. Processing is done in the system by simulated Fatiguing Leaky Integrate and Fire (FLIF) neurons. In the simulation, these neurons are broken into two subnets which correspond roughly to parts of the cochlea and primary au-

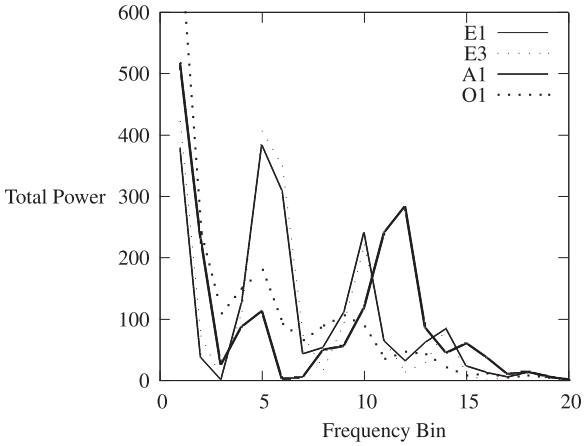

Figure 3: MFCC inputs over the full time averaged for sample vowells. E1 refers to the first instance of E (from keep), and E3 the third. A1 and O1 refer to the first a and u (from father and loot).

ditory cortex. Training is done with a 5-Fold test that is a standard mechanism for machine learning.

Neural Model

The neural model that has been used in this simulation is a fatiguing leaky integrate and fire (FLIF) model. The model has been aligned to biological data (Huyck 2011) so that the model emits spikes at similar times to a biological neuron with the modelled input current. The particular neuron was rat somatosensory cortical neuron, and the behaviour of hair cells and auditory cortical neurons, whether rat, human or gerbil, is likely to be different.

The simulation uses discrete steps, and each neuron has an activation at each time step (initially 0), and this activation is described by equation 1.

$$a_i^t = \frac{a_i^{t-1}}{d} + \sum_{j \in V} w_{ji}, d > 1 \quad (1)$$

The activation at time t, a_i^t is the sum of the new activation it receives, and the activation from the prior time step divided by a decay factor d. The decay factor models the leak. Integration is done by passing activation from all connected firing neurons V weighted by their synaptic weights W_{ji}. If a neuron spikes, it loses all activation, though the incoming activation may cause it to spike again in the next cycle.

The neuron fires if its activation surpasses the threshold as described by equation 2. The threshold is a constant θ added to the current fatigue of that neuron. Fatigue is initially 0, $F_i^0 = 0$, and never goes lower than 0. Fatigue increases by a constant, F_c, when the neuron fires, and decreases by a separate constant, F_r, when the neuron does not fire. Fatigue makes it more difficult for a neuron to fire at a high rate continuously.

$$a_i^t >= \theta + F_i^t \quad (2)$$

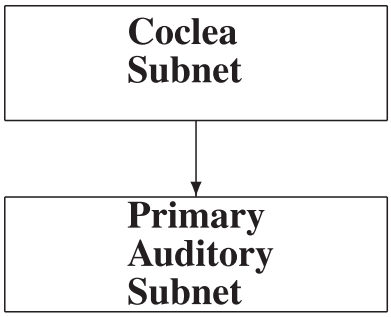

Figure 4: Topology of Intra-Subnet Connections for Vowell Learning.

The basic neural parameters for both subnets used in the simulation were $\theta = 2.2$.. $d = 1.12$, $F_c = 0.045$ and $F_r = 0.01$. These were the parameters derived from biological data (Huyck 2011).

Topology

The topology is the way the neurons are connected. The neurons are broken into two subnetworks, and the two main features are connectivity within subnets and connectivity between the subnets.

There are 400 neurons in the cochlear subnetwork. All are excitatory. They are arranged in rows for the purposes of input (see next subsection). There are no connections between neurons in this subnetwork as is the case in the cochlea (Moser, Brandt, and Lysakowski 2006).

There are 600 neurons in the Primary Auditory (or Phoneme) subnetwork, in a 20x30 array. 30% of these are inhibitory. This subnetwork is preset so that it has three CAs, and there are no connections back to the input (cochlear) subnetwork. The excitatory connections are selected with a distance biased connectivity that has been used in other simulations (e.g. (Huyck 2007)). This connectivity is biologically inspired with excitatory connections locally and to one other part of the net. The inhibitory connections were randomly selected (using the Java random function); each neuron had 40 inhibitory connections, though as there are no self-connections, a neuron could have less if a self-connection was selected. The weights of the connections are preset to make the three CAs, with three bands of 200 neurons. Intra-CA connections are weighted .2 and -.1 for excitatory and inhibitory, and connections between CAs were weighted .01 and -1.

Figure **??** shows the basic form of the connections between the subnets. The only plastic connections were between subnets. Each neuron in the input subnet projected to 20 random neurons in the phoneme subnet. As all of these neurons were excitatory, all of these synapses were excitatory, and were initially given a very small value of .02. During training (and testing) the weights ranged between 0 and .4 and were determined by a Hebbian rule in response to input.

The simulation used a relatively simple correlatory learning rule. Briefly (but see (Huyck 2007) for a more complete

description), this Hebbian rule reflects the likelihood that the post-synaptic neuron fires when the pre-synaptic neuron fires.

$$\Delta^{+} w_{ij} = (A - (w_{ij} * A)) * R \tag{3}$$

$$\Delta^{-} w_{ij} = A * (w_{ij}) * -R \tag{4}$$

The weight is increased (equation 3) when both pre and post-synaptic neurons fire. It is decreased (equation 4) when the pre-synaptic neuron fires but the post-synaptic neuron does not; it is a pre-not-post Hebbian learning rule. R is the learning rate, .001 in these simulations. A is a weighting factor, $A = 0.4$ in these simulations. So if the post-synaptic neuron always fires when the pre-synaptic neuron does, the synapse has a weight near .4, and if it never does, the weight approaches 0. In the simulation, the learning rate is very low (.001) so that the network is not unduly influenced by the most recent learning events.

Input and Training

Training and testing were divided into a series of Epochs. In an epoch, one example of a vowel sound was presented. In the testing epochs, the CA with the most neurons fired was picked as the category of that vowel sound. Activation and fatigue of all neurons were reset to zero after each epoch.

Each voice file was translated into a 20xN vector. With each of the N elements corresponding to roughly .005 seconds of the file. Each of the 20 elements corresponded to a frequency range, and the value of that cell reflected how much power that particular range of frequencies had over that particular time.

The input subnet was directly associated with these 20 frequency ranges, so that every row of neurons corresponded to one frequency; that is 20 neurons were associated with each frequency. The power of a particular frequency was almost always 0, and rarely over 10. Depending on that power, a number of the associated neurons were randomly selected. For example, if the power of the lowest frequency at the 105th time was 8, 8 neurons were selected randomly to receive external activation; if a neuron was selected twice, only 7 were fired.

Each input was fired for three cycles. In the simulated neural model, each cycle is roughly 10ms, and as the input is every 5 ms, this is a flaw in the model.

Each test was a 5-fold test. That is, the system was trained on one set of input files, and tested on five sets (including the training set). The sets consisted of two instances of each of three vowels: u as in loot, i as in keep and a as in father; a, i, and u are standard terms for these vowel sounds. All 30 examples are spoken by one speaker (the author). Each training session alternated between the six training instances, i,u,a,i,u,a 40 times for a total of 120 epochs.

During the training epochs, the correct CA in the phoneme subnet was stimulated. 100 of the 200 neurons were randomly selected and externally activated. Both input neurons and phoneme neurons, when externally activated, received $(1 + r) * \theta$ units of activation. Where r is a random number between 0 and 1 and θ is the firing threshold. Typically this means the neuron fires but in the case of the phoneme neurons, they might not fire due to accumulated fatigue.

After this, learning was turned off for testing. Testing lasted for 600 epochs, going through each of the 30 voice inputs 20 times.

After one full session (720 epochs), a new network was created. As there is randomness in the creation of the net, and the selection of neurons to activate, each net will behave differently. This new network was trained using the next test set.

Results

The model was tested on 100 nets, 20 5-fold tests. The average correct prediction was 92.68%. This test does consider tests on the training set. If the training set is removed from the test set, the average correct prediction is 91.71%. The standard deviation is 6.17%, the best net got 99.38%, and the worst 67.29%. The networks did fail on the training sets during testing, though did get 96.54%.

The networks learned better from some input file sets than from others. The results from the first set ($i1$,$u1$,$a1$,$i2$,$u2$,$a2$) had an average recognition of 92.85%, the second 92.67%, the third 95.74%, the fourth 90.69%, and the fifth 91.43%.

A variant of the system performed better. When $d = 1.11$, $F_c = 0.8$ and $F_r = 0.5$ on the input net, the average performance was 93.31%. However, in this case, many of the test runs had no neurons fire, and in this case the system always guessed u. This raises the question of whether the CAs actually ignited. Remember that the categorisation decision was made symbolically by summing all the neurons fired over all the cycles of a test (typically around 900). In a typical case, 0 of the a neurons fired, 77491 of the i neurons fired, and 18 of the u neurons fired, which quite clearly shows the i CA has ignited. With this second set of parameters, in only slightly over half of the 60000 test cases did over 1000 neurons fire in one CA.

Discussion

This model is another step in the community's developing understanding of neuropsychology. A typical human performs the relatively simple task of vowel categorisation thousands of times in a typical day.

Despite a reasonable performance in this three category task, particularly when only two instances of each category are used to train, as a machine learning algorithm, the performance is relatively weak missing several percent of the examples. While it is likely that the performance could be improved by altering the training regime or using more neurons, sound performance is only one of the goals of this system. A second goal is to perform the task as humans would.

The system is an advancement over that of (Hoshino et al. 2002). Instead of preprocessing the input voice sound to extract formants, the system translates the voice sound to firings that are similar to those that the first set of neurons receives. That is, as a neuro-cognitive model, the input is pretty close. It is likely that a simple statistical algorithm

could account for the variance and perform the task at or near perfection, surpassing the system presented in this paper.

Unfortunately, as a neuro-cognitive model, there are several flaws. These flaws include a form of supervised learning, topological inaccuracy, and training regime inaccuracy. The training mechanism is to activate the correct output CA while the sound is presented. It is far from clear how the correct inputs are learned, but it is clear that the learning brain is not presented with the correct category. From a simulation point of view, a better model would be presented with instances, and would self-organise to categorise them.

Similarly, the topology of the simulation is inaccurate. While the input is a reasonable approximation to the biological hair cells, and the output is some form of approximation of the auditory cortex, biologically, there is a complex series of neural steps between the two. This includes a neural path through the cochlear nucleus, the medulla, the inferior colliculus, and the medial geniculate body. Of course it is difficult to see precise neural behaviour in the human brain in each of these areas because electrode placement is currently the only means of measuring precise neural behaviour and electrodes are invasive. Fortunately, many types of mammals can represent vowels. Less invasive techniques show correlations between those mammals and humans, making it plausible to reconstruct models of vowel recognition from neural behaviour. There is evidence that the neural representation of vowels is based on the first two formants (Ohl and Scheich 1997), but it is not entirely clear how those formants are extracted. That is, it is not clear how the brain translates the signal from hair cells to the first and second formant, and how that signal stimulates the auditory cortex.

Another flaw is the training regime, and the discovery of a correct training regime may be more difficult. It seems likely that like vision (Hubel and Wiesel 1962), hearing requires some input at critical stages. However, unlike vision, it is not clear how to prevent a neonatal mammal from having sound as input, as the mother's heartbeat will cause a sound. It is far from clear how the pattern for vowels is learned. When is it learned? Does it involve synaptic and neural death? What effect does spontaneous activation have? How important is synchronous firing? While many of these questions can be answered with current technology, it requires long-term developmental studies and neonatal animals may need to be included.

Conclusion

This paper has described a neural simulation that learns to categorise three vowel sounds. The inputs is relatively accurate biologically as is the neural model. Using a five fold test, the system categorises 91.71% of the vowels correctly.

While the performance is reasonable, it is not very good from a machine learning perspective. Similarly, the model has several flaws as a neuropsychological model, not least the use of sound normalisation. None the less, and despite its weaknesses, this model is the best neuropsychological model of vowel recognition, that the author is aware of, because it is the only one using simulated neurons starting with reasonable neural input.

There are many plausible next steps with this work. Remaining with simulation, the next step might be the full range of vowels. It would be straight forward to include the full range of vowels in the model by merely increasing the size of the phoneme net. Similarly the training regime could easily be modified to include more vowel sounds, and for that matter a wider range of speakers and a larger number of training and testing instances.

Another way forward is to use unsupervised learning. Here the speech signals would be input, but no output would be provided. This might benefit from some sort of recruitment learning (Diederich, Gunay, and Hogan 2010), but also might be effective without it.

Beyond vowels, other phonemes could be recognised. While vowels depend on the first two formants, other phonemes require different features. Clearly these features are derived from the hair cell firings, but it is not clear what type of intermediary neural processing would be useful. A similar thing could be said for vowels. An exploration of the mechanisms of neural feature extraction would be very useful.

Once all phonemes are recognised with reasonable accuracy, full speech recognition could be the next step. It is hoped that at this step, by basing all processes in neurons, top down and bottom up processing will be able to interact to gain synergy in speech processing. This should provide a good cognitive model; for instance, people report hearing sounds that have been masked out of speech (Warren 1970). The model should also do that. Moreover, this may provide the basis for an improved speech recognition system.

Somewhat orthogonally to this speech recognition task, volume normalisation is a problem that should be addressed. Roughly the same neurons fire in the auditory cortex when a vowel is heard at a low volume or a high volume. How is this volume normalisation managed?

Also, the mammalian hearing system accounts for a wide range of inputs beyond speech. How can other hearing behaviour be managed. For example, how can direction of a sound be determined?

All of this future work on performance can also lead to improved biological models. It is important that inspiration and direction are drawn from an understanding of the actual biological behaviour.

This may lead to exploration of biological data. This would require experts in deriving the biological data. Hopefully, this work will be able ask those neuro-biologists some interesting questions.

References

Diederich, J.; Gunay, C.; and Hogan, J. 2010. *Recruitment Learning*. Springer.

Fettiplace, R., and Hackney, C. 2006. The sensory and motor roles of auditory hair cells. *Nature Reviews Neuroscience* 19–29.

Hebb, D. O. 1949. *The Organization of Behavior*. J. Wiley & Sons.

Holmberg, M., and Gelbart, D. 2006. Automatic speech recognition with an adaptation model motivated by audi-

tory processing. *IEEE Transactions on Audio, Speech and Lnaguage Processing* 14:1:44–49.

Hoshino, O.; Miyamoto, M.; Zheng, M.; and Kuroiwa, K. 2002. A neural network model for encoding and perception of vowell sounds. *Neurocomputing* 44–46:435–442.

Hubel, D., and Wiesel, T. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* 160:106–154.

Huyck, C. 2007. Creating hierarchical categories using cell assemblies. *Connection Science* 19:1:1–24.

Huyck, C. 2011. Parameter values for flif neurons. In *Complexity, Informatics and Cybernetics: IMCIC 2011*.

Moser, T.; Brandt, A.; and Lysakowski, A. 2006. Hair cell ribbon synapses. *Cell Tissue Research* 326:347–359.

Ohl, F., and Scheich, H. 1997. Orderly cortical representation of vowels based on formant interaction. *PNAS* 94:9440–9444.

Peterson, G., and Barney, H. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24:2:175–184.

Robles, L., and Ruggero, M. 2001. Mechanics of the mammalian cochlea. *Physiological Review* 1305–1352.

Sinnott, J., and Mosteller, K. 2001. A comparative assessment of speech sound discrimination in the mongolian gerbil. *Journal of the Acoustical Society of America* 110:4:1729–1732.

Warren, R. 1970. Perceptual restoration of missing speech sounds. *Science* 167:392–393.