CNS 187 Neural Computation - Final Project Report

Ling Li, ling@cs.caltech.edu

December 5, 2000

In this report, I discussed what information in the speech may be crucial to the *Mus silicium*, and what scheme may be used in that organism to recognize single words. Several other issues, including the training and the generalization of the network, are also included.

Some points of this report are: 1) The times of events is the only information input to the organism. Thus the relative times of events, or, the order of those events is crucial to the speech recognition. 2) The organism uses paired α and β cells as delay lines; the γ neurons are used as the coincidence detector. 3) Events in one word are detected in several steps instead of all at once, which makes more than 1 spike of the γ neuron. 4) The information of the network is stored as the decay times of neurons and connection pattern between neurons, thus is robust to cellular variations. 5) The training of the network when more than 1 training instances are used is much more complicated.

Event Times

The basic structure of the *Mus silicium* consists of area A and area W. Area A cells are triggered by the times of events produced by a "front end" which takes a sound file, carries out a spectral analysis of it, and detects events in each of 40 different channels, each channel with a particular preferred frequency and of a particular flavor ("on", "off", "peak").

Thus, all the information passed through the front end is the times of events in 40 channels. This hints that the organism uses only the relative times of those events, or maybe even less, the order of those events, to recognize voices. This seems familiar to us, since we know from [1, 2, 3] that the time can be represented and calculated by normal integrate-and-fire neurons.

For each channel there are 20 cells with different decay times in area A. Every cell has a continuing response to a particular event, supplying the α and β cells in area W with encoded information about the time elapsed since the beginning of the event. Different decay times can be used in different combinations to extract the time information.

Two schemes on how to facilitate the combinations of neurons with different decay times are discussed below. One is the coincidence detect, and the other is the α-β pairs. Though I now think that the former scheme is not suitable for *Mus silicium*, however, it is still worth being written down, and, it has also been discussed a lot by some other students.

Scheme 1: coincidence detect

In area A, the firing rate of one cell A_i with decay time d_i is like

$$f_i \approx \begin{cases} \left(1 - \frac{t-t_i-\delta_i}{d_i}\right) f_{\max}, & \text{when } t_i + \delta_i \le t \le t_i + \delta_i + d_i; \\ 0, & \text{otherwise.} \end{cases}$$

Here t_i indicates the beginning time of an event special to this cell and δ_i is the response rising time; $f_{\max}$ is typically 100 spikes/sec, according to Fig. 4 in [4]. For simplicity, we can combine t_i and δ_i as $\hat{t}_i = t_i + \delta_i$.Thus the excitatory postsynaptic current (EPSC) of this cell is something like

$$I_i \approx \begin{cases} \left(1 - \frac{t-\hat{t}_i}{d_i}\right) I_{\max}, & \text{when } \hat{t}_i \le t \le \hat{t}_i + d_i; \\ 0, & \text{otherwise.} \end{cases}$$

Such currents with different decay times are then fed into different α cells in layer 4 of area W.

Remember that the dynamic function of a leaky integrate-and-fire neuron (here the α or β neuron) is

$$C\frac{dV}{dt} = I^{\text{in}} - \frac{V}{R}.$$

We know that the firing rate is a monotonic function of the input current(Fig. 6 in [4]), and thus the interspike interval time is also determined by the input current. With a high input current, the interspike interval is short and with a low input current, the interval is long. Here we can consider the input current to cell α_i is I_i.[1]

The threshold of a γ cell is set to be a little lower than the sum of EPSPs such that it will fire when and only when all the input cells fire at some same time t. That is, the γ cell will fire only if the interspike times of all the input cells are the same at time t.[2] Thus the γ cell acts as a coincidence detector.[3] At the time of 'coincidence', we have for all cells α_i connected to this γ cell,

$$1 - \frac{t - \hat{t}_i}{d_i} = 1 - \frac{1}{c},$$

where c is a constant ($c \ge 1$). Equivalently,

$$d_i - d_j = -c(\hat{t}_i - \hat{t}_j), \tag{1}$$

for α_i and α_j both are inputs to γ. Thus fixing c and one d_i (for example, set the d_i associated to the largest $\hat{t}_i$ to 0.3ms) we can set up all the other value of d_j's.

With this setting of d_i's, the γ cell can detect special patterns: not only the patterns that the events come at time $\hat{t}_i$, but all the patterns with time $(k\hat{t}_i + s)$, where k and s are some constants and $0 < k \le c$, because

$$d_i - d_j = -\frac{c}{k}\left[(k\hat{t}_i + s) - (k\hat{t}_j + s)\right].$$

So this kind of coincidence detect is invariant to the time translation and time compression/expansion to some extent. This is wonderful in recognizing voice.

However, this scheme can not explain why there is sustained firing of α cells to speech, compared to pure tone. And β cells have no proper function in this scheme.

[1]First, note that the input current I^{in} may not equal the EPSC I_i, since there may exist sub-threshold current. However, we can still consider that the firing rate is a function of I_i. Second, there should not be two or more inputs from area A to a single cell in layer 4 of area W. We can postulate this because in Fig. 5 in [4], the responce of an 'onset' cell has only one peak.

[2]I am still not sure about this point. In fact, I think this is wrong and is one factor why I gave up this kind of scheme. One of the other factors is that in this scheme only α cells are considered and no proper use of β cells is discussed. However, here I took it for granted in order to continue the analysis.

[3]The γ cell has a short membrane time constant of 6ms and is suitable to act as a conincidence detector.

Scheme 2: α-β pairs

Now we are going to take the intracellular connections into consideration.

The paper [4] says that the axons of both α and β cells arborize widely within area W and each makes a total of approximately 75–200 synaptic connections with other cells. Thus the intracellular connections do exits. However, from Fig. 5 in the paper, the response of an 'onset' cell in layer 4 of area W has only one (obvious) peak. Thus we can not expect strong connections between α and β cells. With the assumption that those connections are weak, I postulate that

- The α and β cells are paired in layer 4. The α cell and β cell in one pair receive the same input from area A and output to same cells in area W.

Thus the α cell and the β cell in one pair share the input with a single decay time. However, the EPSCs of α cells have an extremely fast rise time and also decay very fast, while the IPSCs of β cells have a slower rise and decay time (Fig. 6 in [4]). Thus the output of the pair, i.e., the sum of the EPSC of the α cell and IPSC of the β cell in the pair, has a little peak at the end of the decay time (Figure). If the decay time of the input cell in area A is d_i, this pair acts just as a delay line of length d_i.

The γ cell still acts as a coincidence detector. Now we postulate again that

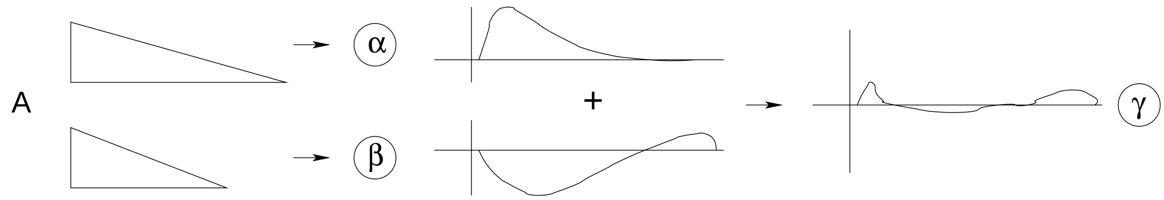
- For each event in the training voice, there's at least one cell A_i in area A and one α-β pair in area W corresponding to that event. The decay times of A_i's are determined by setting $c = 1$ in (1), in order to align all those peaks.

This scheme is illustrated in Figure 1. Obviously, this scheme is insensitive to the time translation of voice. However, it is not invariant to the time compression/expansion.

But we want to ask ourselves whether the invariance to the time compression/expansion important to the speech recognition? The answer is 'no.' In fact, we need the invariance to some sort of voice distortion, or time warp. For example, the relationship between two instances of "one" by different people are not simply time compression or expansion; Some events may happen earlier in one instance and others may happen later. (You may refer to the fourth figure in [3] which shows the spectrograms and feature times of two instances of "one".)

Then, how about the small distortion with the events times? Since the small peaks span a small length of time which is determined by the difference between the decay times of the EPSC and IPSC (about 20ms), the 'coincidence' detect is not so strict. If some peaks are not aligned in a line but still have much overlap, the total current to the γ cell may still be enough to have γ cell fire. This allows the recognition of voices with distortion to some extent. (You may also refer to the sixth figure in [3]. The left part illustrates that with some distortion the total activity is still above the threshold.)

One by-product here is that I can point out there may be one error in Fig. 3 of [4]. The connection from the second α cell to γ_0 should be changed to from the first α cell, since we hypothesis that the α cell and β cell in one pair output to same cells in area W.



(a) The output of an α-β pair.

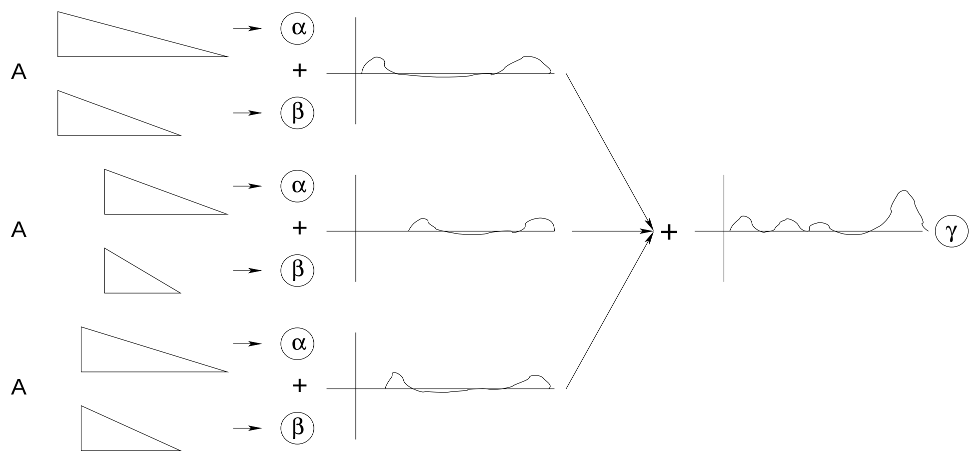


Figure 1: The scheme of α-β pairs. (a) With fast rise and decay time in EPSC and slower rise and decay time in IPSC, the α-β acts as a delay line. (b) With decay times set by (1), the γ cell can detect a certain temporal pattern.

Scheme 2: continued

There exists a problem with the above scheme 2: if a γ cell is to detect the coincidence of all the events, then it will produce only 1 spike. However, the organism actually produces more than 4 spikes to the word "one". So, we have to introduce another hypothesis about the *Mus silicium*:

- Events are detected in several steps instead of all at once. That is, the coincidence detect is applied to a small number of events, say, 5 consecutive events at one time. And several detects are carried out in order to get overall recognition of a word. One possible strategy with some overlapping is: numbering the events in the order of their times, and then aligning events with number $(3i-2)$–$(3i+2)$ to T_i, which may make the γ cell spike at time T_i, $i = 1..13$.

Thus, the γ cell will fire several times when the input pattern really resembles the training sample, and the number of spikes of γ stands for the similarity of the input pattern to the training sample. This is a good way to make the system more robust. Besides, this also answers one of the questions we raised in the class: why do γ neurons "drop" spikes? That is just because some detects fail due to noise or pattern distortion.

Now we go back to the intracellular connections. 75–200 connections per neuron is a big number according to the total number of neurons in layer 4. That means about 38–100 α-β pairs are connected to one pair. This number surpasses the total number of pairs for one word, which is about 40. So the intracelluar connections is related to the interaction between different training patterns, about which I am not knowledgeable.

Training of the network

We have already talked about how to train the network with only one sample. We can summarize here. After the events of 40 channels are produced by the front end, group them with some overlapping in the time order. For each group, set a time T_i as the align time and thus each decay time is set as $d_{ij} = T_i - \hat{t}_{ij}$. If there are overlaps between groups, cells in area A with different decay time but the same frequency may be used.

If more instances for one word are used to train the network, we can treat them as different words, then train the network, and take the result as 'OR' of the outputs of corresponding γ cells. However, if only one γ is supposed to be used, the training becomes much more complicated. We have to find out the relationships between events of different instances and find the most suitable representative from them.

The weights from layer 4 to layer 2+3 can stand for the relative importance of events. However, setting them equal will not degrade the performance of the network seriously.

Thus, all the information of the training pattern(s) is stored as the decay times of the neurons in area A and the connection pattern between area A and area W and within area W. That is, the structure of the network decides what words are to be recognized. So, changing the weights within area W by a random factor of $\pm 50\%$ will only affect the strength of intracellular connection and the relative importance between events, which will not have appreciable effect on the response of the α, β, or γ cells.

References

[1] J. Hopfield. Pattern recognition computation using action potential timing for stimulus representation. *Nature*, **376**:33–6, 1995.

[2] J. Hopfield. Transforming neural computations and representing time. *Proceedings of the National Academy of Sciences*, **93**:15440–4, 1996.

[3] J. Hopfield, C. Brody, and S. Roweis. Computing with action potentials. Advances in Neural Information Processing Systems, **10**:166–72, 1997.

[4] J. Hopfield and C. Brody. What is a moment? 'Cortical' sensory integration over a brief interval (I). `http://shadrach.cns.nyu.edu/~carlos/Organism/Docs/paper.pdf`, Oct. 7, 2000.