A recurrent working memory architecture for emergent speech representation

Mark Elshaw, Roger K. Moore

Abstract

This research considers a recurrent self-organising map (RSOM) working memory architecture for emergent speech representation, which is inspired by evidence from human neuroscience studies. The main purpose of this research is to demonstrate that a neural architecture can develop meaningful self-organised representations of speech using phone-like structures. By using this representational approach it should be possible, in a similar fashion to infants, to improve the performance of automatic recognition systems by aiding speech segmentation and fast word learning.

This RSOM architecture takes inspiration, at an abstract level, from evidence on word representation, the learning approach of the cerebral cortex and the working memory system's phonological loop. The neurocognitive evidence of Pulvermuller (2003) offers inspiration to the RSOM architecture related to how the brain represents words using spatiotemporal cell assembly firing patterns. The cell assembly representation of a word includes assemblies associated with its word form (speech signal characteristics) and others associated with the word's semantic features. Baddeley (1992) notes in his working memory model that the phonological loop is used for the storage and rehearsal of speech based knowledge.

To achieve recurrent temporal speech processing and representation in an unsupervised self-organised manner RSOM uses the extension by Voegtlin (2002) of the Kohonen self-organising map. The training and test inputs for the RSOM model are spoken words extracted from short utterances by a female speaker such as 'do you see the nappy'. At each time-slice the RSOM working memory receives as input the current speech signal slice (27ms) from a moving window and to act as context the activations from the RSOM at previous time-step. From this input a learned temporal topological representation of the speech is produced on the RSOM output layer at each time-step. By examining the sequences of RSOM best matching units (BMUs) for words, it is possible to find that there is a temporal representation of speech in terms of phone-like structures.

By the RSOM architecture developing a representation of words in terms of phones this matches the findings of researchers in cognitive child development on infant speech encoding. Infants have been found to use this phonetic representation approach to aid word extraction and the development of word understanding. The neurocognitive findings of Pulvermuller are recreated in the RSOM model with different BMUs (as abstract cell assemblies) being activate over time as a chain to create the word form representation. In terms of the working memory model of Baddeley the RSOM model recreates functionality of the phonological loop by producing a learned representation of the current speech input using stored weights. Further, by training using multiple observations of the same speech samples this equates to the phonological loop performing rehearsal of speech.

References

D. Baddeley, Working memory, Science, 255(5044) (1992), pp. 556-559.
F. Pulvermuller, The neuroscience of language: On brain circuits of words and language, Cambridge Press, Cambridge, UK, 2003.
T. Voegtlin, Recursive self-organizing maps, Neural Networks, 15(8-9) (2002), pp. 979-991.

Support Material

In the recurrent self-organising map (RSOM) neural architecture (Figure 1), the inputs to the model at each time-step are a 27ms speech slice from a moving window (bottom left) and the activations on the RSOM output layer for the previous time-step (bottom right). By using these inputs and their associated learned weights the RSOM architecture creates a topological temporal representation on the output layer for each speech input slice (top of Figure 1).

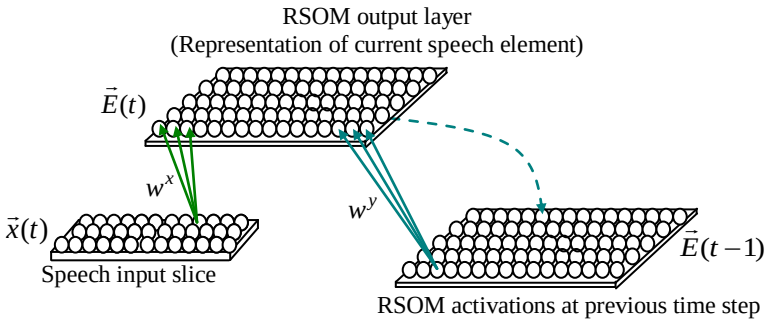

Figure 1 The RSOM memory structure for emergent temporal speech representation based on phones.

In the RSOM model a set of weights are trained so they are associated with the current speech input slice and another set that are associated with the RSOM activations at the previous time-step. The activations on the RSOM are determined using two different sets of Euclidean distance values. These Euclidean distance values $\vec{A}$ and $\vec{B}$ are based on the difference between the speech input slice $\vec{x}(t)$ and associated weights w^x and the activations for the previous time-step of the RSOM $\vec{E}(t-1)$ and their related weights w^y, respectively. $A_k = (||w_k^x - x(t)||)$ and $B_k = (||w_k^y - E(t-1)||)$, where k is the index of units of the RSOM output layer.

To determine the activations for the units of the RSOM $\vec{E}$, $\vec{A}$ and $\vec{B}$ are combined using $E_k = ((\alpha \cdot A_k) + (\beta \cdot B_k))$. The parameters α and β are used to control the impact on the activation values of the current input speech slice and the context. In a similar manner to the standard SOM the weights are trained according to: $\Delta w_{kj}^x = \gamma h_{kk'}(x(t)_j - w_{kj}^x)$ and $\Delta w_{kj}^y = \gamma h_{kk'}(E(t-1)_j - w_{kj}^y)$. Where the learning rate is γ, the neighbour function $h_{k,k'} = e^{(-d_{k,k'}^2 / 2\sigma^2)}$, k' is the BMU and j is index of input data.

By examining the sequences of BMUs created for words on the RSOM map (Figure 2), it is possible to find that the RSOM represents phone-like speech sounds using sub-sequences at specific locations on the map. For instance, the RSOM sub-sequence of BMUs for the speech slices making up the 'S' phone sound are associated with the top left corner of map for the example training session.

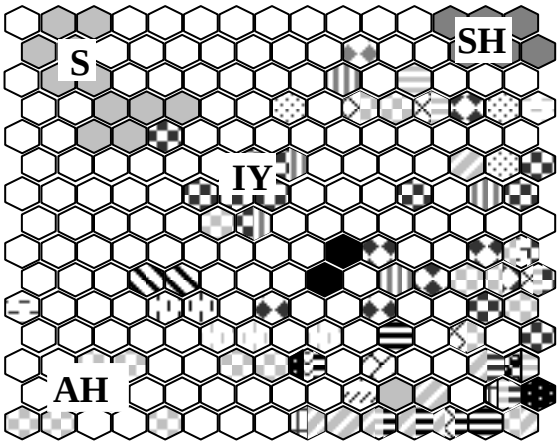

Figure 2 The distribution of BMUs sub-sequence, created by a RSOM from spoke words, associated with specific phone-like speech sounds for an example training session. On the RSOM output layer the phones are representation by different colour patterns with four example phone-region associations for the speech sounds 'S', 'SH', 'AH' and 'IY' shown. (The syntax of phone-like speech sounds equate to those in the DARPA phonetic alphabet).