Hierarchical recurrent self-organising memory (H-RSOM) architecture for an emergent speech representation towards robot grounding

Mark Elshaw[1], Roger K. Moore[1] and Michael Klein[2]

[1] Speech and Hearing Research Group, University of Sheffield, The Department of Computer Science, Regent Court, Portobello, Sheffield, S1 4DP, United Kingdom.

[2] Centre for Language and Speech Technology, Radboud University Nijmegen, 6525 HT Nijmegen, The Netherlands
m.elshaw@sheffield.ac.uk

Abstract

If social robots are to become widespread there is a need for them to offer spoken communication in a manner close to humans. This should be achieved by grounding the speech being recognised by the robot into the real world instead of trying to recognise the speech signal in isolation. This paper examines a hierarchical recurrent self-organising map (H-RSOM) memory model for emergent speech representation, which makes use of evidence from human cognitive processing. The H-RSOM memory architecture is constrained by the structure of the cerebral cortex and working memory, neurocognitive evidence on word representation and infant speech acquisition. By associating the speech signal representation with semantic (visual) features the H-RSOM offers a temporal-distribution speech representation based on phone like-speech sounds and semantic (visual) features. Using this representational approach it should be possible in the future, in a similar fashion to infants, to develop a robotic automatic speech recogniser that offers grounding of words in meaning.

Introduction

If humanoid robots are to be accepted by the general public they must be able to communicate with humans using spoken natural language. Hence, there is a need for speech recognition in an open environment. However, existing speech recognition systems for use in robot communication are overly restrictive, requiring their human users to follow a very strict procedure [den Os et al. 2005]. It is not sufficient to simply try to recognise auditory signals but rather there is a need to achieve understanding of the meaning of what is being spoken and how this relates to the real world [Cangelosi et al. 2007].

Harnad (2003) when considering the grounding problem states that an abstract representation such as spoken words must be grounded or associated to something in the real world to interpret their meaning. Hence, to actually attribute meaning to a spoken word there must be interactions with the world to provide relevance to the abstract representation and so not describe spoken words simply in terms of other spoken words [Roy 2003].

This paper describes a hierarchical recurrent self-organising map (H-RSOM) architecture for emergent temporal speech representation, which has been developed within the European ACORNS (Acquisition of Communication and Recognition Skills) project. This project has the overall goal of developing a novel speech recogniser that does not suffer from the shortcomings described above. The neural network based architecture introduced here takes inspiration and is directed by the growing body of knowledge about human cognitive processing and acquisition of speech in infants.

The capability of the H-RSOM model is demonstrated to move towards robot grounding by developing a representation of spoken words in a temporal emergent manner. To achieve this a learned self-organised neural approach is used. It associates speech signals and semantic (visual) features to provide meaning to the spoken word. Semantic features are used to approximate the visual inputs of an infant learner. Features can be used to represent *type distance* (the degree of similarity between different kinds of objects) as well as giving a realistic representation of the object or action. Semantic features further allow the simulation of phenomena like over-generalisation (calling a cat 'dog') that can be observed during language acquisition.

The cognitive inspiration of the H-RSOM architecture presented in this paper comes from neuroscience evidence on word representation, the structure and organisation of working memory and the cerebral cortex [Baddeley 1992, Doya 1999, Pulvermüller 1999, Werker, and Curtis 2005].

The remainder of this paper is as follows: the **second section** gives the biological inspiration and constraints of the H-RSOM architecture; the **third section** describes previous computational models that relate to this H-RSOM neural architecture; the actual details of H-RSOM that describes the architecture and experimental method used is in the **fourth section**; and the **fifth section** the results and discussion associated with this H-RSOM architecture for emergent speech representation for spoken word grounding for a robot.

Cognitive System Inspiration

The H-RSOM neural architecture takes inspiration, at an abstract level, from various cognitive systems such as neurocognitive evidence on word representation, structure and processing approaches of the cerebral cortex and the working memory system. In addition, the H-RSOM architecture is inspired by the temporal unsupervised self-organised learning found in the cerebral cortex [Doya 1999]. The H-RSOM architecture does not use all features of these cognitive systems but incorporates certain processing characteristics.

Central to the H-RSOM architecture that works towards robot grounding is the inspiration offered by the structure of the brain's working memory. Baddeley (1992) notes in his first model that working memory is split into three main subsystems: (i) the central executive that performs as an attention-control system (ii) the visuospatial sketchpad that manipulates visual inputs which in the H-RSOM architecture is the semantic feature self-organising map (SOM) component; and (iii) the phonological loop which is used for the storage of speech based knowledge and as such inspires the H-RSOM architecture for speech signal representation. Baddeley (2003) later included an episodic buffer in the working memory model that stores information from diverse modalities in the form of an episode and so inspires the associator RSOM component of the H-RSOM model.

Neurocognitive evidence of Pulvermüller [Hauk and Pulvermüller 2002, Pulvermüller 1999, Pulvermüller, 2003] offers inspiration to the H-RSOM architecture related to how the brain represents words using synfire chains. Synfire chains are formed from the spatiotemporal firing patterns of different associated cell assemblies. Synfire chains represent words as temporal sequence of activated cell assemblies (in the cerebral cortex). 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. The semantic factors that influences the cell assemblies that are activated come from various modalities and include the complexity of activity performed, facial expression, the type and number of muscles involved, the colour of the stimulus, movement involved and the tool used.

Related Computational Models

A computational model related to the H-RSOM memory model outlined in this paper is that of Wermter et al. (2003) who explore the use of semantic features in a SOM to achieve the association of language with actions. Although the architecture of Wermter et al. (2003) fails to use a recurrent approach and include speech signals, it does offer a hierarchical approach. The architecture contains a SOM to perform the coarse clustering that relates semantic action verb representations with the appropriate body part. At the next processing level of the Wermter at al. (2003) architecture, there is finer clustering through a SOM for each body part to identify the actual action verbs.

Motivated by the association of semantic features with language representations, Wermter et al. (2005) develop a language memory based approach. This model allows a robot to learn to perform three behaviours 'go', 'pick' and 'lift' based on multimodal inputs that act as semantic features. In this hierarchical memory based architecture there is the association of the motor and high-level vision inputs using the first hidden layer based on Helmholtz machine learning [Dayan et al. 1995]. The activations of the first hidden layer are then associated with the language instruction region input at the second hidden layer based on SOM learning.

Cangelosi et al. (2007) produced a model based on multi-layer perceptron (MLP) neural network to perform the association of language with actions. This is achieved through imitation learning, with the teacher robot performing a set of actions that are associated with linguistic names. From this the student robot learns the actions and then directly grounds the actions in these names. The student robot is able to gain higher-order behaviours by using these names and the learned actions. The MLP has input units associated with language, motor control and visual information and has output motor control and language. For the higher-level behaviours the teacher robot provides the name of two connected actions and the new higher-level name of this and from this the student robot learns the new higher behaviour. The MLP does not incorporate the self-organised nature of the cerebral cortex found in H-RSOM.

A related approach to the H-RSOM architecture that associates speech language with cognitive activities using Modeling field theory (MFT) is outlined by Perlovsky (2005). MFT associates lower-level signals (speech) with higher-level concept-models so that input signals can be seen in terms of real world concepts. In the MFT hetero-hierarchical system, the output is based on concepts produced from the input, with the input-concept association based on representational models and similarity measures. Association-recognition models are updated so they represent better the input and similarity measures are changed to match the uncertainty levels. Overtime the uncertainty reduces and the models become more accurate representation of the input and similarity less fuzzy. Hence, links are produced between speech signal elements such as words and model-concepts to produce grounding between speech and the real world.

Roy (2001) and Roy and Pentland (2002) developed CELL to associate speech with objects using a robot equipped with a camera and microphone. The robot associates symbolic representation of language (utterance) with

semantic features of the utterance (visual representation). CELL represents/memorises 3-dimensional objects depicted by histograms of local characteristics from 2-dimensional representations. In Roy (2001) and Roy and Pentland (2002) CELL approach short-term memory is used to store utterance-shape pairs and produce hypotheses about the association by extracting part of the utterance and linking it with the observed shape. This association is placed in long-term memory which is consolidated over various observations. The focus of interaction between short-term and long-term memory is also found in the H-RSOM with the RSOM activations the working memory and the stored weights long-term memory.

In the approach by Roy (2001) and Roy and Pentland (2002) spoken utterances are depicted as sets of phone probabilities, which are obtained from a spectral depiction using the Relative Spectral-Perceptual Linear Prediction (RASTA-PLP) algorithm. A recurrent neural network examines RASTA-PLP coefficients to predict phone and speech/silence probabilities. The H-RSOM architecture however does offer an approach that is closer to neuroscience evidence on the cortex in that it is unsupervised in nature and develops a representation of speech in an emergent manner instead of using a predefined phone structure.

A further model associated with the H-RSOM model is the working memory approach of O'Reilly and Frank (2006). Despite the H-RSOM memory model described in this paper performing emergent speech representation and the working memory model of O'Reilly and Frank (2006) learning the 1-2-AX task they do share certain common concepts. For instance, they both rely on the learning of sequences of inputs by storing context information. The O'Reilly and Frank (2006) computational working memory architecture models the prefrontal cortex and basal ganglia interaction and how a working memory model is able to perform sequential activities to achieve action selection. This model depends on maintaining depictions in the prefrontal cortex which are gated via the basal ganglia.

Grossberg (2003) developed a related model to the H-RSOM architecture which is also based on interaction between working memory (WM) and long-term memory (LTM). In this approach, by using Active Resonance Theory (ART), speech perception occurs in a self-organised manner using a model based on resonance in the brain. Resonance states occur in the model when bottom-up speech signals interact with top-down beliefs using a matching process. In this model the speech signal activates traces in WM to produce bottom-up patterns. These bottom-up patterns activate list categories in WM through interaction with traces found in LTM. The list categories produce top-down belief signals from LTM which are matched against the active units of WM. Through this matching procedure the WM activations are chosen based on the LTM traces. While the H-RSOM only depends on bottom-up learning the ART model feeds down beliefs at the top-level as well as relying on bottom-up speech signals.

Hierarchical Recurrent Self-Organising Map (H-RSOM) Architecture

The H-RSOM architecture works towards language grounding through the associating of speech signals and semantic (visual) features using an emergent speech representation. In the H-RSOM architecture (Figure 1) the speech signal RSOM (lower right box) is trained using as input the speech slice for the current time-step (27ms) and activations from the speech signal RSOM at the previous time-step. A standard SOM is also trained using semantic features of the words to produce a neural representation of these semantic features (lower left box of Figure 1).

Once training is completed for these two lower components of the H-RSOM, the associator RSOM (top centre Figure 1) is used to associate the speech signal and the semantic (visual) feature representations. Through this association mechanism there is an emergent speech representation that offers the possibility to ground words for a robot application. This is achieved by introducing the activation values associated with each speech time slice for the speech signal RSOM units, with the activations for associator RSOM for the previous time-step and the semantic feature SOM unit activations for the appropriate word.

Standard Self-Organising Map (SOM)

The H-RSOM makes use of standard SOM [Kohonen 1997, Haykin 1994] for the semantic feature component and as basis for the RSOM components. As seen in Figure 2 the non-recurrent standard SOM consists of an input $\vec{x}$ and an output layer $\vec{y}$. Typically all inputs are connected via weights w to all the units in the output layer [Spitzer 1999]. Learning in SOMs is performed by updating the weights between the input layer and the output layer via a form of Hebbian learning.

There are various ways to calculate the activation on the SOM output layer. The approach used in the H-RSOM architecture described in this paper is Euclidean distance. The output representation $\vec{y}$ is determined based on the Euclidean distance between the unit weights and input: $y_k = ||w_k - x||$. Where k is the index of units in the SOM output layer. The best matching unit (BMU) is the one with the smallest Euclidean distance $y_{k'}$. The weights of the SOM are update:

$$\Delta w_{kj} = \gamma h_{kk'} \cdot (x_j - w_{kj}) \quad (1)$$

where γ is the learning rate and j is the index of input units. $h_{kk'}$ is the neighbourhood function. The

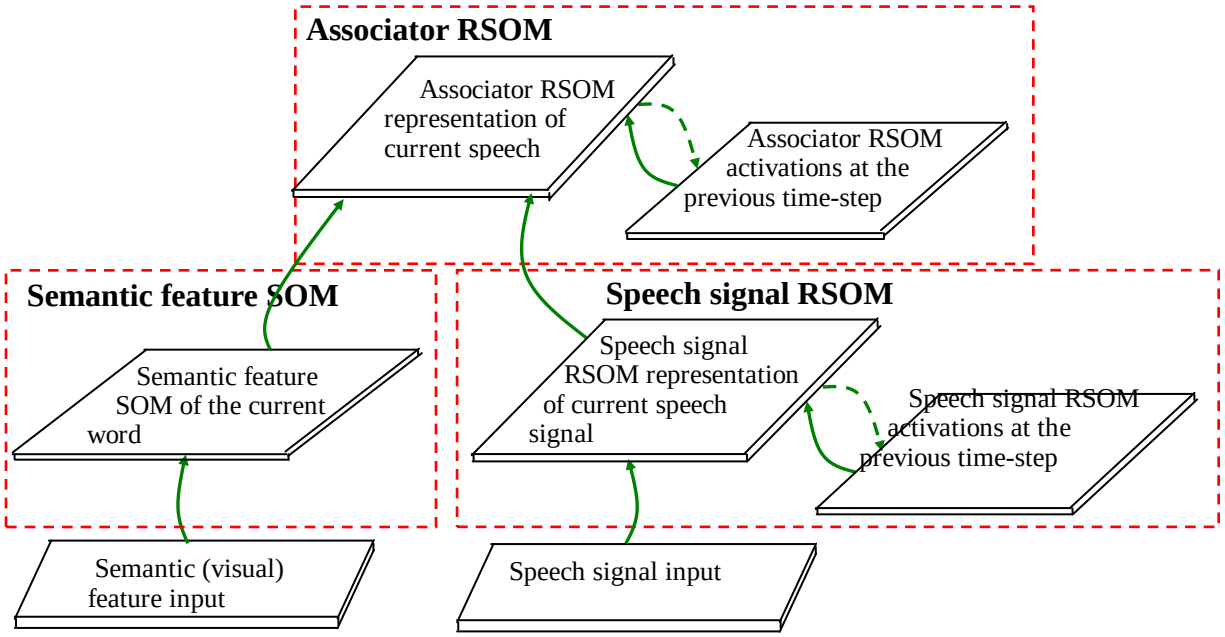

Figure 1 Representation of H-RSOM architecture for emergent speech representation.

neighbourhood function in the H-RSOM architecture in this paper is a Gaussian and is created using Equation (2). Where $d_{k,k'}$ is the distance between unit k and the winning unit k' on the SOM output layer. σ sets the size of the Gaussian with the larger the value the broader the neighbourhood function. The number of units in the neighbourhood usually drops gradually over time.

$$h_{kk'} = e^{(-d_{k,k'}^2 / 2\sigma^2)} \quad (2)$$

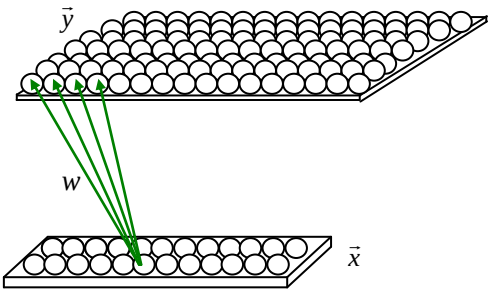

Figure 2 A depiction of a standard SOM, with input $\vec{x}$, weights w and output $\vec{y}$.

A problem associated with the standard SOM is that it is not possible to incorporate temporal information into the model in a straightforward manner. This, however, is critical for emergent speech representation. As a response to this the H-RSOM makes use of the extension of the SOM described by Voegtlin (2002) to incorporate recurrent temporal processing.

Input to H-RSOM for emergent speech representation

The training set of the speech signal RSOM component of the H-RSOM architecture is made up of the individual words extracted from 50 utterances repeated 5 times by a female from the ACORNS English database speaker. These individual words used for training are extracted has they appeared in the utterances from this female speaker. The test data is the same words extracted from 5 new recordings by this female speaker of the 50 training utterances. For the speech signal RSOM network, 703 words are used for training and the same number for testing the network, within which there are 42 distinct words. Hence, the same words are extracted from utterances with difference carrier words.

In the speech signal RSOM the 703 training spoken words are split into speech time slices, using a moving window, which are introduced into the network sequentially. This moving window is 27ms in size that moves along the sample 13.5ms at a time, with an overlap of 50%. The speech slice inputs in the form of logarithmic mel-spectrum values [Holmes and Holmes 2001, Shah, et al. 2004] are introduced with the activations of the speech signal RSOM for the previous time-step.

The semantic feature SOM is trained and tested using semantic (visual) feature inputs for 17 nouns and 12 verbs found in the words used to train the speech signal RSOM. Semantic features are only produced for content words. Function words (such as ‘what’, ‘did’, ‘finally’, ‘today’,

'matches' and 'the') that are used to train the speech signal RSOM are not considered. The nouns considered are 'car', 'water', 'house', 'door', 'couch', 'bath', 'nappy', 'shoe', 'bottle', 'taps', 'book', 'newspaper', 'daddy', 'baby', 'telephone', 'Ewan' and 'mummy'. The verbs are 'like', 'back', 'sits', 'seen', 'coming', 'going', 'join', 'read', 'calls', 'driving', 'changed' and 'take'.

Table 1 Semantic (visual) features for verb inputs.

Semantic Features	Responses
Body Movement	Small/Medium/Large
Interaction with object	Small/Medium/Large
Interaction with agent	Small/Medium/Large
Task Complexity	Small/Medium/Large
Emotion related	Extent (0-1)
Precise of activity	Extent (0-1)
Communication	Extent (0-1)
Change to object	Small/Medium/Large
Cognitive complexity	Small/Medium/Large
Instigated activity	Extent (0-1)

Table 2 Semantic (visual) features for noun inputs.

Semantic Features	Responses
Worn	Extent (0-1)
Food related	Extent (0-1)
Furniture	Extent (0-1)
Inanimate	Extent (0-1)
Communication device	Extent (0-1)
Gender	Male/Female/Neuter
Used by	Child/Adult/Non
Creates noise	Extent (0-1)
Breakable	Fragile/Durable/Strong
Tool	Extent (0-1)
read	Extent (0-1)
animate	Extent (0-1)
man made	Extent (0-1)
Provides information	Extent (0-1)
Texture	Smooth/rough/liquid
technology	Small/Medium/Large
Location	Indoor/Outdoor
Size	Small/Medium/Large

The 17 nouns and 12 verbs semantic feature inputs for the semantic feature SOM are introduced as a single representation per word. The semantic feature input is based on an approach similar to McClelland and Kawamoto (1986) and uses various semantic features. For verbs the features include 'body movement', 'interaction with object', 'interaction with agent', 'task complexity' and 'emotion related'. The noun semantic features include whether the noun is 'worn', 'food related', 'furniture', 'inanimate', 'human', 'communication device', 'gender' and 'creates noise'. Table 1 gives the full set of semantic features for verbs and Table 2 the full set for nouns. Some of the semantic values are given an extent value between 0 and 1. For those features that have multiple possible options such as texture each of the three options have a value that adds up to 1. In this approach the values for the semantic (visual) features are subjective however in the future they can be replaced with more objective values based on measurement such as of colour and movement etc.

The associator RSOM is trained and tested using the activations produced by the speech signal RSOM for the speech time slices and the semantic feature SOM for the set of 17 nouns and 12 verbs. The speech samples for the 17 nouns and 12 verbs are extracted as they appear in the same utterances used for training and testing the speech signal RSOM. This produced 407 word recordings for a training epoch and the same number for testing the associator RSOM.

Training of the associator RSOM (upper centre Figure 1) is performed using 407 words using inputs made up of (i) the activations for each speech time slice produced by the trained speech signal RSOM; (ii) the activation of the associator RSOM at the previous time-step; and (iii) the activations of trained semantic SOM for the appropriate word sample. The speech signal RSOM and the semantic feature SOM are trained separately and once this is complete these networks are used to train the associator RSOM. For each training session to update the weights for the speech signal RSOM, the semantic feature SOM and the associator RSOM the training samples making up an epoch are introduced in random order. The number of epochs used to train these components is fixed at the start of session.

RSOM components for emergent speech representation

Figure 3 gives a more detailed representation of the speech signal RSOM (lower right Figure 1). In this model, the 27ms speech slices making up the word representations are introduced sequentially with the activations for the speech signal RSOM from the previous time-step. In the model a set of weights is trained so it is associated with the current speech input slice and another set of weights trained so it is associated with the speech signal RSOM activations at the previous time-step.

Using a similar structure to the speech signal RSOM in Figure 3 the associator RSOM is trained to produce a representation of speech that associates the speech signal with semantic feature towards robot grounding. However, for the associator RSOM this is achieved with three inputs at each time-step rather than two: (i) the activations of the speech signal RSOM for each speech time slice; (ii) the activation for the associator RSOM at the previous time-step; and (iii) the activations for the appropriate word

from the semantic feature SOM which are presented at the same time as each speech time slice.

The activation values of each unit in the speech signal RSOM are determined using two different Euclidean distance values. These two Euclidean distance values for each unit are based on the difference between the speech input slice and the activation for the previous time slice of the speech signal RSOM and their related weights. The three sets of Euclidean distance values for the associator RSOM units are based on the difference between weights and their related inputs: (i) the activation for the speech signal RSOM; (ii) the activations for the semantic feature SOM; and (iii) the activation for the previous time slice of the associator RSOM. Two parameters for the speech signal RSOM and three parameters for the associator RSOM are used to influence the impact of the Euclidean distance values on unit activation values for these two RSOM components.

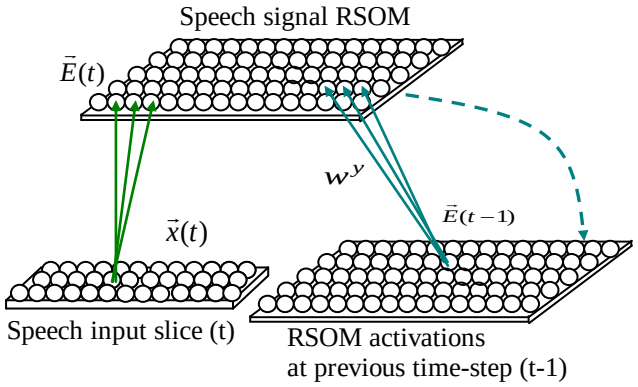

Figure 3 The speech signal RSOM component of the H-RSOM memory structure for emergent speech representation.

The first set of Euclidean distance values for the speech signal RSOM are based on the difference between the input speech slice $\vec{x}(t)$ and weights w^x.

$$A_k = (||x(t) - w_k^x||) \quad (3)$$

The second set of Euclidean distance values $\vec{B}$ for the speech signal RSOM are determined using the difference between the activations of the speech signal RSOM at the previous time-step $\vec{E}(t-1)$ and the associated weights w^y.

$$B_k = (||E(t-1) - w_k^y||) \quad (4)$$

To determine the activation for the units in the speech signal RSOM $\vec{E}$ and hence from this the BMU, A and B are combined (Equation 5). The parameters α and β are used to control the impact of the current input speech slice and the context memory has on the activations for the units, with them set to 0.75 and 2, respectively.

$$E_k = ((\alpha \cdot A_k) + (\beta \cdot B_k)) \quad (5)$$

As with the standard SOM the weights are trained according to:

$$\Delta w_{kj}^x = \gamma h_{kk'} (x(t)_j - w_{kj}^x) \quad (6)$$

$$\Delta w_{kj}^y = \gamma h_{kk'} (E(t-1)_j - w_{kj}^y) \quad (7)$$

The associator RSOM model is trained in a similar way to the speech signal RSOM but includes three inputs rather than two. For the associator RSOM three parameters are used to control impact of three input Euclidean distance values on the activations on the associator RSOM in a similar manner to Equation (5). The three parameters χ (speech signal activations), δ (associator RSOM activations at previous time-step), ε (semantic feature SOM activations) are set to 2.75, 0.75 and 2.0, respectively.

For all the components of the architecture the learning rate γ is set at 0.01. For the neighbourhood function determined using Equation (2) σ starts at 6 and decreasing to 0.5 over the first 2/3 of the epochs and then remains at 0.5 for the 1/3 of the epochs. In the H-RSOM by an empirical study the speech signal RSOM grid is set at 15 units along the x-axis and 15 units along the y-axis, the semantic feature SOM grid is 12 by 12 units and the associator RSOM grid is 18 by 18 units. The number of epochs that the speech signal RSOM, semantic feature SOM and associator RSOM are trained for 1000, 400 and 800 epochs, respectively.

The location of sub-sequences of BMUs on the speech signal RSOM and associator RSOM are examined to establish if the former architecture component creates representations associated with specific speech sounds and the latter component is associated with specific speech sounds and words semantic feature representations.

H-RSOM Architecture Results

Given the nature of the SOM algorithm for each training session a different set of weights and hence BMUs are produced for the inputs to the components of the H-RSOM. However, it is found that the H-RSOM does create meaningful representations across training sessions. This is achieved by the H-RSOM relating of specific speech signal time slices, semantic (visual) features and associations between speech and semantic features with distinct BMU regions. However, these BMU regions are likely to be different for each training session. To assist understanding of these relations, the BMUs regions created for a specific training session are examined below but these should only be seen as an indication of the form the represents take.

The speech signal RSOM is tested using the same words as used for training but recorded at a different time. By examining the sequences of BMUs created for the test

words it is possible to find that the speech signal RSOM represents phone-like speech sounds using BMU sub-sequences. In the rest of this paper the syntax of phone-like speech sounds is equal to those in the DARPA phonetic alphabet. Certain phone speech sounds are represented by sub-sequences of BMUs in a single region of a map. For example, the speech time slices making up the 'SH' sound found in the word 'shoe' are associated with a specific region of the map. For the training session considered here this is the top right hand area of the speech signal RSOM output layer. However, in other cases, phone-like speech sounds are represented by combining sub-sequences of BMUs from different regions of the RSOM.

For the speech signal RSOM the model creates distinct regional associations, based on sub-sequences of BMUs from words, with speech sound simular to phones. For the example training session considered here the top left hand area represents the speech slices making up the sound 'S' at the end or start of words. In the training and test set these include such words as 'matches', 'taps', 'news', 'seen', and 'comes', as well as the 'S' sound inside words such as 'newspaper', 'closer' and 'house'. A region of the map is also associated with 'SH', 'CH', 'JH' and 'K' sounds, which in this example is the dark grey units in top right of the RSOM. These are sounds found in words such as 'fashion', 'shoe', 'shy', 'matches', 'couch', 'join' and 'back'. The sound 'AH' is represented by a sub-sequence of BMUs located in a distinct region of the map, which in this case is the lower left corner of the example network associated with words from the database such as 'telephone' (T EH L AH F OW N), 'Ewan' (Y UW AH N) and 'what' (W AH T).

The speech signal RSOM also creates a region of the map whose units are part of the BMU sub-sequence representations of the 'A', 'I' and 'OW' phone sounds. For the example training session this is at bottom right of the map and are those units that predominately have more than one of these phone sounds associated with them. These phone sounds are particularly 'AE' from words such as 'daddy' (D AE D IY), 'navigating' (N AE V AH G EY T IH NG), and 'bath' (B AE TH), 'AY' found in words such as 'like' (L AY K), 'driving' (D R AY V IH NG), 'finally' (F AY N AH L IY) and 'shy' (SH AY). While 'OW' is associated with words such as 'telephone' (T EH L AH F OW N) and 'broken' (B R OW K AH N) in the training and test set. This is possibly due to these speech sounds at a lower level that the phone level sharing certain sound similarities.

When considering the BMUs for the semantic (visual) feature SOM in most cases the words are located in their own unit on the map. This does not occur in the case of 'comes' and 'going' which share the same BMU as they are very similar words in terms of the activities involved. Furthermore, similar words are also located in close regions of the semantic (visual) feature SOM. For instance, for the example training session higher level cognitive function such as 'like' and 'see' are located in the top left of the SOM output layer. The words associated with humans such as 'daddy', 'mummy', 'Ewan' and 'baby' are found in the lower left hand corner. Also communication approaches 'newspaper' and 'book' are located close together has are the action verbs 'coming' and 'join'.

When combining the activations for the speech signal RSOM and the semantic SOM using the associator RSOM (top centre Figure 1), it is possible to identify units on the associator RSOM (Figure 4) that are associated with specific speech sounds (from speech signal RSOM) for particular words (from semantic features SOM). In addition, in certain cases regions of the associator RSOM are related with specific sounds (from speech signal RSOM) for different words (from semantic (visual) features SOM), and that speech sounds for semantically related words are located in similar regions of the RSOM.

The associator RSOM ouput layer (Figure 4) shows the location of BMU sub-sequences related to phone-like speech sounds (from speech signal RSOM activations) for particular words (from semantic feature SOM activations). The words (from semantic feature SOM) associated with a unit are represented by the colour pattern of the unit and for the speech sounds (from speech signal RSOM activations) the DARPA phonetic alphabet characters on the unit. It is possible to identify that various units of the network are associated with specific phone speech signal sounds and words. For instance, for the example training sesson the 'SH' speech sound for the word 'shoe' (SH UW) is related with the unit 1 on x-axis and unit 1 on the y-axis. This is the light grey unit with a 'SH' on.

When considering the sub-sequence of BMUs for the associator RSOM it is found that the 'S' speech sound for the words 'taps' (T AE P S) and 'house' (HH AW S) are located close together on the map. For the example training session these are found at unit 1 on the x-axis and unit 3 on the y-axis and unit 1 on the x-axis and unit 5 on the y-axis, respectively. It is also possible to identify further regions of the map that are associated with specific speech sounds such as the 'K' sound. The units with a spheres pattern with 'K' on at units 1 on x-axis and 6 and 7 on y-axis are associated with the 'K' sound for 'car' (K AA R) and the units with a 'K' on with diagonal lines up and down pattern which are units 1 on x-axis and 8 and 9 y-axis is the 'K' sound for 'couch' (K AW CH).

The 'T' sound for the words 'telephone' (T EH L AH F OW N) and 'taps' (T AE P S) can also be seen on the associator RSOM to be location in their own individual units but close together on the map for the example training session. These units for the speech sound 'T' for taps' are represented on the map as units that are dark grey with a 'T' on and the units for the sound 'T' for 'telephone' are those that have black dots with a white background with a 'T' on them.

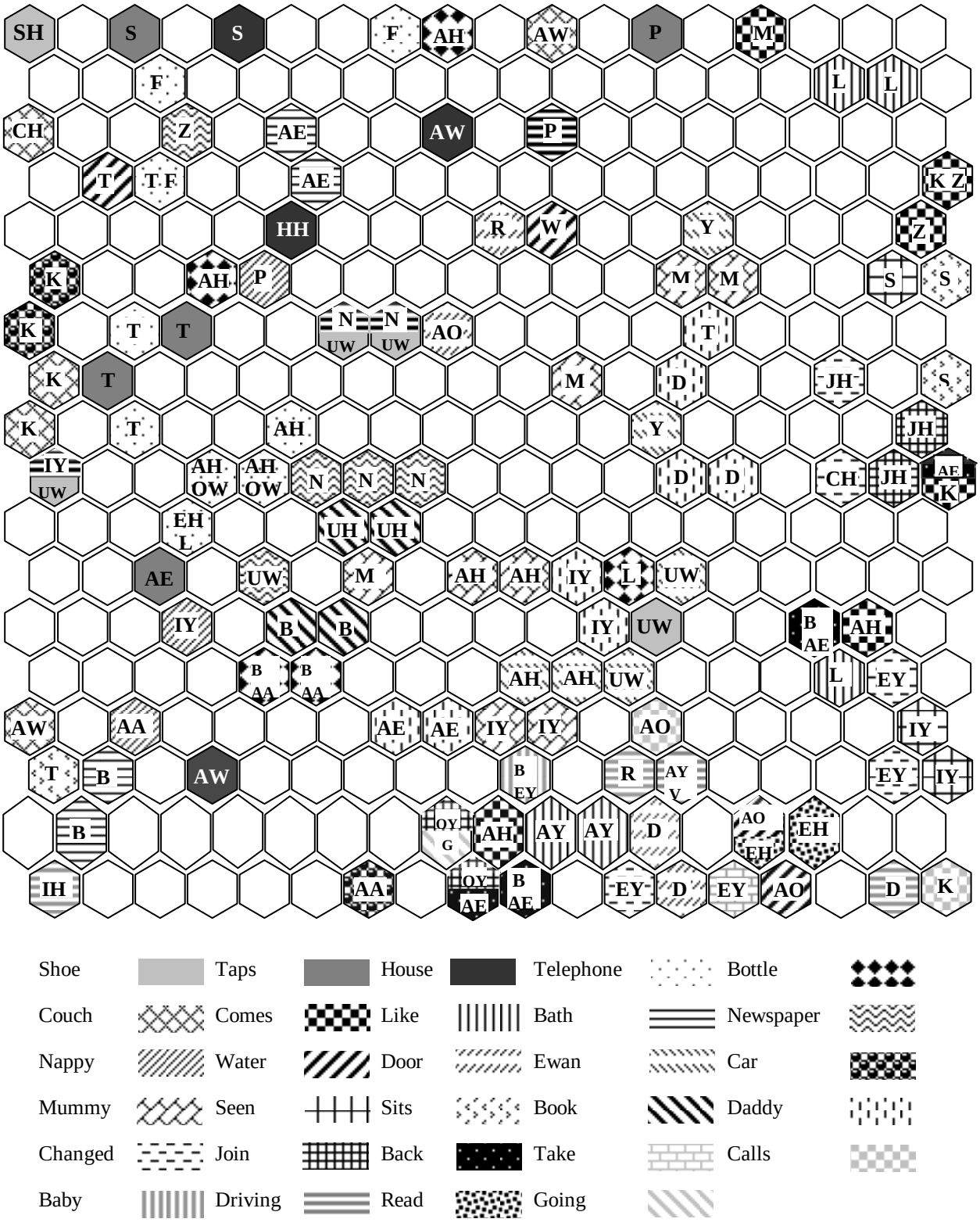

Figure 4 BMU regions of associator RSOM output layer associated with specific phone-like speech sounds and semantic features for words.

It is also possible to see that speech sounds (from speech signal RSOM activations) for semantically related words (from semantic feature SOM activations) are located in near units on the associator RSOM (Figure 4). For instance, for the example training session the sounds 'M' in 'mummy' (M AH M IY), 'D' (D AE D IY) in 'daddy' and 'Y' for 'Ewan' (Y UW AH N) and hence family-human related words are located around x-unit 8 and y-unit 13. This is also the case for family-human related words for sounds such as 'AH' and 'IY' for the word 'mummy' (M AH M IY), 'IY' and 'AE' for the word 'daddy' (D AE D IY) in 'daddy' and 'Y' for 'Ewan' (Y UW AH N) and and 'B EY' speech sound for the word 'baby' (B EY B IY).

This is also the case for communication approach such as 'telephone', 'newspaper' and 'book', with speech sounds associated with these words located close together on the associator RSOM. The phone-like speech sounds that are included in a specific word can be seen to be distributed among different units despite the semantic (visual) features input being the same for full length of word. This indicates that the speech signal and semantic feature representations are combined in such a way that the two sets of activations have an impact on the final representation created by the associator RSOM.

Discussion and Future Work

The H-RSOM model successfully combines self-organising recurrent approaches at different levels of processing that achieves the association of speech signals with semantic (visual) features for word grounding for robot speech recognition. It is anticipated that by making use of such a self-organising temporal representation that speech recognition and human-robot communication would benefit from the neuroscientific inspiration that is incorporated. The H-RSOM architecture produces a representation which is based on associating semantic (visual) features with phone-like speech sounds that could act as building blocks to be combined to produce word meaning and the recognition of words.

In terms of the working memory model of Baddeley (1992) H-RSOM recreates functionality of the phonological loop by producing representations of the current speech signal. The semantic SOM representation of words recreates part of the functioning of the visuospatial sketchpad in the working memory as it gives a representation of visual inputs. The final speech representation of the associator RSOM recreates some of the functionality of the episodic buffer, in an abstract manner, by combining of the visual semantic features and the speech signal. This representational approach based on phone-like speech sounds and semantic features is seen only as the first step towards word grounding. However, scientists working on child development have shown that this type of emergent representation is fundamental for child speech development, understanding and word learning and so a robotic speech recognition system could use this representation in the same way [Kuhl 1993, Werker and Yeung 2005, Kuhl 2004].

The ability of the speech signal RSOM and associator RSOM to develop a representation that discriminates based on phone-like sounds is despite in certain cases, such as between 'P' and 'L', the difference between phone sounds being very small [Kuhl 1993]. Since the H-RSOM model develops a representation of words in terms of phones, it matches the findings of researchers in cognitive child development on infant speech encoding [Kuhl 1993, Werker and Yeung 2005]. It is noted by Kuhl (2004) that infants use and recognise phonetic characteristics of speech and the retention of such speech sounds are critical for the extraction and the development of words.

The H-RSOM architecture also matches the neurocognitive model of Pulvermüller in that different units of the H-RSOM (as abstract cell assemblies) are combined over time in the representation of a word to produce a chain of active units. The H-RSOM representation through active units can be seen to combine the word form (speech signal) and the semantic (visual) features. This was identified by Pulvermüller to give the richer brain representation of words, which offers an approach to achieve word grounding in robots.

Possible future work for this H-RSOM architecture would incorporation of an episodic long-term memory instance based approach such as MINERVA2 [Neath and Surprenant 2002]. This episodic memory could be used to perform word recognition for the robot system based on the emergent representation developed by the H-RSOM model. MINERVA2 is a computational multiple-trace episodic memory model and could be used to produce an overall representation of the sequences of BMUs produced for the words. This would be achieved by comparing the sequence of BMUs for the unknown word with stored sequences of example words to produce an overall word representation based on a distance measure. The H-RSOM can, by means of these names and the learned actions, also be used to create representations for the other input stream. For instance the speech signal RSOM activations might be used in the re-activation of the semantic feature meaning associated with spoke auditory input and vice-versa.

Acknowledgement

This work is part of the ACORNS (Acquisition of Communication and Recognition Skills) project supported by the EU in the FET-IST programme.

References

Baddeley, A. D. 1992 Working Memory. Science 255(5044): 556-559.

Baddeley, A. D. 2002 The Psychology of Memory. In: A. D. Baddeley, B. A. Wilson and M. Kopelman (Eds.) Handbook of Memory Disorders, 2nd Edition. Hove: Psychology Press, 3-15.

Dayan, P., Hinton, G. E., Neal, R., and Zemel, R. S. (1995) The Helmholtz Machine. Neural Computation 7: 1022-1037.

Cangelosi, A., Tikhanoff, V., and Fontanari, J . F. (2007) Integrating Language and Cognition: A Cognitive Robotics Approach. IEEE Computational Intelligence Magazine, August 2007.

den Os, E., Boves, L., Rossignol, S., ten Bosch, L. and Vuurpijl, L. 2005 Conversational Agent or Direct Manipulation in Human-System Interaction. Speech Communication 47(1-2): 194-207.

Doya, K. 1999 What are the Computations of the Cerebellum, the Basal Ganglia, and the Cerebral cortex. Neural Networks 12: 961-974.

Grossberg, S. 2003 Resonant Neural Dynamics of Speech Perception. Technical Report CAS/CNS-TR-02-008.

Harnad, S. 2003 The Symbol Grounding Problem. Encyclopedia of Cognitive Science, London: Macmillan.

Hauk O. and Pulvermüller, F. 2002 Neurophysiological Distinction of Action Words in the Fronto-Central Cortex. Human Brain Mapping 21: 191–201

Haykin, S. 1994 Neural Networks: A Comprehensive Foundation. Macmillian, Toronto: Canada College Publishing Company.

Holmes, J. and Holmes, W. 2001 Speech Synthesis and Recognition. , London: Taylor and Francis.

Kohonen, T. 1997 Self-organizing Maps, Heidelberg, Germany: Springer-Verlag.

Kuhl, P. 2004 Early Language Acquisition: Cracking the Speech Code. Nature Reviews Neuroscience 5(11): 831-843.

Kuhl, P. 1993 Early Linguistic Experience and Phonetic Perception: Implications for Theories of Developmental Speech Perception. Journal Phonetics 21: 125-139.

McClelland J. L. and Kawamoto, A. H. 1996 Mechanisms of Sentence Processing: Assigning Roles to Constituents. In J. L. McClelland & D. E. Rumelhart (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2. Cambridge, MA: MIT Press.

Neath, I and Surprenant A. 2002 Human Memory: An Introduction to Research Data, and Theory. Belmont, California: Wadsworth Pub Co.

O' Reilly, R. and Frank M. 2006 Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. Neural Computation, 18(2): 283-328.

Perlovsky, L.I. 2005 Evolving Agents: Communication and Cognition. Autonomous Intelligent Systems: Agents and Data Mining, (Eds) V. Gorodetsky, J. Liu, V.A. Skormin. Springer-Verlag.

Pulvermüller, F. 1999 Words in the Brain's Language. Cognitive Neuroscience, 22(2): 253-336.

Pulvermüller, F. 2003 The Neuroscience of Language: On Brain Circuits of Words and Serial Order, Cambridge, UK: Cambridge University Press.

Roy, D. 2002 Learning Words and Syntax for a Visual Description Task. Computer Speech and Language, 16(3).

Roy, D. 2003 Grounded spoken language acquisition: Experiments in word learning. IEEE Transactions on Multimedia, 5(2), pp. 197- 209.

Roy, D. and Pentland, A. 2002 Learning Words from Sights and Sounds: A Computational Model. Cognitive Science, 26(1): 113-146.

Shah, J., Iyer, A., Smolenski, B. and Yantorno, R. 2004 Robust Voiced/Unvoiced Classification using Novel Features and Gaussian Mixture Model. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 17-21, Montreal, Canada,

Spitzer, M. 1999 The Mind Within the Net: Models of Learning, Thinking and Acting. Cambridge, MA, USA: MIT Press.

Voegtlin, T. 2002 Recursive Self-organizing Maps. Neural Networks, 15(8-9):979-991.

Werker, J. and Yeung, H. 2005 Infant Speech Perception Bootstraps Word Learning. TRENDS in Cognitive Sciences 9(11): 519-527.

Werker, J. F. and Curtis, S. 2005 PRIMIR: A Developmental Framework for of Infant Speech Processing. Language Learning and Development, 1, 197-234.

Wermter, S., Elshaw, M. and Farrand, S. 2003 A Modular Approach to Self-organisation of Robot Control Based on Language Instruction. Connection Science 15(2-3): 73-94.

Wermter, S., Weber, C., Elshaw, M., Gallese, V. and Pulvermüller, F. 2005 Grounding Neural Robot Language in Action. In Wermter S., Palm G. and Elshaw M. Biomimetic Neural Learning for Intelligent Robots, Heidelberg, Germany: Springer, 162-181.