UNSUPERVISED PHONEME ACQUISITION USING HIERARCHICAL TEMPORAL MODELS

BY

LYDIA LEE MAJURE

B.S., University of Illinois at Urbana-Champaign, 2006

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2009

Urbana, Illinois

Adviser:

Professor Stephen Levinson

ABSTRACT

Unresolved issues in speech processing have prompted new approaches to artificial cognition. This paper outlines autonomous mental development, one such research paradigm. An unsupervised hierarchical temporal model is also demonstrated to learn categories of phonemes from untranscribed speech. This successful demonstration of self-organizing structure from real world data suggests the suitability of this class of models for creating integrated sensorimotor associative memories.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

There are several unresolved issues in speech processing that point towards incompleteness in the approach that has been traditionally used. The missing piece is semantics, which is rooted in embodied experience of the world. This paper outlines a new paradigm in artificial intelligence research — autonomous mental development — and relates it to current issues in speech processing. It also describes hierarchical temporal models (HTM), an associative memory algorithm based on Bayesian belief propagation. A network for unsupervised phoneme acquisition was developed using HTM, which demonstrated learned structure by showing response patterns specific to certain test phonemes, as well as forming stable output patterns at the top node in response to speech. This study forms the basis for future work in large scale associative memories for use in language acquisition and robotics research.

CHAPTER 2

BACKGROUND

2.1 Limits of Traditional Speech Processing

Speech processing has advanced tremendously since its inception. Although the mathematical foundations were laid sixty years ago by information theory, we owe the current state of the technology to increased processor speeds and the development of efficient algorithms. Despite these recent advances of large vocabulary automatic speech recognition (ASR), some major problems in speech remain unsolved.

2.1.1 Robustness

State of the art ASR systems are still highly vulnerable to noise and speaker variation compared to human performance. Even when information loss due to feature extraction is accounted for, human performance is significantly higher in noisy conditions and across different dialects [1]. Spectral differences in speakers' voices, such as between men and women or adults and children, also degrade the performance of ASR. While humans have auditory attentional mechanisms that allow listening to one voice in the presence of others, known as the cocktail party effect, this too is currently out of reach for speech recognition [2].

2.1.2 Supervision

Another issue with ASR is its requirement for transcribed training data. Humans, of course, have no need for this when learning language. Studies on human infants have shown that very young infants detect all phonetic contrasts equally, but between six months and one year phoneme detection specifies for their native language [3]. Research in the area of unsupervised spoken language acquisition has been limited. Self-organizing maps have been shown to learn some phoneme classifications [4]. One study demonstrated unsupervised vowel learning by integrating visual input of the speaker's mouth with the audio signal [5]. This method, while limited, demonstrates an important concept: supervision for language learning can be provided by sensory integration.

2.1.3 Semantics

Semantics is commonly defined as the meaning of words. Cognitively, it is the associations between language and other sensorimotor information. The words "up" and "down" have meaning because one knows what it feels like to be pulled in a certain direction by gravity. Semantics is rooted in embodied experience of the world. To ignore it in speech processing is to ignore the very purpose of language. Humans develop a mental model of the world, and use language to convey their own model and internal state to others. In addition, semantics provides a powerful error correcting mechanism via linguistic and sensorimotor context. Human recognition performance has been

shown to degrade when sentences are semantically incorrect [6]. This indicates that the noisiness of the speech channel is compensated by accounting for word meaning. Automatic semantic analysis has been successfully demonstrated on limited domain problems, such as telephone airline reservation systems [7]. However, the techniques used in these applications cannot be scaled to large vocabulary speech understanding. To solve the problem of semantics in speech processing, it is necessary to learn the meaning of words by interaction with the environment.

2.2 Autonomous Mental Development

As the previously mentioned issues suggest, to study language is to study cognition. Autonomous mental development is characterized by the use of an embodied agent which learns by exploring its environment [8]. Often in such experiments, a human teacher indicates the words for objects or demonstrates tasks. This differs from supervised learning in the traditional sense, in that the agent has a "closed" mind which cannot be directly supplied with the correct answer. Learning takes place by building an internal model of the stimuli and how they are associated. Essentially, this approach mimics the learning behavior of human children. This concept is a relatively old one, originating with Turing in his seminal paper on intelligent computing [9]. In the intervening time, there has been interest in the idea, but it is only recently that the hardware needed to implement it has become available.

2.2.1 Characteristics of adaptive intelligent systems

While the paradigm of autonomous mental development does not imply replicating human brains at every level, certain characteristics are likely to be important in any intelligent system, as they form the core of human cognitive abilities.

Online learning Offline learning requires that all possible stimuli be presented to the agent during training, which leads to inflexible behavior and poor adaptability. However, if online learning can be performed, the agent can continue to expand its mental model as it encounters new stimuli.

Generality The entire human cortex is thought to function the same way computationally, regardless of the task to which a region is dedicated [10]. This indicates that a successful algorithm should be versatile, capable of different sensory modalities and levels of processing.

Sensory integration Since creating a world model relies on integration of information from different senses, any mental architecture should allow processing of multiple sources of information, both internal and external.

Attention selection Attention selection is an important survival trait for all animals. It allows the brain to lighten its computational load by filtering out irrelevant information and focusing on urgent or novel inputs. This also aids learning, as an agent can focus on the new stimuli.

Proprioception Proprioception is the agent's mental model of its own physical state. This includes feedback from effectors and

internal representations of mental processes. Effector feedback permits refinement of motor control and an internal representation of the body. Internal models of the agent's mind may be necessary for planning and decision making. Proprioception might also be essential to learning by imitation [11].

Object invariance A key feature of the mammalian brain is the ability to recognize objects despite large degrees of distortion and noise. Without this ability, an agent cannot function under widely varying perceptual conditions, as is required for operation in a natural environment.

Prediction Prediction enables the detection of anomalous events, which can direct attention or learning. It also allows an agent to plan, by mentally testing different actions and deciding on their likely results.

These characteristics are overlapping, and to date no system other than the human brain has demonstrated all of them. However, they serve to guide machine intelligence research and are criteria by which to evaluate learning algorithms.

2.2.2 Studies in cognitive robotics

Robotics research is increasingly influenced by the idea of autonomous, embodied agents. Some research has focused on replicating basic human cognitive functions such as attention selection, motor mapping, and conditioning [12, 13]. Many studies have used neurally inspired architectures to integrate language and sensorimotor information in mobile robots [14, 15]. Mirror neurons

in particular have attracted attention as a basis for motor learning by imitation [11, 16]. However, despite the supposed biological inspiration for neural networks, in practice they have little to do with neural processing, so there is no reason to believe that they are a superior choice of model. Hierarchical hidden Markov models have been successfully used to form associative memories which autonomously learn language [17, 18]. This work was extended to a robot learning to use noun-verb sentences to describe its actions [19]. An advantage of these models is both their suitability for temporal sequence processing and their hierarchical structure which integrates information across multiple modalities. However, they exhibit weaknesses in spatial pattern recognition. The approach through all these studies is that the robot should have minimal preconceptions about language and its environment. Another commonality is that all treat language as grounded in experience. Cognition is founded on the association between sensorimotor information, so robotics is essential to language acquisition research.

2.3 Hierarchical Temporal Models

Hierarchical temporal models (HTM) is a biologically inspired machine learning algorithm. It is implemented by the Numenta Platform for Intelligent Computing (NuPIC) [20]. The mammalian neocortex is organized into units known as cortical columns in which the neurons are highly interconnected and functionally associated [21]. Each column forms a computational node which learns spatial and temporal patterns in its inputs. Studies on the

cortex have shown that regions are connected in a processing hierarchy, where lower levels recognize simpler features [22]. There are large numbers of both feed-forward and feed-back connection, indicating that prediction is an important component of cortical processing. From these physiological facts and other studies, it seems that Bayesian belief propagation is a good computational model for the processing which occurs in this hierarchy of cortical circuits [23].

2.3.1 Operation of nodes

The fundamental computational unit of HTM is the node. Each node learns both spatial patterns in its input and sequences among those patterns. In NuPIC, learning and inference are separate stages of operation, both of which are outlined below for the case of first order Markov chains. Further detail on the algorithms can be found in [24].

Learning

The first phase of learning in a node consists of the memorization of input patterns from its children. Sufficiently distinct spatial input patterns, henceforth known as coincidences, are stored in memory until the maximum number is reached for that node. During this process, the node maintains a Markov graph of transition probabilities between coincidences. After the coincidences and their transitions are learned, temporal groups are formed by partitioning the graph using agglomerative hierarchical clustering. The similarity metric used for clustering is the transition probability between coincidences. If higher order

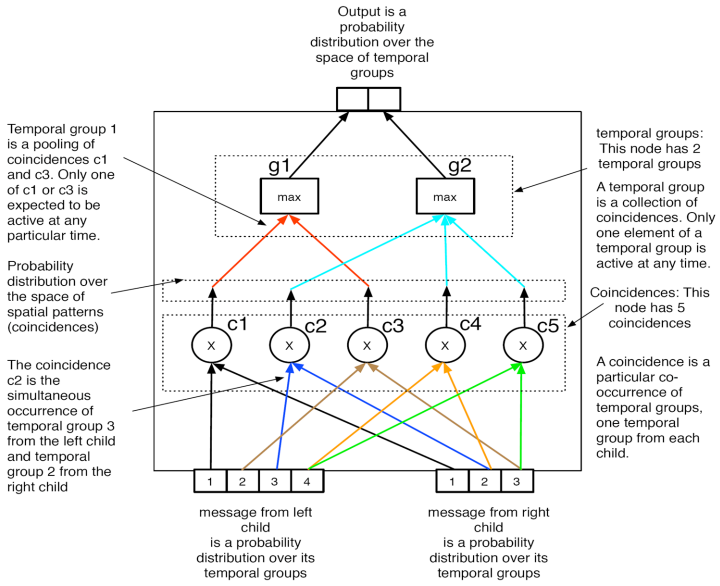


Figure 2.1: Structure of one node in HTM, from [24]

temporal models are used, multiple passes over the data are made, splitting nodes which have a large number of inbound transitions. The resulting probabilities $P(C_t^k|C_{t-1}^k, G_{t-1}^k)$, where C_t^k and G_t^k are the coincidence and group active in node k at time t, are used for inference.

Inference

At each time step, the node receives bottom-up and top-down messages from child and parent nodes, respectively. These are used to calculate its output messages. The bottom-up message is a vector with one component corresponding to each temporal group. Figure 2.1 shows an example of the structure of one node. In the following equations, ${}^{-}e$ represents the time series of messages from below, and ${}^{+}e$ represents the messages from above. The r^{th} component is the probability of the evidence seen at the child node

given the r^{th} group is active at time t:

$$\lambda(g_r) = P(^{-}e_0^t|g_r(t)) \propto \sum_{c_i(t)\in C^k} \alpha_t(c_i, g_r) \tag{2.1}$$

where c_i is the i^{th} coincidence, C^k is the set of coincidences in node k, and $\alpha_t(c_i, g_r)$ is a dynamic programming variable initialized and updated with:

$$\alpha_0(c_i, g_r) = P(^{-}e_0|c_i(t=0))P(c_i(t=0)|g_r) \tag{2.2}$$

$$\alpha_t(c_i, g_r) = P(^{-}e_t|c_i(t)) \sum_{c_j(t-1)\in C^k} P(c_i(t)|c_j(t-1), g_r)\alpha_{t-1}(c_j, g_r) \tag{2.3}$$

The top-down message sent to a child node is a vector with components corresponding to the temporal groups of that child. The message sent to child m is:

$$\pi^m(g_r^m) = \sum_{c_i\in C^k} I(c_i)Bel_t(c_i) \tag{2.4}$$

$I(c_i)$ is an indicator function which is 1 if g_r^m is a component of coincidence c_i. $Bel_t(c_i)$ is the belief in coincidence i at time t, and is calculated as:

$$Bel_t(c_i) = P(c_i(t)|^{-}e_0^t,^{+}e_0) \propto \sum_{g_r\in G^k} P(g_r|^{+}e_0)\beta_t(c_i, g_r) \tag{2.5}$$

The dynamic programming variable β is initialized and updated with:

$$\beta_0(c_i, g_r) = P(^{-}e_0|c_i(t=0))P(c_i(t=0)|g_r,^{+}e_0) \tag{2.6}$$

$$\beta_t(c_i, g_r) = P(^{-}e_t|c_i(t)) \sum_{c_j(t-1)\in C^k} P(c_i(t)|c_j(t-1), g_r)\beta_{t-1}(c_j, g_r) \tag{2.7}$$

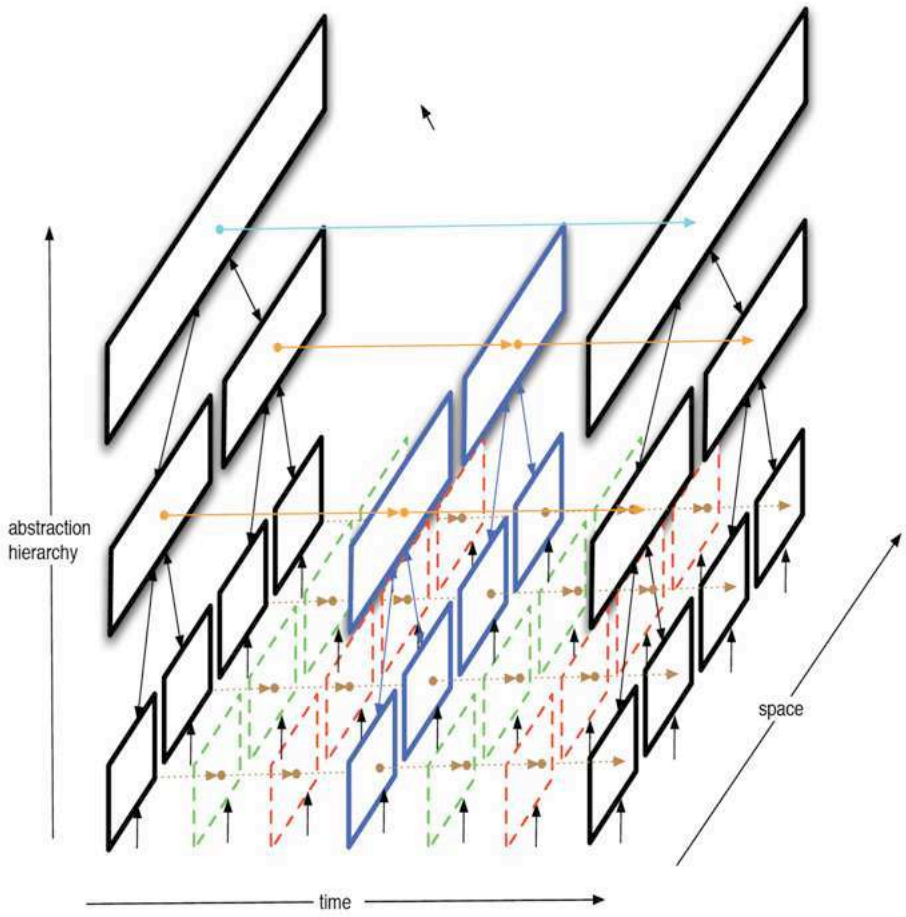


Figure 2.2: Spatial and temporal relationships among nodes, from [24]

Each level of the hierarchy is trained separately, with lower levels placed in inference mode and higher levels turned off. In effect, each node learns an invariant representation of its inputs. The bottom nodes see a segment of the network input, and learn low-level features in the data. In a visual system, for example, the bottom nodes might learn lines of various orientations. Higher levels learn larger and more complex spatial features, and more slowly varying temporal features. Figure 2.2 shows the spatial and temporal relationships among nodes in a hierarchy.

2.3.2 Experimental results

Early demonstrations of NuPIC learned categories for line drawings of objects. The recognition demonstrated significant robustness when presented with scale, translation, and noise distortions [24]. It classified 57% of the test images correctly, compared to 28% using a nearest neighbor classifier on the same problem. A more recent vision demonstration used photographs of four categories of objects from all perspectives. Accuracy was high for novel images, even with significant occlusion. A particularly impressive result of this experiment was recognition of hand-drawn sketches of objects from the categories [20]. An independent study showed success using HTM for content-based image retrieval [25]. Although vision has been the focus of most HTM research to date, some studies in speech processing using HTM have also been done. Experiments have demonstrated gender classification and speaker identification [20]. One study demonstrated a spoken digit recognition system [26]. Although this technology has shown promise, there is much left to investigate with regards to its performance in various tasks.

CHAPTER 3

PHONEME ACQUISITION

3.1 Motivation

This experiment addresses automatic acquisition of phoneme models. As mentioned previously, self-organizing maps have been demonstrated to learn some phoneme classifications, but they lack the ability to process temporal information. What is desired is a model which learns phonemes by differentiating them according to both their spectral and temporal patterns. Although context and semantics are very important to language, this study presumes that some low-level structure can be learned without additional sensorimotor information. This study also explores the suitability of HTM to processing speech data for general domain language learning tasks. The hierarchical feature learning in HTM makes it a good candidate for building general purpose associative memories that integrate the entire sensorimotor system. Demonstrated success in domains where most of the information is temporal, such as speech, would be evidence in favor of trying HTM on this problem.

3.2 Speech Data

It was desired that the network learn from spontaneous speech, so unprompted stories in English were used from a multilingual telephone corpus [27]. For testing, syllables from an articulation index corpus were used so the network's response to particular phonemes could be measured [28]. This data set consists of all possible CV and VC syllables. It also contains a sampling of CVC, VCC, and CCV syllables chosen randomly according to their frequency in the English language. All CV and VC syllables, which were used for evaluation, have recordings from 20 speakers. Both data sets were recorded at 8000 Hz. To generate a spectral representation of the signal, the speech waveforms were converted to Mel-frequency cepstral coefficients (MFCC) [29]. For training, about 10 minutes of speech were used.

3.3 Structure and Parameters of Network

3.3.1 Architecture

A three-layer network was used. Each node in the network consists of one spatial pooler followed by one temporal pooler. Together these implement one unit as described in HTM theory above. At the first level, 13 nodes looked at two components each of the feature vector with a zero appended at the beginning. This input structure was chosen because the first MFCC feature takes on significantly larger values than the others. The zero causes the first node to see only the first coefficient. The second layer consists of 6 nodes, each receiving messages from 3 first-level nodes. The top

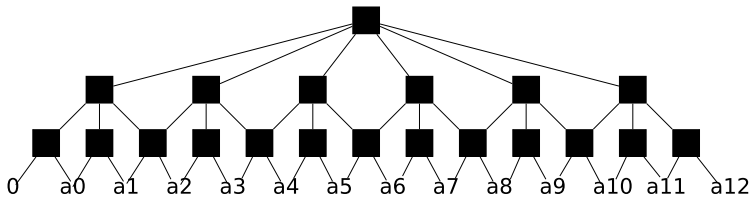


Figure 3.1: Architecture of HTM for phoneme acquisition: Inputs are 13 MFCC features and a constant zero

layer is a single node which sees all of the second-layer nodes. Figure 3.1 shows a simplified schematic of this network architecture.

3.3.2 Learning and inference parameters

There are several parameters that affect the behavior of HTM. The number of coincidences and temporal groups stored by the nodes were chosen to balance processing limitations and descriptiveness of the model. The chosen values were 50 coincidences and 25 temporal groups per node. maxDistance is the maximum Euclidean distance between two coincidences for them to be stored separately during learning, and sigma is the variance of the Gaussian radial basis function used to classify coincidences during inference. Both of these were chosen to account sufficiently for noise without overgrouping. Higher levels of the network were given higher values for these parameters because of their larger receptive fields. sequencerModelComplexity, sequencerWindowCount, and transitionMemory are temporal parameters which relate to the order of temporal learning. The model complexity, a float between 0 and 1, controls the connectivity threshold for splitting nodes in the Markov graph. In general, data with more information in the temporal domain

Table 3.1: NuPIC parameters used for training and testing phoneme acquisition network

	Level 1	Level 2	Top
Number of coincidences	50	50	50
Number of groups	20	25	25
maxDistance	0.25	1	2
sigma	0.5	1	1.5
sequencerModelComplexity	0.75	0.75	0.75
sequencerWindowCount	4	4	4
transitionMemory	4	5	6
largeGroupPenalty	4	6	8

requires a higher model complexity, so 0.75 was chosen for all levels. The window count corresponds to the number of passes made over the data when forming the temporal model. This was set to 4 for all levels, as this was the maximum value for this network to have practical run times. The transition memory is the number of time steps learned in the Markov transition model. Since they should be learning more slowly varying features, higher values were selected for higher-level nodes. Another parameter that affects the behavior of temporal learning is largeGroupPenalty, which controls the number of coincidences that can be put in a single group. Table 3.1 summarizes the parameters used. Another consideration when training the network was ensuring that the lowest-level spatial poolers learn coincidences representative of the data set. It was found that training these nodes on the original data caused the spatial poolers to learn the maximum number of coincidences very quickly, due to high noise. Since this causes poor generalization, the lowest spatial poolers were trained on a randomized version of the input features. When these were placed into inference mode to train the temporal poolers above and the rest of the network, the

in-order training data was used.

3.4 Evaluation

The network was evaluated using the CV and VC syllables in the Articulation Index corpus. The syllables were indexed according to their first phoneme, so examples of the beginning phoneme followed by every other phonotactically plausible phoneme were represented. The network performed inference on each recording, and the top-level group output was converted into a symbolic string by assigning a winning group to each output time step. From the network output for the test syllables, it was obvious that two of the groups corresponded to silence, as they occurred together in long runs at the beginning and end of samples. Using this observation, the symbols corresponding to silence were stripped from the output strings. Group frequencies were collected for each phoneme, gathering statistics over all syllables beginning with that phoneme. In this way, it was intended that the presence of the second phoneme would be smoothed out in the frequency counts, giving a distinct peak at the groups corresponding to the phoneme of interest. The responses of groups to given phonemes were calculated using these frequency counts. Table 3.2 shows the phonemes used for testing and their DARPAbet representations used in the results.

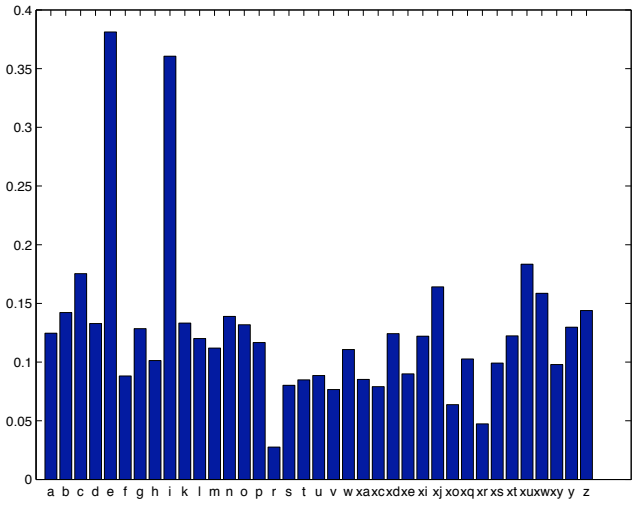
Table 3.2: DARPAbet

DARPAbet	Example	DARPAbet	Example
a	cOt	v	Vow
b	Bee	w	Win
c	cAUght	xa	hUt
d	Dog	xc	CHurch
e	bAIt	xd	THis
f	Fish	xe	bEt
g	doG	xi	hIt
h	He	xj	Judge
i	bEE	xo	bOY
k	Cat	xq	hAt
l	Look	xr	bIRd
m	Man	xs	SHe
n	maN	xt	THink
o	bOAt	xu	pUt
p	Pen	xw	hOW
r	Real	xy	whY
s	See	y	You
t	caT	z	Zoo
u	bOO		

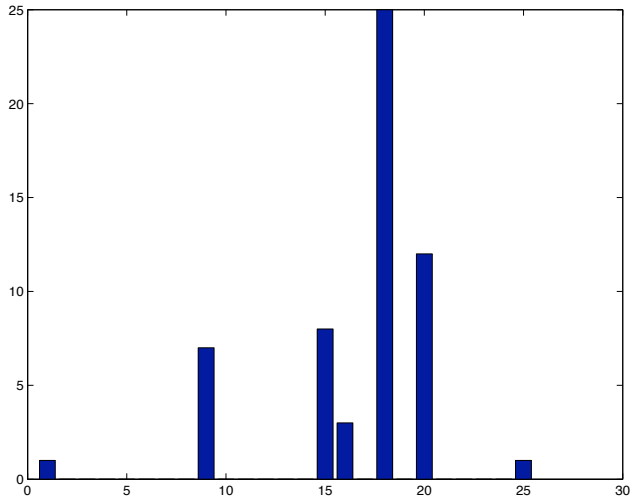
CHAPTER 4

RESULTS

The network clearly developed a distinction between silence and speech, as evidenced by the long strings of identical outputs at the beginning and end of each syllable sample. Another indication that structure is being learned is that the network outputs long runs of the same group, showing that it is recognizing a feature of the speech. However, the most compelling evidence that phonemes are being acquired at some level are the differing responses of groups to phonemes. Although the maximum number of top-level groups was 25, only 14 of the output groups responded to speech signals. This is likely due to an upper limit on the complexity of groups formed given the architecture of the network. Appendix A shows the phoneme responses to these groups. Although none of these groups respond to only one phoneme, there are some that respond much more strongly to a subset of phonemes. For example, group 18 was found to respond to /e/ and /i/ preferentially, as shown in Figure 4.1(a). Figure 4.1(b) shows a histogram of the network response to the syllable /be/, with a strong peak at group 18. Figures A.1 – A.5 in the appendix show the phoneme responses for all 14 groups. In general the most specified responses occured to vowels. Group 4, shown in Figure A.1(a), responded strongly to /a/, /c/, and /xu/, which are all back vowels. Group 10, Figure



(a) Response of group 18



(b) Network response to /be/

Figure 4.1: Comparison of group response statistics and network response to a syllable

A.2(b), responded very specifically to /xr/. Group 12, Figure A.3(a), responded to /o/ and /w/. Group 21, Figure A.5(a), responded to open front vowels. The network is leaning something meaningful about phonetics, although coarsely.

4.1 Discussion

It is clear from this experiment that HTM can acquire temporal structure from a signal without supervision. Since conventional speech recognition has had success using MFCC features, the choice is probably not the most important consideration in improving this network's performance. However, these results suggest that refinements to the network could produce more consistent responses to phonemes. Adding delta coefficients to the feature vector might improve grouping. Due to low resolution of the phoneme acquisition, a larger network or larger group memories might be needed. Another issue is the lack of connection to a complete sensory system. A possible reason for the overemphasis on silence in the network model could be this lack of integration. Humans know how to treat silence in speech because experience tells them that silence correlates with no information being conveyed. Having no context, this network has no way to know that silence is not an information bearing pattern. The coarseness of the correspondence between groups and phonemes also relates to phonemic awareness. In humans, consciously naming and distinguishing sounds occurs only upon exposure to written language. It is a cognitive byproduct of literacy, and children who have not yet begun reading cannot do it. If a layer

were added to this network that associated phonemic labels with the group outputs, it would likely not need much training data compared to conventional speech recognition, since a somewhat invariant structure of the sounds has already been learned. In general, it seems that unsupervised low-level structure learning would make any subsequent supervised classification learning require less training data and time. The performance of the phoneme acquisition would also improve if higher-level linguistic information was incorporated to provide top-down error correction.

CHAPTER 5

CONCLUSION

5.1 Application to Autonomous Mental Development

This study shows that phonetic information can be learned from exposure to untranscribed speech. Even within the limited domain of speech recognition, this could have implications for simplifying the task and reducing the amount of training data. More interesting is that hierarchical temporal memory provides one possible model for associative memory. This study shows that it can extract spatial and temporal structure from real world data. There are some weaknesses in the model that would need to be addressed before implementing an associative memory using HTM in an autonomous robot. First, for cognitive robotics it is essential that learning and inference can occur simultaneously. Another issue is that a node in the network cannot recognize multiple patterns simultaneously, something which becomes very important when networks are applied to realistic data. This is accomplished in the brain by using very sparse representations so that the responses to patterns do not overlap, and multiple objects can be recognized. A similar method could be implemented in the spatial poolers of HTM. Interaction with a noisy, complicated environment also requires attention selection. This is easily

implemented within the framework of HTM, as attention is the top-down influence to respond to a certain subset of incoming signals. The model implemented by NuPIC used in this study will probably prove to be a member of a family of hierarchical associative memory models. Further study is needed on how HTM compares to similar associative memories and how to modify it to increase its descriptive power.

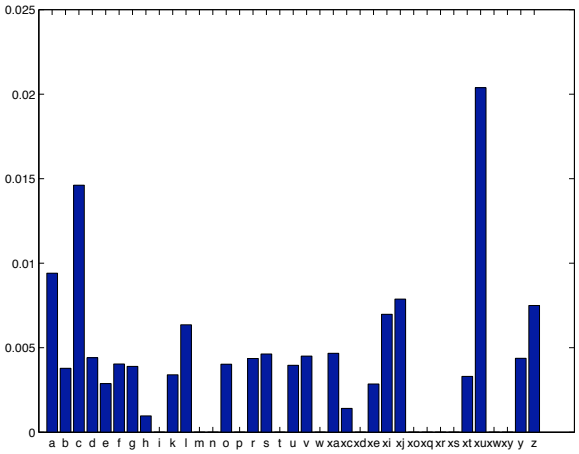
5.2 Questions in Cognitive Robotics

As indicated by the lack of understanding of how HTM relates to other models, there is no unified mathematical description of the dynamic systems that cause cognition. It is likely that these systems could be implemented in a number of ways, both biologically and artificially. Implementing an associative memory that allows a robot to interact and learn language in a complex environment requires addressing all of the characteristics of adaptive intelligent systems that were mentioned previously. Most importantly, sensory patterns, language, and motor skills must all be learned by the same algorithm. This would provide sensorimotor integration, allowing robots to build a rich internal model of the world and illuminating how this occurs in the human brain.

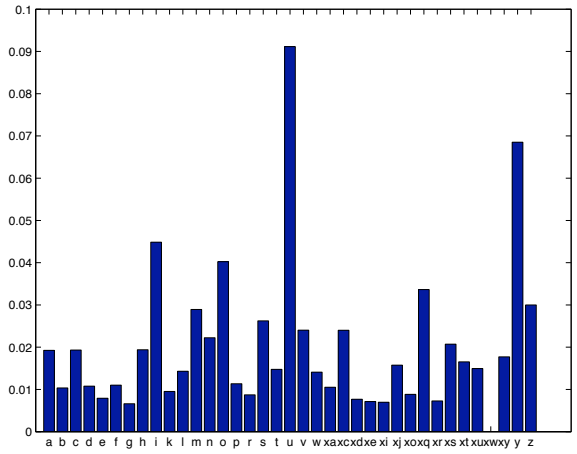
APPENDIX A

GROUP RESPONSES

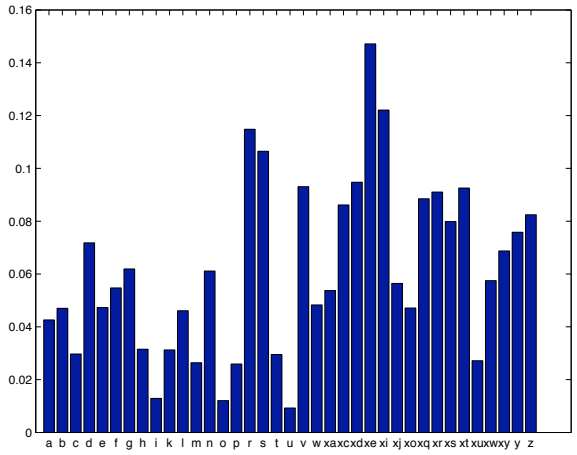
Figures A.1 – A.5 show the responses of the top-level temporal groups to different phonemes.



(a) Group 4

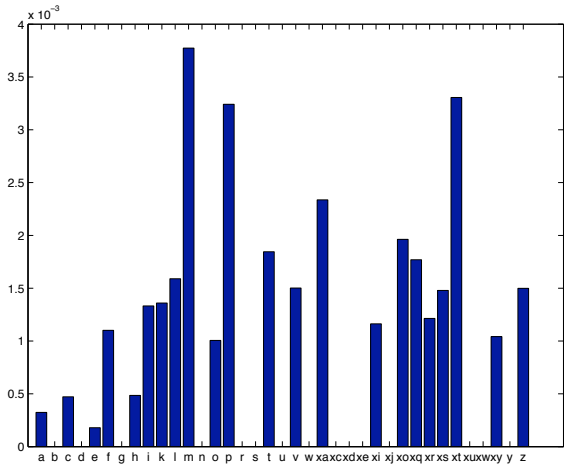


(b) Group 5

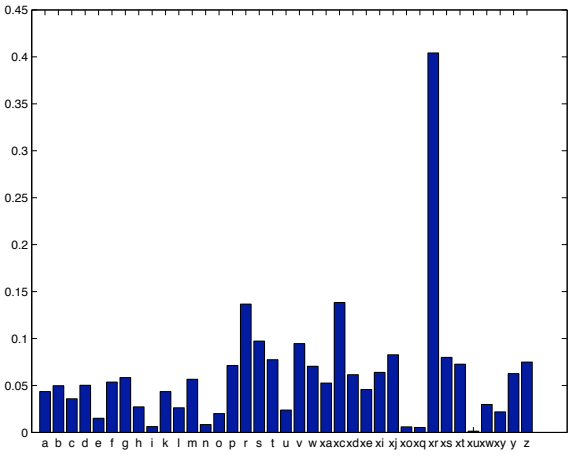


(c) Group 8

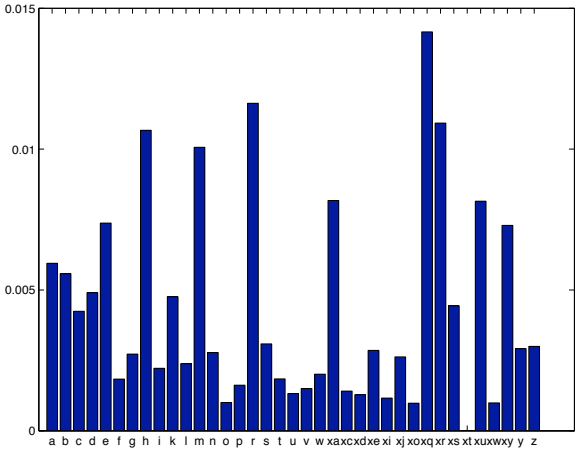
Figure A.1: Responses of groups 4, 5, 8



(a) Group 9

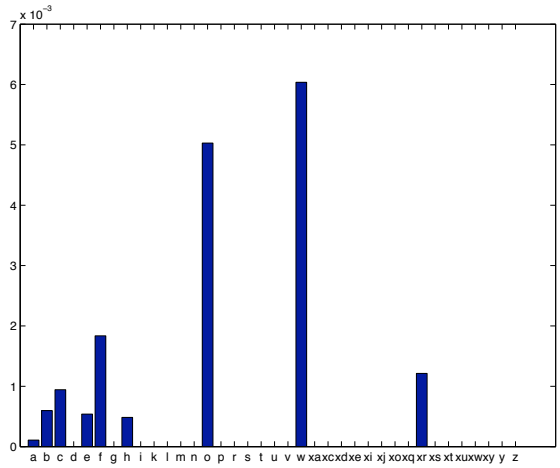


(b) Group 10

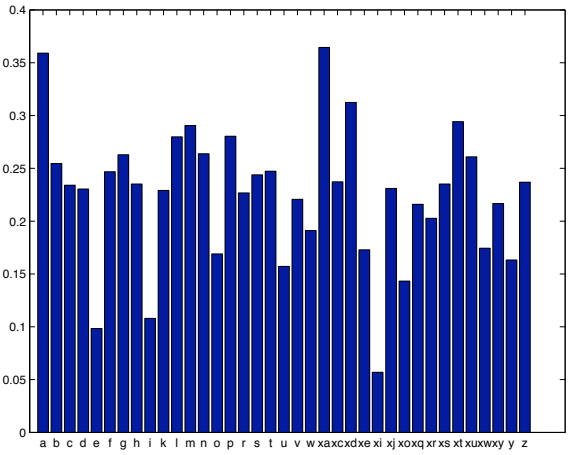


(c) Group 11

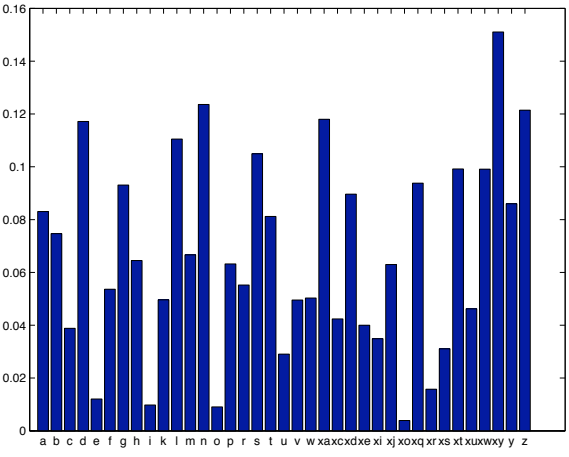
Figure A.2: Responses of groups 9, 10, 11



(a) Group 12

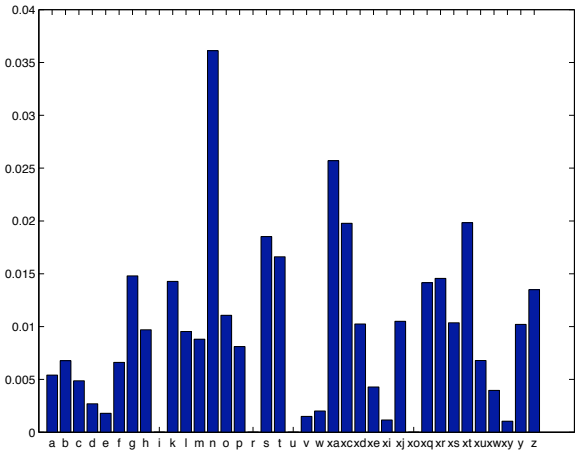


(b) Group 13

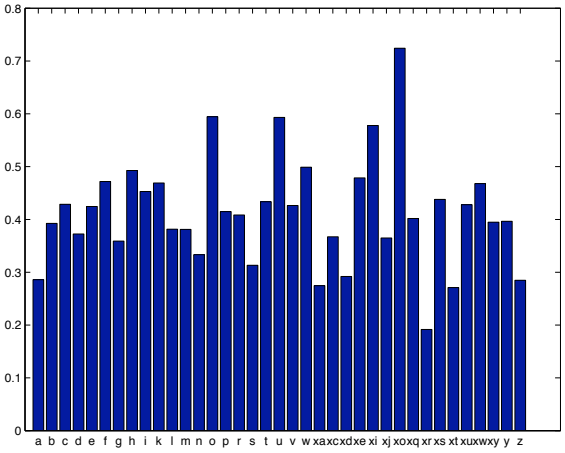


(c) Group 14

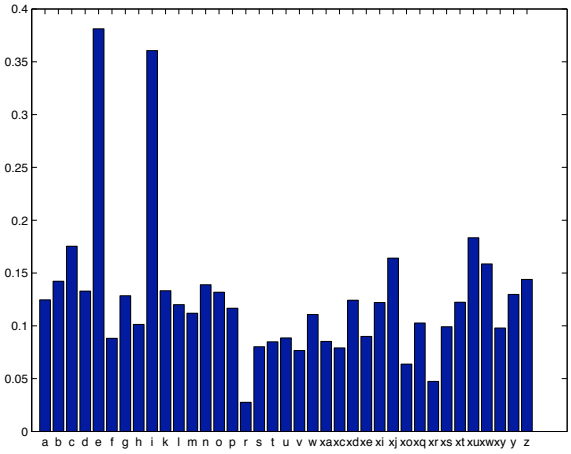
Figure A.3: Responses of groups 12, 13, 14



(a) Group 15

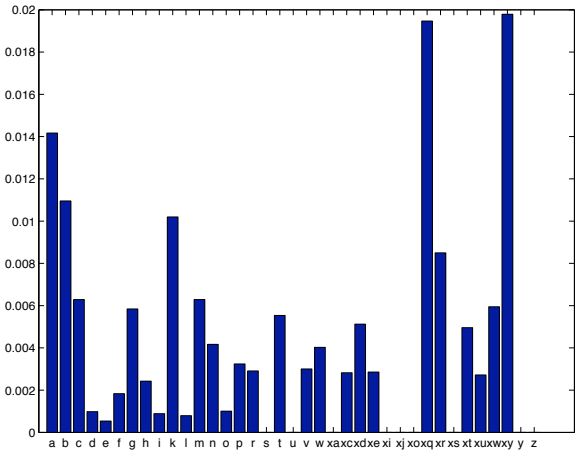


(b) Group 16

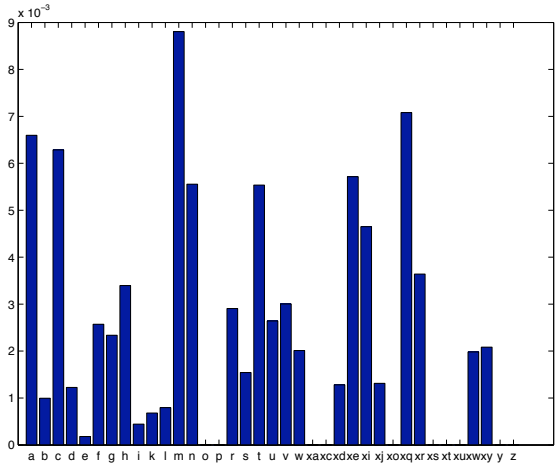


(c) Group 18

Figure A.4: Responses of groups 15, 16, 18



(a) Group 21



(b) Group 23

Figure A.5: Responses of groups 21, 23

REFERENCES

[1] B. Meyer, M. Wachter, T. Brand, and B. Kollmeier, "Phoneme confusions in human and automatic speech recognition," in *Proc. Interspeech*, 2007, pp. 1485–1488.

[2] M. Elhilali and S. Shamma, "A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation," *Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3751–3771, Dec. 2008.

[3] M. Cheour et al., "Development of language-specific phoneme representations in the infant brain," *Nature Neuroscience*, vol. 1, no. 5, pp. 351–353, Sept. 1998.

[4] M. Doniec, B. Scasselati, and W. Miranker, "Emergence of language-specific phoneme classifiers in self-organized maps," in *Proc. Intl. Joint Conf. on Neural Networks*, 2007, pp. 2081–2086.

[5] M. Coen, "Self-supervised acquisition of vowels in American English," in *Proc. Natl. Conf. on Artificial Intelligence*, 2006, pp. 1451–1456.

[6] A. Boothroyd and S. Nittrouer, "Mathematical treatment of context effects in phoneme and word recognition," *Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 101–114, July 1988.

[7] S. Levinson, *Mathematical Models for Speech Technology*. New York, NY: Wiley, 2005.

[8] J. Weng et al., "Autonomous mental development by robots and animals," *Science*, vol. 291, no. 5504, pp. 599–600, Jan. 2001.

[9] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, pp. 433–460, 1950.

[10] V. Mountcastle, "An organizing principle for cerebral function: the unit model and the distributed system," in *The Mindful Brain*. Cambridge, MA: MIT Press, 1978.

[11] M. Asada, K. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, pp. 185–193, 2001.

[12] J. Weng, "On developmental mental architectures," *Neurocomputing*, vol. 70, pp. 2303–2323, 2007.

[13] P. Dominey and J. D. Boucher, "Developmental stages of perception and language acquisition in a perceptually grounded robot," *Cognitive Systems Research*, vol. 6, pp. 243–259, 2005.

[14] X. He, R. Kojima, and O. Hasegawa, "Developmental word grounding through a growing neural network with a humanoid robot," *IEEE Trans. Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 37, no. 2, pp. 451–462, April 2007.

[15] A. Cangelosi, E. Hourdakis, and V. Tikhanoff, "Language acquisition and symbol grounding transfer with neural networks and cognitive robots," in *Proc. Intl. Joint Conf. on Neural Networks*, 2006.

[16] S. Wermter et al., "Towards multimodal neural robot learning," *Robotics and Autonomous Systems*, vol. 47, pp. 171–175, 2004.

[17] N. Iwahashi, "Language acquisition through a human-robot interface by combining speech, visual, and behavioral information," *Information Sciences*, vol. 156, pp. 109–121, 2003.

[18] K. Squire, "Hmm-based semantic learning for a mobile robot," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2004.

[19] M. McClain, "Semantic based learning of syntax in an autonomous robot," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2006.

[20] Numenta Inc., "Numenta platform for intelligent computing," July 2008. [Online]. Available: http://www.numenta.com/.

[21] V. Mountcastle, "The columnar organization of the neocortex," *Brain*, vol. 120, no. 4, pp. 701–722, 1997.

[22] D. Felleman and D. V. Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[23] T.-S. Lee and D. Mumford, "Hierarchical Bayesian inference in the visual cortex," *Journal of the Optical Society of America*, vol. 2, no. 7, pp. 1434–1448, 2003.

[24] D. George, "How the brain might work: A hierarchical and temporal model for learning and recognition," Ph.D. dissertation, Stanford University, Palo Alto, CA, 2008.

[25] B. Bobier and M. Wirth, "Content-based image retrieval using hierarchical temporal memory," in *Proc. 16th ACM Intl. Conf. Multimedia*, 2008, pp. 925–928.

[26] J. van Doremalen and L. Boves, "Spoken digit recognition using a hierarchical temporal memory," in *Proc. Interspeech*, 2008.

[27] Y. Muthusamy, R. Cole, and B. Oshika, *CLSU: Multilanguage Telephone Speech Version 1.2*. Philadelphia, PA: Linguistic Data Consortium, 2006.

[28] J. Wright, *Articulation Index*. Philadelphia, PA: Linguistic Data Consortium, 2005.

[29] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustic, Speech, and Signal Processing*, vol. 28, no. 4, 1980.