Modelling Early Language Acquisition Skills: Towards a General Statistical Learning Mechanism

Guillaume Aimetti
University of Sheffield
Sheffield, UK
g.aimetti@dcs.shef.ac.uk

Abstract

This paper reports the on-going research of a thesis project investigating a computational model of early language acquisition. The model discovers word-like units from cross-modal input data and builds continuously evolving internal representations within a cognitive model of memory. Current cognitive theories suggest that young infants employ general statistical mechanisms that exploit the statistical regularities within their environment to acquire language skills. The discovery of lexical units is modelled on this behaviour as the system detects repeating patterns from the speech signal and associates them to discrete abstract semantic tags. In its current state, the algorithm is a novel approach for segmenting speech directly from the acoustic signal in an unsupervised manner, therefore liberating it from a pre-defined lexicon. By the end of the project, it is planned to have an architecture that is capable of acquiring language and communicative skills in an online manner, and carry out robust speech recognition. Preliminary results already show that this method is capable of segmenting and building accurate internal representations of important lexical units as 'emergent' properties from cross-modal data.

1 Introduction

Conventional Automatic Speech Recognition (ASR) systems can achieve very accurate recognition results, particularly when used in their optimal acoustic environment on examples within their stored vocabularies. However, when taken out of their comfort zone accuracy significantly deteriorates and does not come anywhere near human speech processing abilities for even the simplest of tasks. This project investigates novel computational language acquisition techniques that attempt to model current cognitive theories in order to achieve a more robust speech recognition system.

Current cognitive theories suggest that our surrounding environment is rich enough to acquire language through the use of simple statistical processes, which can be applied to all our senses. The system under development aims to help clarify this theory, implementing a computational model that is general across multiple modalities and has not been pre-defined with any linguistic knowledge.

In its current form, the system is able to detect words directly from the acoustic signal and incrementally build internal representations within a memory architecture that is motivated by cognitive plausibility. The algorithm proposed can be split into two main processes, automatic segmentation and word discovery. Automatically segmenting speech directly from the acoustic signal is made possible through the use of dynamic programming (DP); we call this method acoustic DP-ngram's. The second stage, key word discovery (KWD), enables the model to hypothesise and build internal representations of word classes that associates the discovered lexical units with discrete abstract semantic tags.

Cross-modal input is fed to the system through the interaction of a carer module as an 'audio' and 'visual' stream. The audio stream consists of an acoustic signal representing an utterance, while the visual stream is a discrete abstract semantic tag referencing the presence of a key word within the utterance.

Initial test results show that there is significant potential with the current algorithm, as it segments in an unsupervised manner and does not rely on a predefined lexicon or acoustic phone models that constrain current ASR methods.

Proceedings of the EACL 2009 Student Research Workshop, pages 1–9,
Athens, Greece, 2 April 2009. ©2009 Association for Computational Linguistics

The rest of this paper is organized as follows. Section 2 reviews current developmental theories and computational models of early language acquisition. In section 3, we present the current implementation of the system. Preliminary experiments and results are described in sections 4 and 5 respectively. Conclusions and further work are discussed in sections 6 and 7 respectively.

2 Background

2.1 Current Developmental Theories

The 'nature' vs. 'nurture' debate has been fought out for many years now; are we born with innate language learning capabilities, or do we solely use the input from the environment to find structure in language?

Nativists believe that infants have an innate capability for acquiring language. It is their view that an infant can acquire linguistic structure with little input and that it plays a minor role in the speed and sequence with which they learn language. Noam Chomsky is one of the most cited language acquisition nativists, claiming children can acquire language "On relatively slight exposure and without specific training" (Chomsky, 1975, p.4).

On the other hand, non-nativists argue that the input contains much more structural information and is not as full of errors as suggested by nativists (Eimas *et al*., 1971; Best *et al*., 1988; Jusczyk *et al*., 1993; Saffran *et al*., 1996; Christiansen *et al*., 1998; Saffran *et al*., 1999; Saffran *et al*., 2000; Kirkham *et al*., 2002; Anderson *et al*., 2003; Seidenberg *et al*., 2002; Kuhl, 2004; Hannon and Trehub, 2005).

Experiments by Saffran *et al*. (1996, 1999) show that 8-month old infants use the statistical information in speech as an aid for word segmentation with only two minutes of familiarisation.

Inspired by these results, Kirkham *et al*. (2002) suggest that the same statistical processes are also present in the visual domain. Kirkham *et al*. (2002) carried out experiments showing that preverbal infants are able to learn patterns of visual stimuli with very short exposure.

Other theories hypothesise that statistical and grammatical processes are both used when learning language (Seidenberg *et al*., 2002; Kuhl, 2004). The hypothesis is that newborns begin life using statistical processes for simpler problems, such as learning the sounds of their native language and building a lexicon, whereas grammar is learnt via non-statistical methods later on. Seidenberg *et al*. (2002) believe that learning grammar begins when statistical learning ends. This has proven to be a very difficult boundary to detect.

2.2 Current Computational Models

There has been a lot of interest in trying to segment speech in an unsupervised manner, therefore liberating it from the required expert knowledge needed to predefine the lexical units for conventional ASR systems. This has led speech recognition researchers to delve into the cognitive sciences to try and gain an insight into how humans achieve this without much difficulty and model it.

Brent (1999) states that for a computational algorithm to be cognitively plausible it must:

- Start with no prior knowledge of general language structure.
- Learn in a completely unsupervised manner.
- Segment incrementally.

An automatic segmentation method similar to that of the acoustic DP-ngram method is segmental DTW. Park & Glass (2008) have adapted dynamic time warping (DTW) to find matching acoustic patterns between two utterances. The discovered units are then clustered, using an adjacency graph method, to describe the topic of the speech data.

Statistical Word Discovery (SWD) (ten Bosch and Cranen, 2007) and the Cross-channel Early Lexical Learning (CELL) model (Roy and Pentland, 2002), also similar methods to the one described in this paper, discover word-like units and then updating internal representations through clustering processes. The downfall of the CELL approach is that it assumes speech is observed as an array of phone probabilities.

A more radical approach is Non-negative matrix factorization (NMF) (Stouten *et al*., 2008). NMF detects words from 'raw' cross-modal input without any kind of segmentation during the whole process, coding recurrent speech fragments into to 'word-like' entities. However, the factorisation process removes all temporal information.

3 The Proposed System

3.1 ACORNS

The computational model reported in this paper is being developed as part of a European project called ACORNS (Acquisition of Communication

and Recognition Skills). The ACORNS project intends to design an artificial agent (Little Acorns) that is capable of acquiring human verbal communication skills. The main objective is to develop an end-to-end system that is biologically plausible; restricting the computational and mathematical methods to those that model behavioural data of human speech perception and production within five main areas:

Front-end Processing: Research and development of new feature representations guided by phonetic and psycho-linguistic experiments.

Pattern Discovery: Little Acorns (LA) will start life without any prior knowledge of basic speech units, discovering them from patterns within the continuous input.

Memory Organisation and Access: A memory architecture that approaches cognitive plausibility is employed to store discovered units.

Information Discovery and Integration: Efficient and effective techniques for retrieving the patterns stored in memory are being developed.

Interaction and Communication: LA is given an innate need to grow his vocabulary and communicate with the environment.

3.2 The Computational Model

There are two key processes to the language acquisition model described in this paper; automatic segmentation and word discovery. The automatic segmentation stage allows the system to build a library of similar repeating speech fragments directly from the acoustic signal. The second stage associates these fragments with the observed semantic tags to create distinct key word classes.

Automatic Segmentation

The acoustic DP-ngram algorithm reported in this section is a modification of the preceding DP-ngram algorithm (Sankoff and Kruskal, 1983; Nowell and Moore, 1995). The original DP-ngram model was developed by Sankoff and Kruskal (1983) to find two similar portions of gene sequences. Nowell and Moore (1995) then modified this model to find repeated patterns within a single phone transcription sequence through self-similarity. Expanding on these methods, the author has developed a variant that is able to segment speech, directly from the acoustic signal; automatically segmenting important lexical fragments by discovering 'similar' repeating patterns. Speech is never the same twice and therefore impossible to find exact repetitions of importance (e.g. phones, words or sentences).

The use of DP allows this algorithm to accommodate temporal distortion through dynamic time warping (DTW). The algorithm finds partial matches, portions that are similar but not necessarily identical, taking into account noise, speed and different pronunciations of the speech.

Traditional template based speech recognition algorithms using DP would compare two sequences, the input speech vectors and a word template, penalising insertions, deletions and substitutions with negative scores. Instead, this algorithm uses quality scores, positive and negative, to reward matches and prevent anything else; resulting in longer, more meaningful subsequences.

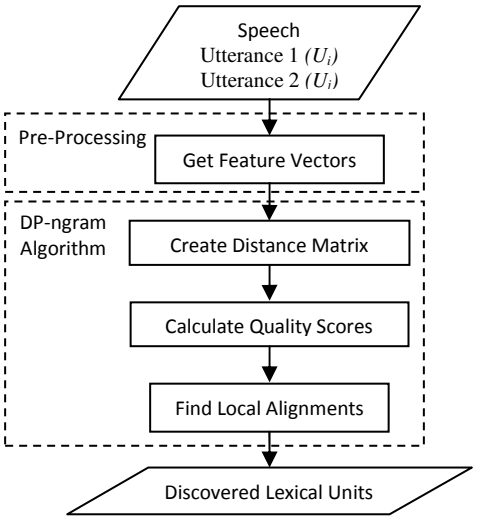

Figure 1: Acoustic DP-ngram Processes.

Figure 1 displays the simplified architecture of the acoustic DP-ngram algorithm. There are four main stages to the process:

Stage 1: The ACORNS MFCC front-end is used to parameterise the raw speech signal of the two utterances being fed to the system. The default settings have been used to output a series of 37-element feature vectors. The front-end is based on Mel-Frequency Coefficients (MFCC), which reflects the frequency sensitivity of the auditory system, to give 12 MFCC coefficients. A measure of the raw energy is added along with 12 differential (Δ) and 12 2nd differential ($\Delta\Delta$) coefficients. The front-end also allows the option for cepstral mean normalisation (CMN) and cepstral mean and variance normalisation (CMVN).

Stage 2: A local-match distance matrix is then calculated by measuring the cosine distance be-

tween each pair of frames (v_1, v_2) from the two sequences, which is defined by:

$$d(v_1, v_2) = (v_1^T . v_2) / (\|v_1\|^T . \|v_2\|) \quad (1)$$

Stage 3: The distance matrix is then used to calculate accumulative quality scores for successive frame steps. The recurrence defined in equation (2) is used to find all quality scores $q_{i,j}$.

In order to maximize on quality, substitution scores must be positive and both insertion and deletion scores must be negative as initialised in equation (3).

$$q_{i,j} = \max \begin{cases} q_{i-1,j} + \left(s_{a_i,\phi} . |d_{i-1,j} - 1| . q_{i-1,j}\right), \\ q_{i,j-1} + \left(s_{\phi,b_j} . |d_{i,j-1} - 1| . q_{i,j-1}\right), \\ q_{i-1,j-1} + \left(s_{a_i,b_j} . d_{i-1,j-1} . q_{i-1,j-1}\right), \\ 0, \end{cases} \quad (2)$$

where,

$$\begin{aligned} s_{a_i,\phi} &= -1.1 \quad \text{(Insertion score)} \\ s_{\phi,b_j} &= -1.1 \quad \text{(Deletion score)} \\ s_{a_i,b_j} &= +1.1 \quad \text{(Substitution score)} \\ d_{i,j} &= \text{frame-frame distance} \\ q_{i,j} &= \text{Accumulative quality score} \end{aligned} \quad (3)$$

The recurrence in equation (2) stops past dissimilarities causing global effects by setting all negative scores to zero, starting a fresh new homologous relationship between local alignments.

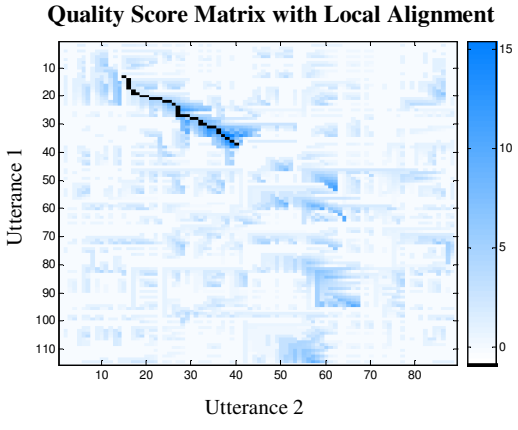

Figure 2: Quality score matrix calculated from two different utterances. The plot also displays the optimal local alignment.

Figure 2 shows the plot of the quality scores calculated from two different utterances. The shaded areas show repeating structure; longer and more accurate fragments attain greater quality scores, indicated by the darker areas within the plot.

Applying a substitution score of 1 will cause the accumulative quality score to grow as a linear function. The current settings defined by equation (3) use a substitution score greater than 1, thus allowing local accumulative quality scores to grow exponentially, giving longer alignments more importance.

By setting insertion and deletion scores to values less than -1, the model will find closer matching acoustic repetitions; whereas a value greater than -1 and less than 0 allows the model to find repeated patterns that are longer and less accurate, therefore allowing control over the tolerance for temporal distortion.

Stage 4: The final stage is to discover local alignments from within the quality score matrix. Backtracking pointers (bt) are maintained at each step of the recursion:

$$bt_{i,j} = \begin{cases} (i-1, j), & \text{(Insertion)} \\ (i, j-1), & \text{(Deletion)} \\ (i-1, j-1), & \text{(Substitution)} \\ (0,0) & \text{(Initial pointer)} \end{cases} \quad (4)$$

When the quality scores have been calculated through equation (2), it is possible to backtrack from the highest score to obtain the local alignments in order of importance with equation (4). A threshold is set so that only local alignments above a desired quality score are to be retrieved. Figure 2 presents the optimal local alignment that was discovered by the acoustic DP-ngram algorithm for the utterances "Ewan is shy" and "Ewan sits on the couch".

The discovered repeated pattern (the dark line in figure 2) is [y uw ah n]. Start and stop times are collected which allows the model to retrieve the local alignment from the original audio signal in full fidelity when required.

Key Word Discovery

The milestone set for all systems developed within the ACORNS project is for LA to learn 10 key words. To carry out this task, the DP-ngram algorithm has been modified with the addition of a key word discovery (KWD) method that continues the theme of a general statistical learning mechanism. The acoustic DP-ngram algorithm exploits the co-occurrence of similar acoustic patterns within different utterances; whereas, the

KWD method exploits the co-occurrence of the associated discrete abstract semantic tags. This allows the system to associate cross-modal repeating patterns and build internal representations of the key words.

KWD is a simple approach that creates a class for each key word (semantic tag) observed, in which all discovered exemplar units representing each key word are stored. With this list of episodic segments we can perform a clustering process to derive an ideal representation of each key word.

For a single iteration of the DP-ngram algorithm, the current utterance (Utt_{cur}) is compared with another utterance in memory (Utt_n). KWD hypothesises whether the segments found within the two utterances are potential key words, by simply comparing the associated semantic tags. There are three possible paths for a single iteration:

1: If the tag of Utt_{cur} has never been seen then create a new key word class and store the whole utterance as an exemplar of it. Do not carry out the acoustic DP-ngram process and proceed to the next utterance in memory (Utt_{n+1}).

2: If both utterances share the same tag then proceed with the acoustic DP-ngram process and append discovered local alignments to the key word class representing that tag. Proceed to the next utterance in memory (Utt_{n+1}).

3: If both utterances contain different tags then do not carry out acoustic DP-ngram's and proceed to the next utterance in memory (Utt_{n+1}).

By creating an exemplar list for each key word class we are able to carry out a clustering process that allows us to create a model of the ideal representation. Currently, the clustering process implemented simply calculates the 'centroid' exemplar, finding the local alignment with the shortest distance from all the other local alignments within the same class. The 'centroid' is updated every time a new local alignment is added, therefore the system is creating internal representations that are continuously evolving and becoming more accurate with experience.

For recognition tasks the system can be set to use either the 'centroid' exemplar or all the stored local alignments for each key word class.

LA Architecture

The algorithm runs within a memory structure (fig. 3) developed with inspiration from current cognitive theories of memory (Jones *et al.*, 2006). The memory architecture works as follows:

Carer: The carer interacts with LA to continuously feed the system with cross-modal input (acoustic & semantic).

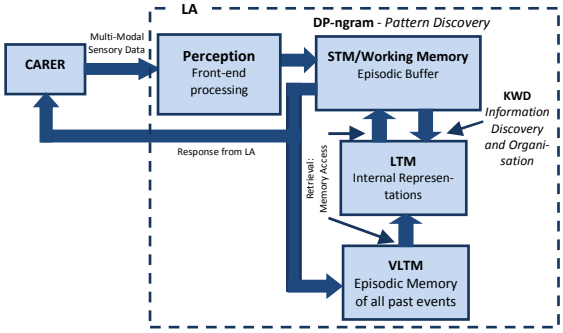

Figure 3: Little Acorns' memory architecture.

Perception: The stimulus is processed by the 'perception' module, converting the acoustic signal into a representation similar to the human auditory system.

Short Term Memory (STM): The output of the 'perception' module is stored in a limited STM which acts as a circular buffer to store *n* past utterances. The *n* past utterances are compared with the current input to discover repeated patterns in an incremental fashion. As a batch process LA can only run on a limited number of utterances as the search space is unbound. As an incremental process, LA could potentially handle an infinite number of utterances, thus making it a more cognitively plausible system.

Long Term Memory (LTM): The ever increasing lists of discovered units for each key word representation are stored in LTM. Clustering processes can then be applied to build and update internal representations. The representations stored within LTM are only pointers to where the segment lies within the very long term memory.

Very Long Term Memory: The very long term memory is used to store every observed utterance. It is important to note that unless there is a pointer for a segment of speech within LTM then the data cannot be retrieved. But, future work may be carried out to incorporate additional 'sleeping' processes on the data stored in VLTM to re-organise internal representations or carry out additional analysis.

4 Experiments

Accuracy of experiments within the ACORNS project is based on LA's response to its carer. The correct response is for LA to predict the key

word tag associated with the current incoming utterance while only observing the speech signal. LA re-uses the acoustic DP-ngram algorithm to solve this task in a similar manner to traditional DP template based speech recognition. The recognition process is carried out by comparing exemplars, of discovered key words, against the current incoming utterance and calculating a quality distance (as described in stage 3 of section 3.2). Thus, the exemplar producing the highest quality score, by finding the longest alignment, is taken to be the match, with which we can predict its associated visual tag.

A number of different experiments have been carried out:

E1 - Optimal STM Window: This experiment finds the optimal utterance window length for the system as an incremental process. Varying values of the utterance window length (from 1 to 100) were used to obtain key word recognition accuracy results across the same data set.

E2 - Batch vs. Incremental: The optimal window length chosen for the incremental implementation is compared against the batch implementation of the algorithm.

E3 - Centroid vs. Exemplars: The KWD process stores a list of exemplars representing each key word class. For the recognition task we can either use all the exemplars in each key word list or a single 'centroid' exemplar that best represents the list. This experiment will compare these two methods for representing internal representations of the key words.

E4 – Speaker Dependency: The algorithm is tested on its ability to handle the variation in speech from different speakers with different feature vectors.

V_1 = HTK MFCC's (no norm)
V_2 = ACORNS MFCC's (no norm)
V_3 = ACORNS MFCC's (Cepstral Mean Norm)
V_4 = ACORNS MFCC's (Cepstral Mean and Variance Norm)

Using normalisation methods will reduce the information within the feature vectors, removing some of the speaker variation. Therefore, key word detection should be more accurate for a data set of multiple speakers with normalisation.

4.1 Test Data

The ACORNS English corpus is used for the above experiments. Sentences were created by combining a carrier sentence with a keyword. A total of 10 different carrier sentences, such as "*Do you see the X*", "*Where is the X*", etc., where *X* is a keyword, were combined with one of ten different keywords, such as "*Bottle*", "*Ball*", etc. This created 100 unique sentences which were repeated 10 times and recorded with 4 different speakers (2 male and 2 female) to produce 4000 utterances.

In addition to the acoustic data, each utterance is associated with an abstract semantic tag. As an example, the utterance *"What matches this **shoe**"* will contain the tag referring to *"**shoe**"*. The tag does not give any location or phonetic information about the key word within the utterance.

E1 and **E2** use a sub-set of 100 different utterances from a single speaker. **E3** is carried out on a sub-set of 200 utterances from a single speaker and the database used for **E4** is a sub-set of 200 utterances from all four speakers (2 male and 2 female) presented in a random order.

5 Results

E1: LA was tested on 100 utterances with varying utterance window lengths. The plot in figure 4 shows the total key word detection accuracy for each window length used. The x-axis displays the utterance window lengths (1–100) and the y-axis displays the total accuracy.

The results are as expected. Longer window lengths achieve more accurate results. This is because longer window lengths produce a larger search space and therefore have more chance of capturing repeating events. Shorter window lengths are still able to build internal representations, but over a longer period.

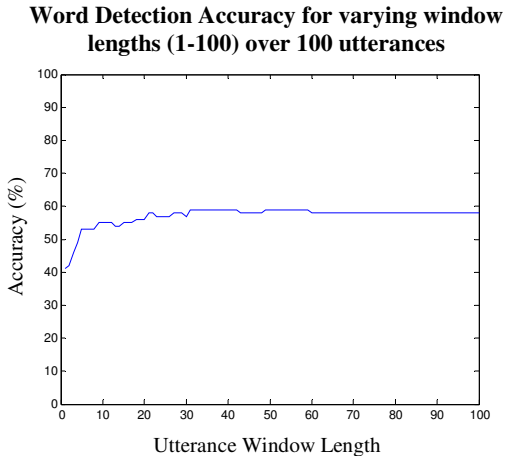

Figure 4: Single speaker key word accuracy using varying utterance window lengths of 1-100.

Accuracy results reach a maximum with an utterance window length of 21 and then stabilize at around 58% (±1%). From this we can conclude

that 21 is the minimum window length needed to build accurate internal representations of the words within the test set, and will be used for all subsequent experiments.

E2: The plot in figure 4 displays the total key word detection accuracy for the different utterance window lengths and does not show the gradual word acquisition process. Figure 5 compares the word detection accuracy of the system (y-axis) as a function of the number of utterances observed (x-axis). Accuracy is recorded as the percentage of correct replies for the last ten observations. The long discontinuous line in the plot shows the word detections accuracy for randomly guessing the key word.

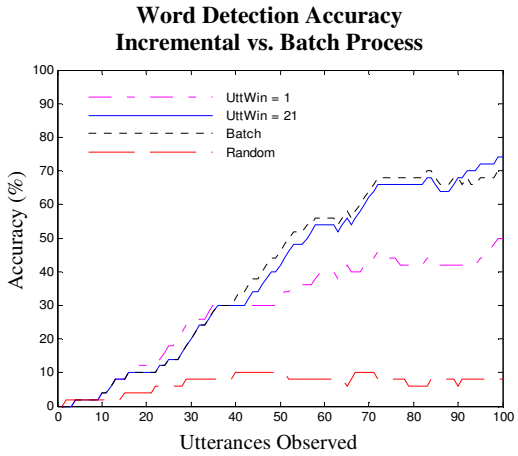

Figure 5: Word detection accuracy LA running as a batch and incremental process. Results are plotted as a function of the past 10 utterances observed.

It can be seen from the plot in figure 5 that the system begins life with no word representations. At the beginning, the system hypothesises new word units from which it can begin to bootstrap its internal representations.

As an incremental process, with the optimal window length, the system is able to capture enough repeating patterns and even begins to outperform the batch process after 90 utterances. This is due to additional alignments discovered by the batch process that are temporarily distorting a word representation, but the batch process would 'catch up' in time.

Another important result to take into account is that only comparing the current incoming utterance with the last observed utterance is enough to build word representations. Although this is very efficient, the problem is that there is a greater possibility that some words will never be discovered if they are not present in adjacent utterances within the data set.

E3: Currently the recognition process uses all the discovered exemplars within each key word class. This process causes the computational complexity to increase exponentially. It is also not suitable for an incremental process with the potential of running on an infinite data set.

To tackle this problem, recognition was carried out using the 'centroid' exemplar of each key word class. Figure 6 shows the word detection accuracy as a function of utterances observed for both methods.

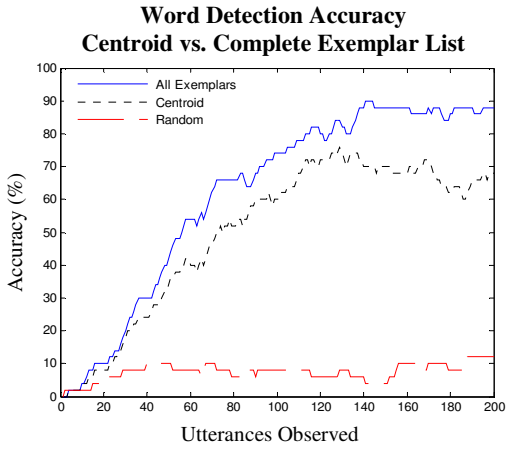

Figure 6: Word detection accuracy using centroids and complete exemplar list for recognition.

The results show that the 'centroid' method is quickly outperformed and that the word detection accuracy difference increases with experience. After 120 utterances performance seems to gradually decline. This is because the 'centroid' method cannot handle the variation in the acoustic speech data. Using all the discovered units for recognition allows the system to reach an accuracy of 90% at around 140 utterances, where it then seems to stabilise at around 88%.

E4: The addition of multiple speakers will add greater variation to the acoustic signal, distorting patterns of the same underlying unit. Over the 200 utterances observed, word detection accuracy of the internal representations increases, but at a much slower rate than the single speaker experiments (fig. 7).

The assumption that using normalisation methods would achieve greater word detection accuracy, by reducing speaker variation, does not hold true. On reflection this comes as no surprise, as the system collects exemplar units with a larger relative fidelity for each speaker.

This raises an important issue; the optimal utterance window length for the algorithm as an incremental process was calculated for a single

speaker, therefore, increasing the search space will allow the model to find more repeating patterns from the same speaker. Following this logic, it could be hypothesised that the optimal search space should be four times the size used for one speaker and that it will take four times as many observations to achieve the same accuracy.

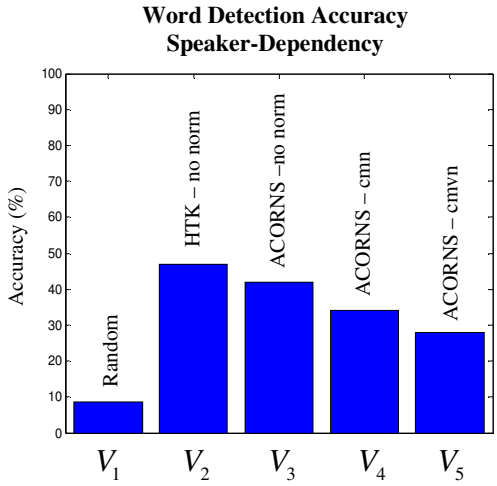

Figure 7: Total accuracy using different feature vectors after 200 observed utterances.

6 Conclusions

Preliminary results indicate that the environment is rich enough for word acquisition tasks. The pattern discovery and word learning algorithm implemented within the LA memory architecture has proven to be a successful approach for building stable internal representations of word-like units. The model approaches cognitive plausibility by employing statistical processes that are general across multiple modalities. The incremental approach also shows that the model is still able to learn correct word representations with a very limited working memory model.

Additionally to the acquisition of words and word-like units, the system is able to use the discovered tokens for speech recognition. An important property of this method, that differentiates it from conventional ASR systems, is that it does not rely on a pre-defined vocabulary, therefore reducing language-dependency and out-of-dictionary errors.

Another advantage of this system, compared to systems such as NMF, is that it is able to give temporal information of the whereabouts of important repeating structure which can be used to code the acoustic signal as a lossless compression method.

7 Discussion & Future Work

A key question driving this research is whether modelling human language acquisition can help create a more robust speech recognition system. Therefore further development of the proposed architecture will continue to be limited to cognitively plausible approaches and should exhibit similar developmental properties as early human language learners. In its current state, the system is fully operational and intends to be used as a platform for further development and experiments.

The experimental results are promising. However, it is clear to see that the model suffers from speaker-dependency issues. The problem can be split into two areas, front-end processing of the incoming acoustic signal and the representation of discovered lexical units in memory.

Development is being carried out on various clustering techniques that build constantly evolving internal representations of internal lexical classes in an attempt to model speech variation. Additionally, a secondary update process, implemented as a re-occurring ‘sleeping phase’ is being investigated. This phase is going to allow the memory organisation to re-structure itself by looking at events over a longer history, which could be carried out as a batch process.

The processing of prosodic cues, such as speech rhythm and pitch intonation, will be incorporated within the algorithm to increase the key word detection accuracy and further exploit the richness of the learners surrounding environment. Adults, when speaking to infants, will highlight words of importance through infant directed speech (IDS). During IDS adults place more pitch variance on words that they want the infant to attend to.

Further experiments have been planned to see if the model exhibits similar patterns of learning behaviour as young multiple language learners. Experiments will be carried out with the multiple languages available in the ACORNS database (English, Finnish and Dutch).

Acknowledgement

This research was funded by the European Commission, under contract number FP6-034362, in the ACORNS project (www.acorns-project.org). The author would also like to thank Prof. Roger K. Moore for helping to shape this work.

References

A. Park and J. R. Glass. 2008. Unsupervised Pattern Discovery in Speech. *Transactions on Audio, Speech and Language Processing*, 16(1):186-197.

C. T. Best, G. W. McRoberts and N. M. Sithole. 1988. Examination of the perceptual re-organization for speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, 14:345-360.

D. M. Jones, R. W. Hughes and W. J. Macken. 2006. Perceptual Organization Masquerading as Phonological Storage: Further Support for a Perceptual-Gestural View of Short-Term Memory. *Journal of Memory and Language,* 54:265-328.

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

D. Sankoff and Kruskal J. B. 1983. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley Publishing Company, Inc.

E. E. Hannon and S. E. Trehub. 2005. Turning in to Musical Rhythms: Infants Learn More readily than Adults. *PNAS*, 102(35):12639-12643.

J. L. Anderson, J. L. Morgan and K. S. White. 2003. A Statistical Basis for Speech Sound Discrimination. *Language and Speech*, 46(43):155-182.

J. R. Saffran, R. N. Aslin and E. L. Newport. 1996. Statistical Learning by 8-Month-Old Infants. *SCIENCE*, 274:1926-1928.

J. R. Saffran, E. K. Johnson, R. N. Aslin and E. L. Newport. 1999. Statistical Learning of Tone Sequences by Human Infants and Adults. *Cognition*, 70(1):27-52.

J. R. Saffran, A. Senghashas and J. C. Trueswell. 2000. The Acquisition of Language by Children. *PNAS*, 98(23):12874-12875.

L. ten Bosch and B. Cranen. 2007. A Computational Model for Unsupervised Word Discovery. *INTERSPEECH 2007*, 1481-1484.

M. H. Christiansen, J. Allen and M. Seidenberg. 1998. Learning to Segment Speech using Multiple Cues. *Language and Cognitive Processes*, 13:221-268.

M. S. Seidenberg, M. C. MacDonald and J. R. Saffran. 2002. Does Grammar Start Where Statistics Stop?. *SCIENCE*, 298:552-554.

M. R. Brent. 1999. Speech Segmentation and Word Discovery: A Computational Perspective. *Trends in Cognitive Sciences*, 3(8):294-301.

N. Chomsky. 1975. *Reflections on Language*. New York: Pantheon Books.

N. Z. Kirkham, A. J. Slemmer and S. P. Johnson. 2002. Visual Statistical Learning in Infancy: Evidence for a Domain General Learning Mechanism. *Cognition*, 83:B35-B42.

P. D. Eimas, E. R. Siqueland, P. Jusczyk and J. Vigorito. 1971. Speech Perception in Infants. *Science*, 171(3968):303-606.

P. K. Kuhl. 2004. Early Language Acquisition: Cracking the Speech Code. *Nature*, 5:831-843.

P. Nowell and R. K. Moore. 1995. The Application of Dynamic Programming Techniques to Non-Word Based Topic Spotting. *EuroSpeech '95*, 1355-1358.

P. W. Jusczyk, A. D. Friederici, J. Wessels, V. Y. Svenkerud and A. M. Jusczyk. 1993. Infants' Sensitivity to the Sound Patterns of Native Language Words. *Journal of Memory & Language*, 32:402-420.

V. Stouten, K. Demuynck and H. Van hamme. 2008. Discovering Phone Patterns in Spoken Utterances by Non-negative Matrix Factorisation. *IEEE Signal Processing Letters*, 131-134.