PhD Thesis Transfer Report

12-Month Meeting

Guillaume Aimetti
Speech and Hearing Research Group
Dept. of Computer Science
University of Sheffield
19/10/2008

Contents

1. Introduction

Conventional Automatic Speech Recognition (ASR) systems can achieve very accurate recognition results, particularly when used in their optimum acoustic environment on examples within their stored vocabularies. However, when taken out of its 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.

There are a number of computational models that attempt to model and explain how children segment speech and discover words (for a detailed review see Brent, 1999). However, most of these simulations take transcripts of speech as their input. The system under development for this project by-passes this stage, working directly on the acoustic signal, opening up a niche area of research within this field.

Current cognitive theories suggest that our surrounding environment is rich enough to acquire language through the use of simple statistical processes that can be applied to all our senses. The system being developed within this project aims to clarify this theory, implementing cognitively plausible algorithms that are general across multiple modalities and have not been pre-defined with any linguistic knowledge.

Developing a speech recognition system implementing processes that may be employed by human language learners should also exhibit similar perceptual properties. Infants are born with universal speech perception abilities, using the statistical regularities in speech to help them find structure. At around 6-8 months the infant loses these universal abilities when it acquires language-specific perceptual abilities. Behavioural data demonstrates the loss of universal perceptual abilities through phonetic categorisation. The system should therefore demonstrate similar developmental milestones.

By the end of the project, it is planned to have an architecture that is capable of acquiring language and communicative skills. If the system drastically fails then it will bring into question current non-nativist theories; that our environment is rich enough to learn language with the use of general statistical mechanisms. However, preliminary results already show that it is possible to segment and build accurate internal representations of important lexical units from speech data.

2. Key Research Questions

1. The main question of focus is whether modeling human language acquisition can help create a more robust speech recognition system?

2. Is our surrounding environment rich enough to learn language without any prior knowledge about speech sounds, words or meanings?

3. What makes up the fundamental units of speech? Do infants begin to learn language by discovering larger more meaningful units of speech such as words or phrases, or smaller units such as phones and syllables?

4. Is the replacement of universal speech perception for language-specific perception skills an innate function in our brain that occurs at a discrete time or a stage-like change that occurs due to a gradual learning process? Are statistical learning processes replaced as the infant acquires semantic knowledge and starts to build linguistic rules for its native language?

5. Early language learners will use cues from multiple modalities to aid speech segmentation and word discovery. Can this process be computationally modelled? And more importantly, how does an infant deal with conflicting cues?

3. Language Acquisition - Critical Literature Review

3.1 Cognitive Theories

3.1.1 Nature vs. Nurture

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). Some nativists believe that it is even possible to acquire language without any input. Research to support this belief was carried out with deaf children showing that they automatically developed 'homesign' with limited exposure to language (Lust, 2006). Other evidence is a topography study revealing that the left hemisphere of a newborn's brain is superior to the right hemisphere for processing speech (Pena et al., 2003). Other nativists have hypothesised that innate linguistic knowledge is attributed with complex parameters that the infant must set, depending on the input, for their native language (Gathercole & Hoff, 2007).

Non-nativists argue that the input contains much more structural information and is not as full of errors as suggested by the nativists. Much research has been carried out showing that young infants use statistical mechanisms to exploit the distribution of patterns heard in speech during the early stages of language development (Eimas et al., 1971; Jusczyk et al., 1993; Best et al., 1995; Saffran, 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 & Trehub, 2005).

The following sections will discuss the key statistical processes thought to be employed by early language learners, and the irreversible, dynamic effects this has on their perceptual abilities.

3.1.2 Statistical Learning Mechanisms

There have been many cognitive experiments carried out with young infants that attempt to solve the 'nature' vs 'nurture' argument (Saffran, 1996; Saffran et al., 1999; Kirkham et al., 2002;

Seidenberg et al., 2002; Saffran, 2003; Anderson et al., 2003; Thiessen & Saffran, 2003; Kuhl, 2004; Hannon & Trehub, 2005). Saffran (1996) carried out experiments on 8-month old infants showing that they use the statistical information in speech as an aid for word segmentation. The stimuli being used was an artificial language consisting of four three-syllable nonsense words which were fed to the infant in random order as a continuous speech stream. The words were synthesized to remove any prosodic features that the infant could use as boundary cues. After only two minutes of familiarisation with the speech stream, the infants were able to segment words using only the transitional probabilities between speech sounds. The results show that preverbal infants possess a powerful method of exploiting the statistical regularities found in speech. Saffran (1996, 2003) concludes that this may due to an innate general statistical learning mechanism; this view is similar to Jeff Hawkins (2004), who also quotes Mountcastle (1978) in his book 'On Intelligence', stating that the same computational tool is used for processing all sensory input. Saffran et al. (1999) further strengthens this hypothesis by carrying out a similar experiment on preverbal infants and adults, replacing the speech stream with a non-linguistic tone stream containing 'tone words'. The statistical distribution of the short tone sequences was the only segmentation cue available to the listeners. The results seem to indicate that the same learning mechanism is employed for both linguistic and non-linguistic input.

Moving away from the auditory domain, and inspired by Saffran's (1996, 1999) results, Kirkham et al. (2002) wanted to show that the theory for a general learning mechanism was also present in the visual domain. This was achieved by creating a similar experiment to Saffran's (1996, 1999) which showed that preverbal infants are able to learn patterns of visual stimuli with very short exposure.

There are theories stating 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 (Seidenberg, 2002). The authors also raise an important issue asking why animals are unable to learn language even though it has been shown that they possess statistical processing skills. They try to answer this question, hypothesising that there are many levels of hierarchical statistical information in language and that it is too complex. Or, animals do not possess innate grammatical knowledge like we do.

3.1.3 Perception of Phonetic contrasts

Preverbal infants younger than 8-months are able to discriminate native and non-native phonetic contrasts with equal ease (Aslin et al., 1981; Best et al., 1988; Eimas et al., 1971; Saffran et al., 1996; Werker & Tees, 1999). But, by the end of their first year, infants, like adults, lose this ability by the age of 12-months for many non-native contrasts (Sheldon & Strange, 1982; Trehub, 1976; Werker et al., 1981; Werker & Tees, 1984). This decline in non-native phonetic discrimination is gradual, with the effects of language specific exposure already visible in vowel discrimination at 4-months of age (Kuhl et al., 1992; Polka & Werker, 1994; see Anderson et al., 2003, for a summary), while consonant discrimination declines at a later stage.

However, Kuhl et al. (2003) show that it is possible for the infant to reacquire the phonetic contrasts of a foreign language through face-to-face interaction; this is not true for prerecorded exposure. With longer training, adults are also capable of improving their discrimination abilities (Logan, Lively & Pisoni, 1991; Werker & Polka, 1993).

Experimental data suggests that phonetic categories are formed by the statistical properties of the speech that the preverbal infant is exposed to. Maye et al. (2002) have shown that the statistical distribution of speech sounds in the input language affects the language learners' discrimination capability. Anderson et al. (2003) confirmed that the frequency of speech sounds in the language input determines the order in which native categories are acquired and non-native contrasts are lost.

Kuhl (2004), states that this transition, from universal to language-specific speech perception skills, occurs at 8 months of age and involves neural commitment. The neural networks in the brain commit to the patterns of input sound heard during the beginning of a young infant's life, optimising a kind of 'mental filter' for its native language. This native language neural commitment process helps the infant acquire all the different phonetic units of its language before starting to learn words.

Results from recent studies have helped enforce this hypothesis (Stager & Werker, 1997; Maye et al., 2002; Hannon & Trehub, 2005; Zhang et al., 2005). Stager & Werker (1997) carried out simple experiments indicating that infants re-organise their perceptual sensitivities when learning to associate meaning to words. The authors habituated 8 and 14-month-old infants with phonetically similar nonsense words 'bih' and 'dih' to two visual objects. The experiments tested the infants' phonetic discrimination by switching one of the nonsense words with the other object. If the infant looked longer at the switched image than usual then it had successfully discriminated

the fine phonetic detail separating the two words. The results showed that the 14-month-old infants could not distinguish the phonetic difference when the word was switched, whereas 8-month-old infants could. Further results by the authors support this hypothesis; that infants only lose the ability to discriminate the finer phonetic information when attempting to associate meaning to words.

This also coincides with infants' decreasing ability to discriminate non-native phones. U-shaped learning occurs, as this ability seems to increase when word-learning task becomes easier at 3 years of age. Functional reorganisation could be fundamental to understanding early language development.

Recent studies by Zhang et al. (2005) have also shown strong results supporting the native language neural commitment hypothesis. Their experiments were carried out on American and Japanese listeners using the stimuli /ra/ and /la/ which are distinguishable in English but not Japanese, where they are considered the same. Using magnetoencephalography, it was also proven that processing non-native speech sounds is considerably less efficient than native speech sounds. Magnetoencephalography (MEG) is used to measure neural activation in the brain and helped analyse the listeners' perception of the speech sounds. Japanese listeners showed a larger area of neural activation with longer durations of activity when discriminating non-native speech sounds than native speech sounds.

Hannon & Trehub (2005) help reinforce the argument for neural commitment and that it is an integral part of a general statistical learning mechanism, showing that statistical language learning theories also apply to music. The authors present evidence for neural commitment to musical rhythms of the listener's native culture. 6-months old infants show a culture-general responding to musical rhythms, which develops into an adult-like, culture-specific, response by the age of 12-months. However, further tests revealed that the 12-month old infants were capable of learning the musical rhythms of other cultures with little exposure, whereas adults had much more difficulty.

Recent behavioural data gives strong evidence for a general statistical learning mechanism, and that neural commitment is a key process for creating a robust and stable communicating system; allowing the gradual emergence and deep-rooting of internal representations of the patterns within the language learners world.

3.1.4 Dynamic Systems View of Language Acquisition

An increasingly popular view, of developmental researchers, is that the brain is a complex dynamic system and behaviour is emergent through self-organization. *"Development, then, can be envisioned as a changing landscape of preferred, but not obligatory, behavioural states with varying degrees of stability"* (Thelen, 1995). This view of development, as a constantly evolving landscape, challenges previous cognitive views that language development occurs as a sequence of innate language skills that are acquired at discrete, arbitrary time-steps.

Behaviour is classed into more or less stable attractor states and changes between these states have a non-linear relationship with environmental input. The timing of developmental changes is controlled by variation in the control parameters, body or environmental changes, rather than some kind of internal clock. Thelen (1982) strengthened this theory, overturning the previously held belief that developmental changes were due to cortical inhibition, by proving that the stepping reflex in newborns disappears due to an increase in non-muscular body mass and then reappears when the legs are strong enough again. This sparked further research into the application of dynamic systems theory to other motor skills, such as the development of motor skills required to reach for an object (Savelsbergh and Van der Kamp, 1993).

Learning can be seen as a shift or bifurcation into a new attractor state by the destabilisation of older stable states (Thelen, 1995). The behaviour of the system becomes more complex with age, with the formation of multiple attractor states. These wider areas encompass certain categories of actions such as walking, jogging and sprinting. Thelen says that the view of development as an evolving landscape is not supposed to prescribe behaviour, but represent a probability of behaviour depending on the control parameters. Figure 1 is a diagram of the attractor landscape for the acquisition of speech production skills of an infant as envisioned by Muchisky et al (1996).

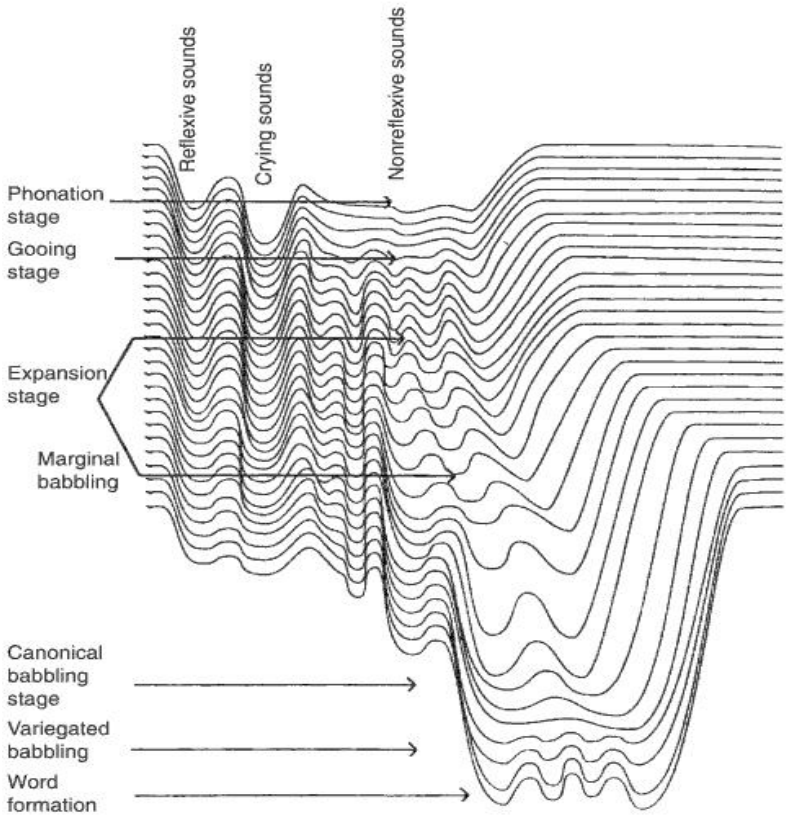


Figure 1. Diagram of the evolving speech production attractor landscape (Muchisky et al., 1996)

The three dimensions represent a) time, b) emergent behaviour, and c) the relative stability of the system at any point in time. Each attractor well is a state of behaviour. The deeper the well of an attractor, the more stable the system is when in that state.

The theory also states that if the system is at or near a phase transition then it will be unstable and fluctuate, have a higher sensitivity to change and return to the attractor state more slowly. Kelso (1995) carried out some finger wagging experiments supporting this theory. Subjects were able to stably reproduce only two simple rhythmical finger movements: in-phase and anti-phase. From the anti-phase condition subjects would transition to in-phase movement at a critical frequency, beyond this point only in-phase is stable. Haken (1985) created a mathematical model, which was able to accurately predict the results of Kelso's finger wagging experiments. The model was able to describe the phase transitions, reproduce the results with minor inference and achieve accurate predictions of the times taken for the phase transition from anti-phase to in-phase to occur. This shows that the dynamic systems perspective can be used to predict behaviours of a system with varying control parameters.

Aslin (1993) and Port (2000) both agree that this perspective is useful for solving general problems but argue that the range of different cognitive behaviours is too great for these kinds of models. Aslin (1993) also states that it is difficult to incorporate non-observable influences such as motivation.

It is becoming commonplace to analyse connectionist models, particularly recurrent neural networks, as dynamic systems. Connectionism and dynamic systems are compatible and having a dynamic perspective of connectionist models may help deepen our understanding of them.

3.2 Development of the auditory system

3.2.1 Hearing in the Womb

The womb is a constantly noisy place. The fetus will be subjected to sounds from the mothers' digestive system, blood rushing round the body. Research has shown that young infants will become calmer when exposed to intrauterine sound and music that the infant was played during pregnancy (Federico, 1999). Federico (1999) suggests that the fetus will give extra attention to his mother's voice, as it is so different from its amniotic environment. At around 8 weeks it will start to develop ears, fully forming by about 24 weeks. Although the fetus hasn't developed a full auditory system by 18 weeks it can already hear the mother's heartbeat and blood passing through the umbilical cord. At 25 weeks it can hear the mother's and other nearby voices and it is believed that by 27 weeks the fetus can even recognise the voice of the mother and partner (Federico, 1999).

3.2.2 Fetal Memory

Tests by Hepper (1996) help confirm the idea of fetal memory. 32 weeks was the earliest observations of fetal memory through classical conditioning and 36 weeks with habituation. Previous experiments have shown habituation as early as 22-23 weeks supporting the existence of fetal memory (see Federico, 1999, for a review of fetal memory experiments).

It is thought that the existence of memory in utero is to help recognition and attachment to the mother (DeCasper & Spence, 1986; Hepper et al., 1996). Nazzi et al. (1998) hypothesise that this may be a vital property of early language acquisition. They carried out experiments showing that newborns, as early as five days old, had a preference for the rhythm class of their mother's native language.

Brent Logan (1987) believes that if the fetus were subjected to more complex rhythms, at a critical period, then its neural connectivity would grow to accommodate these sounds, helping the infant to derive structure from more complex input. An interesting question arises; do the rhythmic intrauterine sounds kick-start the fetus's neural connectivity which gives rise to a general statistical mechanism as mentioned by many developmental psychologists, such as Saffran (1996) and Kuhl (2003).

3.3 Prosody – an Aid for Language Acquisiton

In order to learn language, infants must discover the basic lexical units within speech from their environment. As yet, there does not seem to be a common, universal cue that can be used for all the languages in the world. A common misconception is that words, in natural speech, are separated by silence periods, analogous to the white space between words in written form. When silences are present, they can often be misleading and occur within words.

3.3.1 Focal Attending & Synchronisation

It is thought that the rhythmic properties in speech help us communicate with our environment (Port et al., 1998; Drake et al., 2000; Clark, 2002; Jungers, 2002; Koreman, 2006; Zatorre et al. 2007).

Every language is believed to have a basic level of prosodic regularity (Jungers et al., 2002). With regularity we are able to predict the occurrence of future events, aiding our focal attendance. Clark (2002) mentions the 'readiness principle', which states that we are much quicker and more accurate if something occurs when we are expecting it. In order to do this when communicating in every day speech it is believed that we need to synchronise with our audience (Port et al., 1998; Drake et al. 2000; Clark, 2002; Jungers et al., 2002).

Listeners have a tendency to hear the regularity of stress patterns in speech. The English language has alternating stresses in syllables, contrast of weak and strong, which we use during segmentation (Cutler, 1994). Cutler (1994) references experiments with French and Japanese speakers showing that they use the rhythm class of their native language, syllable and mora-timed, to aid segmentation.

Newborns have the ability to discriminate and categorise simple rhythms; it has been shown that they favour a 2:1 ratio (Drake et al., 2000). With age our rhythmic attendance becomes much more reliable and we are able to synchronise to more complex ratio's (Drake et al., 2000). Patel (2003) looks at the functional and neural architecture of music and language. Neuroimaging tests

reveal that there is an overlap in music and language processing. This brings up an interesting question, when is a newborn able to distinguish music from language? And does this have anything to do with this processing overlap? The author suggests that studying the domains comparatively will help give us a more complete view of the workings of the mind than studying either alone.

3.3.2 Rhythm Class Hypothesis

The rhythm class hypothesis states that newborns are able to extract prosodic information to help classify languages into three distinct classes based on rhythmic properties[1]; syllable-timed, stress-timed and mora-timed. The R hypothesis predicts that newborns are able to discriminate languages from two different rhythm classes but not if they both belong to the same one (see Nazzi et al., 1998; Grabe & Low, 2002, for a more detailed definition). Nazzi (1998) carried out experiments on newborns up to 5 days old, using the variance of their sucking rate for evaluating their performance. The results support the R hypothesis showing that newborns could discriminate two languages from different rhythm classes without ever hearing them before but could not discriminate two non-native languages belonging to the same rhythm class as their native language. It is interesting to note that the author low pass filtered the speech signal in order to reduce the linguistic information whilst keeping its prosodic properties.

Further experiments have been carried out supporting this hypothesis. Ramus et al. (1999) agree with Nazzi's results but believe that there may be room for more rhythm classes. Grabe & Low (2002) observe similar results but state that there is much overlap between stress-timed and syllable-timed languages, for example they found that Catalan and Polish is a mixture of both. Inspired by Kelso's (1995) finger wagging experiments, Kitahara (2000) examined the effects of rhythm in English speech. Subjects repeated 3-5 syllable English phrases to a metronome, to promote rhythmic patterns. The results support the rhythm class hypothesis and the author suggests that this is due to internal dynamic attractor structures.

3.3.3 Visualising Rhythm

Todd & Brown (1996) have created a computational model that recovers the rhythm structure from an acoustic signal and visually represents the output as a 'rhythmogram'. The model is able to find hierarchical structure from the phoneme level up to the structure of a complete poem. The authors use their results to compare two common methods of describing speech rhythm, the

[1] For a detailed definition of rhythm in speech see Evans (1986)

binary tree (Liberman & Prince, 1977) and time-span reduction (Lerdahl & Jackendoff, 1983). The rhythmogram model supports the time-span reduction method; it is a more perceptually realistic representation than binary tree's that conforms to standard metrical phonology.

3.4 Computational Models of Language Acquisition

The focus of this section is on algorithms that could be used by early language learners for speech segmentation and word discovery. 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 they must:

1. Start with no prior knowledge of any language
2. Learn in a completely unsupervised manner
3. Segment incrementally

3.4.1 Neural Networks

Cognitive scientists once held the view that the brain is like a computer, that we are born with an innate program that allows the brain to compute information. However, advances in neuroscience have helped this metaphor become more and more redundant, showing us that there are quite big differences between the two systems. A deeper understanding of the processing and architecture of the brain has allowed computer scientists to model it through 'Artificial Neural Networks'.

There is a lot of interest in these models as they perform well at tasks where we outperform computers, such as recognition and learning (NETtalk, developed by Sejnowski and Rosenberg, 1987; Linsker's self-organising perceptual network, 1988), which is of particular interest to developmental psychologists. Linsker (1988) developed a neural network that develops, through hebbian learning, feature detectors in a layered visual system. The model shows similarities to that of feature detector cells found in the mammalian brain.

Elman (1992) created a recursive neural network able to learn about the sequential dependencies between words in sentences. The model hierarchically clusters words into similar grammatical types and then finds structure within each class. Representations of elements degrade as the input sentences become more complex with hierarchical structure, such as embedded clauses, which is similar to the difficulty that human language users face. The network weights act as attractors in a state space, enabling the system to respond sensibly to novel input.

Neural networks can also be used to simulate brain damage. Hinton et al. (1993) simulate the reading capabilities of people with deep dyslexia. They use a neural network that is capable of reading words, using a small vocabulary, and selectively remove connections to the units representing physical and functional attributes. They found that the model not only reproduced the visual and semantic errors of deep dyslexia but also some of the more subtle characteristics. This supports the theory that semantic memory can be understood as an attractor space and that neural networks can be used to observe the behaviours and functions of the brain.

Although neural networks, at the moment, can only solve simple problems, they are able to give us an insight into the workings of the brain. Many of the models created to simulate simple human behaviours have shown the same development properties that humans have; for example the U-shaped learning curve we see in early language acquisition (Rumelhart & McClelland, 1986, modeled past tense learning), the stage-like transitions in development which is believed to arise from a gradual learning process (McClelland & Jenkins, 1991, devised a model for Seigler's (1981) balance scale task showing that children fell into one of four stages) and they are also good at making generalisations through distributed representations (McClelland et al., 1995, created a simple semantic memory model. The results agreed with Keil's (1979) infant developmental progression findings, that distinctions between concepts proceed from coarse to fine). Generalisations also make the networks resilient to local damage within the network and noise in the input.

3.4.2 DP-ngram

The DP-ngram model was originally developed by Sankoff & Kruskal (1983) to find two similar portions of gene sequences. Nowell & Moore (1995) then modified the model to find repeated patterns within a single phoneme transcription sequence through self-similarity. The model uses dynamic programming to find partial matches, portions that are similar but not necessarily identical, taking into account noise, speed and different pronunciations of the speech. Traditional template based word spotting algorithms using dynamic programming would compare two sequences, the input speech vectors and the word template, penalising insertions, deletions and substitutions with negative scores. Instead, this algorithm uses quality scores, positive and negative, in order to reward matches and prevent anything else, resulting in longer more meaningful sub-sequences.

3.4.3 Segmental DTW

Dynamic time warping (DTW) is a dynamic programming technique used to accommodate temporal distortion within a sequence. Park & Glass (2007 & 2008) have successfully adapted this technique to find matching acoustic patterns between two utterances. The discovered units are then clustered, using an adjacency graph method, and used to describe the topic of the speech data. The experiments were carried out on a corpus of recorded academic lectures, with topics such as poetry, psychology and science.

The algorithm compares two utterances to find multiple warp paths with limited temporal distortion. The temporal constraint limits the path to a diagonal band. This diagonal band allows for a natural partitioning of the search space, with which they are able to find a single optimal alignment for each band using DTW (see fig 2). Each alignment is then refined to a single fragment by trimming areas of high distortion. Fragments that do not meet the minimum length constraint are discarded.

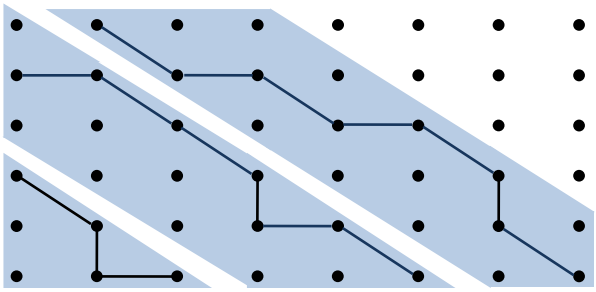


Figure 2. Optimal alignments for each diagonal temporal constraint

3.4.4 Statistical Word Discovery

Statistical Word Discovery (SWD) is an algorithm, developed by ten Bosch & Cranen (2007), that combines acoustic pattern discovery and clustering in order to discover words from low-perplexity speech data (connected digits from the Aurora 2.0 database).

The first stage of the algorithm is to automatically segment the speech. Boundaries are hypothesised by observing local differences of the feature vectors. A sliding window is used to search for local maximums above a threshold within the utterance. The second stage labels all the discovered segments using a k-means clustering algorithm (MatLab). The number of clusters was specified at 25 to cover all identifiable phonemes in the database (20). Then, each segment can be labeled with the integer number of the cluster it belongs to.

The final stage is the word discovery process. DTW is used to find the best-matching subsequence by comparing the label sequence of the current utterance against those of all past utterances. Each utterance includes abstract tags indicating which words are present (the tags do not contain any temporal or acoustic information). The algorithm then hypothesises the associated tags with the discovered subsequences.

The results show that words are emergent and that accuracy increases with experience.

3.4.5 Other Fields using DTW for Pattern Discovery

Dynamic programming techniques are not only resurfacing as a popular method for pattern discovery within speech data, but as a novel approach for analysing the structure of music (Lu et al., 2004) and recognising heart rhythm disturbances (Volkan & Selman, 2005).

The model developed by Lu et al. (2004) focuses on melody similarity; timbre similarity is suppressed in order to help find two similar melodies played on different instruments. After discovering local repeated patterns they are able to find global structure within the whole piece. This model uses an erosion and dilation method to enhance the significant repeated lines and remove short unwanted lines from the similarity matrix. This is a common operation used in grayscale image processing.

Volkan & Selman (2005) use DTW as a novel method for automatically detecting heart rhythm disturbances. Normal rhythm electrocardiogram templates are compared with electrocardiogram's of various arrhythmias (abnormal heart beat). The results show that with DTW analysis it is possible to detect and differentiate between different types arrhythmias.

3.4.6 NMF

Non-negative matrix factorisation (NMF) is a radical approach of modeling language acquisition (Stouten et al., 2007 & 2008; ten Bosch et al., 2007), originally developed as a method for learning the salient parts of a different faces from an image (Lee & Seung, 1999). The algorithm detects words from 'raw' cross-modal input without any kind of segmentation during the whole process and is therefore referred to as a *'word detection algorithm'*. The authors claim cognitive plausibility as the algorithm automatically detects words in an unsupervised fashion, and that the internal representations constantly evolve as more input is presented, referencing current behavioural data showing that language is an emergent property of a communicative learner and its environment.

The ACORNS year 1 database (see section 4.2 for more detail) was used for the experiments carried out by ten Bosch et al. (2007). The utterances are represented as a phone lattice of transitional probabilities. Each utterance is then summarised into an n-dimensional column vector, where n is every possible phone transition (n^2). The probability of every consecutive phone transition within the utterance is accumulated to give a non-negative weighted co-occurrence count. Each column vector (utterance) is appended into a $n \times m$ matrix, where m is the total number of utterances, to create a large input matrix V.

The NMF algorithm then factorises V into vectors W and H. The columns of W represent the recurrence of the different elements within the data, and the columns of H represent the elements that are active within each utterance. The W column can be seen as an abstraction of the information in V, coding recurrent speech fragments into to 'word-like' entities. Recognition of incoming utterances is carried out by comparing the columns of W, the learned speech representations, to predict the correct tags associated with them. It is important to note that this method is not able to predict where the key word lies within the utterance as all temporal information is lost during the factorisation stage.

Figure 3 shows plots of the key word tag recognition accuracy (if the learners' response is the same as the visual tag of the input utterance) against the number of utterances observed. Plot 3.a) displays the recognition accuracy for 8000 Dutch utterances recorded from 4 different speakers fed to the system in a random order, whereas for plot 3.b) the input has been split into 4 separate blocks for each speaker (new speaker at 2000, 4000 and 6000). The blue line shows the accuracy of correct responses divided by the total number of utterances heard, the red line is the accuracy for the past 50 utterances.

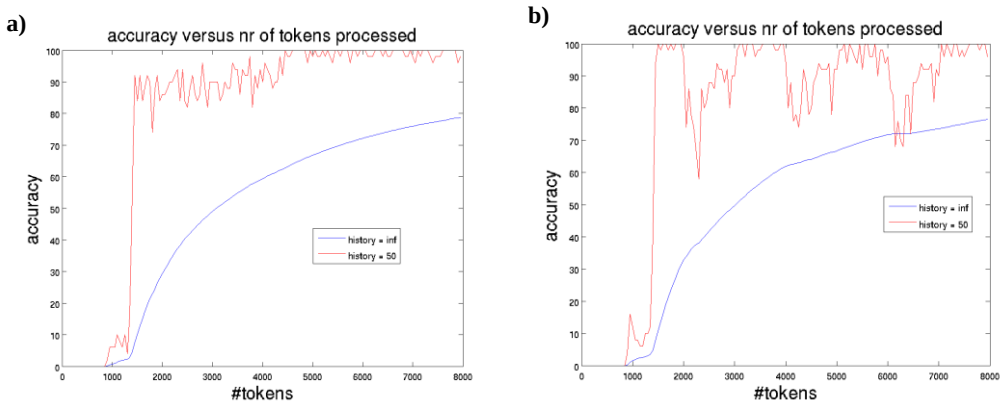


Figure 3. a) Plot of the key word tag recognition accuracy for 8000 utterances fed to the NMF system in a random order, b) fed as 4 separate speaker blocks

The results show that the NMF model is able to dynamically build and adapt internal representations of the key words. It can be seen in figure 3.a) that the learner model does not have any internal representations of the key words within the first few hundred utterances, but accuracy quickly rises once it has started to bootstrap learning. As more data is processed the representations become more accurate. The dynamically evolving internal representations can be seen in figure 3.b) as accuracy results dip for each new speaker block and the algorithm has to rebuild the speaker-dependent representations.

4. Research Conducted

4.1 ACORNS

ACORNS (Acquisition of Communication and Recognition Skills) is an EU funded project aiming to develop an artificial agent (Little Acorns) that is capable of acquiring human verbal communication skills. The main objective of the project 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 the project there are five main research areas:

4.1.1 Front-end Processing

This area will look at the development of new feature representations, in particular trying to encode features affecting speech recognition in phonetic and psycho-linguistic experiments.

4.1.2 Pattern Discovery

Little Acorns (LA) will have to begin life as a newborn, without any prior knowledge of basic speech units. Unlike conventional ASR systems, where they are pre-defined, LA will have to discover patterns from the continuous input that form these basic speech units. Behavioural data shows that infants exploit the repetitive nature of speech to discover these basic units, which are also constantly developing over time.

4.1.3 Memory Organisation and Access

As important patterns are being discovered they will need to be stored in memory. Theories suggest that there are three distinct types of memory; short term, long term and working. A suitable computational model of this memory architecture and plausible processes occurring within them will be developed.

4.1.4 Information Discovery and Integration

Research and development of efficient and effective techniques for retrieving the patterns stored in memory will be carried out.

4.1.5 *Interaction and Communication*

LA will be given an innate need to grow his vocabulary and communicate with the environment. Simple reinforcement learning methods can also be applied by rewarding LA with correct reactions and ignoring or repeating utterances in a different form for incorrect reactions (fig. 4).

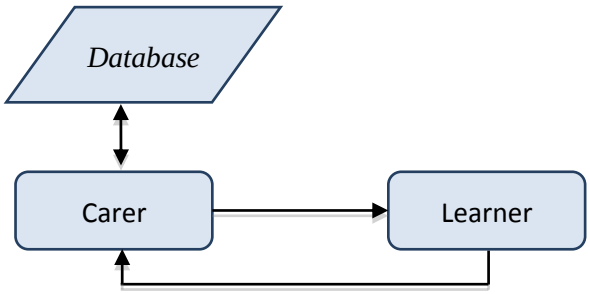


Figure 4. Simple diagram of carer and learner interaction/communication

The four areas of research above will be developed as modules that will inter-operate within an integrated architecture. The framework of the ACORNS architecture is shown in figure 5 and was designed to adhere to current cognitive theories of human communication and memory.

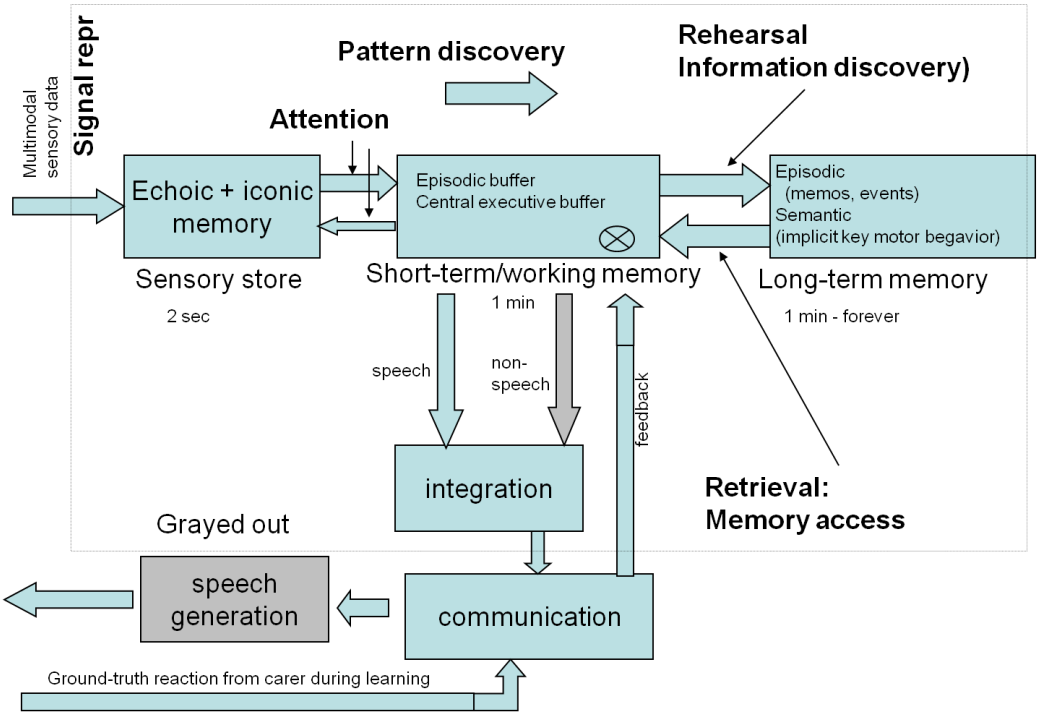


Figure 5. Framework of the ACORNS integrated architecture

4.1.6 Milestones for Little Acorns

The project has set three major milestones for LA. Each milestone corresponds to a year of the three-year project and also to the language development milestones that a young infant should reach in the same time. It is important to note that LA will possess general language acquisition abilities and should be able to learn any single or multiple languages. Therefore all the models being developed will be language independent.

1. LA must learn 10 words from simple repetitive speech produced by one or both of his 'parents'.

2. By the end of the second year LA should have learned a vocabulary of 50 words, along with sub-word units, from his parents and other speakers in his environment.

3. At the end of the project LA should have learned 250 words along with references to concepts stored in memory and relationships between the different concepts.

4.2 ACORNS Speech Corpus

By the end of the project ACORNS will have recorded a corpus of 3 databases, one for each year of LA's life, in 3 different languages (English, Dutch and Finnish). Within the project I have been given the responsibility for organising and recording the English corpus. The first year English corpus has been successfully completed and consists of 4000 utterances. 100 different sentences, each containing one of ten key words, were repeated 10 times and recorded with 4 different speakers (2 male and 2 female). A few examples are listed (key words have been underlined):

1. *"Ewan sits on the couch"*
2. *"What matches this shoe?"*
3. *"Daddy comes back"*
4. *"A shoe is a fashion item"*
5. *"Finally Ewan is there!".....*

The ACORNS database will be used as cross-modal input. Each utterance has an abstract visual tag attached to it. As an example, the utterance *"What matches this shoe"* will contain the tag referring to *"shoe"* (fig. 6). The tag does not give any location or phonetic information about the key word within the utterance.

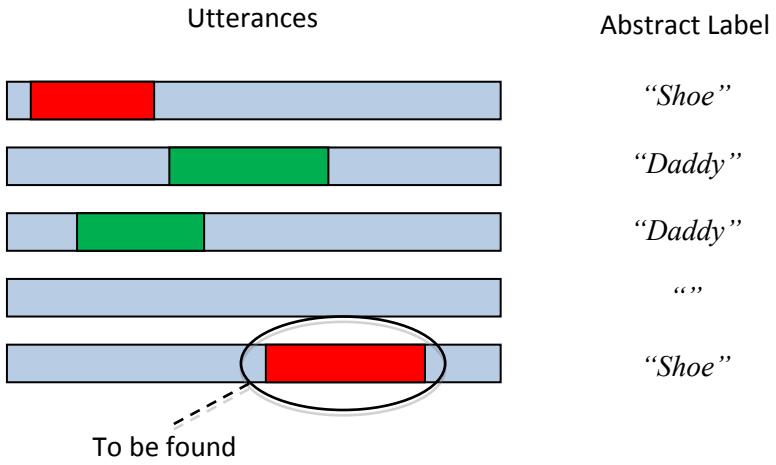


Figure 6. Utterance structure with associated abstract label

The second year corpus is currently being recorded. As stated in section 4.1.6, LA must learn 50 words by the end of his second year from his parents and other speakers. The second year corpus consists of 2000 utterances with a total of 52 key words; each key word is repeated at least 74 times. The 2000 utterances will be recorded for each of the 4 original speakers, and an additional 6 speakers will record a subset of 600 utterances each.

The sentence structure is more complicated, as there are multiple key words within each utterance (between 1 and 3), and there are also simple dialog sentences. The dialog sentences have been included to allow for interaction/communication experiments between LA and the carer (see section 5.1.4 for research plans). Some examples are listed (key words have been underlined):

1. *"Here is a small toy and a dog"* *(regular)*
2. *"Where is the round fish?"* *(regular)*
3. *"Mummy takes the square toy"* *(regular)*
4. *"No, I mean dog"* *(dialog)*
5. *"Fish!"* *(dialog)*

The problem of associating abstract semantic tags with each utterance in the first year database was simple, as there was only ever one key word in every sentence. For the second year database, representing the key words has become an issue and is currently being resolved. The simplest method that has been proposed is to semantic feature vectors for every utterance. Each key word will have its own semantic tag which will have a binary state (present or not present). The

problem arises when key words consist of multiple tags; discussions are being carried out within the group, with which I play an active role.

The third year database is in the planning stage and will depend on the outcomes and problems encountered during the second year experiments.

4.3 Pattern Discovery

The main area of research that is being pursued as part of the ACORNS project and to subsequently obtain a PhD thesis topic is pattern discovery. However, this does not mean that the other areas will be ignored. The DP-ngram model under development already provides an end-to-end system that includes, although simplified, development and research relevant to all five areas of the project.

4.3.1 Baseline - NMF

The first and currently only other complete end-to-end system within the project is the Non-Negative Matrix Factorisation (NMF) approach. This approach is being used as the baseline for all other pattern discovery methods. Successful key word discovery results are achieved with this method but, as described in section 3.4.6 of the literature review, it cannot give any temporal information.

4.3.2 Using prosody with NMF

The researchers within the project working on the NMF algorithm wanted to investigate whether additional prosodic cues, rhythm and pitch, would increase their accuracy results. Having an interest in this area, and looking to employ prosodic cues within my own research, I took on the responsibility of modifying the NMF algorithm to handle prosody and carried out the experiments. It was hypothesised that the addition of prosodic cues to the input stream would increase the accuracy results at a faster rate.

4.3.2.1 *Rhythm Vectorisation*

The first prosodic cue to be employed was rhythm. The 'Rhythmogram' model (Todd & Brown, 1996) is used to detect peaks within the speech signal (fig. 7), looking over a range of different time constants in order to derive hierarchical structure of the onsets of individual events.

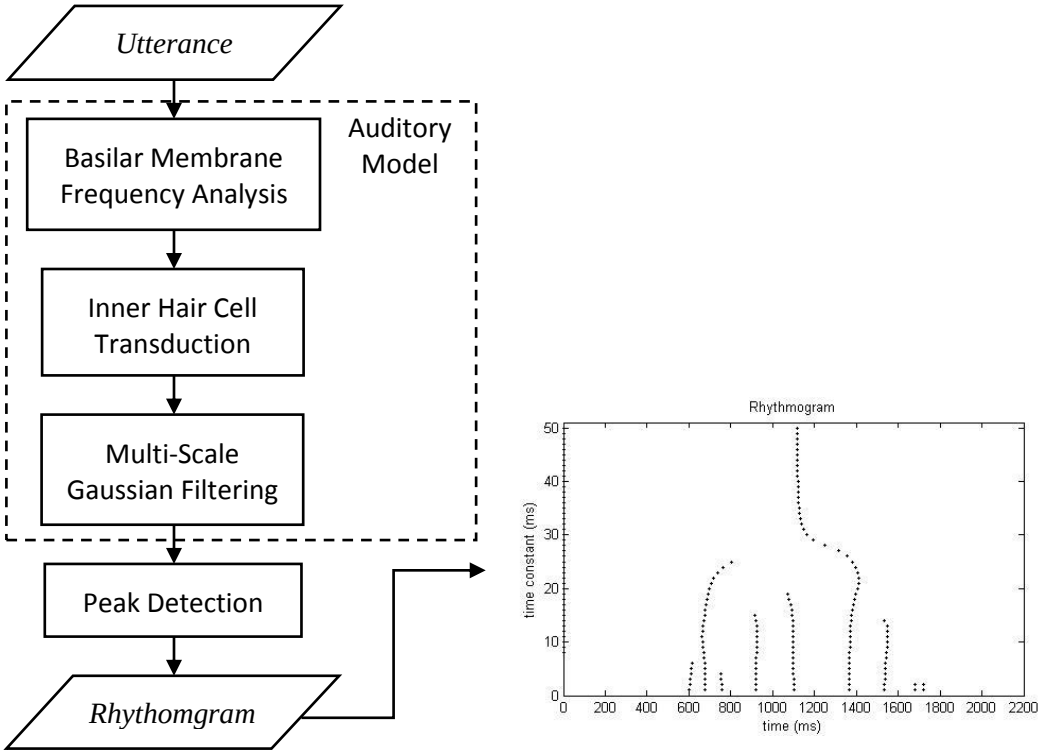


Figure 7. Rhythmogram processes and output

The utterance is fed through an auditory model where the auditory nerve response is modelled and pooled across frequency, then passed to a multi-scale Gaussian low-pass filter system. The rhythmogram output is derived by finding the peaks in the low-pass response or zero crossings of the 1st derivative. An example is provided in figure 7 (right). Time is displayed across the horizontal axis and the different time constants used are on the vertical axis.

The NMF technique requires the data to be decomposed and made available in the form of a data matrix. This implies that an additional stream of information must be encoded in terms of a sequence of vectors. To vectorise the rhythm for NMF we exploited the manner in which syllabic-type events converge to higher levels within the utterance. We achieved this by creating event strings across all channels with varying time constant. So, for each event we collect the time distance of the nearest peak in each consecutive time constant channel, where the root

position is channel 1 (time constant = 1ms). Figure 8 shows the event strings within an example utterance.

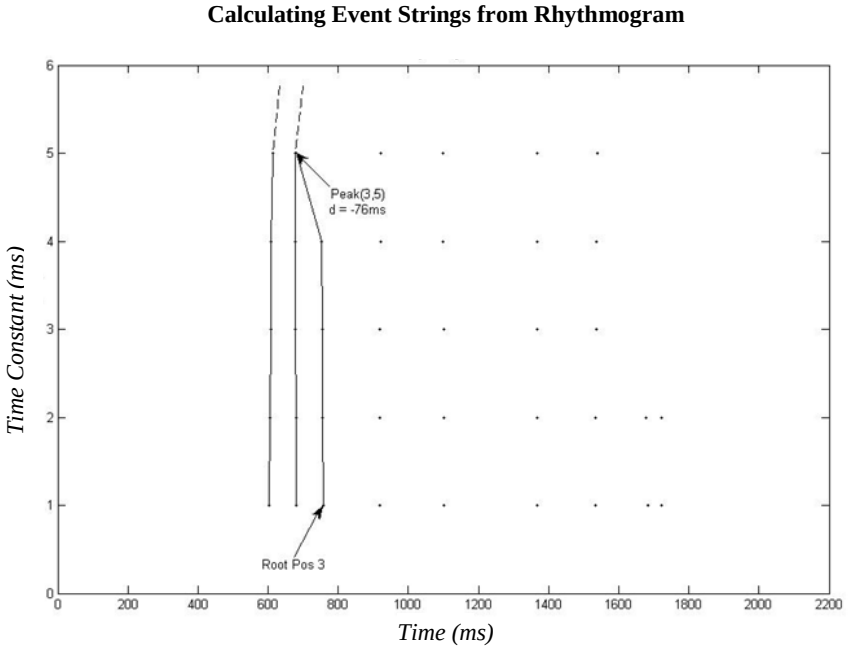


Figure 8. Syllabic-type event string within utterance

The time distances are then quantized and labelled in terms of a value between 1 and 20 (see table 1), with finer resolution for shorter distances. The result is a histogram in terms of the labels 1-20.

Distance (ms)	Mapping	Distance (ms)	Mapping
0	1	-1	11
1	2	-2	12
2	3	-3	13
3	4	-4	14
4	5	-5	15
5→7	6	-6→-8	16
8→10	7	-9→-10	17
11→20	8	-11→-20	18
21→50	9	-21→-50	19
50→∞	10	-51→-∞	20

Table 1. Quantisation of time distances between peaks

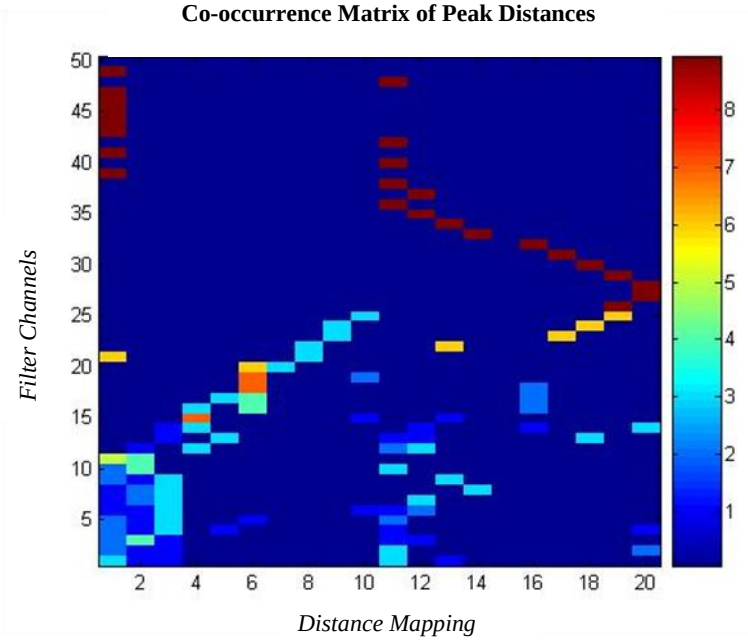


Figure 9. Co-occurrence matrix of quantised peak distances

The final stage of the rhythm vectorisation process is to create a co-occurrence matrix of the quantised peak distances within the utterance (fig. 9). This stream can now be appended to the NMF input.

4.3.2.2 Rhythm VQ results

The first experiment carried out was to simply append the rhythm stream to the spectral input. Figure 10.a) shows the key word recognition accuracy using the same experimental set-up as described in section 3.4.3. The original result (baseline) is in blue and is an average of five attempts to correctly guess the associated key word tag of 1000 incoming utterances. The red plot is the average of five attempts with the rhythm stream appended to the current speech VQ labels. It appears that there is not much difference between the two, with most of the variance occurring in the first 500 utterances.

It was hypothesised that using the rhythm stream would help NMF learn key words faster. Looking at the first 150 utterances it can be seen that the prediction of key words is consistently better than the baseline for all five attempts (fig. 10.b)).

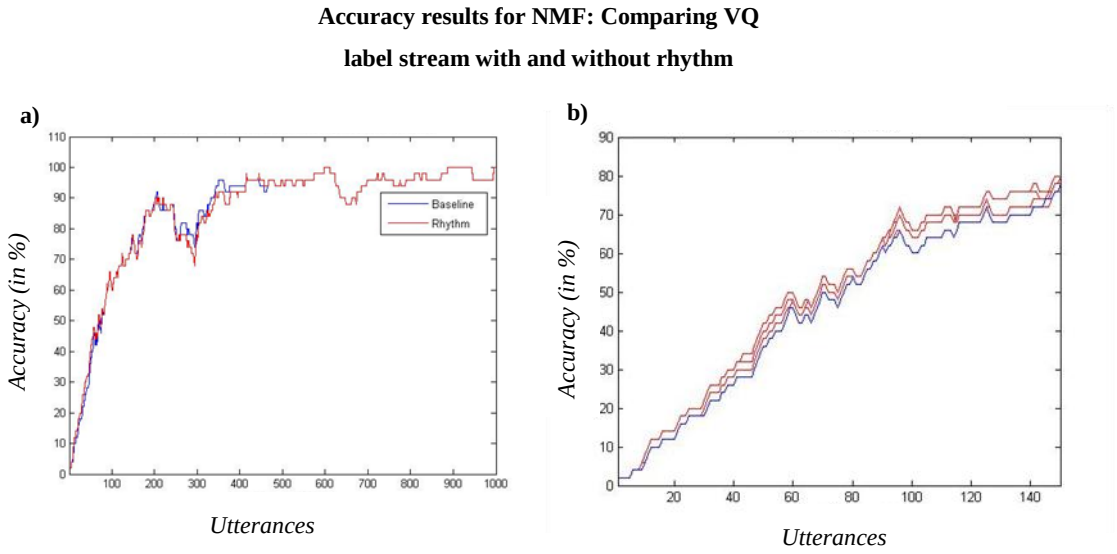


Figure 10. plot of the key word tag recognition accuracy with and without rhythm a) for 1000 utterances b) for 150 utterances

The next set of experiments was carried out with varying weights of the rhythm stream. The plots in figure 11 show the accuracy (%) difference from the baseline taken every 50 utterances for each weighting. The only weighting that is consistently better than the baseline is a factor of 0.1, where there is an improvement of nearly 2% after the first 100 utterances. The biggest accuracy difference is during the first 300 utterances, after this period they all tend towards baseline.

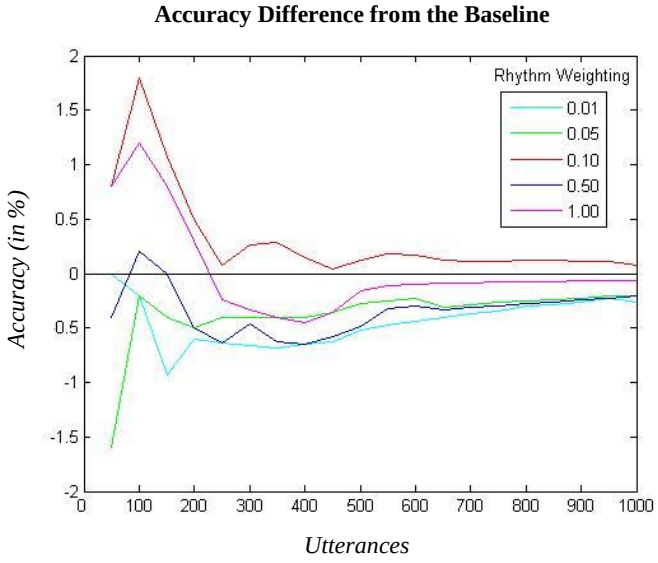


Figure 11. Accuracy difference from the baseline with different rhythm weightings for 1000 utterances

4.3.2.3 Pitch Vectorisation

The second prosodic cue under investigation here is pitch. To vectorise the pitch for NMF we devised a procedure that relates pitch movement at different time instants within the utterance. This is achieved by calculating the delta's of the pitch contour and then accumulating the co-occurrence counts of a user defined lag (tau) value within this stream.

This procedure consists of three steps. In step 1, the pitch contour is calculated. There are two pitch extraction methods being used and compared in these experiments; the ACORNS 'Pitch Estimator' and the 'Subharmonic' function which uses dynamic programming smoothing. Both methods carry out speech detection processes to give voiced/unvoiced values which are used to exclude unvoiced regions.

In step 2, the frequency range of the pitch contour within the voiced regions (i.e. regions consisting of consecutive voiced frames) is then quantized into a 50-channel filter bank (fig 12).

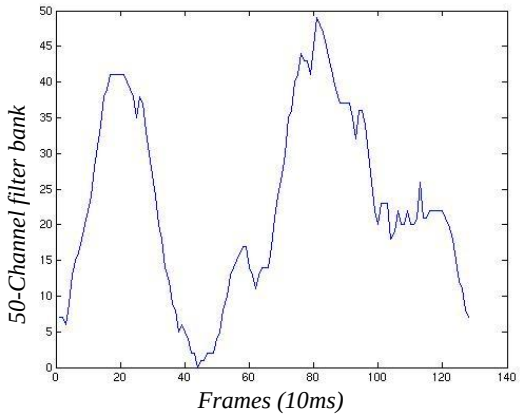


Figure 12. Island of the voiced pitch contour after quantisation

We are now able to accumulate counts of lag-tau co-occurrences using:

$$coocM(q_t, q_{t-tau}) = coocM(q_t, q_{t-tau}) + 1 \quad (1)$$

Where,

$coocM$ = Co-occurrence Matrix
q_t = Label of the quantised pitch stream at time t
tau = Lag of n frames of 10 ms

Figure 13 shows the plot of the co-occurrence matrix within the example utterance. This is the data matrix that can now be appended to the NMF input stream.

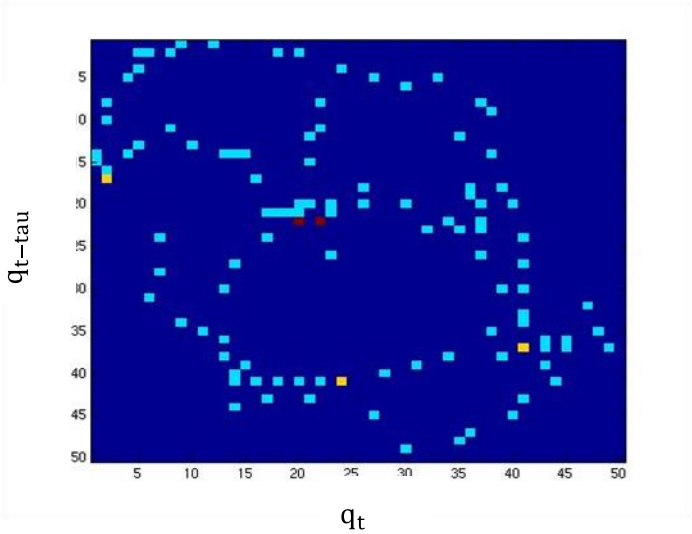


Figure 13. Co-occurrence matrix of pitch movement within utterance

4.3.2.4 Pitch VQ results

The key word recognition accuracy difference from the baseline is plotted in figure 14.

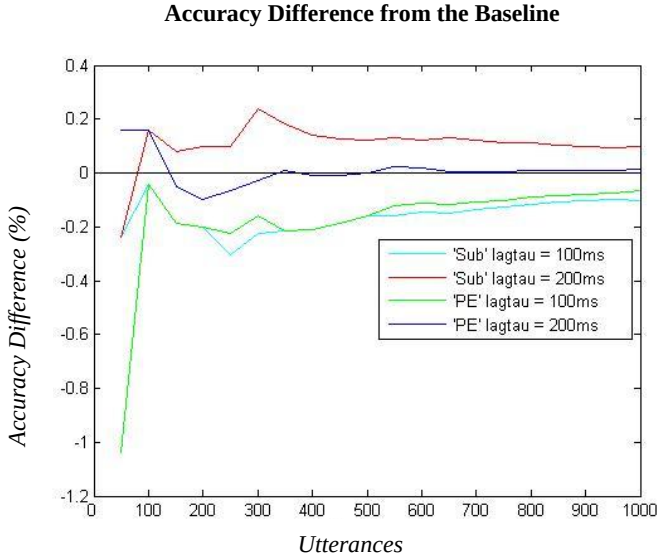


Figure 14. Accuracy difference from the baseline for pitch VQ using 'Subharmonic' and 'PitchEstimator'

The results show that using a longer lag-tau greater accuracy while only the 'Subharmonic' method for calculating the pitch contour was better than the baseline.

4.3.2.5 Conclusions

As hypothesised, the addition of prosodic cues as an aid for word detection helped raise accuracy results during the early learning period. Rhythm had more of an impact than the use of the pitch contour with accuracy almost 2% better than the baseline during the first 100 utterances. After this period the accuracy tends towards the baseline for both cues. Calculating the pitch contour with dynamic programming smoothing slightly enhanced the results, but at the expense of computational complexity.

4.3.3 Acoustic DP-ngram

The two variations of the DP-ngram method (Sankoff & Kruskal, 1983; Nowell & Moore, 1995) described in section 3.4.1 were used to find similar repeating patterns within a simple sequence of discrete symbols. Expanding on these methods, I have developed a version 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 dynamic programming (DP) allows this algorithm to accommodate temporal distortion through dynamic time warping (DTW). Initial test results show that there is significant potential with this approach, as it segments in an unsupervised manner, therefore not relying on a predefined lexicon or acoustic phone models.

What the acoustic DP-ngram model does:

- Finds similar repeating patterns directly from the acoustic signal
    - Works on the assumption that common words and phrases are acoustically similar, finding low distortion alignments between spectral representation of different regions of time in words and phrases
- Uses Dynamic Programming techniques to find partial matches
    - Partial matches allows the algorithm to discover alignments that are longer and more meaningful
- Clusters similar alignments
    - Alignments with the same underlying meaning will be grouped together. The centroid of each group is the ideal representation, which will constantly evolve and become more accurate as more data is processed

- Output can be used to create a symbolic representation of audio stream
    - Compresses the audio stream which can be used to feed back into the algorithm to find hierarchical repeating structure over a longer time range

4.3.3.1 Acoustic DP-ngram Method

Figure 15 shows the simplified architecture of the acoustic DP-ngram method. There are four main stages to the basic process:

1. Two utterances are fed to the DP-ngram algorithm as two sets of feature vectors.
2. A frame-to-frame distance matrix is calculated.
3. Accumulative quality scores are calculated for successive frame steps.
4. Local alignments are then discovered within the quality matrix.

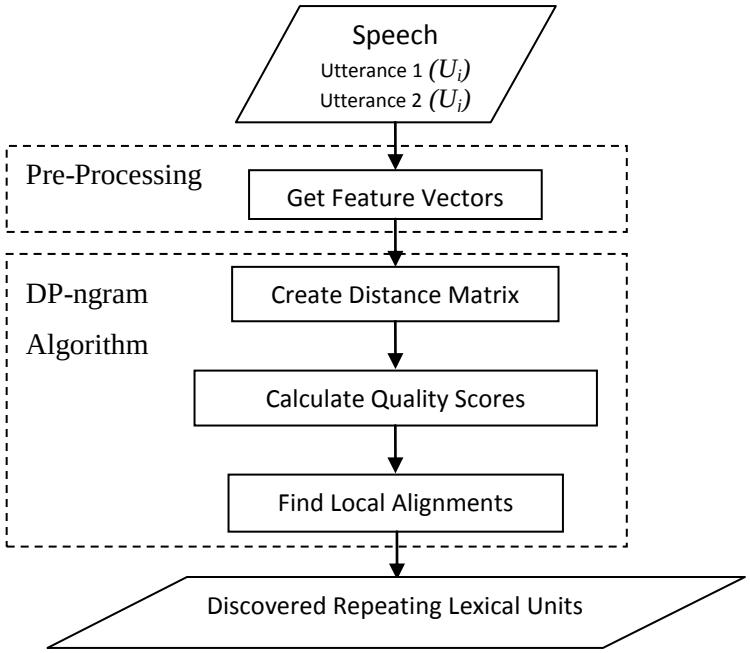


Figure 15. Acoustic DP-ngram architecture

4.3.3.2 Feature Vectors

The test data being used is utterances of speech that have been specifically recorded by the ACORNS project during the first year (as detailed in section 4.2).

The ACORNS MFCC front-end has been used to parameterise the raw speech signal. 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 (ΔΔ) coefficients. The front-end also allows the option for cepstral mean normalisation (CMN) and cepstral mean and variance normalisation.

4.3.3.3 Distance Matrix

A local-match distance matrix is then calculated by measuring the cosine distance between each pair of frames (v_1, v_2) from the two sequences, which is defined by:

$$d(v_1, v_2) = (v_1^T \times v_1)/(||v_1||^T \times ||v_2||) \quad (2)$$

Where,

$T =$ Transpose

Figure 16 is a plot of the frame-frame similarity for two utterances. The distance measure is on a scale of 0-1, where 1 is an exact match.

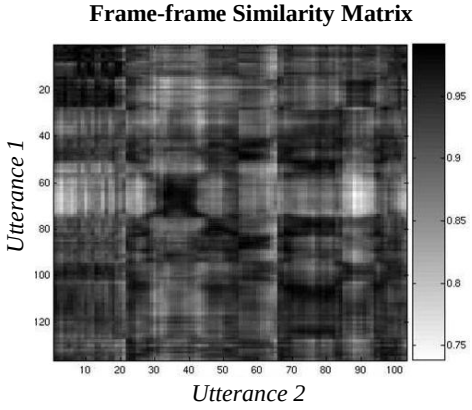


Figure 16. Frame-frame similarity matrix between two utterances

4.3.3.4 Quality Scores

Traditional DP template based recognition systems use negative scores to penalise insertions, deletions and mis-matches to find the shortest distance. This method uses positive scores to reward matches and negative scores to discourage anything else, allowing us to find longer and more meaningful alignments.

The following recurrence is used to find all quality values $q_{(i,j)}$:

$$q_{ij} = max \begin{cases} q_{i-1,j} + \left(s(a_i,) \times \left(\left|d_{i-1,j-1}\right| - 1\right) \times q_{i-1,j}\right), \\ q_{i,j-1} + \left(s(,b_j) \times \left(\left|d_{i-1,j-1}\right| - 1\right) \times q_{i,j-1}\right), \\ q_{i-1,j-1} + \left(s(a_i, b_j) \times d_{i-1,j-1} \times q_{i-1,j-1}\right), \\ 0, \end{cases} \quad (3)$$

Where,

$s(a_i,) = -1.1$	Score for alignment ending with an insertion
$s(,b_j) = -1.1$	Score for alignment ending with a deletion
$s(a_i, b_j) = +1.1$	Score for alignment ending with a substitution
$d_{i,j}$	Cosine distance between frames (i, j)

In order to maximize on quality, substitution scores must be positive and insertions/deletions must be negative. The recurrence also stops past dissimilarities causing global effects by setting all negative scores to zero, therefore starting a fresh new homologous relationship between local alignments. Figure 17 shows the plot of the quality scores after carrying out the recurrence (eqn. (3)).

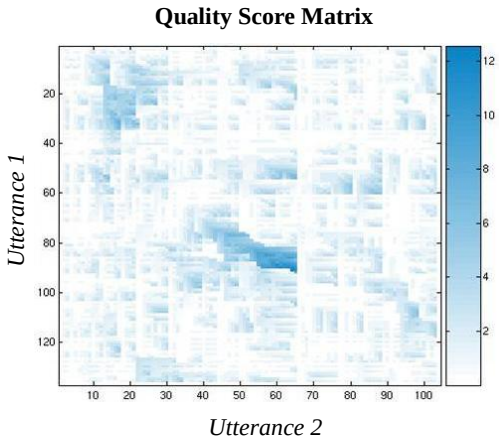


Figure 17. Quality score matrix for two utterances

Applying a substitution score of 1 causes the quality scores to grow as a linear function. The current settings use a substitution score greater than 1 (1.1), thus allowing the quality scores to grow exponentially, giving longer alignments more importance.

Altering the negative insertion/deletion scores greater or less than -1 allows the model to increase or decrease the spread of quality scores, therefore allowing control over the tolerance for

distortion. By setting insertion/deletion scores to values less than -1, the model will find closer matching repetitions, whereas a value greater than -1 allows the model to find repeated patterns that are longer and less accurate.

4.3.3.5 Finding Local Alignments

Backtracking pointers (bt) are maintained at each step of the recursion:

$$bt_{(i,j)} = \begin{cases} (i-1,j), & (deletion) \\ (i,j-1), & (insertion) \\ (i-1,j-1), & (substitution) \\ (0,0) & (initial\ pointer) \end{cases} \quad (4)$$

When the quality scores have been calculated with the recursion defined by equation 3, it is possible to backtrack from the highest score to obtain the local alignments in order of importance. A threshold is set so that only alignments of a desired quality are to be retrieved.

Figure 18 displays the steps taken to find the local alignments in order of importance:

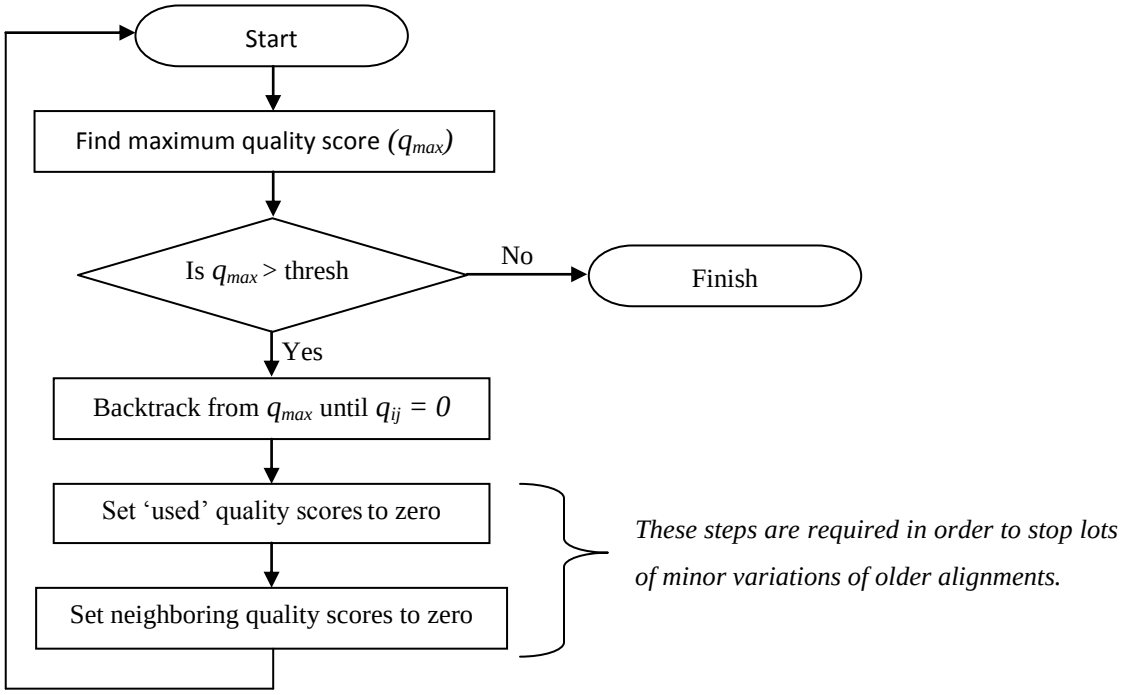


Figure 18. Flowchart of backtracking process to find local alignments

Figure 19 presents the optimal local alignment that was discovered by the DP-ngram model for two utterances:

utt_1 "Finally Ewan was there"

utt_2 "But Ewan does"

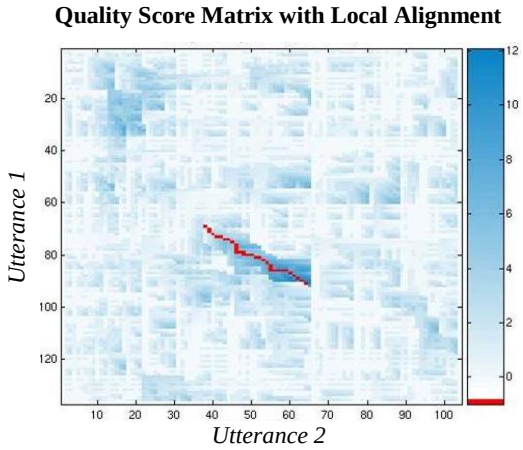


Figure 19. Quality score matrix with optimal local alignment plot

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

4.3.3.6 Key Word Discovery

As discussed earlier in section 4.1.6, the milestone for the end of the first year of the ACORNS project is that LA should have learned 10 key words. A key word discovery (KWD) method has been added to the acoustic DP-ngram algorithm 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 semantic features to build internal representations of key words. Both of these processes combine to achieve a system that is able to discover important word-like units from a cross-modal environment.

KWD is a simple approach that creates a class for each key word, 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}) will be 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 before - create a new key word class and store the whole utterance as an exemplar of it. Do not carry out the DP-ngram process and proceed to the next utterance in memory (Utt_{n+1}).
2. If both utterances share the same tag - proceed with the DP-ngram process and append discovered local alignments to that key word class. Proceed to the next utterance in memory (Utt_{n+1}).
3. If both utterances contain different tags - do not carry out DP-ngram process 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 will allow 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 either use the 'centroid' exemplar or all the stored local alignments for each key word class. Using all the stored alignments gives more accurate results, but the processing required for this method increases exponentially with experience until it is not a viable method. Using the 'centroid' is less accurate as it does not model the variance of the input. Future work, discussed in more detail in section 5.1.2, will be carried to try and build a single ideal representation of each key word class that models the variance.

Another important point to mention is that for the year 1 database there is only one tag associated with each utterance which makes this a binary decision for the KWD process. Utterances from the year 2 database will be more complex and contain multiple key words, but KWD will still run in the similar fashion with the addition of an 'uncertain' class; this is where segments are temporarily stored before a decision has been made on their correct key word class.

4.3.3.7 DP-ngram - Batch Process

For a set of utterances the acoustic DP-ngram method compares every incoming utterance (U_i) with all past utterances $\left(U_{j=1\rightarrow(i-1)}\right)$ to find local alignments. The search space for this process is the blue area in figure 20.

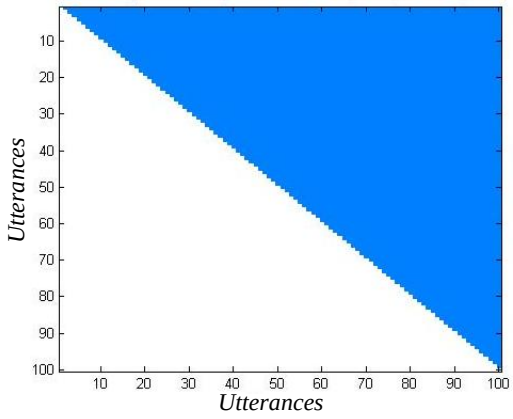


Figure 20. Plot of the completed search space for the DP-ngram as a batch process for 100 utterances

As a batch process the DP-ngram method can only run on a limited number of utterances, with processing complexity increasing towards infinity. This approach does not lend itself to an online language and recognition process. It was necessary to design an implementation of the acoustic DP-ngram method that could potentially handle an infinite number of utterances.

4.3.3.8 Incremental DP-ngram

Running the acoustic DP-ngram method as a batch process on the ACORNS English corpus (4000 utterances) is not viable with current processing and memory capabilities. Therefore an incremental version of the acoustic DP-ngram method has been designed. Also, running the algorithm as an incremental process allows for a more cognitively plausible system (Brent, 1999).

The model is made incremental by restricting the number of past utterances to be compared with the incoming utterance(U_i) with an utterance window $\left(U_{j=UttWindow\rightarrow(i-1)}\right)$. This creates a search space as shown in figure 21.a) & b).

Completed Search Space for Incremental DP-ngram (UttWindow = 20)

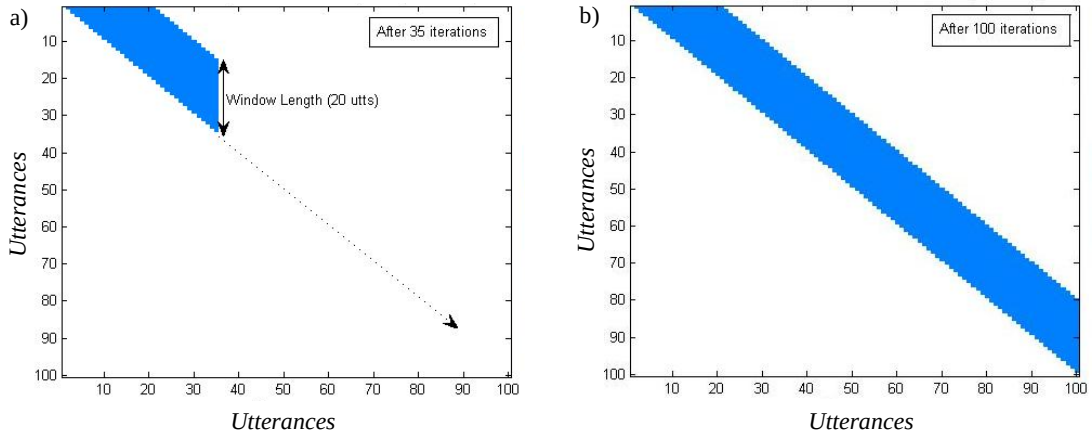


Figure 21.a) Plot of search space for the incremental DP-ngram after 35 utterances with a window length of 20 utternaces. b) Shows a plot of the completed search space after 100 utterances.

Decreasing $UttWindow$ allows the system to run faster at the expense of reducing the search space, therefore running the risk of potentially missing important repetitions. It was necessary to find the minimum $UttWindow$ length that will allow the system to find enough important repetitions in order to build accurate internal representations of the key words (experiment 1 in section 4.3.3.10).

4.3.3.9 Integrating DP-ngram into the LA Architecture

As discussed in section 4.3.3.9 the project has developed a framework that allows interaction and communication between LA and its carer. The framework also includes a memory architecture that attempts to achieve cognitive plausibility.

The DP-ngram algorithm has now been modified to work within this framework. The memory structure being implemented can be seen in figure 22.

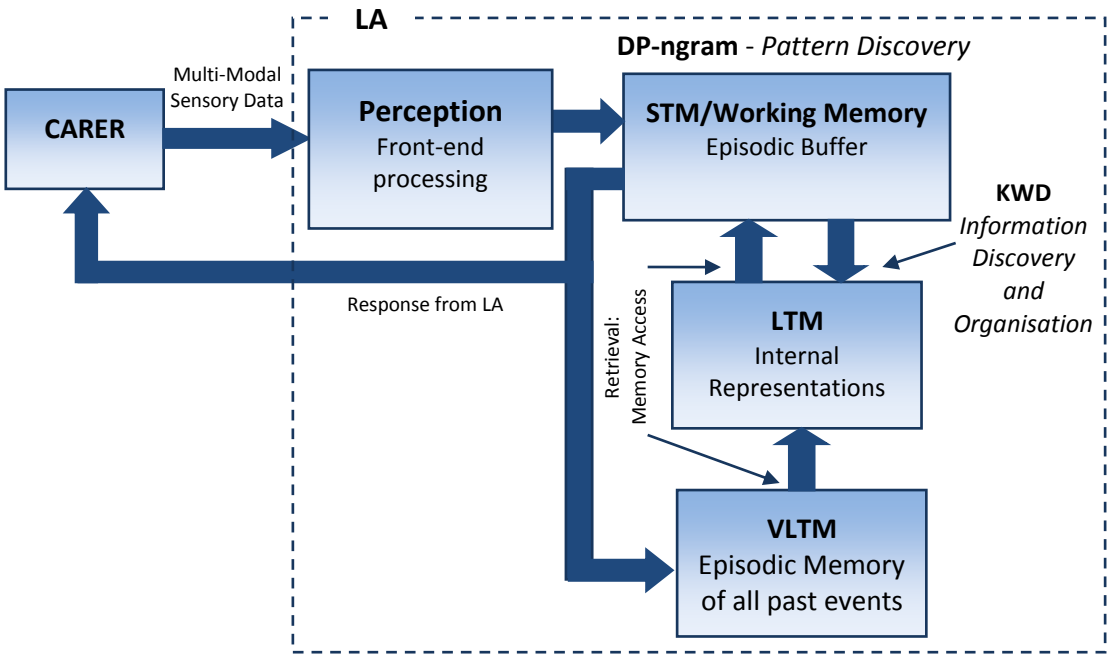


Figure 22. Integration of DP-ngram into the LA architecture

Carer – The carer feeds LA with cross-modal input (acoustic & semantic).

Perception – The stimulus is processed by the 'perception' module which converts the acoustic signal into a representation similar to the human auditory system (mfcc's using ACORNS front-end).

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 can then be compared with the current input to discover repeated patterns using DP-ngram.

Long Term Memory (LTM) – The ever increasing list of discovered units for each word representation are stored in the LTM. Clustering processes can then be applied to find the ideal representation. 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, in the future additional 'sleeping' processes could be carried out on the data stored in VLTM to re-organise internal representations or carry out additional analysis.

4.3.3.10 Experiments

Accuracy of experiments within the ACORNS project is based on LA's response to its carer. The response the carer is looking for is if LA can predict the key word tag associated with the current incoming utterance by only using the speech signal. The acoustic DP-ngram implementation attempts to solve this task using a method similar to traditional DP template based recognition. The recognition process is carried out by comparing exemplars, of discovered key words, against the current incoming utterance using the DP-ngram method to calculate quality scores. Thus, the alignment that produces 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:

1. Finding the optimal utterance window length for incremental DP-ngram

For this experiment, 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.

2. Comparing DP-ngram as a batch and incremental process

The optimal window length chosen for the incremental implementation is compared against the batch implementation of the DP-ngram algorithm.

3. Key Word Discovery - Centroid vs Complete Exemplar List

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 exemplar that best represents the list, the 'centroid' (method described in section 4.3.3.6). This experiment will compare these two methods for representing internal representations of the key words.

4. Speaker-dependency

The algorithm is tested on its ability to handle the variation in speech from different speakers. Different feature vectors from the front end are fed to the system. It was mentioned in section 4.3.3.2 that the ACORNS front-end allowed the option for normalization. The feature vectors used are listed below:

V1 - Default HTK 39-element mfcc's (no normalisation)

V2 - ACORNS 37-element mfcc's (no normalisation)

V3 - ACORNS 37-element mfcc's with Cepstral Mean Normalisation

V4 - ACORNS 37-element mfcc's with Cepstral Mean and Variance Normalisation

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

5. Comparing DP-ngram against the NMF approach

The NMF approach was the first and only end-to-end approach within the ACORNS project and is used as a baseline for ACORNS experiments. The NMF results are based on the ACORNS Y1 Dutch corpus with 4 different speakers. The data set contains 10 key words and the utterances are fed to the system at random.

Test Data

The test data being used for experiments 1 and 2 is a sub-set of the ACORNS Y1 UK corpus – 100 different utterances from a single speaker. Each utterance contains one of the 10 key words and each key word is presented 10 times.

Experiment 3 uses 200 utterances from the ACORNS Y1 UK corpus, but includes all four speakers (2 male and 2 female) presented in a random order. There are still 10 key words present but the number of times they occur within the data set is random.

Experiment 4 uses the same setup as the data set for experiment 3 but the utterances are from the ACORNS Y1 Dutch corpus. Table 2 shows the test data being used for all four experiments.

Experiment	Corpus	Utterances	Speakers	Key words	Occurrences
1	UK Y1	100	1 m	10	10
2	UK Y1	100	1 m	10	10
3	UK Y1	200	1m	10	20
4	UK Y1	200	2 m - 2 f	10	Random
5	NL Y1	200	2 m - 2 f	10	Random

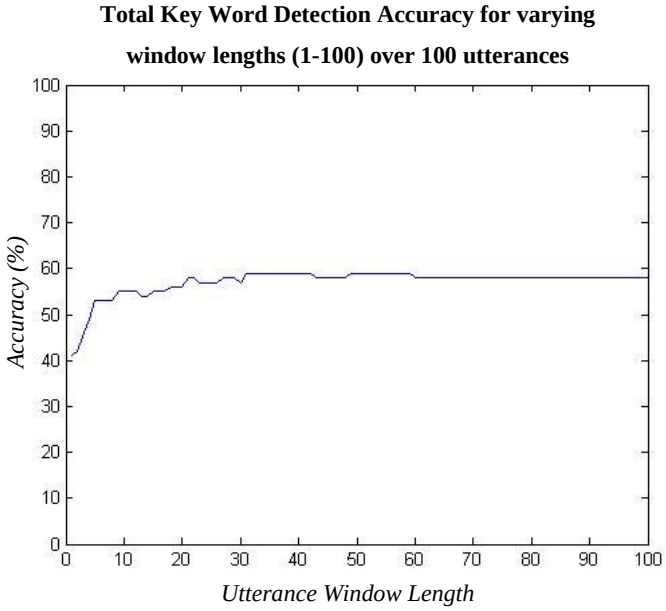
Table 2. Test data for DP-ngram experiments 1-4

4.3.3.11 Key word Recognition Results

1. Finding the optimal utterance window length for incremental DP-ngram

The DP-ngram algorithm was carried out on 100 utterances with varying utterance window lengths. The plot in figure 23 shows the total accuracy result for each window length used. The x-

axis displays the utterance window lengths used (1–100) and the y-axis displays the total accuracy (%).



Fugure 23. Plot of the total key word accuracy using varying utterance window lengths of 1-100. Each trial has been carried out on a test set of 100 utterances from a single

The results are as expected. Longer window lengths achieve more accurate results. This is because longer window lengths achieve 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. Accuracy results reach a maximum with an utterance window length of 21 utterances and then stabilise at around 58% (±1%). This shows us the minimum window length needed to build accurate internal representations of the words within the test set.

The window length used for all proceeding experiments is 21 utterances.

2. Comparing DP-ngram as a batch and incremental process

The plot in figure 23 displayed the total accuracy result for the different utterance window lengths and does not show the gradual word acquisition process. Figure 24 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 red plot shows the accuracy for randomly guessing the key word.

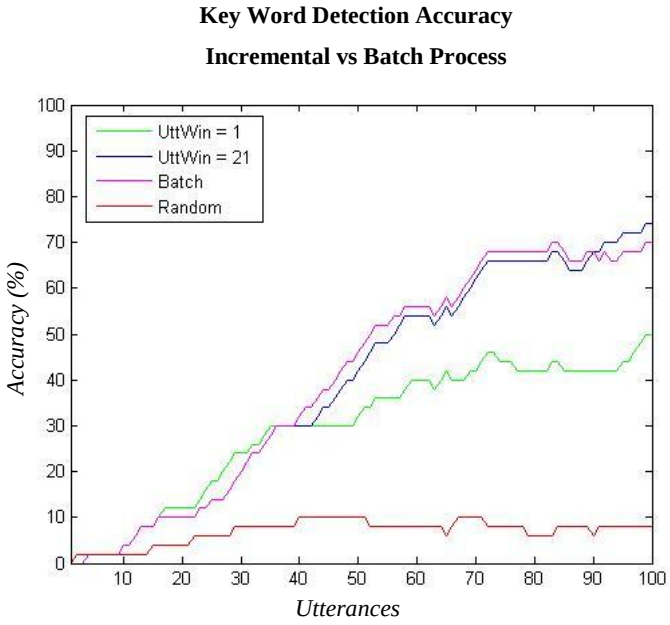


Figure 24. Key word detection accuracy for the DP-ngram algorithm running as a batch and incremental process. Results are plotted as a function of the past 10 utterances observed.

It can be seen from the plot in figure 24 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 I believe the batch process would 'catch up' in time.

Another important result to take into account is that only comparing 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.

3. Key Word Discovery - Centroid vs Complete Exemplar List

Currently the recognition process uses all the discovered exemplars for 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 to run on an infinite data set.

Another method employed was to calculate the 'centroid' for each key word class and use this single exemplar unit for recognition. Figure 25 shows the accuracy as a function of utterances observed for both methods.

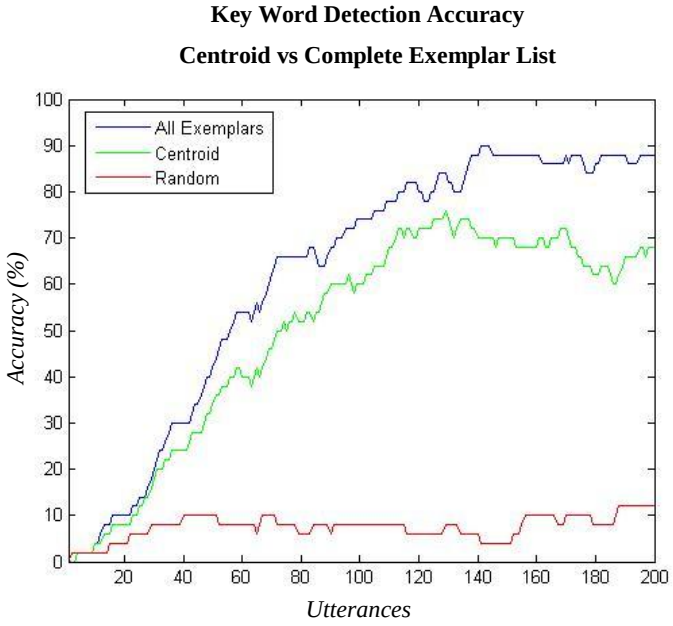


Figure 25. Comparison of key word detection accuracy using centroids or complete exemplar list for recognition.

The results show that the 'centroid' method is quickly outperformed and that the accuracy gap 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 accuracy results of 90% at around 140 utterances, where it then seems to stabilize at around 88%.

4. Speaker-dependency

This experiment has been carried out to test the speaker-dependency of the DP-ngram method. 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, accuracy of the internal representations increases but at a much slower rate than the single speaker experiments.

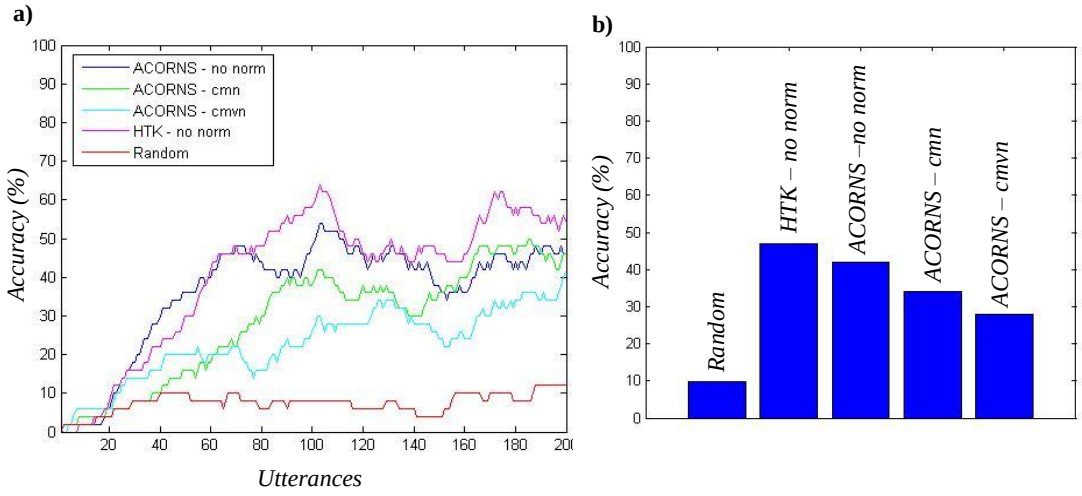


Figure 26. a) Accuracy of the system using different feature vectors as a function of utterances observed. b) Total accuracy after 200 utterances.

The assumption that using normalisation methods would achieve greater word detection accuracy, by reducing speaker variation, does not hold true. On second thought it is not surprising, as the system will be collecting exemplar units for each speaker.

This brings up another issue; the optimal utterance window length for the incremental DP-ngram process was calculated for a single speaker. Increasing the search space will allow the model to find more repeating patterns from the same speaker. With this logic, it could be hypothesized 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 results.

5. Comparing DP-ngram against the NMF approach

The NMF approach is currently the baseline for word discovery experiments within the ACORNS project. Figure 27 plots the key word detection accuracy results for the two methods as a function of observed utterances. The test set consists of Dutch utterances from four different speakers (2 male and 2 female) in a random order.

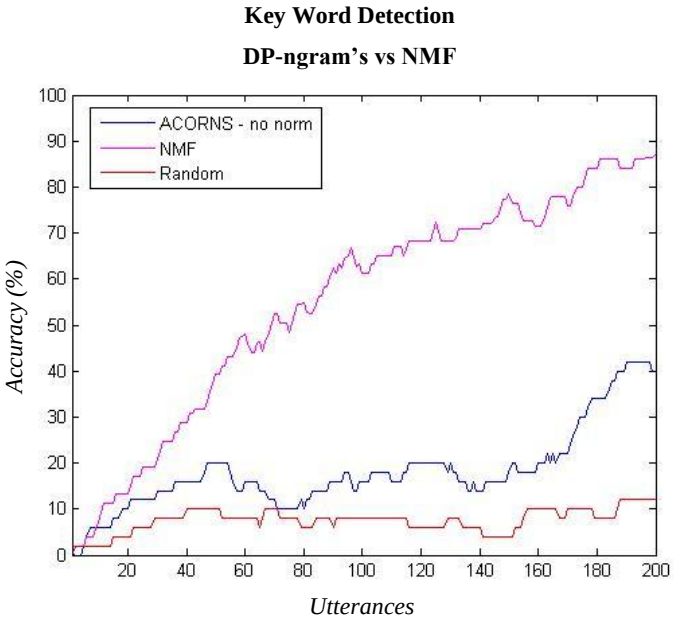


Figure 27. Comparison of key word detection accuracy for NMF and DP-ngram methods over 200 utterances.

Accuracy of the internal representations rise at a much faster rate for the NMF method compared to the DP-ngram method, although there is a sharp rise up to 40% after 170 utterances.

4.3.3.12 Conclusions

Preliminary results show that the environment is rich enough for word acquisition tasks. The DP-ngram method has proven to be a successful approach for building stable internal representations by exploiting the statistical regularities across multiple modalities (semantic and acoustic). The incremental approach shows that the model is still able to learn correct word representations by only processing a limited number of past utterances. It is also apparent that the re-use of internal representations allows the word hypotheses to become more accurate and stable at a faster rate by discovering more exemplar representations that would have been outside of the incremental search space.

Additionally to the acquisition of words and word-like units the system is able to use these representations 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.

The experimental results are promising. However, it is clear to see that the model suffers from speaker-dependency issues. Further research into more suitable feature representations and methods to model the additional variation within the acoustic signal will need to be carried out.

Compared to the baseline, performance of the DP-ngram method falls short. The main reason for this is because of speaker-dependency when working directly with the acoustics. However, it is important to note that the architecture is in its early stages of development and is very adaptable. It is also not dependent on pre-defined parameters to learn, which only optimise the algorithms efficiency for its surrounding environment.

Another advantage of this system, compared to NMF, is that it is able to give temporal information of the whereabouts of important repeating structure. As discussed in section 3.4.6, the NMF approach is only able to detect the presence of the key word units and not where they lie within the utterances.

5. Research Plans

5.1 Pattern Discovery and Organisation

The view that the brain is a self-organising, complex dynamic system is constantly being reinforced (Waddington, 1957; Thelen, 1982 & 1995; Savelsbergh and Van der Kamp, 1993; Kelso, 1995; Newell et al, 2003). Complex systems are made up of many interrelated parts with the same general statistical mechanism, which by themselves are insignificant but as a whole produce rich and complicated structures. Trying to model the whole system adequately is very difficult, whilst only modelling the smaller parts will not allow critical emergent properties to arise.

The proposal outlined in this section is for an architecture aimed towards online language learning and recognition through a cognitively plausible approach.

The first key question of this report asked whether modelling human language acquisition will help create a more robust speech recognition system. Therefore development of the proposed architecture will be limited to cognitively plausible approaches and should exhibit similar developmental properties as early human language learners. It is understood that some of the novel techniques that will be considered will have question marks on their plausibility, but in turn this will raise some important questions, giving us a deeper insight into human language acquisition and creating more robust speech recognisers.

The model being developed will be compared and tested against the current state of the art speech recognition systems and novel word discovery algorithms:

- ACORNS experiments will be used as a baseline (NMF, NMF with prosody) for evaluating the DP-ngram approach within the project.
- Conventional ASR systems will be used to compare the DP-ngram approach for different tasks. Experiments will be designed that highlight the weaknesses and strengths for both methods. It is hypothesised that the DP-ngram approach should excel with the following issues:
    - Out of dictionary issues
    - Language-dependancy

    - Discovering optimal basic units for its environment
    - Pre-defined grammatical rules
- Preliminary results show that we can exploit the information across multiple modalities (abstract semantic features and speech) for word acquisition? The Dp-ngram model will be tested against other novel cross-modal speech segmentation and word discovery algorithms.

The second key question asked how much do humans learn from their environment alone and is it rich enough to learn language without any prior knowledge? The architecture in development aims to answer this question by taking a non-nativist approach. Structure will be derived from the systems cross-modal environment without any innate, pre-defined, linguistic knowledge, using general statistical learning mechanisms that can be applied to all types of modalities.

The main areas of research and development to be carried out have been split into separate work categories listed below. The Gantt chart in section 5.2.4 displays the distribution of the workload for each category. Although the work is listed in a distinct serial fashion, research for each category will naturally overlap and it will be possible to run experiments in parallel.

5.1.1 Automatic Segmentation of Word-Like Units

The Acoustic DP-ngram algorithm is now fully operational and can be used as a platform for further development and experiments. As described in section 3.4, there are a couple of novel automatic segmentation and word discovery approaches similar to the acoustic DP-ngram model. Experiments will be carried out on the same test data used for their experiments and hopefully lead to publishable results.

- Experiments and comparisons with ten Bosch's SWD approach (2007). Louis ten Bosch is a member of the ACORNS project and is keen to compare both methods.
- Experiments and comparisons with Park & Glass' segmental DTW approach (2007 & 2008). Personal contact has been made with Jim Glass to see if I could use the same test data they used, which has been allowed.

As shown in the experiments in section 4.3.3.10, the system suffers from speaker-dependency issues. At the front-end, experiments will be carried out with various feature representations to compare against the ACORNS features. Investigations will also be carried out into different techniques, statistical and exemplar, to handle speaker variation.

5.1.2 Memory Organisation

Recent experiments with the acoustic DP-ngram algorithm have highlighted the issue of organising and structuring the units continuously being discovered. Various clustering techniques will need to be considered in order to create internal representations of important lexical units that are constantly evolving and becoming more accurate with experience.

Another problem to be solved is how do we carry out recognition on the incoming speech signal? Currently the algorithm discovers repeating acoustic segments. With these segments we can decide to build statistical models that capture speech variation or use all the exemplars stored in memory.

The internal representations are currently being updated every time something new is discovered in an incremental fashion. This means that the system is dependent on events co-occurring over a short period of time. A secondary update process could be implemented as a re-occurring 'sleeping phase'. This phase would allow the memory organisation to re-structure itself by looking at events over a longer history as a batch process. Research will be carried out into current cognitive theories of memory re-organisation during sleeping periods and clustering algorithms that could be implemented.

5.1.3 Cross-Modal Processing

The second key question asks if our environment is rich enough to learn language. Key question 5 extends this by asking what modalities are used by early language learners and can therefore be exploited for pattern discovery tasks.

Semantic Features

The ACORNS project associates semantic features to LA's world as crisp abstract tags (as discussed in section 4.2). Utterances in the second year corpus will contain multiple key words and key words will contain multiple semantic features. The DP-ngram algorithm will need to be modified to handle the multiple semantic features contained within each utterance. The system will handle the more complex input in a similar fashion by exploiting the co-occurrence of semantic events.

Prosodic Features

The prosodic information within the acoustic data will be used as a cue to aid pattern discovery and as an attention mechanism. Key question 5 raised the issue of how infants deal with conflicting cues? A weighting system will need to be developed for the different input modalities.

These should be dynamic depending on the environment and how much each modality can be trusted.

Rhythm

- Used by infants as a cue to help detect word boundaries. The output of the rhythmogram model (Todd & Brown, 1996) could be used as a pointer or weight for word or syllable boundaries.
- Rhythm is used as an attention mechanism, whereby the infant can discriminate different languages from different rhythm classes and concentrate only on native language. Could behavioural data of pre-verbal infants be modelled? These results will link well with section 5.1.7 when carrying out experiments on multiple languages.
- Behavioural data has shown that fetal memory exists and that it allows the fetus to build a bond with its mother's voice and her native language (DeCasper & Spence, 1986; Hepper, 1996; Federico, 1999). The ACORNS corpus has been specifically recorded to include parental figures. Experiments will be carried out with the proposed architecture that introduces low pass filtered speech from a parent as it would be heard in utero. Experiments will show if this rhythmic low-perplexity speech lays down the essential building blocks for learning the native language of the mother.

Pitch

- Infants attend to variation in their environment. Adults will highlight words of importance through infant directed speech (IDS). During IDS adults will place more pitch variance on words that they want the infant to attend to. Pitch will be used as a simple attention mechanism, giving more weight to parts of utterances with more pitch variance.

5.1.4 Communicative Behaviour

Parental interaction plays a key role during the early stages of language acquisition. Parents are able to provide corrective feedback and aid the learning process. Will the system achieve higher accuracy rates with some kind of carer feedback during an initial training phase?

The current experiments require LA to reply[2] to the carer that it understands the incoming utterance. The second year experiments, to be carried out on the second year database, will require a different reply (currently being decided). The second year database also includes corrective sentences which will be fed to the system when LA replies incorrectly.

The architecture will need to be modified to handle the questions being posed by the carer and how to deal with a corrective sentence. Interaction with the carer allows the system to make full use of its environment and therefore help shed some light on the second key question.

5.1.5 Hierarchical Analysis

Key question 3 raised an important problem faced in linguistics and speech recognition research, what makes up the fundamental units of speech? Over the years, rather than converging to a single basic unit, the candidates have increased, some successfully being used for conventional ASR systems and others for cognitive plausibility. Current theories of early language acquisition suggest that speech units are an emergent, self-organising property of speech perception. Automatically deriving hierarchical structure in the input will allow us to discover and analyse the relations between different significant units present in speech.

The automatically discovered lexical units obtained from the acoustic DP-ngram model can be used to transcribe past utterances into a sequence of discrete symbols. This lossless compression will greatly reduce the speech data, allowing pattern discovery algorithms, such as the DP-ngram model, to find repeated structure over a larger time scale. Iteratively carrying out this process will discover repeating patterns over larger time scales, which can be used to hierarchically structure the input.

In addition to hierarchically structuring the input, compressing the speech data will allow larger and more important lexical units to be statistically analysed, by calculating transitional probabilities between units. This could then be used to give recognition confidence measures or help predict what the following units could be.

5.1.6 Modelling Phonetic Categorisation

Deriving hierarchical structure will hopefully give an insight into the fundamental units of speech. The fundamental units discovered through this system will be used for recognition

[2] A measure of word discovery accuracy was obtained by asking LA to predict the key word tag associated with the current incoming utterance

experiments that will aim to answer key question 4. Current cognitive theories and behavioural data suggests that the loss of universal perceptual abilities, through phonetic categorisation, occurs as a stage-like change over a gradual learning process and not as an innate function in our brain that occurs at a discrete time.

Experiments will be carried out to test the systems discrimination abilities of the finer phonetic detail within the speech data over time. I am hypothesising that phonetic categorisation is an emergent property of language learning and will only appear as general statistical processes are gradually replaced with automatically derived stochastic linguistic rules.

5.1.7 Multiple Language Learning

If the language acquisition skills employed by an infant, and the computational model, are truly general then it should be possible to learn all languages. Experiments will be carried out with the multiple languages available in the ACORNS database (English, Finnish and Dutch). The system should exhibit similar patterns to behavioural data recorded for second language learners.

Newborns are sensitive to the rhythm class of their mother's native tongue. As discussed in section 5.1.3, an important experiment to carry out would be to see if adding an additional rhythm input affects the systems accuracy for languages of different rhythm classes. The ACORNS database includes both 'syllable-timed' (Finnish) and 'stress-timed' (English & Dutch) languages; further tests with 'mora-timed' languages (e.g. Japanese) will need to be recorded.

5.1.8 Visualising Evolution

Describing early language acquisition as a dynamic system has become very popular within the cognitive science field. As described in section 3.1.4, development is visualised as a constantly evolving ontogenetic landscape. Current literature displays these theoretical landscapes as hand drawn examples. As yet, there does not seem to be anyone who has attempted to visualise the learning stages of a language acquisition computational model in the same way. If it was is possible to plot the emerging behaviour of the system, would it display the same patterns that they hypothesise.

5.2 Other work to be carried out

5.2.1 Experiments

- An important issue that will need to be resolved is speaker-dependency. Experiments with different feature vectors, suitable for multiple speakers, will be carried out and compared.
- Experiments and comparisons with Saffran's behavioural data (1996, 1999)
- Experiments with the Y3 database will be carried out late next year.
    - Acoustic Dp-ngram
    - Prosodic NMF
- The ACORNS researchers in Helsinki are developing a co-occurrence model that uses DTW. Collaboration work can be carried out comparing the two methods.
    - They are particularly interested in extracting hierarchical structure of sub-word units within speech.

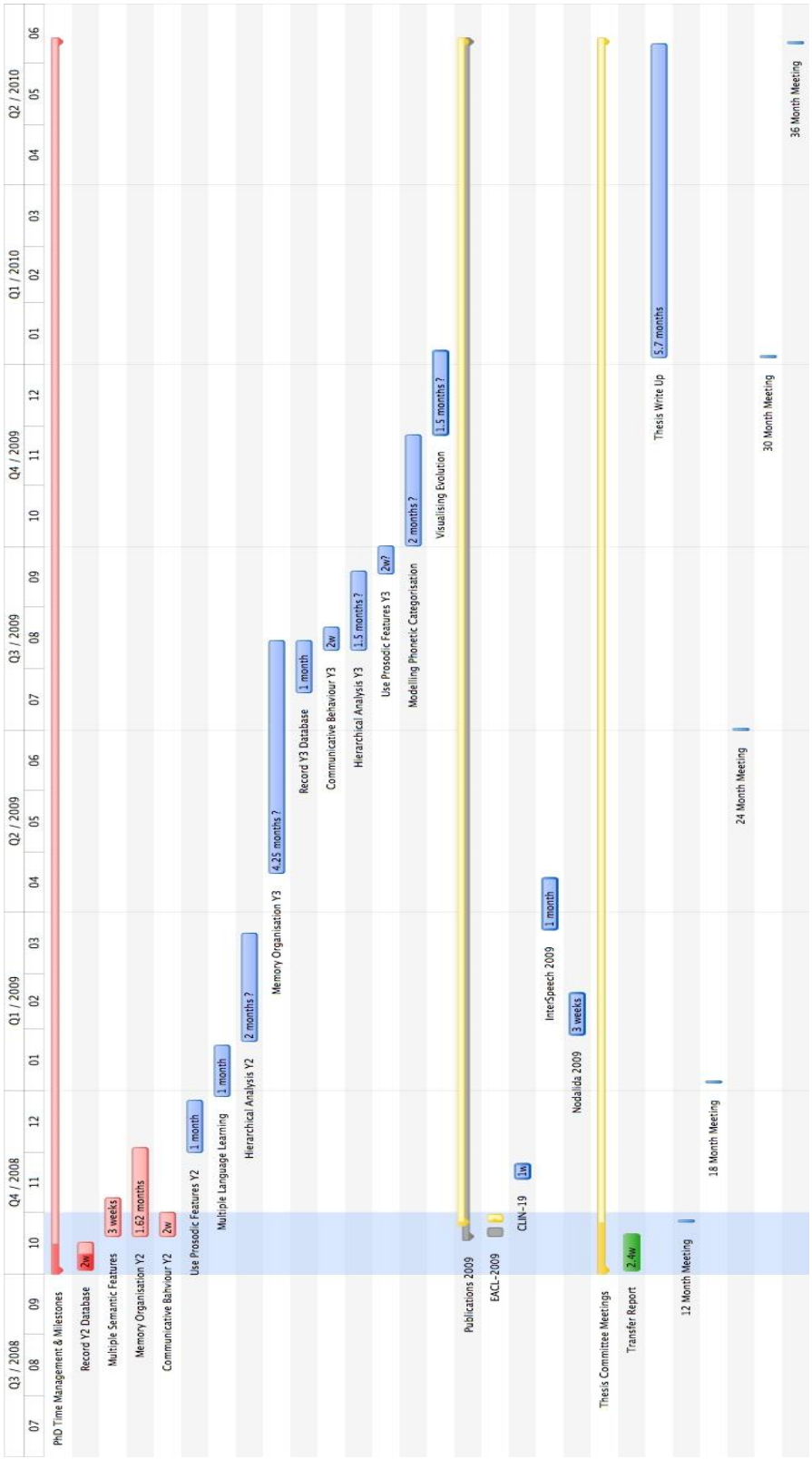
5.2.2 ACORNS Speech Corpus

- Record Y2 database
- Record Y3 database

5.2.3 Possible Publications for 2009

Conference 2009	Deadline	Date	Possible Submission
CLIN-19 **Groningen, NL**	17/11/08	22/01/09	Abstract submission for oral or poster presentation – Preliminary results of hierarchical structural analysis.
EACL-2009 **Athens, Greece**	24/10/08	30/03/09 – 03/04/09	Student workshop. Full paper of on-going research
Nodalida 2009 **Denmark**	12/01/09	15/05/09 – 16/05/09	Full paper of speech segmentation and hierarchical structural analysis.
InterSpeech 2009 **Brighton, UK**	17/04/09	06/09/09 – 10/09/09	Full paper of work carried to date.

5.2.4 Gantt Chart



6. Research Training Program

6.1 Modules

All RTP modules successfully completed.

Computational Models of Mind

Credits =10

Examines computational models of the mind inspired by the architecture of the brain and the distributed nature of biological cognition.

Mathematical and Statistical Programming

Credits =10

Introduction to numerical, statistical, graphical and algebra computer languages and packages (MatLab & SciLab).

Learning and Memory in Young Children

Credits =10

Considers early cognitive development and the assessment of learning and memory in preverbal infants and young children. Focus on human cognitive development, but animal models of learning and memory are considered along with the contributions of adult humans with brain injury to understand memory development.

Speech Technology

Credits =10

Study the principles of the emergent field of speech technology, the typical applications of these principles and assess the state-of-the-art in this area.

Personal and Professional Skills 1

Credits =5

Undertake a range of activities that will benefit continuing professional development and completion of research studies.

- Undertake regular laboratory demonstrating duties

- Prepare submission for and successfully complete application for ethics review procedure
- Set up and manage a website describing research activities
- Attend a national/international conference
- Draw up and maintain an updated CV

7. References

Anderson J.L., Morgan J.L. & White K.S. (2003),'A Statistical Basis for Speech Sound Discrimination' *Language and Speech*, **46**(2-3), 155-182.

Aslin R.N., Pisoni D.B., Hennessy B.L. & Perey A.J. (1981),'Discrimination of voice onset time by human infants: new findings and implications for the effects of early experience' *Child Development* **52**, 1135-1145.

Aslin, R, N. (1993), 'The strange attractiveness of dynamic systems to development', In L. B. Smith & E. Thelen (Eds.), *Dynamic Systems in Development: Applications*, Cambridge, MA: MIT Press, pp. 385-399.

Best C.T., McRoberts G.W. & Sithole N.M. (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.

Brent, M, R. (1999), 'Speech segmentation and word discovery: a computational perspective', *Trends in Cognitive Sciences* **3**(8), 294-301.

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

Christiansen, M. H., Allen, J. & Seidenberg, M. (1998),'Learning to segment speech using multiple cues', *Lang. Cogn. Processes* **13**, 221-268

Clark, H. H. (2002), 'Speaking in time', *Speech Communication* **36**, 5-13.

Cutler, A. (1994), 'The perception of rythm in language', *Cognition* **50**, 79-81.

DeCasper, A. & Spence (1986), 'Prenatal maternal speech influences newborns' perception of speech sounds', *Infant Behavior and Development* **9**, 133-150.

Drake, C., Jone M. R. & Baruch, C. (2000), 'The development of rhythmic attending in auditory sequences: attunement, referent period, focal attending', *Cognition* **77**(3), 251-288.

Eimas P.D., Siqueland E.R., Jusczyk P. & Vigorito J. (1971), **'**Speech perception in infants', *Science* **171**, 303-306.

Elman, J. L. (1991), 'Distributed representations, simple recurrent networks, and grammatical structure', *Machine Learning* **7**, 195-225.

Evans, J. R. (1986), *Rhythm in Psychological, Linguistic, and Musical Processes*. C. C. Thomas. Ill.

Federico, G. (1999), 'Music therapy & pregnancy: Prenatal stimulation', *26th Canadian Conference of Music Therapy*.

Gathercole, V. C. M. & Hoff, E. (2007), In Hoff, E. & Shatz, M. (Eds.), *Blackwell Handbook of Language Development*, Blackwell Publishing, chapter 6. Input and the Acquisition of Language: Three Questions, 107-127.

Grabe, E. & Low, E. L. (2002), 'Duration Variablility in Speech and the Rythm Class Hypothesis', *Papers in Laboratory Phonology 7*, Mouton.

Hannon, E. E. & Trehub, S. E. (2005), 'Tuning in to musical rhythms: Infants learn more readily than adults', *PNAS* **102**(35), 12639-12643.

Haken, H., Kelso, J.A.S., Bunz, H. (1985). 'A theoretical model of phase transitions in human hand movements'. *Biological Cybernetics* 51, 347-356.

Hawkins, J. (2004). *On Intelligence*. New York: Time Books, Henry Holt and Co.

Hepper, P. G. (1996), 'Fetal Memory: Does it exist? What does it do?', *ACTA Paediatrica Supplement* **416**, 16-20.

Hinton, G. E., Plaut, D. C. & Shallice, T. (1993), 'Simulating brain damage', *Scientific American* **269**, 76-82.

Jungers, M. K. Palmer, C. & Speer. S. R. (2002), 'Time after time: The coordinating influence of tempo in music and speech', *Cognitive Processing* **1**, 21-35.

Jusczyk, P. W., Friederici, A. D., Wessels, J., Svenkerud, V. Y. & Jusczyk, A. M. (1993) 'Infants' sensitivity to the sound patterns of native language words', *Journal of Memory & Langauge*, **32**, 402-420.

Keil, F. C. (1979), *Semantic and conceptual development: An ontological perspective*, Cambridge, MA: MIT Press.

Kelso, J. A. S. (1995) *Dynamic Patterns*. Cambridge, MA: MIT Press, chapter 2. The Self-Organization of Brain and Behavior, pp. 46-53.

Kirkham N. Z., Slemmer A. J. & Johnson S. P. (2002),'Visual statistical learning in infancy: evidence for a domain general learning mechanism', *Cognition* **83**, B35-B42.

Kitahara, M. (2000), 'Effects of rhythm patterns in English speech: an account in dynamical systems theory'.

Kuhl P.K, Williams K.A., Lacerda F., Stevens K.N. & Lindblom B. (1992), 'Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age', *Science* **255**, 606-608.

Kuhl, P. K. (2000), 'A new view of language acquisition', *PNAS* **97**(22), 11850-11857.

Kuhl, P. K. (2003), 'Human speech and birdsong: Communication and the social brain', *PNAS* **100**(17), 9645-9646.

Kuhl, P. K. (2004), 'Early Language Acquisition: cracking the speech code', *Nature* **5**, 831-843.

Lee, D. D. & Seung, H. S. (1999), 'Learning the parts of objects by non-negative matrix factorization', *Nature* **401**(6755), 788–791.

Lerdahl, F. & Jackendoff, R. (1983), *A generative theory of tonal music*. Cambridge: MIT Press.

Liberman, M. and Prince, A. (1977), 'On Stress and Linguistic Rhythm', *Linguistic Inquiry* **8**(2), 249-336.

Linsker, R. (1988), 'Self-organisation in a perceptual network', *Computer Magazine* **21**(3), 105-117.

Logan, B. (1987), "Teaching the Unborn: Precept and Practice', *Pre and Perinatal Psychology Journal* **2**(1).

Logan J.S., Lively S.E. & Pisoni, D.B. (1991), **'**Training Japanese listeners to identify English /r/ and /l/: A first report', *Journal of Acoustic Society of America* **89**, 874-886.

Lu, L., Wang, M. & Zhang, H. (2004), 'Repeating pattern discovery and structure analysis from acoustic music data', *Proc. 6th ACM SIGMM Int. Workshop on Multimedia Information Retrieval*, 275-282.

Lust, B. (2006), *Child Language Acquisition and Growth*, Cambridge University Press.

Maye J., Werker J.F. & Gerken L. (2002), **'**Infant sensitivity to distributional information can affect phonetic discrimination', *Cognition* **82**, B101-B111.

McClelland, J. L. & Elman, J. L. (1986), 'The TRACE Model of Speech Perception', *Cognitive Psychology* **18**, 1-86.

McClelland, J. L. & Jenkins, E. (1991), In K. Van Lehn (Ed.), *Architectures for Intelligence*, Hillsdale, NJ: Erlbaum. Nature, nurture, and connections: Implications of connectionist models for cognitive development, pp. 41-73.

McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. (1995), 'Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory', *Psychological Review* **102**, 419-457.

Mountcastle, V. B. (1978), *The Mindful Brain*, MIT, 'An organizing principle for cerebral function: the unit model and the distributed system'.

Muchisky, M. L., Gershkoff-Stowe, E. C., & Thelen, E. (1996), In Rovee-Collier, C. & Lipsitt, L. P. (Eds.), *Advances in Infancy Research* **10**, Ablex Publishing Corp., Norwood, NJ, 'The epigenetic landscape revisited: A dynamic interpretation', 121-160.

Nazzi, T., Bertoncini, J. & Mehler, J. (1998), 'Language Discrimination by Newborns: Toward an Understanding of the Role of Rhythm', *Journal of Experimental Psychology: Human Perception and Performance* **24**(3), 756-766.

Newell, K. M., Liu, Y. & Mayer-Kress, G. (2003), 'A dynamical systems interpretation of epigenetic landscapes for infant motor development', *Infant Behav. & Dev.* **26**, 449-472.

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

Patel, A. D. (2003), 'Language, music, syntax and the brain', *Nature* **6**(7), 674-681.

Park, A. & Glass, J. R. (2007), 'Towards unsupervised pattern discovery in speech', *Proc. ASRU*, 53-58.

Park, A. & Glass, J. R. (2008), 'Unsupervised Pattern Discovery in Speech', *Trans. ASLP* **16**(1), 186-197.

Pena, M., Maki, A., Kovacic, D., Dahaene-Lambertz, G., Koizumi, H., Bouquet, F. & Mehler, J. (2003), 'Sounds and silence: An optical topography study of language recognition at birth', *PNAS* **100**(20), 11702-11705.

Polka L. & Werker J.F. (1994), 'Developmental changes in perception of non-native vowel contrasts', *Journal of Experimental Psychology* **20**, 421-435.

Port, R. F., Tajima, K. & Cummins, F. (1998), 'Speech And Rhythmic Behavior'.

Port, R. F. (2000), 'Dynamical Systems Hypothesis in Cognitive Science', In A. Lockyer (Ass. Ed.), *Encyclopedia of Cognitive Science*, MacMillan Reference Ltd, London.

Ramus, F., Nespor, M. & Mehler, J. (1999), 'Correlates of linguistic rhythm in the speech signal', *Cognition* **73**, 265-292.

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999), 'Statistical learning of tone sequences by human infants and adults', *Cognition* **70**(1), 27-52.

Saffran, J. R., Senghas, A., & Trueswell, J. C. (2000), 'The acquisition of language by children', *PNAS* **98**(23), 12874-12875.

Saffran, J. R. (2003), 'Statistical Language Learning: Mechanisms and Constraints', *Current Directions in Psychological Science* **12**(4), 110-114.

Saffran, J. R. & McMullen, E. (2004), 'Music and Language. A Developemental Comparison', *Music Percieption* **21**(3), 289-311.

Sankoff, D. & Kruskal, J. B. (1983), In Sankoff, D. & Kruskal, J. B., (eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley Publishing Company, Inc. Finding similar portions of two sequences, pp. 293-296.

Savelsbergh, G. J. P. & Van der Kamp, J. (1993),'The development of coordination in infancy', *Advances in Psychology* **97**, 289-317.

Seigler, R. S. (1981), 'Developmental sequences between and within concepts', *Monographs of the Soc. for Research in Child Dev.* **46**.

Seindenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2002), 'Does grammar start where statistics stop?', *SCIENCE* **298**, 552-554.

Sejnowski, T. J. & Rosenberg, C. R. (1987), 'Parallel networks that learn to pronounce English text', *Complex Systems* **1**, 145-168.

Sheldon A. & Strange W. (1982),'The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception', *Applied Psycholinguistics* **3**, 243-261.

Stager, C. L., & Werker, J. F. (1997), 'Infants listen for more phonetic detail in speech perception than in word-learning tasks', *Nature* **388**, 381-382.

Stouten, V., Demuynck, K. & Van hamme, H. (2007), 'Automatically Learning the Units of Speech by Non-negative Matrix Factorisation', *Proc. European Conference on Speech Communication and Technology*, 1937-1940.

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

Ten Bosch, L. & Cranen, B. (2007), 'A computational model for unsupervised word discovery', *INTERSPEECH 2007*, 1481-1484.

Thelen, E., & Fisher, D. M. (1982), 'Newborn stepping: An explanation for a "disappearing reflex"', *Developmental Psychology* **18**, 760- 775.

Thelen, E. & Smith, L. B. (1995), 'A dynamic systems approach to development of cognition and action', *Journal of Cognitive Neuroscience* **7**(4), 512-514.

Todd, N. P. M. & Brown, G. J. (1996), 'Visualization of Rhythm, Time and Metre', *Artificail Intelligence Review* **10**, 253-273.

Trehub S.E. (1976),'The discrimination of foreign speech contrasts by infants and adults', *Child Development* **47**, 466-472.

Volkan, T. & Selman, N. (2005), 'Dynamic time warping as a novel tool in pattern recognition of ECG changes in heart rhythm disturbances', *IEEE Int. Conf. on Sytems, Man, and Cybernetics*, 10-12.

Waddington, C. H. (1957), *The strategy of the genes*, London: George Allen & Unwin.

Werker J.F., Gilbert J.H., Humphrey K. & Tees R.C. (1981),'Developmental aspects of cross-language speech perception' *Child Development* **52**, 349-355.

Werker J.F. & Tees R.C. (1984),'Cross-language speech perception: Evidence from perceptual reorganization during the first year of life', *Infant Behavior and Development* **7**, 49-63.

Werker J.F. & Polka L. (1993) 'Developmental changes in speech perception: New challenges and new directions', *Journal of Phonetics* **21**, 83-101.

Werker J.F. & Tees R.C. (1999),'Influences on infant speech processing. Toward a new synthesis', *Annual Review of Psychology* **50**, 509-535.

Zhang, Y., Kuhl, P. K., Imada, T. Kotani, M. & Tohkura, Y. (2005), 'Effects of language experience: Neural commitment to language-specific patterns', NeuroImage 26, 703-720.

8. Bibliography

Axelrod, S. & Maison, B. (2004), 'Combination of hidden markov models with dynamic time warping for speech recognition', *ICASSP* **1**, 173-176.

Brenowitz, A. E. & Beecher, M. D. (2005),'Song learning in birds: diversity and plasticity, opportunities and challenges', *TRENDS in Neuroscience* **28**(3), 127-132.

Chamberlain, D. B. (1998), 'Prenatal receptivity and intelligence', *Journal of Prenatal and Perinatal Psychology and Health*, **12**(3,4), 95-117.

Davis, B. L. & MacNeilage, P. F. (2000),'An embodiment perspective on the acquisition of speech perception', *Phonetics* **57**, 229-241.

de Boer, B. & Kuhl, P. K. (2003), 'Investigating the role of infant-directed speech with a computer model', *Acoustics Research Letters Online*.

Denardo, E. V., Denardo, E. V., ed. (1982), *Dynamic Programming. Models and Applications*, Prentice-Hall, inc..

Digalakis, V., Rohlicek, J. R. & Ostendorf, M. (1991),'A dynamical system approach to continuous speech recognition', *IEEE Int. Conf. Acoust., Speech, Signal Processing*, **1**, 289-292

Dominey, P. F. & Ramus, F. (2000),'Neural network of natural language: Sensitivity to serial, temporal and abstract strucureof language in the infant', *Language and Cognitive Processes* **15**(1), 87-127.

Draganova, R., Eswaran, H., Murphy, P., Lowery, C. & Preissl, H. (2007),'Serial magnetoencephalographic study of fetal and newborn auditory discriminative evoked responses', *Early Human Dev*. **83**, 199-207.

Echols, C. H., Crowhurst, M. J. & Childers, J B. (!997), 'The perception of rhythmic units in speech by infants and adults', *J. of Memory and Lang.* **36**(2), 202-225.

Evans, J. R. (1986), *Rhythm in Psychological, Linguistic, and Musical Processes*. C. C. Thomas. Ill.

Evans, J. L. (2007), In Hoff, E. & Shatz, M. (Eds.), *Blackwell Handbook of Language Development*, Blackwell Publishing, chapter 7. The Emergence of Language: A Dynamical Systems Account, pp. 128-147.

Fitch, W. T. (2000), 'The evolution of speech: a comparative review', *Trends in Cognitive Science* **4**, 258-267.

Frankel, J., Richmond, K., King, S. & Taylor, P. (2000), 'An Automatic Speech Recognition System using Neural Networks and Linear Dynamic Models to recover and Model Articulatory Traces', in *ICLSP-2000* **4**, 254-257.

Gervain, J., Macagno, F., Cogoi, S., Pena, M. & Mehler, J. (2008), 'The neonate brain detects speech structure', *PNAS* **105**(35), 14222-14227.

Goldinger, S. D. & Azuma, T. (2003), 'Puzzle-solving science: the quixotic quest for units in speech perception', *Journal of Phonetics* **31**, 305-320.

Gollinkoff, R. M. & Hirsh-Pasek, K. (2008), 'How toddlers begin to learn verbs', *TRENDS in Cognitive Science* **12**(10), 397-403.

Gomez, R. L. & Gerken, L. (2000),'Infant artificial language learning and language acquisition', *TRENDS in Cognitive Science* **4**(5), 178-186.

Guenther, F. H. (1994), 'A neural network model of speech acquisition and motor equivalent speech production', *Biological Cybernetics* **72**, 43-53.

Hauser, M. D. & McDermott, J. (2003), 'The evolution of the music faculty: a comparative perspective', *Nature* **6**(7), 663-668.

Howard, R. A. (1966), *Dynamic Programming and Markov Processes*, The Massachusetts Institute of Technology.

Hsu, J., Chen A. L. P. & Chen, H. (2004), 'Finding Approxiamte Repeating Patterns from Sequence Data'.

Koreman, J. (2005), 'Percieved speech rate: the effects of articulation rate and speaking style in spontaneous speech', *J. Acoust. Soc. Am.* **119**(1), 582-596. .

Leader, L. R., Baillie, P., Martin, B. & Vermeulen, E. (1982), 'Fetal habituation in high-risk pregnancies', *Br. J. Obstet Gynaecol*, **89**,441–446.

Lieberman, P., Crelin, E. S. & Klatt, D. H. (1972),'Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee', *American Anthropologist* **74**(3), 287-307.

Linda Polka, S. R. & Mattock, K. (2007), In, Hoff, E. & Shatz, M. (Eds.), *Blackwell Handbook of Language Development*, Blackwell Publishers, Experiential Influences on Speech Perception and Speech Production in Infancy, pp. 154-172.

McNealy, K., Mazziotta, J. C. & Dapretto, M. (2006),'Cracking the language code: Neural mechanisms underlying speech parsing', *J. of Neuroscience***26**(29), 7629-7639.

Moore, R. K. (2007), 'Spoken language processing: piecing together the puzzle', *Speech Communication* **49**, 418-435.

Morgan, J. L. (1996), 'A rhythmic bias in preverbal speech segmentation', *Journal of Memory and Languages* **35**, 666-688.

Nakisa, R. C. and Plunkett, K. (1998), 'Evolution of a rapidly learned representation for speech', *Language and Cognitive Processes*, **13**(2,3), 105-127.

Neto, J. B. (2005), 'Is Music a Language', *Electronic Musicological Review 9*.

Newell, K. M., Liu, Y. & Mayer-Kress, G. (2003), 'A dynamical systems interpretation of epigenetic landscapes for infant motor development', *Infant Behav. & Dev.*, **26**, 449-472.

Ostendorf, M. (1999), 'Moving beyond the 'beads-on-a-string' model of speech', *IEEE ASRU workshop.*

Oudeyer, P. (2005), 'The self-organization of speech sounds', *Journal of Theoretical Biology* 233, 435-449.

Patel, A. D. & Daniele, J. R. (2002), 'An empirical comparison of rhythm in language and music', *Cognition* **87**, 35-45.

Pickering, M. J. & Garrod, S. (2006), 'Do people use language production to make predictions during comprehension', *TRENDS in Cognitive Sciences* **11**, 6.

Porter, R. J. & Hogue, D. M. (1998),'Nonlinear dynamical systems in speech perception and production', *Nonlinear Dynamics, Psychology and Life Sciences* **2**(2), 95-131.

Plunkett, K. (1995) Language acquisition. In M. Arbib (Ed.), *Handbook of Neural Networks*, Cambridge, MA: MIT Press.

Polka, L., Rvachew, S. & Mattock, K. (2007), In Hoff, E. & Shatz, M. (Eds.), *Blackwell Handbook of Language Development*, Blackwell Publishers, chapter 8. Experiential Influences on Speech Perception and Speech Production in Infancy, pp. 154-172.

Prescott, T. (2007, Oct),'Connectionism: an Introduction', Lecture Notes: Cooperative Models of Mind, University of Sheffield, unpublished.

Prescott, T. (2007,Oct),'The Mind Viewed as a self-organising, complex system', Lecture Notes: Cooperative Models of Mind, University of Sheffield, unpublished.

Prescott, T. (2007, Oct),'The mind/brain viewed as a dynamic system', Lecture Notes: Cooperative Models of Mind, University of Sheffield, unpublished.

Prescott, T. (2007, Nov),'Dynamic Systems and Human development', Lecture Notes: Cooperative Models of Mind, University of Sheffield, unpublished.

Sanders, L. D., Newport, E. L. & Neville, H. J. (2002),'Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech', *Nature Neuroscience* **5**(7), 700-703.

Smith, L. B. & Thelen, E. (2003), 'Development as a dynamic system', *TRENDS in Cognitive Science* **7**(8), 343-348.

Sodestrom, M. (2007), 'Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants', *Dev. Rev.* **27**, 501-532.

Spath, H. & Meek, B., (Eds.) (1980), *Cluster Analysis Algorithms for data reduction and classification of objects*, Ellis Horwood Publishers.

Stryker, M. P. (2003), 'Drums Keep Pounding a Rhythm in the Brain', *Neurocience* **291**, 1506-1507.

Tarpy, D. R. (2003), 'The Honey Bee Dance Language'.

Thiessen, E. D. & Saffran, J. R. (2003), 'When Cues Collide: Use of Stress and Statistical Cues to Word Boundaries by 7- to 9-Month-Old Infants', *Developement Psychology* **39**, 706-716.

Trainor, L. (2008), 'The neural roots of music', *Nature* **453**, 598-599.

Trehub, S.E., Trainor, L.J., & Unyk, A.M. (1993), 'Music and speech processing in the first year of life', *Adv. Child Dev. Behav.* **24**, 1-35.

Tuller, B. (2003), 'Computational models in speech perception', *J. of Phonetics* **31**(3,4), 503-507.

Werker, J. F. & Tees (1999), 'Influences on infant speech processing: toward a new synthesis', *Annu. Rev. Psychol.* **50**, 509-535.

Werker, J. F. & Yeung, H. H. (2005), 'Infant speech perception bootstraps word learning', *TRENDS in Cognitive Science* **9**(11), 519-527.

Wolfe, J. (2002), 'Speech and music, acoustics and coding, and what music might be 'for'', *Proc. of the 7th Int. Conf. on Music Perception and Cognition*, 10 - 13.

Zatorre, R. J., Chen, J. L. & Penhune, V. B. (2007), 'When the brain plays music: auditory-motor interactions in music perception and production.', *Nature* **8**, 547-558.

Zhang, Y. & Wang, Y (2007),'Neural plasticity in speech acquisition and learning', *Bilingualism: Language and Cognition* **10**(2), 147-160.

Unknown, (2006), 'Rhythm in Music and Speech'.