UNIVERSITY OF ROCHESTER WORKING PAPERS IN THE LANGUAGE SCIENCES -VOL. 1, NO. 2

McMurray, B., Spivey, M., & Aslin, R. (2000). The perception of consonants by adults and infants: Categorical or categorized? Preliminary results. In K. M. Crosswhite & J. S. Magnuson (Eds.), *University of Rochester Working Papers in the Language Sciences, 1 (2),* 215-256.

The Perception of Consonants by Adults and Infants: Categorical or Categorized?
Preliminary Results

Bob McMurray
Department of Brain and Cognitive Sciences
University of Rochester
mcmurray@bcs.rochester.edu

Michael Spivey
Department of Psychology
Cornell University
mjs41@cornell.edu

Richard Aslin
Department of Brain and Cognitive Sciences
University of Rochester
aslin@cvs.rochester.edu

Abstract

An overwhelming majority of speech perception research has focused entirely on the end product of the perceptual process. Perhaps no other phenomenon in cognitive science is as overstudied with these "endpoint" techniques as the categorical perception of consonants. Recent advances in eye tracking methodologies have allowed us to now look at the intermediate stages of processing in several domains. In this paper we present two studies examining the time course of categorical perception in adults. We demonstrate that, although categorization seems to be present throughout the time-course of categorical perception, it is not immediately discrete. Accompanying simulations suggest that categorical perception may only be a single temporal facet of a more complex, continuously evolving process. Categorical perception has been pervasive in explaining diverse areas of cognition such as speech perception, color perception, music perception, non-human speech perception. Most importantly it has been invoked in explaining infants' speech perception abilities. Given the results presented here, it seems appropriate to expand any study of categorical perception beyond simply the temporal endpoints to the entire time course of infant perception. However, the inadequacy of current infant methodologies to provide identification data for speech stimuli provides the greatest obstacle to achieving this goal with infants. Thus, we present the anticipatory eye movement paradigm, which will allow us to assess identification and categorization in infants. Preliminary data obtained with this methodology suggests that this methodology can provide categorization data and may also provide a glimpse into the temporal dynamics of infant speech perception.

Introduction

Five decades of research in speech perception and phonetics have been based primarily on a single experimental paradigm. In this paradigm, the participant hears a speech sound (or series of speech sounds) and must report the stimulus as belonging to one of a set of possible response

the right theoretically motivated ways, the structure of the human speech recognition mechanism will become apparent.

This approach treats the speech recognition mechanism as a black box, accepting input from the ears and yielding phonemic output to the button-pushing finger (or as is commonly assumed, to the word-recognition system). Much research has demonstrated the viability of this basic hypothesis (see McQueen, 1996 for examples). This approach has been quite valuable, leading to refinements of stimulus generation cration -46.2 -rec.nsTw esi,ec.nsTwuj -46.2 -word

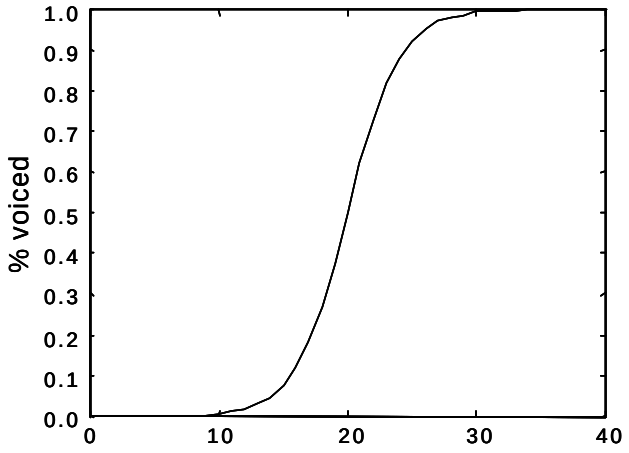

Figure 1: A schematic identification curve. Percentage of stimulus identified as voiced as a function of VOT is plotted, although similar curves have been reported for many other features of speech such as the acoustic cues signaling place of articulation.

stimuli on one side of the boundary is at chance in many situations[1]) has been taken to mean that perception of speech sounds is categorical on some level. In a sense, lower level acoustic information is lost in favor of a discrete category label (Liberman, Harris, Hoffman & Griffith, 1957).

Categorical perception is typically described in terms of an identification curve over an acoustic continuum (Figures 1). For example, the identification curve of a voicing continuum would indicate the percentage of time a stimulus was labeled "voiced" as a function of the voice onset time (VOT) of the stimulus. A steep logistic identification function has been the hallmark of virtually every experiment involving consonant perception.

Another nearly universal feature of speech perception experiments is that this identification curve has only been examined at the end of the perceptual process. We do not know the initial state or the pattern of change as the system settles on a response. Typical response times to individual phonemes in a button-pushing perception experiment are between 500 and 750ms. However, in this same amount of time during *on-line* speech recognition, the system may have to process as many as 20 phonemes. Clearly whatever phonetic category information is used for word recognition is not necessarily the sharp categorization we see in categorical perception experiments. This focus on the *on-line* processing of speech calls into question the pertinence of categorization data collected after a 750ms of processing.

[1] There is some evidence that in certain situations, greater-than-chance discrimination is possible within phonetic categories (Massaro and Cohen, 1983; Samuels, 1977; Pisoni and Lazarus, 1973). However, the sharp category boundary in the identification function is a hallmark of most phonetic categorization experiments. Since the research presented here is concerned mostly with identification, we will leave the debate on discrimination to another paper.

Since the initial findings of categorical perception by humans for speech sounds, categorical perception has been revealed in other domains. Facial discrimination, color perception and musical triad identification and have all been shown to be categorical processes (Beale & Keil, 1995; Bornstein & Korda, 1984; and Howard, Rosen & Broad, 1992; and respectively). It has also been shown for the perception of complex nonspeech sounds (Pisoni, 1977). Finally, categorical perception of human speech sounds has been found to occur for several non-human species (Kluender, Diehl & Killeen, 1987 [quail]; Kuhl & Miller, 1975 [chinchillas]; Kuhl & Padden, 1982 [macaques]).

These diverse findings of categorical perception suggest that categorical perception could be a very general property of the cognitive system (see Harnad, 1987, for numerous examples). However, although all categorical perception experiments show the same steep logistic function as

The temporal dynamics of categorical perception in adults

When participants listen to synthesized speech sounds that span a voice onset time (VOT) continuum between /ba/ and /pa/, stimuli with shorter VOTs are consistently identified as /ba/ and those with longer ones as /pa/. Additionally, using traditional measures, discrimination between different stimuli within a category is typically at chance. Although other studies have shown subjects have limited access to this subcategorical level of repbTw (1 12enw ( ) TjVhman, Blumly in, & 0 -1

The eye-tracking and response deadline methodologies used here were designed to give us a precise picture of the time course of categorical perception and tell us which of these hypotheses provides the best description of the data. The use of head-mounted eye tracking methodologies has allowed us to extend Pisoni and Tash's (1974) exploration of the time course of categorical speech perception by observing evidence of the speech percept in a state that had not yet "settled" into a discrete category. Our goal is to detect the underlying representations that lead to the patterns of reaction times found by Pisoni and Tash.

Experiment 1: The categorical perception of voicing over time, measured by fixation probability

Methods

Subjects were 16 undergraduate students at Cornell University. They were either paid $5.00 or given course credit for their participation. All were native monolingual speakers of American

Results

In order to provide an accurate picture of the temporal dynamics of categorical perception, eye tracker data were analyzed frame by frame (where one frame equals 33.33 ms). Research assistants viewed the video tapes of the experiment and for each frame they were instructed to record whether or not the subject was looking at or saccading to the buttons labeled "/pa/" and "/ba/" or to neither of them. Sound was not recorded on the videotape, so the coders were unaware of which stimulus the subject was hearing.

Averaging these data across subjects and trials for each VOT yielded a picture of the probability of fixation on a particular choice as a function of time. Figure 3a shows the probability of a fixation to "/ba/", "/pa/" or neither as a function of time after subjects heard a /ba/ with a VOT of –50. Figure 3b shows fixation patterns after an ambiguous VOT of +10. The time course of processing following unambiguous stimuli shows qualitative similarities to results from word recognition (Allopenna, Magnuson & Tanenhaus, 1998). This suggests a common underlying process. In particular, looks to multiple objects immediately after presentation of the stimuli suggest parallel activation of responses, with competition as the disambiguating mechanism.

Averaging the percentage of looks to "/ba/" or "/pa/" (ignoring the looks to neither button) at several time bins as a function of VOT allows us to view the time-course of categorization. Figure 4 shows an identification curve created by using the button the subject fixated on last as a response. It also shows subjects' mouse identifications and their reaction time. The pattern of reaction times is qualitatively similar to the pattern found by Pisoni and

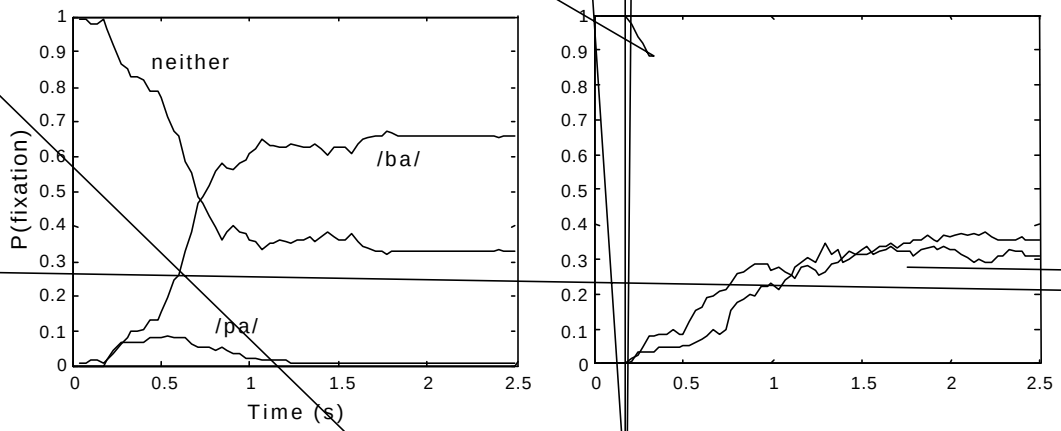

Figure 3: Fixation probability as a function of time for a good/ba/ (VOT=-50, figure 3a) and for an ambiguous stimulus (VOT=10, figure 3b)

parameters that described the logistic function that best fit the identification data for that subject and time. This procedure yielded a new dataset that contained the subject, time slice, and each of the four parameters (essentially a description of the identification function at that time and subject).

By analyzing the effect of time on these parameters we can determine which of our hypothesis is correct. If the no change hypothesis were correct, we would expect to see no effect of time on any of those parameters. If the linear->sigmoid hypothesis were correct we would expect to see a significant increase in slope over time but not in amplitude or the other parameters. If the expanding sigmoid hypothesis were correct, we would expect to see an increase in amplitude but little change in the others. We would not expect to see any change in category boundary or bias.[2]

Four hierarchical regression analyses were performed on each of these parameters to determine how they were affected by time. The first step of the analysis was the addition of 15 dummy codes to the model to capture within subject effects. The second step was the linear effect of time and the third was the nonlinear effect of 1/time (since the scatterplots suggest that the parameters asymptote after a certain amount of time). Scatter plots for each parameter as a function of time are shown in Figure 7 (the dark line represents the mean value).

[2] However these terms suggest that this procedure might be well adapted for better understanding effects like the Ganong effect (Ganong, 1975), where a bias is preferred over the acoustic evidence. Two Consonant Precedence

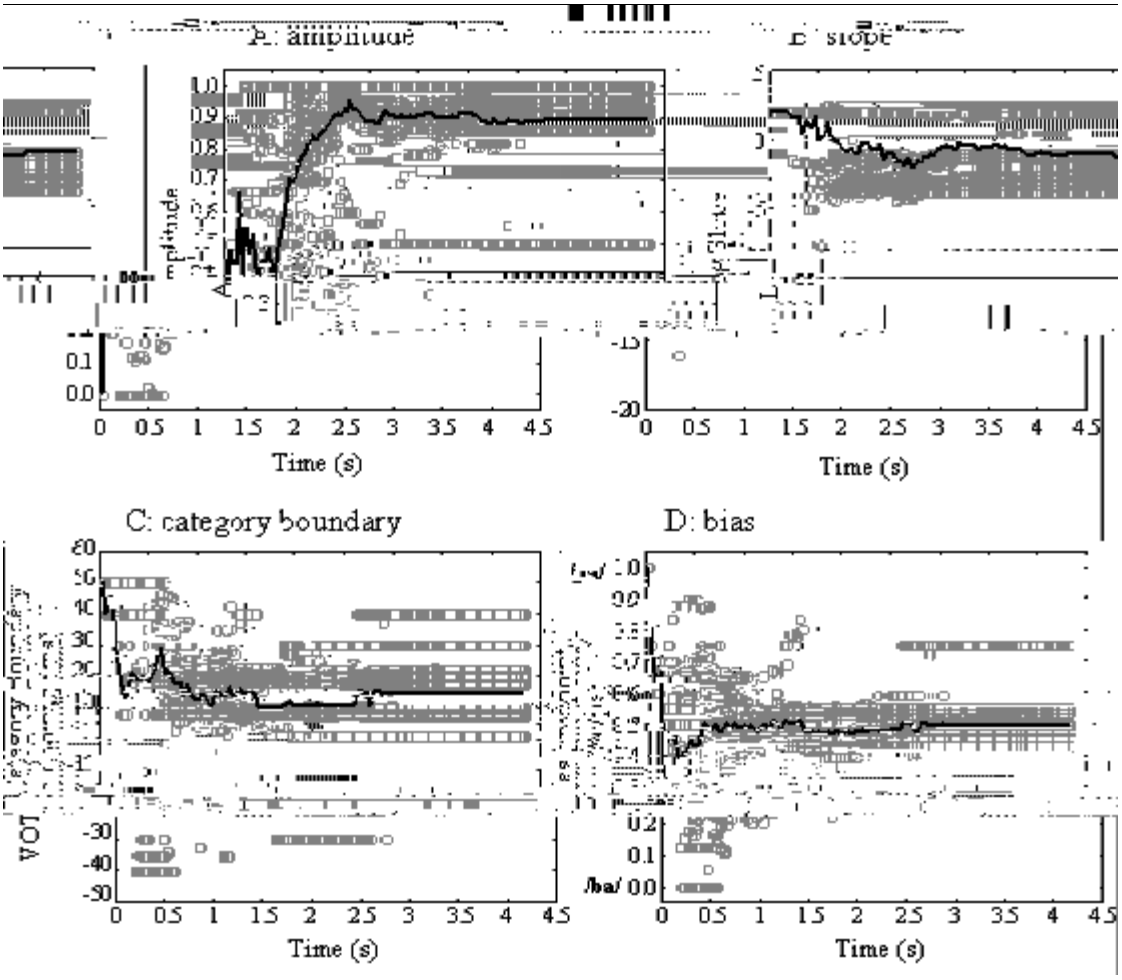

Figure 7: Parameters of the logistic identification function as a function of time. There is substantial change in amplitude over time, but very little in the other three parameters. This suggests the expanding sigmoid hypothesis. Gray circles represent individual data points. The dark line represents the average value over time.

Amplitude (scatterplot and mean shown in Figure 7a) showed a highly significant effect of time. Above and beyond the intrasubject effects, a positive effect of linear time significantly accounted for an additional 11.5% of the variance (β=0.339, t(1945)=19.746, p<.0001). Above and beyond that, 1/time accounted for an additional 8.8% of the variance (total R^2_{model}=0.516). It was significant and negative (β=-.358, t(1944)=18.792, p<.0001)—as time increased, so did amplitude, reaching an asymptote near 1.0.

The second analysis looked at slope as a function of time (scatterplot and mean shown in Figure 7b). Although time and inverse time showed significant effects (time: β=-1.70, t(1945)=10.960, p<.0001, 1/time: β=.189, t(1944)=10.352, p<.0001), they accounted for very little

variance individually (R^2_{time}=.029; $R^2_{1/time}$=.024). Moreover, the slight change in slope that did occur was in the opposite direction of that predicted by the linear->sigmoid hypothesis—identification curves appear to start off steep and flatten out a bit over time.

Regressions for the category boundary (mid_x, Figure 7c for scatter plot and mean) again yielded very small, but significant effects of time and 1/time (time: β=-.068, t(1945)=-3.398, p=.001, 1/time: β=.136, t(1944)=5.623, p<.0001). Although this shift in category boundary was significant, the effect size was extremely small (R^2_{time}=.005, $R^2_{1/time}$=.013) suggesting that overall, there was not much of a shift in category boundary, if there was one at all. An examination of the data within each subject, however, suggested that within a subject there might be a considerable shift in category boundary early in the time-course (although they all arrive at the same boundary by the end of the time-course). To test this intuition, subject by 1/time interaction terms were added and significantly accounted for 28% of the variance (F(15,1929)=72.861,p<=.0001)the cael<fbthee t

t

)

In summary, it seems that each subject has a characteristic starting point with a shifted category boundary and a compensating response bias. Across all subjects, the amplitude of the curve is small initially. Over time, the category boundary and response bias shift towards the center and the amplitude increases. Additionally, any change in slope is likely to be towards a flatter, smoother curve. Of the hypotheses we've proposed, this story fits the expanding sigmoid hypothesis best, with the possibility for some further potentially interesting work on individual differences.

Experiment 2: The categorical perception of voicing over time, measured by a response deadline

Motivation

Since the mid 1990's eye tracking has become an important tool for understanding cognitive processes in which information comes in serially (i.e. word recognition and sentence processing). Many papers have shown that eye-fixations are often very tightly time-

consisted of three components. In the first two, subjects were familiarized with the response deadline to be used during the experiment. Initially, they saw a countdown from 5 seconds followed by the auditory presentation of a prototypical (endpoint) /ba/ or /pa/. Immediately following that, they saw the phrase "respond now". These letters persisted until that block's response deadline at which point they were replaced with the phrase "too-late", and a buzzing sound was heard (the same buzz as if the subjects responded after the deadline).

Following this timing display, subjects were given 10 practice trials with the prototypical sounds. Each practice trial was initiated by a button-press from the subject. This started a countdown from five. At the end of the countdown, one randomly chosen speech sound was presented. Subjects were to press their right button if they heard a /pa/ and the left button if they heard a /ba/. If they responded before the deadline, they were rewarded with a pleasant beep. If they responded too late or not at all, they heard a nasty buzz. During the practice session, the word "Practice" was visible on the top of the screen during all 10 trials.

Testing trials were identical to practice trials with the exception that the word practice was not visible. There were 90 testing trials per block (so the subject heard 10 repetitions of each stimulus item per trial). Subjects were able to respond quickly enough most of the time. In the fastest deadline (400 msec), they responded on time on average 61% of the time. In the slowest deadline (1000 msec), they responded quickly enough 86% of the time.

Results

Identification curves for each response deadline are shown in Figure 8. The data closely resemble the results of Experiment 1, suggesting that the expanding sigmoid is the correct hypothesis. They agree with the previous results to the point that even the noise (e.g., the %/pa/ for a VOT of 50 is greater than that of 60 early in processing) is replicated. They suggest that the eye-tracking methodology and the response deadline are measuring fundamentally the same things: the temporal dynamics of categorization.

The same hierarchical non-linear model was applied to these data as before. Again, a regression analysis was used on the output of this model to determine the effect of time on amplitude, slope, and the midpoint location. Highly similar results were found.

For amplitude, a significant linear effect of time was found (β=.527, t(43)=4.795, p<.0001) over and above within subject effects (R^2_{time}= .263). Moreover, as in Experiment 1, a significant effect of 1/time was also found ($R^2_{1/time}$=.118, β= -1.245, t(42)= -3.653, p=.001). Overall, the model did quite well at predicting amplitude (R^2_{model}=.627). These results replicate the findings of Experiment 1.

For slope, there was no effect of time or 1/time ($F<1$) suggesting that either the response deadline task did not have the sensitivity necessary to detect the slight change in slope we found in Experiment 1 (due to the lack of temporal resolution) or that the regression model used in Experiment 1 was too sensitive (due to the large number of within-subject responses). In either case, this confirms that the linear->sigmoid hypothesis is not likely to be correct.

Similarly, our analysis of the category boundary (mid_x) showed the same lack of effect of either time or 1/time ($F<1$). However, since we did find considerable individual differences in the starting category boundary location, we added <subject X 1/time> interaction terms to the model and found a significant effect ($R^2_{10,32)=2.18UR,p=0486),e}$ (F<1) suggestw (cl used in) TD 0 Tc 0 T-207Tj -325.8 -.04892 in

y ndly tolj 17eybetween(starting category 567 biimeboR=.299R,p=0427imemuchns islocafound in) T

one that fits, we can make a case for the psychological processes that might be responsible for observed temporal dynamics of categorical perception.

Simulation 1: The "no change" hypothesis

The "no-change" hypothesis was instantiated in a simple two layer feed-forward network that learns statistical distributions of its input using competitive Hebbian learning (Rumelhart and Zipser, 1986). This network is a simplified version of the network found in McMurray (in preparation). Forty input and output nodes were used with the input array indicating VOT—lower indexed nodes represented small VOTs and higher nodes large VOTs. Input was in the form of a psuedo-gaussian curve[3] (across the input array) with the mean chosen from a bimodal normal distribution (as per Lisker and Abramson, 1964).

This form of input is compatible with two lines of thought that have begun to emerge. The lateral representation of VOT is compatible with much of the work in population coding which has found topographic representations for a number of dimensions in the visual system including stimulus location and orientation, movement direction and ocular dominance, as well as in the auditory system for interaural time and intensity differences, frequency, and space (see Knudsen, duLac & Sascha, 1987, for a review). Moreover, this selection of each input from a multimodal distribution is compatible with distributional accounts of phoneme learning (Guenther and Gjaja, 1996; Maye and Gerken, 2000; McMurray, in preparation).

The output of the network was "winner take all"—after the output was computed, the node with the highest activation was given all the activation and the rest of them were set to zero (Rumelhart and Zipser, 1986). This is assumed to have happened after many iterations of some sort of lateral competition. Learning occurred after this idealized "competition" using a variant of Rumelhart and Zipser's (1986) unsupervised learning rule and based on the ideas of Hebb (1949).

$$\Delta W_{io} = (I_i * O_o - W_{io}) * \varepsilon \qquad (6)$$

Here, ΔW_{io} refers to the change in weight connecting input node i (I_i) and output node o (O_o). W_{io} refers to the current connection strength and ε is the learning rate (set to .1 for this simulation).

After 5000 training epochs, the network correctly learned to categorize VOTs into one of two categories. This categorization was in the form of one of the 40 output nodes representing voiceless sounds and one representing voiced sounds. Although it may seem that only two nodes were needed to represent two categories, previous research (McMurray, in preparation) has suggested that additional output nodes are necessary for the learning process—without them, the network has a tendency to group all the stimuli under a single output node.

[3] The actual form was a squared Gaussian curve. This introduced some kurtosis into the activation patterns. This creates a greater difference near the endpoints between two curves centered at different locations. In a sense this helped the network generalize the information from the center of the continua to the endpoints by increasing the difference between a prototypical /pa/ curve and /ba/ curve at the end. Neurobiological evidence from the population coding literature suggests some form of Gaussian curve as the activation function for population codes (Knudsen, du Lac, and Esterly, 1987). However, there is not enough data to rule out a squared Gaussian function.

Although this model did show the correct categorization behavior, it showed none of the temporal dynamics we were interested in. Rather it demonstrates the simplest possible network that could learn phoneme information from the distributional properties of its linguistic environment. In short, this network was extremely sensitive to the statistics in its learning environment but had no temporal processing. Because of this, it could not fully model the data.

Simulation 2: The linear to sigmoid hypothesis

To model the linear to sigmoid hypothesis, the Normalized Recurrence Network was used (McRae, Spivey-Knowlton & Tanenhaus, 1998; Spivey & Tanenhaus 1998). This network consists of two input nodes and two output nodes. Input nodes indicated the probability that the VOT is a /pa/ or a /ba/ (with the decision indicated by the output nodes). A very good /pa/ for example might have 0.9 and 0.1 as the activation for its input nodes, where a /ba/ would be 0.1 and 0.9 and an ambiguous stimulus would have 0.5 and 0.5. At each round, the sum of activation at either level can only be 1.0 so each node's activation is divided by the total amount of activation for that level.

WPLhtheracat

This network does categorize correctly (see Figure 9a), since the categories are built into the architecture. It also shows the appropriate increase in reaction time at the category boundary (as per Pisoni and Tash, 1974) since it takes a greater number of iterations for the competi

Simulation 3: The expanding sigmoid hypothesis

Our third hypothesis, the expanding sigmoid was instantiated in a synthesis of the two networks presented previously, the Hebbian Normalized Recurrence Network. Since this hypothesis was the one found to best describe the data, we'll spend more time on this network than the others.

The Hebbian Normalized Recurrence Network combines the representation and learning of the two-layer network of Simulation 1 with the processing algorithm of the normalized recurrence network. Like the network in Simulation 1, the Hebbian normalized recurrence network has 40 inputs and outputs. The input layer is organized topographically, and for any perception event, the input is chosen in the same manner (from a bimodal normal distribution). After activating the input nodes, the network then uses the following algorithm:

1) The inputs are normalized so that they sum to 1.
2) The outputs are computed by multiplying the inputs by the weight matrix and adding that value to the current value.
3) Activation in the output layer is squared.
3) The outputs are normalized so that they sum to 1.
4) The weights are modified using the Hebbian Learning Rule.
5) The output is multiplied by the weight matrix transposed and this value is multiplied by the input vector and added to it.
6) This repeats until the average change in the outputs is less than .00001.

This algorithm is similar in nature to that of the normalized recurrence network, except that the addition of the weight matrix allows for a topographic representation of the input, as well as the possibility for learning. Rather than simply passing activation directly from input to output, this operation is mediated by the weight matrix (or the transpose of the weight matrix if we are passing activation backwards from output to input).

The only deviation from the processing architecture of the normalized recurrence algorithm is in step 3 where the activation in the output layer is squared. The reason this was done was that without some nonlinearity in the output layer's activation function, the network never learns to map regions of the input onto a single category. Some form of lateral inhibition seems like a psychologically plausible choice for this nonlinearity. In Simulation 1, we used winner-take-all learning for this, but if used here, the simulation would stop after only one processing cycle—preventing a model of temporal dynamics. Instead, we had to develop a more gradual form of lateral inhibition that we call quadratic normalization (quadratic because the activations are squared, normalization because the output is normalized to sum to one after squaring). This algorithm is based on a relatively neurologically plausible model of lateral inhibition and we direct the interested reader to Appendix B for a more thorough explanation and derivation.

Results

After training for 3000 stimuli on a bimodal normal distribution with means at 30 and 70, the Hebbian Normalized Recurrence network approximates the expanding sigmoid hypothesis quite

The learning algorithm has been shown to have a close correlate with long term potentiation a process by which synaptic connection strengths increase after concurrent firing by pre- and post-synaptic neurons. We have derived the activation function (quadratic normalization) from a relatively simple model of lateral inhibition (Appendix A). Finally, the competition algorithm is based on Heeger's work (1993) in neural interaction, and so is also supported neurologically. For these reasons, this "breed" of connectionist models may be an excellent way to explore the effects of both learning and processing.

Hebbian Normalized recurrence not only predicts the correct time-course of perception, but also provides testable predictions about the ends of the continua (and potentially about development). It represents a combination of a statistical learning device and a competitive processing device (neither of which could fit the data on their own) and suggests that both of these processes may underlie the categorization of speech sounds.

Categorization in Infants

A major goal of the work we've discussed so far has been to convince the reader that 1) time should be an important dimension for characterizing speech perception and 2) eye-

However, soon sucking becomes less interesting because the auditory stimulus never changes and they start to habituate. At this point the stimulus is changed to something else. If the baby can discriminate this sound from the old one, then the baby will become more interested and his or her sucking rate will increase (dishabitauation), otherwise it will decrease.

These methodologies do not permit a simple linking hypothesis between the data they provide and the underlying infant speech processing mechanism for several reasons. The first is that they rely on the infant's response to changes in a stimulus rather than to a single stimulus. This more closely parallels discrimination than identification data, a much more complex process to attempt to model. These techniques also confound the issues in that they are really studying infant perception *after* they have heard many repetitions of the same stimulus. This, of course, may lead to the possibility of adaptation effects. Finally, these techniques do not permit us to study speech in any sort of ecologically valid way. It is clear that identification of speech sounds (as opposed to discrimination) is the more essential element of word recognition, and the presentation of multiple stimuli may enable the infant to rely on acoustic representations that are not available during on-line (single-presentation) speech processing.

In this last section of the paper, we describe a new technique that employs anticipatory eye movements after a brief training session to measure speech categorization in 4 and 5 month old infants. This new methodology promises to not only overcome the methodological hurdles I have mentioned but also to yield information on the temporal dynamics of infant speech processing.

Methodological background

Clearly habituation and high-amplitude sucking will not provide the sort of data we need to explore infants' categorization abilities. However, two other common infant methodologies offer features that may overcome the problems of high-amplitude sucking and habituation.

The visual expectation paradigm (Canfield, Smith, Brezsnyak & Snow, 1997; Haith, Wentworth & Canfield, 1993) has been used extensively to explore the nature of expectancy in infants. To oversimplify it, when an infant is presented with two alternating side-by-side lights (or images), he or she is able to learn the pattern very quickly (Haith et al, 1993, report 11 trials). On

rewarded if he turns his head after hearing /i/. These sorts of studies are very difficult to do because head turns are both costly (metabolically) to perform and quite noisy. Since the infant is responding based on multiple presentations of the background stimulus, it is much more difficult to compare results to that of adults (who hear only one) and also introduces the possibility of adaptation effects or the building of completely different representational schemes than the ones used in o

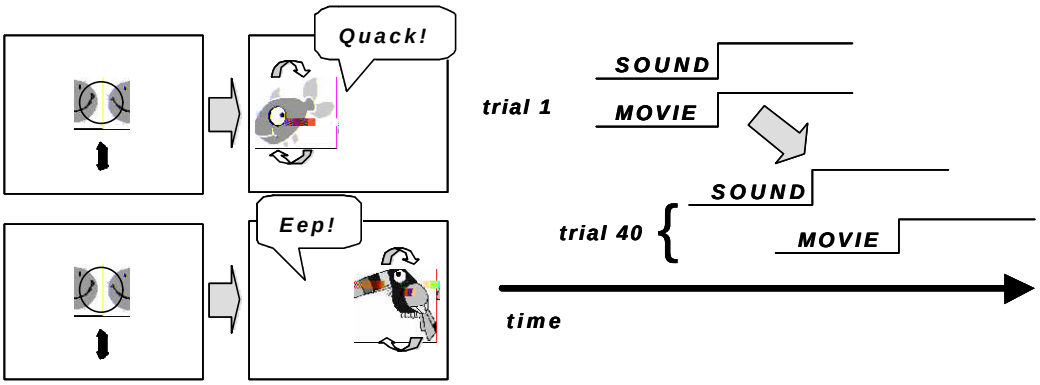

Figure 11: The Anticipatory Eye Movement Paradigm. Each trial begins with a vertically moving 'smiley face' to orient the infant's gaze to the center of the screen. The face is removed and one of two randomly selected sounds is presented simultaneously with a short animation on one side of the screen. The sound is consistently paired with the side on which the animation is presented. As the experiment progresses, the delay between the sound and the movie increases. By the end of the experiment, infants are making anticipatory eye movements during this delay period.

Training consisted of 40 repetitions of this basic cycle. To keep participants interested during training, training was broken up into two blocks. Between the training blocks the infant was given a short break during which the parent was encouraged to talk to and play with the baby. Within each block the training cycles were periodically (every 5 or so trials) broken up by presenting a brief visual display consisting of moving smiley faces, expanding and contracting psychedelic shapes and other images that served to break the monotony.

During training the time interval between the presentation of the sound and the presentation of the movie was gradually increased to a maximum delay (approximately 2 seconds, though it varied from Experiment 3 to Experiment 4). If the infant learned which side of the screen is associated with each sound, he or she should make eye movements toward that side *before* the presentation of the visual stimulus.

After 40 training trials, the infant is tested until he or she becomes too fussy to continue (usually about 10-15 trials). Testing was quite similar to training. The infant heard one of the sounds, and eye movements during the delay between auditory and visual stimulus were recorded.

Recently we have incorporated a computerized eye and head tracking system to increase the accuracy of our eye movement measurements (they used to be coded by observers viewing a video tape of the babies head) and the ease in collecting and scoring data. For an overview of this system please see appendix C.

Experiment 4: A preliminary test of speech sounds categorization

Introduction and Methods

Experiment 3 provided evidence that the anticipatory eye-movement paradigm could be used as a two-alternative-forced-choice measure of sound categorization in infants. Thus, for Experiment 4, we examined, a /pa/->/ba/ continuum. A five-step continuum was created using the Sensimetrics, Inc. implementation of the Klatt synthesizer with VOTs ranging from 0 to 40 msec. Infants were trained on the endpoints. Because we felt that the stimuli might initially be hard to discriminate[5], during training /pa/ was initially presented with an artificially heightened F0 (1200 hz) and /ba/ with an artificially lowered one (800 hz). This gave the infants an additional cue to aid in their discrimination during training. Over the course of training, the fundamental frequencies of these sounds were gradually brought together to reach 1000hz for both so that during testing this cue was no longer helpful.

A preliminary analysis of the timing of the saccades in Experiment 3 indicated that many of the incorrect saccades were actually occurring *after*

t h r a l d o m p r a r e e i e r t o s e n c

Since the ba/pa categorization task was a much more difficult task than the quack/eep categorization task, we employed a criterion measure for including a participant's data: participants had to look in the correct direction last for at least 60% of the endpoint-trials (this criterion is similar, though lower, to the criterion employed in many psychophysical tasks with adults). 50% of our babies (7/14) reached this criterion.

We were able to generate identification curves (Figure 13) similar to those from adult categorical perc

By this account infants are extracting aspiration information about whether or not the sound is voiced early in the stimulus. This suggests a voiceless sound. Later, glottal pulses are detected and the infant arrives at the correct conclusion. For a /ba/, the result of this is that the infants initially perceive the sound as /pa/ (due to aspiration) and correct themselves later. For a /pa/ the two cues are not in conflict, and they correctly respond /pa/ throughout the time-course. Of course,

The anticipatory eye-movement paradigm on the other hand makes use of only a single stimulus presentation, and only the endpoints for training. Any generalization that we see is the result of a natural similarity metric or categorization. Rather than building this similarity into the training procedure, training merely provides reference points with which to explore this similarity structure.

This paradigm has also begun to prove itself useful to look at visual stimuli. In a series of ongoing experiments in our lab, we have replaced the centering stimulus (the smiley faces) with either a cross or a square. The few subjects we have run in this pilot experiment have learned to categorize these shapes successfully. Given this preliminary success, this application of the anticipatory eye-movement paradigm will also allow us to ask a host of questions regarding visual perception. For example, we might train infants to categorize circles and squares and then present them with a continuum. Alternatively, we could train them to categorize circles from "pacmen" (circles with a 90 degree arc removed) and then look at issues of object occlusion.

In short, the methodologies presented here provide a very simple, natural way to look at infants' internal representation of auditory (and potentially visual) stimuli with few of the problems inherent in other methodologies. Moreover its use of eye movements provides a window into the temporal dynamics of infant categorization, which, as we've discussed, may become an important criterion for characterizing infant cognition.

Conclusions

Through the interplay of adult and infant experimentation with simulation, this work directs our attention to an aspect of categorical speech perception that has been all but ignored: temporal dynamics. It is clear that we cannot ignore this aspect of perception much longer for several reasons. In the case of speech, the word recognition system is clearly a system that is "pressed for time" in that it must process a large amount of information in a very short time. Because of this it may rely on incomplete phonetic representations of incoming speech, and the temporal dynamics of sublexical processing may provide a clue as to what those representations may look like. In the case of categorical perception in other modalities, we may find that phenomena that looked categorical in their end-state may be quite different in their intermediate states. The combination of empirical analysis of the temporal dynamics and good computational modeling may tell us what these differences mean psychologically. Moreover, although we have used categorical perception as a testing ground for this approach, it is certainly applicable to many other areas of perception.

From the adult eye movement and response deadline data, it appears that the continuous

sensitivity. A neural network that combined competitive Hebbian learning (Rumelhart & Zipser, 1986) with a "settling" algorithm (Spivey & Tanenhaus, 1998) provided the only satisfactory account of these data. This network also opens the door to further explorations, both in the extreme ranges of VOT, as well as in issues regarding the development of speech perception (McMurray, in preparation).

These findings and their accompanying simulations have broad implications for speech processing and phonetics in general. If it actually takes a few hundred milliseconds to discretize one's percept of a potentially noisy consonant, speech recognition is an even more convoluted process than initially suspected. Before one phoneme is fully categorized, the next few are already being received as input. Mapping such a string of multiple partially active and mutually exclusive phonemic representations onto possible lexical items will no doubt be a massively parallel process. This perspective lends support to feature-based parallel-activation models of speech recognition (e.g., McClelland & Elman, 1986), in which phonetic features are not binary but exhibit graded activation levels. Thus, treating phonetic representations as discrete logical symbols may be useful for idealized instances where noise and interfering signals are absent. However, when speech perception is considered in realistic noisy environments, with the real-time accrual of acoustic-phonetic input being faster than the real-time classification of that input, phonetic representations will have to be treated as probabilistic representations.

On a different note, the anticipatory eye-movement paradigm presented here may be just the tool to begin exploring these issues throughout development. We have shown that this methodology can be used analogously to the two-alternative forced choice task used with adults, and provides a unique window into infant categorization by allowing us to obtain discrete responses for a single stimulus. Moreover, it appears to be quite sensitive to the temporal dynamics of perception and will likely prove a useful tool in looking at two scales of time in perception simultaneously.

As we have shown, the methodological and theoretical developments presented here are built off a wide range of work in adult psycholinguistics, phonetics, and developmental psychology. However, this line of research represents a unique approach to studying cognition and advocates a new focus on the temporal dynamics of perception. Perhaps speech perception is best seen as a dynamical system, and our goal as researchers is to uncover the parameters of this system, their meaning, and the way in which they are set.

Acknowledgements

The authors would like to thank Tobey Doeleman for help with the speech synthesis, Michelle Spence, Rebecca Mabie and Melinda Tyler for coding the data, Harry Reis for help with the curve fitting and Sabine Hunnius for assistance with the magnetic head tracker. We are also grateful to Michael Tanenhaus for help in developing the response deadline methods used here and for providing the Magnetic Head Tracker. Experiment 1 was presented at the ChiPhon panel of the 33rd meeting of the Chicago Linguistics and can be found in McMurray and Spivey (1999). Experiment 3 and 4 were presented as a poster at the 2000 International Conference on Infant Studies (McMurray and Aslin, 2000). Supported by grants from the National Science Foundation (SBR9873427) and the National Institute of Health (HD37082) to R.N. Aslin, a Sloan Fellowship to M. Spivey, and a National Science Foundation Grant (SBR9729095) to M. Tanenhaus.

References

Allopenna, P., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38(4),* 419-439.

Beale, J., and Keil, F. (1995). Categorical effects in the perception of faces. *Cognition, 57,* 217-239.

Bornstein, M.H, and Korda, N.O. (1984) Discrimination and matching within and between hues measured by reaction time: Some implications for categorical perception and levels of information processing. *Psychological Research, 46,* 207-222.

Canfield, R., Smith, E., Brezsnyak, M., and Snow, K. (1997). Information processing through the first year of life: A longitudinal study using the visual expectation paradigm. *Monographs for the Society for Research in Child Development, 62(2).*

Cohen J.D., MacWhinney B., Flatt M., and Provost J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25(2),* 257-271.

Eimas, P., Siqueland, E., Jusczyk, P., and Vigorito, J. (1971). Speech perception in infants. *Science, 22,* 303-306.

Guenther, F., and Gjaja, M. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100,* 1111-1121.

Haith, M., Wentworth, N., and Canfield, R. (1993). The formation of expectations in early infancy. *Advances in Infancy Research, 8,* 251-297.

Harnad S., ed. (1987). *Categorical Perception: The Groundwork of Cognition.* New York: Cambridge University Press.

Hebb, D. (1949). *The Organization of Behavior*. New York: Wiley.

Howard, D., Rosen, S., and Broad, V. (1992). Major/minor triad identification and discrimination by musically trained and untrained listeners. *Music Perception, 10(2),* 205-220.

Humphreys, G. (1981). Flexibility of attention between stimulus dimensions. *Perception and Psychophysics, 30(30),* 291-302.

Jusczyk, P. (1997). *The Discovery of Spoken Language.* Cambridge MA: The MIT Press

Kluender, K. (1994). Speech perception as a tractable problem in cognitive science. In Gernsbacher, M. (Ed.) *The Handbook of Psycholinguistics.* London, UK: Academic Press.

Kluender, K., Diehl, R. and Killeen, P. (1987). Japanese quail can learn phonetic categories. *Science, 237(4819),* 1195-1197.

Knudsen, E., du Lac, S., and Esterly, E., (1987). Computational maps in the brain. *Annual Review of Neuroscience, 10,* 41-65.

Kuhl, P. and Miller, J. (1975) Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190(4209),* 69-72.

Kuhl, P. and Padden, D. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception and Psychophysics, 32(6),* 542-550

Lamberts, K., and Brockdorff, N. (1997). Fast categorization of stimuli with multivalued dimensions. *Memory and Cognition, 25(3),* 296-304.

Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54(5),* 358-368.

Lisker, L., and Abramson, A. (1964). A cross language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384-422.

MacDonald, M., Pearlmutter, N. and Siedenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101(4),* 676-703.

Marslen-Wilson, W. (1975). The limited compatibility of linguistic and perceptual explanations. In *Papers from the parasession on functionalism.* Chicago Linguistic Society.

Massaro, D., and Cohen, M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication, 2(1),* 15-35.

Maye, J. and Gerken, L. (2000). Learning phonemes without minimal pairs. *Proceedings of the Boston University Conference on Language Acquisition, 24.*

McClelland, J. and Elman, J. (1986) The TRACE model of speech perception. *Cognitive Psychology, 18,* 1-86.

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: a time-course analysis. *Journal of Memory & Language, 32(4),* 536-571.

McMurray, B. (in preparation). The Hebbian Categorization Architecture: A neurologically plausible connectionist account of the development of speech perception abilities.

McMurray, B., and Aslin, R. N. (2000). Anticipatory eye movements: A technique for assessing auditory categorization in infants. Poster at the Internation Conference of Infant Studies, 2000. Brighton, UK.

McMurray, B., and Spivey, M. (1999). The categorical perception of consonants: The interaction of learning and processing. *Proceedings of the Chicago Linguistics Society, 34(2).*

McQueen, J. (1996). Phonetic Categorization. *Language and Cognitive Processes, 11(6),* 655-664.

McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1998). Modeling the effects of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 37,* 283-312.

Miller, J. (2000). Mapping from acoustic signal to phonetic category: nature and role of internal category structure. *Proceedings of the Workshop on Spoken Word Access Processes.*

Ohlemiller, K., Jones, L, Heidbreder, A., Clark, W., and Miller, J. (1999). Voicing judgments by chinchillas trained with a reward paradigm. *Behavioral Brain Research, 100,* 185-195.

Phillips, C., Marantz, A., Yellin, E., Pellathy, T., McGinnis, M., Wex1 10.2 TTj 243 0 TD -0.0297 Tc -0.1203 Tw (M., Wex1 10.2 T

Appendix A: A hierarchical nonlinear model used for the analysis of categorization data over time

The identification function found in categorical perception experiments can be described by the logistic function whose formula is given in (1).

$$\text{logistic}(x) = \frac{b_1}{(1 + e^{(-1(b2*VOT + b3))})} + b_4 \qquad (1)$$

For a categorization experiment, b_1 is typically assumed to be one and b_4 to be zero. Then, changing b_2 changes the slope or steepness of the function and b_4 the location of the category boundary. Only varying two of the parameters, however, prevents this function from approximating the identification curves a few hundred milliseconds after stimulus presentation, because the distance between the upper and lower asymptotes is fixed at one. Typical logistic regression models, for example, only use these two parameters. As a result they are unable to account for models like the expanding sigmoid hypothesis.

In this analysis, we have allowed all four parameters to vary so that we may better fit our data, and explore what happens to the identification curve over time. Although the addition of this

fit. In order to perform a maximum likelihood estimation of the best parameters for the data, this Equation (2) was used to predict the mean probability of a Bernoulli distribution (3).

$$P(\text{ looking to /pa/ } | \text{ VOT}) = \text{logistic(VOT)}^{X} * (1 - \text{logistic(VOT)}^{(1-X)}) \quad (3)$$

X is 1 if the subject made an eye movement to /pa/ and 0 if he or she looked to ba. Thus, for each time by subject pair, the likelihood function of the eye-fixation data is given by (4).

$$L = \prod_{VOT} \text{logistic(VOT)}^{Pvot} * (1 - \text{logistic(VOT)}^{Bvot}) \quad (4)$$

Here, P_{vot} is the number of looks to /pa/ at that VOT and B_{vot} is the number of looks to /ba/ at that VOT. The product was taken over the data points (looks) from all VOTs for a given time-slice and subject. The maximum likelihood estimator started by searching a number of points in the parameter space for the starting point with the largest log-likelihood. It then used a constrained gradient descent algorithm to minimize the log of this likelihood function over the parameters of the logistic function: amp, slope[7], mid_x and mid_y. The algorithm was constrained so that all possible values of the logistic function were between 0 and 1 (otherwise, the function could not approximate a probability).

This analysis can be compared to hierarchical linear modeling in the sorts of statistical questions it is able to address for nonlinear functions (in this case, the 4 parameter logistic function although other functions could be used). It represents a new and potentially very powerful way to analyze categorical data.

Appendix B: Why quadratic normalization?

A

activation function, let us first define how inhibition is going to work. We will assume that the inhibited activation of node x is equal to the uninhibited activation minus some function of the other nodes.

$$O'_x = O_x - f(O_{1\ldots n}) \quad (5)$$

We'll also define the proportion of the total activation held by any output node X to be

$$P(O_x) = Ox/A \quad (6)$$

where A is the total activation in the array of nodes. We now shall define inhibition such that the if O_x is receiving the inhibition, and O_y is the inhibitor, then the effect of O_y on O_x will be

$$O_x = O_x - O_x P(O_y) \quad (7)$$
$$O_x = O_x - O_x(O_y/A) \quad (8)$$

Thus the inhibitor will take its proportion of the total activation from the node being inhibited. For example, if O_x = 0.2 and O_y = .8, (and there are only two nodes), then O_y has 80% of the total activation. The effect of O_y on O_x will be to remove 80% of O_x's activation, leaving it with .04. By this definition, the activation of a node after being inhibited by *every* other node will be

$$O'_x = O_x - \sum_y O_x(O_y / A) \quad (9)$$

This equation isn't completely correct, given that a node cannot inhibit itself. Thus we get

$$O'_x = O_x - [\sum_y O_x(O_y / A) - O_x(O_x/A)] \quad (10)$$

Now a little algebraic manipulation

$$O'_x = O_x - [\ ((O_x/A) \sum_y O_y\ ) - O_x^2/A\ ] \quad (11)$$

relatively useful nonlinear activation function. Since in this particular network, we are normalizing the vectors to sum to one at each term, we can eliminate the A in the denominator simplifying the equation even further.

Appendix C: Computer-based eye tracking in infants

The setup

Although computer-based eye-tracking technologies have been around for quite some time, the application of these methods to infants in any kind of non-intrusive, simple fashion has begun only recently.[8] Because of this, it may be of some value to include a description of the methods that we are successfully using.

Since existing head-mounted eye-trackers are too big (and heavy) for infant use (and even if you could get one to fit, and find a baby willing to wear it, it is probably not a good idea to put a $15,000 piece of equipment on a baby's head even if you could), we use a remote eye-tracking system which sits on the table in front of the infant (we used the ASL Pan/Tilt model 504). This small camera emits and detects infrared light enabling us to run the experiment in the dark, when the infant's pupil is largest and there is the least amount of distraction. The camera is able to move in response to the user's commands on a joystick or the computer console to maintain a constant close-up view of the infants' eye. From this close-up view of the eye, the eye tracker locates the pupil and corneal reflection which, after calibration (which will be discussed shortly), allow it to determine where the subject is looking.

The software that operates the eye tracker contains an algorithm that attempts to maintain the eye in the center of the camera's view at all times. This allows it to compensate for small head movements. Unfortunately, the average head movement of an unconstrained infant is much bigger than this allows. Although encouraging the parent to constrain the infant's head movements (by holding them gently below the chin or by holding a pacifier in the infant's mouth) minimizes this head movement, it is often not enough, as many infants do not like having their head constrained. However, the ASL remote eye tracking system includes a simple interface (right out of the box) with several Magnetic Head Trackers (MHTs). This allows the eye camera to use real time information about head position to regain the eye when it is lost. Although we use the Polhemus FastTrak, they all work similarly.

The MHT consists of a small magnetic receiver that we attach to an infant knit cap with Velcro. A larger transmitter is mounted behind the infant in the room. The MHT receiver is able to detect its position (x,y,z) and orientation (elevation, azimuth, roll) relative to the transmitter. This information is relayed to the eye-tracker that can then use it to find the eye when it is not in the view of the camera. Before using the MHT, it must be calibrated for the room and the eye tracker, although this can be done once and saved. This allows the eye-tracker and the MHT to share a

[8] The interested reader may want to visit

common coordinate system. The head tracker must also be calibrated for the subject so that the eye tracker knows how far away the sensor is from the eye.

The layout and connectivity of equipment in our lab is shown in figure 15.

Calibration and Use

At the beginning of the experiment the infant is seated with its parent and the infant cap (with the MHT receiver) is placed on its head. The system seems to work best if the receiver is directly above one of the eyes (the eye we will track). We then play a short movie to engage the infant while the experimenter manually finds the eye with the eye-tracker. At this point the experimenter calibrates the head tracker (by pressing a single key). This records the angle of the eye tracker and the position of the receiver in space, allowing the eye-tracker to compute a vector difference, which it can use to zero in on the eye from head tracker coordinates. At this point the autotrack feature of the eye tracker is enabled and the eye tracker follows the eye for the rest of the experiment.

The experimenter then adjusts pupil and corneal reflection detection thresholds until the computer is able to locate both. The next step is to calibrate the eye tracker. For most eye trackers, calibration consists of having the subject look at specific known points in space and recording the pupil and corneal reflection locations while they are looking there. For most eye tracking applications, nine points are used. However, infants rarely complete such a task. Instead, ASL offers a two point "quick calibration."

To perform this calibration, we show the infant a set of colorful concentric circles that are expanding and contracting at a 1 second frequency. These circles change color frequently and are accompanied by a phase locked frequency-modulated tone. Infants seem to like watching this for up to a minute or more. We first present these circles in the top left of the television and record the pupil and corneal reflection locations. We then move them to the bottom right and do the same. At this point the infant is calibrated and we begin recording data.

built sensors that detect sound (on the unheard right channel) or light (on an occluded portion of the screen), and insert numerical codes into the data stream. Thus, we can mark trial onsets by playing a tone (that the infant will not hear), or displaying a light (that the infant will not see) to these sensors using our experimental control software. We use Psyscope to control the timing of the stimulus and data signals although nothing in this system depends on that choice. In addition, the ASL hardware permits a parallel port connection to the eye tracker that allows a direct connection between the experimental control computer (and software) and the eye tracking data stream.

Apart from an explicit recording of fixation coordinates, the eye tracker also outputs (in real time) fixations as video in the form of a set of crosshairs superimposed on the image of the screen the subject (infant) is watching. This is mixed with the image from an infrared video camera that is focused on the infants' head and recorded on normal VHS tape. From this tape, coders can code fixations from either the eye tracker's crosshairs, or, when the eye track is lost, from a view of the infant's head and eyes. Very shortly we will be adding a third image to this tape: the close-up image of the eye from the eye camera. The infrared camera also allows the experimenter to see the subject while the experimenter is running.

The combination of the MHT with the remote eye tracking system seems to allow us a more or less constant data stream of fixations in screen coordinates. Fixation can be determined analytically, taking much of the guesswork out of eye tracking research with infants. When combined with the anticipatory eye movement methodology it is easy to see the potential power of this technology.

UNIVERSITY OF ROCHESTER WORKING PAPERS IN THE LANGUAGE SCIENCES – VOL. 1, NO. 2

Katherine M. Crosswhite and James S. Magnuson, Editors
Joyce Mary McDonough, Series Editor

Jean Ann and Long Peng: *u2E3ENCE91 head and56c 0.15 Bob (Murraaary McDoi) Tj 93 0:aa( )Mh6p3I.eFo4a6t:aa(TD af0 TD /F0 9.6 Tf 0.0824 ohd and ey*