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 categories. This research has been guided by the underlying belief that if the stimuli are varied in

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 (e.g. synthesized speech) and experimental design that have provided a large body of research from which to build. Indeed, Kluender (1994) has argued that in part because of this approach, speech perception is one of the more tractable problems in cognitive science.

These methodological improvements in fine-grain stimulus generation have not been equally matched by methodological improvements in fine-grain response measures. There has been little development of the standard dependent measure, in which a listener provides a metalinguistic judgment a second after the speech perception event has taken place. The events that occur during the 500-750 milliseconds between auditory transduction and the behavioral response remain an entirely open question.

A similar state of affairs once existed in the field of sentence processing. Throughout much of the 1970's, the dominant method used by researchers to infer the structure of the human sentence processing mechanism was to present a sentence (or series of sentences) and subsequently test the participant's memory for certain aspects (e.g., syntax, semantics) of the sentence. Then, at the 1975 Chicago Linguistic Society meeting, Marslen-Wilson (1975) argued convincingly for developing measures of sentence comprehension that tap representations and processing *during* the comprehension event, rather than after it. This motivation has driven the field of sentence processing for the past two decades, and has resulted in an extremely rich understanding of the possible mechanisms by which a listener/reader integrates linguistic structure and linguistic content in real-time (for recent reviews, see MacDonald, Pearlmutter & Seidenberg, 1994; Tanenhaus & Trueswell, 1995).

In the work presented here, we apply that same logic in experimental design to the study of low-level speech perception. Our goal in this research is to demonstrate that although discrete/symbolic phonetic representations are useful descriptions of the ultimate reportable percept during speech perception, a considerable amount of information is available in the intermediate representations that get computed along the way toward that final state. That is, we do not wish to discount the importance of categorical representations of speech (the existence of sharp categorization functions for speech perception is difficult to contest), we merely wish to convince the reader that "getting there is half the fun."

Why study the time course of speech perception?

Synthesized speech has allowed the field to examine the psycho-acoustic microstructure of speech perception mechanisms. Using such speech, experimenters have been able to generate continua of speech sounds that vary in small steps by a single acoustic feature. Research with these continua has found overwhelming support for a very sharp category boundary between phonemes along many acoustic continua. This sharp boundary (along with the fact that discrimination between

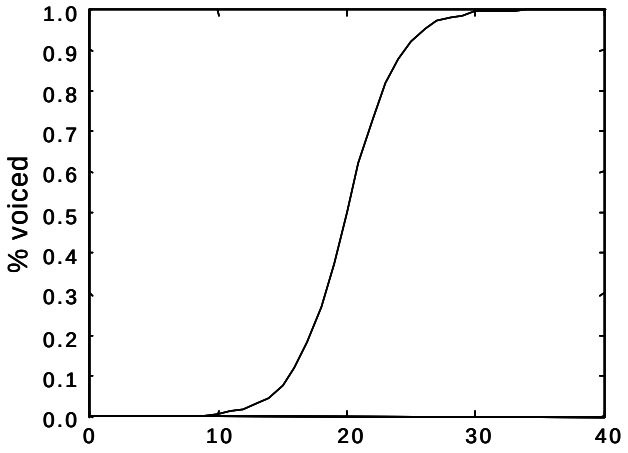

Figure 1: A schematic identification curve. Percentage of stimulus identified as voiced as a function of VOT is plotted, although similar curves have been reported for many other features of speech such as the acoustic cues signaling place of articulation.

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 end-state of categorization, it is possible that they result from different patterns of change over time. By exploring the temporal dynamics of categorical perception in these different modalities and subjects, we may find differences between these categorization abilities. The combination of this type of research with explicit computational models, may allow us to draw conclusions regarding the mechanism or mechanisms behind categorical perception.

More importantly, categorical perception has also been invoked in explaining the speech perception abilities of infants. Although it is unclear at this point how to assess identification, discrimination data have suggested that human infants might at least discriminate speech sounds categorically (Eimas, Siqueland, Jusczyk and Vigorito, 1971; see Werker & Polka, 1993 for a review). These conclusions are tenuous for two reasons. The first is that these studies are based on habituation measures that can only measure discrimination (and may be subject to selective adaptation processes, since the infant is hearing many repetitions of the same stimulus). Unfortunately at this time, there have been no procedures that allow us to obtain identification data from infants or allow a response after the presentation of only a single stimulus (thus allowing a direct comparison with adult data). The second reason to doubt these findings (and more important to our argument here) is that even if we had a measure of identification based on the percept of a single stimulus, it is quite possible that the end-state of perception is quite different than the intermediate stages. These well-replicated findings have encouraged the view that speech perception is a modular or special process. However, the study of the temporal dynamics of perceptual processes may permit a new dimension on which to compare categorical perception in speech and non-speech modes and between infants and adults, if we are able to develop the appropriate methodologies.

In this paper we will present preliminary data from several studies that attempt to elucidate the temporal dynamics of low-level speech processing. We will present data from several adult studies involving the time course of phoneme categorization, and explore several possible models for the mechanism that might account for our data. We will then present preliminary data from a series of infant studies that potentially solve the methodological pitfalls outlined above. Although they do not demonstrate conclusive data on infant speech perception the methodology clearly holds promise.

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 representation (Utman, Blumstein, & Burton, 2000; Massaro and Cohen, 1983; Pisoni and Lazarus, 1973; Samuels, 1977), in many experimental situations, the actual VOT of the individual stimulus appears to be discarded, and all that remains in the percept is category membership.

Nonetheless, something very interesting is going on during those several hundred milliseconds between stimulus offset and the reporting of the percept. The first hint at this came from work by Pisoni and Tash (1974), in which they replicated the basic categorical identification of VOT and also recorded participants' reaction times. Stimuli in the ambiguous region of the VOT continuum elicited longer reaction times than the unambiguous phonemes. Thus, categorization of a particularly ambiguous stimulus takes longer than categorization of a less ambiguous stimulus. This is consistent with the kind of settling, or pattern completion, process seen in attractor networks.

As previously mentioned, the end product of the categorization process is clearly a steep logistic identification function of a continuous acoustic value (such as VOT). The initial shape of the identification function and it's transformation to this steep logistic is much less clear. At this point, we can formulate three hypotheses for the time course of categorical speech perception:

Hypothesis 1: No Change -- The initial state of the speech percept is no different from the end-state, i.e., categorical speech perception is instantaneous and exhibits no temporal dynamics whatsoever.

Hypothesis 2: Linear to Sigmoid -- The initial state of the speech percept may retain the continuous nature of the synthesized stimuli, as in the top left panel of Figure 2. This continuous perception might gradually warp itself over time (t) into a steep sigmoid, as in the top middle and right panels.

Hypothesis 3: Expanding Sigmoid -- The initial state of the speech percept may be completely ambiguous (see the flat line in the bottom left panel of Figure 2), and the halves of the continuum separate themselves over time, retaining their categorical structure (lower middle and right panels of Figure 2).

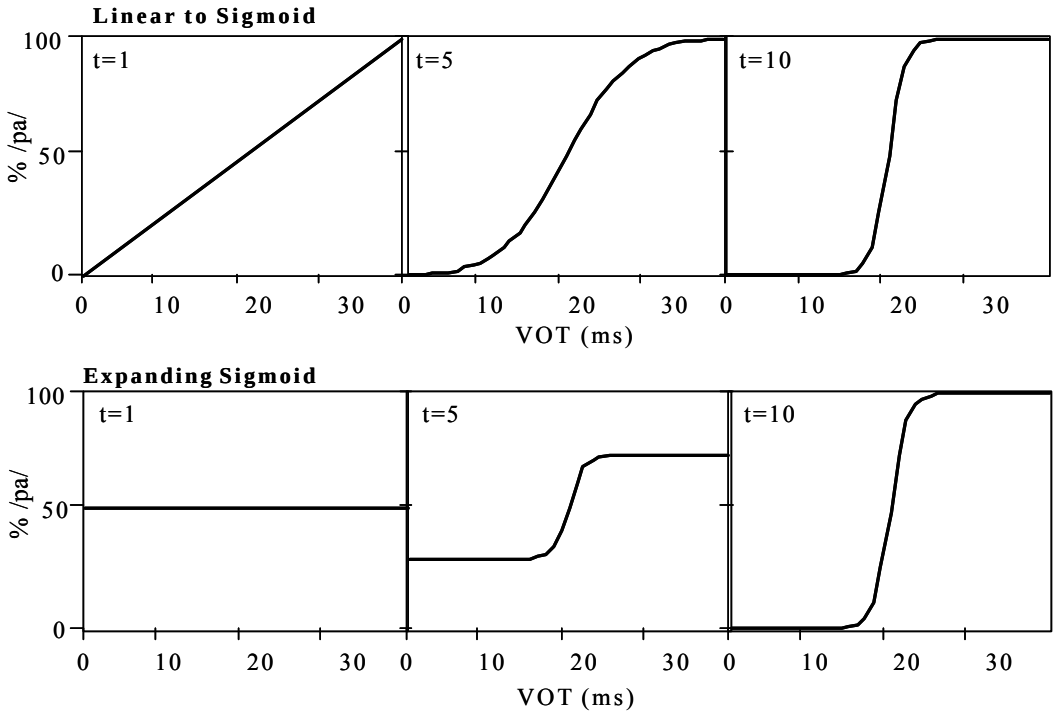

Figure 2: Two possibilities for the timecourse (left-to-right) of categorical voicing identification. Although the final curve is the same for the two models, the initial and intermediate identification curves are quite different.

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 English with normal hearing and normal or corrected to normal vision. In accordance with university human subjects procedures, subjects were informed of the risks of the experiment and consent was obtained prior to beginning the procedure.

Stimuli were synthesized on a Sun workstation with the Delta system from Eloquent Technologies. A 9 stimulus /ba/-/pa/ continuum was created by varying the temporal onset of voicing relative to the onset of the release burst (VOT). VOTs varied from -50 to 60 ms with the category boundary (determined by the subject's responses) lying roughly at 10ms. During the post experiment briefing, none of the subjects reported having any trouble identifying the stimuli as /pa/ or /ba/.

Our methods are based on the eye-tracking methodology of Tanenhaus, Spivey-Knowlton, Eberhard and Sedivy (1995). In this paradigm, subjects are asked to perform motor tasks in response to linguistic instructions. In tasks such as reaching or pointing, eye movements have been shown to indicate what the subject is *about to* reach for or point to. The probability of an eye fixation to a potential target over time has been interpreted as being indicative of activation levels as the system settles on a decision (Allopenna, Magnuson and Tanenhaus, 1998). We adapted this paradigm to use computer-generated images as the targets and mouse control as the motor task (as opposed to real objects and reaching).

Subjects were seated at a computer and the head-mounted eye tracker (to be described shortly) was calibrated. They were told that they were about to hear a series of synthetic speech sounds and that their task was to categorize them as accurately as possible by clicking with the mouse on one of two large squares labeled /ba/ and /pa/ (which were on the screen during the instructions). They were told to relax and take their time and that the labeling of the squares would not change throughout the experiment—/ba/ would always be on the left and /pa/ on the right (so that eye movements would not be induced by the subject searching for the button's locations).

Throughout the experiment, eye position was monitored with an ISCAN eye tracker. The eye tracker consists of two cameras that are mounted on an adjustable helmet. The "eye camera" records an infrared image of the eye. This image is analyzed by a computer to determine the location of the pupil and the corneal reflection. From this, the computer is able to find the position of the eye relative to the head. This information is combined with the view from the "scene camera" (which records the subject's field of view) as a set of cross hairs indicating the subject's point of fixation. Although the eye tracker is able to compute this 60 times per second, the data ware recorded on a HI-8 video recorder at 30 Hz.

Each trial starts with the presentation of a single gray circle in the middle of the screen. The subject fixates on this for two seconds to establish that their gaze is midway between the two buttons. When the circle turns red, the subject clicks on it, establishing that the mouse is at the midway point. At this point, the circle disappears, and two gray squares labeled /ba/ and /pa/ appear on the computer screen. One of the nine stimuli is played through the Macintosh computer speaker and the subject clicks on the button corresponding to what he or she heard. In addition to eye position, reaction time (measured as the time between the stimulus onset and when the mouse was clicked) and the square the subject chose was measured. Each of the stimuli was presented a total of 7 times, with the order of presentation randomized for each subject. The experiment was designed and run using the PsyScope experimental design software (Cohen, MacWhinney, Flatt, & Provost, 1993).

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

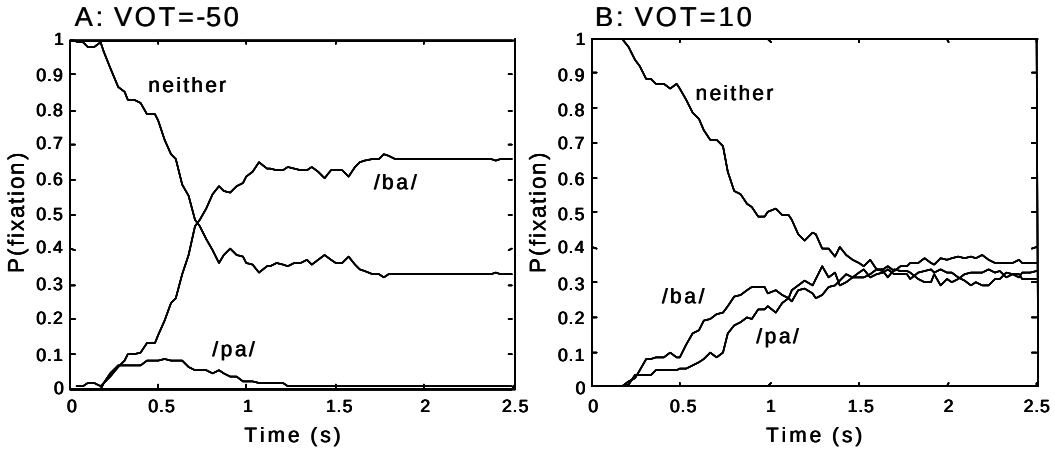

Figure 3: Fixation probability as a function of time for a good/ba/ (VOT=-50, figure 3a) and for an ambiguous stimulus (VOT=10, figure 3b)

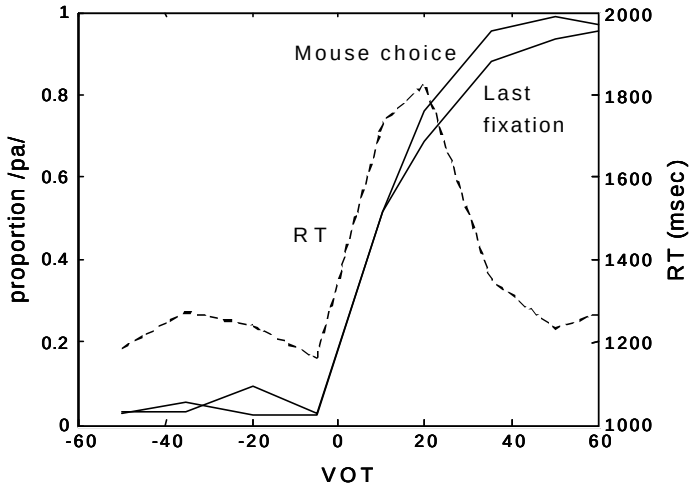

Figure 4: Proportion of /pa/ responses for last eye movements, mouse clicks and reaction time as a function of VOT.

Tash (1974) and the identification curves exhibit the same steepness. In addition, it is clear that the final eye movements are highly correlated with the mouse choice data. A repeated measures logistic regression (with VOT, trial and whether or not the data came from the late eye movements or the mouse choice) found highly significant effects of VOT and trial (yielding a total $\chi^2(22)=1539$, $p<.0001$). More importantly, no effect was found for whether the data were from a mouse choice or an eye movement ($p > .5$) or for the interactions of this variable with the others (all $p>.4$). The two datasets are not different from each other when intra-subject effects and the effect of VOT are partialed out. Late eye movements provide the same information as the mouse choice. Given that both of these events are occurring during approximately the same time window, this provides support for our view that the time of an eye movement tells us something about processing at that time.

Figure 5 shows identification curves created by averaging whether or not each subject was fixating to /pa/ during particular temporal windows (time-bins of 0–400ms, 433-633, 667-867,900-1100, and 1133-1333 msec). The data seem to support the expanding sigmoid hypothesis (we'll show this statistically in a moment). Category membership is maintained throughout the time-course, by which we mean that for each time-bin, within an eventual phonetic category, stimuli with differing VOTs have similar probabilities of being identified as /pa/.

Although we want to stress that under our view categorization is a graded process that takes place over time, it is possible to determine when (across subjects) categorization (or the beginnings of it) has reliably occurred. The earliest time at which a stimuli could be reliably ($p<.05$) identified as /ba/ (given that the subject heard a stimulus with VOT –50, -35 or –20) was 16 frames or 533 ms after stimulus onset (%/pa/ = 32, $t(15)=2.48$, $p=.025$). The earliest time at which a stimulus (with VOT 60, 50 or 35) was identified as /pa/ was 19 frames or 633 ms (%/pa/=68, $t(15)=2.5$, $p=.024$). It takes roughly 200ms to generate saccade and the release burst occurred at 80 ms, so it took the system roughly 250ms to identify a /ba/ after it received enough acoustic information. Similarly, it

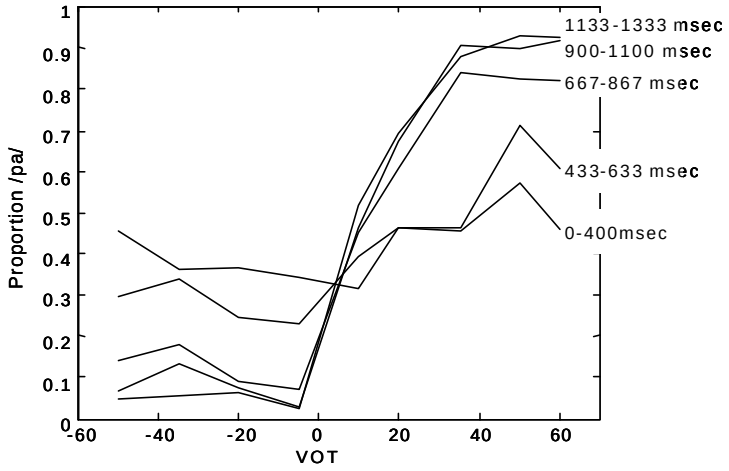

Figure 5: Proportion of /pa/ fixation responses as a function of VOT and time.

took 290 ms to identify a /pa/ (since it would not have enough acoustic information for an extra 60 milliseconds—after the onset of voicing). These values are consistent with a magnetic mismatch field study by Phillips, Marantz, Yellin, Pellathy, McGinnis, Wexler, Poeppel and Roberts (submitted) that found a latency of roughly 200 msec for categorically distinguishing voicing. The longer value for /pa/ is probably the result of our imperfect stimuli (they were underaspirated causing a shift in the boundary towards /ba/), which did not show the same sharp identification as /ba/.

Since it is clear there is an effect of time on the identification function, we must establish which of the remaining two hypotheses best fits the data by determining whether that change is due to a change in its slope or steepness (the derivative at the midpoint), or due to a change in category goodness or amplitude (the difference between the upper and lower asymptotes). To do this, a hierarchical nonlinear model based on the logistic function was used (see Ohlemiller, Jones, Heidbreder, Clark and Miller, 1999, for a simpler example).

Here, we will only describe the analysis conceptually before moving onto the results. The interested reader may wish to see Appendix A for a more complete treatment.

The identification function seen here is well approximated by a logistic function (Figure 6). This function is usually only described by two parameters the slope (how fast it changes from 0 to 1 along the y axis, or more specifically, the derivative at the midpoint) and the category boundary (the point along the x axis where the mid point lies). However, to fully describe all possible logistic functions, two additional parameters are needed—the amplitude (the distance between the upper and lower maximums of the function) and the bias or height (the point along the y axis where the midpoint lies).

To generate the statistics described here, we subdivided our large dataset into smaller datasets containing the fixations and VOTs for a single subject and time-slice (1 frame). We had 16 subjects x 130 time slices, or 2080 such datasets. For each dataset we found the four

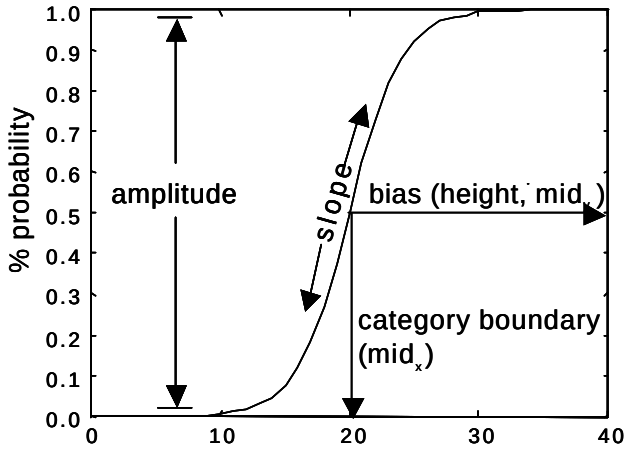

Figure 6: A typical logistic function and it's parameters. The amplitude of this function is 1. The slope is fairly steep. The category boundary is at 20 and the function is equibiased (it's height is .5).

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 or trading relations where secondary factors affect the response probabilities. By analyzing these effects in terms of their effect on category boundary or bias we might reach a more detailed understanding of them than by simply exploring the overall "difference" in the logistic functions that results from them.

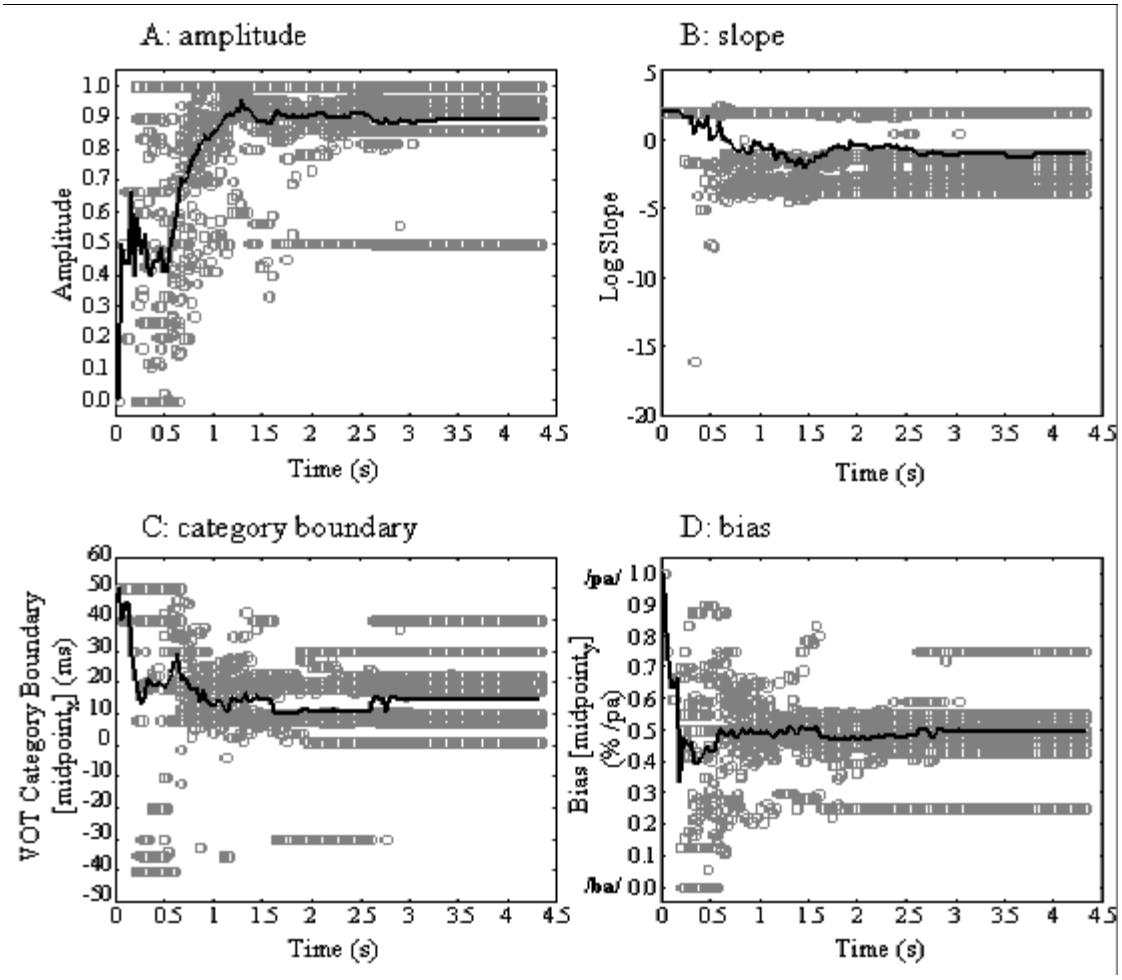

Figure 7: Parameters of the logistic identification function as a function of time. There is substantial change in amplitude over time, but very little in the other three parameters. This suggests the expanding sigmoid hypothesis. Gray circles represent individual data points. The dark line represents 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) over and above the previous model (consisting of subject, time and 1/time).

This large effect may be the result of individual differences in the way subjects process temporal cues to voicing. For example, aspiration, as a high frequency cue to voicing, can be detected relatively early (only a few samples are needed) while the presence of a voicing pulse might take a bit longer to detect. If different subjects weighted these cues differently, we might see evidence of these differences reflected in the time-course of processing. Of course, this tenuous hypothesis cannot be shown conclusively without first examining the temporal dynamics of aspiration and voicing-pulse cued voicing more directly.

The last regression analyzed the height of the logistic function as a function of time (the y coordinate of the midpoint of the function, see Figure 7d for the scatterplot and mean). It is important to note that this variable cannot be anything other than .5 when the amplitude is 1 (otherwise the curve would yield an invalid identification probability greater than 1 or less than 0). When mid_y is not equal to .5, though, it indicates a general bias to chose one category over another. Like mid_x, there were very small but significant effects of time (R^2_{time}= .003, β=.055, t(1945)=3.398, p=.001) and 1/time ($R^2_{1/time}$=.007, β=.105, t(1944)=5.426, p<.0001). However, again, the scatterplot suggests there may be a strong interaction of 1/time by subject (Figure 7d). Each subject's data has its own characteristic hyperbolic function, all of which converge at the same height around frame 40. The slope and the starting point of these functions tend to vary by subject, however. Indeed, the addition of <subject by 1/time> interaction terms accounted for an additional 18.4% of the variance (F(15,1929)=76.4, p<.0001).

Since we see such dramatic within-subject shifts in both category boundary (mid_x) and bias (mid_y) over time, one might suspect that they are related. Interestingly, there seems to be a small, but significant correlation (R=.241, p<.0001) so as the category boundary shifts towards /pa/ (resulting in more stimuli being identified as /ba/), the overall response bias seems to shift towards /ba/. This is consistent with range effects on phonetic categorization. Ohlemiller et al (1999), for example, found that if the VOT range of the continuum was increased on one end, a boundary shift occurred, suggesting that subjects were attempting to maintain equal numbers of responses in the two categories.

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-locked to the serial arrival of information (e.g. Allopenna et al, 1998; Tanenhaus et al, 1995). Moreover, the proposed linking hypothesis between serial fixations and serial information is quite simple and explicit. However, the informativeness of eye tracking as a tool to look at the temporal process of categorization and decision-making in a setting where information does not arrive serially has not been fully explored (Experiment 1 in this paper represents one of the first times that this application has been attempted). We hypothesize that fixations indicate where the subject is *about to* click (or reach) and can tell us something about the probabilistic tendencies of subjects over time. This linking hypothesis is fundamental to our arguments, but has not been explicitly tested.

To deal with this potential objection to our line of research, in the present experiment, we replicate Experiment 1 using a different type of response paradigm. Response deadline methodologies (and the associated speed accuracy tradeoff task) have been quite successful in exploring the temporal dynamics of higher-level language processes (see McElree, 1993 for an example) and visual categorization (Humphreys, 1981; Lamberts & Brockdorff, 1997). In this methodology subjects are given response deadlines that force them to respond before they are fully confident in their decision. We have adapted this methodology to our purposes by giving subjects a number of different deadlines that will map onto different portions of the time-course of perception. Subjects' responses to early deadlines should map onto early fixations. Their responses to later deadlines should map onto later fixations. In this way, we expect a slightly cruder (since the temporal resolution of the procedure is much less than that of the eye tracking methodologies) replication of Experiment 1.

Methods

The same 9-step VOT continuum from Experiment 1 was used in Experiment 2. Eleven subjects were recruited in the same manner. The experiment was designed and run with the PsyScope experimental design software (Cohen et al, 1993).

The experiment consisted of five randomly ordered blocks of trials. Each block had a different response deadline (400 msec, 500 msec, 600 msec, 800 msec and 1000 msec). Each block

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.

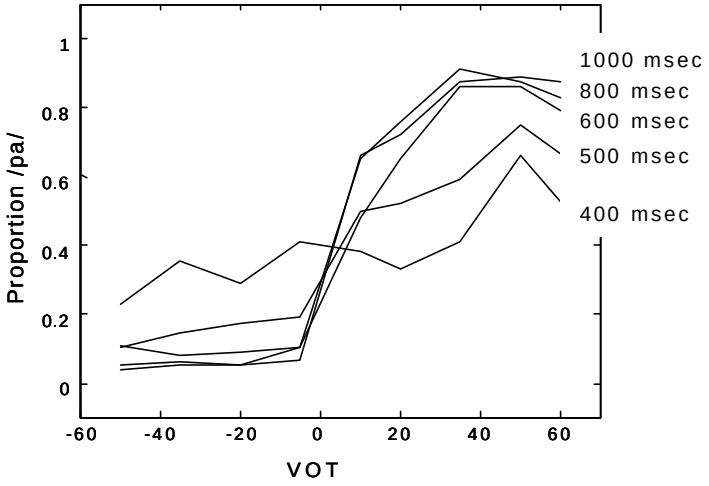

Figure 8: Response deadline data for Experiment 2. Proportion of /pa/ responses as a function of VOT and response deadline. Results mirror those of Experiment 1.

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 ($R^2_{interation}=.269$, $F(10,32)=2.181$, $p=.046$), suggesting that this individual variation is reliable.

The analysis of bias (mid_y) showed no significant effects ($F<1$) of time, or 1/time. Although our <subject X 1/time> interaction terms were not significant, ($F<1$) they did account for almost as much variance as they did in Experiment 1 ($R^2_{interaction}=.142$, compared to $R^2_{tiime}=.040$), suggesting that with more subjects, or with more data per subject, we may find a significant effect. This is supported by a correlation between category boundary and bias ($R=.299$, $p=.027$) much like we found in Experiment 1.

In summary, it appears that the important results of Experiment 1 are replicated. Amplitude increases over time, and there is no effect of slope. Interestingly, the findings of individual differences in the starting points of the category boundaries also seem to be supported. These results provide strong independent support for the use of eye tracking in this domain.

Neural network simulations of the temporal dynamics of categorical perception

Now that we understand more fully what is occurring over time during categorical perception, it is important to begin to address the psychological consequences of these findings. To do this we instantiated each of our three hypotheses in a connectionist network. Since we know the architectures of each network and the kind of representations and processing they use, if we can find

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.

The network then passes activation back and forth from the input to the output layers until it has reached a decision. The algorithm is as follows:

1) Output = Output + Input
2) Output = output/sum(output)
3) Input = Input + Input*output
4) Input = Input/sum(input)
5) Repeat.

The first two steps simply compute the cumulative output and normalize it so that the two nodes sum to one. The third step passes that output back to the input and adds some nonlinearity. The last step normalizes the inputs. In addition to these steps, a small amount of noise was added to the input layer before processing began. This tipped the network towards one decision or the other in cases where the input nodes were exactly equal (something which is extremely unlikely in a biological system).

Most instantiations of normalized recurrence use an absolute threshold of activation to indicate that the network has made a decision (although often this threshold does change with time). That is, the network is said to have picked /pa/ or /ba/ when the activation of one of those nodes crosses some threshold. We chose to model this behavior of the network differently, by stopping the network when the derivative of the output nodes was less than 0.0001—essentially when the network settles on a decision.[4] If we measure the number of iterations until the network reaches such a decision, we have an analogue of reaction time.

[4] We have also run this model with dynamic and fixed thresholds and found the same general results. Although the results are the same, these different stopping criterion have different theoretical motivations. The fixed threshold model, for example, is roughly analogous to the variable rate model of reaction time. The algorithm used here (based on the derivative) is not directly analogous to either the variable rate or variable threshold models of reaction time, and might represent the beginnings of a third model.

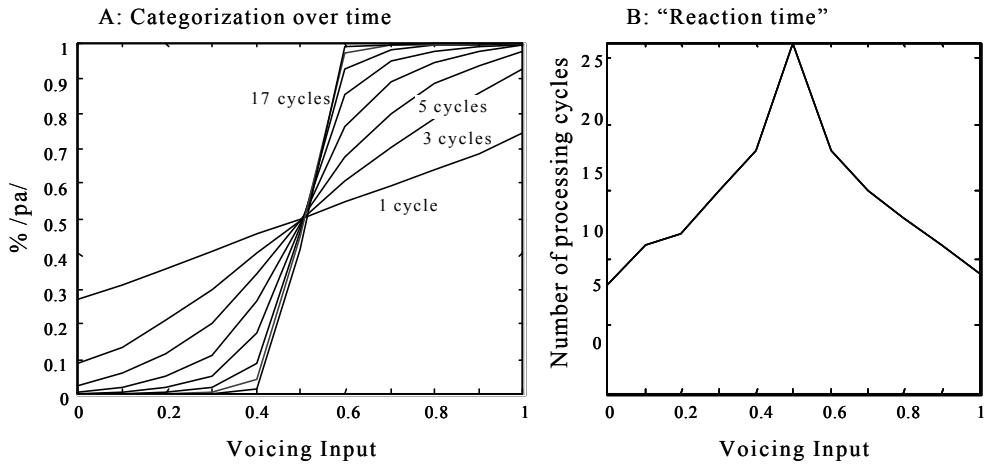

Figure 9: Output of the normalized recurrence network. A) Activ ation in the /pa/ node as a function of the input in the voicing layer at several time slices. The model clearly follows the predictions of the linear to sigmoid hypothesis. B) Number of processing cycles to settle as a function of voicing—a close correlate to reaction time.

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 competition algorithm to resolve an ambiguous input (see Figure 9b). However, since the network is not sensitive to any sort of realistic internal representations (which we believe are derived from the statistical distributions of its input), it does not treat inputs from the same category the same—perceptions start out as a linear function of VOT and are categorized over time (as per the linear to sigmoid hypothesis). This network showed no statistical sensitivity, but did show competitive processing.

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

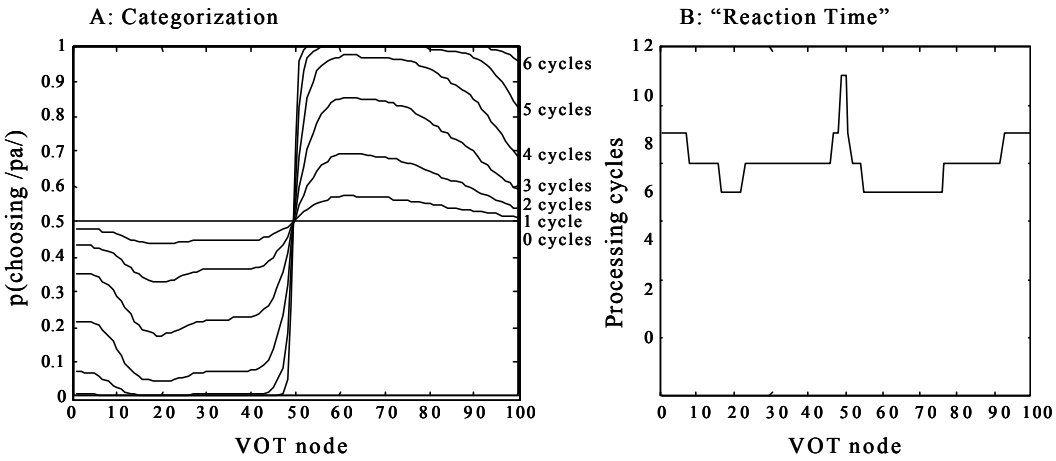

Figure 10: Output of the Hebbian normalized recurrence network. A) Probability of the network choosing /pa/ as a function of input VOT and time. B) Number of processing cycles to settle, a good analogue to reaction time.

well (Figure 10). It is clear that it learns to categorize the input correctly (the identification curve after 6 processing cycles in Figure 10A) and also shows the appropriate spike in processing cycles at the ambiguous region (Figure 10B). In addition, it also seems to have the properties of the expanding sigmoid in that category members are treated similarly throughout the time-course of processing. The only deviation to this pattern (the pattern we saw empirically) is the tendency of the model to show graded category membership at the far ends of the VOT continuum (at the extreme of devoicing and prevoicing) early in the time course. This "prototype effect" has seen some support in the work of Miller, (2000) who found graded representations at the extreme devoiced end of the continuum using a rating task. It represents a testable hypothesis that we have begun to look at by exploring the extreme devoiced and prevoiced ends of the VOT continuum with our eye-tracking measure.

Other simulations with this architecture have shown that it is capable of learning as many categories as appear in the input (as modes in a multimodal distribution), and that it is robust across multiple instantiations and against noise in the input. However, despite the fact that it undergoes learning, its performance early in the training does not map well onto the abilities of infants, so it is clearly not a complete model (it initially discriminates all VOTs from one another rather than grouping some together). By combining this architecture with the structures in McMurray (in preparation) that do predict infant abilities (but do not deal with the temporal dynamics of categorization), it may be quite simple to simulate in one model both the large and small time-scales of learning and categorization.

The normalized recurrence network presented in Simulation 2 can best be thought of as a system of two equally sized and spaced attractor basins, while the Hebbian normalized recurrence network learns those attractor basins and can learn as many as it needs in whatever "shapes" it needs. It shows sensitivity to both the statistical distribution of VOT in the learning environment and competitive processing. This network also shows a high degree of neurological plausibility.

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-tracking and response deadline methodologies make it quite possible to assess how perception changes over time. The addition of this temporal dimension to our characterization of the perceptual system allows us to further define our notion of categorical perception and gives us the ability to explore further perceptual domains that may result in similar representations at the time of a perceptual decision. The comparison of infant and adult perceptual abilities may be the most important of these domains.

Work in infant speech perception has found that infants tend to discriminate only those sounds that lie on either side of an eventual phonetic category (Eimas et al, 1971; see Jusczyk, 1997, for a review). This has led many researchers to propose that infants exhibit a categorical perception mechanism that may be quite similar to that of adults, at least by the time a perceptual decision has been made. However, in light of our temporal view of speech perception, the obvious question becomes, "What about the intermediate stages?" More importantly the methodologies used to make these claims do not provide adequate data to support them because none of these techniques has been able to provide the sort of labeling or categorization data that can be obtained from adults.

Two major techniques have been used to assess infant speech perception abilities. In habituation tasks the infant is repeatedly presented with a stimulus until he or she habituates (gets bored). Habituation is usually measured by looking time to a sound source—as looking time decreases, the baby is habituating. Once the baby is habituated, the stimulus is changed to something new. If the baby can discriminate the new stimulus from the original, he or she should become more interested (dishabituate). If the baby cannot discriminate the two stimuli then he or she will continue to habituate.

High amplitude sucking methodologies are based on habituation techniques but add an element of learning. Babies are given a pacifier to suck on. The pacifier is attached to a computer such that every time the baby sucks with sufficient amplitude the computer plays a speech sound. Infants learn this contingency quite rapidly and begin to suck faster as they hear more sounds.

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 some trials, the baby will begin to make anticipatory eye movements (eye movements to the other light or image location before it is actually turned on).

Unlike habituation studies, this paradigm relies on a naturalistic response—making an eye-movement to a moving object is something that infants do quite frequently. Moreover, each response is given as the result of a single stimulus (after training). Finally, the visual expectation paradigm relies on a very metabolically cheap response—eye movements are faster and take less energy than head movements or sucking responses. Unfortunately, the eye-movement response in this paradigm is not arbitrary (we cannot assign a response to a particular stimulus as we do in a two-alternative-forced-choice task), and the stimuli are entirely visual.

The conditioned head-turn procedure (Kuhl, 1985) has some of the properties we are looking for in a methodology. In this procedure, the infant is conditioned to turn his or her head to the left (or right) in response to a change in a stimulus by presenting a visual reward on that side only when the change occurs. As an example, an infant may hear /a/ /a/ /a/ /a/ /i/ and will only be

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 online perception. Moreover, the closest adult methodology that is analogous to this would be phoneme monitoring (and we are more interested in something resembling a two-alternative-forced-choice task). This methodology does, however, employ something of the arbitrary response that the visual expectation paradigm lacked. Here, the infant is trained to make an unrelated response to an arbitrary stimulus (much like an adult pushing a button).

Although neither of these methodologies provides a perfect measure of categorization (and both ignore the temporal dimension that we are interested in), a combination of the two may be ideal. Eye movements are clearly the right sort of response measure—they are metabolically cheap, and by looking at the timing of the saccade we can get some measure of the temporal dynamics of the system. As we've discussed, this use of eye movements to explore the temporal dynamics of the system is becoming prevalent in the adult literature (Experiment 1, this paper; Tanenhaus, et al, 1995; Allopenna, et al, 1998) and it is starting to become used in work with children as well. Swingley and Aslin's (2000) work on the temporal dynamics of early word recognition (in 18 month olds) is the most recent example.

By using eye movements (instead of head-movements) in a conditioning paradigm we may be able to overcome some of the limitations of the head-turn preference procedure. Since eye movements are much cheaper to produce than head movements, we may be able to condition infants to look to two different locations (as opposed to only one) and we may find it takes much less training. This relatively simple idea of conditioned eye movements represents the heart of the anticipatory eye-movement procedure presented here.

Overview of the methodology

The procedure (Figure 11) starts with the infant seated on a parent's lap viewing a series of movies on a large computer screen. Since we'll be measuring eye movements relative to the center of the screen, each trial begins with a visual stimulus to orient the infant's gaze to the center of the screen. In the experiments we report here, a vertically moving 'smiley face' of a randomly selected color was used. Immediately following the removal of the centering stimulus, one of two randomly selected sounds is presented simultaneously with a short animation (which varied from trial to trial) on one side of the screen. The sound is consistently paired with the side on which the animation is presented.

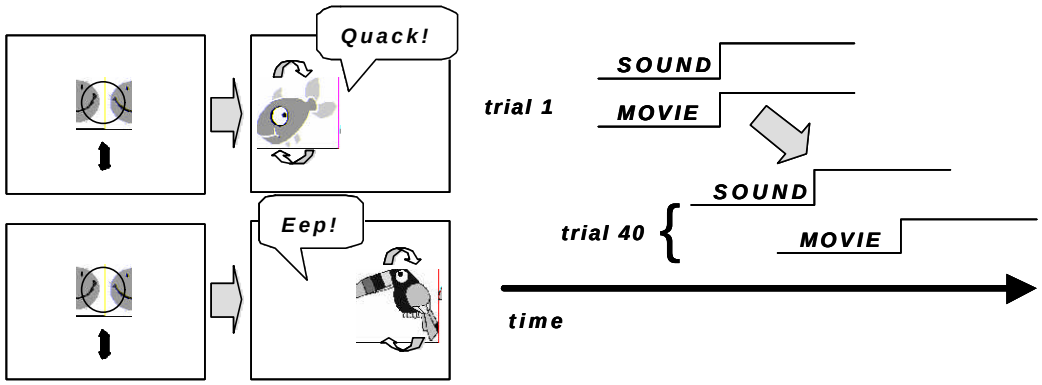

Figure 11: The Anticipatory Eye Movement Paradigm. Each trial begins with a vertically moving 'smiley face' to orient the infant's gaze to the center of the screen. The face is removed and one of two randomly selected sounds is presented simultaneously with a short animation on one side of the screen. 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.

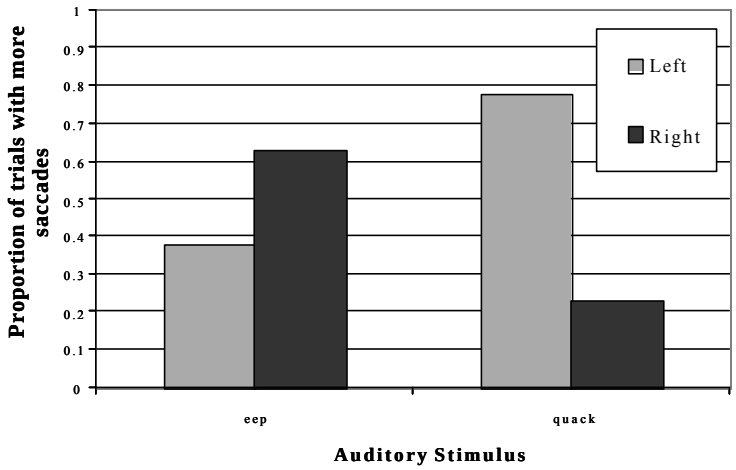

Figure 12: Proportion of trials with more saccades to right or left as a function of the auditory stimulus

Experiment 3: A preliminary test of the anticipatory eye-movement paradigm.

Introduction and Methods

To evaluate the effectiveness of the anticipatory eye-movement methodology, six 4-month-old infants were trained and tested on easily discriminated sounds: the Macintosh "Quack" and "Eep" sounds. Training occurred exactly as described above. The maximum (end-point) delay between the auditory stimulus and the visual stimulus was 2400 msec.

The testing procedure was identical to the training procedure except that the delay between auditory and visual stimuli was held constant at 2400ms (it did not increase like it did in training). Close-up videos of the infant's face were coded by trained observers for gaze direction.

Results

Performance was quite good. On average, infants viewed 11.2 test trials before becoming fussy and dropping out. Given that 70% of these trials contained an eye-movement to either stimulus location this represents just enough within subject data to begin looking at speech continua (here, for example, we would be able to get one response on each step of a 8 step continuum). With improvements in the testing procedure (such as including the psychedelic images from training every 5 blocks) we might be able to get even more data.

For the 70% of trials on which infants made one or more eye movements to either of the two possible locations, the direction of the latest eye movement was correct 85% of the time ($t(5)=6.032$ $p=.002$, see figure 12 for a breakdown by auditory stimulus). Additionally, none of the infants scored lower than chance (50%). This indicates that the infants were learning to expect the visual display (reward) on the side cued by the auditory stimulus.

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* a correct one. This suggested that the infants were moving their eyes to the correct location and subsequently getting bored or impatient that the visual stimulus had not appeared yet. To rectify this potential confound, we shortened the maximum delay between auditory and visual stimuli to 1800 msec.

After training, infants were tested on all five steps of the continuum. For the work presented here, we initially reinforced all the testing sounds. The endpoint sounds were reinforced on the same side as during training. Sounds immediately adjacent to them on the continuum (10ms and 30ms) were reinforced as though they were endpoints. The ambiguous token (20ms) was reinforced randomly to either side. Since we did not expect infants to hear more than two or three exemplars of each sound we did not expect this additional training to make much of a difference.[6] However in current work on this paradigm in visual categorization we do not reinforce testing trials at all to avoid a potential confound.

Results

Trained coders scored videotapes containing a close-up of the baby's head for gaze direction (although we have subsequently moved to a computerized eye/head-tracking system—see appendix C). Assuming some sort of temporal decision-making process, we made the assumption that the latest saccade in a trial would be the most accurate. Thus, for each trial, the direction (right or left) in which the baby looked last was scored as the baby's "decision".

---

[5] Although it has been shown that infants can discriminate endpoint sounds from a VOT continuum (Eimas et al, 1971) these demonstrations have depended on the presentation of multiple stimuli in high amplitude sucking and habituation tasks. Thus these demonstrations may depend on representations not available to the infant who has only heard a single stimuli.

[6] In subsequent (ongoing) experiments we have shifted to a new testing paradigm in which odd-numbered testing trials are always endpoint trials and are reinforced. Even-numbered trials use generalization stimuli and are not reinforced, there is simply a 1800ms delay in which we record eye movements before moving on to the next trial.

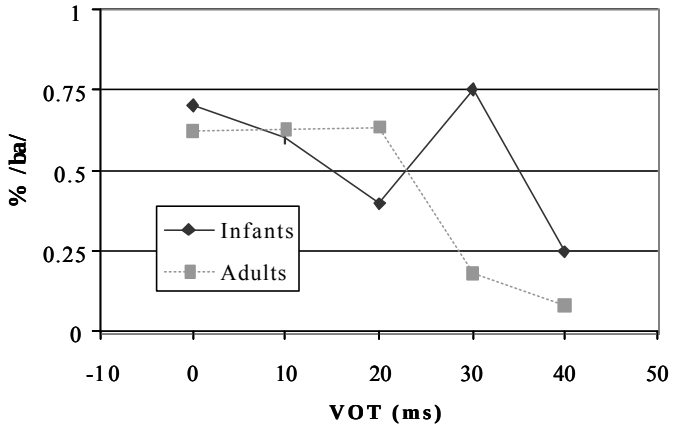

Figure 13: Percentage of stimuli identified as /ba/ for both infants and adults.

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 perception experiments. The solid black line represents data from the 7 babies who met our criterion. The dotted gray line is two-alternative-forced choice data from 9 adult subjects listening to the same stimuli in similar listening conditions.

A comparison with the adult data revealed that the infants performed as well as they could have on the endpoints. Although the babies only identified the 0ms stimuli as /ba/ 70% of the time, and the 40ms stimuli as /pa/ 25% of the time, this was similar to the adult values of 60% and 7%, respectively. To test this intuition, for each infant and adult we computed the percentage of "correct" trials (the subject chose /ba/ after hearing a 0 msec VOT or /pa/ after hearing a 40 msec VOT) for each of the two endpoint stimuli. In a nested-design regression analysis, there was no effect of whether or not a data point came from the adult or infant dataset ($F<1$), or the interaction of that with VOT ($F<1$). This suggests that the inaccuracy of the infants is not due to problems with the methodology, but rather with the synthesized stimuli, which were difficult to distinguish. Future studies will need to make use of better, more natural sounding stimuli.

One's immediate impression of the infants' ID curve as a whole is that it doesn't form the smooth logistic function that most categorization curves show. If we want to assume a logistic function, one of two hypotheses must be correct.

Hypothesis 1: The percentage of 20 msec stimuli chosen as /ba/ is incorrect and should be at approximately 60% (the actual percentage was 40%). This would mean the infants heard

four /ba/'s and one /pa/ (compared to the adults 3 /ba/'s and 2 /pa/'s)—they must have a slightly different category boundary than adults.

Hypothesis 2: The percentage of 40 msec stimuli chosen as /ba/ is incorrect and should be at around 30% (the actual percentage was 75%). This would yield a relatively smooth identification curve, though one with a slope that is too shallow to call the perception "categorical".

In either case there are deviations from the adult data. Although there is no conclusive way to determine which of these two interpretations is correct (more subjects trained on these and other endpoints are needed), the best fitting logistic function with a midpoint of 35 (the first hypothesis) yielded a log-likelihood of –323.3 while the best fitting logistics with a midpoint of 20 (the second hypothesis) yielded a log-likelihood of –333.7. Although not a huge difference on a log scale, this actually represents a difference of 4 orders of magnitude. These crude simulations provide weak support for the first explanation with its shifted (though sharper) category boundary.

The time-course of infant speech perception

We now return to the central topic of this paper: temporal dynamics. It is quite clear that our categorization function is too noisy (due to poor stimuli and not enough infants) to make the sorts of assertions we were able to make about adult categorization in Experiments 1 and 2. There is, however, a tantalizing hint in this experiment that infant-eye-tracking is sensitive enough to assess temporal dynamics (given some more methodological refinements) if we compare the time-course of categorization of the two end-point stimuli.

Figure 14 shows the probability of a fixation towards /ba/, /pa/ and neither as a function of time for the seven subjects who reached the 60% criterion on the endpoints. There is an interesting discrepancy in the two graphs in that after hearing /ba/ (Figure 14A), subjects initially make a large number of false looks towards /pa/ before arriving at the correct response. After hearing /pa/ (Figure 14B), on the other hand, subjects make very few false looks. The obvious explanation is that /pa/ is simply easier to identify than /ba/, however, that would predict a more even temporal distribution of looks to /pa/ than we see.

A closer examination of the stimuli suggests a possible explanation. When the stimuli were originally synthesized, the duration and amplitude of aspiration was held constant across all stimuli (all the stimuli have a medial amount of aspiration), so that VOT would be the only cue available to the listener. However, aspiration is a good cue to voicing in many languages. Moreover, aspiration is carried in the high frequency components of the speech stream and is present from the onset of the sound. The presence of glottal pulses, on the other hand, can only be detected in the lower frequency components of the sound and is not initially fully present. Thus, from a processing standpoint, aspiration can be extracted from the speech stream much faster than voicing because high frequency signals require fewer samples to detect, and (as already mentioned) it may be physically present in its complete form earlier.

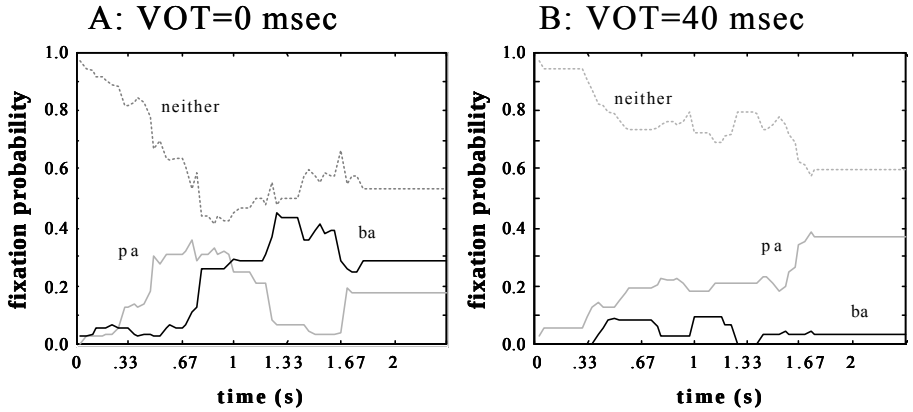

Figure 14: Fixation probability as a function of time after subjects hear an endpoint /ba/ (0ms VOT, seen in A) or an endpoint /pa/ (40ms VOT seen in B).

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, this is a tenuous explanation based on few subjects and on an experiment that was not designed to test this hypothesis. However, future studies examining the differential weighting of acoustic cues may support this claim.

Discussion of the anticipatory eye-movement paradigm

Although the methods presented here are based on training (specifically, operant conditioning) procedures, there is good reason to believe that we are not simply teaching the infants the categories we are trying to study. We are only training the infants on one pair of sounds, and the additional sounds used during testing are novel. Thus, the methodology is looking directly at generalization or similarity. Although a single experiment may not be enough to establish a category boundary, performing multiple experiments with different ranges of stimuli (e.g. one experiment with VOTs from 0 to 40 and one with VOTs from 10 to 50) should tell us whether stimuli are being grouped relative to the endpoints or relative to an abstract phonetic category.

Kuhl, (1985) has extended the conditioned head-turning procedure to look at what she claims is categorization by using more than one stimulus repeating in the background. As an example, her methodology initially trains the infant on a single background stimulus pair (/a/ vs. /i/). Then it slowly adds /a/'s and /i/'s from other speakers at different rates or fundamental frequencies. If the infant is able to learn this discrimination, then, she claims, it has something resembling a phonetic category. Rather than looking at infants' natural categorization of speech sounds, this methodology, is really just teaching the infant the range of variation to ignore in its categorization. Although it is certainly interesting what sorts of factors the infant is able to learn to ignore or use, this is not the question we are interested in.

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 information of VOT for a given speech stimulus may indeed be discarded rather quickly during the perception process. However, complete categorization is not instantaneous. Early on in speech perception, each half of the VOT continuum is treated as belonging *somewhat more* to one category than the other. As time proceeds, this gradual categorization becomes more confident and discrete, until finally displaying the signature step-function of categorical speech perception.

This pattern of data could not be simulated with a neural network that simply incorporated a statistical learning algorithm but no temporal dynamics. Similarly, the results could not be simulated by a hand-coded attractor network that displayed temporal dynamics but no statistical

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., Wexler, K., Poeppel, D., and Roberts, T. (submitted). Auditory cortex access phonological categories: An MEG mismatch study. Obtained from http://www.ling.udel.edu/colin/research/ftp.html.

Pisoni, D. (1977). Auditory identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America, 61,* 1352-1361.

Pisoni, D., and Lazarus, J. (1973). Categorical and noncategorical modes of speech perception along the voicing continuum. *The Journal of the Acoustical Society of America, 55(2),* 328-333.

Pisoni, D. and Tash, J. (1974). Reaction times to comparisons with and across phonetic categories. *Perception and Psychophysics 15(2),* 285-290.

Rumelhart, D., and Zipser, D. (1986). Feature discovery by competitive learning. In Rumelhart, D., McClelland, J. (Eds.) *Parallel Distributed Processing: Exploration in the Microstructure of Cognition, Vol. 1.* 151-193. Cambridge, MA: The MIT Press.

Samuel, A. (1977). The effect of discrimination training on speech perception: noncategorical perception. *Perception & Psychophysics. 22(4),* 321-330.

Spivey, M. and Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24,* 1521-1543.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268(5217),* 1632-1634.

Tanenhaus, M., Trueswell, J. (1995). Sentence comprehension. In Miller, Joanne L., Eimas, Peter D. (eds) *Speech, Language, And Communication. Handbook Of Perception And Cognition (2nd ed.)*, 217-262. San Diego, CA: Academic Press.

Utman, J., Blumstein, S., and Burton, M. (2000). Effects of subphonetic and syllable structure variation on word recognition. *Perception & Pychophysics, 62(6),* 1297-1311.

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

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 two new parameters makes the curvefitting much more difficult, we were able to overcome that by using a constrained gradient descent algorithm. This represents a real improvement over existing statistical techniques in the amount of detail we can extract from our data.

This logistic function can be fully described by four psychologically interesting properties: the slope, the amplitude (amp), the location of the category boundary on the VOT axis (mid_x), and the height of the function (mid_y). Amplitude will give some measure of noise or unbiased competition in the system. Slope provides a measure of the relative discreteness of the categories (the sharpness of the boundary). Mid_x is a direct measure of the category boundary. This measure differs from other category boundary measurements in that it is based on all of the data, not simply the few datapoints near the boundary. The height of the function gives a measure of overall bias in the system. For example, a height of .5 represents an unbiased categorization. A height of .6 on the other hand (paired with an amplitude less than .8—otherwise the maximum would be greater than 1) would show an overall bias (regardless of VOT) towards /pa/ in our data.

Simple algebraic manipulations can change (1) into a form that uses these parameters (2).

$$\text{logistic}(VOT) = \frac{amp}{(1 + e^{4slope/amp\ (midx - VOT)})} + mid_y - \frac{amp}{2} \qquad (2)$$

We extracted a subset of the total dataset for each subject at each 33 msec time slice (the minimum temporal resolution of the eye-tracker). We then fit a logistic function of this form to the data for that particular subject and time slice. By characterizing the degree of change in these four properties over time (across subjects), we can determine which of the hypotheses provides the best

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 Hebbian normalized recurrence network starts its "life" with a random weight matrix. Because of this, it cannot initially count on any help from its knowledge of the inputs in categorizing stimuli—rather, its competition process must do most of the work. Hebbian learning works best when learning mappings between an input array and a singly active output node (hence Rumelhart and Zipser's, 1986, competitive learning scheme). With a completely linear activation function in the output arrays, and no information from the weight matrix, it is very unlikely that even an approximation of this ideal output pattern will result from recurrent processing. By inducing a non-linearity into the computations of the output, the model can resolve to a singly active output node quite easily. Competitive learning, for example, represents one such nonlinear (and in this case discrete) activation function.

The quadratic function is not the typical form of non-linear activation in neural networks—logistic or softmax functions are usually preferred. However, the quadratic function can be shown to arise quite naturally out of a particular model of lateral inhibition. To derive the quadratic

[7] In general the range of slopes that best fit the data were very small (mean = .47) and the effect of slope on the resulting curve (and the log likelihood) was very nonlinear. Thus, we actually minimized over the log of slope so that we could maximize the effectiveness of the [linear] search algorithm.

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

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

Since the sum of all activation in O is equal to A (by definition),

$$A = \Sigma O_y \quad (12)$$
$$O'_x = O_x - [\ (O_x/A)\ A\ - O_x^2/A\ ] \quad (13)$$
$$O'_x = O_x - [\ O_x - O_x^2/A\ ] \quad (14)$$
$$O'_x = O_x^2 / A \quad (15)$$

Thus, if we start with a relatively neurologically plausible inhibition model of inhibition (in which each node inhibits the others as a function of it's proportion of total activation) we can derive a

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 http://www.bcs.rochester.edu/infanteyetrack, the web page of a newly formed consortium devoted to the topic.

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.

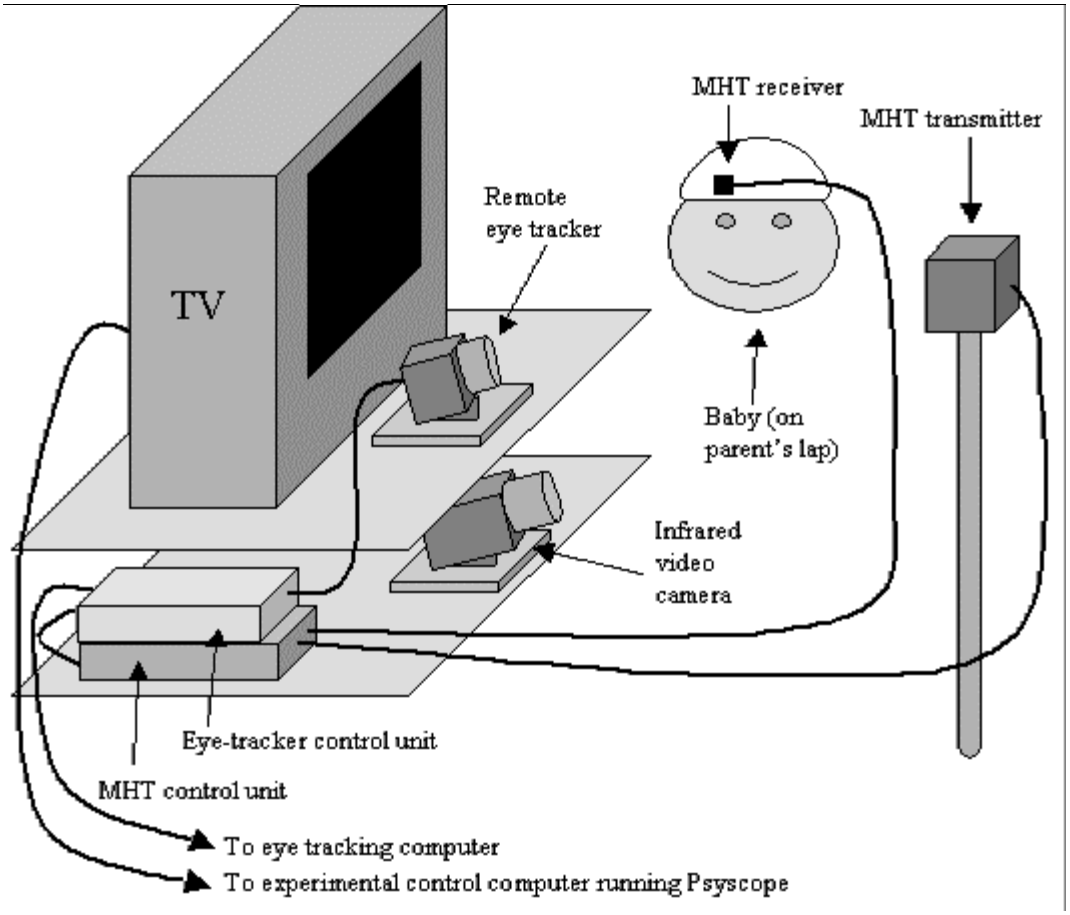

Figure 15: The infant eye-tracking setup. The baby is seated comfortably on the parent'; lap (parent not shown) watching the television. He wears a hat on which we have attached the Magnetic Head Tracker (MHT) receiver. The receiver detects its position and orientation relative to the MHT transmitter and sends this to the MHT control unit. This computes six dimensional coordinates and sends them to the eye tracker control unit which directs the movements of the remote eye tracker. An infrared video camera is positioned below the eye tracker giving a view of the infant's entire face. This allows us to hand code the data during periods when the eye track is lost and also allows the experimenter to see the subject while calibrating the equipment and running the experiment.

During the experiment, the experimenter monitors the corneal reflection and pupil thresholds (to make sure the computer is able to locate them at all times) and verifies that the auto-track is still following the eye (if the baby bumps the hat, for example, the head tracker will need to be recalibrated). The experimenter can also adjust the calibration coordinates to clean up any bias in the calibration (with a squirmy infant and only two points, the calibration can be quite noisy).

Remote eye tracking systems can record eye fixations in screen coordinates (allowing for easy data analysis). To mark the beginnings and ends of trials in the data stream, we use two custom

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