Predicting the consequences of vocalizations in early infancy

Francisco Lacerda and Lisa Gustavsson

Department of Linguistics, Phonetics, Stockholm University, SE-106 91 Stockholm, Sweden

frasse@ling.su.se, lisag@ling.su.se

Abstract

This paper describes a method to study the infant's ability to predict the consequences of its vocalizations and presents the first results of the on-going investigation. The research method uses a voice-controlled device, with which the infant may control the position of a figure on a screen, in combination with an eye-tracking system (Tobii) that simultaneously registers the infant's gaze fixations on the screen where the figure appears. The preliminary results indicate that 12.5 month-old infants seem to be able to predict the consequences of their vocalizations as indicated by the decrease in the mismatch between the infant's gaze position and the location at which the figure is displayed as a function of the infant's F_0. [Work supported by grants from the Bank of Sweden Tercentenary Foundation (MILLE, K2003-0867) and EU NEST program (CONTACT, 5010).]

Index Terms: infancy, language acquisition, predictive behaviour

10.21437/Interspeech.2007-633

1. Introduction

Young infants produce spontaneous vocal and motor actions. Such spontaneous actions appear often to be performed without obvious external goals, in the sense that it seems like the infant is not necessarily attempting to communicate or reaching objects in its immediate neighborhood. From this perspective, these infant's spontaneous vocalizations and gestures are thus better described as exploratory actions that should help the infant establishing a link between its actions and their possible acoustic and motor outputs [1;2]. Thus, in this sense, such spontaneous exploratory behavior although not directed towards communicative or grasping goals is likely to provide the infant with implicit knowledge on the consequences of its vocal and motor acts [3;4]. Indeed, such "purposeless" vocal and motor actions early in life provide the infant with fundamental implicit knowledge on how different gestures map onto the domain of acoustic and motor outputs associated with them [5] and can be seen as the background against which the infant may take its first steps towards goal-oriented actions; the infant will eventually use them as a basis for further communicative acts or gestures clearly intended to act on objects in the infant's immediate environment [6-8]. At some point in its developmental path, the infant will realize some of the consequences of its vocal and motor actions and will start using them to achieve particular interaction goals. It is expected that this shift from "purposeless" exploratory actions towards intentional actions may be paralleled by a shift in the infant's expectations on the outcome of its acts. Indeed, while initially the infant's exploratory actions may be accompanied by a reactive observation behavior that simply observes the outcome of different actions, the infant's shift towards goal-oriented behavior should be linked to the capacity to predict some of the outcomes of its behavior, leading to specific expectations on the outcome of the planned behavior [5].

The aim of the present study is to investigate the circumstances underlying this possible developmental process in infants and young children. The work is carried out within the framework of two research projects – MILLE and CONTACT – where Stockholm University's specific contribution is focused on language and speech development aspects.

The basic methodology used of this study consists in creating experimental situations in which the subjects are offered the opportunity of discovering hidden contingencies between their own vocal (or motor) actions and visual events that have been linked to certain variables associated with those actions. Since the present study is concerned with the infant's ability to learn the consequences of its vocalizations, the subject's fundamental frequency was chosen to be the acoustic variable that would be linked to the visual events. Specifically, the x-coordinate on which an animated actor would appear on the computer screen was directly related to the utterance's pitch. A logarithmic function mapping the relation between the vocalization's pitch and an *a priori* defined reference pitch was used. This function was defined so that -1.6 octaves relative to the reference pitch would map onto a location on the left end side of the screen and +1.6 octaves on the screen's right end side. To reduce the complexity of the situation, only the fundamental frequency was used to control the placement of the visual object on the screen. Thus, the visual object simply moved along a horizontal line at about 50% of the screen's height. Other spectral parameters (like overall intensity and ratio between the vocalization's high-to-low frequency bands' energy content) had been previously considered but were temporarily suspended in order to simplify the current test situation and thereby increase chances that infants might discover and explore the F_0 to x-coordinate's contingency. If subjects succeed in capturing the "hidden link" between their fundamental frequencies and the screen coordinates on which the visual object will be displayed, their gaze fixation points are expected to anticipate (or at least closely follow) the coordinate on which the visual object is going to be displayed. Thus, by studying the relative timing between the appearance of the visual object at a given location and the gaze orientation towards that location, it may be possible to achieve a probabilistic model of how infants develop the capacity to predict consequences of their vocal actions and assess the amount of information necessary in order to achieve such a predictive behavior at different developmental stages. Furthermore, by combining results from the development of predictive behavior on vocalizations with results from experiments on the development of predictive awareness for motor actions (from the parallel experiments on the capacity to predict the consequences of gestures, carried out at Uppsala University in the context of CONTACT-project), a unique opportunity of assessing potential relationships is created offering the prospect of a deeper understanding of the possible interactions between motor and vocal development.

The goal of this paper is to present the general outline of a series of on-going experiments and illustrate the outcome of these experiments with an example from a typical 12.5 months-old subject.

2. Method

2.1. Subjects

At this stage, the experiments were conducted only on a limited number of subjects because the goal still is trimming the methodology to resolve some practical issues that hampered the functionality of previous versions of the software used in the tests. Both infants and adults were tested in this phase of the study. The decision to include adult subjects in the experiment was motivated by the need to establish controlled baseline conditions that might be described in terms of the subject's reported awareness of the consequences of her own actions.

Four infants, 8.5, 10, 12.5 and 13.5 months old, and four adults were tested in this phase of the study. The adult subjects were student volunteers who were shown that things happened on the screen when there were vocalizations but did not receive information on the relationship between the F_0 and the location of the visual object on the screen.

2.2. Experimental setup

The experimental setup for this study used a combination of two systems: A specially designed program to implement voice control of a figure's position on a computer screen and the Tobii Eye-tracking system. The voice control program, JollerTrigger (documentation available online at http://eris.liralab.it/contact/reporting-period-1.shtml, deliverable 4.2), was created as a LabView application. The program estimates the fundamental frequency (F_0) of an input signal picked up by a microphone and displays a figure on a computer screen, at coordinates that are dependent on the instantaneous value of the signal's F_0. The Tobii Eye-tracking system is a non-invasive eye-tracking system that can measure the subject's gaze vector with a nominal precision of up to 0.5 degrees. Both JollerTrigger and Tobii generate log files that are subsequently analyzed to determine how well the subjects succeed in anticipating the consequences of their vocalizations by directing their gaze vector towards the screen position related to the instantaneous F_0 value of the subject's vocalization.

This set up did not use stimuli, in the sense that they were presented as independent variables for which the subjects' responses would be observed. A visual object – a 221×342 mm rectangle that randomly changed colors – was displayed in the center of the screen when no vocal activity was detected under a 4 seconds period but its only function was to re-direct the subject's attention towards the screen center and hopefully elicit new vocalizations. If a vocalization was detected this randomly colored rectangle disappeared from the screen until another 4 seconds period of inactivity occurred. Besides this there was no connection between the appearance of the rectangle and the subject's activities.

Contingent on the fundamental frequency of the subject's vocalizations, another visual object – a 221×342 mm cartoon figure – appears on the screen at an x-location determined by the logarithmic relation between the detected fundamental frequency and a pre-established reference F0 frequency. This reference frequency is always associated with the middle of the screen. F0 frequencies 1.6 octave below the reference frequency result in the placement of the figure's lower left corner at 0 pixels along the horizontal axis and frequencies 1.6 octave above the reference frequency place that lower left corner on the highest available horizontal pixel coordinate, in this case actually drawing the picture outside the screen limits. The center reference frequency was arbitrarily set to 500 Hz. Because it would be important to maintain the frequency-to-position mapping rule throughout the experiment, the 500 Hz value was chosen as a crude estimate of an expected average fundamental frequency that would increase the probability of placing the figure within the limits of the screen given the infant's typical fundamental frequency ranges.

The overall level of the vocalizations is only used to trigger the appearance of the figure and does not affect its position or clearness on the screen.

2.3. Procedure

The subjects sat in a dimmed lit studio, in front a computer screen equipped with a Tobii Eye-Tracking system.

The test session was initiated with the Tobii Eye-Tracking calibration procedure. After the calibration procedure, the JollerTrigger was initiated. In case the infant would not vocalize, the parent was encouraged to produce a vocalization to catch the infant's interest towards the screen and hopefully engage the infant in further vocalizations. The adult subjects were simply told that a figure would appear on the screen if they vocalized.

The subjects were free to experiment with JollerTrigger as much as they wanted. Their gaze vectors were continuously monitored and registered by the Tobii system. The fundamental frequencies of the subject's utterances were registered by JollerTrigger, along with the coordinates of the lower left corner of the visual object being displayed on the screen and a time stamp for subsequent data synchronization across the Tobii and JollerTrigger systems.

The experimental session had no time limit. Data were collected as long as the subject was willing to play with the system, which typically lasted for about 5 to 10 minutes.

The Eye-Tracking data was sampled at 50 Hz throughout the experimental session whereas the JollerTrigger data only was registered when there were vocalizations that triggered the appearance of the visual object on the screen. This means that in general the time stamps from JollerTrigger will fall in between the 20 ms of two consecutive time stamps from the Tobii system. A Mathematica program was subsequently used to line up these time stamps and generate graphical outputs showing the relationships between the position of the JollerTrigger's visual object and the subject's gaze vector towards the screen.

3. Results

The outcome of these experiments consisted in long log-files from both the Tobii system and from JollerTrigger. An example of the data compiled from these files is shown in figure 1. The figure presents an overview of EyeTracking (red line) and JollerTrigger data (green line). The data comes from the 12.5 months-old subject. The green line starts only at about 70 seconds after the onset of the Tobii system because the subject was initially silent and therefore the JollerTrigger system was not activated. The red line shows the instantaneous x-position of the subject's gaze vector, as measured by the Tobii system. To obtain this plot the raw

data collected by the Tobii system was used, instead of the filtered "fixation-data" available via Tobii's software. This was done in order to use as much temporal and spatial resolution as possible, although this requires dealing with a somewhat unstable signal. For each time frame, the fixation data from the left and the right eye, if it existed for both eyes, was combined to calculate the average coordinates of the two fixations. If only one eye had been tracked during that frame, data from that eye alone was used. Estimates of gaze locations falling outside the limits of the monitor were discarded. If no data was available within a time frame, the gaze location for that frame was discarded.

The red line in the graph on figure 1 displays the gaze position along the monitor's X-axis, for each of the time frames obtained from the Tobii system. The gaze position is given in mm from the left hand side of the monitor screen. Only the X-axis data is used here because this was the only dimension that was contingent on the F_0 frequency of the subject's utterances.

The graph's green line shows the position of the visual object when the subject's vocalizations were detected by JollerTrigger. The coordinates at which the visual object is being displayed are registered only when the object is visible, i.e. when the subject vocalizes above the pre-established threshold level. As mentioned above, if the subject is silent for more than four seconds, an attention catching rectangle is displayed in the middle of the screen. This event is not represented in the graph but its occurrence can be inferred from the long time gap between two consecutive points along the green line.

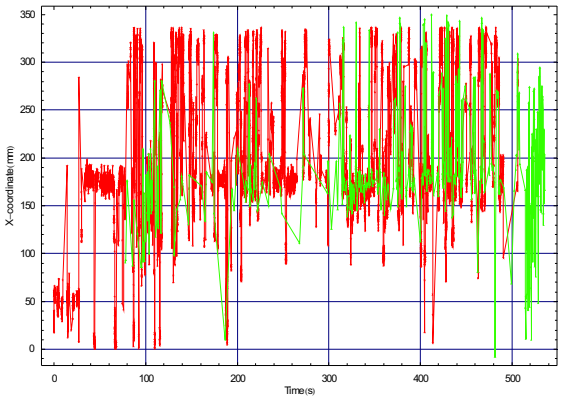

Figure 1: *Overview of the first 9 minutes of a JollerTrigger session with a 12.5 months-old subject. The Y-axis shows the instantaneous horizontal position of the gaze vector and of the visual object (whose position on the screen is controlled by F_0). See text for details.*

Figure 1 shows that the infant may have discovered the attention catcher rectangle after about 30 s in the session, as the gaze clearly moved towards the center of the screen, where the rectangle is displayed. After about another 30 s the first vocalization occurred, as indicated by the appearance of the first point on the green line. The subsequent mismatch between the red and the green lines suggests that the infant's gaze was initially not following very well the position of the figure. However, after about 5 minutes of experience, the situation seems to have changed and the red and green curves tend to follow each other more often.

In an attempt to quantify the proximity between the subject's gaze location and the actual position of the figure, two distance measures were introduced. Both measures were computed every time JollerTrigger had presented the figure on the screen in response to a vocalization. One of the measures looked at the average distance between the gaze points and the actual position of the figure, computed over a

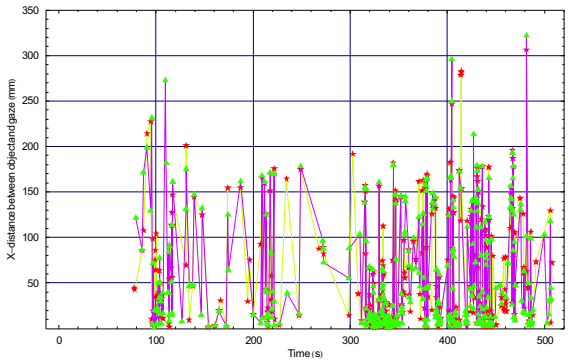

Figure 2: *Mismatch between the gaze location and the actual location of appearance of the figure, as seen through predictive (red stars and green-yellow line) or reactive (green triangles and violet line) windows with the durations specified in the legend.*

time window starting before the appearance of the figure and ending at the time when the figure was displayed on the screen. This will be referred to as the "Predictive time window". The other measure was similar to this one but it used instead a "Reactive time window", with onset at the time of presentation of the figure and extending forward in time for the duration of the time window. Figure 2 shows the evolution of the distances throughout the session above, as seen over a 100 ms predictive window and a 150 ms reactive window.

The data suggests that by about 300 s into the session there may have been a shift towards a closer match between both predictive and reactive gaze. However, while encouraging, these results must be viewed with some caution since at this time in the experiment the infant's vocalizations also tended to place the figure on about the same coordinates where the rectangle had been displayed. Thus, in the worst case scenario, the infant's gaze may be closer to the location of the figure during this period not because the infant is actually predicting the position where the figure is going to appear but rather because the figure appears at the location where the infant's gaze is anyway. This issue needs to be further investigated in a follow up of the present report.

4. Discussion

So far most of the research effort has been canalized to the technical problems that had to be solved in order to be able to conduct the experiments described in the present report. Many unsuccessful experiments were carried out with previous versions of JollerTrigger, in which several acoustic dimensions were initially being used to control different aspects of the appearance and location of the figure. Unfortunately the combinatorial explosion created by the simultaneous use of several simultaneous dimensions seems to curtail the infant's possibility to pick up relevant

contingencies during the practical time window of an experiment session. Obviously the infant's real life experiences are far more complex than those created in a laboratory session, in spite of its apparent intricacy. However, because the infant's mobility is relatively constrained in a lab situation where gaze measurements are required, the typical infant will tend to get bored after a short while and it is extremely difficult to make the infant recover from that state of mind and regain interest on the experiment. This means that although in principle the infant may be able to deal with more complex situations than those created in the laboratory environment, the time window available in a controlled experimental situation is often far too short to give the infant a fair chance to pick up complex underlying contingencies. In addition, it is also often methodologically difficult to divide the experiment session into a number of partial exposures while meeting strict criteria of experiment control. For these reasons the option adopted in the current experiments was to create a simple enough situation that would give the infant a fair chance to discover the consequences of its vocal behavior during the time span of an experiment session. The current results suggest that this may have been a successful research strategy and further analysis of the data gathered so far will provide valuable insight on which technical adjustments should be implemented in the future. So far, the sample of data collected seems to suggest that one-year olds may be able to start predicting the consequences of their vocal actions after some minutes of interplay with an object responding to those vocal actions.

5. Conclusions

Preliminary data from further experiments following this study appears to suggest that at some point around 12 to 14 months of age infants may be able to start predicting the consequences of their vocal actions as revealed by their gaze behavior. If this type of results can indeed be confirmed by the ongoing experiments, there is an important linguistic message to be learned from them since the ability to predict the consequences of one's vocal actions is a core aspect for the emergence of a linguistic referential function. Obviously, a question in line concerns the articulation of these results with those from our colleagues at Uppsala University (http://eris.liralab.it/contact/reporting-period-1.shtml, deliverables D3.1 and D3.2): How well can such a prediction based on vocal gestures be integrated in the general framework of parallel development of vocal and motor actions?

Pursuing this line of research and achieving an adequate model of the infant's early cognitive and linguistic development are the main goals of the CONTACT- and MILLE-projects within which this research is being carried out.

6. Acknowledgements

This work was carried out with the joint effort of all the CONTACT and MILLE project teams at Stockholm University and with the collaboration of parents and students who volunteered to participate in different series of pilot experiments and proper experiment sessions.

Work supported by grants from the Bank of Sweden Tercentenary Foundation (MILLE, K2003-0867) and EU NEST program (CONTACT, 5010).

7. References

[1] C. von Hofsten, "An action perspective on motor development," Trends Cogn Sci., vol. 8, no. 6, pp. 266-272, June2004.

[2] K. Rosander and C. von Hofsten, "Development of gaze tracking of small and large objects," Exp. Brain Res., vol. 146, no. 2, pp. 257-264, Sept.2002.

[3] B. L. Davis and B. Lindblom, "Phonetic Variability in Baby Talk and Development of Vowel Categories," in Emerging Cognitive Abilities in Early Infancy. F. Lacerda, C. von Hofsten, and M. Heimann, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2001, pp. 135-171.

[4] F. Lacerda, C. von Hofsten, and M. Heimann, Emerging Cognitive Abilities in Early Infancy. Hillsdale: Erlbaum, 2001.

[5] C. von Hofsten, "Action in development," Dev. Sci., vol. 10, no. 1, pp. 54-60, Jan.2007.

[6] F. Lacerda and U. Sundberg, "An Ecological Theory of Language Acquisition,", 1 ed Porto, Portugal: Faculdade de Letras da Universidade do Porto & Centro de Linguística da Universidade do Porto, 2006, pp. 53-106.

[7] F. Lacerda, E. Klintfors, L. Gustavsson, E. Marklund, and U. Sundberg, "Emerging linguistic functions in early infancy," in Epigenetic and Robotics, 123 ed. L. Berthouze, F. Kaplan, H. Kozima, H. Yano, J. Konczak, G. Metta, J. Nadel, G. Sandini, G. Stojanov, and C. Balkenius, Eds. Nara, Japan: Lund University Cognitive Studies, 2005, pp. 55-62.

[8] F. Lacerda, E. Klintfors, L. Gustavsson, L. Lagerkvist, E. Marklund, and U. Sundberg, "Ecological Theory of Language Acquisition,", Epigenetics and Robotics 2004 ed Genova: Epirob 2004, 2004.