In Berthouze, L., Kozima, H., Prince, C. G., Sandini, G., Stojanov, G., Metta, G., and Balkenius, C. (Eds.)
Proceedings of the Fourth International Workshop on Epigenetic Robotics
Lund University Cognitive Studies, 117, ISBN 91-974741-3-4

Integration of audio-visual information in 8-months-old infants

Lisa Gustavsson*, Ulla Sundberg, Eeva Klintfors, Ellen Marklund, Lisa Lagerkvist and Francisco Lacerda
Department of Linguistics
Stockholm University
SE-106 91 Stockholm
Sweden
* lisag@ling.su.se

Abstract

The results from a series of perception experiments designed to test 8-month-old infants' ability to derive linguistic information from audio-visual events are reported in this presentation.

Using a visual preference technique, groups of 8-month-old infants were tested on their ability to extract linguistic information implicit in short video sequences where the images displayed different puppets and the audio tracks presented sentences describing the puppets in naturalistic infant-directed speech style. To assess the relative importance of memory and attention factors, the prosodic and syntactic structure of the speech materials was systematically changed across different groups of subjects. The experimental results are interpreted in terms of the emergentistic acquisition model discussed in the paper presented by Lacerda et al. ("Ecological theory of language acquisition").

1. Introduction

Designing general and flexible systems that may be capable of learning from their experience of interaction with the immediate environment is a central problem for the fields of robotics and artificial intelligence. Whereas demands of designing extremely high performance systems that are able to cope with high precision complex tasks can be met with the techniques available nowadays provided the environment variables are well under control, designing a system that is able to achieve general learning from its "living experience" is quite another matter. Indeed, because there seems to be an inescapable trade-off between flexibility and specificity, the design of general intelligent systems requires a deeper understanding of how that balance is solved in natural biological systems. In this vein, the study of the early stages of the human language acquisition process offers a unique opportunity of understanding how intelligent systems may handle and structure information generally available in their environments. To be sure, developmental studies consistently indicate that, after about 12 months of spontaneous and apparently effortless interaction with its ambient language, the young infant typically acquires the linguistically essential referential function and explores shortly thereafter the combinatorial power of word concatenation. Yet, even accounts of this amazing linguistic development often concentrate on surface aspects, like the process's acoustic manifestations, rather than on a deeper search for its possible general underlying principles.

The emergentistic approach outlined in the "Ecological theory of language acquisition" (Lacerda et al., 2004) is an attempt to view the process of early language acquisition as the result of the infant's interaction with its immediate linguistic environment. In the present context the Ecological Theory implies a shift in the focus in system design, from attempts to bypass the long developmental path that eventually lead to an adult competence tend to result in rather rigid systems that can hardly cope with realistic variance in speech, towards systems achieve linguistic structure relying on built-in, linguistically unspecific, general sensory, representational and interactive abilities. In a wider perspective our aim is to design a prototypical system that essentially will be able to learn from multimodal information sources and integrate them to achieve flexible and realistic representations of its environment, mimicking an infant's behaviour. The ambient language of almost every infant is dominated by Infant Directed Speech (IDS), a typical speech style used by adults when communicating with infants (van de Weijer, 1999). IDS highlights the language-specific properties in the speech signal and there is no doubt that this structured speech style is helping the infant on its way towards a spoken language. Whereas communication between adults usually is about exchanging information and emotions, speech directed to infants is of a more referential nature. The adult refers to objects, people and events in the world surrounding the infant (Lacerda, Marklund et al. 2004). Because of this, the sound sequences the infant hears are very likely to co-occur with actual objects or events in the infant's visual field. The expanded intonation contours, the repetitive structure of IDS and the modulation of the sentence intensity (Lacerda and Molin, 2002) are likely to play an important role in helping the infant establishing an implicit and plausible word-object link.

The present poster will present the current results of an investigation on target-word position and stress may influence the infants' ability to establish such audio-visual object links.

2. Method

A Visual Preference procedure, essentially a version of Fernald's Preferential Listening Procedure (Fernald, 1985), was used. After a short exposure to speech materials that are presented in connection with the presentation of visual objects, the infant's looking time towards a target object while listening to sentences referring to that object is compared with the looking time towards a competing non-target object.

1.1 Speech materials

A female speaker of Swedish recorded the speech materials in nine different conditions where main stress and target-word were placed in all possible combinations of sentence initial, medial and final positions. The sentences introduced non-words as possible names of objects (ex: "It is a nice *nnnn*"). The speech materials were produced in IDS-style, which is characterized by modifications such as frequent prosodic repetitions and expanded intonation contours (Fernald, 1989).

1.2 Visual materials

The audio-visual materials were organized in 3 minutes long video films. The objects (puppets) were presented in the films visually and auditory with corresponding names embedded in phrases. Each film consisted of three phases – baseline, exposure and test phase.

• Baseline – In this phase a split-screen of two puppets side by side. The infants' initial visual bias (spontaneous preference) towards the puppets was measured during this baseline phase. The audio track during this phase consisted of an especially composed short instrumental "infant-appropriate" melody (Anna Ericsson, 2004) played throughout the 30 seconds of the baseline phase.

• Exposure – During this phase alternating 20 seconds full screen presentations of one of the two puppets, with audio tracks matched to each of the displayed puppets, were presented 3 times each. The total duration of the exposure was 120 seconds. The infants' looking time towards the each of the puppets was taken as a measure of attention during the exposure phase.

• Test – In the test phase the two puppets were displayed again in a split-screen similar to that of the base-line. This time however the audio track referred to one of the puppets which name was embedded in questions. The infants' gain in looking time towards the target-object, as compared to initial bias towards that future target in Phase 1, was taken as a measure of changed preference (Test-bias gain), was taken as an indication of a sound-meaning connection. Phase 3 lasted for 30 seconds.

1.3 Subjects

The subjects were randomly selected from the National Swedish address database (SPAR) among 8 month-old infants whose parents lived in the Stockholm metropolitan area. A total of 50 infants participated in the study. The subjects were randomly assigned to watch one, two or three of the films. The parents participated voluntarily and were not paid for their participation.

1.4 Procedure

A video camera recorded the infant watching the film. The infant's looking behaviour was subsequently analyzed frame-by-frame to determine the focus of the infant's visual attention on the screen. A video camera recorded a close-up image of the infant watching the film.

The infant was seated on the parent's lap. To reduce the risk of interacting with the infant, the parent listened to music through soundproof headphones during the whole procedure. Each infant's recording was analyzed frame-by-frame (with precision 0.04 sec) as eye movements to left-, right-, front-, or off-screen. The separation between the puppets on the split-screen situations was about 30^{o} and the current manual procedure allows for a resolution of about 10-15 degrees.

3. Results

The video data is currently being analyzed but there are preliminary indications of a differential response attributable to the exposure to the language materials.

4. Discussion

The current experiments are expected to disclose essential aspects underlying the emergence of the linguistic referential function and how target-word and main-stress placement influences it.

Acknowledgements

Research carried out with grants from the Swedish Research Council and the Bank of Sweden Tercentenary Foundation.

Reference List

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Dev., 60,* 1497-1510.

Fernald, A. (1985). "Four-month-old infants prefer to listen to motherese", *Infant Behavior and Development 8*, 181-195.

Lacerda, F. & Molin, J. (2002). Stress judgements by naïve listeners. Fonetik 2002,the XVth Swedish Phonetics Conference, Fysikcentrum, Stockholm, May 29 - 31, 2002. Quartely Progress and Status Report (Department of Speech, Music and Hearing and Centre for Speech Technology, KTH, Stockholm) 44: 145-148.

Lacerda, F., Gustavsson, L and Svärd, N. (2003). Implicit linguistic structure in connected speech. *PHONUM 9*, Umeå, Sweden, 69-72.

van de Weijer, J. (1999). *Language Input for Word Discovery.* MPI Series in Psycholinguistics.