Potential relevance of audio-visual integration in mammals for computational modeling

Eeva Klintfors and Francisco Lacerda

Department of Linguistics
Stockholm University, Stockholm, Sweden
eevak@ling.su.se

ABSTRACT

The purpose of this study was to examine typically developing infants' integration of audio-visual sensory information as a fundamental process involved in early word learning. One hundred sixty pre-linguistic children were randomly assigned to watch one of four counterbalanced versions of audio-visual video sequences. The infants' eye-movements were recorded and their looking behavior was analyzed throughout three repetitions of exposure-test-phases. The results indicate that the infants were able to learn covariance between shapes and colors of arbitrary geometrical objects and to them corresponding nonsense words. Implications of audio-visual integration in infants and in non-human animals for modeling within speech recognition systems, neural networks and robotics are discussed.

Index Terms: language acquisition, audio-visual, modeling

1. INTRODUCTION

This study is part of the "MILLE" project (Modeling Interactive Language Learning, supported by the Bank of Sweden Tercentenary Foundation), interdisciplinary research collaboration between three research groups – Department of Linguistics, Stockholm University (SU, Sweden), Department of Psychology, Carnegie Mellon University (CMU, USA) and Department of Speech, Music and Hearing, Royal Institute of Technology (KTH, Sweden). The general goals of the co-work are to investigate fundamental processes involved in language acquisition and to develop computational models to simulate these processes.

As a first step towards these goals, perceptual experiments with human infants were performed at the SU. These experiments were designed to investigate infants' early word acquisition as an implicit learning process based on integration of multi-sensory (audio-visual) information. The current study specifically provides data on integration of arbitrary visual geometrical objects and "nonsense" words corresponding to attributes of these objects.

2. BACKGROUND

2.1 Research in mammals

As a theoretical starting point for the current study it is assumed that infants do not have a priori specified linguistic knowledge and that the acquisition of the ambient language is guided by general perception and memory processes. These general purpose mechanisms are assumed to lead to linguistic structure through learning of implicit regularities available in the ambient language. Thus, as opposed to the belief that initial innate guidance is a prerequisite for language acquisition [1], [2], the proposed Ecological Theory of Language Acquisition (ETLA) suggests that the early phases of the language acquisition processes are an emergent consequence of the interplay between the infant and its linguistic environment [3]. To be sure, observations based on implicit regularities may indeed lead to wrong assumptions and, just like other experience-based-learning, this type of trial and error use of words tends initially to create situated knowledge [4].

To be able to extract and organize implicit sensory information available from several modalities is an important ability for an organism's success in its environment. Implicit learning processes are presumed to occur also in non-human mammals – a hypothesis being currently investigated by our collaborators at CMU. Explicitly concerning audio-visual integration in non-human animals – Ghazanfar & Logothetis [5] showed using preferential-looking technique that rhesus monkeys (Macaca mulatta) were able to recognize auditory-visual correspondence between their conspecific vocalizations ("coo" or "threat" calls) and appropriate facial gestures (small mouth opening/protruding lips or big mouth opening/no lip protrusion). Just like the perception of human speech, the perception of monkey calls may thereby be enhanced by a combination of auditory signals and visual facial expressions. Bimodal perception in animals was viewed by the authors as an evolutionary precursor of human's ability to make multimodal associations necessary for speech perception. Such studies on non-human mammals are obviously important to examine questions that are impossible, unethical, or extremely difficult to answer with human listeners [6], like the pre and post operative comparisons of rhesus monkeys' performance in auditory-visual (e.g. noise-shape pairs) association tasks supporting the notion that left prefrontal cortex plays a central role in integration of auditory and visual information [7].

2.2 Modeling within speech recognition systems, neural networks and robotics

After about one year's exposure to their ambient language, children typically start to speak to interact with their environment. Despite of its complexity, children soon learn

the linguistic principles of their ambient language. However, this seemingly simple task is not easily transferred to formalized knowledge about language acquisition, nor is it easily integrated within speech recognition or operational models. Part of the problem that speech recognition systems battle with is presumably caused by the focus on the speech signal as the primary component of the speech communication process. Within natural speech communication speech is only one, albeit crucial, part of the process and the latest systems have started to make use of multimodal information to improve the systems' communication efficiency. As an example, Salvi [8] analyzed the behavior of Incremental Model-Based Clustering on infant directed speech data in order to describe acquisition of phonetic classes by an infant. Salvi analyzed the effects of two factors within the model, namely the number of coefficients describing the signal and the frame length of the incremental clustering. His results showed that despite varying amount of clusters, the classifications obtained were essentially consistent. In addition to the model by Salvi the current co-work with the KTH further aims to develop a computational model able to handle information received from at least two different sensory dimensions.

We have recently reported [9] that two neural network models submitted to process encoded video materials that were earlier tested on a group of adult subjects in a simple inference task (similar to the task in the current study), performed well on both classification and generalization tasks. The task of the first type of architecture was simply to associate colors and shapes of the visual objects to the words corresponding to these two attributes, i.e. the model merely reproduced its input at the output level. The second architecture of the model attempted to simulate the adult subjects' ability to apply their just learned concepts to new contexts. The performance of the model was tested with novel data not previously seen by the network (either new colors or new shapes). The performance of both network models was robust and mimicked the adults' results well. Since the outcome of a neural network is dependent on the peculiarities of the input coding, its architecture and specific training constraints, the type of neural network models described here would presumably not mimic well enough the behavior of the infants in the current study. One reason for a vague match of behavior would probably be caused by the unlimited memory of the network which enables it to process data unrestricted as compared with infants' restricted memory capacities. Also neural networks, often using sigmoid activation function for nodes, are non-linear regression models in which small differences in input value may cause large differences in neural computation behavior. Hence, a better way of modeling infant behavior would be to calculate effects of memory constraints when predicting audio-visual integration.

In addition to development of communicative ability leading to more sophisticated use of language, children quickly learn to interact with their environment through observation and imitation of manipulative gestures. To explore whether these motor abilities are developing independently, or if there are fundamentally similar mechanisms involved in development of perception and production of speech and manipulation of gestures, results from the current study and similar other experimental studies are tested within an embodied humanoid robot system. The hypothesis deals with a recent scientific finding pointing at the fact that action representations can be addressed not only during execution but also by the perception of action-related stimuli. Experiments with monkeys have shown that mirror neurons in premotor cortex (area F5) discharge both when a monkey makes a specific action and when it observes another individual who is making a similar action [10], [11] that is meaningful for the monkey. A mirror system is proposed also to exist in humans [12] and F5, the area for hand movements, is suggested to be the monkey homolog of Broca's area, commonly thought of as an area for speech, in the human brain [13]. Both F5 and Broca's area have the neural structures for controlling oro-laryngeal, oro-facial, and brachio-manual movements [14]. Further studies in neuroscience suggest that there are parallels in mechanisms underlying actions such as to manipulate, tear, or put an object on a plate and mechanisms underlying actions recognized by their sound or mechanisms underlying speech systems [15].

For our research group at the SU the importance of modeling language acquisition lies unquestionably within the possibility of experimentally manipulating learning processes on the basis of experimental achievements and theoretical hypothesis formulated by us and other Neuroscience and Child-development partners.

3. METHOD

3.1 Subjects and procedure

The subjects were 160 Swedish infants randomly selected from the National Swedish address register (SPAR) on the basis of age and geographical criteria. Thirty-one infants were excluded from the study due to interrupted recordings (infant crying or technical problems). The remaining 129 infants were divided in two age groups: 26 boys and 27 girls in age-group I (age range 90-135days, mean age 120 days) and 42 boys and 34 girls in age-group II (age range 136-180days, mean age 156 days). The subjects were randomly assigned to watch one of four counterbalanced film sequences. The subjects and their parents were not paid to participate in the study. The infant was seated in a safe baby-chair or on the parent's lap at approximately 60 cm from the screen. The parent listened to masking music through soundproof headphones during the whole session. The infants' eye-movements were recorded with Tobii (1750, 17" TFT) Eye-tracking system using low-intensity infra-red light. The gaze estimation frequency was 50Hz, and accuracy 0.5 degrees. Software used for data storage was ClearView 2.2.0. The data was analyzed in Mathematica 5.0 and SPSS 14.0.

3.2 Materials

The films' structure was: BASELINE (20 s), EXPOSURE1 (25 s), TEST1 (20 s), EXPOSURE2 (25 s), TEST2 (20 s), EXPOSURE3 (25 s), and TEST3 (20 s). The image used to measure infants' pre-exposure bias in BASELINE is shown in Figure 1. During BASELINE the audio played a lullaby to catch the infant's attention towards the screen.

The elements used in the EXPOSURE phases are shown in Figure 2. Each of the elements was shown in 6 s long film sequences. Each object moved smoothly across the screen while the audio played 2 × repetitions of two-word phrases, such as *nela dulle* (red cube), *nela bimma* (red ball), *lame dulle* (yellow cube), *lame bimma* (yellow ball), implicitly referring to the color and shape of the object. The two-word

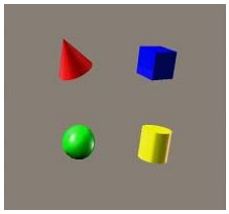

Figure 1: Split-screen displaying four objects shown during the BASELINE along with a lullaby to measure infants' spontaneous looking bias before the three repetitions of exposure and subsequent test phases.

After presentation (EXPOSURE1-3) of the objects (Figure 2) the split-screen was displayed again during TEST1-3 along with questions such as *Vur bu skrett dulle?* (Have you seen the cube?) and *Vur bu skrett nela?* (Have you seen the red one?). The films were counterbalanced regarding choice of words corresponding to colors/shapes of the objects.

phrases were read aloud by a female speaker of an artificial language in infant-directed speech style. The "nonsense" words were constructed according to the phonotactic and morphotactic rules of Swedish. However, the prosody of the phrases did not mimic Swedish prosody of two-word phrases – the words were instead pronounced as if they occurred in isolation, without sentence accent on either one of the words.

Figure 2: An illustration of the elements shown during EXPOSURE1-3. Each object moved smoothly across the screen (in 6 s long film sequences) while the audio played 2 × repetitions of two-word phrases, such as *nela dulle* (red cube), *nela bimma* (red ball) *lame dulle* (yellow cube), *lame bimma* (yellow ball), implicitly referring to the color and shape of the object. The films were counterbalanced regarding the presentation order of the objects during exposure.

During TEST1, TEST2 and TEST3 the image shown in Figure 1 appeared again while questions such as *Vur bu skrett dulle?* (Have you seen the cube?) or *Vur bu skrett nela?* (Have you seen the red one?) were asked. The questions were naturally produced using question intonation.

In each one of the four counterbalanced versions of the film there was one question on shape of an object and another on color of an object. In the test phases the target shapes appeared in new colors and the target colors appeared in new shapes – as compared with the shape-color combinations during exposure phases – in order to address generalization of learned associations into new contexts.

4. RESULTS

We predicted that if infants are capable of extracting the objects' underlying properties, their looking times towards the relevant target color (**Red** or **Yellow**) and target shape (**Cube** or **Ball**) of an object will increase from BASELINE (Pre-exposure bias) to TEST1-3 (Post-exposure).

The results (Figure 3) showed increased looking times towards **Red** upper-left (UL), **Cube** upper-right (UR), **Ball** lower-left (LL) and **Yellow** lower-right (LR) from BASELINE to TEST1-3. The increments in looking time were larger in response to target shapes (**Cube** and **Ball**) as compared with target colors (**Red** and **Yellow**). This was in particular true for age-group II. Furthermore, repeated measures ANOVA on BASELINE to TEST1-3 contrasts revealed significant differences for **Red**, **Cube** and **Ball**.

The looking behavior of age-group II – as indicated by the error bars – was more stable than the looking behavior of age-group I. An analysis on BASELINE to TEST1-3 contrasts with age as a factor, revealed an overall age-group tendency for **Red** ($F = 4.106$, d.f. = 18, $P < 0.058$) indicating disparity in the looking behavior of infants in age-group I and II.

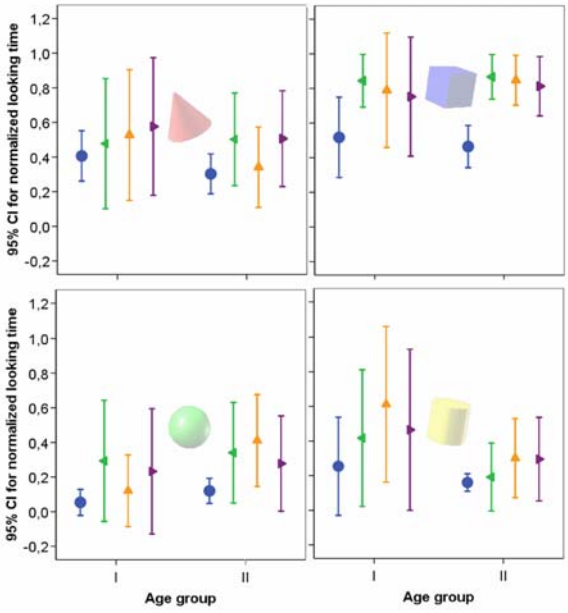

Figure 3: The attribute targets were **Red** (UL), **Cube** (UR), **Ball** (LL) and **Yellow** (LR). The bars in each quadrant from left to right respectively indicate: Pre-exposure bias (normalized looking time at future target during BASELINE), and Post-exposure target dominance (normalized looking time at target during TEST1-3) for age-group I and II respectively.

There were no significant interactions between the times spent looking at target and the choice of words corresponding to the colors and shapes of the objects or the object-position (left or right) on the split-screen. Infants did, however, look longer towards the upper quadrants (target **Red** $F = 37.564$, d.f. = 54, $P < 0.0005$ and target **Cube** $F = 26.758$, d.f. = 63, $P < 0.0005$) than towards the lower quadrants.

5. DISCUSSION

Whereas early studies on language acquisition in infants were focused on production and perception of isolated speech sounds [16], recent experimental studies have addressed the structure of perceptual categories [17], and word extraction from connected speech [18], [19], [20], [21]. The current study further expands the scope of these investigations by addressing the emergence of referential function, integrating

recognition of patterns in auditory with the recognition of patterns in the visual input. In addition, our theoretical outline views early language acquisition as a consequence of general sensory and memory processes. Through these processes auditory representations of sound sequences are linked to co-occurring sensory stimuli and since spoken language is used to refer to objects and actions in the world, the implicit correlation between hearing words relating to objects and seeing (or feeling, smelling or otherwise perceiving) the referents can be expected to underline the acquisition of spoken language.

The materials in the current study were, as opposed to synthetic stimuli, naturally produced two-word phrases, whose linguistic meaning emerges from their situated implicit reference to the shapes and colors of the visual geometrical objects. The adjective-noun pattern in the two-word phrases followed the Swedish syntactical pattern. This situation, in which four new words were mapped onto a known grammatical structure, is of course different from the challenge faced by a pre-linguistic child, but simplifying the experimental materials (to only two shapes and two colors) was necessary for revealing the essential process of audio-visual integration during short exposure (about 1 minute) in young infants (< 6mos). Also, because success in the test phases required generalization of learned associations into new visual contexts, finding the correct visual target cannot be seen as simple “translation process”.

The age range of the subjects was 90-135 days (age-group I) and 136-180 days (age-group II). These age-groups were selected to investigate the extent of the age-related differences in the infants’ capacity to learn covariance between the two different types of attribute targets (shapes and colors) and to them corresponding words. Indeed, despite of the fact that color words typically appear first after about ten months, these young infants showed some evidence of being able to pick up the color references implicit in the current experiment. Thus, although infants in this age range may not have the capacity to handle the concept of color as such, they at least demonstrate, as observed in other mammals, an underlying capacity to associate recurrent visual dimensions with their implicit acoustic labels.

6. ACKNOWLEDGEMENTS

Research supported by the Swedish Research Council (421-2001-4876), the Bank of Sweden Tercentenary Foundation (K2003-0867) and Birgit & Gad Rausing’s Foundation.

7. REFERENCES

[1] N. Chomsky, *Language and mind.* New York: Harcourt Brace Jovanovich, 1968.

[2] S. Pinker, *The Language Instinct: How the Mind Creates Language*, 1 ed. New York: William Morrow and Company, Inc., 1994, pp. 15-472.

[3] 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.

[4] E. J. Gibson and A. D. Pick, *An Ecological Approach to Perceptual Learning and Development* Oxford University Press US, 2006.

[5] A. A. Ghazanfar and N. K. Logothetis, "Neuroperception: Facial expressions linked to monkey calls," *Nature*, vol. 423, pp. 937-938, 2003.

[6] K. R. Kluender, A. J. Lotto, and L. L. Holt, "Contributions of nonhuman animal models to understanding human speech perception," in *Listening to Speech: An Auditory Perspective*. S. Greenberg and W. Ainsworth, Eds. New York, NY: Oxford University Press, 2005.

[7] D. Gaffan and S. Harrison, "Auditory-visual associations, hemispheric specialization and temporal-frontal interaction in the rhesus monkey," *Brain*, vol. Oct; 114, no. Pt 5, pp. 2133-2144, 1991.

[8] G. Salvi, "Ecological Language Acquisition via Incremental Model-Based Clustering,", InterSpeech'2005 - Eurospeech, 9th European Conference on Speech Communication and Technology ed Lissabon: InterSpeech'2005, 2005.

[9] F. Lacerda, E. Klintfors, and L. Gustavsson, "Multi-sensory information as an improvement for communication systems efficiency,", Fonetik 2005 ed Gothenburg: Fonetik 2005, 2005, pp. 83-86.

[10] G. Rizzolatti and M. A. Arbib, "Language within our grasp," *Trends in Neurosciences*, vol. 21, no. 5, pp. 188-194, May1998.

[11] C. Keysers, E. Kohler, M. A. Umilta, L. Nanetti, L. Fogassi, and V. Gallese, "Audiovisual mirror neurons and action recognition," *Experimental Brain Research*, vol. 4, pp. 628-636, 2003.

[12] L. Fadiga, L. Fogassi, G. Pavesi, and G. Rizzolatti, "Motor Facilitation During Action Observation - A Magnetic Stimulation Study," *Journal of Neurophysiology*, vol. 73, no. 6, pp. 2608-2611, June1995.

[13] G. Rizzolatti and M. A. Arbib, "Language within our grasp," *Trends in Neurosciences*, vol. 21, no. 5, pp. 188-194, May1998.

[14] G. Rizzolatti and M. A. Arbib, "Language within our grasp," *Trends in Neurosciences*, vol. 21, no. 5, pp. 188-194, May1998.

[15] E. Kohler, C. Keysers, M. A. Umilta, L. Fogassi, V. Gallese, and G. Rizzolatti, "Hearing sounds, understanding actions: action representation in mirror neurons," *Science*, vol. 297, no. 5582, pp. 846-848, Aug.2002.

[16] P. D. Eimas, E. R. Siqueland, P. Jusczyk, and J. Vigorito, "Speech perception in infants," *Science*, vol. 171, pp. 303-306, 1971.

[17] P. Kuhl, K. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science*, vol. 255, no. 5044, pp. 606-608, Jan.1992.

[18] P. Jusczyk, "How infants begin to extract words from speech," *Trends Cogn Sci.*, vol. 3, no. 9, pp. 323-328, Sept.1999.

[19] P. Jusczyk and R. N. Aslin, "Infants' Detection of the Sound Patterns of Words in Fluent Speech," *Cognitive Psychology*, vol. 29, no. 1, pp. 1-23, Aug.1995.

[20] P. Jusczyk and E. Hohne, "Infants' Memory for Spoken Words," *Science*, vol. 277, no. 5334, pp. 1984-1986, Sept.1997.

[21] J. R. Saffran, R. N. Aslin, and E. Newport, "Statistical learning by 8-month old infants," *Science*, vol. 274, pp. 1926-1928, 1996.