Investigating the Emergence of Speech Communication – A Study on Infants' Ability to Predict Phonetic Information

Klintfors Eeva (eevak@ling.su.se)
Department of Linguistics, Universitetsvägen 10C,
106 91 Stockholm, Sweden

Lacerda Francisco (frasse@ling.su.se)
Department of Linguistics, Universitetsvägen 10C,
106 91 Stockholm, Sweden

Marklund Ellen (ellen@ling.su.se)
Department of Linguistics, Universitetsvägen 10C,
106 91 Stockholm, Sweden

Abstract

The introduction of this paper provides an overview of infants' prediction skills of action goals, as well as their ability to predict perceptual acoustic information. Prediction skills' neurological correlates in general are discussed. A central hypothesis under investigation is that there are commonalities between the development of speech and manipulation. The current research is focused on the communication mode investigating infants' ability to associate images of familiar objects with auditory-stimuli presented both as whole words in intact form and as disrupted (partly noise-replaced) spoken words. The looking behaviour of the infants' was measured with the Tobii eye-tracking device. The results suggested that 11 to 16 month-old infants recognize the target object when the word referring to it was intact, i.e. when the name of the object was presented in its entirety. However, the infants did not seem to recognize the target object when the word referring to it was partially masked so that only its initial phonetic information was presented. These results indicate that young infants are sensitive to the phonetic information of the words and may need more extensive linguistic experience in order to derive full lexical forms from partially masked words. The paper concludes with suggestions for future demonstrations of infant anticipation of speech.

Introduction

Prediction of Other People's Action Goals

Recent research indicates that one-year-old infants are able to predict other people's action goals (Falck-Ytter et al., 2006). In these experiments, proactive goal-directed eye-movements of adults were compared with 6-month-old and 12-month-old infants' gaze behaviour. The subjects were exposed to video presentations showing trials in which toys were moved by an actor's hand into a bucket. The results showed that adults, as well as 12-month-old infants directed their gaze towards the bucket before the toy had reached it, while 6-month-old infants seemed to fail anticipating the goal of the action. The younger infants' inability to predict the actor's intention was suggested to indicate that predictive action perception in children develops simultaneously with other social competences such as imitation, "theory of mind" and communication. Such gestural and linguistic competencies are assumed to emerge typically by about 8-12 months of age. The authors also speculated on the possibility that the 6-month-old infants' inability to predict the actor's action goal could originate from their general inability to predict future events with their gaze. However this assumption was rejected as it conflicts with the results from earlier experiments demonstrating that 6-month-olds could indeed predict the reappearance of temporarily occluded objects.

Falck-Ytter et al.'s (2006) study also suggested that successful goal prediction relies on observing the interaction between the hand of the agent and the object. Indeed, the results from another experimental condition where the objects moved along the same paths while the actor's hand was not visible showed that predictive eye movements were not activated for any of the subject groups (adults, 12-month-old, or 6-month-old infants), suggesting that self-propelled objects are not perceived as performing goal-directed actions.

Neurological Coding System of Mirror-Neurons

The recent discovery of mirror-neurons presents a powerful neurological correlate of predictive perceptual behaviour. Activation of the mirror-neuron system (MNS) was first found in the ventral premotor cortex (area F5) of macaque monkeys (Rizzolatti & Arbib, 1998; Keysers et al., 2003) when a macaque observed another individual (human or monkey) performing an action that could be related to observer's own repertoire of actions. An important aspect is that MNS was only activated when the goal of the action was clear to the animal. In fact, mimicking the gestures of goal-directed actions, without actually using the objects typically involved in the actions (like when pretending to peal an invisible banana) did not lead to activation of the MNS. The behavioural study by Falck-Ytter et al. (2006) is well in line with this assumption, indicating that the absence of an actor impairs the observer's predictive perception of goal-oriented actions.

Furthermore, data from neuro-physiological and brain-imaging experiments indicates that MNS also exists in humans (Fadiga et al., 1995; Grafton et al., 1996) and it has been suggested that it may constitute a neurological base for coding empathy, social understanding and the ability for human communication (Rizzolatti & Craighero, 2004) .

Mirror-Neuron System and Communication

Mirror neurons may account for coding of object-directed actions in which the gestural meaning is intrinsic to the gesture itself, but the question of whether the same system may be able to mediate abstract symbolic representations that are typically involved in human communication must also be addressed. In line with this, a relevant finding by Ferrari et al. (Ferrari et al., 2003) indicated that mirror-neurons of F5 may also code *mouth*-actions. Most of these "mouth mirror-neurons" were observed to become active both during the execution and observation of mouth-actions related to ingestive functions (e.g. grasping, sucking or breaking food) but some of the mirror neurons responded particularly to *communicative* mouth gestures (e.g. lip smacking). Therefore these findings extend the scope of MNS from hand-actions to mouth-actions suggesting that area F5, considered to be the homologue of human Broca´s area, is also involved in communicative behaviour.

From an evolutionary perspective, Ferrari et al. (Ferrari et al., 2003), suggested that understanding words related to mouth-actions may have evolved via activation of *audio-visual* mirror neurons. Mirror neurons initially responding to ingestive behaviour, may have led to a further development in which a set of F5 audio-visual mirror neurons eventually became responsive to the *sound* associated with the original actions, like hearing the sound of ripping a piece of paper without actually seeing the action (Kohler et al., 2002). Pursuing this evolutionary perspective Rizzolatti & Craighero (Rizzolatti & Craighero, 2004) speculated that the human individuals' improved imitation capacities may have enabled the generation of onomatopoetic sounds without actually performing the action that originally generated the sounds. Hence, this capacity might have led to the acquisition of an *auditory* mirror system on the top of the original audio-visual one, progressively independent of it (see also (MacNeilage & Davis, 2000)).

Prediction of Phonetic Information

Earlier research (Samuel, 1996; Warren, 1970; Warren & Obusek, 1971; Warren & Warren, 1970) has shown that adults are able to interpret and reconstruct disrupted speech signals. In these studies adult listeners could identify both words disrupted by noise (Warren, 1970), as well as words disrupted by silence (Warren & Obusek, 1971). In the noise-disrupted case subjects reported that the disrupted words sounded intact. This phenomenon suggests that listeners perceive the word to continue behind the noise and "restore" missing phonemes – a phenomenon known as "phoneme restoration". In the silence-disrupted case, subjects did not perceive the word as intact even though they were able to identify the word.

To examine school children's perceptual ability in noisy environments, Newman (Newman, 2004) studied 5.5 year-old children's ability to use partial phonetic information to identify familiar words. The results of this study showed that children's perceptual ability, just like adults', was better when speech signals were disrupted by noise as opposed to silence. However, compared to adults, children were overall more affected by signal disruptions, suggesting that young children are more dependent on the speech signal, particularly in noisy environments. These results are in agreement with Walley's (Walley, 1988) who showed that children's phoneme restoration demands more phonetic information than adults' (i.e. children need to listen to longer portions of the disrupted word).

A study by Fernald, Swingley, & Pinto (Fernald et al., 2001) suggests that 18 and 21 months-old children are able to recognize words without access to complete acoustic speech signals. In this study the subjects associated pictures with auditory-stimuli presented both as whole words in intact form and as disrupted words in which only the initial 300 ms of the word was heard. Their results showed that children from both age groups could recognize whole words, as well as disrupted words. There were no differences in the two age-groups reaction times to these two types of stimuli. Instead their study indicated that children's word processing ability was related to their lexical development – the children with greater productive vocabularies were more accurate in their recognition, a conclusion that departs somewhat from earlier findings by Fernald, Pinto, Swingley, Wineberg & McRoberts (Fernald et al., 1998) indicating an age dependence on the ability to identify partial words for infants in the age range 15 to 24 months.

The Nature of Infants' Lexical Representations

There is an ongoing discussion on whether infants' lexical representations are of a holistic or of a more specific nature. As an example, the fact that younger infants were not so accurate in identifying partial words was taken as an indication of lacking segmental structure in their lexical representations of words (Fernald et al., 1998). On the other hand, studies concerning language-specific tuning of vowels have shown that 11-12 months-old infants are sensitive to the detailed sound structure of the ambient language, as opposed to a structure of a non-native language (Kuhl et al., 1992; Polka & Werker, 1994). This indicated, according to the authors, that infants have a rather detailed representation of native segments. However, these studies were not aimed at studying word processing explicitly, so there is a possibility that infants do have detailed lexical representations of words, but they do not process words incrementally, i.e. they do not make use of the acoustic information before the offset of a word to the extent that adults do.

Modular or General Perceptual Restoration

There is evidence that the perceptual restoration phenomenon is not restricted to speech, to the modality of spoken language, or to the human species. Indeed, top-down processing has been demonstrated in musical perception (DeWitt & Samuel, 2006), in the interpretation of American Sign Language (Schultz-Westre, 1985) as well as in the perception of starlings' birdsong (Braaten & Leary, 1999).

Yet, due to psychoacoustic and methodological factors, some potential restrictions to the generality of the perceptual restoration phenomenon should be taken into consideration. As an example, for the perceptual restoration effect to take place, the class of the phoneme being masked and the nature of the masker (e.g. type of noise) as well as the amplitude of the masker have to be sufficiently similar (Newman, 2004). Originally a cough was used as a masker of the phoneme /s/ so that both the masker and the phoneme to be masked contained energy at several frequencies (Warren, 1970). Also the type of laboratory task facing subjects may differ in number of ways from e.g. everyday listening to speech in noisy environments. In the study by Newman (Newman, 2004), adults and school children were to detect interruptions presented at high rate (at slowest 2 per second). The interruptions also occurred at constant rate and were accordingly easily predictable by the subjects. Further, the stimuli in the experiment were presented over headphones and the adults were asked to type into a keyboard what they thought they heard, while the children were to repeat the sentences into a microphone. These recordings were later transcribed by an experimenter. In sum, the cognitive load intrinsic to the procedure of typing or repeating might have been rather demanding for the subjects. On contrary, an analysis of infants' looking behaviour like in the studies by Fernald et al. (Fernald et al., 1998; Fernald et al., 2001) makes use of the infants' spontaneous tendency to look at images of objects while hearing names corresponding to them.

Rationale for the Current Study

Although several studies have shown that adults and children possess the ability to identify words from partial phonetic information, less is known about the corresponding restoration ability in young infants. The aim of the present study was to investigate 11-16 months-old infants' ability to identify spoken words on the basis of their initial sounds. The theoretical motivation for this study is inspired by analogue experiments performed on infant's ability to predict other people's action goals interpreted as a foundation of empathy and social understanding. The current study is further designed to investigate whether infants' ability to predict phonetic events might be related to their age or to their productive vocabulary size. Therefore the youngest infants in the current study were chosen as representatives of subjects essentially lacking productive vocabularies.

As opposed to heavy-cognitive-load procedures, the current eye-tracking method (Tobii, http://www.tobii.com), just like the method used in the studies by Fernald et al. (Fernald et al., 1998; Fernald et al., 2001), takes advantage of the infants' spontaneous tendency to look at images of objects while hearing names corresponding to them.

Method

The infants' eye-movements were recorded as the subjects watched short video sequences displaying images of four familiar objects (a watch, a car, a ball and a teddy-bear). The objects were first introduced one at the time, along with a speaker voice presenting the Swedish names (target-nouns) of the objects embedded in carrier phrases. After these presentations, all the four objects were displayed simultaneously, one object per quadrant, while the speaker asked for *one* of the objects.

Subjects

A total of 17 infants participated in this study. Data from two infants were excluded due to interrupted recoding (failure to complete test session), resulting in 15 subjects (8 girls and 7 boys, age range 11.6-16.2 months, mean age 13.7 months). The infants were divided in two age groups, separated by the global median age – one group ranging from 11.6-13.5 months and the other from 13.5-16.2 months. All the subjects were primarily exposed to Swedish in their families. According to parental reports, none of the subjects had hearing abnormalities as revealed by BOEL-distraction test ("Blicken Orienterar Efter Ljud") routinely used by Swedish Child health centres to screen all infants 7-10 months of age. The subjects' receptive vocabularies were assumed to include the target-nouns used in the tests but nevertheless these target-nouns were explicitly presented in the video materials of this study. The infants were not expected to have these target-nouns in their productive vocabularies, although this may not be the case for some of the older infants.

Speech Materials

The target-nouns were selected from SECDI (Swedish Early Communicative Development Inventories) (Eriksson & Berglund, 1996) language assessment data base, which is a Swedish version of the MacArthur Communicative Development Inventories (CDI) based on parental reports. In order to assess age-appropriate test words, the "words and gestures" version of the database designed for children 8-16 months of age was used. In addition, only disyllabic target-nouns were selected and the words to be used in the test had to be matched regarding their initial phonemes. The target-nouns selected were:

- /klok:an/ (Watch)
- /nal:en/ (Teddy-bear)
- /bi:len/ (Car) (occurred only in manipulated version)
- /bol:en/ (Ball) (occurred only in manipulated version)

The disrupted target-nouns were prepared as follows: the first sound of /bi:len/ was extracted and concatenated with brown noise, the first sound of /bol:en/ was extracted and concatenated with brown noise, the first *two* sounds of /bil:en/ were extracted and concatenated with brown noise and the first *two* sounds of /bol:en/ were extracted and concatenated with brown noise resulting in:

- /b(i)/+NOISE (the /b/ carrying traces of /i/)
- /b(o)/+NOISE (the /b/ carrying traces of /o/)
- /bi/+NOISE
- /bo/+NOISE

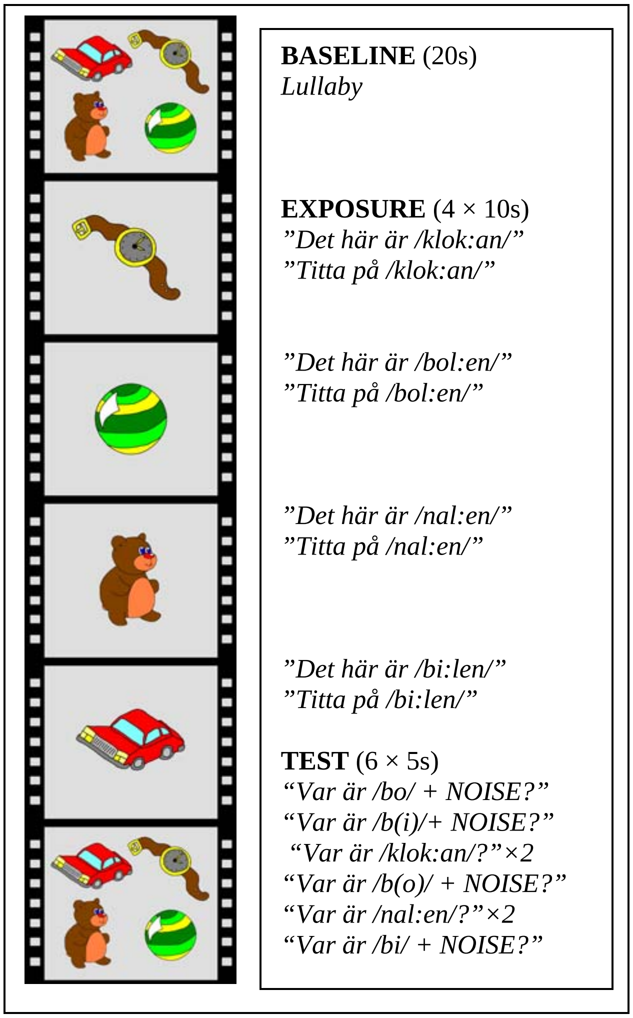

Figure 1: An illustration of the type of familiar objects and audio materials presented in the video materials.

Video Materials

The video materials (1min 30s) consisted of a BASELINE, an EXPOSURE, and a TEST phase, as follows (Figure 1):

- BASELINE (20s): The four objects were displayed to measure infants’ spontaneous preference before EXPOSURE. The audio played a lullaby to catch the infants’ attention towards the screen.
- EXPOSURE (4 × 10s): Each object was presented one at the time. A recorded female voice named the objects in carries phrases, uttered in Infant-Directed Speech style – “*This is a X*” or “*Look at the X*” (where *X* represents the actual target-noun).
- TEST (6 × 5s): Subsequently a split screen of the four objects was presented again. The voice asked for one of the objects – “*Where is the X?”* The target-nouns /klok:an/ (Watch) and /nal:en/ (Teddy-bear) were presented intact and repeated twice (TEST1 & TEST2). The other two target-nouns occurred only in their manipulated versions (TEST /bi/ + NOISE, TEST /b(i)/ + NOISE, TEST /bo/ + NOISE, TEST /b(o)/ + NOISE).

There were four counterbalanced versions of the video materials in which the presentation order of the objects (during EXPOSURE) and questions (during TEST) was systematically varied to control for possible memory effects.

Procedure and Instrumentation

A Tobii 1750 eye-tracker integrated with a 17” TFT monitor was used in the measurements. For each subject the system was calibrated at the beginning of each session. The infant was seated facing the monitor at a distance of approximately 60 cm. The care-giver, listening to masking music through sealed head-phones with active noise reduction, sat by the infant, slightly behind, out of the infant’s visual field.

The equipment uses infrared light to generate gaze measurements sampled at 50 Hz and its average nominal accuracy of gaze estimation is 0.5 degrees. The ClearView 2.2.0 software was used to store gaze data. The data collected was exported from ClearView to Matematica 5.2 and SPSS 14.0 for statistical analysis.

Results

The distribution of the looking times during the BASELINE showed a slight, non-significant, preference for Teddy-bear.

The data was assessed in terms of departures from the BASELINE to TEST expressed in looking times towards each object shown in the split screen. Figures 2-5 show detailed results for each of the target words pooled for all the infants. The results indicate that intact target nouns were recognized in a much more accurate way than the partially masked words. Significant departures/tendencies according to paired samples test (2-tailed) from BASELINE to TEST are shown in Table 1.

Additional phonetic information (i.e. partially masked words represented by their first two sounds as opposed to by one sound only) did not seem to improve the infant’s ability to predict the target word.

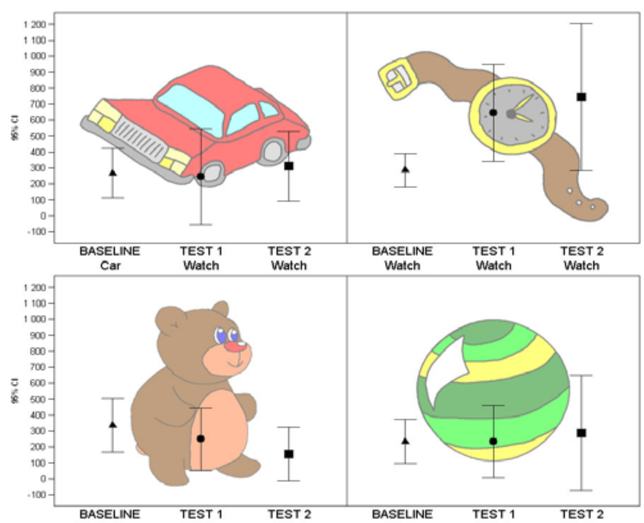

Figure 2: **Target Watch**. CI (95%) from left to right indicate looking time (ms) during BASELINE, TEST 1 (1st rep. of /klok:an/), and TEST 2 (2nd rep. of /klok:an/).

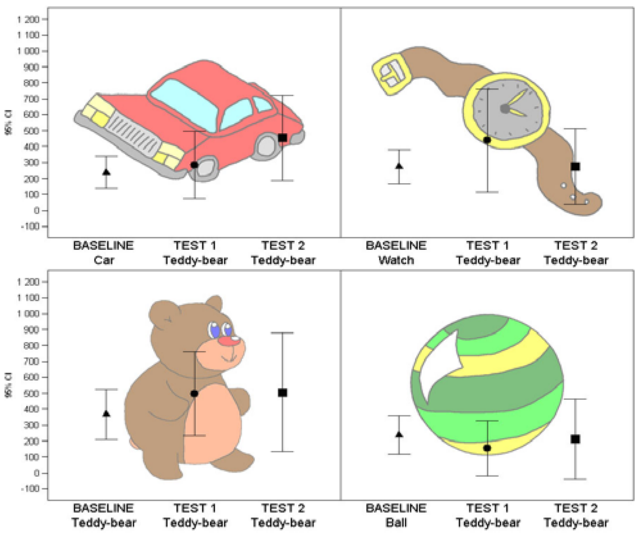

Figure 3: **Target Teddy-bear**. CI (95%) from left to right indicate looking time (ms) during BASELINE, TEST 1 (1st rep. of /nal:en/), and TEST 2 (2nd rep. of /nal:en/).

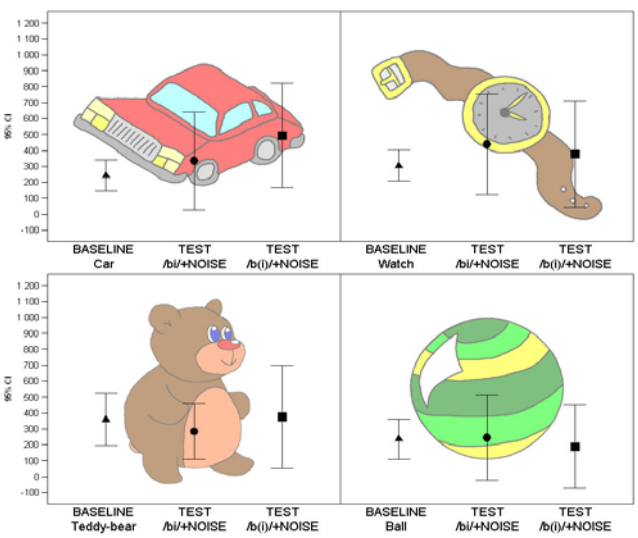

Figure 4: **Target Car**. CI (95%) from left to right indicate looking time (ms) during BASELINE, TEST (/bi/+NOISE), and TEST (/b(i)/+NOISE).

Table 1: Paired samples test (2-tailed). Significant departures/tendencies from BASELINE to TEST for intact and interrupted target nouns are shown in grey.

Intact target nouns		Interrupted target nouns	
TEST 1 /klok:an/ (Watch)	t(13)=2.482, p<0.027	/bi/+NOISE (Car)	t(13)=0.161, p<0.874
TEST 2 /klok:an/ (Watch)	t(12)=2.102, p<0.057	/b(i)/+NOISE (Car)	t(13)=1.814, p<0.093
TEST 1 /nal:en/ (Teddy-bear)	t(14)=2.057, p<0.059	/bo/+NOISE (Ball)	t(13)=-0.011, p<0.991
TEST 2 /nal:en/ (Teddy-bear)	t(12)=0.974, p<0.349	/b(o)/+NOISE (Ball)	t(13)=0.401, p<0.695

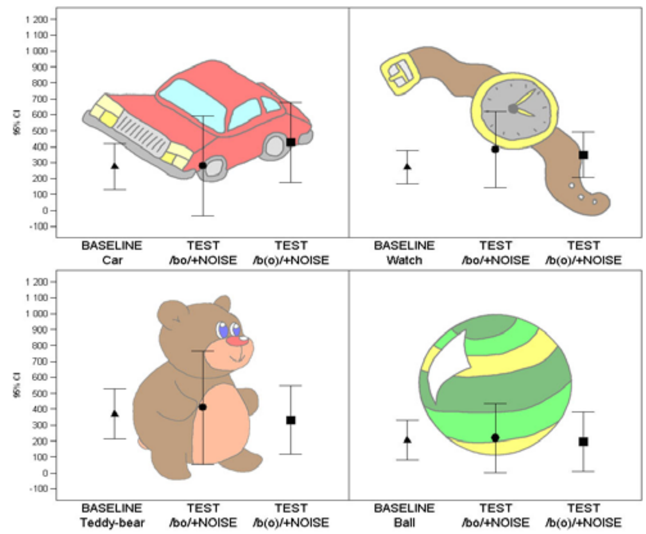

Figure 5: **Target Ball**. CI (95%) from left to right indicate looking time (ms) during BASELINE, TEST (/bo/+NOISE), and TEST (/b(o)/+NOISE).

The two age groups (11.6 to 13.5 and 13.5 to 16.2 months) behaved in about the same way, although there was a slight advantage for the older group in responding to intact words.

Summary and Discussion

Infants seem to possess prediction skills for anticipation of other people's goal-directed behaviour already by the age of 12 months (Falck-Ytter et al., 2006). The mirror-neuron system (MNS) coding for learning to do an action from seeing it done, and functioning as the neurological correlate of these and similar results, is suggested to exist both in monkeys (area F5) (Rizzolatti & Arbib, 1998) and humans (Broca's area) (Fadiga et al., 1995). As an extension, the MNS hypothesis of social cognition (Rizzolatti & Craighero, 2004) proposes this system to constitute a neurological base for coding social understanding and language abilities. Evolutionarily, the coding of hand-actions is progressively assumed to be transferred to mouth-actions in general and further through ingestive behaviour (such as breaking food) to communicative mouth-actions (such as lip smacking) in particular. Accordingly, the original function of *audio-visual* mirror neurons might have transferred to respond to sound only, leading to a separate *auditory* mirror system on the top of the original audio-visual one (Rizzolatti & Craighero, 2004).

Experimental evidence points at commonalities between the development of prediction skills of action goals for motor and speech events. Infants, as young as 18-21 months-old, seem to be able to predict partial acoustic information (Fernald et al., 2001). In addition, these results suggested that recognition of non-complete acoustic speech signals is related to infants' lexical development (productive vocabulary sizes) and/or to their age (Fernald et al., 1998). Further, the "perceptual restoration" phenomena – which comprises that the observer perceives the word/song/signing to continue behind the noise and restore what ever is missing – is suggested to function in a domain- and species-general manner (e.g. to exist also in music, sign-language, and in birdsong).

The current study was set up to investigate young infants' (age range 11-16 months) ability to use partial phonetic information. The choice of age-continuum and the youngest infants' supposedly non-existing productive vocabularies were a result of the endeavour to bring light on the discussion of whether infants' ability to predict acoustic information is correlated to productive vocabulary size and/or age. Caution has been taken concerning potential methodological pitfalls regarding the class of phonemes to be masked and the energy levels of the masker at different frequencies, as well as regarding the type of laboratory task facing the infants. The brown noise used (containing energy at all the frequencies) was simply added – as opposed to superimposed on the phonemes to be masked – to the first phoneme/phonemes after excising the last phonemes of the signal. In this way the masker was prepared to function effectively assuring the perceptual restoration effect to be able to take place. In addition, the interrupted signals were not presented in a predictable or in a cognitively demanding manner. The infants' spontaneous tendency to look at objects while hearing names referring to them was used.

To test the perceptual restoration effect's modularity to speech processing versus it's generality as a part of auditory behaviour, studies on infant's restoration of *non-linguistic stimuli*, as well as studies on *non-human mammals*' restoration of linguistic stimuli, are to be performed in future experiments.

Acknowledgments

Research supported by The Swedish Research Council, The Bank of Sweden Tercentenary Foundation, European Commission, and Birgit and Gad Rausing's Foundation. The authors are grateful to a group of speech-therapist students (Logopedprogrammet 2005) for performing the subject running.

References

Braaten, R. F. & Leary, J. C. (1999). Temporal induction of missing birdsong segments in European starlings. *Psychological Science, 10,* 162-166.

DeWitt, L. A. & Samuel, A. G. (2006). The role of knowledgebased expectations in music perception: Evidence from musical restoration. *Journal of Experimental Psychology, 119,* 123-144.

Eriksson, M. & Berglund, E. (1996). Swedish Early Communicative Development Inventory -- words and gestures. *First Language, 19,* 55-90.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor Facilitation During Action Observation - A Magnetic Stimulation Study. *Journal of Neurophysiology, 73,* 2608-2611.

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neurosci., 9,* 878-879.

Fernald, A., Pinto, J. P., Swingley, D., Wineberg, A., & McRoberts, G. W. (1998). Rapid games in speed of verbal processing by infants in second year. *Psychological Science, 9,* 228-231.

Fernald, A., Swingley, D., & Pinto, J. P. (2001). When half a word is enough: infants can recognize spoken words using partial phonetic information. *Child Dev., 72,* 1003-1015.

Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience, 17,* 1703-1714.

Grafton, S. T., Arbib, M. A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations in humans by positron emission tomography .2. Observation compared with imagination. *Experimental Brain Research, 112,* 103-111.

Keysers, C., Kohler, E., Umilta, M. A., Nanetti, L., Fogassi, L., & Gallese, V. (2003). Audiovisual mirror neurons and action recognition. *Experimental Brain Research, 4,* 628-636.

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science, 297,* 846-848.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606-608.

MacNeilage, P. & Davis, B. L. (2000). Deriving speech from nonspeech: a view from ontogeny. *Phonetica, 57(2-4),* 284-296.

Newman, R. S. (2004). Perceptual restoration in children versus adults. *Applied Psycholinguistics, 5,* 481-493.

Polka, L. & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *J.Exp.Psychol.Hum.Percept.Perform., 20,* 421-435.

Rizzolatti, G. & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences, 21,* 188-194.

Rizzolatti, G. & Craighero, L. (2004). The Mirror-Neuron System. *Annual Review of Neuroscience, 27,* 169-192.

Samuel, A. (1996). Phoneme restoration. *Language and Cognitive Processes, 11,* 647-654.

Schultz-Westre, C. (1985). A visual analogue of phonemic restoration: Sign restoration in American Sign Language. *Milwaukee, WI: University of Wisconsin-Milwaukee.*

Walley, A. C. (1988). Spoken word recognition by young children and adults. *Cognitive Development, 3,* 137-165.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science, 167,* 392-393.

Warren, R. M. & Obusek, C. J. (1971). Speech perception and phonemic restorations. *Perception and Psychophysics, 9,* 358-362.

Warren, R. M. & Warren, R. P. (1970). Auditory Illusions and confusions. *Scientific American, 223,* 30-36.