Parallels in infants' attention to speech articulation and to physical changes in speech-unrelated objects

Eeva Klintfors, Ellen Marklund, Francisco Lacerda

Department of Linguistics, Section for Phonetics, Stockholm University, Sweden

eevak@ling.su.se, ellen@ling.su.se, frasse@ling.su.se

Abstract

The mechanisms of how children develop the capacity to make use of speech articulation cues to support interpretation of the speech signal are not exhaustively explored. The purpose of this study is to investigate if there are parallels in infants' way of attending to speech articulation and their perception of physical changes in speech-unrelated objects. The current research questions grew out from an earlier study in which it was found that perception of speech in infants was based on a match between auditory and visual prominence – as opposed to a match between sound and to it corresponding face. Data suggested that speech perception in infancy may function as described by Stevens power law, and two methodological supplements to test the validness of this hypothesis were made in the current study. First, a non-articulatory test condition was added to investigate infants' perception of speech-unrelated objects. Second, amplitude manipulated stimuli were added to introduce systematic changes in loudness. Results confirmed our hypothesis; the visually prominent articulations were favored, and the same pattern was found in response to non-speech related objects.

Index Terms: infant perception, visual cues, Power Law

1. Background

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. Moreover, it is well established that speech perception in adults is strongly influenced by visual speech cues. In a typical face-to-face conversation adults pay attention to speaker's articulation of speech. However, the mechanisms of how this competence develops in children are not well explored. Studies on infants are important to examine questions that are impossible to answer with adult listeners, such as how perceivers with limited experience of speech interpret visual information that is present simultaneously with the speech signal. That is, the adults' experience of and competence on how speech sounds are articulated is impossible to separate from their perception of face and sound. The research questions in this study are focused exactly around these issues: How is speech perception organized in infancy? What do infants attend to visually while hearing speech? Is speech perception in infancy guided by the same principles as perception of other physical events? Is speech perception in infancy enhanced by visual information of articulation of the speaker?

A number of studies have addressed different aspects of audio-visual speech perception in adult speakers [1-4] as well as its potential ontogenetic origins [5-7]. For instance, Kuhl & Meltzoff [5] showed that 4.5- to 5 month-old infants may pick up the correlation between acoustic and articulatory characteristics of speech sounds. In their experiment infants were exposed to a split-screen displaying two faces articulating /a/ and /i/ respectively, and an audio signal consisting of one of these vowels. Their results showed that infants looked significantly longer at the face matching the vowel sound suggesting that audio-visual integration in speech perception is present at early age. However, a study (here referred to as Study 1) performed at our Phonetic laboratory revealed quite different results [8]. Infants in the age range of 6- to 8-months were exposed to a split-screen displaying four faces articulating /ba/, /by/, /a/, and /y/ respectively, and an audio signal consisting of one of these speech sounds (/by/ or /a/) (Figure 1 to the left). The results showed that infants looked significantly longer at the /ba/-face irrespective of what speech sound was played indicating that the infants looked at the visually *most prominent* stimulus. In addition, organized as a function of visual prominence, a linear trend in looking time towards /ba/, /by/, /a/, and /y/ in descending order was found irrespective of the sound played.

Study 1 diverged from the Kuhl and Meltzoff study in several methodological aspects. First, the infants in our study had a mean age of 7 months (as opposed to 5 months) and their looking at one of the images was reduced to 25% chance level (as opposed to 50%) by presenting four alternatives (as opposed to two) on a single screen. Second, the film started off with a baseline showing a split-screen of four identical still images of an actress's face during which the infant's spontaneous looking towards the quadrants was measured. After the baseline a test-phase, according to the above description, was displayed. Thus, to add validity to the measurements, the infants' looking times were quantified as the net difference (ms) between looking time at the target and non-targets during test-phase and looking time at the target and non-targets during baseline. The infants' eye movements were registered with a modern eye-tracking technology (Tobii) non-existent by the time point of Kuhl and Meltzoff study.

Further relative to the Kuhl and Meltzoff study, two additional test conditions (condition 2 and 3) were created. Test condition 2 (Figure 1 to the right) investigated infants' ability to detect non-speech and visual non-speech information. The non-speech stimulus was the sound of hand clapping and the visual non-speech information was images of the actress clapping hands in four different tempo. Test condition 3 investigated auditory non-speech and articulatory information synchronized in time. The non-speech stimulus was, just like in condition 2, the sound of hand clapping and the visual information consisted of four images of the actress articulating /by/ in four different tempo. Results showed that infants looked significantly longer at the quadrant showing the clapping movements synchronized in time with the clapping sound (condition 2). In the cross-modal test condition (condition 3), infants did not look longer at the quadrant showing the articulation synchronized in time with the clapping sound. Instead a correlation between looking time and frequency of hand clapping tempo was found. The difference in response in these conditions was speculated to depend on infants' already existent experience of the outside world which might have enhanced their capacity to match clapping sound to it corresponding clapping image (condition 2), while it in the cross-modal condition (condition 3) may

have guided their attention to function prominence-based as indexed by most attention towards the fastest clapping image, the next fastest image, etc. in descending frequency order.

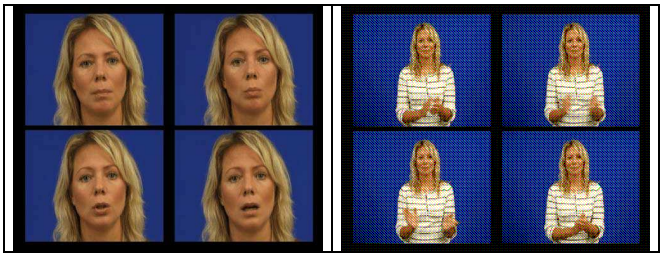

Figure 1: *Left: The actress repeats articulation of /ba/ (upper-left), /by/ (upper-right), /y/ (lower-left) and /a/ (lower-right). The audio repeats the syllable /by/, i.e. the target is shown to upper-right. Right: The actress is clapping hands in 101% (upper-left), 49% (upper-right), 63% (lower-left), and 157% (lower-right) manipulated versions of the original tempo. The audio repeats clapping sound corresponding to 101% of the original, i.e. the target is shown to upper-left.*

1.1. The current study

The results from Study 1 altogether left us with a hypothesis, that infants might be responding on the basis of a general matching of perceptual prominence between visual and acoustic stimuli, as suggested by *Stevens power law* [9-13] rather than on the infants' actual phonetic knowledge of the articulatory and acoustic relationships present in speech. Indeed, because the physical intensity of speech stimuli correlates, in particular vowels, with the degree of jaw opening used to produce those sounds [14, 15], it is possible to establish correct associations between mouth openings and the acoustic intensity of speech stimuli by matching loud acoustic stimuli with visually prominent stimuli, *i.e.* wide mouth openings. The infant's presumed phonetic knowledge could thus be seen as an emergent consequence of general matching processes associating different perceptual continua.

2. Method

To investigate the possibility that speech perception in infancy might be guided by the same principles as perception of other physical events, the infants in the articulatory condition of the current study were exposed to a split-screen displaying the same four articulating faces as in Study 1 (Figure 1 to the left) but this time an audio signal consisting of each of the four sounds /ba/, /by/, /a/ and /y/, as well as a 6 dB attenuated version of the /ba/ and a 6 dB boosted version of the /y/ stimuli. The rationale for including these manipulated sounds was that if the infants' responses indeed are due to matches between the auditory and visual modalities then the intensity manipulations of the acoustic stimuli should induce changes in the looking preferences of the infants. In contrast, if the infants' visual preferences are based on phonetic knowledge (irrespective of whether such knowledge is present at birth or emerges during the first months of life), then their visual preferences should not be affected by the changes in intensity. To further address the issue of the infants' potential phonetic knowledge, an additional non-articulatory condition, in which the articulating faces were replaced by animated circles of different visual prominences was included (Figure 2). In this condition, infants were expected to match the acoustic and the visual stimuli on the basis of their relative prominences.

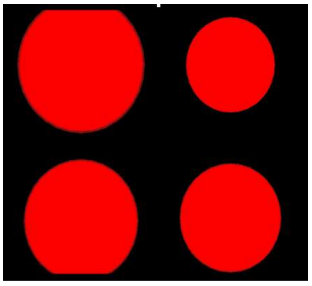

Figure 2: *In the non-articulatory condition the faces were replaced by circles pulsating in synchrony with the acoustic stimuli being played. Each circle started off from a default size and pulsated thereafter between its original size and a pre-established maximum size.*

Thus, two types of structurally identical conditions were created – an articulatory and a non-articulatory condition:

> The film started off with four consecutive BASELINE phases (4 x 10 s) showing the four faces articulating /ba/, /by/, /a/ and /y/ while bird song was played to catch attention towards the screen. The quadrant assignment of each articulation was rotated for every 10 s.
>
> Immediately after the baseline an EXPOSURE phase (20 s) was initiated. During this phase arbitrary squares moved slowly on different trajectories on the screen while the three sounds to be used in the subsequent test phases were repeatedly played in random order. This was done to familiarize the participant with characteristics of the sounds to be used in the test phases.
>
> After exposure three consecutive TEST phases (3 x 20 s) were initiated. In each TEST phase one test sound was presented while the actress' face articulating /ba/, /by/, /a/ and /y/ was displayed in four quadrants. The quadrant assignment of each articulation was changed after 10 s within each block. To assess the infant's sensitivity to the audio signal's intensity, two of the test sounds were manipulated with 6 dB increase (y+) or decrease (ba-) relative to original intensities. The specific test sounds of the four counterbalanced film versions were: /a/ /ba/ /by/ (film A), /a/ /y/ /by/ (film B), /a/ /y+/ /by/ (film C), and /a/ /ba-/ /by/ (film D). During TEST phases the gaze of the participant was expected to be preferentially directed towards the quadrant showing the articulatory gestures matching the sound heard.

In the non-articulatory condition the faces were replaced by circles pulsating in synchrony with the acoustic stimuli being played. Each circle started off from a default size and pulsated thereafter between its original size and a pre-established maximum size. The maximum size of each circle was determined by the articulation shown in the matching speech videos; that is the sizes of the circles ranked from the largest to the smallest (from 1 to 4), the /ba/ articulation was represented by size 1, the /by/ by size 2, the /a/ by size 3 and the /y/ by size 4.

The participants were 30 Swedish infants (16 girls, 14 boys; age range 6 to 8 months; mean age 7 months) randomly selected from the National Swedish address register (SPAR) on the basis of age and geographical criteria. Repeated measures were received from 21 infants, resulting in 51 recording sessions. The participants were first randomly assigned to the articulatory or non-articulatory condition. The 21 participants who were tested twice, were exposed to two

films within the assigned condition. The first film was drawn at random from the four possible films (film A-D) within the condition. The second film was a matched version of the first, in terms of the sound level of auditory stimuli (*i.e.* film A and D, and film B and C were matched versions of each other).

The infant sat on the parent's lap in a dimly lit studio, facing a 17" TFT monitor at approximately 60 cm distance. The parent wore head-phones with active noise reduction and listened to masking music throughout the session to minimize systematic influence on the infant's behavior. The experiment session started with a calibration procedure of the system. The Tobii 1750 eye-tracker system uses a measurement frequency of 50 Hz and has a nominal gaze estimation accuracy of 0.5 degrees. The ClearView 2.2.0 software was used. A program in Mathematica (5.0 and 6.0) was written to structure the raw data files generated by ClearView into relevant quadrants and video stimuli categories before processing. Following this procedure for all the frames in the video film, each participant's cumulative looking times towards the monitor were obtained as a function of the quadrant, the quadrant's visual content and the sound being played. These files were exported to PASW Statistics 18 for statistical analysis.

3. Results

3.1. Analysis of looking behavior during baseline

To investigate whether infants were biased to look at a specific visual stimulus (/ba/, /by/, /a/ or /y/) and/or at a specific quadrant (upper-right, upper-left, lower-left or lower-right) during baseline, the cumulative looking times as a function of the visual stimulus and the quadrant position were measured. Also, the cumulative looking times at the visual stimuli (faces) as a group in the articulatory condition *vs.* the visual stimuli (circles) as a group in the non-articulatory condition during baseline were compared. Analyses of looking behavior during baseline showed that within the articulatory condition, the largest proportion of infants looked at the articulation of /ba/, /by/, /a/ and /y/ in descending order. This is illustrated by the leftmost group of four bars in the upper ("Face") part of figure 3, above the label "Birdsong". In the non-articulatory condition the same looking pattern following objects' degree of prominence was found; proportion of participants looking at circle 1 was largest, and decreased gradually for circles 2, 3, and 4 respectively. This is illustrated by the leftmost group of four bars in the lower ("Circle") part of figure 3, above the label "Birdsong". Multivariate tests indicated significant interaction of quadrants by visual stimulus (/ba/, /by/, /a/, /y/), ($F(3,23)=3.055$, $p<0.049$). For cumulative looking times as function of quadrant position, a significant interaction by position (upper-right, upper-left, lower-left, lower-right), ($F(3,23)=10.297$, $p<0.0005$) was found. Also, a significant interaction of quadrants by test condition (articulatory *vs.* non-articulatory condition), ($F(9,17)=2.702$, $p<0.037$) was found; infants looked more at the visual stimuli as a group in the articulatory condition relative to the group of stimuli in the non-articulatory condition.

3.2 Analysis of looking behavior during test

To investigate potential effects of amplitude manipulations of the stimuli, a comparison of proportion of infants' looking at the different quadrants while listening to /ba/ relative to /ba-/; a version that was amplitude decreased by 6dB, and looking at the different quadrants while listening to /y/ relative to /y+/; a version that was amplitude increased by 6dB, within the articulatory condition, and within the non-articulatory

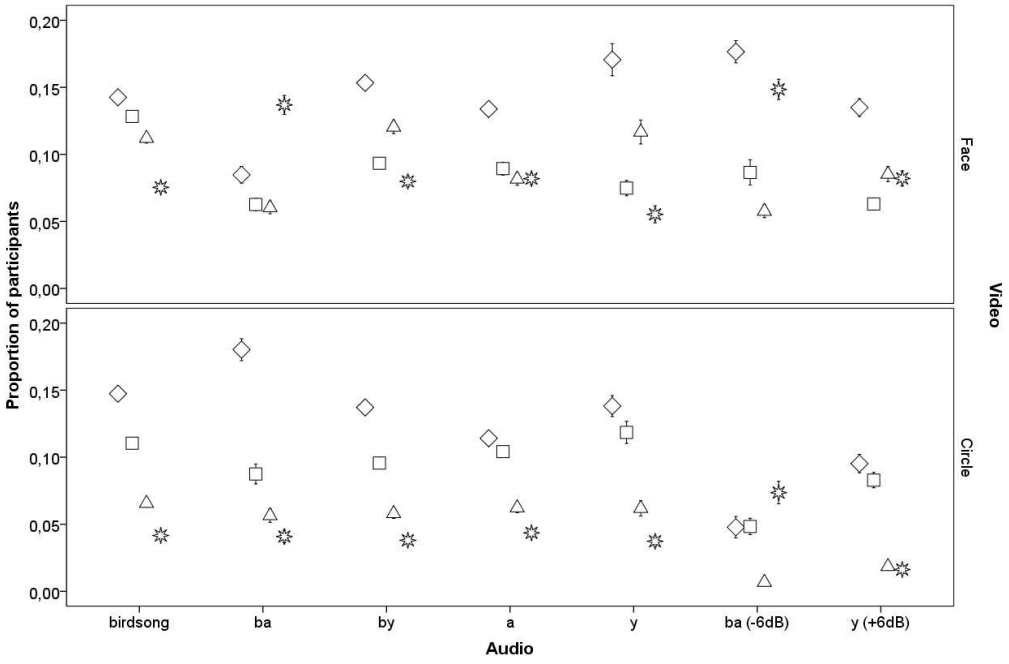

Figure 3: *upper part: Proportion of infants looking at the face articulating /ba/ (diamonds), /by/ (squares), /a/ (triangles), and /y/ (asterisks) during baseline (birdsong), and while the audio repeated one of the test sounds /ba/, /ba-/ (by 6dB attenuated version), /y/, and /y+/ (by 6dB boosted version), or one of the reference sounds /by/ or /a/ in the articulatory condition. Lower part: Corresponding data in the non-articulatory condition in response to circles of varying sizes: size 1 - the largest expanding circle (diamonds), size 2 (squares), size 3 (triangles) and size 4 (asterisks).*

condition was made. Infants' looking behavior while listening to /by/ or /a/, which were included as reference stimuli in each film, were not analyzed in great detail. Thus, the focus of the analysis was to highlight whether infants' looking behavior was guided by face-sound match or by visual prominence, and if similar pattern would hold in the non-articulatory condition.

To begin with the non-articulatory condition, analysis of looking behavior during test revealed that the largest proportion of infants looked at the largest expanding circle (circle 1), followed by 2, 3, and 4 in descending prominence order, irrespective of whether /ba/ or /y/ was played (the lower part of figure 3). Infants looking behavior in the same condition, while listening to the manipulated /ba-/ sound - which amplitude now corresponded to the amplitude of /y/ - revealed that circle 4 now received the largest proportion of infants looking at that quadrant. While the manipulated sound /y+/ - which amplitude now corresponded to the amplitude of /ba/ - was played, the largest proportion of infants looked at one of the two largest expanding circles (1 or 2). Also, most infants looked at circle 1, and gradually fewer at circle 2, 3, 4 respectively while listening to the reference stimuli /by/ or /a/.

Analysis of looking behavior during test within the articulatory condition revealed that most infants looked at the face articulating /y/ while /ba/ was played (the upper part of figure 3). While the amplitude attenuated version /ba-/ was played, the largest proportion of infants looked at one of the faces articulating /ba/ or /y/. Infants looking behavior in the same condition, while the sound /y/ was played, showed that most infants looked at the articulation of /ba/, /by/, /a/ and /y/ in descending order. While the amplitude boosted version /y+/ was played, the same pattern of prominence effect was revealed. Also, most infants looked at the face articulating /ba/, irrespective of which of the reference stimuli /by/ or /a/ was played.

4. Discussion

The results showed that during baseline, as well as during test, the prominence of visual stimuli seemed to be the determinant factor guiding infants' attention to /ba/ (the most prominent bilabial articulation with a large mouth opening), /by/ (bilabial articulation with smaller mouth opening), /a/ (a vowel with a large mouth opening), and /y/ (a rounded vowel with a small mouth opening) in descending order in the articulatory condition. Similarly, the prominence of non-articulatory stimuli (both during baseline and test) seemed to guide infants' attention primarily towards the largest pulsating circle (circle 1), and thereafter towards circle 2, 3, and 4 in descending order. Thus, the results indicate that there are parallels in infants' attention to speech articulation and to physical changes in speech-unrelated objects.

Within the non-articulatory condition the proportion of infants looking at the largest circle was interestingly increased from baseline to the test while /ba/ (the most prominent sound) was played. Also, compared with response to the manipulated version of /ba/ which amplitude now corresponded to the much lower amplitude of /y/, infants' looking preference was changed in favor to circle 4 (the smallest expanding circle). Yet in line with our hypotheses, the greatest proportion of infants looking at circle 1, 2, 3, 4 in descending order while /y/ was played, was not changed in response to the amplitude boosted version of /y/. Thus, since the amplitude of /y+/ corresponded to the much higher amplitude of /ba/, most attention to circle 1 was expected.

In the articulatory condition, we can not find any obvious explanation for the great proportion of infants looking at the face articulating /y/ while listening to /ba/. In the manipulated /ba-/ condition it is adequate that most infants looked at the face articulating /ba/ (*i.e.* corresponding to the actual sound played), and secondly at the face articulating /y/ (*i.e*. the smallest mouth opening corresponding to the amplitude of /ba-/). Also, as expected, the kind of prominence-based looking behavior while listening to /y/ - that is most attention towards /ba/, /by/, /a/, and /y/ in descending order - was also found in response to /y+/ which amplitude now corresponded to the amplitude of /ba/. The manipulation by 6dB corresponds to doubling and dividing the loudness into half of the original stimulus respectively. Manipulation performed in this way seems to have created a detectable change in loudness.

To expand the external validity of the current study, infants response to several types of speech-unrelated objects need to be tested. Also, the current study contained one manipulated version of another stimulus, while confirmation of the relevance of Stevens power law would require use of a continuum of manipulated stimuli ranging from visually and auditorily non-prominent variants to prominent ones. Yet another future study direction is to explore how attention to speech in infancy is manifested neurologically.

The results confirmed the pattern of significant interactions already observed at our laboratory (in Study 1) indicating that visually prominent articulations are favored. Similarly, while exposed to objects that were unrelated to speech, the infants looked primarily at the largest expanding circle. This suggests that early speech perception is guided by factors that in general organize visual attention.

5. References

[1] Green, K. and P. Kuhl, *Integral processing of visual place and auditory voicing information during phonetic perception.* Journal of Experimental Psychology: Human Perception and Performance, 1991. **17**: p. 278-288.

[2] Fowler, C., *Listeners do hear sounds, not tongues.* Journal of Acoustical Society of America, 1996. **99**: p. 1730-1741.

[3] Dekle, D.J., C. Fowler, and M.G. Funnel, *Audiovisual integration in perception of real words.* Perception and Psychophysics, 1992. **51**: p. 355-361.

[4] Remez, R.E., et al., *Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances.* Speech Communication, 1998. **26**: p. 65-73.

[5] Kuhl, P. and A.N. Meltzoff, *The bimodal perception of speech in infancy.* Science, 1982. **218**: p. 1138-1141.

[6] Kuhl, p. and A.N. Meltzoff, *The intermodal representation of speech in infants.* Infant Behavior and Development, 1984. **7**: p. 361-381.

[7] Kuhl, P., K. Williams, and A.N. Meltzoff, *Cross-modal speech perception in adults and infants using nonspeech auditory stimuli.* Journal of Experimental Psychology: Human Perception and Performance, 1991. **17**: p. 829-840.

[8] Lacerda, F., et al. *Emerging linguistics functions in early infancy*. in *The Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. 2005. Nara, Japan: Lund University Cognitive Studies.

[9] Buchsbaum, M. and S.S. Stevens, *Neural events and psychophysical law.* Science, 1971. **172**: p. 502.

[10] Stevens, S.S., *To Honor Fechner and Repeal His Law: A power fucntion, not a log function, describes the operating characteristics of a sensory system.* Science, 1961. **133**: p. 80-86.

[11] Stevens, S.S., *Intensity functions in sensory systems.* International Journal of Neurology, 1967. **6**: p. 202-210.

[12] Stevens, S.S., *Neural events and the psychophysiological law.* Science, 1970. **170**: p. 1043-1049.

[13] Stevens, S.S. and A.L. Diamond, *Effect of glare angle on the brightness function for a small target.* Vision Research, 1965. **5**: p. 649-659.

[14] Fant, G., *Acoustic theory of speech production*. 1960, The Hague, Netherlands: Mouton.

[15] Stevens, S.S., *Acoustic Phonetics*. 1998, Cambridge, Massachusetts: MIT Press