Cross-modal magnitude matching as a precursor of multi-modal speech perception

Ellen Marklund, Anna Ericsson and Francisco Lacerda
Department of Linguistics, Stockholm University

Abstract

Four- to ten-month-old infants (n=58) were examined on their ability to match magnitude across modalities. Their looking behaviour was recorded as they were presented with an intensity modulated auditory stimulus and three possible visual matches. The mean looking times towards a visual target (size envelope matching intensity envelope of the auditory stimulus) and a non-target were calculated. Five-month-olds and seven- to ten-month-olds show a significant preference looking towards the target, as do an adult control group. Four- and six-month-olds do not.

Background

The current paper will briefly cover multimodal perception of speech as well as the power law of phsycophysics. Cross-modal magnitude matching in infancy will be proposed as a result of the latter and a precursor of the former. Lastly, the experiment of the current study will be presented.

Multimodal speech perception

Speech perception is a multimodal phenomenon. It is not the acoustic signal alone that provides the listener with information; if available, the articulatory movements of the speaker's face influence the perception of the speech signal. In noisy environments, listeners are better able to understand what is being said if able to see the speaker (Erber, 1975; Ewertsen and Nielsen, 1971; Sumby and Pollack, 1954) and when presented with conflicting auditory and visual cues, listeners' perception of speech sounds is heavily influenced by what is seen (McGurk and MacDonald, 1976; Traunmüller and Öhrström, 2007). Similarly, perception of the acoustic speech signal can be enhanced or influenced by information presented in other modalities, such as touch (Bernstein and Benoît, 1996).

Multimodal perception of speech is present already in infancy. If presented with conflicting visual and auditory cues, 5-month-olds' perception of syllables is influenced by the visual component much like in adults (Rosenblum, Schmuckler and Johnson, 1997). In addition, several studies have shown that infants are able to identify the articulatory movements related to syllables or speech passages (Barker and Tomblin, 2004; Kuhl and Meltzoff, 1982; Kuhl and Meltzoff 1984; Kuhl and Meltzoff, 1988; MacKain, Studdert-Kennedy, Spieker and Stern, 1983).

In a classic study, Kuhl and Melzoff (1982) demonstrated 4- to 4.5-month-old infants' ability to identify articulatory movements of vowels. The infants heard either /a/ or /i/ repeatedly, while visually presented with recordings of two faces, each articulating one of the syllables. The infants' looking time towards the correct face was significantly above chance, showing that infants at this age are able to connect the articulatory movements with the acoustic signal.

Attempting to replicate and expand Kuhl and Meltzoff's studies with 6- to 8-month old infants, Lacerda and colleagues visually presented four faces articulating syllables, while one syllable was presented auditorily, thereby reducing chance level to 25% instead of 50%. Syllables used in the study were /a/, /y/, /ba/ and /by/, and the gaze-measure used was gain (how much longer subjects looked to a certain area during the test compared to during baseline). The infants, however, did not alter their looking behaviour significantly towards the visually matching target syllable. Instead they looked more to the articulation of /ba/ regardless of what was presented auditorily (Klintfors, 2008; Lacerda, Klintfors, Gustavsson, Marklund and Sundberg, 2005), a behaviour hypothesized to be related to a general cross-modal matching ability compatible with the phsycophysical power law.

The power law of psychophysics

Psychophysics pertains to the relationship between magnitude of physical stimuli and the perceived magnitude. The psychophysical power law as proposed by Stevens states that "equal stimulus ratios produce equal subjective ratios" (1957). This means that if subjects are asked to adjust the level of a stimulus (e.g. the intensity of a tone or the brightness of a light) to half of the original, they will reduce the stimulus magnitude to a certain percentage of the original, regardless of the level on which they start. The relationship between subjective and stimulus magnitude can thus be described as a power function[1] where M_{subj} is the subjective magnitude, M_{stim} is the stimulus magnitude and k is a constant dependent on which units are used.

$$M_{subj} = kM_{stim}^{n}$$

Different sensory impressions have their own characteristic exponent n, derived from experimental data (Stevens, 1966). For instance, the exponent of loudness is 0.3 while for heaviness it is 1.45, resulting in different functions which describe how the perceived magnitude corresponds to the stimulus intensisty and stimulus weight respectively (Stevens, 1957). Since all sensory representations are expected to follow the power law, there is a potential direct proportionality between sensory dimensions. Subjects are indeed able to match sensations in one modality to those in another (Stevens, 1966); if, for instance, subjects are asked to adjust the loudness of a tone to the intensity of vibrations applied to fingertips or to the brightness of a light, they will do so systematically according to the relation between the power functions for each of the two modalities (Stevens, 1962).

Cross-modal magnitude matching and the current study

As reported by Stevens (1962), humans have the ability (and possibly a predisposition) to associate degrees of magnitude across modalities. This could explain the results of Lacerda et al. (2005). Infants looked more towards faces with more noticable articulation, suggesting they did not necessarily connect the syllable they heard with any particular articulation, but instead looked towards the most visually prominent event when presented with any sound (contrasted to the silent baseline), resulting in greater gain for the /ba/-articulation than for the others.

In this light, the general process of cross-modal magnitude matching can be seen as a possible precursor of multimodal perception of speech, serving as a fundament of speech development and first language acquisition. Indeed, while there is no reliable evidence for audio-visual speech perception before the age of 3 months (Burnham, 1998), infants' ability to match intensities across modalities has been demonstrated already at three weeks of age (Lewkowicz and Turkewitz, 1980), suggesting that the cross-modal magnitude matching ability may precede the onset of multimodal perception of speech.

To assess the potential role of general cross-modal magnitude matching in the devopment of multimodal speech perception, it is important to study infants' spontaneous magnitude matching across modalities in non-speech contexts. The present study thus examines 4- to 10-month-olds ability to match intensity modulated white noise to size modulated images of suns.

Method

Participants were 58 infants between 4 and 10 months of age (mean age = 7.3 months), randomly selected from the Swedish national address register, based on date of birth and geographical criteria. A control group of 12 adults also participated in the study.

The experiment consisted of several short film sequences presented in random order, and had a total duration of 140 seconds. The current paper is based on the analysis of one of the two types of film sequences presented in the experiment. The relevant sequence type had a duration of 20 seconds and showed three suns lined up next to each other (see figure 1). The sizes of the suns were modulated while intensity modulated white noise was presented and the size envelope of one of the suns was congruent with the intensity envelope of the noise (the target). There were three occurrences of the relevant film sequence, balanced for target-sun position.

For each version of the film sequence three areas of interest were defined (see figure 2) and the total looking time for each subject within each of those areas were measured. Finally, each subject's average looking times to non-target suns and target suns respectively (regardless of screen position) were calculated.

[1] This description is valid the middle ranges; for magnitude values near the minimum and maximum thesholds, there are known departures from this power law.

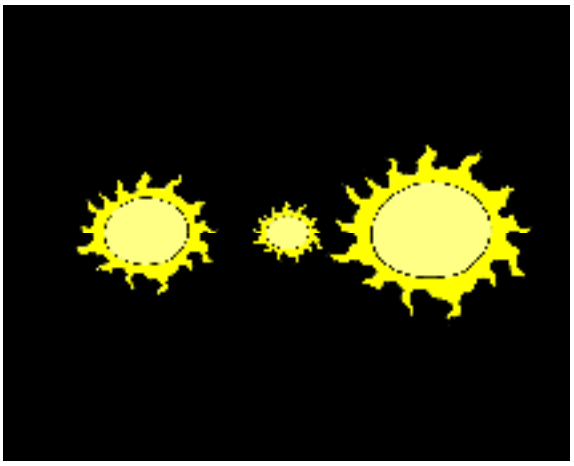

Figure 1. Snapshot of the film sequence. The suns had different size envelopes, one of which matched the noise envelope.

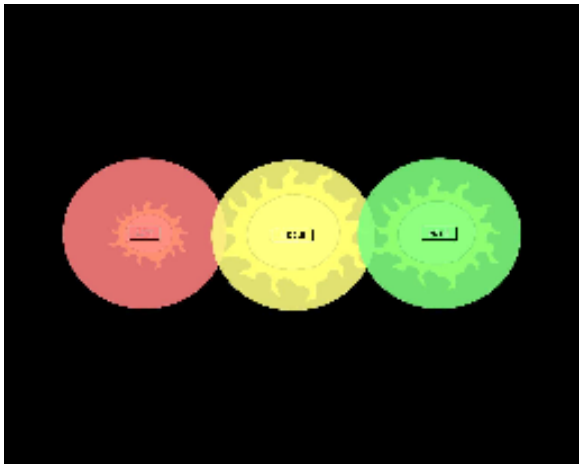

Figure 2. The areas of interest, each covering one sun.

Data recording and pre analysis processing were performed using Tobii T120 and Tobii XL hardware, and Tobii Studio 2.0 software. Data analysis was performed in SPSS 17.0.

Results

The participans were divided into different age groups as shown in table 1.

Table 1. Age groups of the participants.

Age (months)	Number of participants
4	4
5	12
6	5
7	8
8	9
9	9
10	11
Adult (control)	12

The looking behaviour of the infants is shown in figure 3. A significant difference in mean looking time towards target versus non-target was found using a repeated measures ANOVA ($F(1,51)=20.185$, $p<0.0005$). There was no interaction between looking behaviour and age group. When analysed separately, the looking preferences of 4-month-olds and 6-month-olds were not significant.

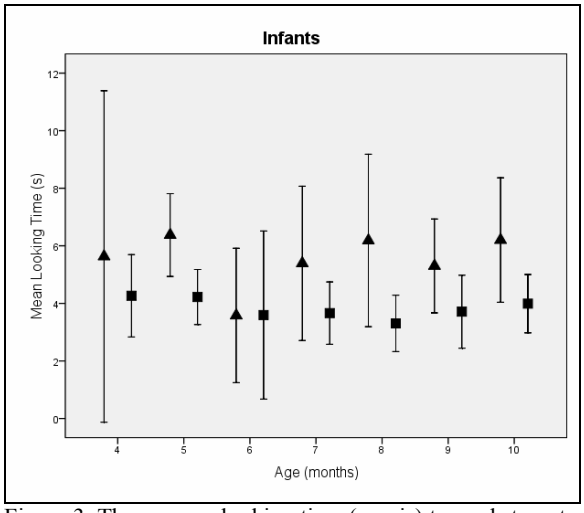

Figure 3. The average looking time (y-axis) towards target (triangles) and non-target (squares) respectively for all age groups (x-axis), with a confidence interval of 95%.

Using a paired samples 2-tailed t-test it was demonstrated that the adults looked significantly longer to the target sun ($t(11)=2.487$, $p<0.03$). For target suns, the mean looking time was 8.3 seconds and for non-target suns it was 6.2 seconds (see figure 4).

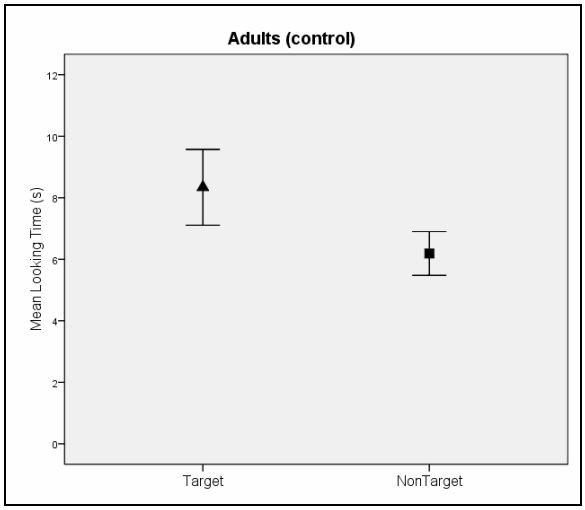

Figure 4. The average looking time (y-axis) towards target and non-target respectively for adults, with a confidence interval of 95%.

Discussion

Adults looks longer to the target sun than they do to the non-target sun and can so be assumed to detect the cross-modal magnitude match.

Infants between the ages of 7 and 10 months show a similar behaviour, as do infants at 5 months of age. The results from both 4- and 6-month-olds are non-significant and might be explained by the low number of subjects in those groups (n=4 and n=5 respectively). Additional data from 4- and 6-month-olds will be collected and added to the present data.

If adding data from more 4- and 6-month-old subjects confirms the present pattern, it may be assumed that the behaviour does not undergo any major changes during the age span covered in the experiment. However, pooling across all age groups suggests a stable looking preference towards the target sun. Expanding the age span covered would also be of interest to determine the youngest age at which cross-modal magnitude matching can be demonstrated.

If, even with added data the target-preference pattern does not emerge, different explanations as to why would be in order for 4- and 6-month-olds. Since 4 months is the youngest age covered in the current study, one possible explanation could be that they are too young for the cross-modal magnitude matching ability to be fully developed. This would, however, contradict the results of Lewkowicz and Turkewitz (1980), where even younger infants were reported to have this ability. One possible (if rather far-fetched) idea as to why 6-month-olds would behave differently than the other age groups is that at around 6 months of age, infants' perception of the world is largely restructured (in terms of caterorization and/or generalization). It is however far more likely that 6-month-olds will display the same pattern as the other age groups.

Once it has been thoroughly established that both multimodal perception of speech and the cross-modal magnitude matching ability is present in infancy, an interesting next step would be to present infants with conflicting cues. Infants' behaviour when presented with e.g. vowels whose intensity has been altered so that a vowel whose intensty is relatively low when occurring naturally is increased (and vice versa), as well as their articulations to match them to, would give answer to which of the two phenomena is more heavily relied upon. If infants in this situation rely more on general magnitude matching, they would prefer the most visually prominent articulation when presented with the loudest syllable. However, if they are more sensitive to the correlation between articulation and vowel quality, they would prefer the face with the correct articulation.

In conclusion, some evidence for cross-modal magnitude matching is found in 5- and 7- to 10-month-old infants and adults. Additional subjects in the groups of 4- and 6-month-olds will probably give more conclusive results for the entire age span investigated. Further experiments with conflicting cues (articulation matching vs. magnitude matching) are suggested.

Acknowledgments

The current study was funded by the Bank of Sweden Tercentenary Foundation as a part of the MILLE-project (K2003-0867).

The authors would like to thank Eeva Klintfors for help with experiment design and data collection and Simon Carlgren, Mathilda Eriksson and Tove Jörgensen for help with data collection. Thanks also to Kelly Smith and Fredrik Myr for proof-reading the paper.

References

Barker BA and Tomblin JB (2004). Bimodal speech perception in infant hearing aid and cochlear implant users. *Archives of Otolaryngology – Head & Neck Surgery,* 130: 582-586.

Bernstein LE and Benoît C (1996). For speech perception by humans or machines, three senses are better than one. In: *Proceedings of ICSLP-96*.

Burnham D (1998). Language specificity in the development of auditory-visual speech perception. In: R Campbell, B Dodd and D Burnham, eds, *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech.* UK: Psychology Press/Erlbaum (U.K.) Taylor & Francis, 27-60.

Erber NP (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, 40: 481-492.

Ewertsen HW and Nielsen HB (1971). A comparative analysis of the audiovisual, auditive and visual perception of speech. *Acta Oto-laryngologica*, 72: 201-205.

Klinfors E (2008). Emergence of words: multisensory precursors of sound-meaning associations in infancy. Doctoral thesis in Phonetics, Stockholm University, Stockholm.

Kuhl PK and Meltzoff AN (1982). The bimodal perception of speech in infancy. *Science*, 218: 1138-1140.

Kuhl PK and Melzoff AN (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, 7: 631-381.

Kuhl PK and Melzoff AN (1988). Speech as an intermodal object of perception. In: A Yonas, ed, *Perceptual development in infancy. The Minnesota symposia on child psychology: Vol. 20.* NJ: Lawrence Erlbaum Associates, Inc., 235-266.

Lacerda F, Klintfors E, Gustavsson L, Marklund E and Sundberg U (2005). Emerging linguistic functions in early infancy. In: L Berthouze, F Kaplan, H Kozima, H Yano, J Konczak, G Metta, J Nadel, G Sandini, G Stojanov and C Balkenius, eds, *Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems.* Sweden: LUCS, 55-70.

Lehiste I and Peterson GE (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, 31: 428-435.

Lewkowicz DJ and Turkewitz G (1980). Cross-modal equivalence in early infancy: auditory-visual intensity matching. *Developmental Psychology*, 16: 597-607.

MacKain K, Studdert-Kennedy M, Spieker S and Stern D (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, 219: 1347-1349.

McGurk H and MacDonald J (1976). Hearing lips and seeing voices. *Nature*, 264: 229-239.

Rosenblum LD, Schmuckler MA and Johnson JA (1997). The McGurk effect in infants. *Perception & Psychophysics*, 59: 347-357.

Stevens SS (1957). On the psychophysical law. *Psychological Review*, 64: 153-181.

Stevens SS (1962). The surprising simplicity of sensory metrics. *American Psychologist*, 17: 29-39.

Stevens SS (1966). A metric for the social consensus. *Science*, 151: 530-541.

Sumby WH and Pollack I (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26: 212-215.

Traunmüller H and Öhrström N (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, 35: 244-258.