Audiovisual integration in binaural, monaural and dichotic listening

Niklas Öhrström, Heidi Arppe, Linnéa Eklund, Sofie Eriksson, Daniel Marcus, Tove Mathiassen and Lina Pettersson
Department of Linguistics, Stockholm University

Abstract
Audiovisual speech perception was investigated in three different conditions: (i) binaurally, where the same sound was presented in both ears, (ii) monaurally, where the sound was presented in one ear randomly, and (iii) dichotically, where the subjects were asked to focus on what was heard in the right ear. The results showed visual influence to be lowered in random monaural presentation as well as in dichotic presentation. Low visual influence to dichotic presentation, as compared with binaural presentation, supports the notion of an attentional component in audiovisual speech processing. Low visual influence in the random monaural presentation may be due to increased attention to the auditory modality because of uncertainty.

Introduction

This paper is concerned with cross modal integration in speech perception and whether or not the visual impact on auditory perception might be related to the amount of accessible attentional resources.

The McGurk effect (McGurk and MacDonald, 1976) shows that optical information about a speaker's speech gesture has an influence upon auditory speech perception, not only at low signal to noise ratios (Sumby and Pollack, 1954; Erber, 1969) but also when the acoustically conveyed speech signal is clear. In the study carried out by McGurk and MacDonald a face articulating /gaga/ presented together with an acoustically presented /baba/, was predominantly perceived as /dada/ by adult listeners. In a later study (Traunmüller and Öhrström, 2007a), incongruent audiovisual front vowels were presented. Perceived vowel openness correlated almost exclusively with vowel openness conveyed through the auditory channel, while perceived lip rounding correlated predominantly with lip rounding conveyed through the visual channel. An auditory (Swedish) /gig/ synchronized with a visual /gøg/ was accordingly perceived as /gyg/. An auditory /gyg/ synchronized with a visual /geg/ was perceived as /gig/.

Ever since the finding of McGurk and MacDonald (1976), the nature of audiovisual integration has been debated. Some theorists have claimed the effect to occur at an early level of speech processing, which means that optical information influences the phonetic percept (e.g. Traunmüller and Öhrström, 2007b, Colin et al., 2002), while others claim it to occur later: E.g. according to the fuzzy logical model of perception (Massaro, 1998), information in each modality is supposed to be processed and evaluated in parallel before integration and decision making (i.e. concept matching) take place.

Automaticity is another issue connected to timing of integration. Intuitively, an early integration approach would leave less space for endogenous attention to have impact on the incoming signal before it is integrated. Signs of automaticity have been shown in many studies (e.g. Green et al., 1991, Rosenblum and Saldaña, 1996, Hietanen et al. 2001). Massaro (1984) also claimed audiovisual integration to be robust to lack of attention. Considering audiovisual perception of emotions, Vroomen et al. (2001) claimed integration to be independent of attention. However, in later studies where distractors were used (Tiippana et al., 2004) and according to the dual task paradigm (Alsius et al., 2005; Alsius et al., 2007), presence of an attentional component has been demonstrated in audiovisual integration.

The present study aims to further investigate the robustness of audiovisual integration in speech perception using a (i) dichotic listening task, in which endogenous attention is kept on what is heard in the right ear, and (ii) a monaural task, in which sound is presented in one ear only, while the listener won't know in which one in advance. In the first task, attentional re-

sources are supposed to be consumed by focusing. The audiovisual integration would be inhibited if dependent of available attentional resources. The second task is concerned with uncertainty, where the listener won't know in which ear the sound will appear next (i.e. possibly attention consuming). Listening to one ear is however equivalent with a decrease in sound intensity of about 3 dB. This could in contrast potentially lead to more auditory confusions and more visual influence.

Method

Participants

In total 30 subjects, 25 female and 5 male, volunteered as perceivers. They were all native speakers of Swedish. They were all right-handed, reported normal hearing and normal or corrected vision. Their mean age was 24.5 years (SD = 5.6 years).

Speech materials

A right ear advantage (REA) test was made to ensure that the listeners' preference would be on the right ear. It was a subset of a test used by Söderlund et al. (2009), originally constructed by Hugdahl (2002). It consisted of the syllables /ba/, /ga/ and /da/ presented in congruent and incongruent dichotic fashion. There were a total of nine REA stimuli, each presented three times in random order.

The stimuli in the following experiments were a further edited subset of the visual, audio and audiovisual stimuli used in Traunmüller & Öhrström (2007a). There were two speakers, one male and one female.

In the first block the visual stimuli showed the speaker while pronouncing the syllables /gig/, /gyg/, /geg/ and /gøg/. Each token was presented twice in random order, thus giving a total of 16 presentations.

Block 2 consisted of auditory and incongruent audiovisual stimuli. A summary of these stimuli is shown in table 1. Each token in the second block was presented twice in random order, thus giving 48 presentations in total.

Block 3 consisted of stimuli corresponding to those in block 2, but presented in one ear at a time. Stimuli were randomized in such a way the listener couldn't predict in which ear the next sound would appear. Each token in block 3 was presented once, giving a total of 48 presentations.

Block 4 consisted of incongruent dichotic auditory and audiovisual stimuli. There were dichotic incongruences concerning vowel openness but not roundedness. The stimuli of block 4 are shown in table 2. Each dichotic token were presented twice, thus giving a total of 48 presentations.

Table 1. Stimuli presented in the second experimental block. A = acoustically presented stimulus, V = optically presented stimulus.

A	V	A	V
/gig/	-	/gyg/	-
/gig/	/gyg/	/gyg/	/gig/
/gig/	/gøg/	/gyg/	/geg/
/geg/	-	/gøg/	-
/geg/	/gyg/	/gøg/	/gig/
/geg/	/gøg/	/gøg/	/geg/

Table 2. Stimuli presented in the forth experimental block. A_{left} = acoustically presented stimulus in the left ear, A_{right} = acoustically presented stimulus in the right ear V = optically presented stimulus.

A_{left}	A_{right}	V	A_{left}	A_{right}	V
/gig/	/geg/	-	/gyg/	/gøg/	-
/gig/	/geg/	/gyg/	/gyg/	/gøg/	/gig/
/gig/	/geg/	/gøg/	/gyg/	/gøg/	/geg/
/geg/	/gig/	-	/gøg/	/gyg/	-
/geg/	/gig/	/gyg/	/gøg/	/gyg/	/gig/
/geg/	/gig/	/gøg/	/gøg/	/gyg/	/geg/

Experimental procedure

Three listeners were participating at a time. They were seated at approximately an arm's length distance from a computer screen and wore somewhat isolating headphones (Deltaco, stereo dynamic, HL-56). They were given instructions in both written and spoken form. The subjects wrote their answers on prepared response sheets in a forced choice design.

In the initial REA-test, the listeners listened to the incongruent dichotic stimuli. The listeners were asked to report what they had heard and choose between <ba>, <da> and <ga>.

The order of the following blocks varied across subjects to avoid context effects. In the experimental blocks, the nine Swedish long vowels appeared as response alternatives.

In block 1 (optic stimuli), the subjects were asked to report what vowel they had seen through speech reading.

In block 2 (binaural stimuli), the subjects were asked to report what vowel they had heard,

while watching the articulating face when shown.

In block 3 (monaural stimuli), the subjects were asked to report what they had heard while watching the articulating face when shown on screen. They weren't aware of in which ear the sound would appear next.

In block 4 (dichotic stimuli), the subjects were asked to report what they had heard in their right ear while watching the articulating face when shown on screen.

Results

According to the initial REA-test, a majority of the subjects responded mostly in accordance with what was presented in the right ear. This tendency was not however overwhelming: on average 53.6% (SD = 9.1%).

In the following, relative visual influence on perceived rounding will be calculated according to equation 1:

Equation 1:

$$\text{Rel.infl.} = (AV_{round} - A_{round}) / (V_{round} - A_{round})$$

AV_{round} = Proportion of audiovisual tokens perceived as a rounded vowel.
A_{round} = Proportion of auditory (only) tokens perceived as a rounded vowel.
V_{round} = Proportion of visual (only) tokens perceived as a rounded vowel.

Example: If an optic /i/, paired with an avoustic /y/ is perceived as rounded to a 60% extent, then $AV_{round} = 0.6$. If the acoustic /y/ in single mode is completely perceived as rounded, then $A_{round} = 1$. If the optic /i/ = is completely perceived as unrounded, then $V_{round} = 0$. The relative visual influence on the perceived rounding would then be 0.6.

Five subjects were excluded in the following analysis because of too small differences, ($|V_{round}-A_{round}| \leq .4$), leading to incomparable results and unreliable measures.

For block 1, the visual responses regarding roundedness are shown in table 3.

For block 2, the responses to auditory and audiovisual binaural stimuli regarding roundedness are shown in table 4. An intended /i/, produced by the female speaker was often categorized as /y/ (42.3%) and even as /ʉ/ in some cases (5.8%). This skewness is also present in block 3 and 4.

For block 3, the responses to monaurally presented stimuli are shown in table 5.

For block 4, the responses to dichotically presented stimuli are shown in table 6. As could be seen in table 2, intended/presented rounding didn't differ across ears.

Table 3. Confusion matrix for visually perceived roundedness (block1). "0"=unrounded, "1"=rounded. Rows: intended, columns: perceived rounding (%).

Stimulus	0	1
0	95.2	4.8
1	2.9	97.1

Table 4. Confusion matrix for perceived roundedness among auditory and audiovisual stimuli, binaurally presented (block2). "0"=unrounded, "1"=rounded responses to visual stimuli. Rows: presented, columns: perceived vowels (%).

Stimulus			
Aud	Vis	0	1
0	*	83.7	16.3
1	*	1.9	98.1
0	1	57.1	42.9
1	0	26.2	73.8

Table 5. Confusion matrix for perceived roundedness among auditory and audiovisual stimuli, monaurally presented (block3). "0"=unrounded, "1"=rounded responses to visual stimuli. Rows: presented, columns: perceived vowels (%).

Stimulus			
Aud	Vis	0	1
0	*	86.5	13.5
1	*	4.3	95.7
0	1	73.6	26.4
1	0	20.7	79.3

Table 6. Confusion matrix for perceived roundedness among auditory and audiovisual stimuli, in dichotic mode (block4). "0"=unrounded, "1"=rounded responses to visual stimuli. Rows: presented, columns: perceived vowels (%).

Stimulus			
Aud	Vis	0	1
0	*	86.5	13.5
1	*	8.7	91.3
0	1	72.8	27.2
1	0	21.4	78.6

The relative visual influence was calculated according to equation 1 for each subject in each condition. The averages across subjects are shown in figure 1. Paired samples t-tests revealed that the visual influence is significantly

lower in the monaural and dichotic condition as compared with the binaural condition: $t(24) = 4.89$, $p < .005$ (2-tailed); $t(24) = 2.71$, $p < .05$ (2-tailed).

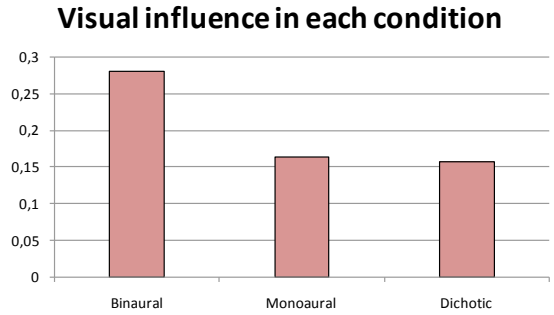

Figure 1. Visual influence on rounding in each condition. Averages across subjects.

Discussion and conclusion

The results of this study are in accordance with earlier studies (Alsius et al., 2005, Alsius et al., 2007, Tiippana et al., 2004), that there is an attentional component involved in audiovisual speech processing. Still, the issue about automaticity in audiovisual speech processing isn't yet clarified: We have shown that integration is inhibited when a competing task consumes attentional resources, but can we disregard from visual information, without looking away, even when attentional resources are available? Just asking the subjects to focus on what is heard vs. seen isn't a satisfactory approach since the two answers will reflect two different percepts, one vocal and one gestural (facial) (Traunmüller & Öhrström, 2007b).

The results in block 3 (monaural stimuli) is of particular interest. Only one ear involved, would intuitively evoke more confusions in auditory mode than binaural stimuli, due to degraded sound input. This would, according to Sumby & Pollack (1954) and Erber (1969), leave more space for visual influence. Instead there were slightly less confusions and visual influence significantly lower than for binaural stimuli. This may be due to the experimental design where, the listeners weren't aware of in which ear the next sound would appear. This uncertainty may cause attention to be drawn to the auditory modality. The visual influence in this study is substantially lower than that obtained in Traunmüller & Öhrström (2007a). This may be due to the experimental design, where visual stimuli were mixed together with auditory and audiovisual stimuli in the same experimental block, forcing the subjects always to attend to the speakers' face.

References

Alsius A, Navarra J, Campbell R, & Soto-Faraco S (2005). Audiovisual integration of speech falters under high attention demands. *Curr Biol*, 15: 839-843.

Alsius A, Navarra J & Soto-Faraco S (2007). Attention to touch weakens audiovisual speech integration. *Exp Brain Res*, 183.3: 399-404.

Colin C, Radeau M, Soquet A, Demolin D, Colin F & Deltenre P (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clin Neurophysiol*, 113.4: 495-506.

Erber NP (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *J Speech Hear Res*, 12: 423-425.

Green KP, Kuhl PK, Meltzoff AN & Stevens EB (1991). Integrating speech information across talker, gender and sensory modality: Female faces and male voices in the McGurk effect. *Percept Psychophys*, 50: 524-536.

Hietanen JK, Manninen P, Sams M & Surakka V (2001). Does audiovisual speech perception use information about facial configuration?. *Eur J Cogn Psychol*, 13.3: 395-407.

Hugdahl K & Davidson RJ (2003). Dichotic listening in the study of auditory laterality. In: Kenneth Hugdahl, eds, *The Asymmetrical Brain*. Cambridge. MA US: MIT Press, 441-475.

Massaro DW (1984). Children's perception of visual and auditory speech. *Child Dev*, 55: 1777-1788.

Massaro DW & Stork DG (1998). Speech recognition and sensory integration. *Am Sci*, 86: 236-244.

McGurk H & MacDonald J (1976). Hearing lips and seeing voices. *Nature*, 264: 746-748.

Rosenblum LD & Saldaña HM (1996). An audiovisual test of kinematic primitives for visual speech perception. *J Exp Psychol Hum Percept Perform*, 22: 318-331.

Sumby WH, Pollack I (1954). Visual contribution to speech intelligibilty in noise. *J Acoust Soc Am*, 26.2: 212-215.

Söderlund G, Marklund E & Lacerda F (2009). Auditory white noise enhances cognitive performance under certain conditions: Examples from visuospatial working memory and dichotic listening tasks. In: *Fonetik 2009*, 160-164.

Tiippana K, Andersen TS & Sams M (2004). Visual attention modulates audiovisual speech perception. *Eur J Cogn Psychol*, 16.3: 457-472.

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

Traunmüller H & Öhrström N (2007b). The auditory and visual percept evoked by the same audiovisual stimuli. In: *AVSP 2007*, L4-1.

Vroomen J. Driver J & de Gelder B (2001). Is cross-modal integration of emotional expressions independent of attentional resources?. *Cogn Affective Behav Neurosci*, 1.4: 382-387.