Audiovisual perception of Swedish vowels with and without conflicting cues

Niklas Öhrström and Hartmut Traunmüller
Department of Linguistics, Stockholm University

Abstract

Auditory, visual and audiovisual syllables with and without conflicting vowel cues (/i y e ø/) presented to men and women showed (1) most to perceive roundedness by eye rather than by ear, (2) a mostly male minority to be less relying on vision, (3) presence of lip rounding to be noticed more easily than absence, and (4) all to perceive openness by ear rather than by eye.

Introduction

It has been known for a long time that lip reading is practiced not only by the deaf, but also by people with normal hearing. Especially in situations with an unfavorable signal to noise ratio, visual cues contribute substantially to speech comprehension (Erber, 1975; Mártony, 1974). Amcoff (1970) investigated lip reading of Swedish vowels and consonants by normal hearing speakers. These investigations showed very good visual perception of labial features.

The topic of audiovisual integration in speech perception became suddenly very popular when McGurk and MacDonald (1976) had published their study in which auditory stimuli had been dubbed on visual stimuli with conflicting cues. They used repeated syllables with stop consonants and a following open vowel. In stimuli with conflicting cues, they observed (1) fusions such as when an auditory [b] presented together with a visual [g] evoked the percept of a [d] and (2) combinations such as when an auditory [g] presented together with a visual [b] evoked the percept of [bg].

While the McGurk-effect appears to work with perceivers with various language backgrounds, some differences have been observed: Japanese listeners show almost no effect of the visual signal when presented stimuli by Japanese speakers. However, they do show an effect when the speaker is a foreigner (Sekiyama and Tohkura, 1993; Hayashi and Sekiyama, 1998).

In addition to cultural differences, there may also be sex differences in audiovisual perception since women have been shown to perform better in lip reading tasks (Johnson et al., 1988). This suggestion was confirmed by Aloufy, Lapidot and Myslobodsky (1996) with speakers of American English, but the sex difference was much smaller among speakers of Hebrew.

Subjects exposed to stimuli with conflicting audiovisual cues typically report "hearing" what they perceive with their eyes. This illusion appears to have a neural foundation: Sams et al. (1991) observed that visual information from lip movements modifies activity in the human auditory cortex.

Audiovisual integration appears to be very robust since it works even when the visual and auditory components are from speakers different in sex (Green et al. 1991).

Although it is not far-fetched to ask oneself whether fusions analogous to those observed by McGurk and MacDonald (1976) would appear in vowel perception, we are not aware of any previous investigation of this kind. However, Summerfield and McGrath (1984) did experiments in which manipulated [bVd] syllables were presented to English subjects together with [i], [ɑ] and [u] faces. They observed that the phoneme boundaries in the auditory vowel space were moved closer towards the position of the vowels presented visually.

It is the purpose of the present study to investigate audiovisual integration in the perception of vowels in a language in which lip rounding is an independent distinctive feature. Since the auditory cues to lip rounding are not very prominent while the visual cues are, it could be expected that perceivers are heavily influenced by vision in perception of this feature. We shall also keep an eye on possible sex differences.

Method

Subjects

The speakers were 2 men (29 and 45 years of age) and 2 women (21 and 29 years), while 10 men (16 to 49 years) and 11 women (18 to 48 years) with normal hearing and vision served as listeners. The speakers were students and researchers from the department of linguistics while the listeners were all phonetically naïve native speakers of Swedish.

Speech material

The materials consisted of four Swedish nonsense syllables /gi:g/, /gy:g/, /ge:g/, /gø:g/. The speakers' faces and the acoustic signal were recorded using a video camera Panasonic NV-DS11 and a microphone AKG CK 93. The recorded material was subsequently edited and dubbed using Premiere 6.0. Each visual stimulus was synchronized with each one of the auditory stimuli. Each visual and each auditory stimulus was also presented alone. In this way, 24 final stimuli were obtained for each speaker. In the perception test these stimuli were each presented twice in random order. Thus, the perception test consisted of 192 stimuli in total. The stimuli were presented in 24 blocks of eight stimuli each, using Windows Media Player.

Procedure

The subjects participated one by one. They carried headphones AKG K135 and were seated with their faces at 60 cm from a computer screen. The height of the faces on screen was roughly 17 cm. The subjects were instructed to visually focus on the speaker's mouth while listening and to write down which vowel they had heard. They used response sheets and were allowed to choose any one of the 9 ordinary Swedish letters that represent vowels. Prior to the experimental blocks, one training block with 8 stimuli was run. The subjects were supervised during the whole session to make sure they were focused on the screen all the time. The entire session lasted about 20 min.

Results

A preliminary analysis of the results suggested that the listeners did not all agree in their behavior. In order to see whether different groups need to be distinguished, stepwise regression analyses were performed for the results of each listener, using the independent factors "auditory openness", "auditory roundedness", "visual openness" and "visual roundedness".

The result showed that "auditory openness" explained most of the variance for all 21 listeners. For the majority (Group 1: 16 listeners), "visual roundedness" explained next to most, i.e., it explained more than "auditory roundedness" did. For a minority (Group 2: 5 listeners), "auditory roundedness" explained more of the variance than "visual roundedness" did.

Table 2. Confusion matrix for Group 1 (10 ♀, 6 ♂). Rows: presented, columns: perceived vowels.

aud	vis	i	y	e	ø	ɛ	ɒ	o
i	*	117	11					
y	*	4	124					
e	*			108	20			
ø	*				122	6		
*	i	74	3	43	7	1		
*	y		77		51			
*	e	21	1	102	1	3		
*	ø		10	1	115		1	1
i	i	127	1					
y	y		128					
e	e			128				
ø	ø				128			
i	y	7	120		1			
y	i	99	28	1				
e	i			127	1			
ø	y				128			
i	e	128						
y	ø		128					
e	ø			19	109			
ø	e			8	59	61		
i	ø	16	108		4			
y	e	107	17	4				
e	y			1	127			
ø	i			6	79	43		

Table 3. Confusion matrix for Group 2 (1♀, 4 ♂).

aud	vis	i	y	e	ø	ɛ	ɒ	o
i	*	37	3					
y	*	2	38					
e	*			40				
ø	*				40			
*	i	14	1	24	1			
*	y	1	25		14			
*	e	6	3	31				
*	ø		10	1	27		2	
i	i	39	1					
y	y		40					
e	e			40				
ø	ø				40			
i	y	31	9					
y	i	8	32					
e	i			40				
ø	y				40			
i	e	39	1					
y	ø		40					
e	ø			31	9			
ø	e				34	6		
i	ø	28	12					
y	e	14	26					
e	y			26	14			
ø	i				39	1		

Subsequently, the two listener groups were kept apart and analyzed separately. The results are shown in Table 2 and 3.

It is of interest to note that Group 2 included 4 of the 10 male listeners (40%) but only 1 of the 11 female listeners (9%).

When the stimuli were presented in (1) auditory mode alone, (2) visual mode alone, and (3) audiovisual mode, the error rates obtained were 8%, 28%, 0.2% for Group 1, and 3%, 39%, 0.6% for Group 2. Cases with conflicting cues are analyzed in Table 4.

Table 4. Response percentages for stimuli with fully conflicting cues (i/ø, y/e, e/y, ø/i). "Fused": always visual roundedness and auditory openness.

	Group 1	*Group 2*
Auditory	22	75
Visual	2	0
Fused	76	25

Within Group 1, the pattern of confusions in roundedness was asymmetric. Table 5 shows that a vowel was rarely identified as unrounded when the lips could be seen as rounded, but when the lips were visibly unrounded, identifications with rounded vowels were not so rare. The table also shows that rounded vowels were more often correctly perceived as rounded.

Table 5. Confusion matrix for roundedness. "0" = "visually unrounded", "1" = visually rounded. Rows: intended, Columns: perceived. Mean number of responses per subject listed.

	Group 1		*Group 2*	
Roundedness	0	1	0	1
0	68	12	38	26
1	3	76	22	42

In addition to the roundedness value (0 or 1), an openness value (0 for [i y], 1 for [e ø o], 2 for [ɛ ɒ]), was assigned to each vowel for numerical evaluation. Based on the numerical values, the mean values of perceived roundedness and openness were computed for each stimulus in each one of the two groups of listeners. In stepwise linear regression analyses, the interaction factors (vis. openness * vis. roundedness and aud. openness * aud. roundedness) were also considered. The result is shown in Table 6. As can be seen, both groups relied on audition in openness perception, with no significant contribution of visual cues. However, in roundedness perception, Group 1 relied on vision, and the contribution of aud. roundedness even failed to attain significance. The other group relied mainly on audition, but did not fully neglect visual cues to roundedness.

Table 6: Result of stepwise regression analyses. Variance explained (r^2) and unnormalized weights of significant independent factors.

	Group 1		*Group 2*	
	round	open	round	open
r^2	.926	.965	.972	.995
Constant	0.18	0.01	0.03	0.00
aud opn	ns	0.99	ns	1.02
aud rnd	ns	ns	0.77	ns
aud opn*rnd	0.27	0.20	ns	ns
vis opn	ns	ns	ns	ns
vis rnd	0.77	ns	0.22	ns
vis opn*rnd	ns	ns	ns	ns

Discussion

As the principal result of the present investigation, we have shown that audiovisual integration of the kind that results in a new, fused percept occurs not only in consonants, but also in vowels. Thus, a visual [y] combined with an auditory [e] was mostly perceived as an [ø], while an auditory [y] combined with a visual [e] was mostly perceived as an [i]. This can be considered as analogous to the presentation of an auditory [b] together with a visual [g] evoking the percept of a [d]. We can see that short segment duration is by no means prerequisite for such fusions to occur.

Instances of double perception analogous to the "combined" results of McGurk and MacDonald (1976) have not been investigated in the present experiment. According to Colin et al. (2002), combinations are less common with voiced consonants than with unvoiced stops while fusions are more common. This would suggest that only fusions should occur with vowels, but in the present experiment some of the listeners informally reported having heard non-standard diphthongs.

Although both lip-rounding and openness are easily visible features, the perceptual weight of the visible cues to openness was found to be insignificant and close to zero among all listeners, while that of visible cues to roundedness was found to be distinctly higher than that of auditory cues in the majority group of listeners. For vowels, it had not been shown before that the perception of some of their features can be dominated by the visual signal. However, this is what had been observed by McGurk and Mac-

Donald (1976) concerning the presence or absence of a labiality feature in stop consonants. The dominant role of the visual input appears to be restricted to labiality features in both consonants and vowels, i.e., to the distinctly most easily visible features.

Since lip rounding is not a distinctive feature within the vowel system of English, such a result could not have been obtained by Summerfield and McGrath (1984), and their data did not suggest any differences of this kind.

The fact that the minority of listeners that did not rely on lip reading was composed predominantly of male listeners is in accord with the reported greater susceptibility of female listeners to visual input and their greater proficiency in lip reading (Johnson et al., 1988), which our results confirm. This has been reasonably explained by sex differences in gazing behavior, women being more likely to look at a speaker's face. In this connection, it is relevant to note that in Japanese culture it is considered more polite not to gaze at a speaker's face, which may explain why both men and women are less sensitive to visual input (Sekiyama and Tohkura, 1993; Hayashi and Sekiyama, 1998).

The insignificant perceptual weight of visual cues to openness may be due to the fact that the openness of the mouth varies with vocal effort and between speakers and also as a function of context, while the visible reflection of lip rounding and labial closure is quite robust to such influences. Moreover, the acoustic cues to labiality are feeble, while the relative position of F_1 is an acoustically strong cue to openness.

It is interesting to note that an auditory [ø] was often perceived as an [ɛ] when combined with a visual [e] or even an [i]. This is reflected in the significant auditory rnd * opn interaction, and it highlights the irrelevance of visual cues for the perception of openness. Only on the basis of acoustics, this behavior can be understood: the frequency position of F_1 in a Swedish [ø] is, typically, higher than in [e], and the F_2 of [ø] is close to that of an [ɛ] (Eklund and Traunmüller, 1997).

The observed asymmetric pattern of confusions in roundedness implies that listeners notice visibly rounded lips as incompatible with the presence of an unrounded vowel, while the incompatibility of unrounded lips with the presence of a rounded vowel goes more often unnoticed. Listeners appear to consider unrounded lips as "unmarked" and therefore as less noteworthy.

References

Aloufi S., Lapidot M., and Mylobodsky M. (1996) Differences in susceptibility to the "blending illusion" among native Hebrew and English speakers. Brain Lang 53, 51–57.

Amcoff S. (1970) Visuell perception av talljud och avläsestöd för hörselskadade. Rapport 7, Lärarhögskolan i Uppsala, Pedagog. inst.

Colin C., Radeau M., Deltenre P., Demolin D., and Soquet A. (2002) The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. Eur J Cogn Psychol 14, 475–491.

Eklund I., and Traunmüller H. (1997) Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. Phonetica 54, 1–21.

Erber N.P. (1975) Auditory-visual perception of speech. J Speech Hear Disord 40, 481–492.

Green K. P., Kuhl P. K., Meltzoff A. N., and Stevens E. B. (1991). Integrating speech informatin across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. Percept Psychophys 50, 524–536.

Hayashi T., and Sekiyama K. (1998) Native-foreign language effect in the McGurk effect: a test with Chinese and Japanese. AVSP'98, Terrigal, Australia. http://www.isca-speech.org/archive/avsp98/

Johnson F.M., Hicks L., Goldberg T., and Mylobodsky, M. (1988) Sex differences in lip-reading. B Psychonomic Soc 26, 106–108.

Mártony J. (1974). On speechreading of Swedish consonants and vowels, STL-QPSR 2-3/1974, 11–33.

McGurk H., and MacDonald J. (1976) Hearing lips and seeing voices. Nature 264, 746–748.

Sams M., Aulanko R., Hämäläinen M., Hari R., Lounasmaa O.V., Lu S-T., and Simola J. (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett 127, 141–145.

Sekiyama K., Tohkura Y. (1993) Inter-language differences in the influence of visual cues in speech perception. J Phonetics 21, 427–444.

Summerfield, A.Q., and McGrath M (1984) Detection and resolution of audio–visual incompatibility in the perception of vowels. Q. J. Exp Psychol -A 36, 51–74.