

Factors affecting visual influence on heard vowel roundedness: Web experiments with Swedes and Turks

Hartmut Traunmüller

Department of Linguistics, University of Stockholm

Abstract

The influence of various general and stimulus-specific factors on the contribution of vision to heard roundedness was investigated by means of web experiments conducted in Swedish. The original utterances consisted of the syllables /gy:g/ and /ge:g/ of a male and a female speaker. They were synchronized with each other in all combinations, resulting in four stimuli that were incongruent in vowel quality, two of them additionally in speaker sex. One of the experiments was also conducted in Turkish, using the same stimuli. The results showed that visible presence of lip rounding has a weaker effect on audition than its absence, except for conditions that evoke increased attention, such as when a foreign language is involved. The results suggest that female listeners are more susceptible to vision under such conditions. There was no significant effect of age and of discomfort felt by being exposed to dubbed speech. A discrepancy in speaker sex did not lead to reduced influence of vision. The results also showed that habituation to dubbed speech has no deteriorating effect on normal auditory-visual integration in the case of roundedness.

Introduction

In auditory speech perception, the perceptual weight of the information conveyed by the visible face of a speaker can be expected to vary with many factors.

- 1) The particular phonetic feature and system
- 2) Language familiarity
- 4) The individual speaker and speech style
- 3) The individual perceiver
- 5) Visibility of the face / audibility of the voice
- 6) The perceiver's knowledge about the stimuli
- 7) Context
- 8) Cultural factors

Most studies within this field have been concerned with the perception of place of articulation in consonants, like McGurk and MacDonald (1976). These studies have shown that the presence/absence of labial closure tends to be perceived by vision. As for vowels, it is known

that under ideal audibility and visibility conditions, roundedness is largely heard by vision, while heard openness (vowel height) is hardly at all influenced by vision (Traunmüller & Öhrström, 2007). These observations make it clear that the presence/absence of features tends to be perceived by vision if their auditory cues are subtle while their visual cues are prominent. Differences between phonetic systems are also relevant. When, e.g. an auditory [g] is presented in synchrony with a visual [b], this is likely to fuse into a [gb] only for perceivers who are competent in a language with a [gb]. Others are more likely to perceive a [g] or a consonant cluster. The observed lower visual influence in speakers of Japanese as compared with English (Sekiyama and Burnham, 2008) represents a more subtle case, whose cause may lie outside the phonetic system.

The influence of vision is increased when the perceived speech sounds foreign (Sekiyama and Tohkura, 1993; Hayashi and Sekiyama, 1998; Chen and Hazan, 2007). This is referred to as the “foreign-language effect”.

The influence of vision varies substantially between speakers and speaking styles (Munhall et al., 1996; Traunmüller and Öhrström, 2007).

The influence of vision also varies greatly between perceivers. There is variation with age. Pre-school children are less sensitive (Sekiyama and Burnham, 2004) although even pre-linguistic children show influence of vision (Burnham and Dodd, 2004). There is also a subtle sex difference: Women tend to be more susceptible to vision (Irwin et al., 2006; Traunmüller and Öhrström, 2007).

The influence of vision increases with decreasing audibility of the voice, e.g. due to noise, and decreases with decreasing visibility of the face, but only very little with increasing distance up to 10 m (Jordan, 2000).

Auditory-visual integration works even when there is a discrepancy in sex between a voice and a synchronized face (Green et al., 1991) and it is also robust with respect to what the perceiver is told about the stimuli. A minor effect on vowel perception has, nevertheless, been observed when subjects were told the sex

represented by an androgynous voice (Johnson, Strand and d'Imperio, 1991).

Auditory-visual integration is robust to semantic factors (Sams et al, 1998) but it is affected by context, e.g. the vocalic context of consonants (Shigeno, 2002). It can also be affected by the experimental method (e.g., blocked vs. random stimulus presentation).

It has been suggested that cultural conventions, such as socially prescribed gaze avoidance, may affect the influence of vision (Sekiyama and Tohkura, 1993; Sekiyama, 1997).

Exposure to dubbed films is another cultural factor that can be suspected to affect the influence of vision. The dubbing of foreign movies is a widespread practice that often affects nearly all speakers of certain languages. Since in dubbed speech, the sound is largely incongruent with the image, habituation requires learning to disrupt the normal process of auditory-visual integration. Considering also that persons who are not habituated often complain about discomfort and mental pain when occasionally exposed to dubbed speech, it deserves to be investigated whether the practice of dubbing deteriorates auditory-visual integration more permanently in the exposed populations.

The present series of web experiments had the primary aim of investigating (1) the effects of the perceiver's knowledge about the stimuli and (2) those of a discrepancy between face and voice (male/female) on the heard presence or absence of liprounding in front vowels.

Additional factors considered, without being experimentally balanced, were (3) sex and (4) age of the perceiver, (5) discomfort felt from dubbed speech, (6) noticed/unnoticed phonetic incongruence and (7) listening via loudspeaker or headphones.

The experiments were conducted in Swedish, but one experiment was also conducted in Turkish. The language factor that may disclose itself in this way has to be interpreted with caution, since (8) the "foreign-language effect" remains confounded with (9) effects due to the difference between the phonetic systems.

Most Turks are habituated to dubbed speech, since dubbing foreign movies into Turkish is fairly common. Some are not habituated, since such dubbing is not pervasive. This allows investigating (10) the effect of habituation to dubbed speech. Since dubbing into Swedish is only rarely practiced - with performances intended for children - adult Swedes are rarely habituated to dubbed speech.

Method

Speakers

The speakers were two native Swedes, a male doctoral student, 29 years (index ♂), and a female student, 21 years (index ♀). These were two of the four speakers who served for the experiments reported in Traunmüller and Öhrström (2007). For the present experiment, a selection of audiovisual stimuli from this experiment was reused.

Speech material

The original utterances consisted of the Swedish nonsense syllables /gy:g/ and /ge:g/. Each auditory /gy:g/ was synchronized with each visual /ge:g/ and vice-versa. This resulted in 2 times 4 stimuli that were incongruent in vowel quality, half of them being, in addition, incongruent in speaker (male vs. female).

Experiments

Four experiments were conducted with instructions in Swedish. The last one of these was also translated and conducted in Turkish, using the same stimuli. The number of stimuli was limited to 5 or 6 in order to facilitate the recruitment of subjects.

Experiment 1

Sequence of stimuli (in each case first vowel by voice, second vowel by face):

e♂e♂, e♀y♂ x, y♂e♀ x, e♂y♂ n, y♀e♀ n

For each of the five stimuli, the subjects were asked for the vowel quality they heard.

"x" indicates that the subjects were also asked for the sex of the speaker.

"n" indicates that the subjects were also asked whether the stimulus was natural or dubbed.

Experiment 2

In this experiment, there were two congruent stimuli in the beginning. After these, the subjects were informed that they would next be exposed to two stimuli obtained by cross-dubbing these. The incongruent stimuli and their order of presentation were the same as in Exp. 1. Sequence of stimuli:

e♀e♀, y♂y♂, e♀y♂, y♂e♀, e♂y♂ n, y♀e♀ n

Experiment 3

This experiment differed from Exp. 1 in an inverted choice of speakers. Sequence of stimuli:

e_♀e_♀, e_♂y_♀ x, y_♀e_♂ x, e_♀y_♀ n, y_♂e_♂ n

Experiment 4

This experiment differed from Exp. 1 only in the order of stimulus presentation. It was conducted not only in Swedish but also in Turkish. Sequence of stimuli:

e_♂e_♂, y_♂e_♀ x, e_♀y_♂ x, y_♀e_♀ n, e_♂y_♂ n

Subjects

For the experiments with instructions in Swedish, most subjects were recruited via web fora:

[Forumet.nu > Allmänt forum](#),
[Flashback Forum > Kultur > Språk](#),
[Forum för vetenskap och folkbildning](#),
[KP-webben > Allmänt prat](#) (Exp. 1).

Since young adult males dominate on these fora, except the last one, where girls aged 10-14 years dominate, some additional adult female subjects were recruited by distribution of slips in cafeterias at the university and in a concert hall. Most of the subjects of Exp. 2 were recruited by invitation via e-mail. This was also the only method used for the experiment with instructions in Turkish. This method resulted in a more balanced representation of the sexes, as can be seen in Figure 1.

Procedure

The instructions and the stimuli were presented in a window 730 x 730 px in size if not changed by the subject. There were eight or nine displays of this kind, each with a heading 'Do you also hear with your eyes?'. The whole session could be run through in less than 3 minutes if there were no cases of hesitation.

The questions asked concerned the following:

First general question, multiple response:

- Swedish (Turkish) first language
- Swedish (Turkish) best known language
- Swedish (Turkish) most heard language

Further general questions, alternative response:

- Video ok | not so in the beginning | *not ok*
- Listening by headphones | by loudspeaker
- Heard well | *not so* | undecided.
- Used to dubbed speech | not so | undecided.
- Discomfort (*obehag, rahatsızlık*) from dubbed speech | not so | undecided.
- Male | Female
- Age in years
- Answers trustable | *not so*.

If one of the negations shown here in italics was chosen, the results were not evaluated. Excluded were also cases in which more than one vowel failed to be responded to.

The faces were shown on a small video screen, width 320 px, height 285 px. The height of the faces on screen was roughly 55 mm (♂) and 50 mm (♀). The subjects were asked to look at the speaker and to tell what they 'heard' 'in the middle (of the syllable)'. Each stimulus was presented twice, but repetition was possible.

Stimulus-specific questions:

Vowel quality (Swedes):

- *i* / *e* / *y* / *ö* / undecided (natural stimuli)
- *y* / *i* / *yi* / undecided (aud. [y], vis. [e])
- *e* / *ö* / *eö* / undecided (aud. [e], vis. [y])

Swedish non-IPA letter: *ö* [ø].

Vowel quality (Turks):

- *i* / *e* / *ü* / *ö* / undecided (natural stimuli)
- *ü* / *üy* / *i* / *ı* / undecided (aud. [y], vis. [e])
- *e* / *eö* / *ö* / undecided (aud. [e], vis. [y])

Turkish non-IPA: *ü* [y], *ö* [ø], *ı* [u] and *y* [j].

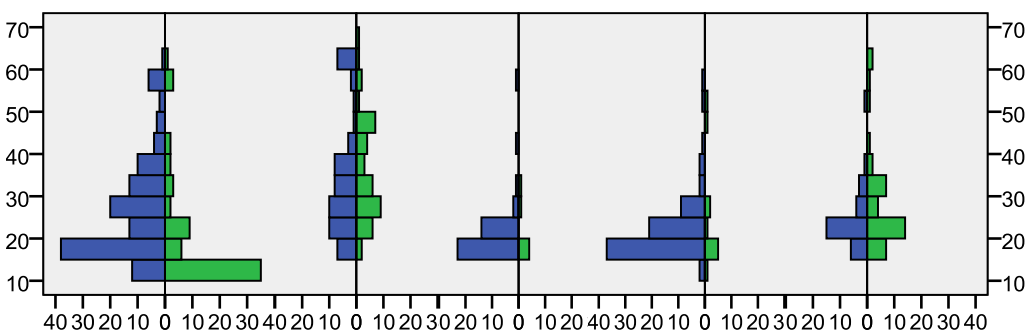


Fig. 1. Population pyramids for (from left to right) Exp. 1, 2, 3, 4 (Swedish version) and 4 (Turkish version). Evaluated subjects only. Males left, females right.

- Female sounding male | male looking female | undecided (when voice♂ & face♀)
- Male sounding female | female looking male | undecided (when voice♀ & face♂)
- Natural | dubbed | undecided (when speaker congruent, vowel incongruent)

Upon completing the responses, these were transmitted by e-mail to the experimenter, together with possible comments by the subject, who was invited to an explanatory demonstration

(<http://legolas.ling.su.se/staff/hartmut/webexperiment/xmpl.se.htm>, ...tk.htm).

Subjects participating via Swedish web fora were informed within 15 minutes or so about how many times they had heard by eye.

Results

The most essential stimulus specific results are summarized in Table 1 for Exp. 1, 2 and 4 and in Table 2 for Exp. 3. Subjects who had not indicated the relevant language as their first or their best known language have been excluded.

It can be seen in Table 1 and 2 that for each auditory-visual stimulus combination, there were only minor differences in the results between Exp. 1, 2, and the Swedish version of Exp. 4. For combinations of auditory [e] and visual [y], the influence of vision was, however, clearly smaller than that observed within the frame of the previous experiment (Traunmüller and Öhrström, 2007), while it was clearly greater in the Turkish version of Exp. 4. Absence of lip rounding in the visible face had generally a stronger effect than visible presence of lip rounding, in particular among Swedes. In the Turkish version, there was a greater influ-

ence of vision also for the combination of auditory [y] and visual [e], which was predominantly perceived as an [i] and only seldom as [yj] among both Turks and Swedes. The response [u] (Turkish only) was also rare. The proportion of visually influenced responses substantially exceeded the proportion of stimuli perceived as undubbed, especially so among Turks.

The results from Exp. 3, in which the speakers had been switched (Table 2), showed the same trend that can be seen in Exp. 1, although visible presence of roundedness had a prominent effect with the female speaker within the frame of the previous experiment.

A subject-specific measure of the overall influence of vision was obtained by counting the responses in which there was any influence of vision and dividing by four (the number of incongruent stimuli presented).

A preliminary analysis of the results from Exp. 1 to 3 did not reveal any substantial effects of habituation to dubbing, discomfort from dubbing, sex or age.

Table 2. Summary of stimulus specific results for Exp. 3 arranged as in Table 1.

Stimulus	Prev. exp.		Exp 3	
	n=42, 20 ♂, 22 ♀	O	n=47 41 ♂, 6 ♀	Nat
y♂ & e♂	79	4	81	66
e♀ & y♀	86	3	34	11
y♀ & e♂	-	2	85	
e♂ & y♀	-	1	28	

Table 1. Summary of stimulus specific results from Exp. 1, 2, and 4: Percentage of cases showing influence of vision on heard roundedness in syllable nucleus (monophthong or diphthong). No influence assumed when response was 'undecided'. O: Order of presentation. Nat: Percentage of stimuli perceived as natural (undubbed) shown for stimuli without incongruence in sex. Corresponding results from previous experiment (Traunmüller and Öhrström, 2007) shown in leftmost column of figures.

Stimulus	Prev. exp. n=42, 20 ♂, 22 ♀	Exp. 1		Exp. 2		Exp. 4		Exp. 4			
		n=185 122 ♂, 63 ♀	Nat	Informed n=99, 57 ♂, 42 ♀	Nat	Sweds n=84, 73 ♂, 11 ♀	Nat	Turks, n=71, 30 ♂, 41 ♀	Nat		
y♀ & e♀	83	4	82	72	81	66	3	81	61	99	64
e♂ & y♂	50	3	26	15	25	11	4	23	9	79	23
y♂ & e♀	-	2	80		65		1	75		94	
e♀ & y♂	-	1	41		42		2	52		83	

For Exp. 4, the result of Chi-square tests of the effects of subject-specific variables on the influence of vision are listed in Table 3.

Table 3. Effects of general variables (use of headphones, habituated to dubbing, discomfort from dubbing, sex and age) on "influence of vision".

	n	Swedes	n	Turks
Phone use	25 of 83	0.02	12 of 68	0.11
Habituated	7 of 81	0.7	53 of 68	0.054
Discomfort	50 of 76	0.4	33 of 63	0.93
Female	11 of 84	0.9	39 of 69	0.045
Age	84	0.24	69	

Use of headphones had the effect of reducing the influence of vision among both Swedes (significantly) and Turks (not significantly).

Habituation to dubbed speech had no noticeable effect among the few Swedes (7 of 81, 9% habituated), but it increased(!) the influence of vision to an almost significant extent among Turks (53 of 68, 78% habituated).

Discomfort felt from dubbed speech had no significant effect on the influence of vision. Such discomfort was reported by 66% of the Swedes and also by 52% of the Turks.

Among Turks, females were significantly more susceptible to vision than males, while there was no noticeable sex difference among Swedes.

Discussion

The present series of web experiments disclosed an asymmetry in the influence of vision on the auditory perception of roundedness: Absence of lip rounding in the visible face had a stronger effect on audition than visible presence of lip rounding. This is probably due to the fact that lip rounding (protrusion) is equally absent throughout the whole visible stimulus when there is no rounded vowel. When there is a rounded vowel, there is a dynamic rounding gesture, which is most clearly present only in the middle of the stimulus. Allowing for some asynchrony, such a visible gesture is also compatible with the presence of a diphthong such as [eø], which was the most common response given by Turks to auditory [e] dubbed on visual [y]. The reason for the absence of this asymmetry in the previous experiment (Traunmüller and Öhrström, 2007). can be seen in the higher demand of visual attention. In this previous experiment, the subjects had to identify randomized stimuli, some of which were presented

only visually. This is likely to have increased the influence of visual presence of roundedness.

The present experiments had the aim of investigating the effects of

- 1) a male/female face/voice incongruence,
- 2) the perceiver's knowledge about the stimuli,
- 3) sex of perceiver,
- 4) age of the perceiver,
- 5) discomfort from dubbed speech,
- 6) noticed/unnoticed incongruence,
- 7) listening via loudspeaker or headphones,
- 8) language and foreignness,
- 9) habituation to dubbed speech.

1) The observation that a drastic incongruence between face and voice did not cause a significant reduction of the influence of vision agrees with previous findings (Green et al., 1991). It confirms that auditory-visual integration occurs after extraction of the linguistically informative quality in each modality, i.e. after demodulation of voice and face (Traunmüller and Öhrström, 2007b).

2) Since the verbal information about the dubbing of the stimuli was given in cases in which the stimuli were anyway likely to be perceived as dubbed, the negative results obtained here are still compatible with the presence of a small effect of cognitive factors on perception, such as observed by Johnson et al. (1999).

3) The results obtained with Turks confirm that women are more susceptible to visual information. However, the results also suggest that this difference is likely to show itself only when there is an increased demand of attention. This was the case in the experiments by Traunmüller and Öhrström (2007) and it is also the case when listening to a foreign language, an unfamiliar dialect or a foreigner's speech, which holds for the Turkish version of Exp. 4. The present results do not suggest that a sex difference will emerge equally clearly when Turks listen to Turkish samples of speech.

4) The absence of a consistent age effect within the range of 10-65 years agrees with previous investigations.

5) The absence of an effect of discomfort from dubbed speech on influence of vision suggests that the experience of discomfort arises as an after-effect of speech perception.

6) Among subjects who indicated a stimulus as dubbed, the influence of vision was reduced. This was expected, but it is in contrast with the fact that there was no significant reduction when there was an obvious discrepancy in sex. It appears that only the discrepancy in

the linguistically informative quality is relevant here, and that an additional discrepancy between voice and face can even make it more difficult to notice the relevant discrepancy. This appears to have happened with the auditory $e_{\text{♀}}$ dubbed on the visual $y_{\text{♂}}$ (see Table 1).

7) The difference between subjects listening via headphones and those listening via loudspeaker is easily understood: The audibility of the voice is likely to be increased when using headphones, mainly because of a better signal-to-noise ratio.

8) The greater influence of vision among Turks as compared with Swedes most likely reflects a “foreign language effect” (Hayashi and Sekiyama, 1998; Chen and Hazan, 2007). To Turkish listeners, syllables such as /gy:g/ and /gi:g/ sound as foreign since long vowels occur only in open syllables and a final /g/ never in Turkish. Minor differences in vowel quality are also involved. A greater influence of vision might perhaps also result from a higher functional load of the roundedness distinction, but this load is not likely to be higher in Turkish than in Swedish.

9) The results show that habituation to dubbed speech has no deteriorating effect on normal auditory-visual integration in the case of roundedness. The counter-intuitive result showing Turkish habituated subjects to be influenced *more often* by vision remains to be explained. It should be taken with caution since only 15 of the 68 Turkish subjects were not habituated to dubbing.

Acknowledgements

I am grateful to Mehmet Aktürk (Centre for Research on Bilingualism at Stockholm University) for the translation and the recruitment of Turkish subjects. His service was financed within the frame of the EU-project CONTACT (NEST, proj. 50101).

References

- Burnham, D., and Dodd, B. (2004) Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology* 45, 204–220.
- Chen, Y., and Hazan, V. (2007) Language effects on the degree of visual influence in audiovisual speech perception. *Proc. of the 16th International Congress of Phonetic Sciences*, 2177–2180.
- Green K. P., Kuhl P. K., Meltzoff A. N., and Stevens E. B. (1991) Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception and 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/>
- Irwin, J. R., Whalen, D. H., and Fowler, C. A. (2006) A sex difference in visual influence on heard speech. *Perception and Psychophysics*, 68, 582–592.
- Johnson K., Strand A.E., and D’Imperio, M. (1999) Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27, 359–384.
- Jordan, T.R. (2000) Effects of distance on visual and audiovisual speech recognition. *Language and Speech* 43, 107–124.
- McGurk H., and MacDonald J. (1976) Hearing lips and seeing voices. *Nature* 264, 746–748.
- Munhall, K.G., Gribble, P., Sacco, L., Ward, M. (1996) Temporal constraints on the McGurk effect. *Perception and Psychophysics* 58, 351–362.
- Sams, M., Manninen, P., Surakka, V., Helin, P. and Kättö, R. (1998) McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication* 26, 75–87.
- Shigeno, S. (2002) Influence of vowel context on the audio-visual perception of voiced stop consonants. *Japanese Psychological Research* 42, 155–167.
- Sekiyama K., Tohkura Y. (1993) Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics* 21, 427–444.
- Sekiyama, K., and Burnham, D. (2004) Issues in the development of auditory-visual speech perception: adults, infants, and children, In *INTERSPEECH-2004*, 1137–1140.
- Traunmüller H., and Öhrström, N. (2007) Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics* 35, 244–258.
- Traunmüller H., and Öhrström, N. (2007b) The auditory and the visual percept evoked by the same audiovisual stimuli. In *AVSP-2007*, paper L4-1.