A preliminary study of audiovisual integration of roundedness in front vowels: limitation due to discrepancy in jaw depression

Niklas Öhrström
Department of Linguistics, Stockholm University

Abstract

Audiovisual integration of vowel roundedness was investigated, as the discrepancy in jaw depression increased. The results show that the relative visual impact on perceived roundedness decreases at larger discrepancies. The results may suggest that this tendency would be stronger among acoustically presented [i] than [y]. To verify this, more research with talking heads may be required. The results confirm earlier findings that audiovisual integration doesn't require unconsciousness among subjects about the dubbing procedure.

Introduction

Information about events and object can be mediated through various channels. An approaching car conveys optic information of an object whose imprint on the retina is magnifying. The simultaneous acoustic information is characterized by magnifying amplitude of motor noise as well as relatively high frequency due to Doppler shift. When estimating the position of parts of one's body, one can rely on interacting visual information (if visible), and proprioceptive information.

Perceiving our surroundings involves taking into account information from these various channels and integrate them to match a single concept according to fig. 1. This is highlighted in speech perception, since visual information about the speaker's speech gestures enhances the intelligibility at low S/N ratios (Sumby and Pollack, 1954; Erber, 1969). The integration can take place at an early level of processing, which means that integration precedes the percept, or later, which means that integration merely is part of the categorization into concepts. Fig. 1 does not take any position on the controversy of late or early integration.

Integration of information, as mentioned above, does not require information, mediated through the various channels, to originate from the same source. The ventriloquist effect is one example, where the source of the sound appears to be the same as mediated through the visual signal, although they are different. However, the effect can be reduced if sound and vision are asynchronous, especially in cases where sound precedes the vision (Slutsky and Recanzone, 2001). One further example of illusions is when proprioception is influenced by a discrepant visual signal. This holds as long as the discrepancy is not too large (Warren and Cleaves, 1971).

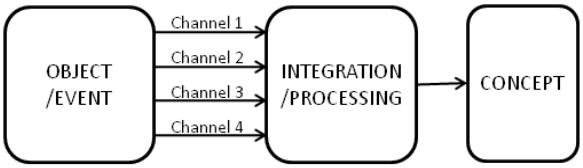

Figure 1. The box to the left represents an object or event. Information can be mediated through various channels. This information is detected by the suitable sense and processed/integrated to match a mental concept.

Another example of cross modal illusions is due to Shams et al. (2000): A number of short presented beeps altered the perceived number of optically presented flashes.

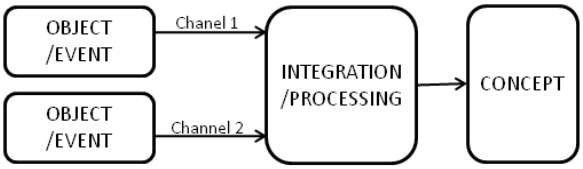

Figure 2. The boxes to the the left represent two different objects or events. Information is mediated through different channels. Information is detected by a suitable sense and processed/integrated to match a mental concept.

Illusions have been useful since they may reveal how the brain interprets and processes information about our surroundings. This process is shown in fig. 2. As remarked earlier, there seems to be a precondition, that objects/events must be similar enough to be integrated.

A well known illusion, relevant to speech perception, is the McGurk effect, (McGurk and MacDonald, 1976): An auditory /baba/ synchronized with a face pronouncing /gaga/ was perceived as /dada/, i.e. a fused percept which neither agreed with the information conveyed acoustically, nor optically. In the reversed situation: An auditory /gaga/ together with visual /baba/, the subjects reported having heard combinations from the two modalities, such as /gabga/ or /gaba/.

The McGurk effect seems to be robust since it persists asynchronies quite well (Munhall et al., 1996). Audiovisual integration of incongruent stimuli does not require the listener to be unaware of the discrepancy (Massaro, 1987). Further proof of the robustness is due to the fact that the McGurk effect still appears, even when a male voice is dubbed onto a female face (Green et al. 1991).

Recent studies have shown that McGurk effect also appears in audiovisual perception of vowels: Traunmüller and Öhrström (2007) showed that an auditory /gyg/ presented together with a visual /geg/ was perceived as /gig/. An auditory /gig/ together with a visual /gøg/ evoked the percept of /gyg/. An auditory syllable /geg/ paired with a visual /gyg/ evoked the percept of /gøg/. Briefly, information about lip rounding is captured by the visual modality, while information about openness is captured by the auditory modality.

The present study aims at investigating the robustness of audiovisual integration of vowel roundedness as the discrepancy in jaw depression increases. It can be assumed that integration will be impeded, at least at large amounts of discrepancy. The perceivers' consciousness of dubbing will also be controlled.

Method

Subjects

16 (8 female and 8 male) subjects volunteered as perceivers. They reported normal hearing and had normal or corrected-to-normal vision. Average age among the subjects was 33.9 years. They were all native speakers of standard Swedish.

Speech material

One female (aged 34 years) served as speaker. The speech material consisted of the open syllables [hV] (V represents a rounded or unrounded front vowel), where jaw depression was to be varied in small steps (i.e. ranging between [i] and [æ]; [y] and [œ]). In order to produce vowels in these small steps in jaw depression, the speaker used different levels of vocal effort (Schulman, 1989). The speaker was asked to produce vowels such as the teeth were clearly visible. During production, the speaker was recorded on video and audio.

Table 1. Auditory, visual and audiovisual stimuli.

Aud	Vis
[hi]	-
[hy]	-
[hi]	[hi]
[hy]	[hy]

Aud	Vis
[hi]	[hy]-[œ]
[hy]	[hi]-[hæ]

Aud	Vis
-	[hy]-[œ]
-	[hi]-[hæ]

Each token was thereafter measured and tagged in terms of maximal distance (in mm) between the upper and lower teeth and distance between the corners of the mouth. The video recordings were subsequently focused around the mouth part and dubbed such as an auditory syllable [hi] was combined with a visual mouth pronouncing [hV] (V represents a front rounded vowel). An auditory syllable [hy] was combined with a visual [hV] (V represents a front unrounded vowel). There was a total of 56 unimodal visual stimuli, 4 unimodal auditory stimuli (each presented twice), 4 congruent audiovisual stimuli (each presented twice) and 112 incongruent audiovisual stimuli according to table 1, thus making up a total of 176 stimuli.

Procedure

The subject sat at an arm's length from a computer screen. The session was divided into two parts. In the first part subjects were presented unimodal visual stimuli only. They were asked which one of the nine Swedish long vowels they had perceived through lip reading.

In the second part of the session, the subjects were presented auditory and audiovisual (congruent and incongruent) stimuli. They were asked which one of the nine Swedish long vowels they had heard. For each audiovisual stimulus, the subjects were asked to judge the stimulus as dubbed or not.

Results

Visual presentation

Stimuli presented in unimodal fashion was to be identified as one of the Swedish long vowels. For further analysis, a criterion was set up: The intended roundedness had to be identified correctly in at least 70% of the cases. Two of the visual stimuli didn't meet this criterion and was therefore excluded.

Auditory presentation

Auditory stimuli were presented together with audiovisual stimuli. In total there were four different auditory stimuli, of which two were intended /i/ and two were intended /y/. The roundedness of /i/ was correctly identified in 93.8% and 90.6% respectively. The roundedness of /y/ was correctly perceived in 71.9% and 68.8% respectively.

Audiovisual presentation

The audiovisual stimuli were either congruent or incongruent. The congruent stimuli were all correctly identified.

The incongruent stimuli were analysed based on the relative impact of visually presented rounding. Fig. 3a and 3b show the relative visual impact on perceived roundedness as a function of discrepancy in jaw depression. As can be seen, the relative visual impact on perceived vowel roundedness is impeded, in cases of larger discrepancies in jaw depression. This tendency seems to be stronger for an acoustically presented [i] than for [y].

Due to the fact that natural stimuli were used, distances in jaw depression were not independent from the distance between mouth corners, as can be seen in fig. 4a and 4b. A lower jaw correlated negatively with distance between the mouth corners.

The subjects' ability to detect whether an audiovisual stimulus was congruent or not was investigated. As can be seen in fig. 5, many of the stimuli, where the incongruent visual signal had an impact on perceived rounding, were also perceived as incongruent/dubbed.

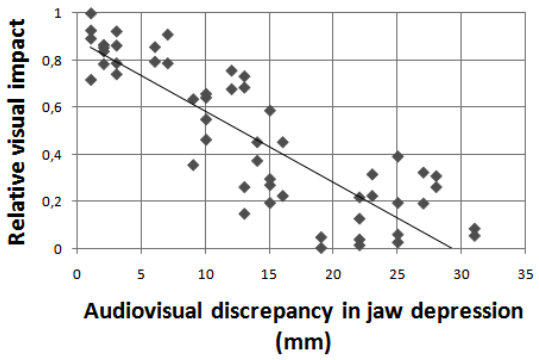

Figure 3a. Acoustically presented [hi], optically presented [hV] (V represents a front rounded vowel). The abscissa relates to the audiovisual discrepancy in jaw depression (in mm). The ordinate relates to the relative visual impact on perception of vowel roundedness.

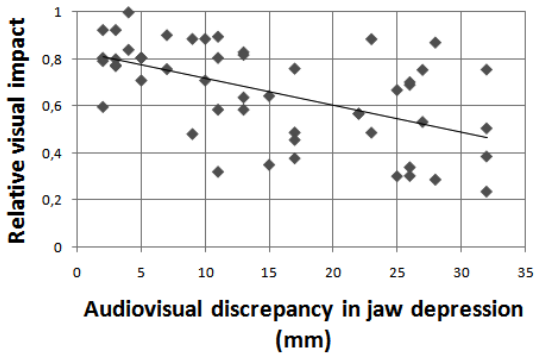

Figure 3b. Acoustically presented [hy], optically presented [hV] (V represents a front unrounded vowel). The abscissa relates to the audiovisual discrepancy in jaw depression (in mm). The ordinate relates to the relative visual impact on perception of vowel roundedness.

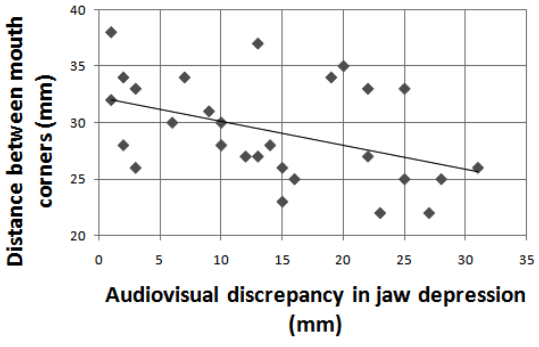

Figure 4a. Dependency between jaw depression and vertical distance between the mouth corners among visually rounded stimuli.

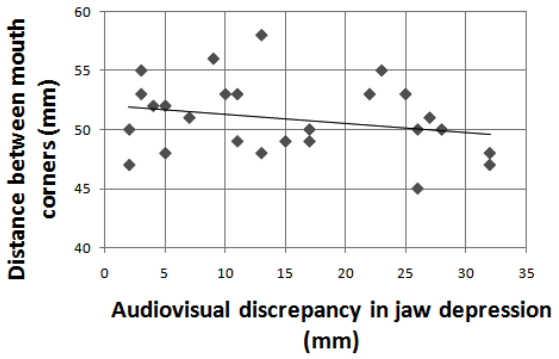

Figure 4b. Dependency between jaw depression and vertical distance between the mouth corners among visually unrounded stimuli.

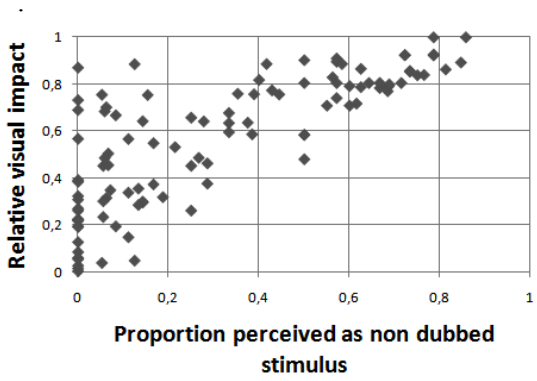

Figure 5. Relation between visual impact on perceived roundedness and detectability of incongruency

Discussion

Analogous to many other illusions, this study show that the visual impact on roundedness is reduced when discrepancy in jaw depression increases. The results may also suggest that this tendency would be stronger among acoustically presented [i] than [y]. This may be true, since the relative visual impact take into account the visual and auditory perception of roundedness in single mode, but such a conclusion may nevertheless be rash, since the two variables, *distance between mouth corners* and *jaw depression* are not independent from each other. Thus, further research is needed. A suggestion would be to proceed with artificial stimuli, using talking heads, in order to fully control these variables.

In this study, the acoustically (unimodally) presented [y] had a perceptual bias towards [i], although it was not reflected in the congruent stimuli. This could be explained by the fact that these auditory stimuli were presented in the same block as the audiovisual ones, of which a majority were incongruent regarding roundedness. However, this potential contextual effect did not have any impact on the auditory perception of /i/. In a future continuation of this study, effort should be put to make /y/ as auditorily rounded as /i/ is auditorily unrounded.

As expected, the subjects' unconsciousness of the dubbing procedure was not a prerequisite for audiovisual integration to occur. This is in line with earlier results by Massaro (1987). This also raises question about visual and auditory separation in other modes of speech perception. As Traunmüller (2006) shows, there are two percepts in speech perception, one auditory (vocal) and one visual (gestural). Will the gestural perception behave in the same way as vocal concerning separation due to discrepancies?

References

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

Green KP, Kuhl PK, Meltzoff AN and Stevens EB (1991). Intergrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Percept. Psychophys.,* 50.9: 524-536.

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

Massaro DW (1987). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, NJ: Lawrence Erlbaum Associates.

Munhall KG, Gribble P, Sacco L and Ward M (1996). Temporal constraints on the McGurk effect. *Percept. Psychophys.,* 58.3: 351-362.

Schulman R (1989). Articulatory dynamics of loud and normal speech. *J. Acoust. Soc. Am.,* 85.1: 295-312.

Shams L, Kimitani Y and Shimojo S (2000). What you see is what you hear. *Nature.,* 408: 788.

Slutsky D A and Recanzone G H (2001). Temporal and spatial dependency of the ventriloquist effect. *Neuroreport*, 12.1: 7-10.

Sumby WH and Pollack I (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.,* 26: 212-215.

Traunmüller H and Öhrström N (2007). Audiovisual perception of openness and lip rounding in front vowels. *J. Phonet.,* 35.2: 244-258.

Traunmüller H (2006). Cross-modal interaction in visual as opposed to auditory perception of vowels., *Working Papers.* 52. 137-140. Dept. Lingustics, Lund University.

Warren DH and Cleaves WT (1971). Visual-proprioceptive interaction under large amounts of conflict. *J. Exp. Psychol.,* 92.2: 206-214.