THE EFFECT OF INCONGRUENT VISUAL CUES ON THE HEARD QUALITY OF FRONT VOWELS

Hartmut Traunmüller and Niklas Öhrström

Department of Linguistics, University of Stockholm, Sweden
hartmut.traunmuller@ling.su.se, niklas.ohrstrom@ling.su.se

ABSTRACT

Swedish nonsense syllables distinguished solely by their vowels [i], [y] or [e], were presented to phonetically sophisticated subjects auditorily, visually and in cross-dubbed audiovisual form with incongruent cues to openness, roundedness or both. Acoustic [y] dubbed onto optic [i] or [e] was heard as a retracted [i], while acoustic [i] or [e] dubbed onto optic [y] were perceived as rounded and slightly fronted. This confirms the higher weight of the more reliable information and that intermodal integration occurs at the level of phonetically informative properties prior to any categorization.

Keywords: Audio-visual integration, McGurk effect, vowel perception.

ZUSAMMENFASSUNG

Es wird gezeigt, dass ein akustisches [y] bei audiovisueller Integration als ein zurückgezogenes [i] gehört wird, wenn synchron mit optischem [i] oder [e] dargeboten, wobei die Zuverlässigkeit der Sinne eine Rolle spielt und die Integration vor der phonetischen Kategorisierung erfolgt.

RÉSUMÉ

Notre étude d'intégration audiovisuelle montre que des [y] acoustiques, présentés en synchronie avec des [i] ou [e] optiques, sont entendus comme des [i] légèrement retirés. Ceci confirme l'importance d'informations fiables et que l'intégration intermodale doit arriver avant la catégorisation.

1. INTRODUCTION

A previous investigation [4] showed that in audiovisually presented front rounded and unrounded vowels with incongruent cues to openness and/or roundedness, listeners perceived openness (vowel height) nearly always by ear alone. In contrast, most listeners, with the exception of a mostly male minority, perceived roundedness by eye rather than by ear even under auditorily ideal conditions. This resulted in fused percepts such as when an acoustic /ge:g/ dubbed onto an optic /gy:g/ was predominantly heard as /gø:g/. Since the acoustic cues to openness are prominent, while those to roundedness are less reliable, and the opposite is true for the optic cues, this lends support to the hypothesis that perceivers gauge the sensory cues for the presence of specific features according to the relative reliability of the information available in the different modalities [2, 5].

Subsequent experiments [3] with vowels in monosyllabic utterances presented auditorily, visually and bimodally with incongruent cues to openness and/or roundedness revealed that incongruent audiovisual stimuli evoke two different percepts: an auditory percept that may be influenced by vision and a visual percept that may be influenced by audition. The two percepts tend to disagree with each other to some extent when there is incongruence between the modalities, but in any case, the strength of the influence of the unattended modality showed between-feature variation that appears to reflect the reliability of the information.

In the mentioned previous experiments [4, 3], the subjects had to identify the vowels heard or seen in a categoric (phonemic) fashion by clicking on the letter that represents the vowel in Swedish. It was, however, observed informally that the quality of vowels heard when a rounded acoustic stimulus (intended /y/ or /ø/) was presented synchronously with an unrounded optic stimulus (intended /i/ or /e/) was quite noticeably different from that of the natural vowels to which listeners attached the same phonetic labels, /i/ or /e/. There appears to arise a subphonemic difference in the front/back dimension that can be related to articulation (considered in relation to the lips, the tongue is further back in rounded vowels) as well as to auditory and acoustic properties (F_2 and F_2', the single upper formant in two-formant stimuli, are lower in rounded vowels). A quantitative analysis of this phenomenon may contribute to a proper understanding of information processing in audiovisual speech perception [2, 1]. The present report is not concerned with the visual percepts evoked by the same incongruent stimuli.

2. METHOD

2.1. Speech material

The Swedish nonsense syllables /gi:g/, /gy:g/ and /ge:g/ from the previous experiment [3] were re-used. Their acoustic properties are detailed in [4]. There were two experimental sessions. The 6 possible incongruent auditory-visual combinations, in which synchronization had been based on the release burst of the first consonant, were used in both. In session 1, each acoustic stimulus was also presented alone - in session 2, each optic stimulus.

2.2. Speakers

There were 4 speakers: S1 (male, 45 years), S2 (male, 29 years), S3 (female, smiling, 29 years), and S4 (female, long-necked, 21 years).

2.3. Listeners/viewers

The 8 subjects who served as perceivers (2 male, aged 27 and 60 years, and 6 female, aged 20, 21, 23, 25, 34 and 59 years) were chosen among the 14 who had participated in the previous experiment [3]. Since not much can be learned about detail in audio-visual integration from subjects with a low susceptibility to optic input, it was attempted to avoid these, but subject selection had to be based on informal observation during the first experiment, before the data had been analyzed. In this way the four who were least sensitive to optic input and two with average sensitivity had been excluded. All the subjects were native speakers of Swedish who had passed at least a basic course in phonetics. They were familiar with the IPA-chart for vowels. All reported normal hearing. Their vision was normal or corrected to normal.

2.4. Procedure

The subjects wore headphones AKG K25 and were seated with their faces at an arm's length from a computer screen. Each one of the 36 stimuli was presented once in quasi-random order, using Windows Media Player. The height of the faces, shown in the right half of the screen, was roughly 12 cm. The subjects were instructed to look at the speaker when shown. In session 1 they were asked to rate the dimensions of the vowel they heard - its roundedness, lip spreading and position in a vowel rectangle reminiscent of the IPA chart. There were two such rectangles, one for unrounded and one for rounded vowels. In session 2 they were asked to rate the same dimensions of the vowel they *saw*. Stimulus presentation was controlled individually by the subjects, who were allowed to repeat each stimulus as often as they wished. They gave their responses by clicking on an electronic response sheet in the left half of the screen. There were 6 choices concerning roundedness: not rounded (0.0), noticeably rounded (0.25), half rounded (0.5), rounded with deficit (0.75), rounded (1.0), and rounded with surplus (1.25). There were 3 choices for lip spreading: not spread (0.0), noticeably spread (0.5), and clearly spread (1.0). In openness, there were 18 response alternatives, the 2nd corresponding to IPA [i] and [y] (0.0), the 6th to [e] and [ø] (1.0) and the 10th to [ɛ] and [œ] (2.0). In session 1, there were 11 backness response alternatives, the 2nd corresponding to the front vowels [i], [e], [ɛ] or [y], [ø], [œ] (0.0) and the 6th to the central vowels (1.0). In session 2, there were only 7 alternatives, the central vowels corresponding to the 4th (1.0). Since the subjects knew that the stuimuli were the same as in the previous experiment, they could expect a rather skewed vowel distribution. Prior to each experiment proper, three stimuli were presented for familiarization. Each of the two sessions lasted for no more than 20 minutes.

3. RESULTS

For the displays and analyses presented in the following, the ratings obtained for each stimulus on the dimensions of roundedness *rnd*, lip spreading *spr,* openness *opn* and backness *bac* were averaged over the 8 subjects.

In Fig. 1a, the average rating of *opn* is plotted against that of *rnd* for all purely auditory stimuli. As for *opn*, the categories were well separated and the ratings obtained with different speakers agreed very well with each other (opn_A = 0.03 to 0.10 for [i], 0.03 to 0.13 for [y] vs. 1.00 to 1.13 for [e]). In contrast, there was a great deal of speaker-related variation in the ratings of *rnd* and of *spr* as well. On average, the intended [i] and [e] of long-necked S4 were even heard as more rounded and less spread than the intended [y] of smiling S3.

In the ratings of the purely optic stimuli (Fig. 1b), there was very good separation and agreement across speakers in the roundedness ratings (rnd_V = 0.00 for all [i], ≤ 0.03 for [e] vs. 0.81 to 1.10 for [y]), and also in the ratings of lip spreading ($spr_V \leq$ 0.13 for [y] as compared with spr_V = 0.25 to 0.63 for [i] and [e]). In contrast, there was much overlap

between the openness categories (opn_V = 0.44 to 1.13 for [i], 0.03 to 1.06 for [y] vs. 0.85 to 1.38 for [e]). On average, the [i] of S4 was even seen as slightly more open than the [e] of S3.

In the Figs. 2a and 2b, heard openness and roundedness of the bimodal stimuli, which all were incongruent, are plotted against the ratings for openness and roundedness obtained with purely acoustic presentation. Fig. 2a makes it evident that the openness ratings for the audiovisual stimuli are highly correlated with those for the purely acoustic stimuli (r = 0.986, ρ = 0.80). No such correlation is visible for roundedness in Fig. 2b (r = –0.26, ns; ρ = –0.05, ns). This means, in fact, that the acoustic signal had no significant influence on the roundedness ratings obtained from these listeners.

Figure 1: Average perceived openness *opn* and roundedness *rnd* of the acoustic stimuli (**a,** listening only, -$_A$) and of the optic stimuli (**b,** lip reading only, -$_V$). Speakers: S1, male: □; S2, male: ◊; S3, female, smiling: ▲; S4, female, long-necked: ●.

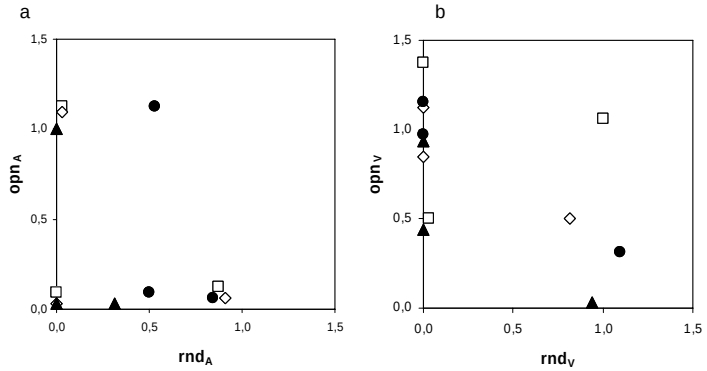


Figure 2: Heard openness (a), roundedness (b) and backness (c) in the incongruent stimuli (*opn, rnd, bac*) as a function of openness and roundedness perceived by listening only (opn_A, rnd_A). Acoustic vowels: /i/ −, /y/ ●, /e/ ■.

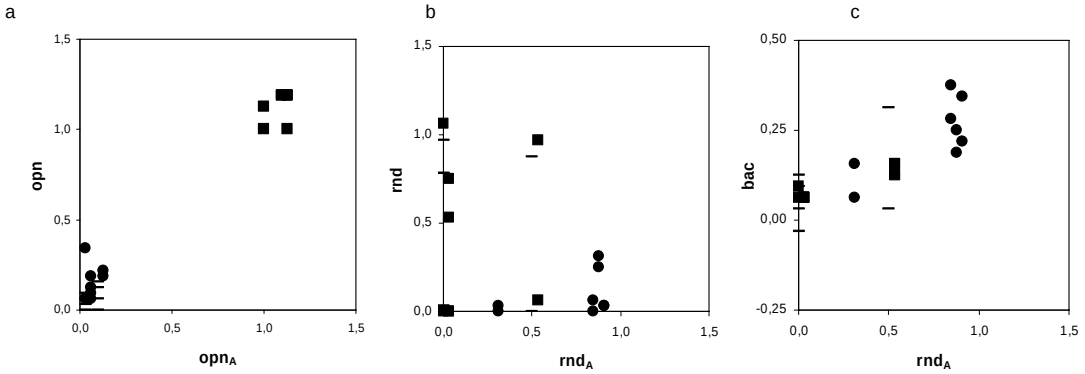
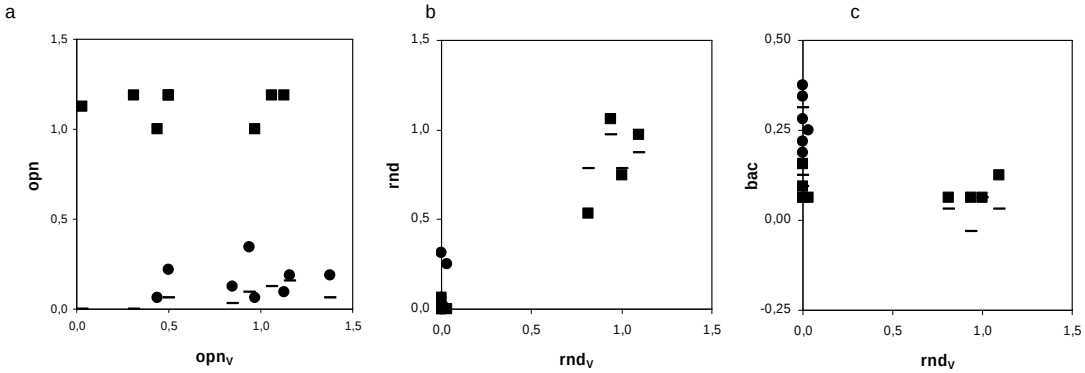


Figure 3: Heard openness (a), roundedness (b) and backness (c) in the incongruent stimuli (*opn, rnd, bac*) as a function of openness and roundedness perceived by lip reading only (opn_V, rnd_V). Acoustic vowels: /i/ −, /y/ ●, /e/ ■.



Figs. 3a and 3b show the same data as Figs. 2a and 2b but now plotted against the ratings obtained with purely optic presentation. Fig. 3b shows clearly that the roundedness ratings for the audiovisual stimuli are correlated with those for the purely optic stimuli (r = 0.93, ρ = 0.79). No such correlation is visible for openness in Fig. 3a (r = –0.23, ns; ρ = 0.03, ns): The optic signal had no significant influence on the rating of openness although it completely dominated the rating of roundedness by these listeners.

In Fig. 2c, heard backness (*bac*) is plotted against roundedness perceived with purely acoustic stimuli (rnd_A). The correlation (r = 0.81, ρ = 0.71) is highly significant. Fig. 3c shows, in addition, a significant correlation ($p < 0.01$, two-tailed) with rnd_V, the roundedness perceived with purely optic stimuli (r = –0.55, ρ = –0.59). Neither bac_A nor bac_V contributed significantly to the heard backness of the incongruent stimuli.

Here are the results of stepwise linear regression analyses, where "bac_{AV}" stands for the interaction $bac_A * bac_V$:

$opn = 0.05 + 1.00\ opn_A$ ($r^2 = 0.97$)

$rnd = 0.05 + 0.82\ rnd_V$ ($r^2 = 0.92$)
$rnd = -0.03 + 0.86\ rnd_V + 0.47\ bac_V$ ($r^2 = 0.95$)

$bac = 0.06 + 0.24\ rnd_A$ ($r^2 = 0.66$)
$bac = 0.06 + 0.25\ rnd_A - 0.20\ rnd_{AV}$ ($r^2 = 0.74$).

No other main factors and interactions contributed significantly.

4. DISCUSSION

Fig. 1a demonstrates that the acoustic signal is not always sufficient to convey the distinction between /y/ and /i/. However, it might have been sufficient in case the stimuli had been blocked by speaker, so that the listeners would have had a better chance to tune in to each speaker's voice.

The results obtained with the incongruent stimuli confirm the previous finding that listeners rely on the acoustic signal in openness perception, while they rely mainly on the optic signal in the perception of roundedness. This holds at least for the subjects selected for this experiment.

Since the subjects rated the dimensions of the stimuli in a sub-phonemic manner, the results show directly that audio-visual integration occurs prior to any phonetic categorization, no later than at the level of the phonetically informative properties that describe the stimuli. This could already be inferred from several previous studies in which the (categoric) perception of consonants had been investigated. For an overview see [1].

Of particular interest is the influence that the roundedness of the acoustically and the optically presented stimuli had on the auditory backness rating of the incongruent stimuli. The position of the tongue and the jaw (in relation to the scull) is, by definition, the same in any front vowels that are distinguished from each other solely by their roundedness. However, in the IPA-chart, and in similar displays, phoneticians use to place the rounded vowels along a line that corresponds to a slightly retracted tongue in unrounded vowels. While this might be just a convention, it could be motivated on the basis of the articulatory gestures involved: If considered in relation to the lips, the tongue is further back in rounded vowels. It could also be motivated on the basis of auditory properties: The formant frequencies, in particular F_2 (and F_2'), are lower in the rounded vowels, as they would be if the tongue was retracted. The present results show that vowels whose auditory cues agree with those of front rounded vowels are actually heard as less fronted than front unrounded vowels (Fig. 2c). They also show that this must be due to auditory rather than articulatory associations. Visible rounding of the lips that was not accompanied by formant lowering had even a significant effect in the opposite sense (Fig. 3c).

5. ACKNOWLEDGMENT

This research has been supported by grant 421-2004-2345 from the Swedish Research Council.

6. REFERENCES

[1] Green, K. P. 1998. The use of auditory and visual information during phonetic processing: implications for theories of speech perception. In: Campbell, R., Dodd, B., Burnham, D. (eds), *Hearing by Eye II: Perspectives and Directions in Research on Audiovisual Aspects of Language Processing.* Hove (UK): Psychology Press, 3-25.

[2] Schwartz. J.-L., Robert-Ribes, J., Escudier, P. 1998. Ten years after Summerfield: A taxonomy of models for AV fusions in speech perception. In: Campbell, R., Dodd, B., Burnham, D. (eds), *Hearing by Eye II.* Hove (UK): Psychology Press, 85-108.

[3] Traunmüller, H. 2006. Cross-modal interactions in visual as opposed to auditory perception of vowels. *Working Papers* 52, 137-140. Dept. of Linguistics, Lund Univ.

[4] Traunmüller, H., Öhrström, N. 2007. Audiovisual perception of openness and lip rounding in front vowels. *J. Phonet.* 35, 244-258.

[5] Wada, Y., Kitagawa, N., Noguchi, K. 2003. Audio-visual integration in temporal perception. *Int. J. Psychophysiol.* 50, 117-124.