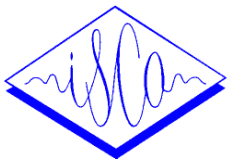


AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4–7, 2003

Effect of audiovisual primes on identification of auditory target syllables

Ville Ojanen, Jyrki Tuomainen, Mikko Sams

Laboratory of Computational Engineering, Helsinki University of Technology, Finland
viloja@lce.hut.fi

Abstract

We studied the representations underlying audiovisual integration using a priming paradigm. Audiovisual primes, preceding auditory targets, were either incongruent (auditory /ba/ & visual /va/) or congruent (auditory /va/ & visual /va/, auditory /ba/ & visual /ba/). The targets were /ba/ or /va/. The intensity of the prime´s auditory component was either 50 dB or 60 dB. Identification speed of the target /ba/ was strongly affected by the nature of the prime. The effect of the incongruent audiovisual prime depended on the intensity of its acoustic component. Our results suggest that processing of visible articulatory movements influence auditory speech processing.

1. Introduction

Speech perception is often audiovisual. We often both hear speech and see some of the corresponding articulatory gestures. Seeing a congruent articulating face improves perception of acoustic speech stimuli. The improvement is the stronger the worse is the signal-to-noise ratio of the acoustic stimuli [1]. In addition to improving identification, watching articulatory gestures may even modify auditory percepts phonetically. For example, simultaneously presented acoustic /ba/ and visual /ga/ are usually perceived as /da/ [2]. This "McGurk effect" is a strong auditory illusion, which occurs even when the acoustic syllables are perfectly identified unimodally.

A key issue in the research on audiovisual speech perception is at which processing levels the acoustic and visual inputs interact [3]. It is open whether 1) audiovisual speech perception is based on modality specific representations, 2) either auditory or visual modality serves as a common representation space to which the other modality is mapped or 3) there are common amodal representations. We studied the representations involved in audiovisual integration by looking at the effects of a prior exposure to congruent and incongruent audiovisual prime stimuli to the identification speed of auditory targets.

Several studies have shown that the preceding auditory context has a striking effect on the identification of following speech stimuli. Studies on the so called phonetic context effect show, that acoustically identical syllables are perceived differently depending on context. For example, a syllable which is perceptually ambiguous in isolation - sometimes perceived as /ga/, sometimes perceived as /da/ - is strikingly more often identified as /ga/ when the syllable /al/ precedes it. The same syllable is frequently perceived as /da/ when preceded by /ar/ [4].

In selective adaptation studies subjects are repeatedly presented with stimuli at one end of a synthetic speech continuum, for example /ba/. Then subjects have to identify an item on the /va/-/ba/ continuum. Selective adaptation shifts the phoneme boundary towards the /va/ end of the continuum. In another words, after exposure to many prototypical /ba/ stimuli, the listener identifies more of the following stimuli as /va/. Ambiguous adaptors, intermediate between two categories, do not give rise to selective adaptation effects [5]. Previous studies have suggested that selective adaptation is not affected by the visual component of an audiovisual stimulus and is due to the auditory representations alone [6,7]. However, Bertelson and coworkers [8] suggest that also the visual component of an audiovisual stimulus has an effect on the identification of the target stimulus. Exposure to an ambiguous syllable (intermediate between /aba/ and /ada/), dubbed onto a face articulating /aba/ increased the proportion of /aba/ responses despite the adaptor being ambiguous. The effect was suggested to be purely visual and not to rise from auditory representations or their modification by the visual component of the audiovisual stimulus.

We studied the nature of the representations underlying audiovisual speech perception using congruent and incongruent audiovisual primes and auditory targets. To vary the effectiveness of the visual component on the audiovisual prime, the acoustic component was presented at two different intensities. Instead of auditory identification scores, we measured reaction times (RTs) to the acoustic target stimuli. We analysed the data only from those subjects, who strongly integrated the acoustic and visual speech, as evaluated on the basis of the strength of the McGurk effect.

2. Methods

2.1. Stimuli

Meaningless consonant-vowel syllables /va/ and /ba/, uttered by a male Finnish speaker, were videotaped in an anechoic chamber. The visible area was from the bottom of the talkers nose to the middle of his chin, with no background visible (Fig. 1). Sound files (digitized at 44 100 Hz) and video clips (frame rate 25 Hz) were extracted from the digital video. The duration of the acoustic /va/ was 315 ms and that of the acoustic /ba/ 328 ms. The duration of the corresponding visual utterances was 1 s. In audiovisual tokens the mouth opening began 300 ms prior to the acoustic utterance onset.

2.2. Baseline experiment

A two-choice auditory identification task was run for eighteen subjects (15 females and 3 males, mean age 18

years, range 16-19 years). Auditory /va/ and /ba/ (N=20+20, ISI = 1.5 s) stimuli were presented to the subjects in random order. Subjects were instructed to identify the stimulus by pressing a respective button as soon as possible. The order of the response buttons was reversed for half of the subjects.

2.3. Audiovisual experiment

Twenty-seven high-schools students (19 females and 8 males; mean age 18 years, range 16-19 years) served as subjects. Data from one male subject was excluded on the basis of reported fatigue and inability to perform the experimental task as instructed. All subjects reported normal hearing and normal or corrected vision, and were native speakers of Finnish. Subjects were paid 10 euros for their participation in the 1-h experiment.

Stimuli were presented in three blocks: 1) McGurk block, 2) Incongruent block, and 3) Congruent block. In the McGurk block, incongruent audiovisual stimuli (auditory /ba/ combined with visual /va/ and vice versa) were presented to the subjects. The interval between two consecutive stimuli was 1.5 s plus RT (927±90 ms; mean±SEM). The intensity of the auditory component of the audiovisual stimulus was 60 db or 50 db. In the Incongruent and Congruent blocks an audiovisual prime was followed by an auditory target stimulus. In the Incongruent block, the auditory target followed the onset of the audiovisual prime by 1.2 s. The interval from the end of the acoustic component of the prime to the target onset was 580 ms. The ISI between two consecutive stimulus pairs was 2520 ms plus RT. The prime was incongruent (auditory /ba/ combined with visual /va/ and vice versa). The target stimulus was auditory /ba/ or /va/. The intensity of the auditory component of the audiovisual prime was 60 or 50 db. The intensity of the target stimulus was always 55 db. In the Congruent block there were two congruent priming stimuli (audiovisual /ba/ and /va/). It was otherwise identical to the incongruent block. Each stimulus condition was repeated 20 times in the three stimulus blocks.

The experiment was conducted in an acoustically shielded booth with background noise of about 30 db. Stimuli were presented on a computer screen 50 cm away from the subjects shoulder. The visual stimuli subtended a visual angle of 11,3°. The auditory stimuli were presented through two loudspeakers located symmetrically in front of the subject on both sides of the monitor.

Figure 1. An example of the visual speech stimuli.

2.4. Statistical analyses

Wrong responses and outlying RT's (longer than mean ± 3.5 standard deviations) were excluded from each subject's data prior to statistical analyses. The data from the McGurk block was analysed with the Fishers exact test. The data from the two priming blocks were analysed separately for the /ba/ and /va/ targets with repeated measures ANOVAs (factors: Prime type, Auditory intensity). Significant interactions were further analysed with a Fisher LSD Post-hoc test. The effects of the incongruent primes containing auditory /va/ and visual /ba/ are not reported here.

2.5. Procedure

The session always began with the McGurk block to have a measure of the strength of audiovisual integration at both auditory intensity levels for each subject. Subjects were instructed to pay attention to the auditory utterances and to fixate on the articulating mouth. After each stimulus the subject was to press one of three buttons corresponding to what they heard: /ba/, /va/ or something else (for example combinations /bva/ or /vba/). The utterances were told to be sometimes incongruent and it was emphasised that the subjects should always respond according to what they heard and that RT's were not measured. Subjects who identified the acoustic stimuli well (more than 50% of all responses, N=9), indicating that they did not integrate the auditory and visual components, were excluded from the further analysis.

The following blocks consisted of the prime-target pairs. The subjects were instructed to pay attention to the first audiovisual stimuli similarly as in the McGurk block but respond only to the second auditory stimulus. They were asked to press one of the two buttons after the second stimulus as fast as possible but correctly. There were approximately 1 min. breaks between the blocks.

3. Results

3.1. The baseline experiment

The baseline experiment revealed a clear difference in the RTs to the auditory targets. Mean RT to target /va/ was 697±42 ms and to target /ba/ 744±42 ms. The difference was statistically significant, $t(17)=4.4$, $p<0.0004$. Auditory /ba/ was also more often misidentified than /va/. The mean error rate for /ba/ was 19%±4 and 4%±1 for /va/. The difference was significant, Sign test, $z=2.6$, $p<0.01$.

3.2. The audiovisual experiment

3.2.1. The McGurk effect

Figure 2 shows the mean proportions of visual, auditory and combination responses to the incongruent audiovisual stimuli. The mean proportion of visual responses was 83±5% in the 60-db condition and 78±6% in the 50-db condition (range 45%-100%). A low proportion of correct identifications of the acoustic stimuli (and high proportion of "visual" responses) shows that the two stimuli were strongly integrated. The effect of the auditory intensity level was not statistically significant in any of the response categories.

3.2.2. The priming effects

The mean RTs to the targets /ba/ and /va/ after the incongruent and the two congruent prime stimuli are shown in Fig. 3 and Table 1. Statistically significant effects were observed only for the target /ba/. The mean error rates in

identification of the targets /ba/ and /va/ were 12±3% and 2±1%.

The identification speed of the target /ba/ in the 60-db condition was not differently affected by the three audiovisual primes. However, in the 50-db condition there were clear differences. The mean RT to /ba/ was 794±47 ms after congruent audiovisual /ba/ prime, 747±52 after congruent audiovisual /va/ prime and 707±41 after incongruent prime (auditory /ba/+ visual /va/, perceived as /va/).

Intensity of the auditory component of the audiovisual prime affected RTs only after the incongruent prime. RT was 81 ms faster in the 50 db than in the 60 db condition. The interaction of the factors Prime type x Auditory intensity was significant, $F(2,32)=3.3$, $p<0.05$. Significant differences between the specific means (post-hoc comparisons) are listed in Table2.

The mean error rates at the different stimulus conditions showed similar pattern of results. The mean error rate to target /ba/ after the congruent audiovisual /ba/ prime was 14%±4 in the 60 db condition and 12%±4 in the 50 db condition. The mean error after the congruent audiovisual /va/ prime was 7%±4 in the 60 db condition and 2%±1 in the 50 db condition. After exposure to the incongruent prime stimulus the mean error rate was 24%±6 in the 60 db condition and 11%±2 in the 50 db condition.

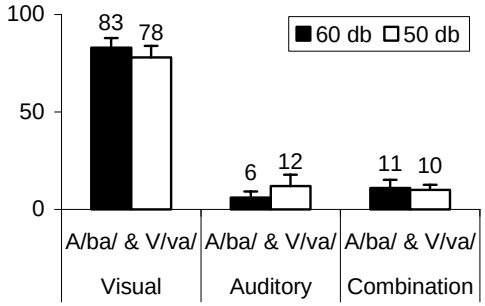
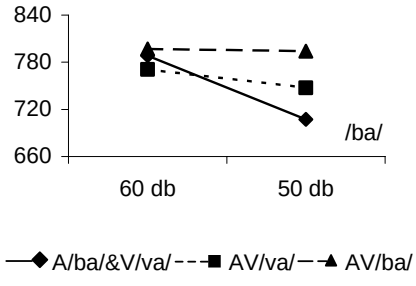


Figure 2. Mean proportions of visual, auditory and combination responses to the incongruent audiovisual stimuli.



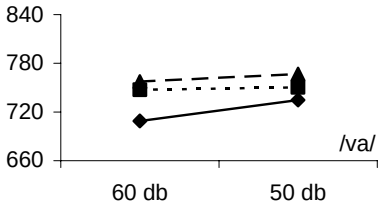


Figure 3. The mean RT's to the targets /ba/ and /va/ at the two auditory intensity levels of the incongruent and he two congruent prime stimuli. Statistically significant effects were observed only for the target /ba/, not for /va/.

Table 1. The mean RT (±SEM) to target /ba/ after the different primes.

	Prime type	*db*	*RT*
1	A/ba/&V/va	60	788±57
2	A/ba/&V/va	50	707±41
3	AV/va/	60	771±58
4	AV/va/	50	747±52
5	AV/ba/	60	797±70
6	AV/ba/	50	794±47

Table 2. Significant differences in RTs after various primes.

#	*Prime type*	*db*	1	2	3	4
1	A/ba/&V/va	60				
2	A/ba/&V/va	50	p<0.01			
3	AV/va/	60		p<0.03		
4	AV/va/	50				
5	AV/ba/	60		p<0.01		p<0.03
6	AV/ba/	50		p<0.01		p<0.04

4. Discussion

We studied the representations underlying audiovisual interactions in speech perception using a priming paradigm. Three rather puzzling effects were observed. First, identification speed of the target /ba/, but not of target /va/, was strongly affected by the nature of the preceding audiovisual stimuli. Second, the identification of the target /ba/ was slower after exposure to audiovisual /ba/ than /va/. Third and most important, only after exposure to the incongruent prime, the identification speed of target /ba/ varied as the intensity of the prime's acoustic component varied. Similar pattern of results were found in the error data. In the following we discuss various factors that might have caused the results.

Most of the effects of the preceding stimulus have been demonstrated with target stimuli at and close to the category boundary [9]. The effects are generally smaller to well-identified stimuli. The baseline experiment showed that /ba/

was more difficult and slower to identify than /va/. Physical differences between the two syllables might explain partly the differences in identification. However, /ba/ was most probably also phonetically more ambiguous than /va/. We suggest that the differences in the priming effects for the targets /va/ and /ba/ are due to the differences in phonetic ambiguity of the stimuli.

Rather surprisingly, the identification of the target /ba/ was the slowest after exposure to audiovisual /ba/. One would expect that an exposure to a stimulus would enhance its processing when the stimulus is repeated. However, with speech stimuli the situation is very different. Due to phonetic context effects and selective adaptation, the perceived identity of an ambiguous speech sound can change. For example, an ambiguous consonant is heard as a /t/ when following good examples of /d/, and as a /d/ when following good examples of /t/. We suggest that in our experiment a somewhat ambiguous /ba/ was heard as a clearer /ba/ (faster RTs and less errors) following /va/ and as a poorer example of /ba/ following /ba/.

Importantly, there was a clear effect of auditory intensity only after exposure to the incongruent prime. In the 60-db condition the effect of the incongruent prime was similar to those of the other prime stimuli. In the 50-db condition, however, the difference was striking. RT was about 90 ms faster after the incongruent prime than after the audiovisual /ba/ prime.

Somewhat unexpected, the effect of auditory intensity did not influence the response distribution in the McGurk block. This might be due to at least two things. First, the stimulus presentation in the McGurk and priming blocks was very different. The effects of the auditory intensity appeared in delayed RTs to a second stimulus due to the previous one. Various factors, such as the ISI and subject's task, were different in the McGurk and priming blocks. Also RT's and identification scores might reflect different aspects of the underlying processes.

In the light of the possible representations underlying audiovisual integration, the observed effects could be due to 1) common amodal representations, 2) independent operations on modality specific representations, or 3) auditory or visual representation spaces to which the other modality is mapped.

If the priming effects were due to a common amodal representation, the effects of the audiovisual /va/ and incongruent prime should have been similar, since the incongruent prime was also perceived as /va/. Even though there were no significant differences between the RT's after the incongruent and audiovisual /va/ conditions at either intensity level, the difference in RT's between the 60 and 50 db conditions was significant only after the incongruent prime. However, since the present incongruent audiovisual stimuli produced only about 80% visual responses, the possibility that an amodal representation would explain the obtained priming effects can not be ruled out.

It seems that the observed priming effects can not be due to operations on unimodal representations only. RT's in the 50 db condition were significantly slower after the audiovisual /ba/ than the incongruent prime, even though the primes contained the identical auditory component. The effects can not be due to operations on the visual representations either, since the RT's in the incongruent prime condition varied significantly as the intensity of the auditory component varied. Moreover, no such effect was observed with audiovisual /va/ prime containing an identical visual component.

The effects were probably due to interactions in the processing of the auditory and the visual components of the incongruent prime. One possibility is that processing the visible articulatory movements influenced auditory speech processing. A possible explanation of the results is that the visual component of the incongruent prime exerted more influence on the auditory processing when the intensity of the auditory stimulus was 50 db than 60 db and directly changed the nature of the auditory processing. Such direct visual influence on the auditory processing is plausible on the basis of neurophysiological studies demonstrating that visual speech activates the auditory cortex [10,11,12].

5. Conclusions

An incongruent audiovisual prime, giving rise to the McGurk effect, influenced the identification of the following auditory target in a way that is difficult to explain on the basis of independent processing of its auditory and visual components. The effect seems to be due to interactions in the processing of the auditory and the visual components of the incongruent prime. Our results suggest that visual speech might have an access to auditory representations. This interpretation of the results supports a model, where the auditory modality serves as a common representation space to which the visual modality is mapped. However, the results do not rule out the possibility that amodal phonetic representation underlie, at least partly, the present priming effects.

6. Acknowledgements

This study was supported by the Academy of Finland grant 44897 to the Center of Excellence of Computational Science and Engineering and grants 49881 and 49900 and to M. Sams.

7. References

[1] Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America, 26(2), 212-215.

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

[3] Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), Hearing by Eye: The Psychology of Lip-reading (pp. 3-51). London: Lawrence Erlbaum Associates.

[4] Mann, V.A., 1980. Influence of preceding liquid on stop-consonant perception. Percept. Psychophys. 28, 407-412.

[5] Sawusch, J.R. & Pisoni, D.B. (1976) Response organization in selective adaptation to speech sounds. Percept Psychophys, 20, 493-498.

[6] Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. Percept Psychophys., Oct;30(4), 309-314.

[7] Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in pseech perception using a compelling

audiovisual adaptor. Journal of the Acoustical Society of America, 95, 3658-3661.

[8] Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. Psychological Science, in press.

[9] Repp, B.H., 1982. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychol. Bull. 92, 81-110.

[10] Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., & Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett, 127(1), 141-145.

[11] Mottonen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. Brain Res Cogn Brain Res, 13(3), 417-425.

[12] Calvert, G.A., Bullmore, E., Brammer, M.J., Campbell, R., Iversen, S.D., Woodruff, P., McGuire, P., Williams, S., David, A.S. (1997). Silent lipreading activates the auditory cortex. Science, 276, 593-596.