Visual attention modulates audiovisual speech perception

Tiippana, Kaisa; Andersen, Tobias; Sams, Mikko

Published in:
European Journal of Cognitive Psychology

Link to article, DOI:
10.1080/09541440340000268 10.1080/09541440340000268

Publication date:
2004

Document Version
Early version, also known as pre-print

Visual attention modulates audiovisual speech perception

K. Tiippana, T. S. Andersen and M. Sams

Helsinki University of Technology, Laboratory of Computational Engineering,

P.O. Box 9203, 02015 HUT, Finland

Final manuscript EJCP44.02

Dated 25 February, 2003

Short title: Visual attention and audiovisual speech

Corresponding author: Kaisa Tiippana

Helsinki University of Technology

Laboratory of Computational Engineering

P.O. Box 9203

02015 HUT

Finland

Tel. –358-9-451 5326

E-mail: kaisa.tiippana@hut.fi

ABSTRACT

Speech perception is audiovisual, as demonstrated by the McGurk effect in which discrepant visual speech alters the auditory speech percept. We studied the role of visual attention in audiovisual speech perception by measuring the McGurk effect in two conditions. In the baseline condition, attention was focused on the talking face. In the distracted attention condition, subjects ignored the face and attended to a visual distractor, which was a leaf moving across the face. The McGurk effect was weaker in the latter condition, indicating that visual attention modulated audiovisual speech perception. This modulation may occur at an early, unisensory processing stage, or it may be due to changes at the stage where auditory and visual information is integrated. We investigated this issue by conventional statistical testing, and by fitting the Fuzzy Logical Model of Perception (Massaro, 1998) to the results. The two methods suggested different interpretations, revealing a paradox in the current methods of analysis.

INTRODUCTION

As is now widely acknowledged, speech perception is audiovisual. Seeing the articulatory gestures of a talker influences the auditory speech percept. This occurs not only in noisy conditions (Erber, 1969; Sumby & Pollack, 1954), but also when the auditory speech signal is clear. McGurk and MacDonald (1976) created incongruent audiovisual speech by dubbing a voice saying /ba/ to a face articulating /ga/, and vice versa. They found that even though auditory speech alone was perfectly recognized, combining it with incongruent visual speech altered percepts so that subjects frequently reported hearing /da/ in the former, and /bga/ in the latter case. This is known as the McGurk effect. Since its discovery, it has been used in a multitude of experiments investigating audiovisual integration of speech (e.g. Grant & Seitz, 1998; Green, 1996) since its strength reflects the extent of visual influence in speech perception.

The question addressed in this study is whether audiovisual speech perception is affected by endogenous attention. Few studies have addressed this issue, and most of them only implicitly.

The original McGurk and MacDonald (1976) study together with several other subsequent reports suggest that integration of auditory and visual information is strongly automatic since the McGurk effect occurs even when the observer is aware of how the stimuli are constructed. Also, Driver (1996) has shown that visual information affects auditory speech localisation even when this results in poorer

performance in a speech shadowing task, leading him to the conclusion that audiovisual integration arises automatically.

Furthermore, in a study on audiovisual speech perception in children, Massaro (1984) found that audiovisual percepts were very similar both when a child just reported what the talker said, and when he also had to report whether the talker's mouth moved. The second task was used to ascertain that the child was paying attention to the visual information. Supposing that visual attention was stronger in the latter case, this result suggests that it does not influence perception of audiovisual speech.

The automatic nature of audiovisual speech perception is a widely held tenet. Integration of heard and seen speech is often considered to occur in an obligatory and automatic manner (e.g. Calvert, Campbell, & Brammer, 2000; Colin et al., 2002). Still, the issue of automaticity is not as clear-cut as it may first seem.

Easton and Basala (1982) have reported that in an audiovisual word recognition task, subjects exhibited no visual influence when they were instructed to base their judgments of incongruent audiovisual speech on what they heard. In other words, there was no McGurk effect, contrary to numerous other reports. On the other hand, in a reverse condition, where subjects reported what they saw, there was significant auditory influence. The former result would suggest that visual influence can be voluntarily extinguished, while the latter result implies that auditory influence is unavoidable. However, Dekle, Fowler and Funnell (1992) have pointed out that the former result may have been due to inappropriate stimulus choice, lacking clear differences in the visual place of articulation. Therefore, in their follow-up

experiments they used improved stimuli, and demonstrated strong visual influence in an auditory judgment condition. Also, in the reverse condition there was some influence of auditory speech on judgments of visual speech. These results support the idea that there is an automatic component to audiovisual speech perception since interaction between audition and vision could not be avoided even when observers were trying to ignore one modality. On the other hand, there seems also to be an attentional component since the response patterns to the McGurk stimuli were different depending on whether the subject was attempting to respond according to audition or vision.

In a similar vein, Massaro (1998) has shown that by instructing subjects to respond according to either auditory or visual information only, their responses to audiovisual consonants are biased towards the instructed modality. Furthermore, the influence of incongruent visual speech on vowel recognition is greater when subjects are asked to watch and listen and to report what was spoken, than when they are told about the dubbing and asked to report only what they heard (Summerfield & McGrath, 1984).

In summary, this review of existing experimental results has provided contradictory evidence regarding the role of attention in audiovisual speech perception. Moreover, previous studies investigated shifts of attention between modalities, with the only exception of Massaro (1984) who manipulated the degree of visual attention. In order to study intramodal endogenous attention further, we designed an experiment where visual attention was modulated using a distractor stimulus presented together with the talking face, and the McGurk effect was measured in two conditions: with full attention at the talking face, and with attention directed towards the distractor. A

weaker McGurk effect in the latter case would mean that visual attention influenced audiovisual speech perception.

METHODS

Subjects

Subjects were 17 native speakers of Finnish. They reported normal hearing and had near visual acuity of at least 6/6. Three subjects were omitted from result analysis since they never integrated, always giving only auditory responses, thus being uninformative for the purposes of this study. The remaining 14 subjects had an age range of 19-37, mean 25 years.

Stimuli

Video clips of a Finnish female talker uttering consonants /k/, /p/ and /t/ embedded in /eCe/ context were edited to create the experimental stimuli in an extended factorial design. This resulted in 15 different stimuli consisting of 6 unisensory stimuli: 3 auditory and 3 visual stimuli, and 9 audiovisual stimuli: 3 congruent audiovisual stimuli for which the auditory and visual utterances were the same, and 6 incongruent audiovisual stimuli for which the auditory and visual utterances differed. There were two experimental conditions in which the speech stimuli were presented:

In the ‘Face’ condition, during the presentation of unisensory auditory stimuli, the visual stimulus was a blank screen of the average colour of the video clips. During the

presentation of unisensory visual stimuli, there was no sound and subjects had to lipread what the face was articulating. During the presentation of audiovisual stimuli, both speech sounds and an articulating face were presented.

In the 'Leaf' condition, a visual distractor was presented together with the speech stimuli. The visual distractor was a partially transparent leaf[1]. During the presentation of visual and audiovisual speech stimuli, the leaf floated across the talker's face starting from bottom middle, going slightly to the right across the mouth, and finishing at the talker's temple on the left. The leaf was next to the mouth when the talker spoke, overlapping it slightly without covering it. The leaf was also spinning around slowly. The movement gave the impression that the leaf was floating in the wind. The distractor movement was identical for all stimuli and easy to follow. Fig. 1 shows a still image from a stimulus together with a trace marking the distractor path. This kind of visual distractor was chosen to manipulate visual object-based attention in a situation where both the face and the distractor were near fixation as the talker spoke. During the presentation of unisensory auditory stimuli, the visual distractor moved exactly as described above but now there was no face, so that the leaf was floating across a blank screen.

Figure 1 about here

Speech sounds were delivered at a level of approximately 50 dB(A) through two loudspeakers positioned on each side of a computer screen where the visual stimuli

were presented. The height of the face on the screen was 16 cm, and it was viewed at a distance of about 130 cm.

Procedure

The experiment consisted of two conditions in which the state of visual attention was different. In both conditions, the stimuli were presented in a pseudo-random order, with ten repetitions of each of the 6 unisensory and 9 audiovisual stimuli. After a stimulus was presented, the subject wrote down which consonant(s) the talker uttered. Subjects were free to write down any consonant(s). Each stimulus was preceded by a warning tone to alert the subject to look back at the screen.

In the ‘Face’ condition, the subject was instructed to pay attention to the auditory speech and to the talking face. This is the baseline situation used in most experiments investigating the McGurk effect.

In the ‘Leaf’ condition, the visual distractor (a floating leaf) was presented together with the speech stimuli. The subject was instructed to pay attention to the auditory speech and the distractor, and to ignore the face. When there was no sound, he/she should still attend the leaf and not the face, and guess what the face uttered.

The presentation of the ‘Face’ and ‘Leaf’ condition was counterbalanced across subjects.

RESULTS

Responses were averaged across subjects and grouped in five categories: k, p, t, combination, and other. Combinations were responses where both visual and auditory consonants were reported as a cluster. Analyses of variance (ANOVA) for repeated measures[2], with stimulus and condition as factors, were performed on the percentages of auditory responses separately for unisensory auditory stimuli, as well as for congruent and incongruent audiovisual stimuli. Incongruent stimuli were further separated into auditory dominance and McGurk stimuli. Responses corresponding to the correct visual utterance were analysed for unisensory visual stimuli.

Unisensory auditory stimuli were almost perfectly recognized in both 'Face' and 'Leaf' conditions, with the percentage of correct responses ranging between 96-100%. Neither factor, stimulus [$F(2,26)=1.43$, $p>0.1$] or condition [$F(1,13)=0.70$, $p>0.1$], was statistically significant. This is not particularly surprising since the difference between conditions was that the subjects watched a blank screen in the former, and followed the distractor in the latter condition. Changes in the visual attention condition did not thus influence auditory percepts.

Unisensory visual stimuli were less well recognized, as shown by Fig. 2. ANOVA showed a main effect of stimulus [$F(2,26)= 34.7$, $p<0.001$], reflecting differences in the recognition rate between stimuli. Visual /k/ had the lowest recognition rate at 41% and 37% for 'Face' and 'Leaf' condition, respectively, being commonly classified as 't' (44% and 49% of responses, respectively). Recognition rate was 88% and 94% for visual /p/, and 84% and 61% for visual /t/ in 'Face' and 'Leaf' condition, respectively.

Subjects tended to respond ‘t’ not only to visual /t/ but also to /k/. This may be due to the fact that in a Finnish articulation of /t/, the visibility of the tongue tip behind the teeth, when present, is a distinguishing feature from the articulation of /k/ since this feature is never present in /k/ (Pesonen, 1968; Sovijärvi, 1963). However, this distinctive tongue position is not always visible in /t/, either. Thus, whenever subjects detected the tip of the tongue behind the teeth, they responded /t/ and therefore were quite accurate at recognizing the visual /t/ stimulus. However, when they did not observe this feature, the articulation could be either /t/ or /k/, and subjects gave both responses.

Figure 2 about here

The main effect of condition was not significant across visual stimuli [$F(1,13)=1.44$, $p>0.1$]. However, interaction between stimulus and condition was significant [$F(2,26)=6.25$, $p<0.01$]. Analysis of simple effects showed no effect of condition for /k/ [$F(1,13)=0.70$, $p>0.1$] and /p/ [$F(1,13)=0.66$, $p>0.1$] but a significant effect for /t/ [$F(1,13)=5.32$, $p<0.05$]. Thus, attention had no effect on the recognition of visual /k/ and /p/, but /t/ was less well recognized when attention was distracted from the face. In the ‘Leaf’ condition, the most common error response (21%) for visual /t/ was ‘k’. The detectability of the tongue position as a distinguishing feature between /t/ and /k/ can be explain this so that in the distracted attention condition, subjects were poorer at detecting the tip of the tongue behind the teeth, and so confused /t/ with /k/ more

often. In contrast, it seems that for visual /p/ the distinguishing feature of lip closure was not disturbed even in the distracted attention condition since the high recognition rate remained unchanged. On the other hand, visual /k/ remained equally confusable with /t/ in both conditions.

For congruent audiovisual stimuli, the recognition rate was 94-100% without any significant effect of stimulus [$F(2,26)=1.82$, $p>0.1$] or condition [$F(1,13)=1.41$, $p>0.1$]. Since auditory performance alone was already at a ceiling level, there was no room for improvement with the addition of congruent visual speech.

Incongruent audiovisual stimuli formed two groups: auditory dominance stimuli and McGurk stimuli. Auditory dominance means that responses were determined by auditory speech. The McGurk stimuli, in contrast, showed visual influence so that responses were predominantly something other than the response to the auditory speech alone.

Auditory dominance was present for incongruent presentations of /k/ with /t/. These stimuli did not show visual influence, responses being almost entirely auditory (93-99%). As with other stimuli giving only auditory responses, there was no effect of stimulus [$F(1,13)=2.17$, $p>0.1$] or condition [$F(1,13)=0.33$, $p>0.1$]. A likely reason for auditory dominance is that visual /k/ and /t/ share features and to that extent are confusable with each other, and therefore when combined with a clear auditory signal, the percepts were determined by the auditory component.

There were four McGurk stimuli: auditory /p/ + visual /k/, auditory /p/ + visual /t/, auditory /k/ + visual /p/, and auditory /t/ + visual /p/. Fig. 3 shows that in the 'Face' condition, visual influence was considerable since the percentage of auditory responses was very low. Also, it shows that visual influence was weaker in the 'Leaf' condition since there were more auditory responses. ANOVA showed main effects of stimulus [$F(3,39)=27.4$, $p<0.001$] and condition [$F(1,13)=47.0$, $p<0.001$]. The first main effect was due to there being fewer auditory responses for stimuli containing auditory /p/ than for stimuli containing visual /p/. That is, presentation of auditory /p/ with visual /k/ or /t/ showed more visual influence than presentation of auditory /k/ or /t/ with visual /p/. The second main effect occurred because there were more auditory responses when the distractor leaf was attended than when attention was directed at the face. Thus, the effect of visual speech was smaller when attention was distracted from the face. Interaction between factors was not significant [$F(3,39)=1.26$, $p>0.1$], indicating that the effect of attention was similar for all stimuli.

Figure 3 about here

In summary, ANOVA confirmed that visual attention influenced audiovisual speech percepts for the McGurk stimuli. This attention effect was very strong, resulting in 17-42 percent point change in auditory responses. For other stimuli attention had no effect, the only exception being visual /t/.

In the 'Face' condition, there was no visual distractor. To check whether the distractor would influence the results when the face was attended, a control experiment was

conducted with three subjects who performed the 'Face' condition both with and without the distractor. The two sets of results were nearly identical. Linear regression between the two sets of results (comparing response distributions for the 15 stimuli with or without distractor) gave a very good fit and a slope close to 1 (slope=0.96, intercept=0.08, r^2=0.98). This nearly perfect correspondence between the results indicates that the presence or absence of the visual distractor did not have an effect. This is in line with previous studies that have used disctractors in tasks involving manipulations of attention with face stimuli. Palermo and Rhodes (2002) studied the influence of divided attention on holistic face perception and found that, when a target face was attended, performance was not affected by whether distracting flanker faces were present or absent. Vroomen, Driver and de Gelder (2001) studied the effect of attention on audiovisual integration of emotional expressions using a dual-task paradigm, and in three experiments with different visual and auditory distractor tasks found that the face influenced voice judgements in the same way both when there was no distractor and when the distractor was present but not attended.

Eye movements were not monitored in these experiments, so that it was not possible to ensure objectively that subjects followed the face or the leaf as instructed. However, had they not done so and instead fixated the face when the leaf should have been followed, the two conditions would have shown identical results (like in the control experiment described above). Since this was not the case, subjects must have been quite consistent in obeying the instructions. Adherence to instructions was also checked verbally by asking subjects between blocks whether they had kept looking at the leaf, which they confirmed. Consequently, subjects' eye movements were probably different between conditions. Paré, Richler, ten Hove and Munhall (2003)

have shown that fixations concentrate on the eyes, mouth and nose area in a speech perception task which was very similar to our 'Face' condition. In our 'Leaf' condition, the distractor followed a path very close to these areas (see the dashed line in Fig. 1), so that the eyes were probably scanning similar areas in both conditions. Furthermore, visual speechreading performance is little affected by slightly eccentric viewing, at least up to 7 degrees (Paré et al., 2003; Smeele, Massaro, Cohen, & Sittig, 1998), and the eccentricity of the disctractor path from the centre of the mouth was maximally 3 degrees during speech in our experiments. Therefore, possible differences in eye positions should not have caused differences in speechreading performance between conditions. However, a possible caveat is that eye movements were saccade-fixation sequences in the former condition and smooth pursuit movements in the latter. It is possible that this difference may have affected the results. For example, following the leaf may have disturbed the processing of the face since its image was now moving across the retina. In this case, visual speech perception should have been poorer in the 'Leaf' condition. Since this was the case for one visual stimulus only, the results cannot be accounted for by differences in eye movement patterns.

DISCUSSION

The McGurk effect was measured in two conditions. In one condition, the talking face was attended, and in the other, attention was directed towards a leaf moving across the face. The strength of the McGurk effect was reduced in the latter case where attention was distracted from the face. This reduction was large, resulting in 29 percent point increase in auditory responses on average in the distracted attention condition. This

result means that audiovisual speech perception was strongly influenced by visual attention.

Previous studies have demonstrated that manipulation of crossmodal attention by instructing subjects to respond according to either audition or vision modifies audiovisual speech percepts (Dekle et al., 1992; Massaro, 1998; Summerfield & McGrath, 1984). The current study shows that shifts of intramodal attention can also produce changes in audiovisual speech percepts, and that these changes can be very prominent.

Our finding appears to disagree with that of Massaro (1984) who concluded that visual attention did not affect audiovisual speech processing. The difference in results is probably due to the different manipulations of attention. Massaro (1984) introduced an additional task (‘tell whether the mouth moved’) to focus a child’s attention to the visual speech, and found no change in speech percepts. In contrast, we manipulated visual attention by distracting it from the face, and found a clear attention effect. The baseline condition in both experiments was to watch the talking face. Since attention tends to be drawn to targets at fixation, visual attention was presumably quite strong in the baseline condition. Consequently, in Massaro’s design, there was little room for enhancement, and thus the effect of attention did not show. Meanwhile, in our design, the distraction of attention gave plenty of range for change so that the attention effect emerged.

Our finding that shifts of attention in one modality altered audiovisual speech percepts raises the question which processing stage is affected. One possibility is that

intramodal attention influences processing of that modality before auditory and visual information is integrated. In our experiment, this would mean that visual performance would be poorer when attention was distracted from the face, and consequently audiovisual responses would be less affected by vision. An alternative possibility is that attention influences the integration process. This would be the case if changes in attention altered audiovisual but not unisensory responses.

According to conventional statistical testing (ANOVA), unisensory recognition accuracy was largely unaffected by shifts in visual attention. Auditory performance was not affected, and visual recognition was unaffected for two out of three stimuli. Audiovisual percepts, however, did change with attention. That is, there were more auditory responses in the distracted attention condition for the McGurk stimuli. This was the case for all McGurk stimuli, including those in which the visual component was unaffected by changes in attention. Since audiovisual responses were affected even when unisensory responses were not, the implication is that attention influenced the audiovisual integration stage.

This is in line with Treisman's Feature Integration Theory (Treisman, 1986; Treisman & Gelade, 1980) which emphasizes the importance of attention in object perception. According to it, object features are processed preattentively, but focused attention is necessary to bind these features together. The theory has mostly been explored in the visual modality but it can be applied to multisensory perception. In the current case this would mean that attention is required to combine visual and auditory speech features. If attention is disrupted, audiovisual integration is less efficient even though the processing of unisensory features remains intact.

Another method for investigating which processing stage attention affects is a modelling approach. Massaro (1998) has used this kind of an approach utilizing his Fuzzy Logical Model of Perception (FLMP). The FLMP is a post-phonetic integration model with a statistically optimal integration rule, and it can be formulated as:

$$P(R_i \mid A,V) = \frac{P(R_i \mid A) \times P(R_i \mid V)}{\sum_{j=1}^{N} P(R_j \mid A) \times P(R_j \mid V)}$$

where $P(R_i \mid A,V)$ designates the probability of responding with the i^{th} response category, R_i, given auditory, A, and visual, V, stimulation.

According to Massaro, if the FLMP fits two different data sets equally well, it means that the same integration rule is applied to both sets, and any differences in audiovisual responses are due to differences in unisensory processing. He applied the model to audiovisual speech perception data collected with two sets of instructions: respond according to what you hear (i.e. attend to auditory speech) or respond according to what you see (i.e. attend to visual speech). Since the model accounted for his experimental results obtained under both auditory and visual instructions, he concluded that instructions and intentional set (attention) affect a unisensory processing stage before integration occurs.

We also adopted the modelling approach, and fitted the FLMP to our data. The free parameters of the model, the unisensory response probability density functions $P(R_i \mid A)$ and $P(R_i \mid V)$, were determined by numerically minimizing the root-mean-

squared error (RMSE) between the estimates and the data, as advocated by Massaro (1998). The model was fitted separately to both attention conditions. The FLMP fits were excellent, the RMSE being just 0.026 and 0.020 across all stimuli and response categories for 'Face' and 'Leaf' condition, respectively. The fits are shown together with the data for visual stimuli in Fig. 2, and for the McGurk stimuli in Fig. 4. Fig. 4 shows that, when auditory /p/ was presented with visual /k/ or /t/, the most common visually-influenced responses were so-called fusions or visual responses 't', and when auditory /k/ or /t/ was presented with visual /p/, the most common visually-influenced responses were combinations. Furthermore, the former type of stimuli give a stronger McGurk effect, i.e. more visually-influenced responses, than the latter type of combination stimuli, in line with previous studies with English stimuli and subjects (e.g. Green, Kuhl, Meltzoff, & Stevens, 1991; McGurk & MacDonald, 1976). Figures also show the attention effects. For example in Fig. 4, a decrease in visually-influenced responses was accompanied by an increase in auditory responses in the 'Leaf' condition. The model described the entire pattern of results very accurately, accommodating the changes in response distributions due to shifts in attention.

Figure 4 about here

In addition to small RMSE values, the error distributions of the FLMP fits for the two conditions were very similar ($p>0.92$; Wilcoxon rank sum test, two-tailed), confirming that the model described both 'Face' and 'Leaf' conditions equally well for the McGurk stimuli. This suggests that attention did not influence integration

since according to Massaro and Cohen (2000), if distraction of visual attention had affected the integration mechanism, the FLMP's optimal integration rule should have given poorer fits in the 'Leaf' condition.

The implication of the FLMP analysis is that the integration process remained optimal and thus unaffected by attention, in agreement with Massaro's (1998) report. Consequently, the changes in audiovisual percepts with attention would reflect changes in information processing before integration. In other words, unisensory response distributions should differ in the two attention conditions. However, statistical testing indicated that this was the case only for visual /t/. There was no effect of attention for auditory stimuli, or visual /k/ and /p/, and still there was a strong attentional effect for the McGurk combinations containing these stimuli.

A mathematical explanation for this conflict is the nonlinear nature of the FLMP which works so that even small changes in unisensory response probabilities can produce large changes when combined. Such nonlinear behaviour has been found in neurophysiological studies where auditory and visual stimuli give weak responses when presented alone but a multiplicative response when presented together in space and time. Neurons in the deep layers of the superior colliculus in the midbrain exhibit this type of response enhancement (see e.g. Stein, Wallace, & Meredith, 1995), and the enhancement has even been modelled by Bayes' rule (Anastasio, Patton, & Belkacem-Boussaid, 2000) which is equivalent to the integration rule of the FLMP. In humans, multiplicative response enhancements to audiovisual speech have been detected for example in the superior temporal sulcus of the cerebral cortex (Calvert et al., 2000). It thus seems that processes similar to those described by the FLMP may

well be implemented in the brain. If so, the brain must be utilizing response differences that are smaller than those that reached statistical significance in our behavioural experiments.

Our attempts at determining the level of attentional influence thus revealed a paradox in the current methods of analysis. Conventional statistics gave support to the hypothesis that attention affects audiovisual processing at the integration stage. In contrast, the modelling approach suggested that attention influences unisensory visual processing. Because of this discrepancy, it was not possible to determine the level at which attention influences processing. This paradox calls for further experimental investigations and development of the analytical methods currently in use.

Discussion thus far rests on the assumption that the level of attention was the same throughout each condition, ‘Face’ and ‘Leaf’. However, an alternative account for the lack of attention effect for visual stimuli, but a clear effect for audiovisual McGurk stimuli is that visual attention may have been stronger for the former than for the latter stimuli in the ‘Leaf’ condition. It might be that with unisensory visual stimuli, subjects divided or shifted their attention between the face and distractor since the speech perception task could not be performed without information from the face. But when audiovisual stimuli were presented, subjects could attend the distractor only since it was possible to do the speech task using just auditory information. To resolve this issue would also require additional studies into the cause of the audiovisual attention effect.

In summary, our results show that audiovisual speech perception was influenced by endogenous visual attention. This study provides evidence that attentional resources are needed in audiovisual speech perception, which has often been considered an automatic process (Calvert et al., 2000; Colin et al., 2002; Driver, 1996; Massaro, 1984; McGurk & MacDonald, 1976) in line with several other instances of multisensory perception that have been shown to be independent of attention, such as integration of emotion information from face and voice (Vroomen, Driver & de Gelder, 2001) and the ventriloquist effect (visual biasing of auditory location) (Bertelson, Vroomen, de Gelder, & Driver, 2000; Vroomen, Bertelson, & de Gelder, 2001). Contrary to this view, we believe that attention is an influential factor in audiovisual speech perception, and that attentional effects should be kept in mind for example as far as the ongoing search for neural and cognitive correlates of integration is concerned, where it would be important to ensure that subjects' attention to both visual and auditory speech stimuli is maintained during experiments in order to induce as strong multisensory interactions as possible.

REFERENCES

Anastasio, T. J., Patton, P. E., & Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation, 12,* 1165-1187.

Bertelson, P., Vroomen, J., de Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics, 62*(2), 321-332.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649-657.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology, 113,* 495-506.

Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception & Psychophysics, 51,* 355-362.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature, 381,* 66-68.

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics, 32,* 562-570.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12,* 423-425.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104*, 2438-2450.

Green, K. P. (1996). The use of auditory and visual information in phonetic perception. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications* (Vol. 150, pp. 55-77). Berlin: Springer.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50*, 524-536.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development, 55*, 1777-1788.

Massaro, D. W. (1998). *Perceiving talking faces*. Cambridge, Massachusetts: MIT Press.

Massaro, D. W., & Cohen, M. M. (2000). Tests of auditory–visual integration efficiency within the framework of the fuzzy logical model of perception. *Journal of the Acoustical Society of America, 108*, 784-789.

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

Palermo, R., & Rhodes, G. (2002). The influence of divided attention on holistic face perception. *Cognition, 82*, 225-257.

Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics, in press.*

Pesonen, J. (1968). *Phoneme communication of the deaf.* Helsinki: Suomalaisen Kirjallisuuden Kirjapaino Oy.

Smeele, P. M., Massaro, D. W., Cohen, M. M., & Sittig, A. C. (1998). Laterality in visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 1232-1242.

Sovijärvi, A. (1963). *Suomen kielen äännekuvasto (Phoneme descriptions of the Finnish language).* Jyväskylä: K. J. Gummerus Oy.

Stein, B. E., Wallace, M. T., & Meredith, M. A. (1995). Neural mechanisms mediating attention and orientation to multisensory cues. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 683-702). Cambridge, MA: M. I. T. Press.

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

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology, 36A*, 51-74.

Tiippana, K., Andersen, T. S., & Sams, M. (2001). *Visual attention influences audiovisual speech perception.* In Massaro, D.W., Light, J., Geraci, K. (Eds.) Proceedings AVSP2001, International Conference on Auditory-Visual Speech Processing, Santa Cruz, CA: Perceptual Science Laboratory (September 7-9, 2001, Aalborg, Denmark).

Treisman, A. (1986). Properties, parts and objects. In K. R. Boff & L. Kaufman & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance. Cognitive Processes and Performance* (Vol. 2, pp. 35.31-35.62). New York: John Wiley & Sons.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97-136.

Vroomen, J., Bertelson, P., & de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics, 63*(4), 651-659.

Vroomen, J., Driver, J., & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive, Affective, & Behavioral Neuroscience, 1*, 382-387.

ACKNOWLEDGEMENTS

We thank Prof. Dom Massaro for suggesting the FLMP analysis. Prof. Jouko Lampinen and Dr. Aki Vehtari have contributed to the implementation of the FLMP analysis.

This study was funded by the Academy of Finland (project number 43957) and the European Union Research Training Network "Multi-modal Human-Computer Interaction" (HPRN-CT-2000-00111).

Note: These results have been published in a preliminary form in the Proceedings of the Audio-Visual Speech Processing conference, 2001, Aalborg, Denmark (Tiippana, Andersen, & Sams, 2001).

FOOTNOTES

[1] The original leaf image was taken from Adobe Photoshop® picture gallery, with permission. Adobe and Photoshop are trademarks of Adobe Systems Incorporated.

[2] Greenhouse-Geisser adjusted *p*-values were also calculated in order to guard against false positives due to violations of the sphericity assumption in repeated-measures designs (Greenhouse & Geisser, 1959). Since the adjustment did not change any conclusions about statistical significance, unadjusted *p*-values are reported.

FIGURE CAPTIONS

Figure 1. A still image from /epe/ video clip at the peak of the mouth opening during the first /e/. The position of the distractor leaf is shown together with a trace marking the distractor path, which was always the same. The dashed line marks the path during speech and dotted line during the rest of the stimulus. The leaf has been outlined here for clarity. In the actual clips it was clearly visible without obscuring the details of the mouth because of colour differences.

Figure 2. Response distributions for unisensory visual stimuli when face or leaf was attended. Data in 'Face'/'Leaf' condition are shown as squares/triangles. Solid/dotted lines indicate the FLMP fits which are described in the Discussion. Standard deviations are also plotted but they are often smaller than the data markers.

Figure 3. Percentages of auditory responses and standard deviations for McGurk stimuli when face or leaf was attended.

Figure 4. Response distributions for McGurk stimuli when face or leaf was attended. Other details as in Fig. 2.

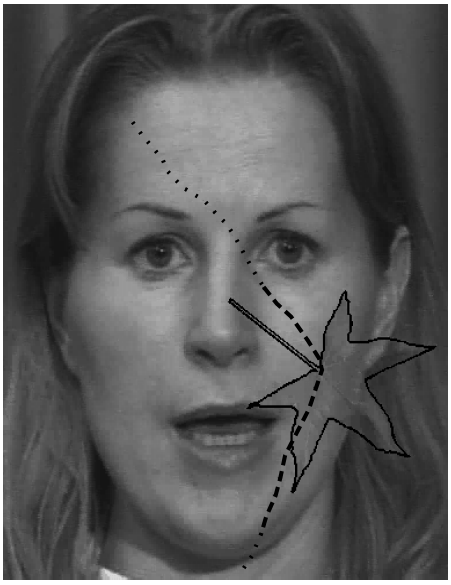

Figure 1.

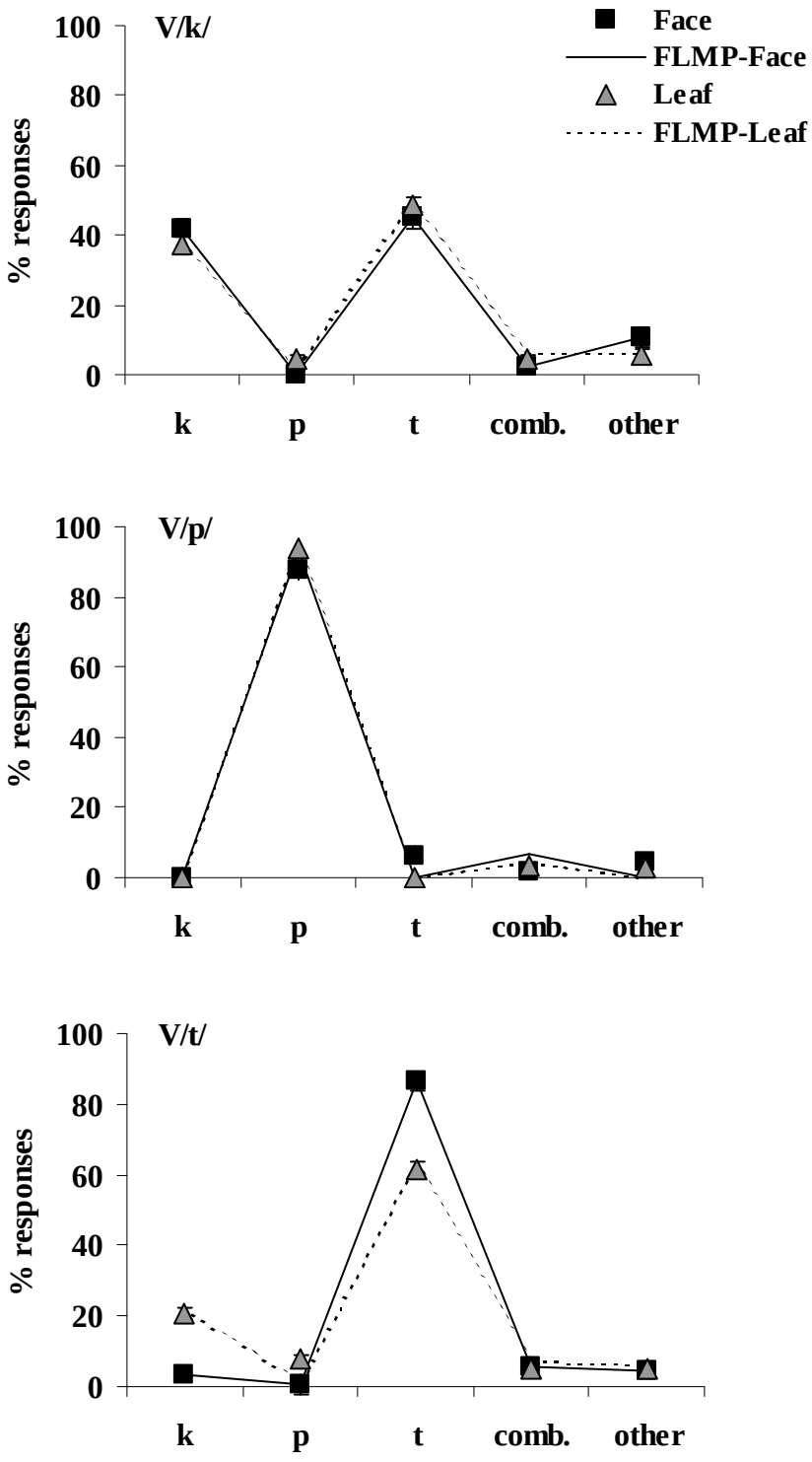
V/k/
Face
FLMP-Face
Leaf
FLMP-Leaf
% responses
100
80
60
40
20
0
k
p
t
comb.
other
V/p/
% responses
100
80
60
40
20
0
k
p
t
comb.
other
V/t/
% responses
100
80
60
40
20
0
k
p
t
comb.
other

Figure 2.

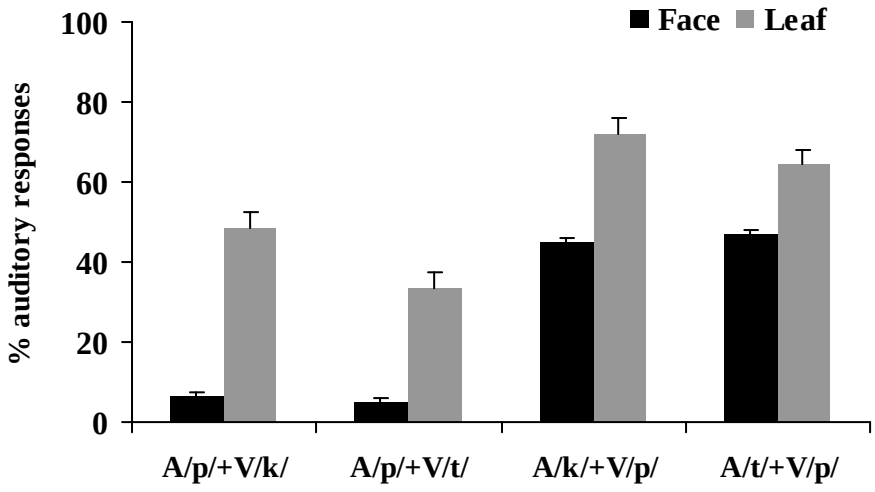
Face
Leaf
100
80
60
40
20
0
% auditory responses
A/p/+V/k/
A/p/+V/t/
A/k/+V/p/
A/t/+V/p/

Figure 3.

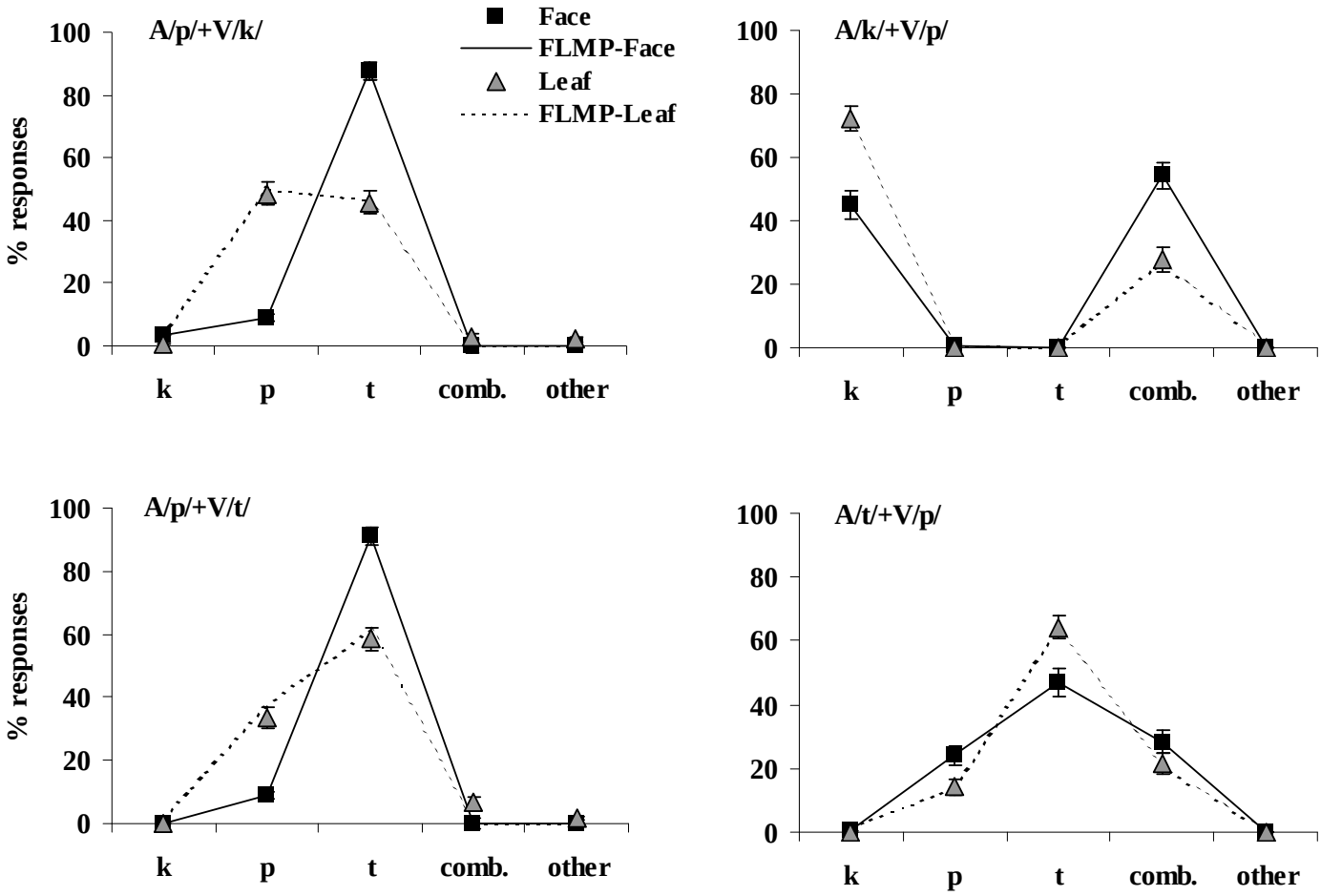
A/p/+V/k/
Face
FLMP-Face
Leaf
FLMP-Leaf
% responses
100
80
60
40
20
0
k
p
t
comb.
other
A/k/+V/p/
A/p/+V/t/
A/t/+V/p/

Figure 4.