

## VISUAL ATTENTION INFLUENCES AUDIOVISUAL SPEECH PERCEPTION

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

Helsinki University of Technology, Laboratory of Computational Engineering, P.O. Box 9400, 02015 HUT, Finland

### ABSTRACT

The purpose of this study was to investigate whether integration of audiovisual speech occurs automatically so that information from heard speech and seen articulatory movements of the talking face are combined without any voluntary effort. The McGurk effect, where seeing discrepant visual speech changes the auditory speech percept, was used as a tool since it reflects the extent of audiovisual integration. The McGurk effect was measured in two conditions which manipulated the subjects' attention to visual speech. In both conditions, the subjects' task was to attend to auditory speech and report what the talker said. In the 'Attend Face' condition, subjects were instructed to pay attention also to the talking face, presented in synchrony with auditory speech. In the 'Ignore Face' condition, subjects were instructed to ignore the talking face and to pay attention to a visual distractor presented in the same location as the face. The proportion of auditory responses was higher in the latter condition, indicating that the influence of visual speech was weaker when the face was not attended. This result suggests that integration of audiovisual speech is not entirely automatic. The mechanism underlying this attentional effect was investigated by fitting the Fuzzy Logical Model of Perception (FLMP) [1] to the results, and the good fit of the model implies that attention influences unimodal information processing before integration across modalities takes place.

### 1. INTRODUCTION

Integration of auditory and visual speech is said to occur when seeing articulatory gestures influences the auditory speech percept. This is most obvious in noisy conditions, where seeing the talking face improves recognition rate [2]. Another situation demonstrating audiovisual integration is the McGurk effect, where incongruent visual speech modifies the auditory speech percept, e.g., when an auditory syllable /ba/ is presented together with a visual syllable /ga/, the percept is typically /da/ [3].

There is some evidence that the integration process is quite automatic. Driver [4] has shown that merely fixating lip movements that are incongruent with the auditory signal can deteriorate speech recognition performance. Also, there are several casual reports

that the McGurk effect persists even when one knows how the stimuli have been manipulated. It thus seems clear that integration is not entirely under voluntary control. But is it then a totally automatic process that occurs as soon as a talking face is seen together with speech sounds? This is argued against by Massaro's [1] finding that by instructing subjects to respond according to either auditory or visual information only, the responses to audiovisual speech stimuli were biased towards the respective modalities, particularly when the other modality gave ambiguous information.

In the current study, we investigated the effect of visual attention on audiovisual integration. The McGurk effect was used as a research tool since its strength reflects the extent of audiovisual integration. The McGurk effect was measured in two conditions. In the baseline condition subjects attended to the talking face. In the distracted attention condition subjects ignored the face and attended to a visual distractor presented together with the face so that both were at fixation. If audiovisual integration is automatic, the McGurk effect should be similar in both conditions, i.e. independent of the attentional state. If audiovisual integration is under some voluntary control, the McGurk effect should be weaker when the face is unattended.

### 2. METHODS

The stimuli were video clips of a Finnish female talker uttering consonants /k/, /p/ and /t/ in /eCe/ context. Audiovisual stimuli were combined into an extended factorial design providing 6 unimodal stimuli (3 auditory, A, and 3 visual, V), 3 congruent audiovisual stimuli (A=V), and 6 incongruent audiovisual stimuli (A≠V). The height of the face on the computer screen was 7 deg, and the sound level was about 50 dB(A).

The experiment consisted of two parts, 'Attend Face' and 'Ignore Face' conditions, which are described below. In both conditions, each stimulus was presented ten times in a pseudo-random order, giving 150 observations per condition per subject.

The subjects were 17 native Finnish-speakers, aged 19-37, who reported normal hearing and had near visual acuity of at least 6/6. Their task was to write down which consonant(s) the talker uttered.

Results from 14 subjects (9 male) were included in the analysis. Three subjects were omitted since they never integrated, always giving only auditory responses. Half of the remaining 14 subjects started with the ‘Attend Face’ condition and half with the ‘Ignore Face’ condition.

**2.1. ‘Attend Face’ condition**

In the ‘Attend 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.

**2.2. ‘Ignore Face’ condition**

In the ‘Ignore Face’ condition, the subject was instructed to pay attention to the auditory speech and a visual distractor, and to ignore the face. The visual distractor was a partially transparent leaf flying across the talker’s face. The leaf started from the bottom middle as the clip started, going slightly to the right (shown in Fig. 1), and finishing at the talker’s temple on the left. The leaf was also spinning around slowly. The movement looked like the leaf was floating in the wind, and was easy to follow. When the talker spoke, the leaf was next to the mouth, overlapping it slightly without covering it, though. The subjects were asked to follow the leaf with their gaze.

This distractor task was chosen to provide a shift in visual object-based attention in a situation where both the face and the distractor were near fixation as the talker spoke. It has been shown that visual speechreading performance is little affected by eccentric viewing (only 4% decrease between 0 and 7 deg eccentricity for unsped speech) [5]. Therefore, we expected the effect of slighty eccentric viewing of the face to be negligible.



**Figure 1:** A still frame from a stimulus (first /e/ in /epe/) showing the face and the visual distractor (a floating leaf).

Eye movements were not monitored to check whether subjects really followed the leaf and not the face. However, if subjects had disregarded the instructions, the two conditions would have shown identical results which was not the case (see 3.2.2).

The presence or absence of the visual distractor did not have any effect on the results when the face was attended (tested on 3 subjects using a full set of 150 stimuli with and without distractor,  $r>0.98$ ). Thus, in the ‘Attend Face’ condition, there was no visual distractor.

**3. RESULTS**

The analysed responses formed five categories: k, p, t, combination, and other. Combination responses were those where both visual and auditory stimuli were reported as a cluster.

Differences between ‘Attend Face’ and ‘Ignore Face’ conditions were tested for statistical significance by performing analysis of variance (ANOVA) for repeated measures on the percentages of auditory responses (with the exception of unimodal visual stimuli for which visual responses were analysed).

**3.1. Unimodal and congruent audiovisual stimuli**

Table 1 summarizes unimodal auditory and visual, and congruent audiovisual results in ‘Attend Face’ and ‘Ignore Face’ conditions. Unimodal auditory and congruent audiovisual stimuli were almost perfectly recognized with no significant differences between conditions [A:  $F(1,13)=0.70, p=0.42$ ; AV:  $F(1,13)=1.41, p=0.26$ ].

Stimulus	Attend Face	Ignore Face
A/k/	97%	96%
A/p/	97%	99%
A/t/	100%	100%
V/k/	41%	37%
V/p/	88%	94%
V/t/	84%	61%
A/k/+V/k/	99%	94%
A/p/+V/p/	100%	100%
A/t/+V/t/	100%	99%

**Table 1:** Speech recognition of unimodal auditory and visual stimuli, and of congruent audiovisual stimuli: percentage of correct responses when face or distractor was attended.

Unimodal visual stimuli were more difficult to recognize, and performance was rather variable. ANOVA showed a main effect of stimulus [ $F(2,26)=5.49, p<0.001$ ]. Visual /k/ had the lowest

recognition rate, being commonly classified as /t/ (about 47% of responses). There was no significant effect of condition across stimuli ( $F(1,13)=1.44$ ,  $p=0.25$ ). However, the interaction between stimulus and condition was significant [ $F(2,26)=6.25$ ,  $p<0.01$ ]. This was because the effect of conditions was different for different stimuli. Analysis of simple effects showed no effect of condition for /k/ [ $F(1,13)=0.70$ ,  $p=0.42$ ] and /p/ [ $F(1,13)=0.66$ ,  $p=0.43$ ] but a significant effect for /t/ [ $F(1,13)=5.32$ ,  $p<0.05$ ].

### 3.2. Incongruent audiovisual stimuli

Incongruent audiovisual stimuli formed two groups: so-called auditory dominance stimuli and visually influenced, McGurk stimuli.

#### 3.2.1. Auditory dominance stimuli

Incongruent combinations of /k/ and /t/ showed auditory dominance, giving almost entirely auditory responses as shown in Table 2. Consequently there was no effect of attention ( $F(1,13)=0.33$ ,  $p=0.57$ ).

AV stimulus	Attend Face	Ignore Face
A/k+V/t/	93%	96%
A/t+V/k/	99%	99%

**Table 2:** Incongruent audiovisual speech with auditory dominance: percentage of auditory responses when face or distractor was attended.

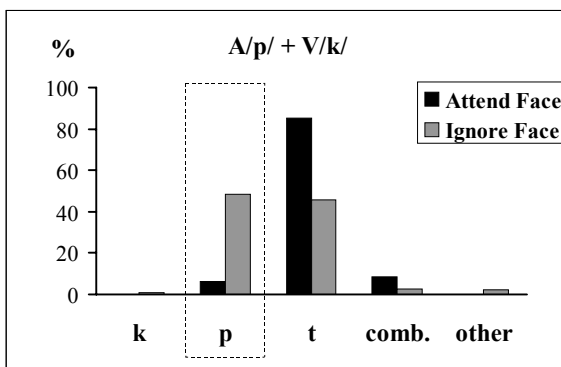
#### 3.2.2. McGurk stimuli

The McGurk stimuli provided an index for the extent of audiovisual integration so that the greater the visual influence, i.e. the lower the percentage of auditory responses, the greater the integration effect. Figs 2-5 show results for the McGurk stimuli in the two attention conditions.

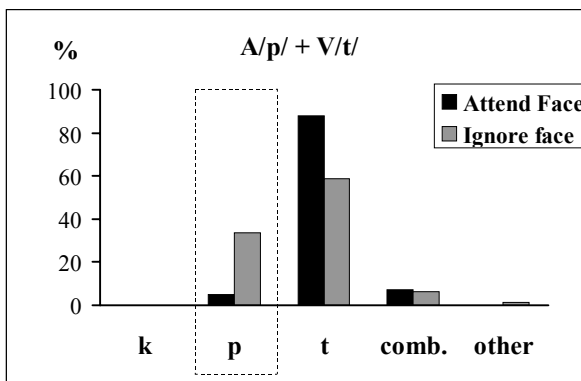
Fig. 2 shows results for a typical fusion stimulus A/p/+V/k/, where non-auditory responses mainly consisted of /t/ percepts which can be considered kind of a fusion between auditory /p/ and visual /k/.

In Fig. 3 non-auditory responses to stimulus A/p/+V/t/ were mostly visually-based ‘t’s.

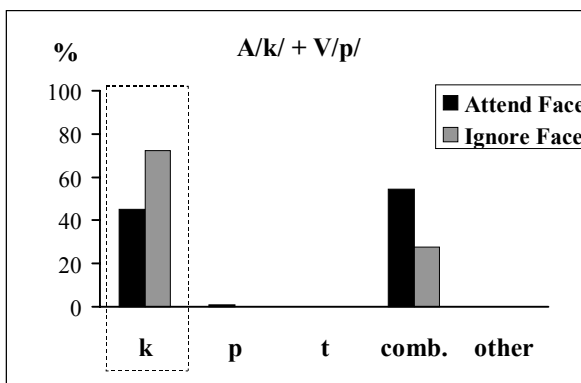
Figs 4 and 5 show results for combination stimuli where visual /p/ was presented with auditory /k/ or /t/. Typical non-auditory responses for these stimuli were combinations of A and V, i.e. /pk/ and /kp/ or /pt/ and /tp/. In Fig. 5 visual responses /p/ were also quite frequent.



**Figure 2:** Response distribution for McGurk fusion stimulus: auditory /p/ and visual /k/. Black bars show the percentages of responses when the talking face was attended, and grey bars when the visual distractor was attended. Responses were grouped as ‘k’, ‘p’, ‘t’, ‘combination’ (comb.) which includes combinations of auditory and visual stimuli, and ‘other’ which comprises all other responses. Auditory responses are highlighted by the dotted box.



**Figure 3:** Response distribution for McGurk visual dominance stimulus: auditory /p/ and visual /t/. Other details as in Fig. 2.



**Figure 4:** Response distribution for McGurk combination stimulus: auditory /k/ and visual /p/. Other details as in Fig. 2.

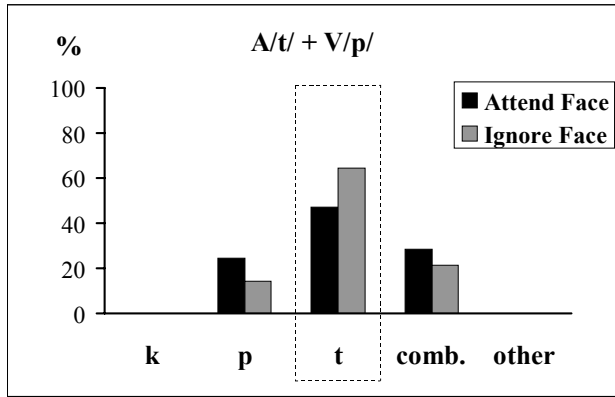


Figure 5: Response distribution for McGurk combination stimulus: auditory /t/ and visual /p/. Other details as in Fig. 2.

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 main effect of stimulus occurred because there were fewer auditory responses for fusion-type stimuli (Figs 2 and 3) than for combination stimuli (Figs 4 and 5), indicating that visual influence was stronger for the former. The main effect of condition was due to the fact that there were significantly more auditory responses when attention was directed towards the distractor and the face was ignored than when the face was

attended. Thus, the effect of visual speech was smaller when attention was distracted from the face. Interaction was not significant [ $F(3,39)=7.84, p=0.30$ ] indicating that distraction of attention had a similar effect on all stimuli.

### 3.3. FLMP fitting

The Fuzzy Logical Model of Perception [1] was fitted to the results to investigate the possible mechanism underlying the reduced integration effect for McGurk stimuli in the distracted attention condition. The fundamental idea underlying the FLMP is that the Bayesian integration rule always remains unchanged, and any changes in the integration process affect information processing before unimodal inputs are integrated into an audiovisual percept.

Another alternative would be that attention affects the actual audiovisual integration mechanism. In such a case, the FLMP would describe the results poorly.

The FLMP fits were excellent (RMS error 2.1% across conditions). As expected, all conditions where attention had no effect were well described by the model. In addition, unimodal visual results and data for the McGurk stimuli were also well fitted, as demonstrated by Fig. 6. This suggests that the effect

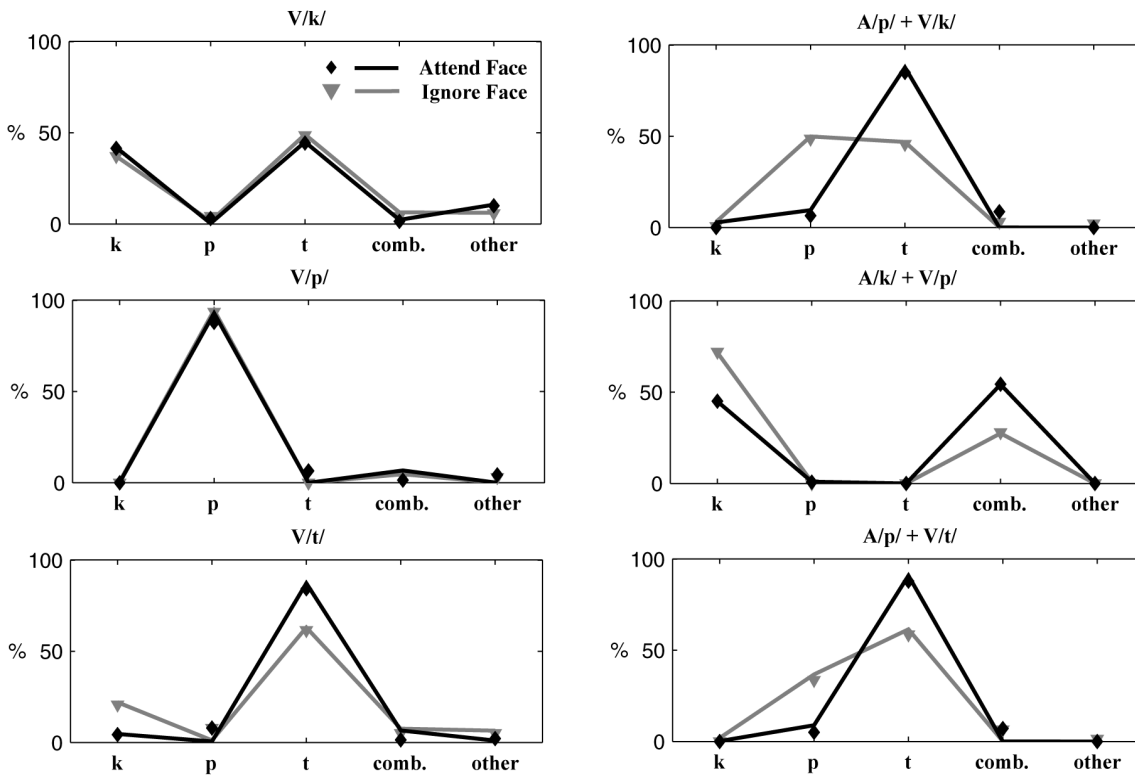


Figure 6: FLMP fits to unimodal visual results (left) and to McGurk stimulus results (right) in the two attention conditions. A/p+V/t is not shown. Data are shown by symbols, fits by lines.

of attention occurs early in the process, before audiovisual integration takes place.

#### 4. DISCUSSION

To our knowledge, this is the first study to explicitly address the issue of whether distraction of endogenous visual attention in one locus in space can influence integration of audiovisual speech. It is often implicitly assumed that all that is needed for integration to occur is to look at a talking face while listening to speech. However, our main finding shows that the McGurk effect became weaker when visual attention was directed towards a distractor stimulus, even though the talking face was clearly visible in the fixated area. This implies that the process of audiovisual integration is influenced by the subjects' attentional state.

Our results are in agreement with Massaro's [1] finding that instructing subjects to base their responses to audiovisual stimuli purely on either audition or vision biases their responses towards the instructed modality. The results were fitted with the FLMP, and in line with Massaro's study, the analysis suggested that attention modulates the information available to each modality before integration takes place.

It is remarkable that the ANOVA and FLMP can suggest different interpretations. Given the small, non-significant changes between conditions in ANOVA for all unimodal responses but for those of visual /t/, one might assume that it is the mechanism of integration that is affected by attention. This assumption is, however, based on linearity of the integration rule. In contrast, the FLMP is non-linear by the normalization term (see e.g. [6] in this volume). Because of this nonlinearity even minor changes in unimodal responses between conditions suffice for the FLMP to accommodate the bimodal responses in both attention conditions. According to the FLMP, the mechanism of integration is unaffected by the subject's attentional state, the changes in the unimodal response probabilities – however small – being the underlying cause for the changes in perception.

Not surprisingly, manipulation of visual attention had no effect on congruent audiovisual speech stimuli whose perception could be accounted for by the auditory stimulus alone. However, it could be asked why two of the incongruent stimuli, i.e. combinations of /k/ and /t/ did not show visual influence. We propose that this is because visual /k/ and /t/ share many features and are thus easily confused. Thus, when combined with a clear auditory signal, the response is determined by the auditory component.

In the ambiguous unimodal situation, subjects preferred to respond /t/ over /k/ for both visual stimuli. This could be because /t/ occurs about twice as frequently as /k/ in the Finnish language [7], and perhaps also because there were more stimuli producing a /t/ percept in the current experiment (41% /t/ vs 27% /k/ of total responses).

The predominance of 't'-responses to incongruent stimuli with auditory /p/ presented together with visual /k/ or /t/ can also be explained in line with the above proposal that the preferred visual response was 't'. In the framework of the FLMP, interpretation /t/ wins because it has the highest audiovisual support value given by the product of strong visual and intermediate auditory support.

To conclude, since distraction of visual attention from the talking face reduced the influence of visual speech in audiovisual speech perception, integration of auditory and visual speech involves processes that are not entirely automatic. These processes are likely to influence information processing at a stage preceding cross-modal integration.

#### 5. REFERENCES

1. Massaro, D.W., *Perceiving talking faces*, MIT Press, Cambridge, Massachusetts, 1998.
2. Erber, N.P. "Interaction of audition and vision in the recognition of oral speech stimuli", *J. Speech Hear. Res.* 12: 423-425, 1969.
3. McGurk, H., and MacDonald, J. "Hearing lips and seeing voices", *Nature*, 264: 746-748, 1976.
4. Driver, J. "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading", *Nature*, 381: 66-68, 1996.
5. Smeele, P. M. T., Massaro, D. W., Cohen, M. M. and Sittig, A. C. "Laterality in visual speech perception", *J. Exp. Psychol. Hum. Percept. Perform.*, 24: 1232-1242, 1998.
6. Andersen, T. S., Tiippana, K., Lampinen, J. and Sams, M. "Bayesian Modeling of Audiovisual Speech Perception in Noise", *Audiovisual Speech Perception Conference Proceedings*, 2001.
7. Pääkkönen, M. "A:sta ö:hön. Suomen yleiskielen kirjaintilastoja", *Kielikello*, 1: 3-8, 1991.

*Acknowledgement:* 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).