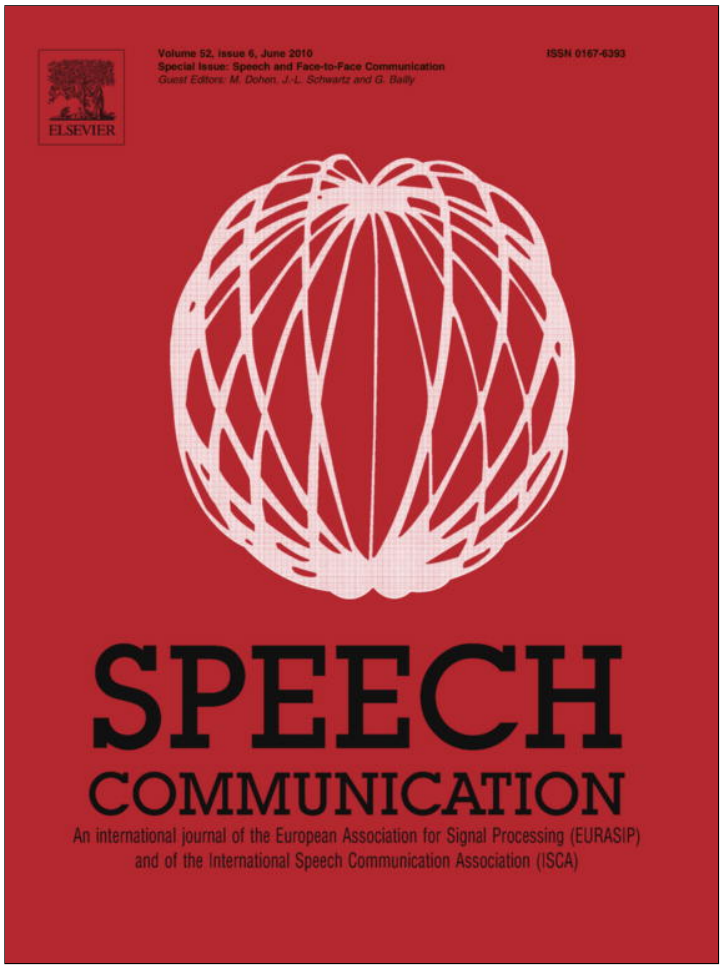

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 525–532
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

The word superiority effect in audiovisual speech perception

Mathilde Fort [a,*], Elsa Spinelli [a,b], Christophe Savariaux [c], Sonia Kandel [a,b]

[a] Université Pierre Mendès France, Laboratoire de Psychologie et NeuroCognition (CNRS UMR 5105), BP 47, 38040 Grenoble Cedex 9, France
[b] Institut Universitaire de France, 103, bd Saint-Michel, 75005 Paris, France
[c] Université Stendhal, GIPSA-lab, Dpt. Parole et Cognition (CNRS UMR 5216), BP 25, 38040 Grenoble Cedex 9, France

Received 30 March 2009; received in revised form 29 December 2009; accepted 9 February 2010

Abstract

Seeing the facial gestures of a speaker enhances phonemic identification in noise. The goal of this study was to assess whether the visual information regarding consonant articulation activates lexical representations. We conducted a phoneme monitoring task with word and pseudo-words in audio only (A) and audiovisual (AV) contexts with two levels of white noise masking the acoustic signal. The results confirmed that visual information enhances consonant detection in noisy conditions and also revealed that it accelerates the phoneme detection process. The consonants were detected faster in AV than in A only condition. Furthermore, when the acoustic signal was deteriorated, the consonant phonemes were better recognized when they were embedded in words rather than in pseudo-words in the AV condition. This provides evidence indicating that visual information on phoneme identity can contribute to lexical activation processes during word recognition.

Keywords: Audiovisual speech; Lexical access; Speech perception in noise; Word recognition

1. Introduction

When we speak with someone, most of the time, we are in a face-to-face situation. Except when speaking on the phone or hearing the radio, conversations take place in an audiovisual context. Moreover, the environment in which these conversations take place is often noisy. Several studies have shown that the information on the speaker's orofacial gestures enhances phoneme identification, especially in noisy situations (Benoît et al., 1994; Erber, 1969; Sumby and Pollack, 1954; see Green, 1998 for a review). In French, Benoît et al. (1994) showed that under noisy conditions, consonant and vocalic phonemes embedded in VCVCVC nonsense words were better identified in audiovisual than in auditory only presentations. We may thus assume that visible orofacial gestures boost phonemic units' activation during audiovisual speech perception in a noisy environment. The purpose of the study was to assess whether visual information not only enhances phoneme identification in noise but also contributes to the process of word recognition.

Most of the researches in the field of spoken word recognition studied lexical access in an auditory context (Cutler et al., 1987; Frauenfelder et al., 1990; Ganong, 1980; Samuel, 1981; Warren, 1970). Findings such as the word superiority effect (Cutler et al., 1987), Ganong effect (Ganong, 1980) or phonemic restoration (Samuel, 1981; Warren, 1970), suggest that lexical information influences phoneme perception. For example, with a phoneme monitoring task, Cutler et al. (1987) observed that a consonant (e.g. /b/) was detected faster in a word (e.g. ***b**elle*, i.e. beautiful) than in a pseudo-word (e.g. ***b**erre*). This "word superiority effect" suggests that lexical activation can influence phoneme perception even in situations where the acoustic signal is clear.

* Corresponding author. Address: Université Pierre Mendès France, Laboratoire de Psychologie et NeuroCognition, BP 48, 38040 Grenoble Cedex 9, France. Tel.: +33 4 76 82 56 30; fax: +33 4 76 82 78 34.
E-mail addresses: mathilde.fort@upmf-grenoble.fr (M. Fort), elsa.spinelli@upmf-grenoble.fr (E. Spinelli), christophe.savariaux@gipsa-lab.inpg.fr (C. Savariaux), sonia.kandel@upmf-grenoble.fr (S. Kandel).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.02.005

One of the first studies investigating word recognition processing in an audiovisual context was conducted in Finnish (Sams et al., 1998) with a McGurk paradigm (McGurk and MacDonald, 1976). This effect occurs when an acoustic stimulus /ba/ is presented simultaneously with the articulation of /ga/ in the video signal. Most of the time, it results in the perception of /da/. Numerous studies have replicated these findings, suggesting that during audiovisual speech perception, acoustic and visual signals integrate and may even produce perceptual illusions (see Colin and Radeau, 2003 for a review). On this basis, Sams et al. (1998) displayed an auditory word (e.g. ***p**annu*, stove) simultaneously with another word that was presented visually (e.g. ***k**annu*, pitcher). The audiovisual integration should result in the perception of a pseudo-word (e.g. ***t**annu*). In another condition, the authors displayed an auditory pseudo-word (e.g. ***p**iili*) simultaneously with a visual presentation of another pseudo-word (e.g. ***k**iili*). The audiovisual integration should result in the perception of a word (e.g. ***t**iili*, brick). The results revealed that the McGurk effect was not stronger for word responses than for pseudo-word responses. In other words, there was no word superiority effect. The authors concluded that lexical knowledge did not bias audiovisual speech perception at the stage of phonetic perceptual processing.

Brancazio (2004) argued however that the reason why Sams and colleagues did not observe a lexical effect was, among others, that in their study the stimuli differed as to various parameters other than lexical status. To justify his assessment, Brancazio examined this issue avoiding the potentially confounding variables in Sams et al.'s experiment. He combined the McGurk effect with a Ganong paradigm (Ganong, 1980). In this paradigm the participants had to identify a phoneme /t/ or /d/ that varied along a synthesized t–d continuum. When auditory stimuli in the continuum formed words (e.g. ***t**ask*) and pseudo-words (e.g. ***d**ask*), the proportion of /t/ response was systematically higher than /d/. There was a word superiority effect indicating that phoneme perception was biased in favour of words. In Brancazio's study carried out in English, the participants had to identify /b/ and /d/ in two conditions. In the first one, a word was displayed in the acoustic signal (e.g. ***b**eg*) dubbed into a visual pseudo-word (e.g. ***d**eg*). In the second one, a visual word (e.g. ***d**esk*) was dubbed into an acoustic pseudo-word (e.g. ***b**esk*). The results showed that the lexical bias was stronger in the visual word condition than in the auditory word condition. This suggests that lexical activation not only influences auditory perception but also visual processing during word recognition.

A recent study (Barutchu et al., 2008) provides evidence in line with Brancazio's research. They investigated lexical influences on the McGurk effect in words and pseudo-words. For instance, in the word condition, the auditory word ***b**et* was presented with the visual word ***g**et*. In the pseudo-word condition, a visual pseudo-word ***g**ez* was dubbed into an auditory pseudo-word ***b**ez*. They observed more visual responses – i.e. consistent with the visual signal (***g**et* or ***g**ez*) – for words than for pseudo-words. Consequently, these results also suggest that visual speech processing can be influenced by lexical knowledge.

In sum, Barutchu et al. (2008) and Brancazio (2004) showed that visual information on phoneme identity contributed to lexical access whereas the results of Sams et al. (1998) did not yield any word superiority effect. All these studies used the McGurk paradigm which placed the participants in a situation of perceptual conflict. This may introduce ambiguity in phoneme identification because visual and auditory information are not congruent. Thus, to avoid the conflict between auditory and visual information in our research we examined this issue with another paradigm that is widely used to study the auditory spoken word recognition: the phoneme monitoring task.

To our knowledge, only a few studies have investigated word recognition processes in audiovisual speech perception without using the McGurk paradigm (Buchwald et al., 2009; Kim et al., 2004). First, Kim et al. (2004) used a priming procedure combined with a naming task. They studied whether the presentation of a speaker's orofacial gestures as a prime (visual speech prime without the auditory information) would facilitate the processing of a written target. In a first condition, they displayed word primes in visual speech that was followed by a written target which could be identical (e.g. *back*/*back,* identical condition) or unrelated (e.g. *sharp*/*back,* unrelated condition). In a second condition, they displayed pseudo-word primes in visual speech and the following written target could be identical (e.g. *scay*/*scay*) or unrelated (e.g. *nunth*/*scay*). When the stimuli were words, the authors found a facilitatory priming effect in the identical condition compared to the unrelated condition. They did not observe any facilitatory or inhibitory effect when the stimuli were pseudo-words. With the same paradigm, a recent research reported that participants identified spoken words in noise more accurately when the words were preceded by a visual speech prime of the same word compared with a control condition (Buchwald et al., 2009). They also observed that most of the incorrect responses were phonetically close to the target-words. These findings show that the information in the visual speech prime influences both correct and incorrect identifications. These two studies suggest that the information in the visual speech prime contributes to lexical processing by activating the linguistic forms that match the visual signal. In more general terms – as in Brancazio (2004) and Barutchu et al. (2008) – these studies suggest that visual information contributes to lexical access in audiovisual speech perception.

The purpose of the present research was to examine whether the visual cues that contribute to phoneme identification (Benoît et al., 1994) are also involved in the activation of lexical representations during word recognition process. We also used an experimental paradigm where the visual and auditory information were congruent. The

participants were placed in a noisy environment, which is a situation found in everyday life.

We used a phoneme monitoring task which is a paradigm widely used to investigate the lexical influences on auditory speech processing (Gow, 2003; LoCasto et al., 2007; see also Connine and Titone, 1996 for a review). This task not only provides information on correct identification but also reaction time measures that shed some light on the online process of word recognition.

We conducted a phoneme monitoring task with words and pseudo-words displayed in audiovisual (AV) and auditory alone (A) situations. The stimuli were mixed with noise in the acoustic signal to avoid ceiling effects on correct detection scores. We expected to replicate Benoît et al.'s (1994) results with words: correct responses should be higher in AV than in A especially in noisy conditions. Following the rationale in the word recognition domain we should observe higher scores for words than for pseudo-words (i.e. a word superiority effect). Finally, assuming that the information provided by the speaker's orofacial gestures contributes to the activation of lexical units during word recognition, we predicted that the AV advantage would be higher for words than for pseudo-words.

2. Experiment 1

2.1. Method

2.1.1. Participants

Eighty-one native French speakers ranging in age from 18 to 51 years (mean age = 23 years) participated in the experiment. They all had normal or corrected-to-normal vision and reported no auditory disorders.

2.1.2. Stimuli

The stimulus set was composed of 74 disyllabic word/pseudo-word pairs. Thirty-four pairs were target-present trials (i.e. the target phoneme was in the carrier item, see Appendix) and 40 pairs were target-absent trials. The stimuli in each word/pseudo-word pair were identical except for the last vowel (e.g. /ʃa**p**o/, hat vs. /ʃa**p**y/).

2.1.2.1. Target-present trials. For the 34 pairs of target-present trials (or carrier items), the target phoneme was located at the onset of the second syllable (e.g. the target /**p**/ in /ʃa**p**o/ or /ʃa**p**y/) so that each member of a pair activates the same number of lexical candidates until the target phoneme appears (Marslen-Wilson, 1990). Having the target at the end of the stimulus, instead of the beginning, also increases the probability of observing a strong lexical effect (Frauenfelder et al., 1990). We used seven consonant target phonemes: three labials (/p/, /f/, /v/) and four dentals (/d/, /t/, /s/, /z/). In the carrier item, the following vowel could be either rounded (/o/, /u/, /y/, /œ/) or stretched (/i/, /e/). Half of the word/pseudo-word pairs were contrasting for articulatory gestures. One member of the pair could end in a rounded vowel whereas the other ended in a stretched vowel (e.g. /tʀupo/, flock vs. /tʀupi/). In the other half both members of each pair ended in rounded or stretched vowels (e.g. /ʃapo/ vs. /ʃapy/). The mean word frequency for the carrier words was 45.88 pm (LEXIQUE, New et al., 2001). Half of them were considered as frequent ($F > 10$ occurrences per million) and the other half were not ($F < 10$ occurrences per million).

2.1.2.2. Target-absent trials. The 40 word/pseudo-word pairs of target-absent trials (e.g. /tɔʀty/, turtle vs. /tɔʀti/) were constructed using the same phonemes as described for the carrier items. However, these pairs were always preceded by a non-matching target phoneme (e.g. the target /**p**/ in /tɔʀty/ or /tɔʀti/). The mean word frequency for the target-absent words was 42.99 (New et al., 2001). Half of them were considered as frequent ($F > 10$ occurrences per million) and the other half were not ($F < 10$ occurrences per million).

2.1.2.3. Stimuli recording. The stimuli were recorded in a sound proof room by a trained male native French speaker with a green background. Only the head and top part of the neck of the speaker was visible. He had to start making each utterance with his mouth closed and was instructed to avoid blinking during the stimulus pronunciation. A tri-CCD SONY DXC-990P camera and an AKG C1000S microphone were used to make the recording. The recording was digitalized with the Dps Reality v 3.1.9 software to obtain mpeg video files. The soundtrack extracted from the video was used for the auditory only (A) condition in order to have exactly the same acoustic signal in the A and AV conditions.

Target phonemes were recorded in a sound proof room with a Marantz PMD 670 digital recorder in order to obtain wave files. They were pronounced by a 22-year-old female native French speaker in a schwa context (e.g. target /pə/ for the carrier items /ʃapo/ and /ʃapy/). Thus, two different speakers were chosen to produce the speech material (target phonemes and carrier items) to make sure that phoneme detection would not be due to speaker specificity.

We used Matlab 7.1 software to generate the noise and to add it to each spoken utterance. We used two Signal to Noise Ratios[1]: −9 dB vs. −18 dB. As each utterance energy was dependent on its vowel and consonant type (e.g., plosive, fricative) we calculated the mean strength for each stimulus and then added white noise so that the stimuli could have the same Signal to Noise Ratio throughout the whole presentation.

The stimuli were spread out over four experimental lists corresponding to the four presentation conditions:

[1] The Signal to Noise Ratio, often written S/N or SNR, is a measure of signal strength related to background noise. The ratio is usually measured in decibels (dB). We used the following formula: $SNR = 20 \log_{10}(V_s/V_n)$ in which V_s and V_n are respectively the original signal amplitude and the noise amplitude.

A −9 dB; A −18 dB; AV −9 dB; AV −18 dB. Each list contained 8 or 9 pairs of target-present trials and 10 pairs of target-absent trials. Consequently, each target-present or target-absent trial was presented only once to each participant. The presentation condition of each list was counterbalanced between the participants.

2.1.3. Procedure

The participants were tested individually. They sat at 50 cm from a LCD screen (Neovo 17 X-17A) in a darkened sound proof room. Video stimuli were presented at 25 frames/s. The auditory component of the stimuli was provided at a 44,100 Hz sampling rate by two SONY SRS-88 speakers located on both sides of the screen. The experiment was performed using E-Prime 2.0 software (Psychological Software Tools, Pittsburgh, PA). The participants were given a set of oral instructions explaining that they would first hear a consonant target phoneme and then a word or a pseudo-word (carrier) in which they had to detect this target. They were told that the target phoneme could or could not be in the carrier utterance. A Go/No Go response task was employed: participants had to press the space bar of a keyboard as quickly as possible when they heard the target phoneme in the carrier item and do nothing if they did not hear it, using only one hand to give their answer. The participants were told to detect the consonant target phoneme regardless of its orthographic representation.[2]

For each participant, half of the carrier items appeared in AV with the video of the speaker moving (half at −9 dB, half at −18 dB). In the other half (the A condition) the stimuli were only displayed auditorily (half at −9 dB, half at −18 dB). The experiment was divided into two blocks, namely A and AV. The block order was counterbalanced across participants. Between each block, a black screen informed the participants of a change in the presentation modality. For the AV condition participants were told to watch and listen carefully to the stimuli in order to avoid focusing on one modality more than another (cf. Amano and Sekiyama, 1998; Tiippana et al., 2004). Within each block, the participants perceived the first part of the stimuli in one SNR condition (e.g. −9 dB) and the second part in the other SNR condition (e.g. −18 dB). The order of each SNR condition was counterbalanced across participants. Moreover, half of the items within the four conditions contained the target phoneme (target-present trials) and half did not (target-absent trials). Within each condition, the order of the stimuli was randomised. A 10 stimuli-long training session preceded the test.

2.2. Results and discussion

Mean response latencies and percentages of correct phoneme detection were calculated for each participant and each item pair. Two participants were removed from the analyses because they did not respond in the A condition at −18 dB. A 2 (modality: A vs. AV) × 2 (lexical status: word vs. pseudo-word) × 2 (Signal to Noise Ratio: −9 dB vs. −18 dB) within participants ANOVA was conducted by both participants (*F1*) and items (*F2*). We discarded from the analyses the data that were for every condition above or below two standard-deviations (SD) from the mean (2.3% of the data).

2.2.1. Percentage of correct phoneme detection

Table 1 presents the percentage of correct phoneme detection for words and pseudo-words in A and AV for the two noise conditions (−9 dB vs. −18 dB).[3] The analyses revealed a strong AV advantage, $F1$ $(1, 78) = 180.87$, $p < .001$; $F2$ $(1, 33) = 46.68$, $p < .001$. The analyses also showed that the scores were higher at −9 dB than at −18 dB, $F1$ $(1, 78) = 199.34$, $p < .001$; $F2$ $(1, 33) = 67.54$, $p < .001$. These results replicate Benoît et al.'s (1994) findings. The performance was also enhanced when the target phonemes were embedded in words, $F1$ $(1, 78) = 8.23$, $p < .005$; $F2$ $(1, 33) = 3.06$, $p < .01$.

The interaction between lexical status and modality was significant, $F1$ $(1, 78) = 10.34$, $p < .005$; $F2$ $(1, 33) = 6.96$, $p < .05$. The word advantage for phoneme detection was greater in AV than in A, $F1$ $(1, 78) = 23.83$, $p < .001$; $F2$ $(1, 33) = 10.21$, $p = .01$ and both $Fs < 1$, respectively. There was also a significant interaction between lexical status and noise, $F1$ $(1, 78) = 6.95$, $p = .01$; $F2$ $(1, 33) = 4.69$, $p < .05$. In AV, planned comparisons showed that the word superiority effect was higher at −18 dB than at −9 dB, $F1$ $(1, 78) = 24.72$, $p < .001$; $F2$ $(1, 33) = 11.48$, $p < .01$ and $F1$ $(1, 78) = 4.37$, $p < .05$; $F2$ $(1, 33) = 2.54$, $p = .12$, respectively.

2.2.2. Response latencies

Table 2 presents the response latencies for words and pseudo-words in A and AV for the two noise conditions (−9 dB vs. −18 dB).

[2] In French, for instance, the phoneme /f/ can be written "f" or "ph".

[3] To make sure that the participants did not develop any response strategies, we also computed a d' for each stimulus pair, using this formula $d' = z\,(\mathrm{CD}) - z\,(\mathrm{FA})$ in which z represents the inverse of the normal cumulative distribution and CD and FA refers respectively to the mean probability of correct phoneme detection and false alarms. A 2 (modality: A vs. AV) × 2 (Signal to Noise Ratio: −9 dB vs. −18 dB) × 2 (lexical status: word vs. pseudo-word) within participants ANOVA was conducted by participants. We replicated the results obtained by analysing only the correct detection scores. The analyses on d' revealed a strong AV advantage, F $(1, 78) = 212.7$, $p < .001$. The scores were also higher at −9 dB than at −18 dB, F $(1, 78) = 119.3$, $p < .001$. There was a word superiority effect, F $(1, 78) = 10.7$, $p < .005$. The interaction between lexical status and modality was significant, F $(1, 78) = 5.5$, $p < .05$. Planned comparisons revealed that the word advantage was greater in AV than in A at −9 dB, $F\,(1, 78) = 4.91$, $p < .05$ in AV vs. $F\,(1, 78) < 1$ in A; and at −18 dB, $F\,(1, 78) = 6.2$, $p = .01$ in AV vs. $F\,(1, 78) < 1$ in A. There was also a significant interaction between modality and noise, F $(1, 78) = 10.1$, $p < .005$, suggesting that the AV advantage for phoneme detection was greater at −18 dB than at −9 dB.

Table 1
Percentage of correct phoneme detection as a function of modality, Signal to Noise Ratio (in Decibels, dB) and lexical status, in Experiment 1. Numbers in parentheses represent the standard deviation.

Modality of presentation	Words	Pseudo-words	Word superiority effect
−9 dB			
Audio alone	68.2 (20.4)	69.4 (19.26)	−1.2
Audiovisual	90.9 (10.54)	88.1 (11.9)	2.8*
−18 dB			
Audio alone	50.1 (17.38)	49.3 (18.16)	0.8
Audiovisual	76.1 (19.13)	65.3 (18.32)	10.8**

* Word superiority effect significant by participants.
** Word superiority effect significant by participants and by items.

Table 2
Mean response latencies (in ms), as a function of modality, Signal to Noise Ratio (in Decibels, dB) and lexical status in Experiment 1. Numbers in parentheses represent the standard deviation.

Modality of presentation	Words	Pseudo-words	Word superiority effect
−9 dB			
Audio alone	808.3 (108.4)	808.9 (204.5)	0.6
Audiovisual	671.3 (121)	673.8 (95.5)	2.5
−18 dB			
Audio alone	844.6 (244.1)	899.2 (175.8)	54.6
Audiovisual	734.9 (113.7)	759.1 (138.2)	24.2

The analyses revealed a significant main modality effect in favour of the AV condition, $F1$ (1, 78) = 32.27, $p < .001$; $F2$ (1, 33) = 62.32, $p < .001$. There was a significant main effect of the Signal to Noise Ratio, $F1$ (1, 78) = 31.01, $p < .001$; $F2$ (1, 33) = 8.51, $p < .01$. The participants were faster at detecting a consonant phoneme at −9 dB than at −18 dB. Contrary to our expectations, the lexical effect was not significant, $F1$ (1, 78) = 1.02, $p = .27$; $F2$ (1, 33) = 1.44, $p = .23$. No interaction was significant, all $F1$ (1, 78) < 1.

In sum, Experiment 1 revealed that the participants were faster and had higher scores in AV than in A only conditions. They were also faster and performed better when the Signal to Noise Ratio was at −9 dB than at −18 dB. This is in line with Benoît et al.'s (1994) study conducted with non-word stimuli. More interesting for the purpose of our study was that the scores were higher when the participants had to detect the consonant phonemes embedded in words than in pseudo-words. This word superiority effect was even stronger in the AV condition. It should be pointed out however that we observed the word superiority effect only for correct phoneme detection and not for latencies. Moreover, we were not able to replicate the lexical effect in the auditory only condition, as observed in many studies on word recognition (e.g. Cutler et al., 1987). We thus re-conducted Experiment 1 in a non-noisy environment.

3. Experiment 2

3.1. Method

3.1.1. Participants

Thirty-seven native French speakers ranging in age from 18 to 32 years with a mean age of 21.8 years participated in the experiment. They all had normal or corrected-to-normal vision and reported no auditory disorders.

3.1.2. Stimuli and procedure

They were the same as in Experiment 1 but without noise.

3.2. Results

Mean correct phoneme detection percentages and response latencies were calculated for each participant and for each item pair. A 2 (modality: A vs. AV) × 2 (lexical status: word vs. pseudo-word) within participants ANOVA was conducted by participants ($F1$) and items ($F2$). We discarded from the analyses 1% of our data that was above or below two standard-deviations (SD) from the mean. Table 3 presents the response latencies for words and pseudo-words in the A and AV conditions.

The analyses conducted on response latencies revealed a main lexical effect, $F1$ (1, 36) = 8.21, $p < .01$; $F2$ (1, 33) = 11.48, $p < .01$. As in other studies using a phoneme monitoring task, the participants were faster at detecting the target phonemes in words than in pseudo-words. However, we neither obtained a main modality effect nor an interaction between the two factors (all $Fs < 1$). The analyses on correct phoneme detection[4] did not yield any significant effect (all $Fs < 1$).

4. General discussion

The goal of this study was to show that the visual information provided by the speaker's articulatory gestures contributes to lexical activation during word recognition. We conducted a phoneme monitoring task with words and pseudo-words in Audio only (A) and Audiovisual (AV) contexts with two levels of noise masking the acoustic signal (Experiment 1) and without noise (Experiment 2).

Our results replicated Benoît et al.'s (1994) findings. Phoneme detection scores were higher in AV than in A, especially in noisy conditions (Experiment 1). The audiovisual benefit could be explained by the fact that under

[4] We also computed a d' for each stimulus pair and conducted a 2 (modality: A vs. AV) × 2 (lexical status: word vs. pseudo-word) ANOVA by participants. As for the correct phoneme detection, neither main effects nor interaction between the two factors were significant (all $Fs < 1$).

Table 3
Percentage of correct phoneme detection (CR, in %), and mean Response latencies (RT, in ms) as a function of modality and lexical status in Experiment 2. Numbers in parentheses represent the standard deviation.

Modality of presentation	Words	Pseudo-words	Word superiority effect
RT			
Audio alone	478 (93)	491 (98.8)	13
Audiovisual	475 (128)	493 (125)	18
Mean	476	492	16**
CR			
Audio alone	95.1 (7.1)	96.1 (7.4)	−1
Audiovisual	95.9 (7)	94.4 (8.31)	1.5
Mean	94.8	94.7	0.1

** Word superiority effect significant by participants and by items.

deteriorated acoustic conditions, visual and acoustic signals complement each other (Summerfield, 1987). The auditory information (e.g. place of articulation) that has been masked by the noise is available in the visual signal and can be recovered by seeing the lips, teeth, tongue and jaw movements (Miller and Nicely, 1955; Robert-Ribes et al., 1998). The data on latencies indicate that phoneme detection was faster in AV than in A when the acoustic signal is deteriorated. This suggests that the information on the speaker's orofacial gestures not only enhances phoneme identification in noise (Benoît et al., 1994), but it also accelerates phoneme detection.

The results also revealed a main lexical effect. Consonant phonemes were detected better when they were embedded in words rather than in pseudo-words. This word superiority effect indicates that phoneme detection can be influenced by lexical knowledge. In the noisy situation (Experiment 1), the lexical effect was mostly present in the AV condition. This suggests that the lexical effect is not due to auditory information only. Indeed, these results indicate that the presence of visual information not only facilitates phoneme detection but also contributes in the process of word recognition, especially when the auditory information is deteriorated. Our results are in line with Brancazio (2004) and Barutchu et al. (2008) and provide complementary data indicating that the processing of facial information accelerates phoneme perception and enhances lexical activation in a noisy environment. In Experiment 1, we did not observe a lexical effect in the A modality. We do not have a plausible explanation for this lack of results.

In the without-noise conditions (Experiment 2), consonant phonemes were detected faster when they were embedded in words than in pseudo-words. We observed a main lexical effect on response latencies but no significant interaction with the presentation modality. The lack of effect was essentially due to ceiling effects in both A and AV. Indeed, when the conditions for speech perception were optimal (i.e. when the acoustic signal was clear) the auditory information was enough for recognizing the words efficiently (see Spinelli and Ferrand, 2005, for a review on auditory word recognition studies). Further research should be carried out to determine whether the visual information enhances the lexical activation in every face-to-face situation or only when the auditory information is deteriorated or unavailable.

Models of spoken word recognition such as TRACE (McClelland and Elman, 1986) or MERGE (Norris et al., 2000) describe lexical access in the auditory modality only. However, our data showed a word superiority effect in the AV modality, suggesting that visual information plays a role in lexical access. None of these models incorporate the orofacial gestures as a source of information in their architectures. How could models like TRACE and MERGE account for our results if they included visual information?

TRACE assumes that during the perception of an isolated utterance (i.e. a word or a pseudo-word) some activation first spreads from the sensory input to the featural level. Next, the activation flows to the phonemic stage, where the phonemic decisions are made. When the stimulus is a word, the activation scattering then spreads to the lexical level. To account for the word superiority effect on phoneme detection, TRACE assumes that the activation which reaches the lexical stage flows back to the phonemic level. In this model, the activation can spread bidirectionally between the pre-lexical (featural or phonemic) levels and the lexical units. Thus, the phonemic level receives more activation for a phoneme embedded in a word than for a phoneme embedded in a pseudo-word. Consequently, the word superiority effect observed in our experiment would result – according to a top-down view of spoken word recognition – from a feedback from high-level lexical representations to low-level phonemic units. In the AV condition, pre-lexical units would receive activation from the visual and the auditory inputs whereas in the A modality, the activation flow would only emerge from the auditory input. According to this hypothesis, the phonemic stage would receive more activation in AV than in A. This mechanism could explain why visual information enhances and accelerates the phoneme detection process.

MERGE differs from TRACE with respect to the direction of the activation flow between the pre-lexical and lexical stages. MERGE assumes that activation spreads unidirectionally from pre-lexical to lexical nodes. There is no feedback from lexical to pre-lexical stages. To account for the word superiority effects, MERGE integrates a phoneme decision stage that is independent of the other two nodes. This level is entirely devoted to phonemic decision processes and it is not permanently connected to the other nodes. The connections from the lexical nodes to the phoneme decision nodes are operational only when the listener has to make phonemic decisions (e.g. during a phoneme monitoring task). According to MERGE, during the perception of an isolated word, the activation would spread from the input (or pre-lexical) nodes to the lexical nodes and to the

phoneme decision nodes simultaneously. Then, activation flows from the lexical nodes to the phoneme decision nodes. This excitatory connection between the lexical and phoneme decision nodes accounts for the word superiority effects on phoneme detection for the auditory modality. If MERGE included a visual input in its architecture, lexical nodes should receive more bottom-up support in AV than in A.

Determining the top-down or bottom-up nature of the lexical influence on phonemic process is beyond the scope of our study and further research is needed to determine how visual processing interacts with lexical activation. The timing of audiovisual integration in lexical access still remains an open question. One possibility is that the visual information directly activates lexical representations. Alternatively, the visual information could influence a pre-lexical stage. Although the present study does not provide an answer to this question, other studies with different paradigms are in progress to provide insights as to the locus of the effect of the visual information during lexical access.

Acknowledgements

We would like to thank Jean-Luc Schwartz for recording the video files and Coriandre Vilain for his very useful advice. We are also grateful to Emilie Sylvestre for proofreading this paper.

References

Amano, J., Sekiyama, K., 1998. The McGurk effect is influenced by the stimulus set size. In: Proceedings of the Auditory-Visual Speech Processing Conference. Terrigal, Australia, December 4–7, pp. 43–48.

Barutchu, A., Crewther, S., Kiely, P., Murphy, M., 2008. When /b/ill with /g/ill becomes /d/ill: evidence for a lexical effect in audiovisual speech perception. Eur. J. Cognitive Psychol. 20 (1), 1–11.

Benoît, C., Mohamadi, T., Kandel, S., 1994. Effects of phonetic context on audio-visual intelligibility of French. J. Speech Hear. Res. 37 (5), 1195–1203.

Brancazio, L., 2004. Lexical influences in audiovisual speech perception. J. Exp. Psychol. Hum. Percept. Perform. 30 (3), 445–463.

Buchwald, A.B., Winters, S.J., Pisoni, D.B., 2009. Visual speech primes open-set recognition of spoken words. Lang. Cognitive Proc. 24 (4), 580–610.

Colin, C., Radeau, M., 2003. Les illusions McGurk dans la parole: 25 ans de recherches. Ann. Psychol. 104, 497–542.

Connine, C.M., Titone, C., 1996. Phoneme monitoring. Lang. Cognitive Proc. 11 (6), 635–645.

Cutler, A., Mehler, J., Norris, D., Segui, J., 1987. Phoneme identification and the lexicon. Cognitive Psychol. 19, 141–177.

Erber, N.P., 1969. Interaction of audition and vision in the recognition of oral speech stimuli. J. Speech Hear. Res. 12 (2), 423–425.

Frauenfelder, U.H., Segui, J., Dijkstra, T., 1990. Lexical effects in phonemic processing: facilitatory or inhibitory. J. Exp. Psychol. Hum. Percept. Perform. 16 (1), 77–91.

Ganong III, W.F., 1980. Phonetic categorization in auditory word perception. J. Exp. Psychol. Hum. Percept. Perform. 6 (1), 110–125.

Gow, D.W., 2003. Feature parsing: feature cue mapping in spoken word recognition. Percept. Psychophys. 65 (4), 575–590.

Green, K.P., 1998. The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In:

Appendix.

Target-present trials used in Experiment 1 (noisy conditions) and Experiment 2 (without noise). Letters in bold represent target phonemes.

Words	Phonetic form	Frequency	Pseudo-words	Phonetic form	Words	Phonetic form	Frequency	Pseudo-words	Phonetic form
a**ff**ût	[a**f**y]	1.42	a**f**é	[a**f**e]	fu**s**il	[fyzi]	48.61	fu**s**ou	[fy**z**u]
a**t**out	[a**t**u]	5.74	a**t**o	[a**t**o]	gla**c**is	[gla**s**i]	0.02	gla**ss**u	[gla**s**y]
a**v**oue	[a**v**u]	61.56	a**v**o	[a**v**o]	in**d**ou	[ẽ**d**u]	0.03	in**d**é	[ẽ**d**e]
bé**v**ue	[be**v**y]	0.36	bé**v**o	[be**v**o]	lan**d**au	[lã**d**o]	1.01	lan**d**i	[lã**d**i]
bi**s**ou	[bi**z**u]	18.4	bi**s**é	[bi**z**e]	man**t**eau	[mã**t**o]	39.97	man**t**u	[mã**t**y]
bou**ff**i	[bu**f**i]	1.16	bou**f**u	[bu**f**y]	me**ss**ie	[me**s**i]	0.69	me**ss**é	[me**s**e]
ca**d**eau	[ka**d**o]	125.79	ca**d**u	[ka**d**y]	na**z**i	[na**z**i]	7.11	na**z**é	[na**z**e]
cha**p**eau	[ʃa**p**o]	54.91	cha**p**u	[ʃa**p**y]	ni**v**eau	[ni**v**o]	50.7	ni**v**i	[ni**v**i]
choi**s**it	[ʃwa**z**i]	170.48	choi**s**é	[ʃwa**z**e]	oi**s**eau	[wa**z**o]	77.73	oi**s**i	[wa**z**i]
cla**p**ot	[kla**p**o]	0.16	cla**p**i	[kla**p**i]	per**d**u	[pɛʀ**d**y]	36.46	per**d**o	[pɛʀ**d**o]
con**f**us	[kõ**f**y]	11.02	con**f**é	[kõ**f**e]	pro**f**it	[pʀo**f**i]	14.29	pro**f**é	[pʀo**f**e]
con**v**ie	[kõ**v**i]	2.52	con**v**é	[kõ**v**e]	ré**c**it	[ʀe**s**i]	9.89	ré**ss**é	[ʀe**s**e]
cou**t**eau	[ku**t**o]	58.15	cou**t**i	[ku**t**i]	ren**d**u	[ʀã**d**y]	508.81	ren**d**eux	[ʀã**d**œ]
dé**f**i	[de**f**i]	12.24	dé**f**u	[de**f**y]	so**s**ie	[so**z**i]	5.36	so**z**ou	[so**z**u]
de**v**is	[dø**v**i]	0.94	de**v**é	[dø**v**e]	sur**t**out	[syʀ**t**u]	1.89	sur**t**i	[syʀ**t**i]
di**ss**out	[di**s**u]	3.94	di**ss**é	[di**s**e]	sur**v**ie	[syʀ**v**i]	11.64	sur**v**ou	[syʀ**v**u]
en**v**ie	[ã**v**i]	213.96	en**v**é	[ã**v**e]	vau**d**ou	[vo**d**u]	2.93	vau**d**o	[vo**d**o]

Campbell, Dodd, B., Burnham, D. (Eds.), 2004, Hearing by Eyes II: Advances in the Psychology of Speechreading and Auditory-Visual Speech. Psychology Press, Hove, England, pp. 3–26.

Kim, J., Davis, C., Krins, P., 2004. Amodal processing of visual speech as revealed by priming. Cognition 93 (1), B39–B47.

LoCasto, P.C., Connine, C.M., Patterson, D., 2007. The role of additional processing time and lexical constraint in spoken word recognition. Lang. Speech 50, 54–75.

Marslen-Wilson, W.D., 1990. Activation, competition and frequency in lexical access. In: Altmann, G.T.M. (Ed.), Cognitive Models of Speech Processing. The MIT Press, Cambridge, MA, pp. 148–172.

McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cognitive Psychol. 18, 1–86.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27 (2), 338–352.

New, B., Pallier, C., Ferrand, L., Matos, R., 2001. Une base de données lexicales du français contemporain sur internet: LEXIQUE. Ann. Psychol. 101, 447–462, http://www.lexique.org.

Norris, D., McQueen, J.M., Cutler, A., 2000. Merging information in speech recognition: feedback is never necessary. Behav. Brain Sci. 23 (3), 299–325.

Robert-Ribes, J., Lallouache, T., Escudier, P., Schwartz, J.L., 1998. Complementary and synergy in bimodal speech: auditory, visual and audiovisual identification of French oral vowels in noise. J. Acoust. Soc. Am. 103, 3677–3689.

Sams, M., Manninen, P., Surakka, V., Helin, P., Kättö, R., 1998. McGurk effect in Finnish syllabes, isolated words, and words in sentences: effects of word meaning and sentence context. Speech Commun. 26, 75–87.

Samuel, A.G., 1981. Phonemic restoration: insights from a new methodology. J. Exp. Psychol. Gen. 110, 474–494.

Spinelli, E., Ferrand, L., 2005. Psychologie du langage: l'écrit et le parlé, du signal à la signification. Armand Colin, Paris.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Summerfield, Q., 1987. Some preliminaries to a comprehensive account to audio-visual speech perception. In: Campbell, B.D.A.R. (Ed.), Hearing by Eye: The Psychology of Lipreading. Erlbaum, Londres, pp. 3–51.

Tiippana, K., Andersen, T.S., Sams, M., 2004. Visual attention modulates audiovisual speech perception. Eur. J. Cognitive Psychol. 16, 457–472.

Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167 (917), 392–393.