I don’t see what you are saying: reduced visual influence on audiovisual speech integration in children with Specific Language Impairment

Aurélie Huyse¹, Frédéric Berthommier², Jacqueline Leybaert¹

¹Center for Research in Cognition and Neuroscience (CRNC), Université Libre de Bruxelles (ULB), Belgium
²GIPSA-Lab/DPC, Grenoble, France

ahuyse@ulb.ac.be, frederic.berthommier@gipsa-lab.grenoble-inp.fr, leybaert@ulb.ac.be

Abstract

The impact of language impairment on audio-visual integration of speech in noise is examined here by testing the influence of the degradation of the auditory and the visual speech cue. Fourteen children with specific language impairment (SLI) and 14 age-matched children with typical language development (TLD) had to identify /aCa/ syllables presented in auditory only (AO), visual only (VO) and audio-visual (AV) congruent and incongruent (McGurk stimuli) conditions, embedded either in stationary noise (ST) or amplitude modulated noise (AM), in a masking release paradigm. Visual cues were either reduced (VR) or clear (VCL). In the AO modality, children with SLI had poorer performance than TLD children in AM noise but not in ST noise, leading to a weaker masking release effect. In VO modality, children with SLI had weaker performance both in VCL and VR conditions. Analyses revealed reduced AV gains in children with SLI compared to control children. In the McGurk trials, SLI children showed a decreased influence of visual cues on AV perception in the SLI group compared to the TLD group. Data analysis in the framework of the Fuzzy-Logical Model of Perception suggested that children with SLI had preserved integration abilities; the differences with TLD children were rather due to differences in the unisensory modalities. An increased weight of audition in the VR condition compared to the VCL condition was observed in both groups, suggesting that participants awarded more weight to audition when the visual input was degraded.

Index Terms: multisensory speech perception, speech language impairment, McGurk effect, masking release

1. Introduction

Audio-visual integration of speech is a developmental process. Regarding the McGurk effect, studies indicated that the ability to fuse AV cues appear early, around 4 months of age [1]. It increases around 6-8 years [2], although the visual influence is stronger in adults, probably due to better lip-reading capacities [3]. Later in life, McGurk fusions remain intact in older adults, suggesting that AV integration is not affected by aging [4]. This trajectory can be diverted in some populations, leading to an atypical pattern of AV speech integration. In case of conflicting AV stimuli, cochlear implanted deaf individuals rely on vision, the most reliable channel for them [5]. Another clinical population of interest is children with specific language impairment (SLI), who have expressive and receptive oral language deficits in presence of normal nonverbal intelligence and hearing ability and in the absence of any obvious neurological problem [6]. Children with SLI have atypical AV speech perception because of abnormal speech processing both in AO [7] and VO modality. However, little is known about how children with SLI perceive audiovisual speech [8,9,10].

In a previous study, we compared children with SLI and children with TLD regarding their ability to perceive /aCa/ syllables in AO, VO, and AV congruent and incongruent conditions [11]. The syllables were presented either with ST or AM noise. Compared to children with TLD, children with SLI were less accurate in AO as well as in VO conditions. In response to McGurk stimuli, children with TLD showed more fusion responses in AM noise than in ST noise as a consequence of masking release effect; by contrast, children with SLI did not show this effect. In accordance with other studies [8,9,10], our data suggest that children with SLI might have atypical AV speech perception in unimodal and bimodal modalities.

The aim of the present study was to determine whether SLI deficits are due to difficulties in extracting cues in AO and VO modalities or to difficulties in the process of AV integration per se. This question was not answered in previous studies though it might have practical and theoretical implications. Practically, answering this question could shed some light on how best to conduct language rehabilitations with SLI children. If children with SLI differ from children with TLD in their ability to integrate AV speech, the focus of rehabilitation programs should be put on the training of integration, rather than on the training of the auditory and visual abilities solely [12]. Theoretically, atypical integration would indicate that SLI entails a more general speech processing impairment, affecting both auditory and visual speech components.

To answer those questions, performance of children with SLI and age-matched control children with TLD was analyzed in AO, VO, congruent AV and incongruent AV (McGurk) conditions. Throughout the experiment, stimuli were embedded in two types of noise: ST and AM. The visual speech cues were either unmodified (VCL condition) or reduced (VR condition) in order to examine the impact of visual degradation on AV speech integration.

The main hypotheses are the following. In VCL condition, children with SLI are expected to have poorer identification scores than control children but a normal masking release...
Children with SLI are also expected to be worse at VO and to have lower benefit in AV condition compared to AO condition. Regarding the McGurk effect, children with SLI are expected to be less influenced by the visual input than TLD children. This should lead them to give more auditory-based responses and less fusion responses. In the VR condition, we expect a rise in the number of auditory-based responses in the TLD group. If children with SLI are indeed less influenced by visual speech, their response pattern should not be profoundly modified in the VR condition compared to the VCL condition.

2. Method

2.1. Participants
Twenty-eight children, all native speakers of French, participated in this study. Fourteen of these children (7 boys and 7 girls, mean age = 9.58 years, SD = 5 months) were identified as having TLD and 14 (11 boys and 3 girls, mean age = 9.83 years, SD = 5.21 months) belonged to the SLI group. Those children were recruited from two special needs schools, where psychologists and speech therapists have established the diagnosis of SLI. Further assessment made in the present study included phonology (a word repetition task from the L2MA battery [14]), receptive vocabulary (the French adaptation of the Peabody Picture Vocabulary Test [15]) and reading skills (regular, irregular words and pseudowords reading of ODEDYS [16]). Children with SLI performed poorer than control children in each of these tasks (p < .05 at least) (see Table 1).

<table>
<thead>
<tr>
<th>Measure</th>
<th>SLI</th>
<th>TLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonology</td>
<td>56</td>
<td>98.8</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>42.4</td>
<td>118.3</td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular words</td>
<td>64.6</td>
<td>95</td>
</tr>
<tr>
<td>Irregular words</td>
<td>42.5</td>
<td>71.4</td>
</tr>
<tr>
<td>Pseudowords</td>
<td>62.9</td>
<td>77.9</td>
</tr>
</tbody>
</table>

Values for Phonology indicate percent correct; values for Vocabulary indicate standardized scores; values for reading indicate percent correct.

2.2. Stimuli
Stimuli were composed of vowel-consonant-vowel syllables with the consonants /p, t, k, s, f, ŋ/ interposed between two /a/ vowels. A male speaker of French was videotaped from the bottom of the nose to the chin while saying these syllables. The production of each stimulus began and ended in a neutral position, with the mouth closed. Videos (Quicktime movie files, 21 by 21 cm) were displayed centered on a 15-inch MacBook Pro laptop on a black background. Three productions of each /aCa/ stimulus were digitally recorded and the productions of each /aCa/ stimulus were digitally recorded and then assigned to the AV, AO and VO trials. Stimuli were delivered through Sennheinser HD 121 Pro headphones.

The congruent AV stimuli included digital audio-video files of the speaker saying and articulating the /aCa/ stimuli. For the AO stimuli, an image of the speaker, appearing neutral and with mouth closed was presented along with the auditory stimulus. For the VO stimuli, the audio was turned off. Incongruent AV syllables (McGurk stimuli) were created by matching the onset of audio files with non-corresponding video files. We used three repetitions of the two following stimuli: A/apa/V/aka/ (expected fusion /ata/) and A/afa/V/ala/ (expected fusion /asa/).

Auditory noise
Each signal was digitized at a 22050Hz sampling frequency. Stimuli were presented in 3 conditions: quiet (i.e. no noise added), with ST noise and with AM noise. Modulation in amplitude was achieved by using a white Gaussian noise low-pass filtered at 500 Hz (WGNf). The expression describing the sine-wave modulator, \( m(t) \), was

\[
m(t) = [1 + \cos(2\pi f_m t)] * WGNf
\]

where the 1st-order modulation frequency \( f_m \) was 8 and 128 Hz. The noise was then added to the signal. The signal-to-noise ratio was fixed at -23dB (prior to the 500Hz filtering). This SNR was determined in a preliminary experiment so as to yield a consonant identification performance of about 30% correct under ST noise (in AO condition).

Visual reduction
In half the experiment, the quality of the visual component was reduced, by varying the contrast of the video around the mean intensity of the image X, for each RGB color of the image.

\[
Y = \text{mean}(X) + \left( \frac{X - \text{mean}(X)}{R} \right)
\]

The contrast \( R \) varies at a period of 4Hz according to the following function:

\[
R = k * 10 \^ (1 + 0.5 * \cos(\phi + f(t)))
\]

where the parameter \( k \) is set to 4 and where \( f(t) \) represents the modulation frequency. \( f(t) \) was set at 4 Hz. Since the total duration of the syllable (the mouth movements) exceeded 250msec, a contrast modulation of 4 Hz always generated periods of masking and periods of unmasking within a single item. The starting phase of the modulation \( \phi \) was randomized in each interval between 0 and \( \pi \). This random phase was added in order to prevent the visual modulation to be synchronized with the auditory modulation. Since \( \phi \) was random, visual reduction did not affect each sample of a single phoneme equally.

Half of the stimuli were presented in VCL condition and the other half in VR condition. Each condition consisted of 180 stimuli: 6 syllables x 3 repetitions x 3 modalities (AO, VO, AV) x 3 types of noise (Quiet, ST, AM) + 2 McGurk syllables.
x 3 repetitions x 3 types of noise. Four blocks of 45 VCL items and 4 blocks of 45 VR items were created. We alternated one VR block with one VCL block.

2.3. Procedure

The experiment took place in a dimly-lit quiet room. The monitor was positioned at eye level, 70 cm from the participant’s head. The session began with a training session composed of one VCL block followed by one VR block. Participants were asked to identify the syllable. Children with TLD reported it aloud while children with SLI were also offered the possibility to point to 6 different images corresponding to the syllables. All participants were informed of the composition of the stimulus set (but not of the presence of the McGurk stimuli) and had the response options written on a paper in front of them during the training session but not during the experimental session. In both groups, half the participants began the experimental session with a VR block and the other half began with a VCL block. Responses were noted by the experimenter. Total duration of the experiment was about 40 minutes.

3. Results

3.1. Speech perception under VO and AO optimal conditions

Speech identification scores (% correct) in quiet and with a VCL input, appear in Table 2. A repeated-measure analysis of variance (ANOVA) with modality (A, V, AV) as within-subject factor and group (SLI, TLD) as between-subject factor yielded a significant main effect of group \( F(1,26) = 5.6, p = .03 \), with TLD children having overall higher scores than SLI children. There was also a significant main effect of modality \( F(2.52) = 76.09, p < .0001 \) : both groups performed better in A and AV condition than in V condition. The interaction between group and modality was not significant \( (p = .49) \).

Table 2. Percent correct responses of children with SLI and TLD under VO and AO optimal conditions (standard deviation are in brackets)

<table>
<thead>
<tr>
<th></th>
<th>SLI</th>
<th>TLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO</td>
<td>92.4 (4.9)</td>
<td>100 (0)</td>
</tr>
<tr>
<td>VO</td>
<td>49.04 (6.8)</td>
<td>63.9 (3.4)</td>
</tr>
<tr>
<td>AV</td>
<td>93.3 (4.9)</td>
<td>100 (0)</td>
</tr>
</tbody>
</table>

3.2. Speech Perception in Noise

Speech identification scores (% correct) in noise and computed scores of masking release are presented in Table 3. Results in noise modulated at 8Hz and noise modulated at 128Hz were averaged for more clarity.

First, results in AO were analyzed. A repeated-measure ANOVA was run with visual condition (VCL, VR) and noise (ST, AM) as within-subject factors and group (SLI, TLD) as between-subject factor. Analyses revealed a significant effect of noise \( F(1,26) = 306.24, p < .0001 \), with AM noise leading to higher scores than ST noise. There was also a significant interaction between noise and group \( F(1,26) = 5.39, p = .03 \). Results of both groups significantly differed in AM noise \( p = .02 \), where children with TLD had higher scores than SLI children. There was no effect of group in ST noise \( p = .57 \). A significant masking release (MR) effect was obtained in both groups. MR was significantly smaller in the SLI group compared to the TLD group (23.15% and 29.6% respectively, with VCL and VR being averaged).

Second, results in the VO modality were analyzed. A repeated-measure ANOVA was run with visual condition (VCL, VR) and noise (ST, AM) as within-subject factors and group (SLI, TLD) as between-subject factor. There was a significant effect of group \( F(1,26) = 64.71, p < .0001 \), with VCL condition leading to higher performance compared to the VR condition. There was also a significant effect of group \( F(1,26) = 9.94, p = .004 \), with the TLD group having higher global performance than the SLI group. There was no significant interaction between group and visual condition.

The audio-visual gain (AVG) (Sumby & Pollack, 1954) was computed in both ST and AM noises using the following formula:

\[
AVG = \frac{(AV – A)}{(100 – A)}
\]

Thus, AVG measures the gain performance in the AV modality in comparison to performance in AO modality, normalized relative to the amount by which speech intelligibility could have improved above AO scores.

A repeated-measure ANOVA was run with visual condition (VCL, VR) and noise (ST, AM) as within-subject factor and with group (SLI, TLD) as between-subject factor. Analyses revealed a significant main effect of group \( F(1,26) = 16.98, p < .0001 \) with weaker AVG in the SLI group compared to the TLD group. There was also a significant main effect of visual condition \( F(1,26) = 29.97, p < .0001 \) with VCL condition leading to higher AVG than VR condition, and a main effect of noise \( F(1,26) = 6.28, p = .02 \), with ST noise leading to higher AVG than AM noise. There was no significant interaction including the group variable.

3.3. McGurk Effects

A /afa/ dubbed on V /a/ . The number of auditory /afa/, fusion /asa/ and visual /a/ responses was computed. Data are presented first in the VCL condition in order to examine the impact of noise on response pattern in both groups. Then, results are presented in the VR condition, in order to evaluate the impact of visual reduction.

In VCL condition: In ST noise, SLI children gave 38% auditory, 14% fusion and 21% visual responses; TLD children gave 12% of auditory, 24% fusion, and 50% visual responses. In AM noise, SLI children gave 44% auditory, 20% fusion and 24% visual responses; TLD children gave 30% auditory, 44% fusion, and 24% visual responses. Group differences were analyzed with separate ANOVAS for ST and AM noise. In ST noise, SLI children gave significantly more auditory \( p = .01 \) and less visual responses \( p = .02 \) than TLD children. In AM noise, TLD children made significantly more fusions than SLI children \( p = .03 \).
Table 3. Percentage correct (standard deviation in parentheses) of children with SLI and TLD children in ST and AM noise

<table>
<thead>
<tr>
<th></th>
<th>Auditory-only (AO)</th>
<th>Visual-only (VO)</th>
<th>Audio-visual (AV)</th>
<th>Audio-visual gain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SLI</td>
<td>Control</td>
<td>SLI</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td>43.3 (3.3)</td>
<td>48 (2.3)</td>
<td>48.6 (3.1)</td>
<td>49.2 (3.2)</td>
</tr>
<tr>
<td></td>
<td>68.3 (3.2)</td>
<td>76.9 (2.3)</td>
<td>68.9 (3.9)</td>
<td>79.4 (2.0)</td>
</tr>
<tr>
<td></td>
<td>44 (3.8)</td>
<td>58.7 (3.6)</td>
<td>32.9 (2.6)</td>
<td>37.7 (4.2)</td>
</tr>
<tr>
<td></td>
<td>44 (3.5)</td>
<td>59.7 (2.7)</td>
<td>29 (3.3)</td>
<td>42.5 (3.4)</td>
</tr>
<tr>
<td></td>
<td>73.8 (4.4)</td>
<td>90.1 (2.1)</td>
<td>57.9 (4.6)</td>
<td>71.4 (2)</td>
</tr>
<tr>
<td></td>
<td>77.4 (3)</td>
<td>93.1 (1.2)</td>
<td>73.4 (4.2)</td>
<td>89.9 (1.3)</td>
</tr>
<tr>
<td></td>
<td>56 (6.5)</td>
<td>81.2 (3.6)</td>
<td>16.4 (8.7)</td>
<td>42.3 (4.8)</td>
</tr>
<tr>
<td></td>
<td>28.7 (8.7)</td>
<td>67.9 (6.8)</td>
<td>9.8 (11)</td>
<td>43.9 (10.1)</td>
</tr>
<tr>
<td></td>
<td>25 (2.5)</td>
<td>29 (3.4)</td>
<td>20.3 (9.8)</td>
<td>30.2 (9.4)</td>
</tr>
</tbody>
</table>

In VR condition: In ST noise, SLI children gave 40% auditory, 19% fusion and 29% visual responses. TLD children gave 21% auditory, 14% fusion, and 62% visual responses. In AM noise, SLI children still gave 43% auditory, 13% fusion and 20% visual responses. TLD children gave 39% auditory, 36% fusion and 19% visual responses. Group differences were analyzed with separate ANOVAs for each level of the noise factor. In ST noise, SLI children gave significantly less visual responses than TLD children (p = .02). In AM noise, there was a significant group difference in the number of fusion, with TLD children making more /asa/ responses than SLI children (p = .01).

Next, we compared responses made in the VR condition to those made in the VCL condition. There was no significant impact of visual condition on auditory, fusion, and visual responses, neither in ST nor in AM noise.

A /apa/ dubbed on V /aka/. The number of auditory /apa/, fusion /ata/, and visual /aka/ responses was computed.

In VCL condition: In ST noise, SLI children gave 14% auditory, 12% fusion, and 38% visual responses; TLD children gave 7% auditory, 26% fusion, and 52% visual responses. In AM noise, SLI children gave 23% auditory, 21% fusion, and 36% visual responses; TLD children gave 14% auditory, 43% fusion, and 31% visual responses. No group difference was found in ST noise. However, in AM noise, TLD children made more fusion responses than SLI children (p = .04).

In VR condition: In ST noise, SLI children gave 33% auditory, 12% fusion and 24% visual responses. Children with TLD gave 24% auditory, 28% fusion and 26% visual responses. In AM noise, SLI gave 44% auditory, 23% fusion, and 19% visual responses; TLD children gave 44% auditory, 32% fusion, and 18% visual responses. Analyses showed no significant group effect either in ST or in AM noise.

Next, for each group, responses made in VCL and VR conditions were compared. In the SLI group, there was no significant interaction between visual condition and responses in ST noise, suggesting no significant impact of visual reduction. By contrast, this interaction was significant in AM noise [F(2,26) = 3.88, p = .03]. Further analyses revealed that VR led to an increase of auditory responses and a decrease of visual responses. In the TLD group, there was a significant impact of VR in ST noise [F(2,26) = 3.30, p = .04] as well as in AM noise [F(2,26) = 18.61, p < .0001]. In ST noise, the visual responses decreased (p = .02). In AM noise, the visual responses increased (p = .003) and the auditory responses increased (p < .0001).

3.4. Modeling results

In order to disentangle unsensory effects from integration effects, WFLMP was applied to our data. The standard form of the FLMP [17] is a post-phonetic integration model with a statistically optimal integration rule. It can be expressed as

\[
P(R_{i,A,V}) = \frac{P(R_{i,A})^{\lambda_A} P(R_{i,V})^{\lambda_V}}{\sum_j P(R_{i,A})^{\lambda_A} P(R_{i,V})^{\lambda_V}}
\]

In this expression, R and R_j are response categories, A and V are auditory and visual stimuli, P(R_{i,A}), P(R_{i,V}) and P(R_{i,A,V}) are auditory, visual and audiovisual response probabilities, respectively. Since the FLMP entails a fixed integration rule, a good fit of data to the FLMP means that any differences in AV responses are due to differences in unsensory processing before the AV integration occurs.

The WFLMP [12] where inputs from audition and vision are weighted, is defined by

\[
P(R_{i,A,V}) = \frac{P(R_{i,A})^{\lambda_A} P(R_{i,V})^{\lambda_V}}{\sum_j P(R_{i,A})^{\lambda_A} P(R_{i,V})^{\lambda_V}}
\]

In this expression, \( \lambda_A \) and \( \lambda_V \) are subject-dependent factors used to weight the auditory and visual inputs in the computation of the audio-visual responses. For each subject, a lambda value is defined between 0 and 1. Then, \( \lambda_A \) and \( \lambda_V \) are computed from lambda by: \( \lambda_A = \text{lambda}/(1-\text{lambda}) \) and \( \lambda_V = (1-\text{lambda})/\text{lambda} \), with thresholds maintaining \( \lambda_A \) and \( \lambda_V \) between 0 and 1. Consequently, P_{AV} varies from a value close to P_A when lambda is close to 0, to a value close to P_V when lambda is close to 1, through a value identical to the FLMP prediction when lambda is close to 0.5, with \( \lambda_A \) and \( \lambda_V \) both equal to 1 [Schwartz, 2010].

Here, we used the root mean square error (RMSE) as assessment criterion. It is computed by taking the squared distance between observed and predicted probabilities of responses, averaging them over all categories C, and all experimental conditions E_i, and taking the square root of the result:
\[
\text{RMSE} = \left( \frac{1}{n_c} \sum_{i=1}^{n_c} (P_{\text{act}}(C_i) - P_{\text{pred}}(C_i))^2 \right)^{1/2}
\]

In this equation, observed probabilities are in lower case and predicted probabilities in upper case.

Figure 1 shows that regression lines of both SLI and TLD groups were flat in the VCL condition, suggesting that lambda was not a significant predictor of RMSE. This was confirmed by a linear regression analysis (\( p = .31 \) in the SLI group and \( p = .06 \) in the TLD group). In the VR condition, the slopes of the regression lines were steeper, suggesting a relationship between lambda and RMSE. Linear regression analyses confirmed these observations in both groups (SLI group: \( R^2 = 0.04, F(1,265) = 10.26, p = .002 \); TLD group: \( R^2 = 0.02, F(1,265) = 6.10, p = .01 \); RMSE decreased as lambda increased (i.e. as lambda approached values close to \( P_c \)). This result suggests an increased weight of audition in the VR condition compared to the VCL condition. In order to test this hypothesis, we performed a one-way analysis of covariance with RMSE as dependent variable, visual condition (VCL, VR) and group as factors and lambda as covariate factor. We found significant main effects of visual condition \( F(1,75) = 20.79, p < .0001 \), lambda \( F(1,75) = 23.86, p < .0001 \) and group \( F(1,75) = 48.67, p < .0001 \). The interaction between lambda and visual condition, which shows the impact of lambda on RMSE according to each visual condition, was marginally significant \( F(1,75) = 3.87, p = .053 \). Further analyses revealed a significant impact of lambda in the VR condition \( p < .005 \) but not in the VCL condition \( p = .40 \). These results suggest that the model fitted the data better when no specific weight was added in the VCL condition (as in the FLMP) and when the auditory weight was increased in the VR condition.

**Figure 1. Variation of the RMSE as a function of the lambda parameter tuning fusion in the WFLMP for the SLI group and the TLD group, in the VCL and VR conditions.**

### 4. Conclusions

Data from the present study show a clear tendency for SLI children to be less influenced by visual speech cues than children with TLD. They had lower AV gains for AV congruent stimuli. When faced with McGurk stimuli, they generally made more auditory responses and fewer fusion and visual responses. Results of the WFLMP simulations suggest that the impact of language impairment occurs prior to the stage of AV integration. Children with SLI differ from control children in the way they extract cues from the auditory and visual modalities, leading to differential AV pattern. Since children with SLI do not seem to benefit from visual speech cues similarly to control children, we assume that specific language impairments reflect a general impairment of speech processing, rather than a disorder of audition only. However, their ability to integrate auditory and visual cues seems to be preserved. This latter result has important practical implications. The fact that children with SLI have preserved integration abilities should lead healthcare workers to take the visual speech cue into account in their daily practice. However, since SLI affects the ability of lip-reading, speech
therapy should first include a stage of assessing those abilities and then a stage of rehabilitating them. Teachers should also adapt the classroom and their communication style in a way that allows a better view of their faces, so that children can have access to visual speech cues. Besides, care givers and teachers should know that children with SLI have enhanced difficulties processing auditory speech in noisy backgrounds. Providing the opportunity to perform lip-reading (after a first stage of rehabilitation) can be a way to get round this difficulty.

5. Acknowledgements

Aurélie Huyse was Aspirant FNRS. The work was supported by grant 2.4539.11 from the Fonds National de la Recherche Scientifique (FRS-FNRS, Belgium) to J. Leybaert & C. Colin

6. References


