An Exploration of Why Preschoolers Perform Differently Than Do Adults in Audiovisual Speech Perception Tasks

Renée N. Desjardins, John Rogers, and Janet F. Werker

The University of British Columbia, Vancouver, British Columbia, Canada

Preschoolers’ perception of audiovisual speech is considerably less influenced by visual information than adults’. We test the hypothesis that experience correctly producing consonants plays a role in developing the underlying representation which mediates the perception of visible speech. We divided preschoolers into two groups: those who made substitution errors and those who did not. Using a newly developed methodology, we tested substituters, nonsubstituters, and adults in an auditory-only condition, a visual-only condition, and an audiovisual condition. There were no differences among groups in the auditory-only condition. Overall, children still showed less visual influence than adults. Among the children, substituters were poorer at lip-reading in the visual-only condition and showed less visual influence on the incongruent audiovisual tokens than did nonsubstituters. These results support our hypothesis that experience correctly producing consonants plays a role in the elaboration of the underlying representation.

There is a growing literature documenting the influence of visual information (i.e., lip-read information) on the perception of speech in hearing adults. Although there are large individual differences in the extent to which visual information influences speech perception, all of us can and do lip-read much more than we might believe (Summerfield, 1991). Much of the current interest in audiovisual speech perception stems from the surprising finding that even under optimal listening conditions, watching a speaker’s lip and mouth movements influences what speech sound is heard. In the initial demonstration of
this phenomenon, McGurk and MacDonald (1976) found that pairing a video of a face articulating /ga/ in synchrony with a soundtrack of /ba/ resulted in reports of a “heard” /da/. Subsequent research has revealed other percepts that arise from incongruent audiovisual speech signals. For example, an auditory /ba/ paired with a visual /da/ results in a percept of /da/ (Manuel, Repp, Studdert-Kennedy, & Liberman, 1983). When the syllable “heard” corresponds to the visible articulation, the resulting percept is often referred to as “visual capture.” It is in visual capture that the impact of the visible articulation of speech on the resulting percept is most obvious.

The remarkable feature of speech percepts showing an influence from the visual information is that the perceiver is entirely unaware of the mismatch between the auditory and visual signals and perceives a single, coherent syllable. (Some mismatched auditory and visual syllables result in percepts that maintain both signals. For example, an auditory /ga/ and a visual /ba/ results in a /bga/ percept.) Following the reports of these perceptual illusions, several studies appeared in the literature documenting the conditions under which these effects hold (e.g., Green, in press; Green & Gerdeman, 1995; Green, Kuhl, & Meltzoff, 1988; Green, Kuhl, Meltzoff, & Stevens, 1991; McGrath & Summerfield, 1985; Saldaña, Rosenblum, & Osinga, 1992; Summerfield, 1979; Werker, Frost, & McGurk 1992).

Several hypotheses have been offered to explain the phenomenon of audiovisual speech perception (Fowler & Rosenblum, 1991; Liberman & Mattingly, 1985; Massaro, 1987, 1989; Summerfield, 1991). One explanation is that heard and seen speech are readily integrated because both give information about a speaker’s intended production. And, it is claimed that audiovisual speech perception is possible because there is an innate link between production and perception that accounts for speech perception more generally (Liberman & Mattingly, 1985). An inference that can be made from this hypothesized link between perception and production is that the underlying representation of the gesture contains information about both the acoustic characteristics and visible characteristics of that phonetic unit. Indirect support for this possibility is provided by studies suggesting that the same areas of brain are involved in producing as perceiving speech (Sams et al., 1991; Ojemann, 1983; cf. Ojemann, 1991). Alternatively, there may be two distinct representations of the gesture which are linked: one containing information about the acoustic characteristics and the other containing information about the visible articulation. One question, then, is whether the information regarding what the gesture looks like is initially well specified in the representation or whether experience listening to, producing, and observing speech produced is necessary to develop the representation more fully. One way to tease apart these possibilities is to examine the perception of audiovisual syllables in young children as they have had much less experience with speech than adults.

There is now a sizable literature that suggests speech perception in children shows less influence of visual information than does speech perception in
adults. In the original demonstration by McGurk and MacDonald (1976), in addition to tests with adults, the responses of 3- to 4-year-old and 7- to 8-year-old children were recorded. Children and adults were tested on four auditory-visual mismatches: audio /ba/—visual /ga/, audio /ga/—visual /ba/, audio /pa/—visual /ka/, and audio /ka/—visual /pa/. The instructions were to watch the television monitor and repeat what the female model said. On all four stimulus pairs, children reported hearing syllables reflecting an influence of the visual information significantly less frequently (about half as often) than did adults.

Using a five-step /ba-da/ continuum of synthesized segments and a face articulating either /ba/ or /da/, Massaro (1984; Massaro, Thompson, Barron, & Laren, 1986) also demonstrated that preschool children (4 to 6 years of age) are less influenced by the visual syllable than are adults. Subjects saw a video recording of a man whose head filled two thirds of the video screen and were asked to identify whether the “man on the TV” said either /ba/ or /da/. In one study (Massaro, 1984), both children and adults indicated their responses by pressing either a button labeled “BA” or a button labeled “DA.” For the children, the button labeled BA was accompanied by a drawing of a ball, and the button labeled DA was accompanied by a drawing of a duck. In a subsequent study (Massaro et al., 1986) both children and adults indicated their responses orally. In both studies, children showed visual capture only one quarter as often as adults.

More recently, Hockley and Polka (Hockley, 1994; Hockley & Polka, 1994) examined the developmental trend of increasing visual capture with increasing age. Four different age groups of children (5-, 7-, 9-, and 11-year-olds) and adults were tested in four different conditions: auditory alone, visual alone, congruent audiovisual, and incongruent audiovisual. Productions of /ba/, /va/, /θa/, /da/, and /ga/ were made using a male talker. For the incongruent audiovisual tokens, an auditory /ba/ was paired with a visual /va/, /θa/, /da/, and /ga/. The procedure was similar to that of previous studies: children were asked to tell an experimental assistant what the man on the TV was saying. Breaks were taken as deemed necessary by the assistant.

The results from the Hockley and Polka work revealed a general trend for increasing accuracy in the visual-only and congruent audiovisual conditions with increasing age of participants (Hockley, 1994; Hockley & Polka, 1994). In the audiovisual condition, the proportion of auditory capture—a perpect influenced by only the auditory component—decreased with age while the proportion of visual capture increased with age. Indeed, in the 5-year-old group, the proportion of responses across the four incongruent syllables showing visual influence was approximately 6% while in the adult group the proportion was approximately 55%.

Taken together, these studies suggest that visual information has less of an influence on the perception of speech for children than for adults. These findings suggest that the visible articulation may not be initially well specified.
in the representation, and that experience listening to, observing, and/or producing speech may be necessary to develop the representation more completely.

Cross-linguistic research with adults supports the notion that experience plays a role in audiovisual speech perception. Werker, Frost, and McGurk (1992) found that speakers of Canadian French with some knowledge of English tended to report an audio /ba/—visual /ða/ as either a /da/ or a /ta/ unlike native English speakers who tended to report hearing /ða/. Werker et al. suggest that the French speakers—for whom /ða/ does not have phonemic status and thus may be poorly represented—were assimilating the visible interdental, /ða/ to the closest phonemic place of articulation, thus reporting an alveolar/dental sound, /da/ or /ta/. Indeed bilingual French-English speakers reported significantly more visual capture to the interdental /ða/ than French speakers who were only beginning to learn English. These findings attest to the significant role experience plays in the perception of visible speech.

A recent study on audiovisual speech perception comparing Japanese and English speakers indicates that the influence from visual information might be less strong in Japanese speakers, particularly when Japanese syllables are used (Sekiyama & Tokura, 1993). The authors suggest that one interpretation for these findings is that the visual information is less salient for Japanese speakers because it is impolite in the Japanese culture to look directly at another person when he/she is speaking. Thus it is possible that Japanese speakers have less experience with visible speech than English speakers (but see Massaro, Cohen, Gesi, Heredia, & Tsuzaki (1993), for results suggesting both the auditory and the visual information to be important even for Japanese speakers).

Young children may be like the French speakers who were beginning to learn English or the Japanese speakers listening to Japanese: both may have inadequate representations of the articulation of certain syllables due to the lack of sufficient experience. What remains difficult to distinguish, however, is whether young children show less influence of visual information than adults simply because they lack experience observing visible speech or because they lack experience producing the relevant speech sounds and gaining feedback from their own productions. We can be fairly certain that the Japanese speakers lacked experience only in observing visible speech. Presumably the French speakers in the Werker et al. study lacked experience both hearing and seeing English spoken as well as producing the English sounds. Unlike the Japanese speakers, preschoolers in English-speaking families have had several years hearing and seeing their native language spoken, but perhaps not as much experience producing those same sounds.

To assess these possibilities we decided to test explicitly the role of production in the perception of visible speech. To accomplish this we tested preschoolers who vary in their ability to produce certain speech sounds. Some articulatory difficulties, such as syllable reduction—reducing CCV (conso-
nant–consonant–vowel) or CVC syllables to CV or VC form—indicate atypical development and may reflect lower intelligence and neurological abnormalities (Smit & Bernthal, 1983). However, other articulatory difficulties, particularly substitution errors, are often seen in otherwise normally developing preschoolers. Indeed it is not uncommon for preschoolers to make substitution errors when attempting to produce consonant clusters, and many children continue to make substitution errors for “difficult” consonant singletons up to the age of 4 or 5 years (Ingram, 1989). Thus, a typical preschool population is composed of some children who would be considered substituters and others who are producing consonant singletons correctly. If experience producing speech correctly plays a role in establishing the underlying representation of both heard and seen speech, or of seen speech alone if it is represented separately from heard speech, then substituters should be less influenced than nonsubstituters by visual information.

The primary purpose of this study was, therefore, to examine the impact of articulatory experience on perception of visible speech. To accomplish this, we chose to administer to a preschool sample a short measure of articulatory proficiency which would allow us to divide preschoolers into two groups: “substituters” and “nonsubstituters.” Substituters, nonsubstituters, and adults—who provided a ceiling level of performance against which we could compare the preschoolers—were tested on their perception of visible speech. Specifically, we expected that the nonsubstituters would be more adult-like than the substituters in their ability to lip-read in a visual-only condition and more adult-like in their perception of incongruent audiovisual syllables. All subjects were also tested in an auditory-only condition. Inclusion of this condition would allow us to ascertain whether experience producing correct articulations impacts the representation of both seen and heard speech or whether it impacts only the representation of seen speech.

We were concerned, however, about the potential validity of any results we might obtain with preschoolers because children likely find the typical audiovisual speech task boring. In fact, it has been noted in previous studies (e.g., Hockley, 1994) and in our own pilot work that children tend to become disinterested and fidgety in the standard task. Thus the standard task may not be sufficiently sensitive to reveal children’s capabilities. For this reason we developed a modified task in the hopes of achieving greater interest and motivation on the part of children. We reasoned that a task which involves more child–experimenter interaction and which involves a more fun response would more optimally engage the child. Instead of simply asking children to repeat what “the man on TV” said, the children were taught names for toy characters that corresponded to the syllables used in the experiment. After attending to the syllable, children were asked to select the toy named by the man on TV. To make the task more engaging, children were praised at the end of each trial and were given a gold star after every five trials. By changing the task to be more age-appropriate, we hoped that children’s performance
might reflect their true ability and thus allow us to investigate more easily why children appear to have underspecified representations of visible speech.

In summary, the present study combines two goals: (1) the examination of the relationship between articulatory proficiency and the perception of both acoustic and visible speech and (2) the development of a child-friendly audiovisual speech perception task to facilitate children’s attention and optimal performance. If an unengaging procedure is at the root of previous findings that children show less visual influence than adults, our new procedure may result in our preschoolers showing relatively more visual influence—relative to adult levels for the same stimuli. Over and above any effects of procedural changes, if an inadequately specified representation of the visual information is present as a result of incorrect articulation, children who tend to make substitution errors may show even less visual influence than children who make no substitution errors.

METHODS

Participants

Sixteen children (11 boys, 5 girls) from a local preschool ranging in age from 3 years, 3 months to 5 years, 5 months participated ($M = 4$ years 3 months, $SD = 8$ months). An additional 10 children were eliminated from the final sample for the following reasons: (a) the child had not memorized the names of the stimuli (8), or (b) the child chose to withdraw from the experiment before testing was complete (2). All of the children were developing normally according to teacher report. Each child was awarded a ‘‘Junior Scientist Degree’’ certificate.

Sixteen adults (8 males, 8 females) ranging in age from 21 to 29 years ($M = 24$ years, $SD = 2$ years) volunteered for the experiment. None of the adults reported any hearing loss.

Stimuli

Speech stimuli. A video recording (Panasonic AG 1960 Pro Line video recorder) was made of the lower portion—mouth and chin—of a male speaker articulating /ba/, /va/, /da/, and /da/. These syllables were chosen because they are all frontal places of articulation and /va/, /da/, and /da/ typically result in visual capture when paired with an auditory /ba/. Open syllables were selected in order to avoid any association with names familiar to the children (e.g., ‘‘Bab’’ may be heard as ‘‘Bob’’ or ‘‘Baby’’). Each syllable was approximately 700 ms in length and syllables were presented on a 12-inch Sony television monitor at approximately 65 dB SPL for both children and adults.

For the auditory-only and visual-only conditions, each of the four syllables was copied three times to produce a final tape consisting of 12 randomly ordered tokens in the auditory-only condition and 12 randomly ordered tokens
in the visual-only condition. For the auditory-only condition, a black screen was shown (instead of the face) so that only the voice could be heard. For the visual-only condition, the audio was turned off so that only the articulating mouth and chin could be seen.

Incongruent audiovisual stimuli were made by pairing an auditory /ba/ with each of the remaining visual tokens: /va/, /ða/, and /da/. The stimuli were made by lining up the release of the auditory /ba/ with the opening of the closure (i.e., parting of the lips) of the visible articulation. The final tape for this condition contained, in random order, five instances of each of the three incongruent tokens and five instances of each congruent audiovisual /ba/ resulting in a tape of 20 syllables.

Toys. Four plastic toys approximately 4 inches high were given names corresponding to the four syllables used in the test conditions: a dinosaur called “Ba,” a gorilla called “Va,” a monster called “Tha,” and a spaceman called “Da.”

Test Conditions

Visual-only condition. This condition was included in order to assess how accurately the participants were able to lip read when no auditory information is available.

Auditory-only condition. This condition was included as a control to ensure that all subjects were able to identify with accuracy the test syllables when no conflicting visual information was present. As well, this condition allowed us to determine whether articulatory experience impacts on the representation of acoustic information in the absence of visual information.

Audiovisual condition. In this condition, each of the visual tokens of the test syllables was paired with an auditory /ba/ which resulted in one congruent pairing, audio /ba/–visual /ba/, and three incongruent pairings, audio /ba/–visual /va/, audio /ba/–visual /ða/, and audio /ba/–visual /da/. The proportion of visual capture reported by participants in this condition was of primary interest.

Procedure

Preschoolers. The procedure followed with the preschoolers was subdivided into four segments: training, articulatory assessment, visual-only, and auditory-only testing and audiovisual testing.

Training. The four plastic toys to be used in testing were left with the two preschool teachers 1 month prior to the commencement of testing. The toys were brought out every day at reading time and the teachers would remind the children of the name of each creature. One month following these tasks, the experimenter returned to the preschool to assess whether the children had learned the names of the stimuli; any child that had not learned the names of the toys was eliminated from the experiment at this time.

Articulatory assessment. In order to assess whether children made substitu-
tion errors, children were asked to repeat four sentences that corresponded with a picture in a children’s book:

1. The boy is having a hot dog.
2. The dog is playing with the ball.
3. The tree is very thick.
4. The other ball is over there.

One of the most common errors is to substitute other consonants for /th/ sounds. (The test sentences contain seven instances of a voiceless interdental /θ/ and one instance of a voiced interdental /ð/; /th/ will be used to indicate both the voiceless and the voiced interdental.) Thus the test sentences employed four instances of /th/ in sentence-initial position, three instances of /th/ in word-initial position, and one in word-medial position. As well, the test sentences included three instances of /b/ (word-initial), two instances of /d/ (word-initial), and two instances of /v/ (one word-initial, one word-medial).

Children who made two or more errors on the highlighted consonants were categorized as substituters. As expected, virtually all errors occurred on /th/. Children who made no articulatory errors were considered nonsubstituters. Two children were unable to be classified unambiguously by this method. Thus, these two children were engaged in further conversation to determine if the children would show correct pronunciation or substitution errors. In both cases the children showed substitution errors.

Visual-only and auditory-only testing. One week following the articulatory assessment, children were presented with the visual-only condition followed by the auditory-only condition. Children were tested individually in a sectioned-off area of the preschool. The child was seated approximately 3 feet in front of a video monitor. On a table to the side and slightly in front of the child were the four plastic toys: /ba/, /va/, /ða/, and /da/. During testing, a screen was placed in front of the toys to avoid distracting the child’s attention from the video monitor. The experimenter was seated off to the side with his back to the monitor while facing the child. Testing proceeded one trial at a time. When the experimenter could see that the child was clearly looking toward the TV monitor, he would press a button on the VCR to present the next stimulus. When the counter on the VCR had advanced to the appropriate location, he pressed pause. In this way the experimenter was blind to the visible articulation but was able to ensure that the child was watching the screen before he pressed the button on the VCR to present the next stimulus. If the child looked away during any one of the trials, that trial number was noted and then repeated for the child at the end of the experiment.

After each syllable was presented, the tape was paused, the screen was removed from in front of the toys, and the child was asked to identify the toy that corresponded to the stimuli presented on the monitor: “OK, now which one of these toys did that guy just name?” The toy selected by the child was recorded rather than the syllable said by the child to avoid confusion
resulting from possible articulatory errors. For example, a child might say /da/ instead of /ða/ because of difficulty producing a /θ/, but still select the correct toy. Once the child made his/her choice, he/she was verbally reinforced regardless of the toy selected, then we proceeded to the next syllable. As an additional incentive, after completing a succession of five test trials, each child received a stick-on gold star which they were allowed at that time to put on their certificate (some children were more than happy to put them on their clothes or face).

The same procedure was followed for the auditory-only condition except that the speaker’s face was no longer visible and the sound was turned on.

Audiovisual condition. One week later, the experimenter returned to conduct testing in the audiovisual condition. The procedure was identical to that of the previous conditions except that the tape used was composed of the one congruent syllable and the three incongruent syllables. The lower portion of the speaker’s face was visible on the monitor and the sound was turned on.

Adults. The procedure for the adult participants was similar to that for the children, however, adults simply responded verbally to the syllables. They were instructed first to give the consonant of the syllable they just heard and then the syllable itself in order to make sure the experimenter heard the adult’s response correctly. Adults were also told to select from one of four alternatives: /ba/, /va/, /ða/, or /da/. We chose to limit adults’ response set in order that their responses would conform to the choices given the children. Adults were tested in the auditory-only, visual-only, and audiovisual conditions in the same session.

Scoring. A response was scored as correct if the child selected the toy corresponding to the heard syllable in the auditory-only condition or to the seen syllable in the visual-only condition. Proportion correct scores were calculated for each subject for each condition. For the audiovisual condition, a response was considered to represent visual capture if the child selected the toy corresponding to the visual portion of the audiovisual tokens. For example, if the audio /ba/-visual /va/ syllable was presented and the child selected the gorilla (‘Va’), this would be scored as visual capture. Proportion visual capture scores for each syllable were calculated in this way. Proportions of other kinds of responses (blends and auditory capture) were also noted. Adults’ oral responses were scored in the same fashion.

RESULTS

Articulatory Assessment

Nine of the 16 children (6 boys, 3 girls) were assigned to the category substituters and 7 children (5 boys, 2 girls) were assigned to the category nonsubstituters as a result of the assessment. An unpaired \( t \) test, \( t(14) = -1.995, p = .07 \), indicated that although the substituters (\( M = 4 \) years 0 months, \( SD = 8 \) months) were on average slightly younger than the nonsubstitu-
tuters (\(M = 4 \text{ years 7 months}, SD = 5 \text{ months})\), the difference was not statistically significant.

However, to rule out the possibility that any differences between the substituters and the nonsubstituters in any of the conditions were merely due to the substituters being, on average, slightly younger than the nonsubstituters, the analyses were also carried out using a reduced sample. To form the reduced sample, we removed the 4 youngest children (3 boys, 1 girl) from the substituter group and the oldest child from the nonsubstituter group (1 boy). This resulted in a sample of 11 children of overlapping ages: 5 children in the substituter group aged 4 years, 4 months to 5 years, 2 months, and 6 children in the nonsubstituter group, aged 4 years, 2 months to 4 years, 9 months. An unpaired \(t\) test, \(t(9) = 0.436, p = .67\), indicated that the substituters (\(M = 4 \text{ years 7 months}, SD = 4 \text{ months}\)) and the nonsubstituters (\(M = 4 \text{ years 6 months}, SD = 3 \text{ months}\)) in the reduced sample did not differ in age.

**Auditory-Only Condition**

Proportion correct scores for each of the four auditory tokens were submitted to a mixed analysis of variance with Group (substituters, nonsubstituters, and adults) as a between factor and Syllable (/ba/, /va/, /da/) as a within-subjects factor. Neither the main effect for Group, \(F(2,29) = 2.120, p = .14\), nor the main effect for Syllable, \(F(3,87) = .923, p = .43\), were significant, suggesting that substituters (\(M = .935, SD = .192\)), nonsubstituters (\(M = .952, SD = .149\)), and adults (\(M = .875, SD = .210\)) were equally accurate at reporting the heard syllables and that there were no differences in accuracy between the syllables. Likewise, there was no significant interaction between Group and Syllable, \(F(6,87) = 1.619, p = .15\), suggesting that all three groups were equally accurate in identifying the different syllables. (See Fig. 1 for comparisons of syllables and groups in the auditory-only condition.) However, as we were interested in whether there were significant differences between adults and nonsubstituters and between nonsubstituters and substituters on each individual syllable, planned contrasts were carried out. As can be seen in Table 1, the adults did not differ from the nonsubstituters on any of the four syllables. Likewise, the nonsubstituters did not differ from the substituters on any of the four syllables.

The failure to find significant differences on the syllables in the auditory-only condition indicates that both preschool groups and the adult group are able to identify the heard syllables and confirms that preschoolers are able to use the appropriate toy to identify the test syllable.

**Visual-Only Condition**

The proportion correct scores for each of the four visual tokens were submitted to a mixed analysis of variance with Group as a between factor and Syllable as a within-subjects factor. There was a significant main effect for Group, \(F(2,29) = 23.341, p = .0001\). A planned contrast comparing adults
and non-substituters, $F(1,29) = 5.204, p = .03$, indicated that the adults ($M = .927, SD = .183$) were significantly better lip-readers than the nonsubstituters ($M = .762, SD = .350$). Moreover, a second planned contrast comparing nonsubstituters and substituters, $F(1,29) = 12.943, p = .0012$, indicated that nonsubstituters ($M = .762, SD = .350$) lip-read more accurately than substituters ($M = .472, SD = .351$).

There was also a significant main effect for syllable type, $F(3,87) = 10.541, p = .0001$, suggesting that participants were not equally accurate on all the

![Fig. 1. Mean proportion correct in the auditory-only condition.](image)

<table>
<thead>
<tr>
<th>Syllable</th>
<th>Comparison</th>
<th>SS ($df$)</th>
<th>$F^*$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ba/</td>
<td>S vs NS</td>
<td>0.872 (1)</td>
<td>2.725</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>0.024 (1)</td>
<td>0.075</td>
<td>n.s.</td>
</tr>
<tr>
<td>/va/</td>
<td>S vs NS</td>
<td>0.030 (1)</td>
<td>0.094</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>1.101 (1)</td>
<td>3.440</td>
<td>n.s.</td>
</tr>
<tr>
<td>/ða/</td>
<td>S vs NS</td>
<td>0.234 (1)</td>
<td>0.731</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>0.667 (1)</td>
<td>2.084</td>
<td>n.s.</td>
</tr>
<tr>
<td>/ða/</td>
<td>S vs NS</td>
<td>0.526 (1)</td>
<td>1.644</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>0.000 (1)</td>
<td>—</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

$F^*$ ratios were calculated with both the interaction error term (Syllable $\times$ Subjects/groups) and with an aggregate of the interaction and between groups error terms. Both methods of calculating the $F$ ratio yielded the same levels of significance for all comparisons.
TABLE 2
Significance Levels of Tukey’s HSD Comparisons between Mean Proportion Correct Identification of the Different Visual Syllables in the Visual-Only Condition

<table>
<thead>
<tr>
<th>Visual syllable</th>
<th>/ba/</th>
<th>/va/</th>
<th>/ða/</th>
<th>/da/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ba/</td>
<td>—</td>
<td>n.s.</td>
<td>n.s.</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>/va/</td>
<td>—</td>
<td>—</td>
<td>n.s.</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>/ða/</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>/da/</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

syllables. A planned orthogonal contrast was computed to test the hypothesis that a linear decrease in accuracy occurs as place of articulation moves further back in the mouth. This contrast yielded a significant value, $F(1,87) = 24.764$, $p < .0001$, confirming that as place of articulation moves back in the mouth, lip-reading becomes more difficult. The mean proportion correct scores suggest that participants were most accurate on /ba/ ($M = .906, SD = .211$), followed by /ða/ ($M = .792, SD = .347$) and /va/ ($M = .760, SD = .319$), followed by /da/ ($M = .594, SD = .386$). However, as shown in Table 2, follow-up comparisons using Tukey’s HSD procedure indicated that the mean proportion correct for /da/ was significantly less than the mean proportion correct for /ba/, /va/, and /ða/, while none of the other means differed from each other.

The analysis of variance also revealed a nonsignificant interaction between Group and Syllable, $F(6,87) = 1.650$, $p = .143$, indicating that the pattern of responding to the syllables did not differ systematically between the three groups (as seen in Fig. 2). However, as we were interested in whether there were significant differences between adults and nonsubstituters and between non-substituters and substituters on each individual syllable— which might be obscured in the two-way interaction— planned contrasts were carried out. Although overall adults are significantly better at lip-reading than nonsubstituters, when each syllable is considered separately, as can be seen in Table 3, adults were only significantly better than nonsubstituters at lip-reading /ða/. Nonsubstituters were better than substituters at lip-reading three of the four syllables: /va/, /ða/ and /da/. The only syllable on which there was no difference between the two preschool groups was /ba/; both preschool groups showed high levels of accuracy lip-reading this syllable. These patterns indicate that nonsubstituters are more adult-like in their ability to lip-read than are substituters.

Using the reduced preschool sample and the adults, the same analyses were carried out as with the full sample. A mixed analysis of variance with Group as a between factor and Syllable as a within-subjects factor yielded a significant main effect for Group, $F(2,24) = 46.803$, $p = .0001$, as did the analysis with the full sample. A planned contrast comparing adults and nonsubstituters
failed to show a significant difference, $F(1, 24) = 1.701$, $p = .20$, unlike in the full sample analysis. A second planned contrast comparing nonsubstituters and substituters, $F(1, 24) = 50.322$, $p = .0001$, replicated the significant difference obtained in the full sample analysis.

A significant main effect for Syllable, $F(3, 72) = 15.469$, $p = .0001$, was

\[\text{TABLE 3} \]
Substituters (S) vs Nonsubstituters (NS) and Nonsubstituters vs Adults (A):
SS and F Values for Each Syllable in the Visual-Only Condition

<table>
<thead>
<tr>
<th>Syllable</th>
<th>Comparison</th>
<th>SS (df)</th>
<th>$F^\alpha$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ba/</td>
<td>S vs NS</td>
<td>0.688 (1)</td>
<td>1.439</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>0.238 (1)</td>
<td>0.498</td>
<td>n.s.</td>
</tr>
<tr>
<td>/va/</td>
<td>S vs NS</td>
<td>4.305 (1)</td>
<td>9.006</td>
<td>&lt;.01</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>1.311 (1)</td>
<td>2.740</td>
<td>n.s.</td>
</tr>
<tr>
<td>/da/</td>
<td>S vs NS</td>
<td>4.305 (1)</td>
<td>9.006</td>
<td>&lt;.01</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>2.416 (1)</td>
<td>5.054</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>/da/</td>
<td>S vs NS</td>
<td>6.712 (1)</td>
<td>14.042</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>1.272 (1)</td>
<td>2.661</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

$^\alpha F$ ratios were calculated with both the interaction error term (Syllable $\times$ Subjects/groups) and with an aggregate of the interaction and between-groups error terms. Both methods of calculating the $F$ ratio yielded the same levels of significance for all comparisons except for the S vs NS comparison for visual /da/, $F(1, 87) = 11.396$, $p < .01$. 
noted, as in the analysis with the full sample. However, unlike with the full sample, a significant interaction between Group and Syllable was observed, $F(6,72) = 2.920, p = .013$. Planned contrasts between adults and nonsubstituters and between nonsubstituters and substituters for each of the syllables were carried out with the reduced sample. In this analysis, adults did not differ from non-substituters on any of the four syllables. Nonsubstituters differed from substituters in the accuracy with which they lip-read three of the four syllables, /va/, /ða/, and /da/. In general, this analysis with the reduced sample echoes the results of the analysis using the full sample: nonsubstituters are more adult-like in their ability to lip-read than are substituters. The only differences between the results of this analysis and those with the full sample is that with a smaller sample (and reduced power) the overall difference between nonsubstituters and adults disappeared, and on the syllable /ða/ the difference between adults and nonsubstituters no longer reached significance.

**Audiovisual Condition**

*Visual capture.* The primary dependent variable in this condition was the proportion of visual capture—percepts reflecting only the visual portion of the syllable—reported by the participants. The proportion visual capture for each of the four audiovisual tokens (one congruent, three incongruent) was submitted to a mixed analysis of variance with Group as a between factor and Visual Syllable as a within-subjects factor. There was a significant main effect for Group, $F(2,29) = 30.506, p = .0001$. A planned contrast comparing the adults and nonsubstituters, $F(1,29) = 9.092, p = .0053$, indicated that the adults ($M = .828, SD = .254$) reported more visual capture than the nonsubstituters ($M = .671, SD = .279$). A second planned contrast, comparing the nonsubstituters and substituters, $F(1,29) = 13.954, p = .0008$, indicated that the nonsubstituters ($M = .671, SD = .279$) reported significantly more visual capture than the substituters ($M = .456, SD = .330$).

There was also a significant main effect for Visual Syllable, $F(3,87) = 22.994, p = .0001$, suggesting that there were differences among syllables in the proportion visual capture. As shown in Table 4, follow-up comparisons using Tukey’s HSD procedure indicated that participants reported significantly more visual capture to audio /ba/—visual /va/ ($M = .637, SD = .298$)
Fig. 3. Mean proportion of percepts corresponding to the visual component in the audiovisual condition. A single asterisk indicates group differences significant at the $p < .05$ level and two asterisks indicate significant differences at the $p < .01$ level.

than to audio /ba/ ± visual /da/ ($M = .475, SD = .372$) and more visual capture to audio /ba/ ± visual /da/ ($M = .731, SD = .272$) than to audio /ba/ ± visual /da/ ($M = .475, SD = .372$).

The analysis of variance also revealed a nonsignificant interaction between Group and Visual Syllable, $F(6, 87) = 2.160, p = .06$, indicating that the pattern of responding overall to the syllables did differ between the three groups. The overall pattern can be seen in Fig. 3. As we were primarily interested in how each group performed on each individual syllable, planned contrasts were carried out to determine whether there were significant differences between adults and nonsubstituters and between nonsubstituters and substituters on each of the syllables. As can be seen in Table 5, for the audiovisual /ba/, the adults did not differ from the nonsubstituters and the nonsubstituters did not differ from the substituters; all performed at close to ceiling levels. For the audio /ba/ ± visual /va/, the adults showed significantly more visual capture than did the nonsubstituters, but the nonsubstituters did not differ significantly from the substituters. For the audio /ba/ ± visual /da/, adults did not differ significantly from the nonsubstituters in proportion of visual capture; however, the nonsubstituters showed significantly more visual capture than the substituters. For the audio /ba/ ± visual /da/, adults reported significantly more visual capture than the nonsubstituters who reported significantly more visual capture than did the substituters.

As in the visual-only condition, the pattern of responding in the audiovisual condition indicates that nonsubstituters are more adult-like in the proportion
TABLE 5
Substituters (S) vs Nonsubstituters (NS) and Nonsubstituters vs Adults (A):
SS and $F$ Values for Each Syllable in the Audiovisual Condition

<table>
<thead>
<tr>
<th>Syllable</th>
<th>Comparison</th>
<th>SS ($df$)</th>
<th>$F^*$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ba/</td>
<td>S vs NS</td>
<td>1.545 (1)</td>
<td>1.178</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>0.438 (1)</td>
<td>0.334</td>
<td>n.s.</td>
</tr>
<tr>
<td>/va/</td>
<td>S vs NS</td>
<td>2.638 (1)</td>
<td>2.011</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>5.539 (1)</td>
<td>4.222</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>/ða/</td>
<td>S vs NS</td>
<td>15.528 (1)</td>
<td>11.835</td>
<td>&lt;.01</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>0.010 (1)</td>
<td>0.008</td>
<td>n.s.</td>
</tr>
<tr>
<td>/da/</td>
<td>S vs NS</td>
<td>6.712 (1)</td>
<td>5.116</td>
<td>&lt;.05</td>
</tr>
<tr>
<td></td>
<td>NS vs A</td>
<td>13.779 (1)</td>
<td>10.502</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

$^a$ $F$ ratios were calculated with both the interaction error term (Visual Syllable $\times$ Subjects/groups) and with an aggregate of the interaction and between groups error terms. Both methods of calculating the $F$ ratio yielded the same levels of significance.

of visual capture reported than are substituters. In fact, adults and nonsubstituters are virtually the same in the proportion of visual capture reported for the audio /ba/—visual /ða/, whereas substituters reported significantly less visual capture for the same audiovisual pair.

Using the reduced preschool sample and the adults, the same analyses were carried out as with the full sample. A mixed analysis of variance with Group as a between factor and Syllable as a within-subjects factor yielded a significant main effect for Group, $F(2,24) = 18.742$, $p = .0001$, as did the analysis with the full sample. A planned contrast comparing adults and nonsubstituters, $F(1,24) = 7.519$, $p = .0014$, revealed a significant difference between the two groups as was revealed in the full sample analysis. Likewise, a planned comparison between nonsubstituters and substituters, $F(1,24) = 8.422$, $p = .0078$, replicated the significant difference observed in the full sample analysis.

A significant main effect for Syllable, $F(3,72) = 14.978$, $p = .0001$, was obtained, as in the analysis with the full sample. However, as in the full sample analysis, the reduced sample analysis did not yield a significant interaction between Group and Syllable, $F(6,72) = 1.324$, $p = .26$. Planned contrasts between adults and nonsubstituters and between nonsubstituters and substituters for each of the syllables were still carried out with the reduced sample to test our a priori hypotheses. Adults showed significantly more visual capture for audio /ba/—visual /da/ than did nonsubstituters whereas in the analysis with the full sample, adults showed, in addition, significantly more visual capture on audio /ba/—visual /va/. As in the full sample analysis, nonsubstituters showed significantly more visual capture than did substituters for two of the three incongruent audiovisual pairings: audio /ba/—visual /ða/ and audio /ba/—visual /da/. Thus, the analyses with the reduced sample mirror
the findings with the full sample for all but one contrast and in fact provide slightly stronger support for the conclusion that nonsubstituters are more adult-like in the proportion of visual capture reported than are substituters.

Some studies of audiovisual speech perception use a broader definition of visual influence. In addition to visual capture, they include responses which show visual influence (e.g., Hockley, 1994; McGurk & MacDonald, 1976). The additional responses that could be considered as showing visual influence in our study are the following: hearing an audio /ba/—visual /da/ as /va/, or an audio /ba/—visual /da/ as either /va/ or /da/. Thus, to examine overall visual influence, rather than just visual capture alone, we recomputed the analysis of overall visual influence as the dependent variable. This analysis yielded the same pattern of results to that of the visual capture analysis with one small exception. The significant difference between adults and nonsubstituters on the audio /ba/—visual /da/ disappeared in this analysis.

**Auditory capture.** It is possible that children showed less visual capture than adults because they were simply attending to only the auditory component. Indeed the results of previous research (e.g., Hockley, 1994; Hockley & Polka, 1994; Massaro, 1984; Massaro, Thompson, Barron, & Laren, 1986) reveal that children show more auditory capture than adults. To explore this possibility we analyzed the proportion of auditory capture responses.

The proportion of auditory capture responses for each of the four audiovisual tokens (one congruent, three incongruent) was submitted to a mixed analysis of variance with Group as a between factor and Visual Syllable as a within-subjects factor. There was a significant main effect for Group, $F(2,29) = 6.101, p = .0061$. A planned contrast comparing the adults and nonsubstituters, $F(1,29) = 5.485, p = .0263$, indicated that the adults ($M = .284, SD = .415$) reported less auditory capture than the nonsubstituters ($M = .371, SD = .388$). A second planned contrast, comparing the nonsubstituters and substituters, $F(1,29) = 3.10, p = .58$, indicated that the nonsubstituters ($M = .371, SD = .388$) did not differ from the substituters ($M = .394, SD = .359$) in the proportion of auditory capture responses.

The main effect for Syllable was significant, $F(3,87) = 143.700, p < .0001$, indicating there were significant differences in the proportion of auditory capture reported to the four syllables. Follow-up comparisons using Tukey’s HSD procedure indicated that there was significantly more auditory capture to audio /ba/—visual /va/ than to either audio /ba/—visual /da/ ($p < .01$) or audio /ba/—visual /da/ ($p < .01$).

The interaction between Group and Visual Syllable was significant, $F(6,87) = 6.385, p = .0001$, indicating that the pattern of responding to the syllables differed between the three groups. As we were primarily interested in how each group performed on each individual syllable, planned contrasts were carried out to determine whether there were significant differences between adults and nonsubstituters and between nonsubstituters and substituters on each of the syllables. Adults showed significantly less auditory capture than
nonsubstituters on only one syllable: audio /ba/ -- visual /va/, $F(1,87) = 16.523, p < .001$. Nonsubstituters showed less auditory capture than substituters on only one syllable: audio /ba/ -- visual /da/, $F(1,87) = 8.865, p < .01$.

Using the reduced preschool sample and the adults, the same analyses were carried out as with the full sample. A mixed analysis of variance with Group as a between factor and Syllable as a within-subjects factor yielded a nonsignificant main effect for Group, $F(2,24) = 3.101, p = .06$, whereas this effect was significant in the analysis using the full sample. A planned contrast comparing the adults and nonsubstituters, $F(1,24) = 4.462, p = .0452$, revealed a significant difference between the two groups as did the full sample analysis. A second planned contrast comparing nonsubstituters and substituters, $F(1,24) = .018, p = .89$, showed no significant difference between the two groups, as in the full sample analysis.

There was a significant main effect for Syllable, $F(3,72) = 153.556, p = .0001$, and a significant interaction between Group and Syllable, $F(6,72) = 6.510, p = .0001$, as in the analysis with the full sample. Planned contrasts between adults and nonsubstituters and between nonsubstituters and substituters for each of the syllables were still carried out with the reduced sample to test our a priori hypotheses. An identical pattern of results was obtained with the reduced sample as with the full sample. Adults showed less auditory capture than nonsubstituters on only one of the incongruent pairs, audio /ba/ -- visual /va/, $F(1,72) = 15.344, p < .001$. Nonsubstituters showed less auditory capture than substituters on only one of the incongruent pairs, audio /ba/ -- visual /da/, $F(1,72) = 8.497, p < .01$.

In summary, the adults showed less auditory capture than the nonsubstituters, but this effect was carried entirely by the audio /ba/ -- visual /va/. There was no overall difference between nonsubstituters and substituters in proportion of auditory capture, however, substituters were more likely than nonsubstituters to report auditory capture for audio /ba/ -- visual /da/.

**DISCUSSION**

The purpose of this study was to begin to examine why children appear to be much less influenced by visible speech than adults. The phenomenon of less visual influence among children than adults has been reported repeatedly in the literature, but there is no clear understanding as to why it exists. Our study represents the first attempt not just to report this fact, but to explain it more fully. One possibility we needed to address first, however, was that in using the standard task, previous researchers had perhaps underestimated the amount that visual information influences speech perception in children. To this end we modified the methodology to make it more engaging with the expectation that it might provide a more accurate measure of the extent of visual influence.

We did find our procedure to be indeed child-friendly. The children remained eager and interested in playing our game and would have happily
continued playing even after testing was completed. Using toys whose names corresponded to the test syllables provided an interesting activity for children: picking up the toy named by “the man on the TV.” Moreover, having children select a toy rather than repeating what they heard avoided potential problems resulting from articulatory difficulties. For example, in the standard task a child’s inability to produce /ða/ may lead to the experimenter recording that child’s percept to record a /da/ when the child may have been trying to say /ða/. This problem is controlled in our procedure. The fact that there were no differences between groups in the auditory-only condition—children were actually slightly more accurate than the adults—confirms that our procedural change yields reliable results. And, because children performed so well in the auditory-only condition, we are more confident that any differences in performance between children and adults in the other two conditions (those which depend upon visual influence) are meaningful.

The results we obtained in the audiovisual condition were very revealing. Although we replicated the previously reported finding of greater auditory capture and less visual capture in children than adults, the overall amount of visual capture was greater for the children in our study than in most previous studies. The preschool children reported levels of visual capture—averaged across the incongruent audiovisual stimuli—at 57% of the levels reported by adults for the same stimuli. In previous studies the proportion of visual capture reported by children tested in these conditions ranged from less than 10% of adult levels to approximately 50% of adult levels (Hockley, 1994; Massaro, 1984; Massaro, Thompson, Barron, & Laren, 1986; Polka & Hockley, 1994). That more visual capture was reported by children in our study than in most of the previous studies makes us confident that our new methodology is providing a sensitive index of visual influence in children.

Even with our new methodology, however, preschoolers still reported much less visual influence than adults. Thus, we pursued the primary goal of this research: to begin to explore why it is that children are less influenced by the visual information in speech than are adults. The hypothesis we tested is that preschool children have not yet had the opportunity to specify fully their representations of visible speech because they have had less experience with speech than have adults. One form of experience that should, theoretically, be related to the completeness of the representation is experience correctly producing speech. This is a form of experience that can be isolated by testing normally developing children who are the same age, but who differ in their ability to correctly produce consonants. Thus, we selected a sample of preschoolers, divided them into two groups on the basis of whether they are substituters or nonsubstituters, and compared the extent of visual influence on their perception of speech.

Overall, our results indicate that, among preschoolers, substituters are poorer lip-readers and tend to be less influenced by the visual component of an incongruent audiovisual syllable than are nonsubstituters. In the visual-
only condition, substituters correctly identified the test syllables 47% of the time whereas nonsubstituters correctly identified the test syllables 76% of the time; this is a substantial difference between the groups and clearly indicates that the substituters are poorer lip-readers than nonsubstituters. When we reduced the sample size to make the groups more age equivalent, the substituters correctly identified the test syllables 52% of the time, whereas nonsubstituters correctly identified the test syllables 88% of the time. Thus the difference in lip-reading ability between the two groups is not due to an overall age difference.

In the audiovisual condition the substituters report percepts of visual capture slightly over half as often as do nonsubstituters (34% of time for substituters and 59% of the time for nonsubstituters). When the reduced sample was used, thereby eliminating overall age differences between the groups, substituters still reported much less visual capture than did nonsubstituters (37 and 60%, respectively). As in the visual-only condition, we see a substantial difference between the two groups of preschoolers in the extent of visual influence. Importantly there was not a corresponding difference between the two groups of preschoolers in the proportion of auditory capture. This shows that the substituters were not merely attending to only the auditory component but rather were also influenced by the visual information. However, the influence of the visual information was less consistent for the substituters than for the nonsubstituters. This is not surprising given that the substituters were poorer lip-readers than the nonsubstituters in the visual-only condition. Taken together, these findings are entirely compatible with the hypothesis that the underlying representation of visible speech is mediated by children’s own ability to produce consonants correctly.

Interestingly, although on the articulatory test most of the errors made by substituters were made on the /th/, substituters in fact showed less visual influence than non-substituters on three of the four places of articulation. This pattern suggests that it does not matter which consonant(s) children have difficulty producing correctly. Children who are still making errors on single consonants at age 4—a time when many children produce single consonant correctly—may have only recently acquired the correct pronunciation of other consonants. The lack of a sufficiently detailed representation of the visual information may not correspond precisely to the articulations that the child can accurately perform at any given time, but instead may be influenced by the length of time the child has been accurately producing consonants more generally.

The interpretation that it is length of time correctly producing consonants that predicts visual influence (rather than the current correctly produced consonant inventory) could help account for the finding of consistent differences between nonsubstituters and adults. Recall that nonsubstituters showed less visual influence overall than did adults, with significant differences on two of the three places of articulation. It is possible that the children who are
now nonsubstituters were making at least some substitution errors when they were younger. Adults, on the other hand, have presumably been producing consonants correctly for a number of years. It could be that the more extensive experience correctly producing consonants that accounts for the sizable difference in visual influence between adults and nonsubstituters. Because adults have had more experience producing consonants correctly, they may have more detailed representations than nonsubstituters. Similarly, nonsubstituters have had more experience producing consonants correctly than substituters and may thus in turn have more detailed representations.

Additional support for the idea that experience producing speech develops the underlying representation comes from a study comparing adults who are unable to produce speech due to severe cerebral palsy with typical adults. Siva, Stevens, Kuhl, and Meltzoff (1995) found that the cerebral palsy group, unlike our substituter group, showed similar levels of visual influence compared to the adult controls when tested on an auditory /aba/-visual /aga/, i.e., they reported perceiving /ada/ or /aða/. However, when tested on an auditory /aga/-visual /aba/, the two groups differed in their responses. As expected, the adult controls reported a “combination” percept of /abga/. The cerebral palsy adults, on the other hand, reported a percept influenced by only the auditory stimulus, /aga/. It is not so surprising that the cerebral palsy adults behave more like adult controls in some conditions whereas our substituters do not. The cerebral palsy adults have had considerably more experience than both our preschool groups watching and listening to speech produced by others and as well the cerebral palsy adults may have been influenced by reading skills whereas the preschool participants likely had little or no experience reading. Of importance, however, when a more difficult task was introduced—the pairing of an auditory /aga/ with a visual /aba/-unambiguous differences in visual influence were found for the cerebral palsy adults compared to the controls; the cerebral palsy adults’ perception of audiovisual speech is affected by their inability to produce speech. Thus, these results provide converging evidence that experience producing speech influences the ability to perceive visible speech.

Although the data reported in this paper support the hypothesis that experience producing speech impacts on the representation of visible speech in children, our research does not address the question of whether this experience is actually necessary for the establishment of an underlying representation that includes visual information or whether there is some initial representation of the visual information even without such experience. A consideration of the research that has been conducted with even younger children can help address this question. Although only a few studies with even younger children have been conducted to date, the existing studies provide fairly convincing evidence that visible speech influences speech perception very early in infancy. Using methodologies appropriate for testing infants (e.g., habituation) researchers have found that young infants’ perceptions of incongruent audiovi-
visual stimuli are consistent with those of adults (e.g., Desjardins & Werker, 1995; Rosenblum, Schmuckler, & Johnson, in press). For example, 4-month-old infants who do not yet produce consonants even in their babbling show something akin to visual capture under certain circumstances. In illustration, in one study one group of infants was familiarized to an audiovisual /bi/ and a second group was familiarized to an audiovisual /vi/; both groups of infants were tested on an audio /bi/-visual /vi/. The logic of the habituation procedure is that infants should show a recovery in looking time only to a stimulus that is perceived as different from the familiarization stimulus. Female infants familiarized to an audiovisual /bi/ show a recovery in looking time to the audio /bi/-visual /vi/ but infants familiarized to an audiovisual /vi/ do not (Desjardins & Werker, 1996). This would suggest that female infants of 4 months, like adults, perceive the audio /bi/-visual /vi/ as /vi/—their percept is captured by the visual /vi/.

Clearly, the habituation methodology used in the Desjardins and Werker (1996) study does not allow us to assess the magnitude of the visual influence in infants relative to adults. It allows us to conclude that infants’ percepts are influenced by both listening to and observing speech produced, but it does not allow us to determine how detailed infants’ underlying representations of visible speech are relative to preschoolers or adults. Nevertheless, the fact that infants’ percepts are influenced by visual information even though they cannot yet produce consonants suggests that experience producing consonants cannot be absolutely essential for establishing an initial representation of visible speech. It would suggest instead that the role experience producing consonants plays is one of filling in details in the underlying representation or recalibrating existing representations. A similar notion is suggested by Meltzoff and Kuhl (1994) who argue that babbling functions to consolidate, rather than establish, auditory-articulatory mappings.

Limitations and Directions for Further Research

We believe the current study provides compelling support for the notion that experience producing speech contributes to the representation used in speech perception, particularly visible speech perception. There are, however, some caveats that need to be considered. First, the suggestion could be made that by dividing our preschoolers into articulatory types, we may have inadvertently created a “more advanced” and a “less advanced” group of children. Indeed, the fact that our nonsubstituters are older (although not significantly so) than our substituters could be interpreted as evidence that more advanced children simply perform more like adults than less advanced children. In an attempt to control for these marginal age differences, we created a reduced sample by removing the four youngest subjects from the substituter group and the oldest subject from the nonsubstituter group. A virtually identical pattern of results was obtained with the reduced sample. Thus although we acknowledge the possibility that the nonsubstituters may be, in some other
unspecified way, more advanced than substituters, we are convinced that it is unlikely to be the primary explanation for our findings for several reasons. First, developmental level was controlled to some degree by requiring that all children—both substituters and nonsubstituters—be able to learn and remember the names for our four figures with only a short presentation period each school day. Actually, we had originally intended to include a separate sample of 3-year-old children, but found the task of learning the names of the characters to be too difficult for almost all children this young. Thus the fact that all children in the current study could successfully learn the characters’ names ensures some uniformity in developmental level. Second, although no standard tests were administered to the children in the preschool, the teachers reported that all children who participated in our study were progressing normally in all ways. Finally, it should be remembered that the substituters performed almost identically to the nonsubstituters in the auditory-only control condition. If an overall developmental difference were impacting negatively on performance of children in the substituter group, it should have influenced their performance in the auditory-only control condition as well. These reasons taken together lead us to be confident that the differences observed between the substituters and the nonsubstituters were likely due to differing experience correctly producing speech.

A second caveat concerns our finding that experience producing speech impacted performance on the tasks requiring lip-reading, but not on the auditory control task. It is possible that if a more difficult auditory speech perception task was used, the substituters may have performed more poorly than the nonsubstituters. For example, in research comparing auditory speech perception in children who were good and poor readers, Brady, Shankweiler, and Mann (1983) found differences between the groups only when background noise was introduced in the speech perception task. If the representations of heard and seen speech are separate, but linked, then even on a more challenging auditory speech perception task, such as that used by Brady et al., the substituters and non-substituters might perform similarly suggesting that experience producing speech correctly impacts only the representation of visible speech. If, on the other hand, the representations of heard and seen speech are in fact one representation, a more challenging task might reveal that substituters perform more poorly than nonsubstituters suggesting that experience producing speech correctly impacts the representation of both heard and seen speech. In future research it would be of interest to disentangle these two possibilities.

A final caveat concerns our suggestion that it is length of time correctly producing consonants that impacts on the detail and usefulness of the representation of visible speech. Our data firmly support our a priori hypothesis that experience correctly producing consonants impacts on the representation of visible speech, but our suggestion concerning length of time correctly producing consonants is post hoc and should be considered only speculation at this
time. The interpretation we have offered implies that children who are now classified as substituters will show increased influence from visible speech as their production improves. We cannot know, however, without longitudinal data whether preschoolers in the substituter group will actually show more accuracy in lip-reading and gradually catch up to nonsubstituters over the course of development or whether they will remain different. It is conceivable, for example, that these children who are now substituters will show some increased influence from visual information as their production improves, but that they will never show the same degree of influence as children who are not substituters. Such a possibility is of interest to consider because it could help to explain the much greater variability seen among adults in lip-reading proficiency than in auditory speech perception (see Summerfield, 1991 for a review). To date, the only factor found in adults to relate to lip-reading proficiency is evoked-response latency to a bright flash of light (Shepard, 1982). Although this research has been interpreted as suggesting a speed-of-processing influence on visible speech perception, it is difficult to evaluate that interpretation; thus we are still unable to explain the tremendous variability seen in adults in lip-reading proficiency. If future research reveals that children who are substituters in the late preschool years never show the same levels of visual influence as do people who were non-substituters by age 4, then we may be in a better position to begin to understand not only the individual differences in adult proficiency, but also the mechanisms that underlie audiovisual speech perception.

Summary

The research reported in this paper represents the first attempt to explain why it is that preschoolers are less influenced by the visual information in speech than are adults. In this endeavor, we developed a new procedure for testing children which yields higher levels of visual influence (relative to adults) than have been reported in some previous studies. And, it is a procedure that children clearly enjoy. We are hopeful this procedure will be of use to other researchers investigating audiovisual speech perception in children and in other populations who are unable to use standard methods to indicate their choices.

Most importantly, in our new age-appropriate methodology, preschool children who made substitution errors were poorer at lip-reading and showed less visual influence on their perception of audiovisual tokens than did preschoolers who do not make substitution errors. This research thus provided the first direct support for the hypothesis that experience—in particular, experience producing speech—does play a role in the elaboration of the representation of visible speech. We believe this finding not only helps understand age related differences in audiovisual speech perception, but also points to interesting and theoretically important directions for future research.
REFERENCES


Received: June 4, 1996; revised: February 24, 1997