(2003) Journal on Developmental and Learning Disorders, 7,111-144.

PERCEIVING SPEECH BY EAR AND EYE:

Multimodal Integration by Children with Autism

Dominic W. Massaro, Ph.D. and Alexis Bosseler

Abstract. *We tested whether children with autism integrate information from multiple sensory modalities in speech identification of spoken syllables. An expanded factorial design was used in which information from the face and voice was presented either unimodally or bimodally, and either consistent with one another or not. After training the children speechreading to enhance the influence of visible speech from the face, we repeated the identification task. Children behaved similarly in the two replications, except for a larger influence of the visible speech after training in speechreading. The fuzzy logical model of perception (FLMP) was contrasted with a single-channel model (SCM) because they represent comparable integration and nonintegration models, respectively. The model descriptions revealed that the FLMP gave a significantly better description of performance than the SCM, supporting the interpretation that children with autism integrate vocal and facial information in speech perception. Given these positive findings, we propose multimodal environments for learning language.*

The goal of the present research is to gain some understanding of speech perception in individuals with autism. Autism is a pervasive developmental disorder, which apparently has increased from affecting approximately 1 in every 500 children (Cowley, 2000) to 1 in 300 (M.I.N.D. Institute, University of California, Davis, http://www.dds.ca.gov/Autism/Autism_main.cfm). Although the etiology of autism varies, individuals diagnosed with autism must exhibit a) delayed or deviant language and communication, b) impaired social and reciprocal social interactions, and 3) restricted interests and repetitive behaviors (American Psychiatric Association, 1994). The language and communicative deficits are particularly salient, with large individual variations in the degree to which autistic children develop the fundamental lexical, semantic, syntactic, phonological, and pragmatic components of language (Tager-Flusberg, 1999). For the roughly one-half of the autistic population who develop some form of functional language (Tager-Flusberg, 2000; Lord, Rutter, & LeCouteur, 1994; Prizant, 1983), the onset and rate at which the children pass through linguistic milestones are often delayed (e.g. no single words by age 2 years, no communicative phrases by age 3; American Psychiatric Association, 1994). Given their limitations in language processing, a better understanding of their speech perception should be particularly valuable. We begin with our extant view of the psy-

chological processes involved in speech perception, followed by a consideration of how children with autism might be at a disadvantage in this domain.

Speech Perception

We define speech perception as the process of imposing a meaningful perceptual experience on an otherwise meaningless speech input. The empirical and theoretical investigation of speech perception has blossomed into an active interdisciplinary endeavor, including the fields of psychophysics, neurophysiology, sensory perception, psycholinguistics, linguistics, artificial intelligence, and sociolinguistics. In any domain of perception, one goal is to determine the stimulus properties responsible for perception and recognition of the objects in that domain. The study of speech perception promises to be even more challenging than other domains of perception because it crosses all of these disciplines.

Our perception and understanding of speech is a multimodal process, influenced by what we hear (the sound of the speakers voice) and what we see of the face and accompanying gestures (Massaro, 1998). Research has repeatedly shown that pairing the auditory speech with visual speech from the face produces a percept that is more accurate and less ambiguous relative to presenting either of these modalities alone (Massaro, 1984; Summerfield and McGrath, 1984). Viewing the speaker's face increases the intelligibility of what is being said, especially when the auditory information is degraded by noise (Sumby & Pollack, 1954) or hearing loss (Erber, 1969). For example, viewing the speaker's face can improve intelligibility of the spoken message as much as 15 dB in the speech to noise ratio (Sumby & Pollack, 1954).

Viewing the speaker's face to augment the spoken message is not limited to situations in which the auditory input is degraded. Perhaps the most compelling demonstration of the impact of visible speech on perception of the spoken message is the McGurk effect (McGurk & Macdonald, 1976). In this classic demonstration, participants were presented a film of a young woman saying /aga/ that was dubbed with the sound /aba/. The participants often reported hearing /ada/, putatively a fusion of the place of articulation features of /aga/ and the manner and voicing features of /ba/ (we provide an alternative explanation after our theoretical framework is developed). When the dubbing process was reversed (an auditory /aga/ dubbed onto /aba/ lip movements) participants sometimes reported hearing /abga/, a combination of the two syllables. Similar results were found with /pa/ and /ka/. This McGurk effect provides evidence that speech perception is a bimodal process, influenced by both the sight and sound of the speaker. A theoretical account of bimodal speech perception must describe how each source of information is evaluated, whether or how the sources are combined or integrated, and how classification decisions are made.

The Fuzzy Logical Model of Perception (FLMP)

In the course of our research, we have found that the Fuzzy Logical Model of Perception (FLMP) to be a universal principle of perceptual cognitive performance that accurately models human pattern recognition (Massaro, 1998). People are influenced by multiple sources of information in a diverse set of situations. These sources of information are often ambiguous and any particular source alone does not usually specify completely the appropriate interpretation.

The three processes involved in perceptual recognition are illustrated in Figure 1 and include evaluation, integration, and decision. These processes make use of prototypes stored in long-term memory. The evaluation process transforms these sources of information into psychological values, which are then integrated to give an overall degree of support for each speech alternative. The decision operation maps the outputs of integration into some response alternative. The response can take the form

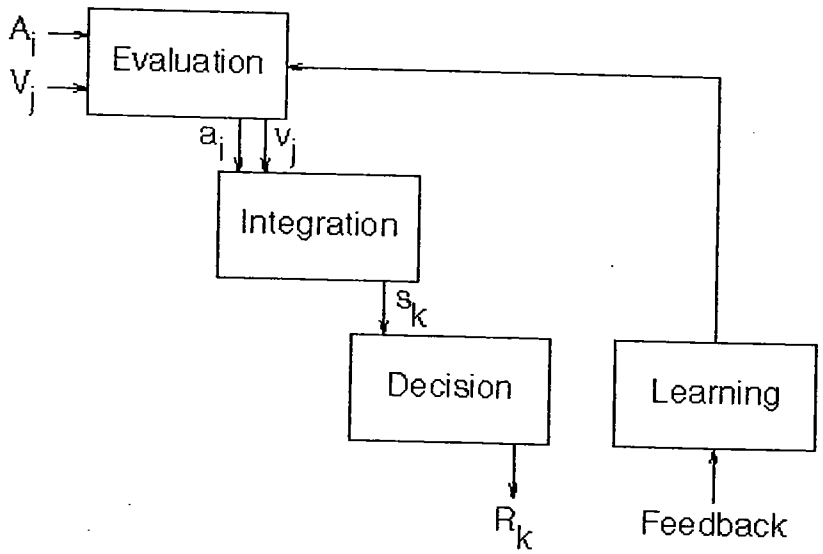

FIGURE 1. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by uppercase letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j) These sources are then integrated to give an overall degree of support, s_k, for each speech alternative k. The decision operation maps the outputs of integration into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. The feedback is assumed to tune the prototypical values of the features used by the evaluation process.

of a discrete decision or a rating of the degree to which the alternative is likely. The assumptions central to the model are: 1) each source of information is evaluated to determine the continuous degree to which that source specifies various alternatives, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated to provide an overall continuous degree of support for each alternative, and 4) perceptual identification and interpretation follows the relative degree of support among the alternatives.

Although speech perception has traditionally been viewed as a unimodal process, it appears to be a prototypical case of multimodal perception. As described in the introduction, experiments in face-to-face communication have revealed conclusively that our perception and understanding are influenced by a speaker's face, as well as the actual sound of the speech (Massaro, 1998). Research has shown that the results are well-described by the FLMP, which is an optimal integration of the two sources of information (Massaro, 1998). A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. For example, if the ambiguous auditory sentence, *My bab pop me poo brive,* is paired with the visible sentence, *My gag kok me koo grive,* the perceiver is likely to hear, *My dad taught me to drive.* Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro, 1998).

Development of Bimodal Speech Perception

The multimodal nature of speech perception has been observed across development and aging. Young infants have been shown to be sensitive to the correspondence between the auditory and visual speech (Kuhl & Meltzoff, 1982; Rosenblum, Schmuckler, & Johnson, 1997). Nonetheless, several studies reveal that the perceptual judgments of young children show less of a visual influence when compared to adults (Massaro, 1984; Massaro, Thompson, Barron, & Laren, 1986; McGurk & McDonald, 1976). Massaro (1984) compared performance of preschool children to college students and found that the major difference between the two groups was the overall contribution of the visual speech: the adults showed a greater visual influence relative to the children. Even though the preschool children were less influenced by the visible speech, they appeared to integrate the audible and visible speech in the same manner as adults. For both groups, the FLMP gave a significantly better description of performance than a nonintegration model. Given this support for integration in young children, it is valuable to ask whether autistic children also integrate audible and visible speech in speech perception.

Multimodal Integration in Children with Autism

A critical assumption of the FLMP is that perceivers integrate information from the face and the voice in speech perception. It has long been suggested, however, that

individuals with autism are impaired in both their face processing (Dawson et. al, 2002; Rogers & Pennington, 1991; Williams et al., 2001), and their ability to integrate information across modalities (i.e. Bryson, 1970; de Gelder, Vrooman, & Van der Heide, 1991; Lelord, Laffont, Jusseaume, & Stephant, 1973; Martineau, Garreau, Roux, & Lelord, 1987; Waterhouse, Fein, & Modahl, 1996). However, the exact nature of these impairments has not been specified and the evidence is limited. For example, children with autism tend to avoid the face to face contact with others required by shared attention (Happe, 1996) and, therefore, would naturally have less experience with visual information from the face. It follows that we would expect autistic children to be less influenced by the face in bimodal speech perception but it is possible that they integrate the two sources in the same manner as normally-developing children.

It is therefore essential to distinguish between how much information is obtained from a sensory input and how information from multiple inputs is processed. Given the FLMP framework, we make a formal distinction between "information" and "information processing." The sources of information from the auditory and visual channels make contact with the perceiver at the evaluation stage of processing. The reduction in uncertainty affected by each source is defined as information. In the description given by the FLMP, for example, the degrees of support for each alternative from each modality correspond to information. These values represent how informative each source of information is. Information processing, on the other hand, refers to how the sources of information are processed. In the FLMP, this processing is described by the evaluation, integration, and decision stages (see Figure 1).

Fortunately, there is an ideal experimental paradigm and theoretical analysis that allow us to distinguish information from information processing, and to determine whether integration occurs. The experiment involves the independent manipulation of two sources of information in an expanded factorial design. It allows an assessment of the influence of the many potentially functional cues, and whether or how these cues are combined to achieve speech perception (Massaro, 1998). This systematic variation of the properties of the speech signal and quantitative tests of models of speech perception test how different sources of information are evaluated and how they are actually used.

The goal of the present investigation was to assess speech processing abilities in children with autism. We asked whether any observed differences are the consequence of information or information processing. This distinction allows us to determine whether any decrement in performance is the result of an impairment in crossmodal integration (information processing) or the inability to discern and utilize the auditory and visual information (information). To address these questions our experimental design was carried out in three stages. We began with an identification test, utilizing an expanded factorial design. Given that the children appeared to integrate crossmodally but were influenced only somewhat by the visual speech, we carried out a training regiment in which the children were trained in speechreading. We then repeated the identification task. We predicted that this training would result in an increase in the influence of visible speech in the identification experiment, and that their crossmodal integration would not change.

Method

Participants

Seven children diagnosed with autism, 1 female, 6 males, ranging in age from 7 to 11 (M= 9.87; SD=1.6) years were recruited from two different day programs for children with autism in Santa Cruz County. Prior to the start of our investigation, we requested parent permission from all of the children enrolled in the two school programs. Out of 12 children, permissions were given for these 7 children. Six of the children were native speakers of American English; one child was a native speaker Spanish, but spoke English fluently. Four of the children attended a private school and the other three attended a special day program at a local public school in the Santa Cruz area. All 7 of the participants were familiar with the experimenter (the junior author) prior to our investigation and were unaware of the goals of the study. Six of the seven children were capable of speech. The Appendix gives a detailed description of each child.

Identification in the Expanded Factorial Design

Stimuli

All stimuli were presented by our computer-animated talking head, Baldi. Baldi's speech and emotion are generated by a parametrically controlled polygon topology (Massaro, 1998). The advantage of using the talking head derives from its ability to mimic natural speech, by incorporating coarticulation and being trained by natural speech measurements (Massaro, 1998; Ouni et al., 2003). The stimuli were the consonant-vowel (CV) syllables /bi/ and /di/ and the vowel (V) syllable /i/. The synthetic visible speech was controlled and aligned with the synthetic audible speech (Black & Taylor, 1997). The duration of the test syllables was 472 ms for /bi/, 385 ms for /i/, and 448 ms for /di/. The intensity of the syllables was 64.4 dB-A. It should be noted that /di/ and /i/ look very similar in visible speech.

Figure 2 gives a diagram of the expanded factorial design used in this experiment. The synthetic auditory and visual stimuli were presented unimodally and bimodally in an expanded factorial combination, giving a total of 15 conditions. There were 3 auditory conditions, 3 visual conditions and 3 x 3 or 9 bimodal conditions. Each of the 15 conditions was sampled randomly without replacement in a block of trials.

The identification task was presented before and after training in speech reading. In the Pre-training test, there were 2 blocks across 5 days for a total of 10 observations under each of the 15 conditions. In the Post-training test, there were 4 blocks across 5 days for a total of 20 observations under each of the 15 conditions.

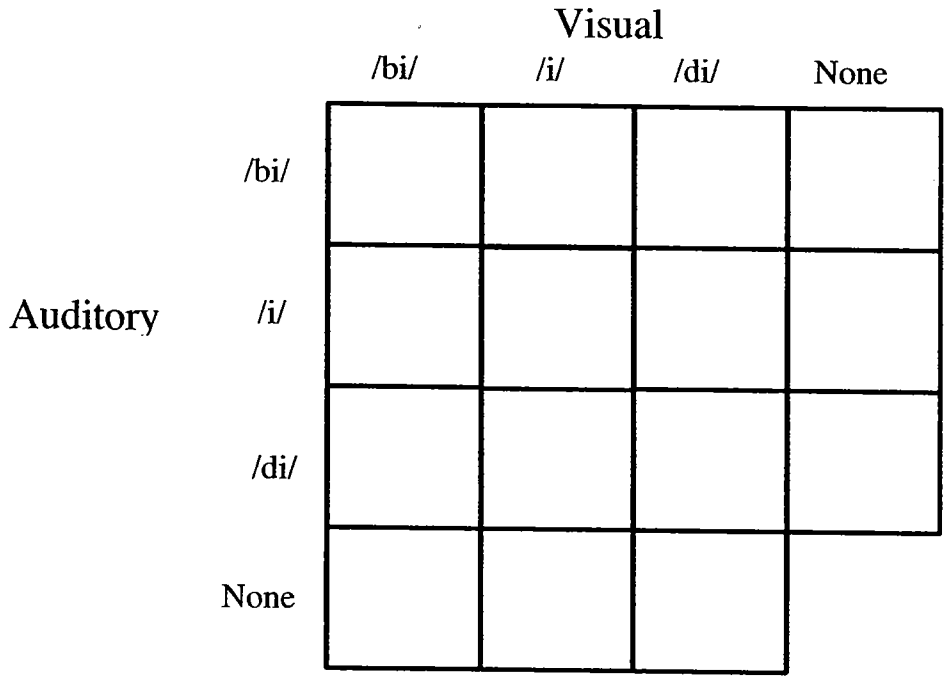

FIGURE 2. Expanded factorial design used in the Pre-training and Post-training identification task. The auditory and visual stimuli were /bi/, /i/, and /di/, presented unimodally or bimodally.

All stimuli was developed on a 600 MHz Pentium III with 128 MB memory and running a Gforce 256 AGP-V6800 DDR graphics board running Microsoft Windows NT 4 and a Graphic Series view Sonic 20" monitor. The Pre-training task was run on the machine just described, whereas Training and Post-training tasks were run on a Toshiba Satellite 5005-S504 laptop which has a 1 GHz Pentium III with 512 MB memory and Nvidia GeForce2Go graphics running Microsoft Windows 2000 professional. The auditory speech was delivered via Harman/Kardon internal speakers or Plantronics PC Headset model SR1. Each student had the option to respond with either an external mouse (Logitech M-CAA42) or a touch screen (KEYTEC Magic Touch). Each child used the same response method throughout the experiment. All sessions occurred at a computer workstation located in each school during Pre-training. Both Training and Post-training sessions were conducted individually at the student's desk.

Procedure

The children were tested individually at their school. They were instructed to listen and watch Baldi on the screen, and to identify the consonant of the syllable "that the speaker said" as either B or D. The participants made their responses by clicking

on a labeled area on the screen directly below the test window. The experiment was participant driven: the computer waited for the child to make a response before proceeding to the next trial. The investigator sat to the left of the child during the duration of the experiment, redirecting the child's attention to the task if the child became distracted and to supply uninformative motivational rewards for responses throughout the investigation.

Training in Speechreading

Stimuli

Baldi was also used to generate the consonant-vowel (CV) syllables, /di/, /vi/, /zi/, or /bi/. Figure 3 shows a view of Baldi at the onset of the articulation for each of the four syllables.

Procedure

Training in speechreading began with a bimodal presentation of each syllable. The intensity of the auditory speech was programmed by a Text-To-Speech (TTS) graphical editor (GUI) for SABLE, which is currently supported via mark up commands. The auditory intensity used during training was based on the student's performance during the previous training session. If the student attained a passing score in a given training session, the level of auditory input would be reduced in the next session, whereas the auditory input was increased if the student did not pass. The intensity of the auditory speech was set at one of 9 levels in which the intensity

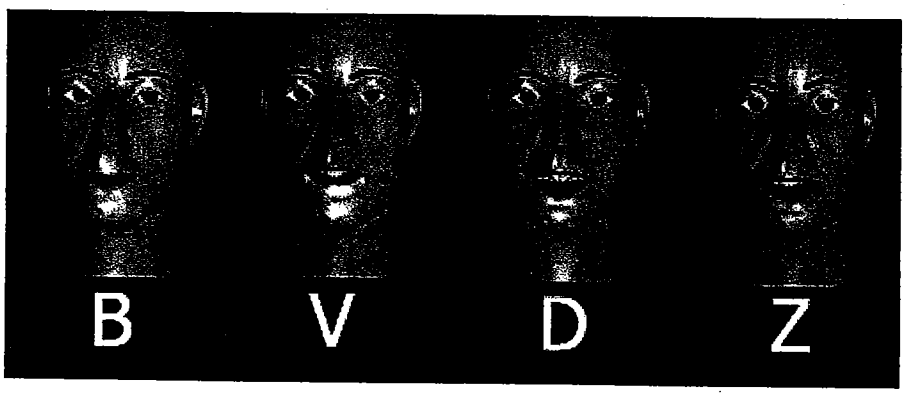

FIGURE 3. Illustration of the four syllables at the onset of articulation in the speechreading training task

varied between 0 (no auditory information) and 1 (the auditory speech at 64.4 dB-A). Table 1 gives the auditory levels used at each stage of training.

The students were instructed to watch Baldi and indicate the syllable that was spoken. A 200 ms beep sounded prior to the presentation of the test stimulus to indicate the start of each trial. Following the test presentation, response buttons appeared in the upper left hand corner of the screen. Responses were made by activating a button labeled B, D, V, or Z presented in a 2 by 2 configuration, using the mouse or touch screen. Placement of the labels was randomized across trials. Immediately following the student's response the buttons were removed and the next trial began. Feedback was given for correct responses in the form of "stickers" and verbal praise given by Baldi.

Before each training session, a test session was presented. Each syllable was presented visually without sound in 3 blocks of trials, generating a total of 12 trials (3 observations for each of the four syllables). Following completion of the 12 test trials, an accuracy score was calculated. If the student attained 100% identification accuracy during the assessment, the student was congratulated and the program automatically exited. If the student did not reach this criterion of 100%, then the program progressed to training (see Figure 4). The criterion level during training was 80%, which had to be met for the child to advance to the reduced level of auditory input.

Students were given the option to select the color of Baldi after every 3 blocks (12 training trials). The students completed 3 sessions per week, which lasted approximately 30 minutes each. The students were given a 3-minute break between training sessions. A "choice board" would appear on the screen and the student selected from a variety activities and/or food items. Upon completion of the break, the experimenter would resume the training session. Training occurred for approximately 15 weeks or until the student was able to identify all stimuli with 100% accuracy on the assessment test (without sound) for 2 consecutive sessions.

Table 1. The auditory training levels across the 9 stages of training.

Training stage	Auditory level	dB-A
1	0	0
2	5%	39.9
3	7%	41.9
4	10%	42.9
5	20%	48.9
6	30%	50.9
7	40%	52.9
8	50%	54.4
9	100%	64.4

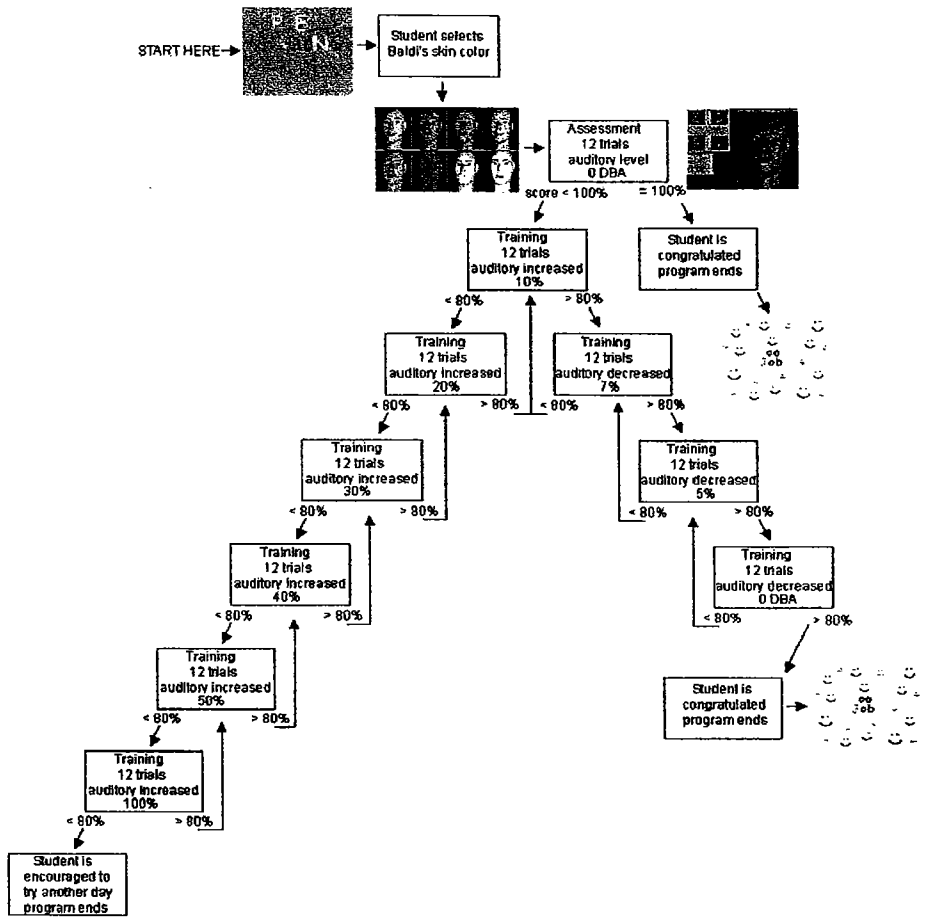

FIGURE 4. Illustration of the procedure used in the training of speechreading.

Results and Discussion

Expanded Factorial Design: Pre-training Test

An identification judgment for each stimulus was recorded. The mean observed proportion of identifications was computed for each participant for the unimodal and bimodal conditions by pooling across all 10 replications of each condition.

The proportion of /di/ responses for each of the trial types was computed for each participant. The top panel of Figure 5 gives the observed (points) proportion of /di/ judgments as a function of the auditory and visual stimuli in the unimodal and bimodal conditions. The children were clearly influenced by both the auditory and visual speech in both the unimodal and bimodal conditions. Six of the points are circled in the top panel of Figure 5. The top three points correspond to the conditions

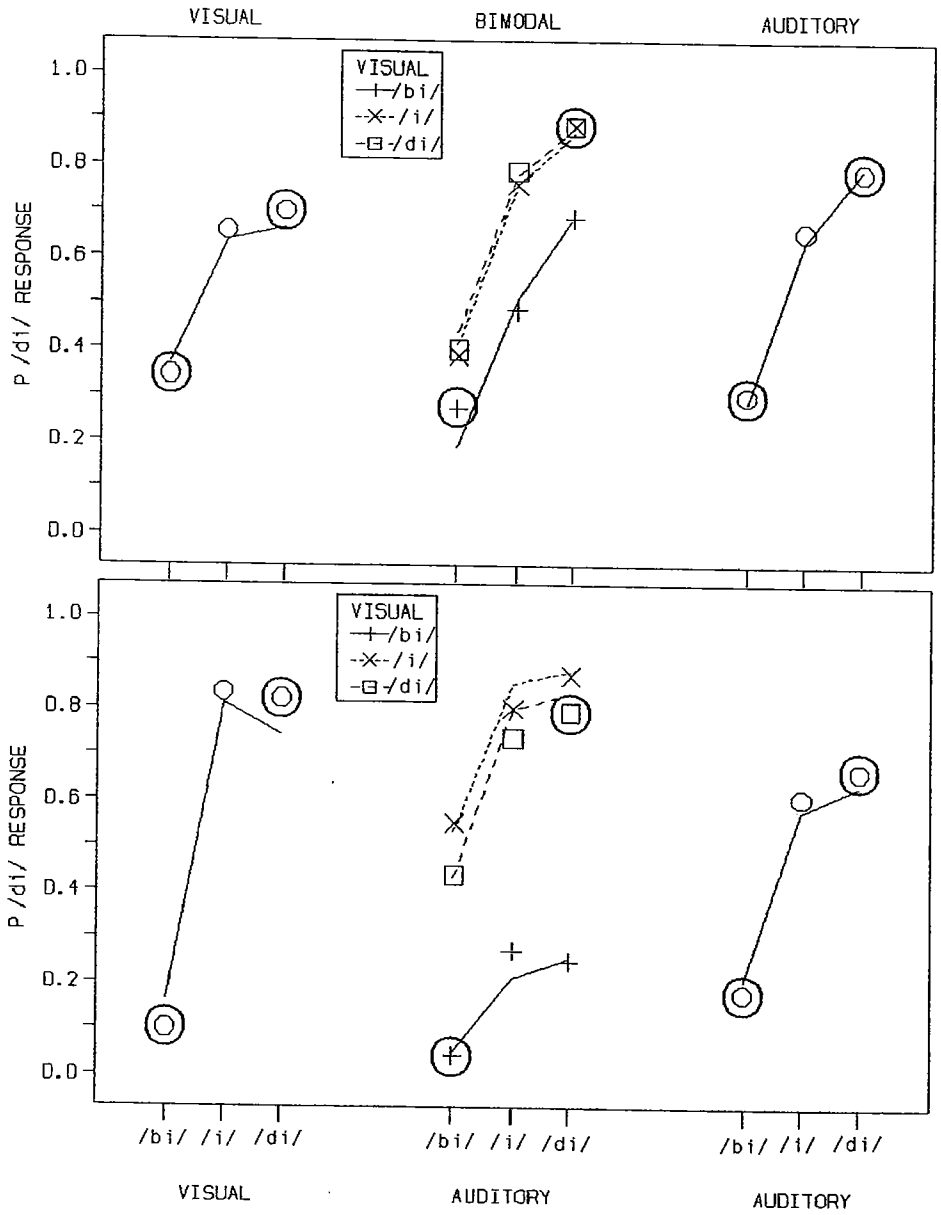

FIGURE 5. Predicted (lines) and observed (points) proportion of /di/ judgments as a function of the auditory and visual stimuli in the unimodal and bimodal conditions. The Pre-training and Post-training identification conditions are given in the top and bottom panels, respectively.

auditory /di/, visual /di/, and bimodal /di/. The outcome that the proportion of /di/ responses was higher in the consistent bimodal condition than in the two corresponding unimodal conditions is strong evidence that the children were integrating the auditory and visual speech (see Massaro, 1998). If only a single source of information

were being used on bimodal trials, the proportion of judgments could not be more extreme than either of the unimodal proportions. The more extreme judgments could only result from some combination (integration) of the two modalities. The same interpretation can be given for the bottom three points for the syllable /bi/ (the proportion of /bi/ judgments is simply one minus the proportion of /di/ judgments in this two alternative task).

These conclusions are supported by separate analyses of variance carried out on the visual, auditory, and bimodal conditions. Under the unimodal conditions, there was a significant effect for the auditory factor $F(2, 12) = 12.277$, $p < 0.01$ and the visual factor $F(2, 12) = 33.879$, $p < .01$ for the auditory and visual conditions, respectively. In the bimodal condition, there were significant main effects for both the auditory factor, $F(2, 12) = 14.55$, $p < .01$, and the visual factor, $F(2, 12) = 6.96$, $p < .01$. However, the interaction between the two variables in the auditory-visual condition did not reach statistical significance, $F(4, 24) = 0.941$, $p = 0.54$.

Training in Speechreading

One possible explanation for the small visual effect is a difficulty in speechreading. This hypothesis coincides with previous findings that for children in general, the auditory input provides more information than the visual input (see Massaro, 1984; Massaro et al., 1986). Recall that Massaro (1984) found that children were not as proficient as adults in their abilities to accurately identify the visual information under the unimodal visual conditions and the children were also less influenced by the visual information in the bimodal conditions. Thus, the ability to process visible speech (and auditory speech) must be accounted for in order to address the question of whether individuals with autism integrate information from these two modalities.

Previous findings show an improvement in speechreading through training. Our question was whether children with autism could be trained to speechread more accurately. We developed and implemented a computer-based speech reading lesson focused on the visible aspects of speech using the consonant-vowel syllables /bi/, /di/, /vi/, and /zi/. These CV syllables were selected because they are reasonably distinctive from one another and because /bi/ and /di/ corresponded to the syllables used in our experiment addressing the integration question.

Assessment data were captured daily for each student and continued until the student was able to maintain 100% identification accuracy across 2 consecutive days of assessments or 15 weeks of training. Other differences in the number of training sessions across students are the result of absences and availability. Given these differences in the number of training sessions across students, we pooled the data across sessions to give an equal number of training blocks across the 7 children. Student 6 reached the passing criterion (100% identification accuracy on the assessment for two consecutive sessions) in 7 training sessions, and data for the remainder of the students was pooled into 7 blocks.

Figure 6 shows the average identification accuracy across the 7 blocks of trials. As can be seen in the figure, identification accuracy improved systematically across blocks. An analysis of variance with the proportion of correct identification as the dependent variable and the block as the independent variable revealed that this increase was significant, $F(6, 36) = 17.079$, $p < .01$. Overall the students made substantial gains from block 1 ($M = .37$, $SD = .09$) to block 7 ($M = .77$, $SD = .14$), $F(1, 6) = 30.624$, $p < .01$.

A second ANOVA, comparing the proportion of correct identifications in block 1 ($M = .37$) to block 2, revealed an 18% increase in accuracy and that this increase was significant, $F(1, 6) = 12.862$, $p < 0.01$. This result indicates that the children made substantial gains in speech reading performance after just one block of training. To assess gains after the initial block of training, we conducted an additional ANOVA in which block 1 was eliminated from the analysis, revealing that the improvement across blocks remained significant, $F(5, 30) = 13.214$, $p < 0.01$. These results indicate that accuracy continued to increase as a function of training. Table 2 gives the individual performance for each student. As can be seen in the table, each student showed a substantial improvement in speechreading across the training sessions.

We then assessed accuracy for each syllable and its change as a function of training. As shown in Table 3, accuracy increased for all syllables and the identification of /bi/ and /vi/ was almost perfect by the end of training. The syllables /di/ and /zi/ showed less improvement. These differences are reasonable because the syllable /bi/

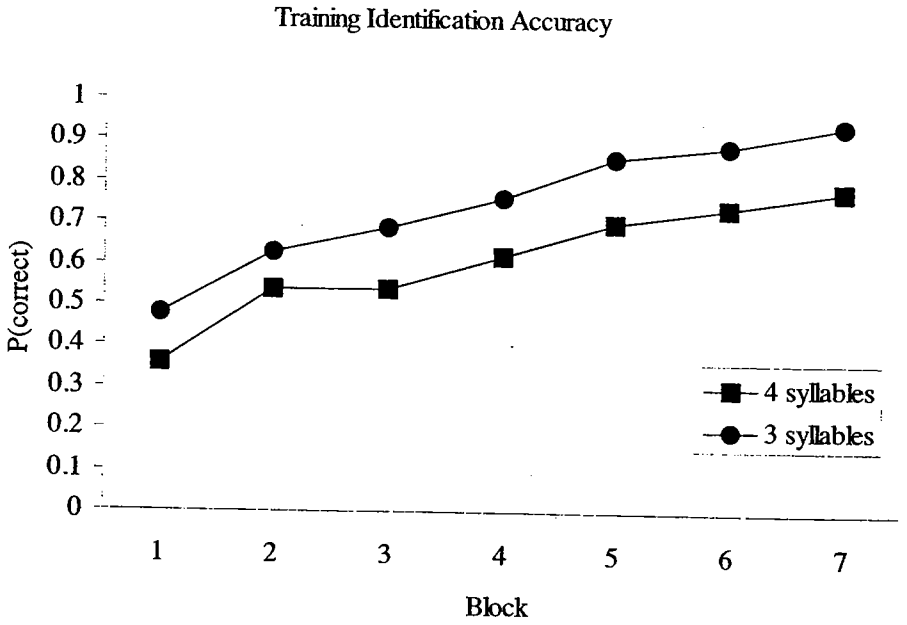

FIGURE 6. Accuracy of identification in the training experiment across blocks. The two curves correspond to accuracy computed for 3 and 4 alternatives, respectively.

Table 2. Average proportion correct across the syllables /bi/, /di/, /vi/, and /zi/ for each of the seven blocks of training.

	Student						
Block	1	2	3	4	5	6	7
1	0.38	0.42	0.38	0.38	0.33	0.17	0.43
2	0.56	0.60	0.38	0.58	0.58	0.58	0.48
3	0.56	0.63	0.31	0.54	0.50	0.67	0.63
4	0.73	0.80	0.44	0.63	0.42	0.75	0.55
5	0.75	0.77	0.42	0.75	0.75	0.83	0.68
6	0.85	0.80	0.63	0.63	0.58	1.00	0.67
7	0.76	0.90	0.58	0.73	0.71	1.00	0.74

Table 3. Average proportion correct for each of the syllables across the seven blocks of training.

	Syllable			
Block	/bi/	/di/	/vi/	/zi/
1	0.49	0.14	0.35	0.42
2	0.55	0.38	0.72	0.55
3	0.79	0.27	0.71	0.37
4	0.79	0.35	0.83	0.41
5	0.88	0.40	0.87	0.52
6	0.93	0.53	0.88	0.60
7	0.84	0.68	0.88	0.49

can be distinguished by the closing of the lips and /vi/ by the bottom lip tuck whereas the syllables /di/ and /zi/ are very similar in visible speech except for duration (see Figure 3). Given the similarity between /di/ and /zi/, we also scored accuracy when /di/ and /zi/ were treated as one category. As shown in Figure 6, this pooling increased the overall level of performance, $F(1, 6) = 37.890$, $p < 0.01$, producing almost perfect performance by Block 7.

Expanded Factorial Design: Post-training Test

As in the Pre-training task, an identification judgment for each stimulus was recorded and the mean observed proportion of /di/ identifications was computed for each participant for the unimodal and bimodal conditions by pooling across all 20 replications of each condition. The bottom panel of Figure 5 gives the observed (points) proportion of /di/ judgments as a function of the auditory and visual stimuli in the unimodal and bimodal conditions. As in the Pre-training task, the children were influenced by both the auditory and visual speech in both the unimodal and bimodal conditions. One can also observe that the visual influence was much larger

in the Post-training than in the Pre-training. The six points circled in the bottom panel of Figure 5 illustrate integration of the auditory and visual speech, following the same logic we gave for the Pre-training task.

As in the Pre-training task, separate analyses of variance were carried out on the auditory, visual, and bimodal conditions. Under the unimodal conditions, our results revealed a significant effect for the auditory factor (2, 12) = 19.410, $p < 0.01$ and the visual factor F (2, 12) = 188.647, $p < 0.01$ for the auditory and visual conditions, respectively. Auditory-visual performance reveled significant main effects for both the auditory factor and the visual factor, F(2, 12) = 20.55, $p < .001$, F(2, 12) = 20.55, $p < .01$, respectively. However, the interaction between the two variables in the bimodal condition did not reach statistical significance, F(4, 24) = 1.091, p = 0.38.

Pre-training versus Post-training Performance in the Expanded Factorial Design

A combined analysis across the Pre-training and Post-training conditions was carried out to determine if there were any differences attributable to training. For the unimodal auditory condition, there was a main effect for the auditory factor, F(2, 12) = 36.690, $p < 0.01$, but this did not interact with training, F (2, 12) = 0.165, p = 0.84. For the unimodal visual condition, there was a main effect for the visual factor, F (2, 12) = 324.277, $p < 0.01$, and a significant interaction with training, F (2,12) = 17.678, $p < 0.01$. As can be seen in Figure 5, the proportion of correct visual identifications for the syllable /bi/ and /di/ increased respectively from .66 and .70 in Pre-training to .90 and .82 in Post-Training.

For the bimodal trials, both the auditory and visual information had significant effects on performance, F (2, 12) = 29.415, $p < 0.01$, for the auditory and, F (2, 12) = 98.229, $p < 0.01$, for the visual. There was an interaction between experiment and the visual factor, F (2, 12) = 6.280, $p < 0.01$, but not for the auditory factor, F (2, 12) = 1.505, p = 0.34. The analysis also revealed that there was no interaction between the auditory and visual factors, F(4, 24) = 1.105, p = 0.377, or for the three-way interaction of auditory factor, visual factor, and experiment, F (4, 24) = 0.843, p = 0.513.

Model Tests of Integration: The FLMP

We now derive the predictions of integration and nonintegration models in order to test whether the autistic children integrated the auditory and visual speech. Consider a simplified situation in which perceivers are given either or both auditory and visual speech information, and asked to decide whether the speaker said /bi/ or /di/. Applying the FLMP to this situation, information from each modality is assumed provide independent and continuous support for these two response alternatives. It is assumed that the perceivers have prototypes corresponding to /bi/ and /di/, and evaluate the incoming signals in terms of these prototypes. For simplicity,

we specify the prototypical features representing auditory /bi/ and /di/ in terms of the onset of the second and third formants (F_2-F_3), whereas the prototypical features representing visual /bi/ and /di/ are given in terms of the amount of lip closure at the onset of the syllable. Following this description, /bi/ and /di/ are represented in memory as

/di/: falling F_2-F_3 and lips apart

/bi/: rising F_2-F_3 and lips closed

At the evaluation stage of processing, the input from each modality is evaluated independently to determine to what extent it matches the prototype descriptions. Independence means that the value assigned to one modality is independent of the value assigned to the other. The degree of match is represented in terms of truth values in fuzzy logic, which can vary continuously between 0 (false) and 1 (true). For example, an apple, a date, and an olive would be good, ambiguous, and relatively poor members of the category fruit.

To illustrate the predictions with just two response alternatives, rising F_2-F_3 can be represented as (1-falling F_2-F_3) and lips closed as (1-lips apart). Assume that the auditory input matches falling F_2-F_3 to degree .8 and the visual input matches lips apart to degree .4. Given just the auditory input, only the degree of match to the auditory feature would be relevant.

/di/: falling F_2-F_3 = .8

/bi/: (1-slightly falling F_2-F_3) = .2

Given the relative goodness for response, the probability of a /di/ response would be .8/(.8+.2) = .8

Analogously, given just the visual input, the probability of a /di/ response would be .4. Given both of these auditory and visual inputs, then we have

/di/: falling F_2-F_3 and lips apart = .8 and .4

/bi/: (1-slightly falling F_2-F_3) + lips closed = .2 and .6

The integration stage of processing involves multiplying the truth values determined at the evaluation stage for each prototype. In this example, the total support for the two alternatives would be

/di/: .8 * .4 = .32

/bi/: .2 * .6 = .12

The decision stage, leading to perceptual identification and interpretation, is based on the relative degree of support between these two alternatives. In this case, P(/di/) is equal to the support for /di/ divided by the sum of the support for /di/ and /bi/. Thus, the probability of a /di/ response, P(/di/) is equal to

P(/di/) = .32/(.32 + .12) = .32/.44 = .73

As can be seen in our example, both sources contribute to perceptual identification, but the degree of influence depends on the relative degree of ambiguity of each source. The auditory support for /di/ is less ambiguous than the visual support for /bi/ and, therefore, the auditory source has a larger influence.

Consider another example in which the two sources of information are relatively consistent: assume that the visual source now supports /di/ to degree .7, while the auditory support remains at .8.

/di/: .8 * .7 = .56
/bi/: .2 * .3 = .06
Using the relative goodness rule, the predicted probability of a /di/ response is
P(/di/) = .56/(.56 + .06) = .56/.62 = .90
We see that the predicted probability of a /di/ response is larger in the bimodal condition than in either unimodal condition.

More generally, the FLMP is formalized in terms of the following equations. In a two-alternative task with /bi/ and /di/ alternatives, the degree of auditory support for /di/ can be represented by a_i, and the support for /bi/ by $(1 - a_i)$. Similarly, the degree of visual support for /di/ can be represented by v_j, and the support for /bi/ by $(1 - v_j)$. The probability of a response to the unimodal stimulus is simply equal to the feature value. For unimodal auditory trials, the predicted probability of a response, P(/di/) is equal to

$$P(/di/) = \frac{a_i}{a_i + (1 - a_i)} \tag{1}$$

For unimodal visual trials, the predicted probability of a response, P(/di/) is equal to

$$P(/di/) = \frac{v_j}{v_j + (1 - v_j)} \tag{2}$$

For bimodal trials, the predicted probability of a response, P(/di/) is equal to

$$P(/di/) = \frac{a_i v_j}{a_i v_j + (1 - a_i)(1 - v_j)} \tag{3}$$

These equations will be implemented in the test of the FLMP against the current results.

Single Channel Model (SCM)

Given that previous research has been interpreted in terms of the lack of integration in children with autism (see Discussion Section, Previous Research), it is worthwhile to consider how this lack of integration would play out in bimodal speech perception. According to non-integration models, the categorization of a speech event is the result of a single influence (Massaro, 1998). Given a perceptual event in which multiple sources of information are available, pattern recognition is determined by only one of these sources. If, indeed, children with autism do not integrate auditory and visual speech, then a non-integration model should give a better description of their performance. We formalize a nonintegration model that provides a fair contrast with the FLMP in terms of their a priori ability to adequately predict results from this type of experiment (see Massaro et al., 2001).

The Single Channel Model (SCM) assumes that only one source of information received from the two modalities is used on any trial. This model predicts that given

a bimodal auditory-visual speech event, the participant will use the auditory information with the probability p and the visual information with the probability (1 – p). Given our earlier example with auditory and visual input and /bi/ and /di/ alternatives, we can derive the probability of identification under the different conditions. In the application of the SCM, it is assumed that the perceiver always uses the appropriate modality when only a single modality is presented. Assume that the probability of a /di/ identification, P(/di/), given a specific auditory event is used is .8. The predicted P(/di/) is also .8 on unimodal auditory trials because it is assumed that the perceiver always uses the appropriate modality when only a single modality is presented. Similarly, assume that that P(/di/), given a specific visual event is used is .4. The predicted P(/di/) is also .4 on unimodal visual trials because it is assumed that the perceiver always uses the appropriate modality when only a single modality is presented. On bimodal trials, the response is determined by the probability of using one modality rather than the other. The value of p varies between 0 and 1 and corresponds to the probability of using the auditory modality. The probability of using the visual modality is simply 1 – p. If we assume p = .7 in our example, the predicted P(/di/) would be equal to the probability of using the auditory modality times the probability of a /di/ identification of the auditory modality plus the probability of using the visual modality times the probability of a /di/ identification of the visual modality.

$$P(/di/) = .7 * .8 + .3 * .4 = .68$$

More generally,

$$P(/di/) = pa_i + (1 - p)v_j \quad (4)$$

where p is the probability of using the auditory modality, a_i is the probability of a /di/ identification of the auditory modality, (1 – p) is the probability of using the visual modality, and v_j is the probability of a /di/ identification of the visual modality. On any trial, pattern recognition of a multimodal event is a consequence of only one of the modalities (Massaro, 1987, 1998).

Model Tests

To test auditory visual integration in speech perception, we tested the FLMP and the SCM against each of the individual participant's results. As described in Massaro (1998, Chapter 2), the FLMP requires 6 free parameters: three parameters for each auditory and visual stimulus to fit the 15 data points of the 3 x 3 expanded factorial design. These parameters symbolize of the degree to which these modalities match a prototypical /di/. The SCM requires 6 analogous parameters and a seventh corresponding to the probability of using the auditory modality. The two models were fit to the individual results and to the mean results across the seven participants. Separate fits were carried out for the Pre-training and Post-training tasks.

The program STEPIT (Chandler, 1962) determined the quantitative predictions of the models. Each model is represented as a set of unknown parameters and prediction equations. STEPIT adjusts the parameter values of the model iteratively, min-

imizing the root mean squared deviation (RMSD) between the predicted and observed points. The RMSD provides an index of each model's goodness-of-fit (Massaro, 1998).

In the Pre-training task, the RMSDs of the FLMP ranged from .07 to .12, with an average RMSD of .09. The fit of the mean participant gave an RSMD of .03. The RSMD of the SCM ranged from .08 to .16, an average of .11, and the mean participant RMSD of .05.

In the Post-training fit, the RMSD of the FLMP ranged from .04 to .08, with an average of .06. The fit of the mean participant gave an RMSD of .04. The RMSD of the SCM ranged from .04 to .1, gave an average of .08, and the mean participant RMSD of .04.

The lines in Figure 5 give the average predictions of the FLMP. As can be seen in the figure, the integration model is able to describe the results fairly accurately. The three circled points in each of the panels show that the results and the model's predictions both show a benefit of having consistent auditory and visual speech relative to either source presented alone.

Figure 7 gives the individual RMSDs for the FLMP plotted as a function of the individual RMSDs for the SCM for both the Pre-training and Post-training tasks. The points that fall below the diagonal line show a better fit of the FLMP over the SCM. An analysis of variance was carried out on these RMSD values, with Pre-training and Post-training and Model as independent variables. The FLMP gave a significantly better fit of individual performance than did the SCM, $F(1, 6) = 7.368$, $p < .05$. Given that the FLMP and SCM represent integration and nonintegration models, respectively, we can tentatively conclude that the children were integrating auditory and visual speech.

The RMSDs were significantly smaller in the Post-training than in the Pre-training task, $F(1,6) = 8.135$, $p < .05$. The reason for this difference is primarily due to having twice than number of observations in Post-training than in Pre-training (Massaro, 1998, Chapter 10). Sampling variability decreases with increases in the number of observations. Although the fit of the models were better for Post-training than Pre-training, the advantage of the FLMP did not interact with training, $F(1,6) = 0.362$, $p = .57$.

General Discussion

Our experiments provide some evidence that children with autism are influenced to some extent by speech information in the face, can be taught to improve their sensitivity to visible speech, and do integrate cross-modally in speech perception. We tested an integration model (the FLMP) against a non-integration model (the SCM) against the identification results from an expanded-factorial design in which the auditory and visual speech were presented alone or together. Although the influence of visible speech was relatively small in the first Pre-training test, we succeeded in training the children to speechread to allow a stronger test of integration when there was

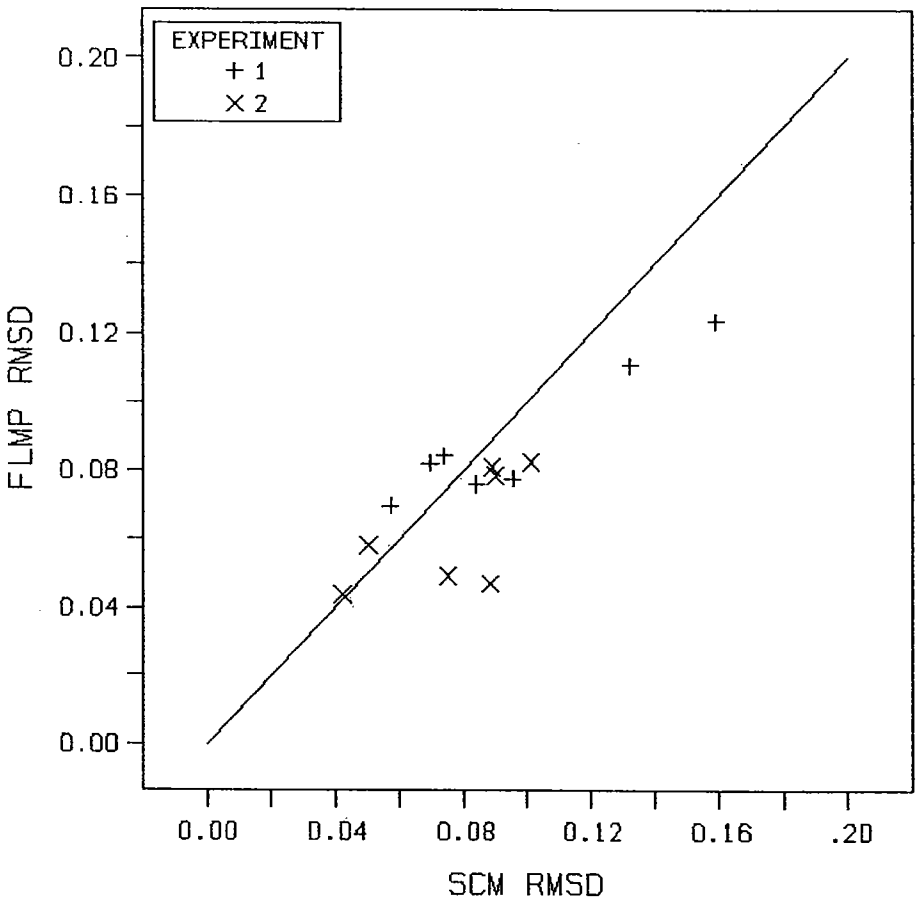

FIGURE 7. RMSD values for the fit of the FLMP as a function of the RMSD values for the SCM for each of the seven children in the Pre-training (Experiment 1) and Post-training (Experiment 2) tasks. The diagonal line gives equivalent RMSD values for the two models.

a larger influence of visible speech. The FLMP gave a significantly better fit than the SCM across these two replications of the expanded-factorial design.

A possible limitation of our investigation is the absence of a control group. Even though we did not test normally-developing children in the current study, however, there is an existing literature that makes such comparisons possible. Our question in the present study was whether children with autism integrate auditory and visual information in a speech perception task. Having now answered this question in the affirmative, we can ask how this outcome compares with that of normally-developing children. Our previous research (Massaro, 1987, Chapter 8) found that the FLMP gave a significantly better description of performance than the SCM across a wide range of development (3.5 to 9.5 years). Thus, the current results taken in conjunc-