Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Research in Developmental Disabilities 25 (2004) 559–575
Research in Developmental Disabilities

# Visual–auditory integration during speech imitation in autism

Justin H.G. Williams[a,*], Dominic W. Massaro[b], Natalie J. Peel[a], Alexis Bosseler[b], Thomas Suddendorf[c]

[a]*Department of Child Health, University of Aberdeen, Royal Aberdeen Children's Hospital, Aberdeen AB25 2ZD, UK*
[b]*Department of Psychology, University of California—Santa Cruz, Santa Cruz, CA 95064, USA*
[c]*School of Psychology, University of Queensland, Brisbane, Qld 4072, Australia*

Received 12 August 2003; received in revised form 11 December 2003; accepted 13 January 2004

**Abstract**

Children with autistic spectrum disorder (ASD) may have poor audio–visual integration, possibly reflecting dysfunctional 'mirror neuron' systems which have been hypothesised to be at the core of the condition. In the present study, a computer program, utilizing speech synthesizer software and a 'virtual' head (Baldi), delivered speech stimuli for identification in auditory, visual or bimodal conditions. Children with ASD were poorer than controls at recognizing stimuli in the unimodal conditions, but once performance on this measure was controlled for, no group difference was found in the bimodal condition. A group of participants with ASD were also trained to develop their speech-reading ability. Training improved visual accuracy and this also improved the children's ability to utilize visual information in their processing of speech. Overall results were compared to predictions from mathematical models based on integration and non-integration, and were most consistent with the integration model. We conclude that, whilst they are less accurate in recognizing stimuli in the unimodal condition, children with ASD show normal integration of visual and auditory speech stimuli. Given that training in recognition of visual speech was effective, children with ASD may benefit from multi-modal approaches in imitative therapy and language training.


*Keywords:* Autistic disorder; Speech perception; Audio–visual integration; Imitation; Mirror neurons

* Corresponding author. Tel.: +44 1224 552471; fax: +44 1224 663658.
*E-mail address:* Justin.Williams@abdn.ac.uk (J.H.G. Williams).

0891-4222/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ridd.2004.01.008

## 1. Introduction

### 1.1. Imitation, autism and mirror neurons

Children with autistic spectrum disorder (ASD) are well-recognized to have difficulties with elicited imitation and facial imitation (Rogers, 1999; Williams, Whiten, & Singh, 2004; Williams, Whiten, Suddendorf, & Perrett, 2001). Yet, they are more likely to show excessive imitation in the form of echolalia or stereotyped speech that contributes to the diagnosis within standard research instruments (Lord, Rutter & Le Couteur, 1994). Speech may be copied exactly, with repetition of tones and intonation (Williams et al., 2004). However, imitation often seems to be unimodal in that autistic children either mimic the sound or sight of the action but not an audio–visually integrated form. Similarly, their social communication is characterized by speech without associated gesture or other non-verbal communication, and their pretend play may involve copying an action but not the sounds that go with it (Lord et al., 2000). Such observations raise the question of whether the imitation deficit associated with autism, involves difficulty integrating visual and auditory information.

Williams et al. (2001) suggested that the imitative problems in autism may be related to abnormal function of 'mirror neurons' (MNs). These neurons code for the same action, whether it is perceived or enacted (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). Gallese and Goldman (1998) considered that the 'theory of mind' deficit associated with autism might be due to a 'simulation' deficit. This suggests that people with autism have difficulty understanding the thoughts of others, because they find it hard to imagine themselves in another person's position by relating others' behavior to their own neural codings for similar behavioral memories. Williams et al. (2001) suggested that such a simulation process could be dependent upon a neurocognitive mechanism allied to that necessary for imitation, and that a developmental delay in such a neural mechanism, involving MN dysfunction, could be the common factor that is core to autistic disorders.

Recently, Kohler et al. (2002) showed that mirror neurons in the F5 area of the monkey brain responded to the sound as well as the sight of actions. Iacoboni et al. (1999) have suggested that cells in the superior temporal sulcus have MN properties, and both of these areas have also been associated with cross-modal binding and audiovisual integration in speech perception (Calvert, 2001; Calvert et al., 1997; Calvert, Campbell, & Brammer, 2000; Mistlin & Perrett, 1990). The superior temporal sulcus is also a region that has been implicated in the psychopathology of autism because of its role in detecting the attentional direction of other individuals and understanding mental states communicated by eye movements (Baron-Cohen et al., 1999, 2000; Emery & Perrett, 2000). Together, these arguments lead to the hypothesis that autism involves dysfunctional neural systems that are concerned with the supra-modal representation of actions, including those concerned with audio–visual integration of stimuli. In this study, we tested this hypothesis within the context of speech perception.

### 1.2. Audio–visual speech perception

We define speech perception as the process of imposing a meaningful perceptual experience on an otherwise meaningless speech input (Massaro, 1984, 1987, 1998).

The stimulus input for speech is a continuous stream of sound (and facial and gestural movements in face-to-face communication) produced by the speech production process. Somehow, this continuous input is transformed into a more or less meaningful sequence of events, and we perceive mostly a discrete auditory message composed of words, phrases, and sentences.

There is now a large body of evidence indicating that multiple sources of information are available to support the perception, identification, and interpretation of spoken language (Massaro, 1998). Researchers have repeatedly shown that pairing noisy 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 (Calvert, Brammer, & Iversen, 1998; Massaro, 1984; Summerfield & McGrath, 1984).

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 with 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/. 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/.

Although more recent research has shown that the original results and interpretation were not exactly correct, there is strong evidence that speech perception is a bimodal process, influenced by both the sight and sound of the speaker (Massaro, 1998). We hypothesised that children with autism would not be susceptible to the McGurk effect, for two possible reasons. Firstly, they may have a problem with sensory integration of auditory and visual speech if such a process is dependent upon 'mirror neurons' which are purportedly dysfunctional. Secondly, children with autism are known to have some difficulty reading facial expression (Davies, Bishop, Manstead, & Tantam, 1994; Schultz et al., 2000, 2003). It therefore seems quite possible that they may also have difficulty lip reading (called speech-reading because it involves more than just the lips) for similar reasons, which could place a limitation in the amount of visual information they decode. These two effects may be distinguished by examining speech-reading ability on its own as well as in the context of the bimodal speech perception task described in the following.

### *1.3. The Fuzzy Logical Model and the Single Channel Model of speech perception*

The question of whether children integrate sound and facial information when perceiving speech is best framed as two opposing hypotheses. The first states that the individual will use information from sight and sound to perceive speech, and that the amount of influence each source has on what is heard, is determined by how well it matches the prototypical representation of that speech sound. Massaro (1998) has mathematically modelled this using fuzzy logic to create the Fuzzy Logical Model of Perception (FLMP). The second hypothesis, modelled as the Single Channel Model of perception (SCM), states that information from only one source, either sound or sight, will be used on any one trial.

In this latter model, the probability of a correct response will be determined by the likelihood of using either sound or sight for a given speech input and the likelihood of making an accurate judgment given that input. In normal individuals, the FLMP is a better predictor than the SCM of responses to consistent, ambiguous, and conflicting speech stimuli, indicating that audio–visual integration occurs (Massaro, 1998). The details of this approach with respect to our results are discussed further in the following.

We recently sought to ascertain whether children with autism integrate information. Firstly, a group of children with ASD were compared to a control group in their performance on a speech perception task in which they had to identify spoken syllables. Secondly, a group of children with ASD were trained in speech-reading to see if this improved their overall performance. If it did, this would be further evidence that they were utilizing visual information in making their responses.

## 2. Method

### 2.1. Subjects

Subjects were three groups of children. The first group was 15 children with autistic spectrum disorder who had at least phrase speech. Most were verbally fluent. They attended schools in Scotland and have all been assessed on the Autism Diagnostic Interview—Revised (Lord et al., 1994) and Autism Diagnostic Observation Schedule—Generic (Lord et al., 2000) to ensure that they meet full research diagnostic criteria for an autistic spectrum disorder. Here, they are referred to as the autistic spectrum disorder group (ASD group). The second group was 15 controls of similar age recruited from local schools in Scotland. All of the subjects were aged 5–13 years and all were male (details are shown in Tables 1 and 2). Children were assessed on the British Picture Vocabulary Scale (BPVS, Dunn et al., 1997), which was administered according to the manual. Parents of controls and subjects were asked to complete the self-report social and behavioral communication questionnaire (SBCQ) (Berument, Rutter, Lord, Pickles, & Bailey, 1999) which is designed to rate the number of symptoms of autism as perceived by parents. The third group participated in a speech-reading training program. This group consisted of five children with ASD, one female and four males, ranging in age from 8 to 13. These children were recruited from a school in Santa Cruz, California, whose students had all been diagnosed with autism according to DSM-IV criteria. All had some speech.

### 2.2. Procedure

A computer-animated head combined with speech synthesizer software (Fig. 1; Baldi) provided the test stimuli (Massaro, 1998). Its visible speech movements are carefully based on recognized patterns of facial and tongue movements known to be associated with different forms of pronunciation. It is proving a useful tool for teaching both deaf and autistic children to speech-read and develop their language skills (Bosseler & Massaro, 2003; Massaro, Bosseler, & Light, 2003). This software was installed on a portable computer and individual testing was carried out in children's schools or homes. Children

Table 1
Characterisitics of the ASD and control participants

| ASD group | | | Control group | | |
|---|---|---|---|---|---|
| Age (months) | Estimated VIQ | SBCQ | Age (months) | Estimated VIQ | SBCQ |
| Scottish children | | | | | |
| 76 | 99 | 22 | 74 | 102 | 6 |
| 83 | 88 | –[a] | 85 | 82 | –[a] |
| 90 | 114 | 37 | 85 | 93 | 5 |
| 96 | 75 | 24 | 92 | 94 | 3 |
| 98 | 81 | 22 | 95 | 133 | 2 |
| 98 | 76 | 20 | 105 | 97 | 8 |
| 104 | 72 | 21 | 106 | 104 | 5 |
| 104 | 87 | 36 | 107 | 115 | 4 |
| 108 | 85 | 25 | 107 | 105 | 0 |
| 117 | 81 | 22 | 111 | 120 | 7 |
| 121 | 73 | 29 | 117 | 91 | 6 |
| 137 | 78 | 32 | 117 | 94 | 7 |
| 158 | 103 | 29 | 125 | 88 | 6 |
| 160 | 87 | 17 | 127 | 111 | 3 |
| 160 | –[a] | 27 | 134 | 115 | 3 |
| Santa Cruz children | | | | | |
| 138 | Mr[a] | –[a] | | | |
| 131 | –[a] | –[a] | | | |
| 157 | 57[b] | –[a] | | | |
| 108 | 38[c] | –[a] | | | |
| 96 | MMR[a] | –[a] | | | |

[a] Data not available. MR: mental retardation, MMR: moderate mental retardation.
[b] Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1989).
[c] Psychoeducational Profile—Revised (PEP-R) (Schopler, Reichler, Bashford, Lansing, & Marcus, 1990). This score represents the developmental age equivalent (in months).

sat in a quiet room at a desk or table. Stimuli were displayed on the 14″ TFT computer screen and sound was provided through the computer's inbuilt speakers.

The stimuli were the consonant–vowel (CV) syllables /ba/, /tha/, and /da/. These stimuli were presented unimodally and bimodally in an expanded factorial combination, giving a total of 15 conditions (see Table 3). Each of the 15 conditions was sampled randomly without replacement in a block of trials. The children were tested for 20 blocks of trials, for a total of 20 observations for each of the 15 experimental conditions. Testing took place in

Table 2
Average age and British Picture Vocabulary Scale (BPVS) score for the two groups of Scottish subjects

| | Mean age (months) | S.D. | BPVS (standardized score) | S.D. |
|---|---|---|---|---|
| ASD group | 105.8 | 17.1 | 86.4 | 12.9 |
| Controls | 114.0 | 12.7 | 102.9 | 13.8 |
| *F* | 0.949 | | 10.15 | |
| *P* | 0.338 | | 0.004 | |

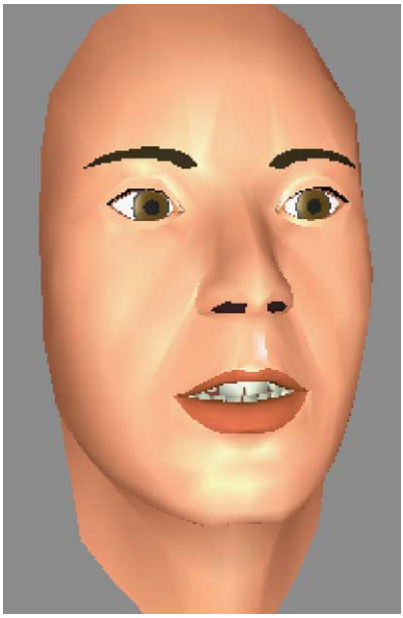

Fig. 1. 'Baldi'. A computer-animated head combined with speech synthesiser software.

four 15-min sessions. The experimenter ensured that the participant was looking at the screen for each stimulus. On being shown the stimulus, they were asked to make a vocal response as to which sound they had heard, before moving on to the next trial. Independent of the accuracy of responding, the experimenter provided encouraging comments to maintain participants' attention.

### 2.3. Speech-reading training programme

The third group of five students from Santa Cruz underwent training in speech-reading, using the same consonant–vowel (CV) syllables /ba/, /tha/, and /da/ used in the identification testing. Training in speech-reading involved the addition of variable amounts of auditory speech to the visual syllable, increasing if the student failed test sessions but

Table 3
The 15 conditions of stimuli presentation in the identification task

| | Visual stimuli | | | |
|---|---|---|---|---|
| | Ba | Da | Tha | Nil |
| Auditory stimuli | | | | |
| Ba | x | x | x | x |
| Da | x | x | x | x |
| Tha | x | x | x | x |
| Nil | x | x | x | |

reducing it if they passed them. Feedback was given for correct responses in the form of ''stickers'' and verbal praise given by Baldi.

For each test session, three blocks were presented generating a total of 12 trials (three observations for each of the four conditions). Following completion of the 12 test trials, an accuracy score was calculated. To advance to a reduced level of auditory input the participant had to achieve 80% correct. The students completed three sessions per week, which lasted approximately 30 min. To maintain attention, participants could select the color of Baldi after every three blocks and, after every 12 correct identifications, they were given a 3 min break. During the break, a ''choice board'' would appear on the screen and the student selected from a variety activities and/or food items. Training and assessment were continued for 26 sessions or until the student was able to maintain 100% identification accuracy across two consecutive sessions. As might be expected, the performance across the students varied, and none of the children achieved 100% identification within the 26 sessions. Individual data were grouped into 13 blocks of two sessions each.

## 3. Results

We first present the results of the group of 15 children with ASD compared to the 15 children in the control group. A full data set was obtained for all 30 participants. Fig. 2 gives the FLMP predicted (lines) and observed (points) probabilities /ba/, /da/, and /tha/ as a function of unimodal auditory (left panel), bimodal (middle panel) and unimodal visual (right panel conditions) for the Scottish ASD children in the top graph and for controls in the bottom graph. Separate analyses of variance were carried out on the unimodal visual, unimodal auditory, and bimodal conditions, with test stimulus and modality as the independent variables, and proportion of correct responses as the dependent variable.

### 3.1. Unimodal conditions

Table 4 gives the percentage of correct identifications and standard deviations in the unimodal auditory and visual conditions. In the auditory condition /ba/ was a more difficult stimulus to identify than the other two stimuli for both groups ($F$ (2, 56) = 48.013, $P <$ 0.001). In the visual condition /da/ was the most difficult to identify ($F$ (2, 56) = 15.34, $P <$ 0.001). The control group was significantly more often correct than the ASD group for the auditory stimuli ($F$ (1, 28) = 4.02, $P = 0.052$) and the visual stimuli ($F$ (1, 28) =6.00, $P =$ 0.02). In neither case was the interaction between syllable difficulty and group significant (visual: $F$ (2, 56) =1.025, $P = 0.367$; auditory: $F$ (2, 56) = 0.397, $P = 0.680$). Age and BPVS raw score predicted both auditory and visual accuracy in the control subjects, but not in the ASD group (Table 5).

### 3.2. Bimodal conditions

The bimodal results were analyzed in terms of accuracy in identifying the auditory stimulus. As expected, there was a significant effect of the auditory stimulus ($F$ (2, 56) =

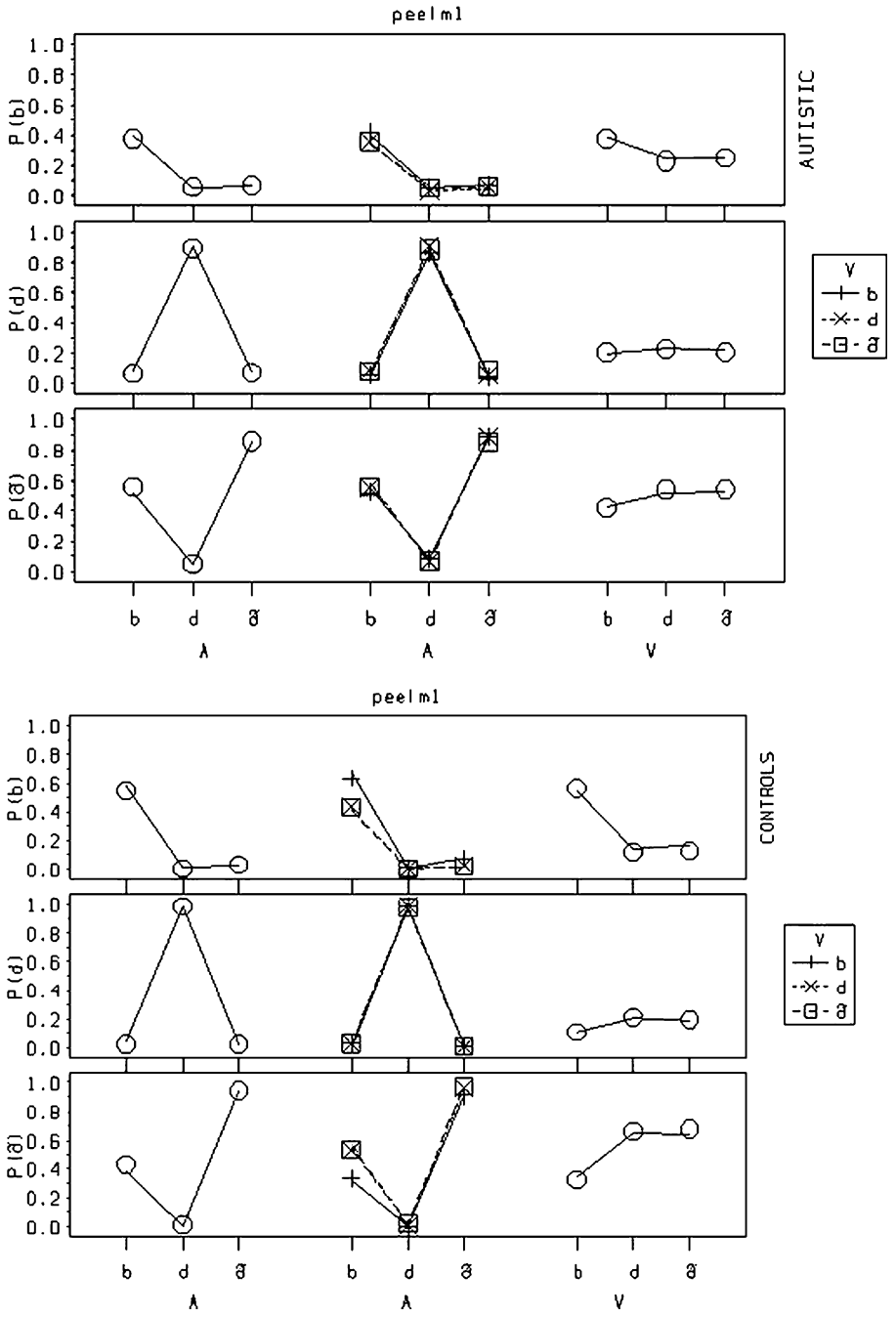


Fig. 2. Probabilities of responses /ba/, /da/, and /tha/ as a function of unimodal auditory (left panel), bimodal (middle panel) and unimodal visual (right panel conditions) for the ASD children in the top graph and for controls in the bottom graph. Predictions of the FLMP are shown as lines and observations are shown as points.

Table 4
Percentage of correct identifications and standard deviations (S.D.) in the unimodal auditory and visual conditions

| Stimulus | Auditory only | S.D. | Visual only | S.D. |
|---|---|---|---|---|
| Controls | | | | |
| /ba/ | 54.67 | 0.3681 | 56.33 | 0.3187 |
| /da/ | 98.33 | 0.0408 | 21.67 | 0.1531 |
| /tha/ | 94.33 | 0.0651 | 67.67 | 0.2492 |
| ASD | | | | |
| /ba/ | 37.67 | 0.3610 | 37.67 | 0.3052 |
| /da/ | 89.33 | 0.1545 | 22.67 | 0.1557 |
| /tha/ | 86.00 | 0.1639 | 54.33 | 0.2796 |

62.21, $P < 0.001$), the visual stimulus ($F$ (2, 56) = 5.13, $P = 0.009$), and the interaction of the auditory and visual stimulus ($F$ (4, 112) = 11.02, $P < 0.001$). Most importantly, however, group interacted with the interaction of the auditory and visual stimulus ($F$ (4, 112) = 4.62, $P = 0.002$). This interaction reflects the accuracy on auditory /ba/ being more influenced by the visual stimulus for the control group than for the ASD group (see Table 6 and Fig. 2). For controls, when the visual stimulus was consistent with the auditory /ba/, then accuracy increased from about 44% to 63%, whilst for the ASD group it went from just 36% to 42%. However, we considered that these differences were likely to be accounted for by the decreased visual accuracy among the ASD group relative to the controls, as evident in the unimodal visual condition. The ability to use visual information is likely to depend upon visual accuracy, as shown in Fig. 3, which gives the strength of visual effect as measured by increased accuracy when visual stimulus is congruent as opposed to incongruent for the /ba/ auditory stimulus. The Pearson correlation coefficient was 0.591, $P = 0.001$. Therefore, to look for a remaining group effect we calculated partial correlations between group status and visual effect controlling for visual accuracy. With the ASD group outlier (see Fig. 3) removed, this partial correlation coefficient was not significant for either SBCQ score or group status using two-tailed tests. In conclusion, these

Table 5
Age and verbal ability (as measured using BPVS) as predictors of visual and auditory accuracy: Pearson correlation coefficients (following removal of outlier)

| | Visual accuracy | | Auditory accuracy | |
|---|---|---|---|---|
| | $r$ (Pearson) | $P$ | $r$ | $P$ |
| BPVS | | | | |
| Control | **0.532** | **0.041** | **0.577** | **0.024** |
| ASD | 0.262 | 0.411 | 0.182 | 0.572 |
| Age | | | | |
| Control | **0.552** | **0.033** | **0.553** | **0.032** |
| ASD | 0.106 | 0.719 | 0.217 | 0.455 |

Table 6
Proportion correct based on identification of the auditory stimulus

| Auditory | Visual | Control | S.D. | ASD | S.D. |
|---|---|---|---|---|---|
| /ba/ | /ba/ | 0.6300 | 0.3390 | 0.4200 | 0.3683 |
| | /da/ | 0.4400 | 0.3124 | 0.3633 | 0.3456 |
| | /tha/ | 0.4333 | 0.2950 | 0.3633 | 0.1941 |
| /da/ | /ba/ | 0.9900 | 0.0207 | 0.8633 | 0.1015 |
| | /da/ | 0.9900 | 0.0280 | 0.9067 | 0.1924 |
| | /tha/ | 0.9733 | 0.0372 | 0.8833 | 0.1232 |
| /tha/ | /ba/ | 0.9100 | 0.1242 | 0.8867 | 0.1232 |
| | /da/ | 0.9667 | 0.0673 | 0.8867 | 0.2120 |
| | /tha/ | 0.9700 | 0.0455 | 0.8567 | 0.3683 |

results indicate that decreased visual accuracy results in a decreased McGurk effect. Hence there is no evidence for diminished integration in this sample of children with ASD relative to controls.

### 3.3. Speech-reading training results

The third group of five children with ASD participated in a speech-reading training program across 13 blocks of trials. An analysis of variance was carried out with blocks as the

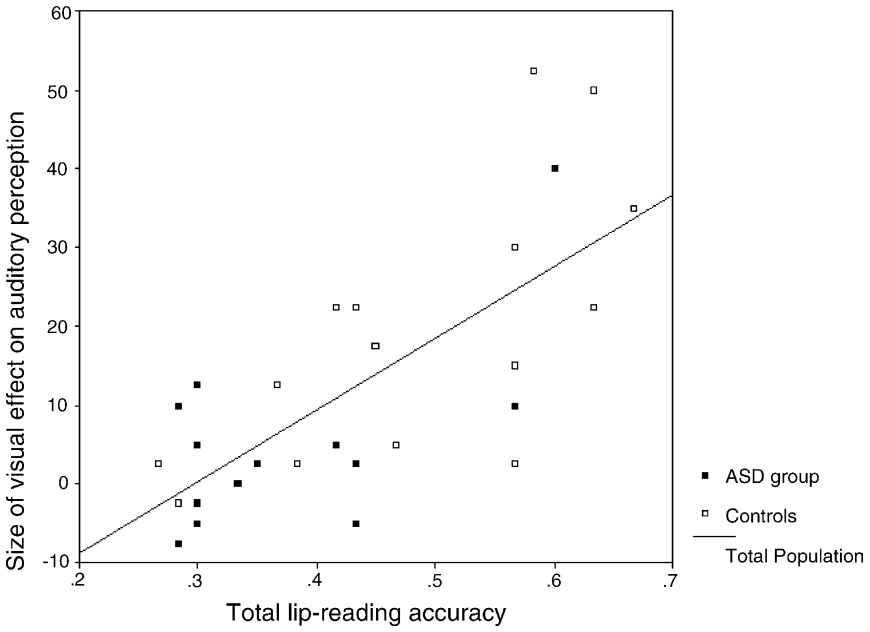


Fig. 3. The strength of visual effect as measured by increased accuracy when visual stimulus is congruent as opposed to incongruent for the /ba/ auditory stimulus. Pearson's correlation co-efficient = 0.591, $P = 0.001$.

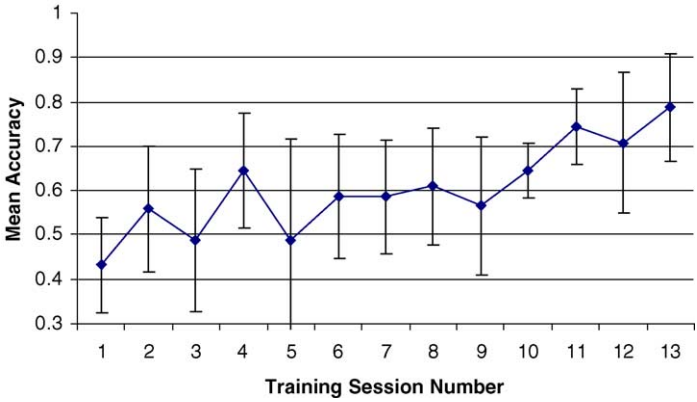


Fig. 4. Improvement across training sessions during speech-reading training sessions (mean accuracy for five individuals ±1S.D.).

independent variable and accuracy averaged across the three syllables as the dependent variable. Training was highly successful in that performance improved from 43% correct in the first block of trials to 79% correct in the 13th block, $F(12, 48) = 2.876$, $P < 0.005$ (see Fig. 4).

## 4. Pre-training identification results

The third group of five children with ASD were also tested in the identification task before and after training. Fig. 5 gives the FLMP predicted (lines) and observed (points) probabilities of responses /ba/, /da/, and /tha/ as a function of unimodal auditory (left panel), bimodal (middle panel) and unimodal visual (right panel conditions) for the five Santa Cruz ASD children (a) before speech-reading training (b) after speech-reading training. An analysis of variance was carried out on the proportion of accurate responses for each of the three stimulus alternatives. Under the unimodal conditions, auditory and visual accuracy differed according to the stimulus syllable (auditory factor: $F(4, 16) = 8.74$, $P < 0.01$ and the visual factor $F(4, 16) = 15.05$, $P < 0.01$). This was similar for the bimodal conditions (auditory factor, $F(4, 16) = 17.07$, $P < 0.01$, and the visual factor, $F(4, 16) = 6.82$, $P < 0.001$). The interaction between the two variables in the auditory–visual condition was not significant, $F(8, 32) = 1.80$, $P = 0.11$.

## 5. Post-training identification results

As can be seen in the bottom panel of Fig. 5, the results for post-training were similar to pre-training with one exception. Under the unimodal conditions, the results revealed a significant effect for the auditory factor $F(4, 16) = 19.03$, $P < 0.01$ and the visual factor $F(4, 16) = 30.41$, $P < 0.01$ for the auditory and visual conditions, respectively. Auditory–visual performance revealed significant main effects for both the auditory factor, $F(4, 16) = 21.22$, $P < 0.01$, and the visual factor, $F(4, 16) = 20.53$, $P < 0.001$. The results differed

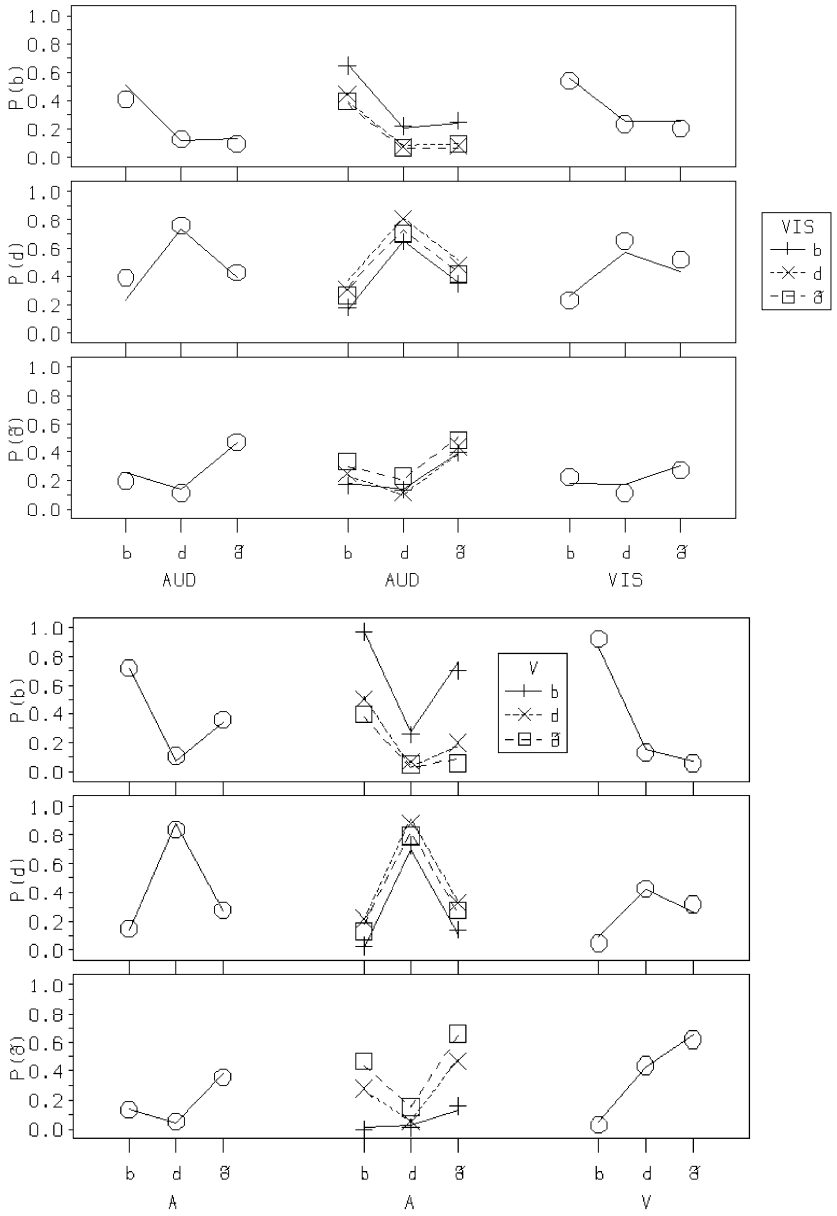


Fig. 5. Probabilities of responses /ba/, /da/, and /tha/ as a function of unimodal auditory (left panel), bimodal (middle panel) and unimodal visual (right panel conditions) for the five Santa Cruz ASD children (top graph) before speech-reading training (bottom graph) after speech-reading training. Lines show predictions of the FLMP and points show observations.

from the pre-training results in that the interaction between the two variables in the auditory–visual condition reached statistical significance, $F\ (8, 32) = 3.05$, $P = 0.011$.

### 5.1. Pre-training versus post-training performance in the expanded factorial design

A combined analysis comparing the pre-training and post-training conditions was carried out to determine if there were any differences attributable to training. The dependent measure was the average proportion correct across the three stimulus alternatives. In the unimodal auditory condition, there was a main effect for the auditory factor, $F(1, 4) = 36.690$, $P < 0.01$, but this did not interact with training, $F\ (1, 4) = 0.165$, $P = 0.84$. For the unimodal visual condition, there was a main effect for the visual factor, $F\ (1, 4) = 324.277$, $P < 0.01$, and a significant interaction with training, $F\ (1, 4) = 17.678$, $P < 0.01$. The proportion of correct visual identifications for the syllables increased from 0.53 to 0.66 with training. This shows that the training was effective in increasing the accuracy of visual speech (but not auditory speech identification), suggesting that the auditory–visual interaction that appeared in the post-training phase resulted from the larger influence of visual information.

### 5.2. Model testing

To test whether speech integration occurred, we contrasted the Fuzzy Logical Model of Perception with a Single-Channel Model (SCM), which represent integration and non-integration models, respectively. As discussed above, tests of these models have the potential to determine whether children with autism integrate vocal and facial information in speech perception. If audio–visual integration fails to occur, the SCM should provide a significantly better description than FLMP, assuming that the visual effect is large enough in the first place to have a meaningful influence.

As described in Massaro (1998, Chap. 2), the FLMP requires six free parameters: three parameters for each auditory and visual stimulus to fit the 15 data points of the $3 \times 3$ expanded factorial design. These parameters symbolize of the degree to which these modalities match the prototypical test stimulus. The SCM requires six 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 participants. Separate fits were carried out for the children in the Scottish experiment and for the five Santa Cruz children for pre-training and post-training tasks.

The program STEPIT (Chandler, 1969) determined the quantitative predictions of the models' fit to each of the participants individually. Each model is represented as a set of unknown parameters and prediction equations. STEPIT adjusts the parameter values of the model iteratively, minimizing the root mean squared deviation (RMSD). The RMSD provides each model's goodness-of-fit between the predicted and observed points (Massaro, 1998).

As can be seen in Fig. 2, there was no significant difference between the accuracy of the descriptions given by the two models for either group of Scottish children. One possible explanation is that the visual effects were small in these groups. When one independent variable in the expanded factorial design does not have a substantial influence on performance, then the data are not usually informative enough to conclude that either model was used over the other.

The visual effect was larger in the training group. In the pre-training task, the RMSDs of the FLMP ranged from 0.045 to 0.073, with an average RMSD of 0.0584. The RMSD of the SCM ranged from 0.054 to 0.106, with an average RMSD of 0.0688. These differences were not statistically significant, $F(1, 4) = 1.226$, $P = 0.331$.

In the post-training fit, the RMSD of the FLMP ranged from 0.04 to 0.08, with an average RMSD of 0.06. The fit of the mean participant gave an RMSD of 0.04. The RMSD of the SCM ranged from 0.04 to 0.1, gave an average RMSD of 0.08, and a mean participant RMSD of 0.04. An analysis of variance was carried out on the RMSD values from the fits of the FLMP and SCM. The FLMP had lower RMSDs than the SCM, and this difference approached statistical significance, $F(1, 4) = 6.128$, $P = 0.068$. Given the very small number of participants in this training condition, the analysis may not have reached statistical significance because of low power.

To explore this possibility further, we chose the two participants from the Scottish ASD group who were most accurate in identifying the visible speech in the unimodal condition. This meant that our analysis was confined to the seven participants that showed the greatest visual effect. When we added these participants to the comparison the fit of the FLMP was significantly better than that given by the SCM, $F(1, 6) = 7.725$, $P = 0.031$. The average RMSDs across the individual fits were 0.0489 for the FLMP and 0.0875 for the SCM.

In conclusion, these results suggest that when the visual effect of speech was large enough to have meaningful influence, participants with ASD were able to integrate audio–visually in speech perception.

## 6. Discussion

The experiment did find some evidence for a McGurk effect in the control subjects, which was not evident in the ASD subjects (see Fig. 2). However, as can be seen in Table 4, this visual effect may be largely accounted for by group differences in speech-reading ability. After controlling for visual accuracy, this difference between control and ASD group was lost. Furthermore, the model tests showed no differences between the groups in terms of whether or not integration occurred. Therefore, it seems likely that the group difference on the McGurk effect arose because poor visual accuracy in the ASD group meant that visually acquired information was less useful, rather than there being a deficit in audio–visual integration (see Massaro & Bosseler, 2003, for supporting results).

This interpretation receives support from the training experiment. Training improved visual accuracy and also increased the size of the visual effect on bimodal speech perception. This in itself does not necessarily argue for integration as individuals could still be using visual information in some cases and auditory in others. However, it is strong evidence that visual information is used in speech perception by individuals with ASD. We tested whether integration occurred by comparing the FLMP and SCM models against the results from the seven subjects showing the largest influence of visual speech. As with normal children, the Fuzzy Logical Model was a better predictor for these data than was the SCM, indicating that individuals with autistic spectrum disorder integrate normally.

Accuracy on the speech-reading task was predicted by BPVS performance and age in the control group but not in the ASD group. One possible interpretation of this curious pattern of associations may be that in normal individuals with ASD, the ability to perceive speech accurately is somehow functionally connected to word knowledge, whereas these two abilities are 'de-coupled' in individuals with ASD. This may be consistent with recent neuroimaging findings from a study of imitation in autism. Williams et al. (2004) found evidence that in contrast to controls, people with autism were less likely to show additional activity when perceiving an action during imitation compared to action perception alone, suggesting that primary perception in autism is less influenced by pre-existing cognitive representation. Therefore for people with ASD, an experience of speech perception may be less likely to activate a cognitive representation of the word it conveys. A question for future research may be to address how much such 'de-coupling' is the result of impaired primary representation (such as reduced accuracy in speech recognition). If this can be improved by activities such as described here, this could possibly facilitate the development of higher cognitive abilities.

The visual effect among the Scottish children was weaker than found in earlier studies using Baldi, raising the question as to why this might be the case. One possibility is that in contrast to adults who have performed this experiment, and the five Santa Cruz children trained with Baldi, the Scottish children in this experiment started with no prior experience. A few children complained that Baldi was too artificial. Part of this artificiality might have been due to the experimental manipulation of the auditory and visual speech. Six of the nine types of bimodal trials were actually inconsistent in that different auditory and visual syllables were paired. It could be that a greater period of familiarization with ecologically valid speech from Baldi might facilitate perception of the visible speech. This suggestion is born out by the effects of the training exercises that increased the size of the visual effect with the Santa Cruz children.

This study employed children across the IQ range, and in particular the FLMP/SCM model testing utilized results from a combined group of Scottish and Santa Cruz children used a group with a wide range of IQ. It may be that more specific sub-groups of children with types of ASD, have more specific deficit, which this study was too weak to detect, either because of a weak McGurk effect in the higher functioning group or the small group size in the ASD group. This study then warrants replication in other groups, with carefully characterized sub-types of ASD.

In conclusion then, individuals with autistic spectrum disorder do show a reduced McGurk effect, compared to control subjects, but this is more likely to be the result of poor speech-reading ability than abnormal audio–visual integration. This study may point to an important practical application. Baldi has been successful in the teaching of speech, vocabulary, and grammar to children with autism (Bosseler & Massaro, 2003). If mirror neurons serve a cross-modal integratory function, then perhaps counter-intuitively, greater success may be achieved teaching imitative tasks that are multi-modal rather than unimodal in content. Also, training children with autism to speechread, may facilitate development of visual perceptive ability, by utilizing the integration of the auditory and visual information. Whether this might have beneficial knock-on effects for the development of language is another question to be answered by future research.

## Acknowledgements

This project was conducted as part of a research program funded by the Health Foundation. We are especially grateful to Aberdeen City Council and local schools for facilitating this project, and the children and their families from Aberdeen and Santa Cruz who were such willing participants.

## References

Baron-Cohen, S., Ring, H. A., Bullmore, E. T., Wheelwright, S., Ashwin, C., & Williams, S. C. (2000). The amygdala theory of autism. *Neuroscience and Biobehavioral Reviews*, *24*, 355–364.

Baron-Cohen, S., Ring, H. A., Wheelwright, S., Bullmore, E. T., Brammer, M. J., Simmons, A. et al. (1999). Social intelligence in the normal and autistic brain: An fMRI study. *European Journal of Neuroscience*, *11*, 1891–1898.

Berument, S. K., Rutter, M., Lord, C., Pickles, A., & Bailey, A. (1999). Autism screening questionnaire: Diagnostic validity. *British Journal of Psychiatry*, *175*, 444–451.

Bosseler, A., & Massaro, D. W. (2003). Development and evaluation of a computer-animated tutor for vocabulary and language learning for children with autism. *Journal of Autism and Developmental Disorders*, *33*, 653–672.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, *11*, 1110–1123.

Calvert, G. A., Brammer, M. J., & Iversen, S. D. (1998). Crossmodal identification. *Trends in Cognitive Science*, *2*, 247–253.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. et al. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*, 593–596.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649–657.

Chandler, J. P. (1969). Subroutine STEPIT—Finds local minima of a smooth function of several parameters. *Behavioral Science*, *14*, 81–82.

Davies, S., Bishop, D., Manstead, A. S., & Tantam, D. (1994). Face perception in children with autism and Asperger's syndrome. *Journal of Child Psychology and Psychiatry*, *35*, 1033–1057.

Dunn, L. I. M., Dunn, L. M., Whetton, C., & Burley, J. (1997). *The British Picture Vocabulary Scale* (2nd ed.). Windsor, Berks: NFER-Nelson.

Emery, N. J., & Perrett, D. I. (2000). How can studies of monkey brain help us to understand 'theory of mind' and autism in humans? In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding other minds—perspectives from developmental cognitive neuroscience* (2nd ed., pp. 274–306). Oxford: Oxford University Press.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*(Pt. 2), 593–609.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*, 493–501.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*, 2526–2528.

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, *297*, 846–848.

Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Jr., Leventhal, B. L., DiLavore, P. C. et al. (2000). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism Developmental Disorders*, *30*, 205–223.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview—Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism Developmental Disorders*, *24*, 659–685.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development*, *55*, 1777–1788.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale NJ: Erlbaum.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, Massachusetts: MIT Press.

Massaro, D. W., & Bosseler, A. (2003). Perceiving speech by ear and eye: multimodal integration by children with autism. *The Journal of Developmental and Learning Disorders*, *7*, 111–146.

Massaro, D. W., Bosseler, A., & Light, J., 2003. Development and evaluation of a computer-animated tutor for language and vocabulary learning. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS '03), CD-ROM*. Barcelona, Spain.

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

Mistlin, A. J., & Perrett, D. I. (1990). Visual and somatosensory processing in the macaque temporal cortex: The role of 'expectation'. *Experimental Brain Research*, *82*, 437–450.

Rogers, S. J. (1999). An examination of the imitation deficit in autism. In J. Nadel & G. Butterworth (Eds.), *Imitation in Infancy* (pp. 254–283). Cambridge: Cambridge University Press.

Schultz, R. T., Gauthier, I., Klin, A., Fulbright, R. K., Anderson, A. W., & Volkmar, F. (2000). Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and Asperger syndrome. *Archives of General Psychiatry*, *57*, 331–340.

Schultz, R. T., Grelotti, D. J., Klin, A., Kleinman, J., Van der, G. C., Marois, R. et al. (2003). The role of the fusiform face area in social cognition: Implications for the pathobiology of autism. *Philosophical Transactions Royal Society of London B Biological Sciences*, *358*, 415–427.

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio–visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology. A*, *36*, 51–74.

Williams, J. H. G., Waiter, G. D., Perrett, D. I., Murray, A. M., Gilchrist, A., & Whiten, A. (2004). Imitation in autism: A systematic review and a neuroimaging study. *Symposium on Imitation and Autism*, Tampa, Florida. Biennial Meeting. Society for Research in Child Development.

Williams, J. H. G., Whiten, A., Suddendorf, T., & Perrett, D. I. (2001). Imitation, mirror neurons and autism. *Neuroscience and Biobehavioral Reviews*, *25*, 287–295.