Developmental Changes in Visual and Auditory Contributions to Speech Perception

DOMINIC W. MASSARO, LAURA A. THOMPSON, BRIGID BARRON, AND ELIZABETH LAREN

University of California, Santa Cruz

Previous research has demonstrated that listeners make use of both auditory and visual information in bimodal speech perception. Preschool children appear to evaluate and integrate the two sources of information in the same manner as adults, but the children are less influenced by the visual source. The current experiments test a possible explanation of this difference and extend the question to younger children. It is possible that children are poorer lip-readers than adults and thus have less complete information about the visual source. Children and adults were tested with both auditory and visual sources and were also required to identify speech events on the basis of only the visual source. In addition to replicating the previous findings in the bimodal situation, the current experiments found that children are poorer lip-readers than adults. Furthermore, a positive correlation was observed between lip-reading ability and the size of the visual contribution to bimodal speech perception. A fuzzy logical model of speech perception provided a good quantitative description of the results even with the assumption that the visual information was equivalent in both the bimodal and lip-reading conditions. The results also contradict the categorical perception of speech events and any nonindependence in the evaluation of auditory and visual information in speech perception. © 1986 Academic Press, Inc.

This research is concerned with the evaluation and integration of information in bimodal speech perception. The primary interest is the developmental trend in the contribution of visual and auditory information. In addition to the well-known influence of auditory information, it has been amply demonstrated that visual information can have a direct influence on the perceptual experience of spoken speech (McGurk & MacDonald, 1976; Sumby & Pollack, 1954; Summerfield, 1979). The present experiments

The present research was supported in part by NIMH Grant MH-35334. The authors thank Michael M. Cohen for help at all stages of the project, the comments of the PEP people, on the research endeavor, Melanie Mayer for use of the research van, the staff at University Child Care, and most importantly, the children and adult subjects. Send requests for reprints to Dominic W. Massaro, Program in Experimental Psychology, University of California, Santa Cruz, CA 95064.

0022-0965/86 $3.00

extend two previous experiments (Massaro, 1984) which explored developmental aspects of integrating visual and auditory information in speech perception.

Auditory and visual information were independently varied in a speech perception task (Massaro, 1984). Young children and adults identified synthetic speech sounds ranging from /ba/ to /da/ combined with a videotaped /ba/ or /da/ or no articulation. Each subject was instructed to watch and listen to the person speaking on a TV monitor and indicate whether he said /ba/ or /da/. The proportion of /da/ responses increased as the level of the auditory stimulus went from the most /ba/-like to the most /da/-like. The identification responses were also influenced by the visual stimulus, with fewer /da/ responses for the /ba/ visual stimulus than for the /da/ visual stimulus. The significant interaction of these two variables indicated that the effect of the visual variable was larger at the more ambiguous levels of the auditory variable.

Although the comparisons between the two groups of subjects indicated no group differences for the auditory variable, the children showed only half the influence shown by adults for the visual variable. This smaller influence of the visual variable for the children was highly consistent across subjects. One possible explanation of the smaller effect for the children is attentional. Children may attend less to visual than auditory inputs and, thus, show a smaller effect for the visual source (Welch & Warren, 1980). To increase the attentional demands of the visual stimulus, the children were also asked to indicate whether or not the speaker's mouth was moving during the speech event. The additional task of identifying whether or not the speaker's mouth moved had no influence on the phonetic identification of the speech event. The relative influence of the visual source and the integration of this source with the auditory information did not change when the children were also required to pay attention to the visual event.

The results of the Massaro (1984) study were used to test two issues in children's perceptual categorization of events. The first issue concerns whether perception of speech is categorical or continuous. If perception is categorical, subjects are limited to categorical information about the occurrence of a stimulus event. Continuous perception, on the other hand, provides subjects with information about the degree to which some particular event represents a given category (Massaro & Cohen, 1983a). The second issue concerns whether children tend to perceive events holistically rather than in terms of their component parts. These contrasting hypotheses were incorporated into formal models of perceptual categorization and tested against the results. The reader is referred to Massaro and Cohen (1983b) and Massaro (1984) for a complete rationalization and presentation of the contrasting models.

The outcome of the Massaro (1984) studies provided unambiguous evidence against both categorical and holistic perception of the auditory and visual information in speech perception. The results from both adults and children were adequately described by a fuzzy logical model of perception. According to this model, the speech event is transduced by the sensory systems, and various perceptual features are derived. The evaluation of a feature provides continuous information rather than simply categorical information. The features are integrated and assessed in terms of prototype definitions represented by complex fuzzy logical propositions (Massaro & Oden, 1980). Integration of features results in the least ambiguous source having the most impact on the perceptual judgment. The merit of each potential prototype is evaluated relative to the summed merits of the other potential prototypes (Luce, 1959). The relative goodness of a prototype gives the proportion of times it is selected as a response.

Applying this model to auditory and visual speech, both sources are assumed to provide independent evidence for the alternatives /ba/ and /da/. One important auditory cue distinguishing /ba/ from /da/ is the onsets of the second and third formants, and an important visual cue is the degree of opening the lips at the onset of the speech sound. Given independent specifications for the auditory and visual sources, the value of one source cannot change the value of the other at the prototype matching stage. The model described the results in the same way for the children and adults, allowing for differences in the influence of the visual variable.

Even though the fuzzy logical model works equally well for children and adults, we are still left with the fact that the children showed less of an effect for the visual factor. It is possible that children are less sophisticated lip-readers than adults. For example, the cues children use to distinguish a visual /ba/ from a /da/ may be less complete. If the visual variable is less informative for young children than for adults, its influence in bimodal speech perception will be smaller. The present study tests whether children are poorer lip-readers than adults and whether there is a positive correlation between lip-reading ability and the visual influence given bimodal speech.

EXPERIMENT 1

In the first experiment, adults and children were tested in a bimodal condition and a visual-only condition. This provided a comparison between the size of the visual effect in the presence of auditory speech and the subject's ability to identify accurately the two types of visual articulation when no sound is present.

Method

Subjects

Eleven children, 9 males and 2 females, between the ages of 4-6 and 6-10 (mean = 6-0) served as subjects in the children's group. They were recruited from the University Child Care Center. The children were from families of faculty, staff, students, and other residents of the University neighborhood. The adult group was composed of 11 adults, 6 males and 5 females, ranging in age from 16 to 32 (mean =25). They were recruited from within the University community.

Stimuli

The speech events were recorded on a videotape. The speaker was seated in front of a wood panel background, illuminated with fluorescent light. His head filled about two-thirds of the screen. On each trial, the speaker said /ba/ or /da/. An experimental tape was made by copying the original videotape and replacing the natural soundtrack with synthetic speech. Computer-controlled analog-to-digital and digital-to-analog converters were used to monitor the natural sound track and replace it with synthetic speech. The synthetic speech syllable was completely synchronized with the onset of the original syllable.

To create the synthetic speech, the speaker's /ba/s and /da/s were analyzed using linear prediction to derive a set of parameters for driving a software formant serial resonator speech synthesizer (Klatt, 1980). By altering the parametric information regarding the first 80 ms of the syllable, a set of five 400-ms syllables covering the range from /ba/ to /da/ was created. During the first 80 ms *F*1 went from 300 to 700 Hz following a negatively accelerated path. The *F*2 followed a negatively accelerated path to 1199 Hz beginning with one of five values equally spaced between 1125 and 1625 Hz from most /ba/-like to most /da/-like, respectively. The *F*3 followed a linear transition to 2729 Hz from one of five values equally spaced between 2325 and 2825 Hz. This auditory continuum contains somewhat more ambiguous speech sounds relative to the ones used in the Massaro (1984) study.

The children were tested in a research van (Mayer, 1982) located outside the Child Care Center. The adults were tested in a sound-attenuated room. All subjects viewed a 12-in. television monitor, which presented both the auditory and visual dimensions of the speech stimuli. The subjects sat 2 to 3 ft from the monitor. The audio was set at comfortable listening level (70 dB-A visual peak reading with fast scale from Bruel and Kjaer Type 2203 sound level meter).

Procedure

During each trial of the bimodal condition, one of the five auditory stimuli on the continuum from /ba/ to /da/ was paired with one of the two visual stimuli, a /ba/ or a /da/ articulation. A 250-ms bell preceded each trial in the bimodal condition. The silent interval between the bell and the onset of the speech sound ranged from 1175 to 1375 ms. Trials in the visual-only condition used the same tape; however, the sound on the television monitor was turned down completely so that only the visual information was presented (without a warning tone). In both conditions, the subjects had about 6 s to make a response before the next trial.

In the bimodal condition, children were instructed to watch and to listen to the "man on the TV" and to tell the experimenter whether he/she heard the sound /ba/ or the sound /da/. Before the visual-only condition, each child watched the experimenter's mouth as she demonstrated silent articulations of the two alternatives. In this condition, children were instructed to report whether the speaker's mouth made /ba/ or /da/. In both conditions, the experimenter sat next to the monitor and facing the child to determine whether he or she was watching the screen at the time of the speech presentation. If this criterion was not met, the trial was disregarded. The children made their response by oral report. The adults were told to report what they heard the speaker say in the bimodal condition and to be sure that they were watching the speaker on every trial. For the visual-only condition, they were told to lip-read. The adults wrote their responses on an answer sheet and were left alone during the experiment. All subjects were tested for two sessions. There were four blocks of 20 trials during each test session. Subjects always began a session with a bimodal block, and alternated between bimodal and visual-only blocks. Athough the bimodal block was always given first, these two conditions were counterbalanced in Experiment 3 without producing any differences in the results. After each session, the child was given choice of a toy.

Results and Discussion

The results from each group of subjects were analyzed separately for the two experimental conditions. In the bimodal condition, the proportion of /da/ responses as a function of each of the 10 trial types was computed for each subject. An analysis of variance revealed significant main effects for both the auditory factor and the visual factor, $F(4, 80) = 42.43, p < .001$, $F(1, 20) = 162.51, p < .001$, respectively. Figure 1 displays the effects of these variables on subjects' responses separately for each subject group. The effect of the visual factor is larger for the most ambiguous levels of the auditory continuum, giving a significant interaction of the auditory and visual factors, $F(4, 80) = 8.25, p < .001$.

As can be seen in Fig. 1, there were large differences between the children and adult subjects. The visual variable had a much larger influence on the adults' than on the children's judgments, $F(1, 20) = 26.30, p < .001$. The difference between the proportions of /da/ identifications given a visual /da/ and a visual /ba/ was 0.82 for adults and 0.35 for children. This result provides a replication of the Massaro (1984) study. In contrast to the effect of the visual variable, the auditory variable was much more influential for the children than for the adults, $F(4, 80) = 11.00, p < .001$. The difference for the most /da/-like and most /ba/-like speech sounds was .68 for the children and .24 for the adults.

One might argue that the large differences between children and adults are difficult to interpret without some measure of performance to only

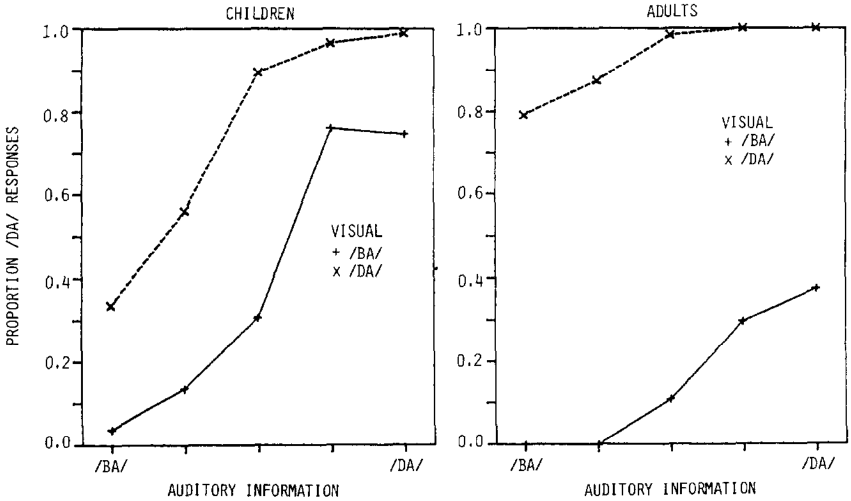

FIG. 1. Observed proportion of /da/ identifications as a function of the auditory and visual levels in the bimodal speech perception condition.

the auditory syllables. If children were much more sensitive to the auditory continuum than were adults, this might reduce the effect of the visual source for the children relative to the adults. Massaro (1984) found no difference between children and adults when the auditory continuum was presented without the speaker moving his mouth. Thus, the group differences with respect to the size of the visual effect appear to be due to differences in the processing of the visual source.

An analysis of variance was performed on the average proportions of correct identifications in the visual-only condition. The adult subjects were much better at lip-reading than children, $F(1, 20) = 13.84$, $p < .001$. The average proportion of correct responses ranged from 0.86 to 1.0 (mean = 0.96) for the adults and from 0.57 to 0.95 (mean = 0.79) for the children. This finding lends credence to the supposition that the developmental decrease in the size of the visual effect during bimodal speech is due to children's decreased ability to lip-read, as compared to adults. There was also a significant main effect of the visual stimulus since subjects were slightly better at identifying visual /ba/ than visual /da/, $F(1, 20) = 7.51$, $p < .025$.

A correlation between subjects' proportion correct in the visual-only condition and the size of their visual effect in the bimodal condition was computed for each subject group and for the combined results of both groups. The size of each subject's visual effect was obtained by subtracting the proportion of /da/ responses given a /ba/ visual stimulus from the proportion of /da/ responses given a /da/ visual stimulus. A larger difference between these two probabilities yields a larger visual effect. Figure 2 is a scatter plot of each subject's accuracy in the visual-only

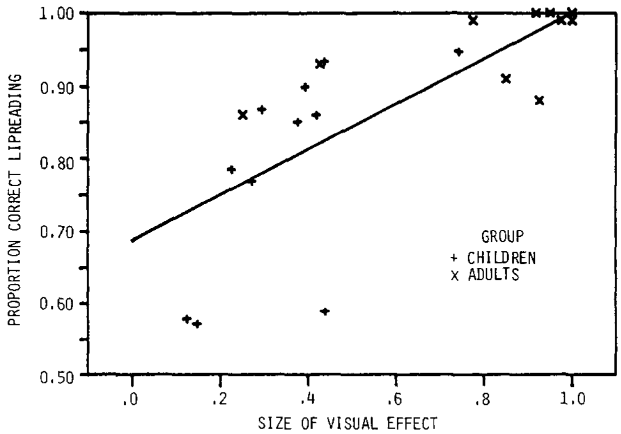

FIG. 2. Scatter plot of each subject's lip-reading accuracy in the visual-only condition against the size of the visual effect in the bimodal condition.

condition against the size of the visual effect in the bimodal condition. The most interesting finding concerns the strong relationship between accuracy of identifying visual /ba/ and visual /da/ and the extent of the influence of the visual component during bimodal speech. The correlation between these variables was significant for both the children's group, $r = .641$, $p < .05$, and for the adult group $r = .670$, $p < .05$, and for the combined results, $r = .75$, $p < .01$. Figure 2 gives the regression line for the combined results. Thus, the obtained results clearly show a strong positive relationship between the ability to lip-read and the extent of the visual influence in bimodal speech perception.

EXPERIMENT 2

This experiment tests three quantitative models of performance in the bimodal speech task. A third level of the visual variable in which the speaker did not move his mouth was included to increase the number of experimental conditions and provide a stronger test of the models. The study extends the Massaro (1984) study to younger children and evaluates additional explanations of the decreased visual influence for young children. The average age of the children in the Massaro (1984) study ranged between 4-9 and 6-9 (mean = 5-11), and it is possible that categorical and/or holistic perception would be found with younger children. A younger group of children was tested ranging in age from 2-5 to 5-3 (mean = 4-2) to assess to what extent the nature of perceptual categorization differs across this age range and contrast it with the performance of the older children tested by Massaro (1984).

To further explore an attentional hypothesis as a contribution to the smaller visual effect in the children, we used two indices that may correlate with individual differences in visual effect. The first was an independent experimenter rating on the task capability of each child. The rating was on a scale from 1 to 10 representing the child's ability to follow instructions and focus on the task. The second indicator was a calculation of the number of missed trials (trials that could not be counted because the subject was not looking at the speaker). In addition, we assessed the correlations of the size of the visual effect with age and sex.

Method

Subjects

The subjects were 23 children, 14 males and 9 females, ages 2-5 to 5-3 years (mean = 4-2), from the Child Care Center at the University of California, Santa Cruz.

Stimuli and Procedure

The stimuli were identical to those used in the bimodal conditions of Experiment 1 and included trials in which the speaker did not move his lips during presentation of the auditory speech. Thus, each trial represented a /ba/, /da/, or no articulation paired with one of the five speech sounds used in Experiment 1. The 15 experimental conditions were sampled randomly without replacement in blocks of 15 trials. The procedure was identical to that used in Experiment 1. Some children also hit one of two buttons on opposite sides of a box in addition to the oral response. One button was marked with the letters BA and a picture of a ball and the other with DA and a duck. The buttons could be hit without looking at them and did not distract the children from looking at the television monitor. Children usually had one or two sessions of 30 trials on a given day. The children were tested at various times during a 6-week period. The subjects were tested for a total of 240 trials, giving up to 16 observations at each of the 15 unique experimental conditions.

To obtain a measure of the subjects' task capability, each experimenter independently rated the child's ability to do the task. The ratings were on a scale of 1 (*worst*) to 10 (*best*). Generally, the ratings were based on attentiveness and ability to focus on the task.

Results and Discussion

For each of the 23 subjects, the proportion of /da/ responses was determined at each of the 15 (5 auditory × 3 visual) possible stimulus conditions. As can be seen in Fig. 3, both the auditory and the visual components of the stimuli significantly influenced the proportion of /da/ responses. The average proportion of /da/ responses with the /ba/ visual articulation was 0.46, whereas this same value with a /da/ visual articulation was 0.75, or a significant visual effect of 0.29, $F(2, 44) = 76.71$, $p < .001$. The proportion of /da/ responses also increased systematically as the speech sound varied from the /ba/ to the /da/ end of the synthetic

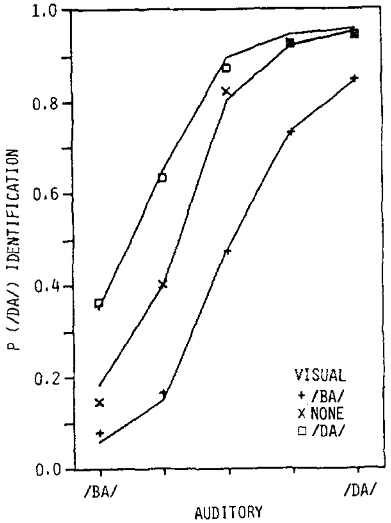

FIG. 3. Observed (points) and predicted (lines) proportion of /da/ identifications as a function of the auditory and visual levels of the speech event. The predictions are for the fuzzy logical model of perception.

speech continuum, $F(4, 88) = 149.47, p < .001$. The interaction between the two variables was highly significant, $F(8, 176) = 14.83, p < .001$. The size of the visual effect can be described in terms of the difference in the proportion of /da/ responses given a visual /da/ minus those given a visual /ba/. At an unambiguous auditory setting (Level 5, most /da/-like), the visual effect is only 0.09. At an ambiguous auditory level (Level 3), this same difference is 0.47, roughly five times the effect.

The magnitude of the visual effect can be compared directly to that observed in the Massaro (1984) study since the experimental design was equivalent. Only two of the five levels of the auditory stimulus were used in both studies, so the comparison must be limited to these two conditions. The size of the visual effect for these two conditions in Massaro (1984) averaged 0.66 for the adults and 0.40 for the children. In the present study, the children gave an average visual effect of 0.34 for these same two conditions. Thus, there is additional evidence that children show a smaller influence of the visual variable than do adults.

To assess for any practice effects, the data were also analyzed as a function of first and second trial blocks, with 120 trials per block. There was no main effect of practice, and practice did not interact with the other variables in the experiment.

Post hoc correlations were carried out to assess the relationship between the magnitude of the effects of the auditory and visual sources and four predictor variables. The magnitude of a given source was set equal to the difference between the response proportions to the extreme values of that source. For example, the magnitude of the visual source was equal to the proportion of a /da/ response, $P(/\mathrm{da}/)$, given the visual /da/ minus $P(/\mathrm{da}/)$ given the visual /ba/. The magnitude of the auditory source was defined analogously by subtracting $P(/\mathrm{da}/)$ given the first auditory level from $P(/\mathrm{da}/)$ given the fifth (most-/da/) auditory level. The four predictor variables were age, sex, the number of discarded trials that the child was not looking at the television monitor, and an average rating of the experimenters between 1 and 10, reflecting the child's task capability. Although both age and the ratings significantly correlated $-.42$ and $-.62$ with the number of missed trials, neither of these two variables were significantly correlated with the magnitude of either of the auditory or visual sources. None of the four prediction variables significantly correlated with the magnitude of the auditory source.

There was a significant ($p < .02$) negative correlation of $-.48$ between the number of missed trials and the magnitude of the effect of the visual source. Thus, children who tended to miss trials because they failed to look at the screen showed a smaller effect of the visual articulation, even when they were looking at the screen. However, this correlation does not seem to illuminate the reason for the smaller effect. If a child obtains less information from the visual variable, it is only natural to be less motivated to look at the screen. An attentional explanation does not seem to be sufficient to account for the reduced size of the visual variable for young children. Massaro (1984) also provided evidence against an attentional explanation, since children were able to indicate also whether or not the speaker's lips were moving, and the effect of the visual variable did not change.

Tests of Three Models

Massaro and Cohen (1983b) and Massaro (1984) formalized three models of bimodal speech perception. In the categorical model, the listener has only categorical information representing the auditory and visual dimensions of the speech event. This model implies that separate categorical (phonetic) decisions are made to the auditory and visual sources of bimodal speech (MacDonald & McGurk, 1978). Given only the two alternatives /ba/ and /da/ in the present task, separate /da/ or /ba/ decisions would be made to both the auditory and visual sources, and the identification response would be based on these separate decisions. Given categorical information, there are only four possible outcomes for a particular combination of auditory and visual information: /da/–/da/, /da/–/ba/, /ba/–/da/, or /ba/–/ba/. If the two decisions to a given speech event agree, the identification response can follow either source. If the two decisions disagree, it is reasonable to assume that the subject will respond with the decision of the auditory source on some proportion p of the trials, and respond with the decision of the visual source on the remainder

$(1 - p)$ of the trials. In this conceptualization, the magnitude of p relative to $(1 - p)$ reflects the relative dominance of the auditory source.

The probability of a /da/ identification response, $P(D)$, given a particular auditory/visual speech event, A_iV_j, would be

$$\begin{aligned} P(\mathrm{D}{:}A_iV_j) &= \{1a_iv_j\} + \{pa_i(1 - v_j)\} \\ &\quad + \{(1 - p)(1 - a_i)v_j\} + \{0(1 - a_i)(1 - v_j)\} \\ &= pa_i + (1 - p)v_j \end{aligned}$$

where i and j index the levels of the auditory and visual stimuli, respectively. The a_i value represents the probability of a /da/ decision given the auditory level i and v_j is the probability of a /da/ decision given the visual level j. Each of the four terms in the equation represents the likelihood of one of the four possible outcomes of the separate decisions multiplied by the probability of a /da/ identification response given that outcome. In Experiment 2, five auditory levels are factorially combined with three visual levels. In this model, each unique level of the auditory stimulus would require a unique parameter a_i, and analogously for v_j. Since the parameter p reflects a decision variable, its estimated value would be constant across all stimulus conditions. Thus, a total of nine parameters must be estimated for the 15 independent conditions in the bimodal task.

Applying the fuzzy logical model to the present task using auditory and visual speech, both sources are assumed to provide independent evidence for the alternatives /ba/ and /da/. Defining the important auditory cue as the onsets of the second and third formants (*F*2–*F*3) and the important visual cue as the degree of initial opening of the lips, the prototypes are

/da/ : Slightly falling *F*2–*F*3 & Open lips

/ba/ : Rising *F*2–*F*3 & Closed lips.

Given a prototype's *independent* specifications for the auditory and visual sources, the value of one source cannot change the value of the other source at the prototype matching stage. In addition, the negation of a feature is defined as the additive complement (Zadeh, 1965). That is, we can represent Rising *F*2–*F*3 as (1 − Slightly falling *F*2–*F*3) and Closed Lips as (1 − Open lips),

/da/ : Slightly falling *F*2–*F*3 & Open lips

/ba/ : (1 − Slightly falling *F*2–*F*3) & (1 − Open lips).

The integration of the features defining each prototype is evaluated according to the product of the feature values. If a_i represents the degree to which the auditory stimulus A_i has Slightly falling *F*2–*F*3 and v_j represents the degree to which the visual stimulus V_j has Open lips, the outcome of prototype matching would be

$$/\text{da}/ : a_i v_j$$

$$/\text{ba}/ : (1 - a_i)(1 - v_j).$$

If these two prototypes were the only valid response alternatives, the pattern classification operation would determine their relative merit leading to the prediction that

$$P(\text{D}{:}A_i V_j) = \frac{a_i v_j}{a_i v_j + [(1 - a_i)(1 - v_j)]}.$$

Given five levels of A_i and three levels of V_j in the present task, the predictions of the model require eight parameters (five a_i values and three v_j values).

According to the dependence model, the visual and auditory sources are not evaluated independently since the stimulus event is perceived holistically. According to this view (Shepp, 1978; Smith & Kemler, 1978), independent dimensions might be present in the stimulus environment but not in the processing of the subject. It is very difficult to formalize and test the holistic model, unless a particular type of dependence between the sources is specified exactly. If no type of dependence is assumed, it is necessary to estimate a unique parameter for each unique set of experimental conditions. Thus, the holistic model would require as many parameters as there are independent conditions. This violation of parsimony might be sufficient for some to reject the holistic model as a meaningful description of performance. Rather than rejecting it without test, however, two tests of the holistic hypothesis are proposed. If the contribution of one source is dependent on the value of the other, any model assuming independent contributions of each source must fail. To the extent that the independence assumption of the fuzzy logical model gives an adequate description of the results, we have evidence against the holistic hypothesis. A second test is to assume a particular form of dependence. Massaro and Cohen (1977) found a linear dependence between voicing amplitude and duration of the fricative for members of a fricative-vowel continuum going from /si/ to /zi/. Thus, it is reasonable to test this form of dependence between the auditory and visual sources of information. Given a good description of the fuzzy logical model and a poor description of the dependence model, we have evidence against the hypothesis of holistic processing.

The dependence hypothesis is formalized using analogous operations to those involved in the fuzzy logical model. Thus, the /da/ alternative has a prototype description, but in this case the description is simply in terms of a holistic source of information, which is the product of the auditory and visual sources

$$/\text{da}/ : (\text{Slightly falling } F2\text{–}F3 \times \text{Open lips}) = (a_{ij})$$

where a_{ij} is the product of the auditory and visual sources:

$$a_{ij} = a_i v_j.$$

Given one holistic source, the prototype description for /ba/ is equal to 1 minus the prototype for /da/

$$/\text{ba}/ : 1 - (\text{Slightly falling } F2\text{–}F3 \times \text{Open lips}) = 1 - (a_{ij})$$

This dependence formalization assumes that only a single, multiplicatively combined (holistic) feature is available for prototype matching. The prototype matching and the pattern classification operations would be identical in form to those assumed in the fuzzy logical model.

Given five levels of A_i and three levels of V_j in the present task, the predictions of the dependence model require eight parameters (five a_i values and three v_j values).

The predictions of the categorical model, the dependence model, and the fuzzy logical model were tested against individual subject's performance. The quantitative predictions of the three models were derived for the proportion of /da/ responses for each subject for each of the 15 conditions using the program STEPIT (Chandler, 1969). A model was represented to the analysis program STEPIT as a set of prediction equations and a set of unknown parameters. Given a model, STEPIT finds a set of parameter values which come closest to predicting the observed data. Initially, all parameters were set to 0.5. The parameters of each model were adjusted iteratively to minimize the sum of the squared deviations between the 15 observed and predicted proportions of /da/ responses for each subject. The measure of deviation of each model to the results is the square root of the average squared deviation, called the root mean squared deviation or simply deviation, between the observed and predicted results.

Figure 4 gives the average predicted results of the categorical model; this model gave a poor description of the observed results with an average deviation of 0.111. Figure 5 gives the analogous predictions for the dependence model, which also failed to capture the pattern of observed results with a goodness of fit of 0.123. The fuzzy logical model provided a much better description, as can be seen in Fig. 3 and the deviation of 0.048. Two analyses of variance were carried out on the individual deviations contrasting both the categorical and dependence models against the fuzzy logical model. The fuzzy logical model gave significantly lower deviations compared to both the categorical, $F(1, 22) = 64.01, p < .001$, and the dependence, $F(1, 22) = 74.22, p < .001$, models. It should be stressed that the good description of the fuzzy logical model cannot be simply due to a large number of free parameters. The categorical and

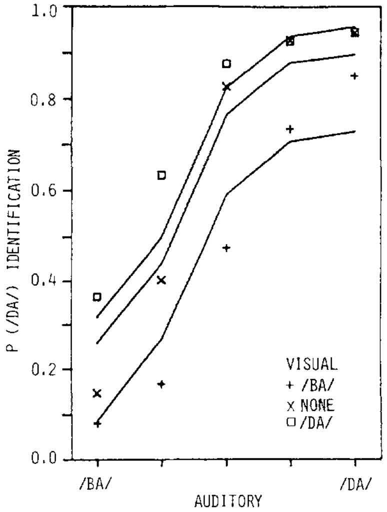

FIG. 4. Observed (points) and predicted (lines) proportion of /da/ identifications as a function of the auditory and visual levels of the speech event. The predictions are for the categorical model.

dependence models required as many or more parameters and gave significantly poorer descriptions of the results.

To assess whether the descriptions of the three models were related to subject variables, the deviation values were entered in the correlation analyses with the independent variables described earlier. The description of the models did not correlate with age, sex, or task capability rating.

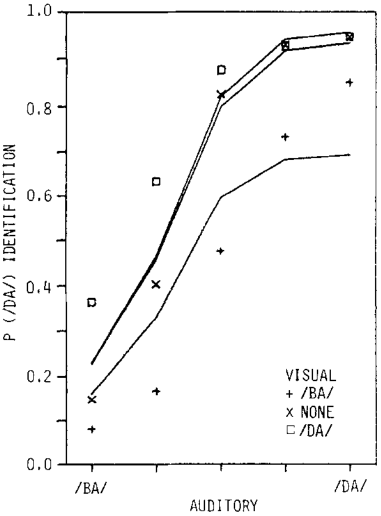

FIG. 5. Observed (points) and predicted (lines) proportion of /da/ identifications as a function of the auditory and visual levels of the speech event. The predictions are for the dependence model.

There is no hint of a developmental trend, in terms of relative adequacy of the three models to describe the observed results. Thus, we have evidence for the processes postulated by the fuzzy logical model for children as young as 2-5 and no evidence that speech perception is more categorical or holistic for younger children. These correlation analyses provide additional support for the fuzzy logical model over the categorical and dependence models.

There were positive correlations ($r = .78$ and $r = .67$, $p < .001$) between the deviation values for the categorical and dependence models and the size of the effect of the visual variable. Both the categorical ($r = -.49$, $p < .01$) and dependence ($r = -.53$, $p < .01$) models also gave somewhat better descriptions of a subject's results to the extent the subject missed trials, which probably reflects the negative correlation ($r = -.48$, $p < .02$) between the number of missed trials and the size of the effect of the visual variable. These same correlations were not significant for the fuzzy logical model. These correlations substantiate the good description of the fuzzy logical model and the poor descriptions of the categorical and dependence models and can be interpreted as follows. If a model inaccurately describes the combination of visual and auditory information in speech perception, it will give a poorer description of the results to the extent a subject shows a large effect of the visual variable. If a model is accurate in its description, increasing the size of the effect of a variable does not decrease the accuracy of its description.

The significant negative correlation ($r = -.67$, $p < .001$) between the effect of the auditory variable and the goodness-of-fit values of the fuzzy logical model shows a better description given by this model for subjects showing a larger effect of the auditory variable. Thus, the fuzzy logical model gives an even better description of the results for subjects showing a large effect of the auditory variable. This correlation might only reflect less variance in the judgments when the proportions are very close to 0 or 1 as would more likely be the case with a large effect of the auditory variable. The goodness of fit of the categorical and dependence models did not correlate with the effect of the auditory variable.

EXPERIMENT 3

Experiment 3 was carried out to replicate the results of Experiment 1 and to provide a stronger test of the fuzzy logical model. If children do not lip-read as well as adults, then the visible speech must be less informative in the bimodal condition. In terms of the fuzzy logical model, the amount of /da/-ness given by a /da/ articulation would be less for children than for adults. To test this hypothesis, we retested some of the children from Experiment 2 and included a lip-reading condition in which no sound was presented. If lip-reading performance is totally responsible for the reduced size of the effect of the visible speech, then the information available in lip-reading should be identical to that used for the visual source in the bimodal condition. In the most parsimonious form of the fuzzy logical model, a single value of visual /da/-ness should be available in both the visual-alone and bimodal conditions.

Method

Subjects

Twenty-one children were tested: 15 of these children had participated in Experiment 2. There were 9 males and 12 females and the age range was 2-5 to 5-11 with a mean of 4-5.

Stimuli

The videotape used in Experiment 2 was edited to produce two new tapes. The neutral trials were eliminated from both tapes. For the lip-reading task, the speech sounds were eliminated from the tape. For the bimodal task, the tape was identical to that used in the second experiment, except for the absence of "no articulation" trials. Thus, subjects lip-read without sound or identified the speech event given both the sound and the visual articulation. Subjects were tested for 20 trials in each of these two conditions on a given day for a total of 4 days. The order of presentation of the two conditions was counterbalanced across subjects and across days. All other procedural details were identical to those in Experiment 2.

Results and Discussion

The proportion of /da/ identifications was computed for each subject for both the lip-reading and bimodal conditions. Figure 6 gives the proportion of /da/ identifications as a function of the two variables in the bimodal and visual-alone conditions. In the bimodal condition, the proportion of /da/ responses was influenced by both the visual and auditory variables and their interaction, $F(1, 20) = 92.40$, $F(4, 80) = 55.41$, and $F(4, 80) = 11.46$, $p < .001$. The results replicate those of Experiments 1 and 2 as can be seen in a comparison of Figs. 1, 3, and 6. The size of the visual effect was 0.36, which is comparable to the 0.35 and 0.29 effects for the children in Experiments 1 and 2.

As can also be seen in Fig. 6, lip-reading performance revealed that subjects could differentiate the /ba/ and /da/ articulations, $F(1, 20) = 74.75$, $p < .001$. Subjects identified /ba/ correctly 0.76 of the time and /da/ correctly 0.73 of the time.

An analysis was also carried out on performance as a function of the first and second halves of the experiment. It is possible that experience on the task would modify lip-reading performance and/or the size of the visual effect. However, there was no effect of experience in either the lip-reading or the bimodal conditions.

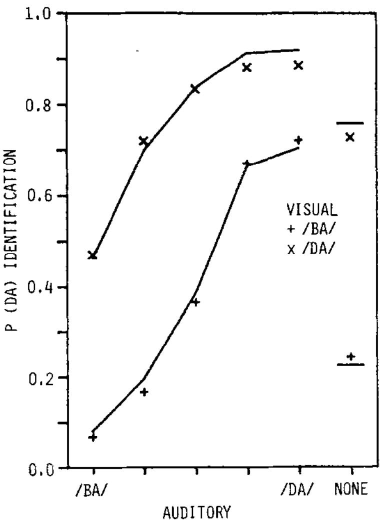

FIG. 6. Observed (points) and predicted (lines) proportion of /da/ identifications as a function of the auditory and visual levels of the speech event in the bimodal condition and in the visual-only condition.

Fifteen subjects participated in both Experiments 2 and 3. The correlation analyses were reported to assess the consistency in results across the two experiments. The size of the visual variable correlated .844 and the auditory variable .676 across the two experiments. Hence, subjects were very consistent in performance across the two studies.

Test of the Fuzzy Logical Model

To assess more formally whether lip-reading performance could predict the magnitude of the visual variable in the bimodal condition, we asked whether the visual /da/-ness computed from the lip-reading condition could be used as the visual parameters in the bimodal condition. Applying the fuzzy logical model to the lip-reading task, we assumed that the subject derives visual information and evaluates the degree to which that information supports the alternative /ba/ and supports the alternative /da/. It is reasonable to assume that the amount of /ba/-ness is equal to 1 minus the amount of /da/-ness. Hence, the probability of responding /da/ in the lip-reading task can be thought of corresponding to the amount of /da/-ness (given by the lips) divided by the sum of the /da/-ness and /ba/-ness values. If v_j is the visual /da/-ness, then the probability of a /da/ response given a particular lip movement (L_j) would be

$$P(\text{da}:L_j) = \frac{v_j}{v_j + (1 - v_j)} = v_j$$

The equation shows that the lip-reading accuracy can be taken as a direct index of visual information. The probability of a /da/ identification corresponds to the visual /da/-ness, whereas the probability of a /ba/ identification corresponds to the visual /ba/-ness.

Given variability in the results, we cannot expect the lip-reading task to provide a perfect estimate of the visual information in the bimodal task. To reduce the effect of this variability, it is possible to test the same hypothesis by estimating the two visual parameters based on performance in both the lip-reading and the bimodal task. Thus, five auditory and two visual parameters are estimated to predict 12 identification probabilities (10 from the bimodal task and 2 from the lip-reading task). The average fit of this model to the individual results is given in Fig. 6. The deviation for the average results was 0.0312 and the average deviation for the 21 individual-subject fits was 0.0655. These goodness of fit values are very similar to the analogous deviation values of 0.0341 and 0.0627 when two visual and five auditory parameters were estimated to describe the bimodal speech results. An analysis of variance on the deviation values revealed no significant difference for the two models, $F(1, 20) = 1.93, p > 0.178$. Hence, the quality of the fit was not significantly reduced with the constraint that the same visual information was available in the lip-reading and bimodal conditions.

The test of the fuzzy logical model simultaneously supported both the model and the lip-reading explanation of the group differences. The contribution of visual information in the bimodal condition is equivalent to the visual information used in lip-reading without sound. Children are poorer lip-readers than adults and also show a smaller visual effect in the bimodal condition. The same visual information could be used in the fuzzy logical model to describe performance in both the lip-reading and bimodal conditions.

GENERAL DISCUSSION

The present set of experiments was conducted to investigate why young children's perceptions of bimodal speech are less influenced by the visual component of speech than adults' perceptions are. The results clearly argue in favor of the explanation that children are poorer lip-readers than adults. Not only were children significantly worse than adults at discriminating a /ba/ from a /da/, given just the visual information, but both the children and the adult groups showed a significant positive correlation between ability to lip-read and the extent of the visual influence during bimodal speech.

Evidence was also obtained against the possibility that children's incomplete visual information is due to their inability to attend to the visual information. Neither age nor an index of children's task capability correlated significantly with the size of the visual effect. Those subjects who missed more trials because they were not looking at the screen did show a smaller effect of the visual component during bimodal speech. However, this is not sufficient evidence in favor of an attentional explanation since it is possible that those children who do not read lips very well would tend not to look at the speaker's mouth.

Other evidence against the attentional hypothesis comes from the finding in Experiment 3 that the quality of the visual information for children is the same in the lip-reading and the bimodal identification task. If a child shows a reduced effect of the visual source in bimodal speech perception because of attentional limitations, then we might expect better lip-reading performance given only the visual source than expected from bimodal identification. Since this was not the case and increasing attentional demands in the Massaro (1984) study had no effect, lip-reading rather than attentional differences seems the preferable explanation of the developmental differences.

To help illuminate the developmental aspects involved in the perception and integration of visual and auditory speech, a separate group of even younger children was tested over an extended time period. These children performed almost as well as the older children on the lip-reading task, and their performance did not change with more experience with the task. Therefore, the ability to lip-read seems to involve a fairly long-term process, with performance similar to adults occurring sometime after the child's 6th year. Although children have less lip-reading ability than do adults, there is no evidence for any change in this ability during the 3 years before first grade. In other unpublished experiments, we have found that fourth graders are significantly better lip-readers so that the developmental increase in the use of visual information must occur sometime after schooling begins. Further research is necessary to chart a more complete course of development with respect to lip-reading ability.

An account of the developmental aspects involved in speech perception would be incomplete without an information-processing framework. It enables us to determine whether developmental differences are due to differences in the way information from one source is evaluated and/or differences in the integration with information from the other source. Our results were tested against three formal models by quantifying two extant developmental issues: whether speech information is continuous or categorical, and whether information from the visual and auditory sources is perceived independently or holistically. The data from both the children and the adults in the current study were well described by the fuzzy logical model, which assumes that the perceiver utilizes continuous and independent sources of information. Further, both children and adults integrate the two sources of information so that the least ambiguous source has the most impact on the perceiver's interpretation of a speech event.

In terms of the fuzzy logical model, the fact that the size of the visual effect is smaller for children than for adults can be explained by assuming a developmental difference in the cue value of the visual source. Since children are poorer than adults at lip-reading, the quality of their visual information (cue value) is reduced. Support for this interpretation was found by comparing two versions of the fuzzy logical model against the results from the third experiment. This comparison included models assuming the same, or different, visual information in the lip-reading and bimodal conditions. The model fit was no better when it could predict just the bimodal condition relative to predicting both the bimodal and lip-reading conditions. Hence, it can be forcefully argued that the quality of the information derived from the visual input is approximately the same in both cases, so that children may be said to ''read lips'' the same way for speech events with or without speech sounds.

The evidence for independent and continuous auditory and visual sources of information in bimodal speech perception stands in marked contrast to two well-known beliefs. First, there has been a tradition for theories of perceptual development to view the child progressing from holistic processing to dimensional processing (James, 1890; Shepp, 1978; Smith & Kemler, 1978; Werner, 1957). However, the experimental results used to support such a view might be better interpreted in terms of different strategies affecting task performance rather than in terms of basic differences in perceptual processes (Kemler & Smith, 1979; Massaro, 1984). Second, the phenomenon of categorical perception has traditionally been identified with speech perception. There is now convincing evidence, however, for the availability of continuous information within speech categories in both infants (Aslin, Pisoni, Hennessy, & Perey, 1981; Eimas & Miller, 1980; Miller & Eimas, 1983) and adults (Carney, Widin, & Viemeister, 1977; Massaro & Cohen, 1983a; Pisoni & Lazarus, 1974; Samuel, 1977). Given that there is considerable evidence for the independent evaluation of continuous sources of information, the fuzzy logical model offers a productive framework for the study of perceptual development in a variety of domains.

REFERENCES

Aslin, R. N., Pisoni, D. B., Hennessy, B. L., & Perey, A. J. (1981). Discrimination of voice-onset-time by human infants: New findings concerning phonetic development. *Child Development*, **52**, 1135–1145.

Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, **62**, 961–970.

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

Eimas, P. D., & Miller, J. L. (1980). Discrimination of information for manner of articulation. *Infant Behavior and Development*, **3**, 367–375.

James, W. (1890). *Principles of psychology* (Vol. 2). New York: Holt.

Kemler, D. G., & Smith, L. B. (1979). Accessing similarity and dimensional relations: Effects of integrality and separability on the discovery of complex concepts. *Journal of Experimental Psychology: General*, **108**, 133–150.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971–995.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

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

Massaro, D. W., & Cohen, M. M. (1977). The contribution of voice-onset-time and fundamental frequency as cues to the /zi/-/si/ distinction. *Journal of the Acoustical Society of America*, **22**, 373–382.

Massaro, D. W., & Cohen, M. M. (1983a). Categorical or continuous speech perception: A new test. *Speech Communication*, **2**, 15–35.

Massaro, D. W., & Cohen, M. M. (1983b). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 753–771.

Massaro, D. W., & Cohen, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Advances in basic research and practice* (Vol. 3). New York: Academic Press.

Mayer, M. J. (1982). A mobile research laboratory. *Behavioral Research Methods & Instrumentation*, **14**, 505–510.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, **24**, 253–257.

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

Miller, J. L., & Eimas, P. D. (1983). Studies on the categorization of speech by infants. *Cognition*, **13**, 135–165.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, **55**, 328–334.

Samuel, A. G. (1977). The effect of discrimination training on speech perception: Noncategorical perception. *Perception & Psychophysics*, **22**, 321–330.

Shepp, B. E. (1978). From perceived similarity to dimensional structure. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization*, Hillsdale, NJ: Erlbaum.

Smith, L. B., & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, **10**, 502–532.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212–215.

Summerfield, Q. (1979). Use of visual information in phonetic perception. *Phonetica*, **36**, 314–331.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, **88**, 638–667.

Werner, H. (1957). The conception of development from a comparative and organismic point of view. In D. Harris (Ed.), *The Concept of development*. Minneapolis: Univ. of Minnesota press.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, **8**, 338–353.

RECEIVED: December 31, 1984; REVISED: April 11, 1985.