Children's Perception of Visual and Auditory Speech

Dominic W. Massaro

University of California, Santa Cruz

MASSARO, DOMINIC W. *Children's Perception of Visual and Auditory Speech.* CHILD DEVELOPMENT, 1984, **55**, 1777–1788. Preschool children's evaluation and integration of visual and auditory information in speech perception was compared with that of adults. Subjects identified speech events, which consisted of synthetic speech syllables ranging from /ba/ to /da/ combined with a videotaped /ba/, /da/, or no articulation. Both variables influenced the identification judgments for both groups of subjects. The results were used to test current views of the development of perceptual categorization and speech perception. Tests of quantitative models indicated that both preschool children and adults had available continuous and independent sources of information. The results were well described by a fuzzy logical model of perception, which assumes that the perceiver integrates continuous and independent sources of information and determines the relative goodness of match to prototype definitions in memory. The only developmental difference was less of an influence of the visual source of information for children relative to adults. 1 explanation is that the children simply attended less to the visual source. A second experiment eliminated the attentional explanation by showing identical results when the children were also required to indicate whether or not the speaker's mouth was moving.

This research is concerned with the evaluation and integration of information in speech perception. The experimental techniques of information integration theory (Anderson, 1981, 1982) are used, and quantitative models of performance are tested (Oden & Massaro, 1978). The goal is to study developmental aspects of speech perception while also testing various accounts of developmental patterns in perceptual categorization.

The domain of my study is the contribution of visual and auditory information to speech perception. Although speech perception is usually thought of as strictly an auditory process, it appears to be visual as well. Viewing the speaker can enhance understanding, especially when the auditory signal is degraded. Three decades ago, Sumby and Pollock (1954) demonstrated that perceiving the face of a speaker significantly improved adults' understanding of speech in noise, and Summerfield (1979) found similar results when the test message was embedded in an irrelevant passage of prose. The visual influence is not limited to situations with degraded auditory inputs. As reported by McGurk and MacDonald (1976), the visual input from the speaker can change the perceptual experience of an auditory speech event. When a labial speech sound /ba-ba/ was dubbed onto the visual articulation of a velar stop consonant /ga-ga/, subjects viewing and listening to the videotape often heard /da-da/.

In a previous study using college students as subjects, Massaro and Cohen (1983b) independently varied auditory and visual information in a speech perception task. Subjects identified as /ba/ or /da/ speech events consisting of high-quality synthetic speech syllables ranging from /ba/ to /da/ combined with a videotaped /ba/ or /da/ or no articulation. Although subjects were instructed specifically to report what they heard, viewing the visual articulation made a large contribution to identification. There were significant effects of both the visual and auditory sources of information and an interaction between these variables. The contribution of one source was larger to the extent the other source of information was ambiguous.

The first experiment replicates the Massaro and Cohen (1983b) study with preschool children and adults to test two current hypoth-

This research was supported in part by NIMH grant MH-35334. The author would like to thank Michael M. Cohen for help at all stages of the project, Marjorie Horton and Laura Thompson for comments on the paper, the comments of the PEP people on the research endeavor, Melanie Mayer for use of the research van, the staff at University Child Care, the adult subjects, and most important, the children who found the experience as much fun as I did. Send correspondence to Dominic W. Massaro, Program in Experimental Psychology, Clark Kerr Hall, University of California, Santa Cruz, California 95064.

[*Child Development*, 1984, **55**, 1777–1788. © 1984 by the Society for Research in Child Development, Inc. All rights reserved. 0009-3920/84/5505-0008$01.00]

eses concerning the development of perceptual categorization. The first question concerns whether a stimulus variable is perceived categorically or continuously. Much of the speech research with infants has been interpreted as supporting the categorical perception of certain phonetic contrasts (Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Gleitman & Wanner, 1982). In contrast, recent research with adult subjects has demonstrated that listeners have available continuous information corresponding to the degree to which a speech event represents a given perceptual category (Carney, Widin, & Viemeister, 1977; Massaro & Cohen, 1983a, 1983b; Pisoni & Lazarus, 1974; Samuel, 1977). For example, Massaro and Cohen (1983a, 1983b) demonstrated that the auditory /ba/-/da/ continuum was perceived continuously rather than categorically. In contrast, this same dimension is supposedly perceived categorically by infants (Eimas, 1974). Thus, it is possible that preschool children would tend to produce categorical results, whereas adults would not, given the /ba/-/da/ continuum in the present experiment. To test this hypothesis, it is necessary to formalize and test specific models of categorical and continuous perception.

The categorical model would assume that 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. 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:

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

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 this task, 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 p reflects a decision variable, its value would be constant across all stimulus conditions. Thus, a total of nine parameters must be estimated for the 15 independent conditions in the task.

The assumption of continuous perception will be tested within the context of a fuzzy logical model of perception (FLMP). According to the FLMP, perceptual categorization is carried out in three operations. The first operation is feature evaluation, during which the stimulus is transduced by the sensory systems and various perceptual features are derived. The features are assumed to be continuous rather than categorical. Thus, the outcome of featural evaluation is not categorical but is represented by a continuous variable reflecting the degree to which each relevant feature is present in the speech stimulus. These continuous values are assumed to be analogous to the truth values in the theory of fuzzy sets (Zadeh, 1965), which explains the first term of the model's name.

The second operation is prototype matching that involves the integration of the features. The featural information is combined following the rules given by prototype definitions in long-term memory. A prototype defines a perceptual unit of speech in terms of arbitrarily complex fuzzy logical propositions (Massaro & Oden, 1980). The outcome of prototype matching determines to what degree each prototype is realized in the speech event.

The third operation is pattern classification in which the merit of each potential prototype is evaluated relative to the summed merits of the other potential prototypes (Luce, 1959). This relative merit gives the proportion of times a prototype would be selected as a response. An important property of the model is that one cue has its greatest effect when the second is at its most ambiguous level. The most informative cue has the greatest impact on the judgments.

Applying the model to the 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 and the important visual cue as the degree of initial opening of the lips, the prototypes are

/da/ : Slightly falling F2-F3 and open lips
/ba/ : Rising F2-F3 and closed lips,

where F2-F3 represent the onsets of the second and third formants. 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. That is, we can represent rising F2-F3 as (1 − slightly falling F2-F3) and closed lips as (1 − open lips),

/da/ : Slightly falling F2-F3 and open lips
/ba/ : (1 − slightly falling F2-F3) and (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 F2-F3 and v_j represents the degree to which the visual stimulus V_j has open lips, the outcome of prototype matching would be:

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

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

$$P(D|A_iV_j) = \frac{a_i v_j}{a_i v_j + [(1 - a_i)(1 - v_j)]}.$$

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

The second hypothesis to be tested concerns the independence of the two sources of information. According to the independence view of the FLMP, the auditory and visual inputs provide independent sources of information about the speech event. A contrasting assumption claims that the visual and auditory sources are not evaluated independently but that 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. Shepp (1978) and his colleagues (Shepp, Burns, & McDonough, 1980) and Smith and Kemler (1977, 1978) have proposed that there is a developmental trend from holistic processing to dimensional processing. Preschool children supposedly process some stimuli holistically, whereas adults do not. If this hypothesis is correct for visual and auditory speech events, then holistic (nonindependent) processing should be found for preschool children but not for adult subjects.

The test of the independence assumption can be viewed as equivalent to a test of the FLMP. On the other hand, 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 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 independence model and a poor description of the dependence model, we have evidence against the hypothesis of holistic processing.

The dependence hypothesis will be formalized using the same operations involved in the FLMP. 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

/da/ : (Slightly falling F2-F3 × 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 one minus the prototype for /da/

/ba/ : 1 − (Slightly falling F2-F3 × open lips)
= 1 − (a_{ij})

This dependence formalization assumes that only a single, multiplicatively combined (holistic) feature is available for prototype matching.

Given these prototype definitions and the pattern classification operation, the dependence model makes the following prediction:

$$P(D|A_iV_j) = \frac{a_{ij}}{a_{ij} + (1 - a_{ij})} = a_{ij}.$$

Following the logic used for the FLMP, the dependence model requires the same number of parameters (eight) as does the FLMP with independent features.

In addiiion to testing the two hypotheses of developmental trends in categorical/continuous and independent/holistic perception, a number of other developmental questions can be asked. First, will young children be similarly influenced by the visual information as were the college students in the Massaro and Cohen (1983b) study? Although McGurk and MacDonald (1976) found less of an influence of the visual event for 3–8-year-old children than for adults, no explanation for the difference was offered or is apparent.

A second question concerns the integration of different sources of information in speech perception. The integration of the visual and auditory sources could not be assessed in the McGurk and MacDonald (1976) study because the two sources were not manipulated in a factorial design by varying the quality (or degree of ambiguity) of the auditory source. Qualitatively different integration rules could describe the identification performance of adults and young children. As an example, young children may add rather than multiply the two sources of information. Using the methods of information integration theory and functional measurement, Anderson and Cuneo (1978) asked children to judge the area of rectangles. The height and width of rectangular cookies were varied in a factorial design, and children rated how happy they would be to have that much cookie to eat. In contrast to the more veridical multiplying rule used by older children, 5-year-olds gave results consistent with an adding rule.

Finally, the FLMP has not been tested against the performance of young children. If the model gives an adequate description of the results, it provides a framework for testing for developmental trends. Assuming that the model is valid, the parameter estimates for the contributions of the auditory and visual sources provide direct indexes of their influence. Significant differences in the parameter estimates will represent developmental differences in the influence of a source.

Experiment 1

Children were tested and compared with adults in a speech identification task. Subjects were asked to view a speaker on a TV monitor and to indicate whether the speech event was /ba/ or /da/. Five levels of auditory information going from /ba/ to /da/ were factorially combined with three levels of visual information: /ba/, no articulation, and /da/. The results will be used to test (*a*) continuous versus categorical perception, (*b*) independent versus holistic processing of features, (*c*) the nature of integrating two features in speech perception, (*d*) the relative influence of visual and auditory features, and, most important, (*e*) whether the answers to these questions change with development.

Method

Subjects.—The subjects were 11 children, ranging in age from 4-9 to 6-9 (mean = 5-11), and 11 adults between 18 and 38 (mean = 25). The children were recruited from the University Child Care Center and the adults from the university community.

Stimuli.—A color videotape was made of the author seated in front of a wood panel background, illuminated with ordinary fluorescent fixtures in the ceiling. His head was centered in the video field and filled about two-thirds of the screen. On each trial the author said either /ba/ or /da/ or nothing, as cued by a video terminal under computer control. The three cues were randomized in blocks of 15 stimuli, with five presentations of each cue in each block of 15.

The original audio track was replaced with 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 msec of the consonant-vowel syllable (CV), a set of five 400-msec CVs covering the range from /ba/ to /da/ was created. During the first 80 msec, F1 went from 300 Hz to 700 Hz following a negatively accelerated path. The F2 followed a negatively accelerated path to

1,199 Hz, beginning with one of five values equally spaced between 1000 and 2000 Hz from most /ba/-like to most /da/-like, respectively. The F3 followed a linear transition to 2,729 Hz from one of five values equally spaced between 2,200 and 3,200 Hz. All other characteristics of the synthetic CVs were identical for the five test stimuli. Additional details of the video recording and the speech synthesis are given in Massaro and Cohen (1983b).

An experimental tape was made by copying the original tape and replacing the original sound track with the synthetic speech. The presentation of the synthetic speech was synchronized with the original audio track on the videotape and gave the strong illusion that the synthetic speech was coming from the mouth of the speaker. To accomplish this synchronization, the audio signal was monitored electronically, and each syllable was replaced with a synthetic speech syllable. Each block of 15 trials contained one replication of the 15 unique conditions created by the factorial combination of three visual articulations times five possible speech sounds.

Procedure.—Subjects were tested in a research van (Mayer, 1982) located outside of the child-care center. The subjects viewed a color 12-inch TV monitor, which presented both the video and audio. The subjects sat about 2–3 feet away from the TV with the loudness level of the speech at a comfortable listening level (70 dB-A). A 250-msec bell preceded each trial. The silent interval between the bell and the onset of speech event varied between 1,175 and 1,375 msec. The subjects had about 6 sec to make their response before the next trial.

Each subject was instructed to watch and listen to the speaker on the TV and to indicate whether he said the sound /ba/ or the sound /da/. The experimenter (the author) observed the children to insure that they were watching the speaker at the time of the speech sound. Less than 4% of the trials did not meet this criterion and were eliminated from the data analysis. The children made their response by hitting one of two buttons. One button was labeled with both the letters "BA" and a drawing of a ball and the other button with "DA" and a drawing of a duck. The experimenter observed the child and recorded which button was hit. One child did not learn the buttons easily and simply responded verbally. Each subject was tested for 90 trials, giving a total of up to six observations for each subject at each of the 15 experimental conditions. The children were tested in sessions of 30 trials with, at most, two sessions per day separated by a break. The adults were left alone and wrote their responses on an answer sheet. The adults were tested for a single session of 90 trials.

Results and Discussion

The proportion of /da/ responses was computed for each subject for each of the 15 conditions. The left and right panels of Figure 1 give the results for the children and adult subjects, respectively. The proportion of /da/ responses as a function of the five levels along the auditory speech continuum is shown for the visual /ba/, /da/, or no-articulation conditions. The average proportion of /da/ responses increased significantly as a function of the level of the auditory stimulus, from .11 for the most /ba/-like to .94 for the most /da/-like, $F(4,80) = 278, p < .001$. There was also a large effect on the proportion of /da/ responses as a function of the visual stimulus, with .45 /da/ responses for /ba/, .62 for no articulation, and .78 for /da/, $F(2,40) = 49, p < .001$. The interaction of these two variables was also significant, $F(8,160) = 17.7, p < .001$, since the effect of the visual variable was smaller at the less ambiguous regions of the auditory dimension.

Tests of the models.—The categorical model, the dependence model, and the independence FLMP were tested against the results of the 15 independent experimental conditions. 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. Initially, all parameters were set to .5. The parameters of each model were adjusted iteratively to minimize the squared deviations between the 15 observed and predicted proportions of /da/ responses for each subject. Thus, the criterion for fitting each model to the results is the average squared deviation, called the root mean squared deviation (RMSD), between the observed and predicted results. Given a model, STEPIT finds a set of parameter values that come closest to predicting the observed data.

Figure 2 gives the average predicted results of the categorical model for the children and adults. As can be seen in the figure, the model gave a poor description of the observed results for both groups of subjects. Figure 3 gives the analogous results for the dependence model. This model also failed to capture the pattern of results. The FLMP provided a much better description, as can be

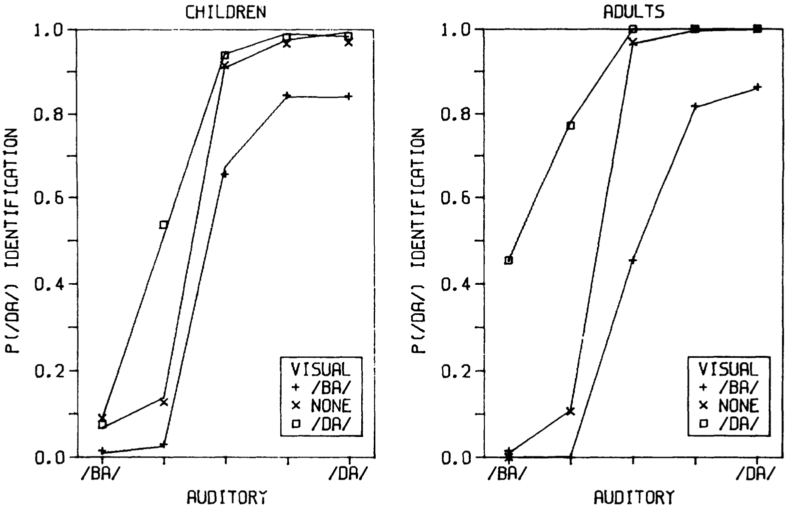

FIG. 1.—Observed (points) and predicted (lines) proportion of /da/ identifications as a function of the auditory and visual levels of the speech event. The predictions are for the fuzzy logical model of perception (FLMP). The left panel gives the results for preschool children and the right panel for adults.

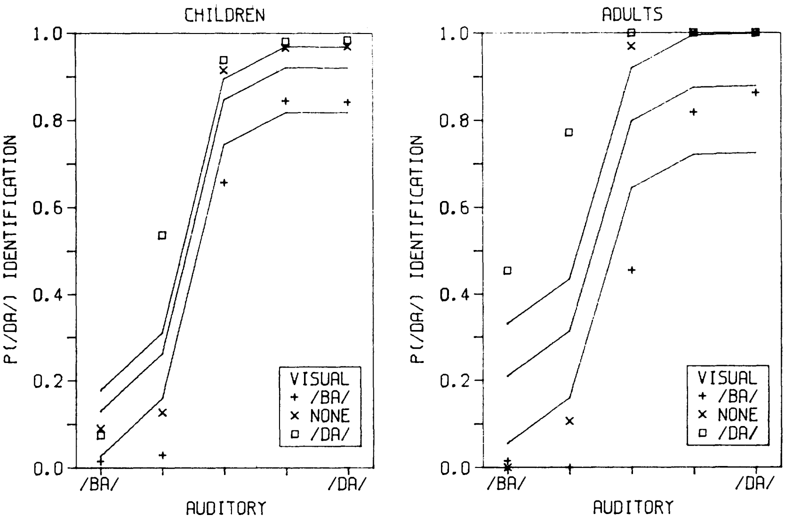

FIG. 2.—Observed (points) and predicted (lines) proportion of /da/ identifications as a function of the auditory and visual levels of the speech event. The predictions are for the categorical model. The left panel gives the results for preschool children and the right panel for adults.

seen in Figure 1. Table 1 gives the average RMSD values for the three models for each group of subjects. Two analyses of variance were carried out on the individual RMSDs contrasting both the categorical and dependence models against the FLMP model. Subject group was treated as an independent variable in each analysis. The FLMP model gave significantly lower RMSDs compared with both the categorical model, $F(1,20) = 207, p < .001$, and the dependence model, $F(1,20) = 141, p < .001$. The effect of subject group and the interaction of subject group with the type of model were also significant in both analyses. Table 1 indicates that the advantage of the FLMP was greater for adults than for children. However, separate analyses of the children's results indicated that the descriptions of both the categorical and dependence models were significantly poorer than that given by the FLMP, $F(1,10) = 30$, and $F(1,10) = 69$, $p < .001$. The somewhat better showing of

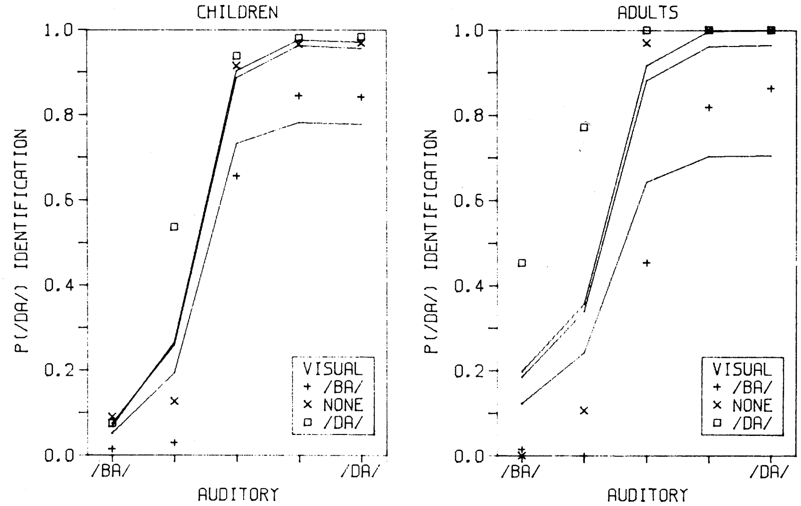

FIG. 3.—Observed (points) and predicted (lines) proportion of /da/ identifications as a function of the auditory and visual levels of the speech event. The predictions are for the dependence model. The left panel gives the results for preschool children and the right panel for adults.

TABLE 1

AVERAGE ROOT MEAN SQUARED DEVIATIONS (RMSD) BETWEEN THE PREDICTED AND OBSERVED RESULTS FOR THE THREE MODELS

Model	Children	Adults
Discrete (9)	.130	.195
Dependence (8)	.123	.206
FLMP (8)	.026	.010

NOTE.—The number in parentheses corresponding to each model represents the number of free parameters estimated to predict the 15 independent observations.

the categorical and independence model for the children might simply be related to the smaller overall influence of the visual source for the children than for the adults. To the extent one source has a small effect, the results provide a less stringest test of any model describing how the two sources are integrated.

These results also allow the rejection of an addition or averaging of the separate sources of information. A weighted adding model gives the following prototypes,

$$/\text{da}/ : wa_i + (1 - w)v_j$$
$$/\text{ba}/ : w(1 - a_i) + (1 - w)(1 - v_j),$$

where w is the weight given the auditory dimension and $(1 - w)$ is the weight given the visual dimension. Given the relative merit rule of the pattern classification operation, the additive model predicts that

$$P(D|A_iV_j) = wa_i + (1 - w)v_j,$$

which is equivalent to the prediction of the categorical model when p is set equal to w. Thus, we can reject the adding model for evaluating and integrating sources of information in speech perception. In contrast to results in other domains (Anderson & Cuneo, 1978), there is no evidence for different integration rules across development in the perceptual categorization of speech.

It should be stressed that the good description of the FLMP cannot simply be because of a large number of free parameters. The categorical and dependence models required as many or more parameters and gave significantly poorer descriptions of the results. Thus, we can reject the categorical and dependence models in favor of the FLMP model for both groups of subjects.

The FLMP model gave equally good descriptions of the individual performance of children and adult subjects. The RMSD between predicted and observed performance averaged .026 for the children and .010 for the adults. The somewhat larger average RMSD for the children was primarily because of the highly variable performance of a single child; thus the group difference in the RMSD values did not approach statistical significance, $F(1,20) = 1.34$, $p = .26$. This result provides evidence consistent with the hypothesis of continuous and independent featural information for both preschool children and adults.

Also of interest in this study are quantitative differences between the children and adult subjects. There was no group main effect and group did not interact with the auditory variable. Figure 4 gives the effect of the visual variable for each subject in the two groups. Larger changes of /da/ identifications across the three visual conditions indicate a larger influence of the visual variable. The children showed only about half of the influence shown by adults, $F(2,40) = 4.6$, $p < .001$. The smaller influence of the visual variable for the children was highly consistent. Eight of the 11 children showed a smaller effect of the visual variable than 10 of the 11 adults. In addition, the attenuation of the visual influence shown by the children interacted with the auditory variable, $F(8,160) = 3.5$, $p < .001$. As shown in Figure 1, the group differences in the effect of the visual source were larger at the /ba/ side of the auditory continuum.

Given the good description of the FLMP, it is reasonable to evaluate the parameter values as indexes of performance in the task. The two age groups and the five levels of the auditory source were independent variables in an analysis of variance on the auditory parameter values. There was a main effect of the auditory level, $F(4,80) = 595$, $p < .001$, but it did not interact with age group, $F(4,80) = 1.26$, $p = .38$. As can be seen in Table 2, the parameter values for the auditory variable were very similar for the two groups. An analogous analysis on the visual parameter values gave both a significant effect of the visual level, $F(2,40) = 161$, $p < .001$, and a significant interaction of visual level with age, $F(2,40) = 3.42$, $p < .05$. The parameter values for the /ba/ and /da/ levels of the visual variable were less extreme for the children than for the adults, reflecting the smaller effect of this variable for the children. When the model was fit to the average results of each of the two groups simultaneously with the restriction that the five auditory parameters be identical for the two groups, the RMSD increased to only .030. This represents a stronger test of the model by predicting 30 independent data points with just 11 free parameters. Accordingly, the model accurately describes the result in the same way for the two groups allowing only for differences in the influence of the visual variable.

Experiment 2

One possible explanation of the smaller visual effect for children is an attentional one. It has been proposed that the influence of a modality in perceptual judgment is related to the attention given to that modality (Welch & Warren, 1980). Children may attend less to the visual than to the auditory source and thus show a smaller influence from the visual source. The attention explanation predicts that the influence of a source will be positively related to the attention given the source. To increase the attentional demands of the visual source, the children were also

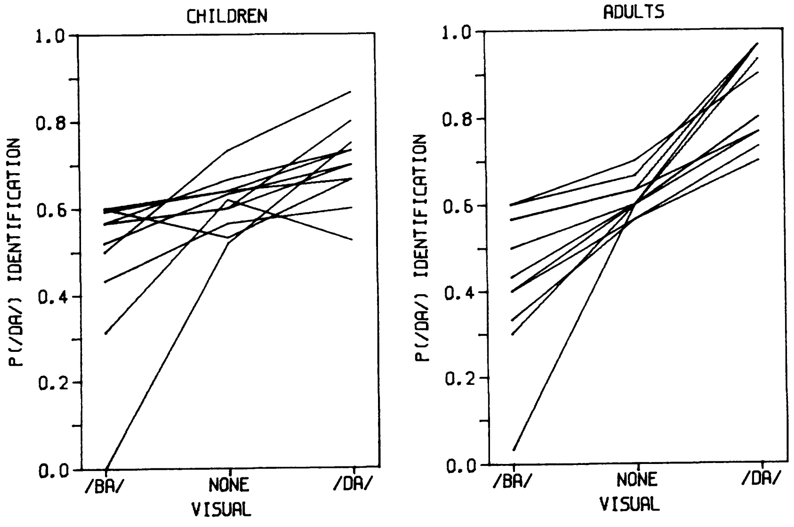

FIG. 4.—The proportion of /da/ identifications for individual subjects as a function of the visual level of the speech event. The left panel gives the results for preschool children and the right panel for adults.

TABLE 2

AVERAGE PARAMETER VALUES FOR THE CHILDREN AND ADULT SUBJECTS FOR THE PREDICTIONS GIVEN BY THE FLMP IN FIGURE 1

A. AUDITORY (a_i)

	/ba/ 1	2	3	4	/da/ 5
Children	.030	.097	.885	.980	.982
Adults	.003	.035	.943	.996	.999

B. VISUAL (v_j)

	/ba/	Neutral	/da/
Children	.128	.605	.921
Adults	.025	.737	.996

required to indicate whether or not the speaker's lips moved during the speech event. We might expect that this requirement would increase the amount of attention given the visual source. To the extent differences in attention are responsible for the smaller effect of the visual variable, the children should show a larger visual influence in this dual task relative to the single identification task in Experiment 1.

Method

Subjects.—Eight of the children in Experiment 1 were tested in Experiment 2. Two children had left the child-care center, and a third chose not to participate.

Procedure.—Experiment 2 included the 90 trials from Experiment 1 plus 30 additional trials using the same task. These 120 trials were followed by 120 additional trials requiring two judgments on each trial: (1) whether the sound was /ba/ or /da/ and (2) whether or not the speaker's mouth moved. After they made their /ba/ or /da/ key press, the children verbally reported whether or not the speaker's mouth moved during the speech sound. This judgment is informative since the speaker's mouth did not move on a third of the trials. All other procedural details were identical to the single judgment task.

Results and Discussion

The accuracy of identifying whether or not the speaker's mouth moved was computed. A percentage correct score was derived for each subject for each of the 15 conditions for the last two trial blocks. Overall performance averaged 86% correct and did not vary with any of the independent variables. Only two children seemed to have any difficulty with the task, averaging 51% and 72% correct. Another child averaged 87% correct, and the remaining five children all performed at 94% correct or better. Thus, children were able to perform this second task, and the experimental question is whether this task influences phonetic identification.

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 was identical in the two conditions. Figure 5 gives the proportion of /da/

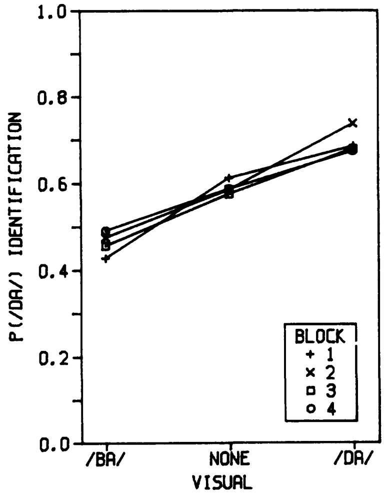

FIG. 5.—The proportion of /da/ identifications for the children as a function of the visual level of the speech event and the four trial blocks. No additional task was required in the first two trial blocks, whereas the task during the last two trial blocks also required the children to identify whether or not the speaker's mouth was moving.

identifications as a function of visual level of the speech event and the four trial blocks. An analysis of variance was carried out as before, but now including the four trial blocks of 60 trials each as factors. No additional task was required in the first two trial blocks, whereas the last two trial blocks required the children also to identify whether or not the speaker's mouth was moving. Four trial blocks rather than two were used to increase the sensitivity of the analysis to practice effects. Trial blocks did not interact with the visual variable or any of the higher-order interactions. Thus, the developmental or age difference resulting from the visual variable in Experiment 1 is probably not attentional in nature. The only significant interaction with trial blocks involved the auditory variable, $F(12,84) = 2.3, p < .025$, which reflected about a 10% increase in the influence of the auditory variable across the four trial blocks. Similar results were obtained when the two subjects who had difficulty with identifying whether the mouth moved were eliminated from the analysis.

General Discussion

This study replicates the finding that children's speech perception shows a smaller influence of visual articulation relative to adults (McGurk & MacDonald, 1976). The present results further show that the difference is unlikely to be an attentional one. The visual influence was not increased when the children were also required to indicate whether or not the speaker's lips were moving. In all other respects, the children behaved identically to adults in the speech identification task. There was absolutely no evidence indicating group differences in the nature of the operations involved in perceptual categorization. More specifically, it was possible to reject the hypotheses that preschool children perceive speech categorically or holistically. Both groups of subjects appear to have continuous and independent featural information from the auditory and visual sources, to integrate the two sources so that the least ambiguous source has the most influence on the judgment, and to classify the speech event on the basis of the relative match of the possible alternatives.

The results stand in sharp contrast to conclusions reached in previous research within other paradigms. Infant speech perception has been interpreted as being categorical (Eimas et al., 1971; Gleitman & Warner, 1982). One can always resolve the apparent contradiction by assuming that speech perception develops from categorical to continuous and becomes continuous before schooling begins. However, the infant research can be interpreted in terms of possible nonlinearities in the discrimination of some speech dimensions rather than in terms of categorical perception (Massaro & Oden, 1980). By nonlinearity is meant that the same amount of stimulus change gives rise to different amounts of perceptual change at different places along the stimulus continuum. Thus, infants might be better at discriminating some changes better than others, but this does not mean that they have only categorical information available. Perception would still be continuous in that small changes along the stimulus dimension produce noticeable changes in perception. Evidence for this view comes from more recent studies that have found within-category discriminations with infant subjects. There is now evidence that infants can make within-category discriminations along nasal-stop (Eimas & Miller, 1980), voiced-voiceless (Aslin, Pisoni, Hennessy, & Perey, 1981) and stop-glide (Miller & Eimas, 1983) continua.

Preschool children are claimed to process certain stimuli holistically (Shepp, 1978; Smith & Kemler, 1977, 1978). To accommodate the present results, one can always claim that preschool children only perceive some stimulus events holistically; other events would be perceived in terms of independent attributes, features, or dimensions. Rather than take this easy tack, however, it is possible to reinterpret the previous studies used to support holistic perception. Consider the perceptual categorization of objects varying in size and brightness (Smith & Kemler, 1978). Older children and adults will tend to group two stimuli together if they have the same size, even if they differ greatly in brightness. Younger children, on the other hand, will group two stimuli together if they differ by relatively small amounts on both dimensions. Rather than accepting these results as implying holistic perception for young children, however, the dimensions of size and brightness might be evaluated independently at all developmental levels. The different results may simply reflect different strategies in the grouping task at the different developmental levels. Consistent with this interpretation, Kemler and Smith (1979) found that young children could treat size and brightness as independent dimensions to learn a higher-order conceptual rule.

It is somewhat surprising that the children showed a smaller visual influence, since just the opposite might have been expected. In contrast to our findings, the research on nonverbal intersensory discrepancy suggests that the influence of vision on judgment in audition does *not* increase with increasing age. Warren and Pick (1970) asked subjects to point with an unseen hand to where the loudspeaker sounded while looking at it through a wedge prism. The prism displaced the visual image of the loud speaker laterally by about 11 degrees. The prism influenced the auditory judgment relative to a nonvisual control condition, but this influence decreased (although nonsignificantly) with increasing age from second to sixth grade to adult. We found significantly less influence of the visual source for children than for adults, and the difference between the verbal and nonverbal task remains to be illuminated.

The current results seem to constrain the possible interpretations of a recent study by Kuhl and Meltzoff (1982). Five-month-old infants were shown to recognize cross-modal correspondences of the vowels /i/ and /a/. The infants viewed a film showing two side-by-side images of a talker articulating /i/ and the same talker articulating /a/, in synchrony, with one of the two vowel sounds. The infants looked longer at the face matching the sound than at the nonmatching face. These infants must have recognized both the auditory and visual speech events and the correspondence between them. The present task assessed the relative contribution of auditory and visual sources of information, whereas infant study assessed only the ability to detect the correspondences between the two sources. The smaller visual effect for children than for adults in the current study suggests that, although children might detect the correspondence between visual and auditory speech, the auditory component probably has the larger influence on the perception of speech categories in language acquisition.

Another relevant finding is that children have been shown to be less sensitive than adults to auditory information in speech perception. For example, Zlatin and Koenigsknecht (1975) found increasing sensitivity to voice onset time differences with age, and Krause (1982) found similar results for vowel duration differences. In both cases, larger acoustic differences were needed for children than for adults to discriminate the phonetic feature of voicing. Given less sensitivity to auditory differences, we might expect an increased influence of the visual articulation. In this study, however, the children appeared to be equally sensitive to the auditory differences, but were less influenced by the visual source. The different speech contrasts tested in these studies are unlikely to be responsible for the different findings. Although these discrepancies remain unresolved, the current research offers a promising approach to the study of developmental patterns in perceptual categorization. Additional research within the current framework will allow a more comprehensive understanding of the development trends in the evaluation and integration of information in speech perception.

References

Anderson, N. H. *Foundations of information integration theory.* New York: Academic Press, 1981.

Anderson, N. H. *Methods of information integration theory.* New York: Academic Press, 1982.

Anderson, N. H., & Cuneo, D. O. The height + width rule in children's judgments of quantity. *Journal of Experimental Psychology: General,* 1978, **107**, 335–378.

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

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

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

Eimas, P. D. Auditory and linguistic processing of cues for place of articulation by infants. *Perception and Psychophysics,* 1974, **16**, 513–521.

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

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. Speech perception in infants. *Science,* 1971, **171**, 303–306.

Gleitman, L. R., & Wanner, E. Language acquisition: The state of the state of the art. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art.* Cambridge: Cambridge University Press, 1982.

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

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

Krause, S. E. Vowel duration as a perceptual cue to postvocalic consonant voicing in young children and adults. *Journal of the Acoustical Society of America,* 1982, **71**, 990–995.

Kuhl, P. K., & Meltzoff, A. N. The bimodal perception of speech in infancy. *Science*, 1982, **218**, 1138–1141.

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

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

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

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

Massaro, D. W., & Cohen, M. M. Phonological context in speech perception. *Perception and Psychophysics*, 1983, **34**, 338–348. (c)

Massaro, D. W., & Oden, G. C. Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and Language: Advances in Basic Research and Practice.* New York: Academic Press, 1980.

Mayer, M. J. A mobile research laboratory. *Behavior Research Methods and Instrumentation*, 1982, **14**, 505–510.

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

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

Oden, G. C., & Massaro, D. W. Integration of featural information in speech perception. *Psychological Review*, 1978, **85**, 172–191.

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

Samuel, A. G. The effect of discrimination training on speech perception: Non-categorical perception. *Perception and Psychophysics*, 1977, **22**, 321–330.

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

Shepp, B. E., Burns, B., & McDonough, D. The relation of stimulus structure to perceptual and cognitive development: Further tests of a separability hypothesis. In F. Wilkening, J. Becker, & T. Trabasso (Eds.), *Information integration by children.* Hillsdale, N.J.: Erlbaum, 1980.

Smith, L. B., & Kemler, D. G. Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology*, 1977, **24**, 279–298.

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

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

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

Warren, D. H., & Pick, H. L. Intermodality relations in blind and sighted people. *Perception and Psychophysics*, 1970, **8**, 430–432.

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

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

Zlatin, M. A., & Koeningsknecht, R. A. Development of the voicing contrast: Perception of stop consonants. *Journal of Speech and Hearing Research*, 1975, **18**, 541–553.