Learning to use an artificial visual cue in speech identification[a)]

Joseph D. W. Stephens[b)] and Lori L. Holt
Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

(Received 5 January 2010; revised 20 May 2010; accepted 22 July 2010)

Visual information from a speaker's face profoundly influences auditory perception of speech. However, relatively little is known about the extent to which visual influences may depend on experience, and extent to which new sources of visual speech information can be incorporated in speech perception. In the current study, participants were trained on completely novel visual cues for phonetic categories. Participants learned to accurately identify phonetic categories based on novel visual cues. These newly-learned visual cues influenced identification responses to auditory speech stimuli, but not to the same extent as visual cues from a speaker's face. The novel methods and results of the current study raise theoretical questions about the nature of information integration in speech perception, and open up possibilities for further research on learning in multimodal perception, which may have applications in improving speech comprehension among the hearing-impaired. © *2010 Acoustical Society of America.* [DOI: 10.1121/1.3479537]

PACS number(s): 43.71.Rt, 43.71.An, 43.71.Es, 43.71.Ft [MSS] Pages: 2138–2149

I. INTRODUCTION

A basic property of speech perception is its dependence upon multiple sources of information. Perceptual interpretation of a speech segment is driven by multiple bottom-up sources, including auditory and visual information (e.g., McGurk and MacDonald, 1976), as well-as higher order information such as the probable identity of the current word (e.g., Ganong, 1980). The use of multiple information sources in speech perception also has important consequences in situations where acoustic information is degraded. For example, speech comprehension in noisy conditions is dramatically improved when listeners are allowed to view the speaker's face (Sumby and Pollack, 1954; Grant and Seitz, 2000). Visual speech cues also improve comprehension in hearing-impaired individuals and cochlear implant users (Lachs *et al.*, 2001; Massaro and Cohen, 1999; Tyler *et al.*, 1995) and the elderly (Walden *et al.*, 1993).

Relatively little is known about the extent to which visual influences in speech perception depend on learning. Computational models of information integration in speech perception (e.g., FLMP: Oden and Massaro, 1978; TRACE: McClelland and Elman, 1986; and Merge: Norris *et al.*, 2000) posit associative links between various information sources and speech categories, which can presumably be learned through experience (as in the case of lexical categories). These models share the common assumption that multiple sources are not merely used jointly but rather are *integrated* into a unified percept. They also assume that the mechanism for information integration is *not* affected by experience, in that the combination of information sources results from basic architectural features of each model that remain constant. However, learning mechanisms have been incorporated into these models to account for experience-based "tuning" of associations between existing information sources (visual, lexical) and phonetic categories (Massaro *et al.*, 1993; Mirman *et al.*, 2006; Norris *et al.*, 2003). Such learning mechanisms might also be expected to enable the incorporation of a completely novel information source in speech perception (e.g., Massaro and Chen, 2008), although this prediction has not been extensively studied.

Very few empirical data are available to document the formation of new links between auditory and visual information in speech perception; in part, this is due to the intractability of decoupling auditory and visual speech information in the experience of young infants. Studies with adult listeners have demonstrated "recalibration" of phonetic representations based on relatively brief audio-visual experience (Bertelson *et al.*, 2003; Samuel and Kraljic, 2009). Further, some evidence suggests that orthography can influence speech perception (Massaro, 1999; van Atteveldt *et al.*, 2004), which implies that new sources of visual information can be learned. However, other evidence suggests that the use of novel sensory information in speech perception may not depend on learning (Fowler and Dekle, 1991; but see Massaro and Chen, 2008).

It remains an open question whether adults can learn to use completely novel visual cues in a manner similar to natural visual speech. Bernstein *et al.* (2004) found that simple detection of auditory speech in noise was improved by the presence of novel artificial visual stimuli (with no need for learning), although the improvement was not as great as with natural visual speech. Massaro and colleagues (Massaro, 1998, ch. 14; see also Massaro *et al.*, 2009) have also found that perceivers can learn to use artificial visual cues generated from speech acoustics as a supplement to simultaneous, natural visual speech. The current study went further than

[a)]Portions of this work were presented at the 46th meeting of the Psychonomic Society and the 151st meeting of the Acoustical Society of America.

[b)]Author to whom correspondence should be addressed. Present address: Department of Psychology, North Carolina Agricultural and Technical State University. Electronic mail: jdstephe@ncat.edu

previous research by directly testing whether listeners can learn to use an arbitrary new source of visual information in speech perception. A set of dynamic visual cues for speech categories was created and a training paradigm was devised to provide participants with the greatest possible opportunity to learn the novel visual cues and associate them with corresponding speech categories. The influence of these novel visual cues on speech perception was then tested and compared to the influence of natural visual cues from a speaker's face. If current computational models of speech perception are correct, then to the extent that listeners can learn to interpret such visual cues, they should combine them with auditory speech in a manner similar to natural visual cues.

II. METHOD

A. General method

Adult participants were exposed to artificial, temporally-dynamic, visual stimuli synchronized with auditory vowel-consonant-vowel (VCV) utterances in the context of a video game. The artificial visual stimuli were computer-animated videos of a "speech robot" with moving parts whose positions specified phonetic categories (Fig. 1). Across multiple sessions, participants were trained to identify voiced consonants (/b/, /d/, /g/) based on these visual cues.

Specific details of experimental stimuli and procedures are provided below. The overall form of the experiment was as follows. The experiment began with a unimodal, *visual identification* pre-test in which participants attempted to judge the consonants associated with the artificial visual stimuli. Participants then completed several daily sessions of *audiovisual training* in which they were exposed to artificial visual signals synchronized with corresponding auditory speech (the exact number of training sessions depended on each individual's performance). Finally, participants completed four post-tests: unimodal *visual identification* (identical to pre-test); *factorial audiovisual identification*; and two tests of *audiovisual mismatch identification in noise* using artificial visual stimuli and natural visual speech cues (i.e., a human face), respectively.

B. Participants

Twelve adult English speakers with native-language competence, no reported hearing impairment, and normal or corrected-to-normal eyesight, participated in the experiment (six female; age range=23–33 years, mean=26.5 years). Participants gave informed consent prior to the experiment and procedures (including experiment length) were approved by Carnegie Mellon University's Institutional Review Board. After the experiment, participants were debriefed regarding the nature of the experiment.

Participants were compensated for their time via payments made at regular intervals during the experiment. Installment payments included $6 for every session that had been completed. In addition, $1.50 for each session was paid to participants as a bonus upon completion of the entire experiment, and an extra bonus was awarded based on performance during training. Overall pay for the experiment across the 12 participants ranged from $105 to $192 and the median pay was $117.

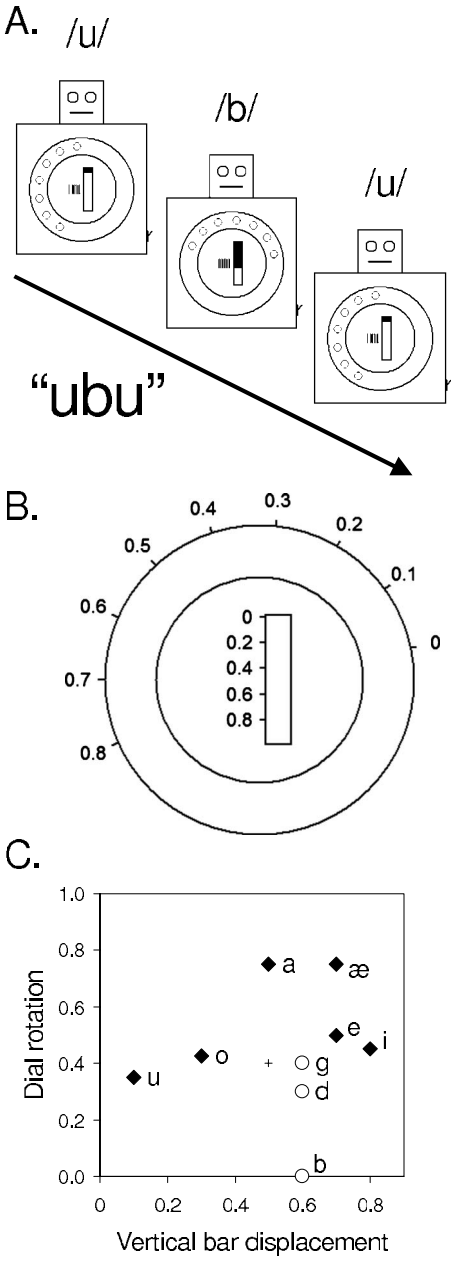

FIG. 1. (a) Still-frame images of the speech robot. The robot had two moving parts that transitioned smoothly between positions specifying phonetic categories. (b) The positions of the moving parts were defined according to parameter values that corresponded to locations on the robot display. (c) The parameter values used in the experiment. The consonants differed only in the rotational component.

C. Stimuli[1]

1. Distributions

Stimulus materials were constructed to represent 12 different VCV utterances based on the combinations of three consonants (/b/, /d/, /g/) and four vowels (/i/, /æ/, /a/, /u/). These utterances were represented by overlapping distributions of cues in a two-dimensional audiovisual space. An example of this two-dimensional stimulus space is displayed

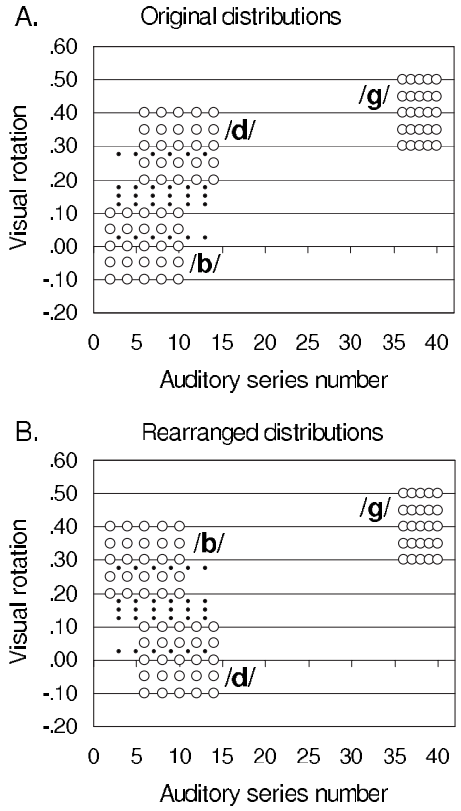

FIG. 2. Depiction of audiovisual category distributions for the /aba/, /ada/, and /aga/ stimuli used in the task. The vertical axis represents the parameter setting of the robot's rotating component for each stimulus. The horizontal axis represents auditory stimuli along a 40-member morphed series from /b/ to /d/ to /g/. Open symbols are stimulus combinations used during training. Black dots represent the combinations used in the factorial audiovisual identification task. (a) Distributions reflecting an arrangement of visual and auditory cues based on the parameters depicted in Fig. 1. (b) Rearranged audiovisual category distributions used for half of the participants. The visual cues (vertical axis) were reversed for the /b/ and /d/ categories.

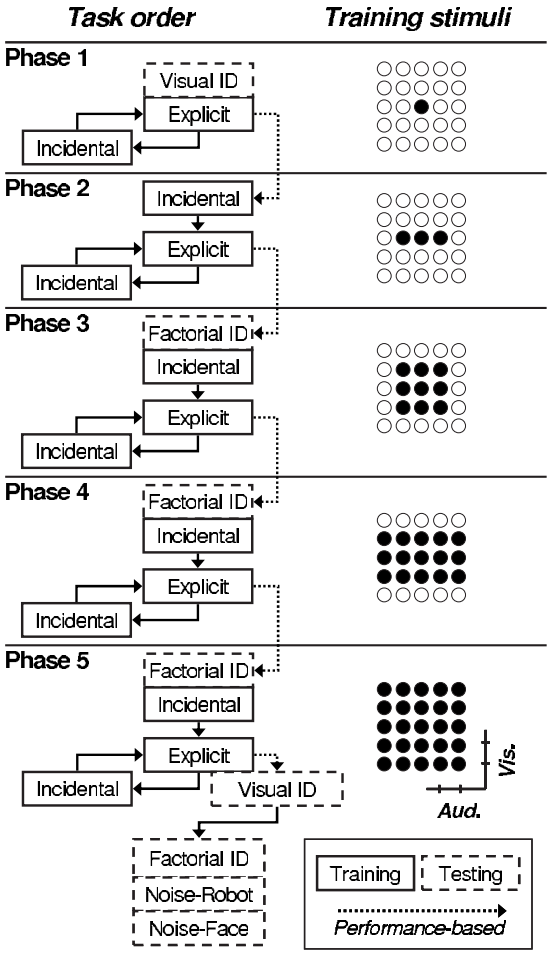

FIG. 3. Schematic representation of task sequence in the experiment and the expansion of the training set across phases (see text for details). Adjoining boxes represent tasks that were performed within a single session. For the stimuli, each of the VCV utterances used in training (/aba/, /ugu/, etc.) could be represented by any of 25 audiovisual combinations. In the first training phase, only the central combination within each of these distributions was used; the training set was gradually broadened until all 25 combinations were used for each utterance.

in Fig. 2 for stimuli in the /a/ vowel context. The three categories are only separable when both the auditory and visual dimensions are taken into account; thus, accurate categorization depended on the use of information from both modalities. This arrangement mirrors the category-conditional independence of audiovisual speech in the natural environment (Massaro, 1998, ch. 4; Movellan and McClelland, 2001).

The two-dimensional stimulus space was manipulated between participants by reversing the arrangement of animated visual cues corresponding to the /b/ and /d/ categories for half of the participants. The rearrangement of audiovisual categories for stimuli in the /a/ context is also shown in Fig. 2. This made it possible to test whether the effects of visual information depended on the relationship of each category to the other categories. For example, in the classic McGurk effect (McGurk and MacDonald, 1976), the combination of auditory /b/ and visual /g/ is interpreted as /d/, which may reflect a perceptual compromise based on overlapping auditory features for /b/ and /d/ and overlapping visual features for /d/ and /g/. In the current experiment, the "original" distributions represent this type of structure in the environment, whereas the "rearranged" distributions represent an alternate structure in which /b/ and /g/ have overlapping visual characteristics. This might be expected to result in a novel form of the McGurk effect in which combinations of auditory /d/ and visual /g/ are interpreted as /b/.

A further aspect of the audiovisual distributions used in the present study was that they were gradually broadened over the course of training. Training was thus divided into five phases: in the initial phase, participants were trained only on the audiovisual combinations at the center of each category distribution; in the final phase, the entire distributions were used for training. In this way, audiovisual identification became progressively more difficult as participants became more skilled at interpreting the visual stimuli. The expansion of the distributions across training phases is illustrated in Fig. 3.

2. Auditory stimuli

In order to create distributions of audiovisual combinations with variability in the auditory dimension, auditory stimuli were drawn from a set of "morphed" natural speech tokens.

a. Auditory stimulus creation. Morphed stimuli were created by adjusting LPC filter coefficients (Atal and Hanauer, 1971) to create a range of consonants that varied from /b/ to /d/ to /g/ in each of the four vowel contexts. An adult, Midwestern American male speaker (JDWS) produced three repetitions of each of the 12 VCV combinations used in the stimulus set. The tokens were recorded digitally on a personal computer using Computer Speech Laboratory (CSL; Kay Elemetrics Corp., Lincoln Park, NJ) with 16-bit precision at a sampling rate of 11.025 kHz. The tokens were isolated and saved separately as monaural PCM .wav files, and matched in RMS power prior to further processing.

Within each vowel context, the tokens for each consonant that were most compatible in pitch and temporal properties (i.e., speaking rate, burst length, and duration) were selected as series endpoints and edited to produce further temporal alignment (i.e., by deleting or duplicating pitch periods, etc.). An LPC analysis was performed on each of these edited natural endpoint tokens using the autocorrelation algorithm (Markel and Gray, 1976) implemented in the computer program Praat (version 4.3.19; Boersma, 2001). The /d/ tokens from each vowel context were inverse-filtered by their LPC coefficients to extract approximate voicing sources for each /d/ endpoint token. The resulting four source waves (one for each vowel context) were saved and used in the subsequent resynthesis of all stimuli within a corresponding vowel series.

To create series ranging perceptually between endpoint consonants, new LPC filters were created by incrementally adjusting coefficients in 20 equal steps between each endpoint (/b/ to /d/ and /d/ to /g/). After each set of LPC filters was created, the filters were applied to the source wave derived from the /d/ token with the corresponding vowel, so that all members of each VCV series were based on the same voicing source. Subsequent to this resynthesis all 160 VCV stimuli were RMS-matched. A more detailed description of these morphing procedures and stimulus characteristics is provided in Stephens (2006, Chapter 6).[2]

All auditory stimuli used in the experiment were given a slight echo (a 30-ms delay), which made them sound more stereotypically "robotic" without altering frequency characteristics or adversely affecting intelligibility.

b. Auditory stimulus selection. In order to create overlapping audiovisual distributions, stimuli were selected from the morphed series so that the /b/ and /d/ tokens lay near the category boundary. Appropriate stimuli were identified via a pilot study in which 27 participants (native, monolingual speakers of English with no hearing impairment) gave identification responses ("B," "D," and "G") to all 160 morphed VCV stimuli. From the VCV series containing /a/, additional tokens spanning the category boundary between /aba/ and /ada/ were also selected for use in the *factorial audiovisual identification* task (also displayed in Fig. 2). One reliable token each of /aba/, /ada/, and /aga/ was also selected for use in the *audiovisual mismatch identification in noise* task (in the case of /aga/, not enough reliable /aga/ tokens were found in the pilot study to select a novel /aga/ stimulus for the audiovisual mismatch identification task, so an /aga/ token was selected that was infrequently presented during training).

3. Artificial visual stimuli

Videos of the speech robot were created with temporal characteristics that corresponded to those of the auditory speech tokens. The movements of the robot were defined by the parameters depicted in Fig. 1, with /b/, /d/, and /g/ differing in the position of the rotating component at consonant onset. This parameter space was intended to be an arbitrary re-mapping of the structure of natural visual speech categories (cf. Montgomery and Jackson, 1983). The parameters for the consonants depicted in Fig. 1 represent the visual cues at center of the audiovisual distributions. Variations in the visual stimuli within each category were created by adjusting the parameter for the rotating component for each consonant, as depicted in Fig. 1. As described above, the visual stimuli for /b/ and /d/ categories were inverted for half of the participants.

When animated, the robot's moving parts transitioned linearly between positions defining the vowel and consonant in each utterance. Transitions were timed according to the lengths of the initial and final vowels and the timing of consonant bursts in the corresponding acoustic tokens. The visual transition from consonant to vowel began approximately 67 ms before the consonant burst, to take advantage of the finding that slight audio lags facilitate audiovisual integration (Munhall *et al.*, 1996). Each visual stimulus was saved as a digital video file (.avi format) with a frame rate of 30 fps.

This procedure was used to create visual stimuli for the various tasks in the experiment. For *audiovisual training*, 48 videos were created to correspond to the visual dimensions of the bimodal distributions within each vowel context. For the *visual identification* task, the 12 videos representing the center of each consonant category in each vowel context were used as well as six additional videos that were created in order to test generalization. These stimuli represented robot movements for each consonant in two novel vowel contexts (based on parameters intended to correspond to /e/ and /o/ in the visual stimulus space of Fig. 1). For *factorial audiovisual identification*, five additional visual stimuli were created with a special focus on the ambiguous region between /b/ and /d/ (also depicted in Fig. 2). For the *audiovisual mismatch identification in noise* task, videos representing the center of each category in the /a/ context were used.

4. Natural visual stimuli

Natural visual speech stimuli were used in the final task of the experiment, a repetition of *audiovisual mismatch identification in noise*. Stimuli were created by videotaping the lower half of the face of a speaker (JDWS) producing /aba/, /ada/, and /aga/ in a normal manner. The digital videos were given the same temporal characteristics as the artificial visual stimuli by selectively deleting or duplicating individual frames. The videos were limited to the lower half of the face in order to provide a closer parallel to the artificial stimuli, in which the moving parts occupied the majority of the visible surface of the robot. Video size was also scaled such that the

mouth occupied approximately the same area of the screen as the moving parts of the speech robot.

D. Procedure

1. General procedure

The experiment was carried out across daily 40–60 min experimental sessions. Nine participants carried out daily sessions on laptop computers borrowed from the laboratory (Gateway Computer Corp., Irvine, CA), and three participants completed the daily sessions on a desktop computer in the laboratory. All portions of the experiment were executed using Presentation software (Neurobehavioral Systems, Inc., Albany, CA), and tasks were designed so that participants were automatically guided through the appropriate sequence of sessions. This was achieved through the use of log files that were created after each session and set the relevant parameters (e.g., experiment phase, current task) that were read by the program the next time it was launched. Participants were also provided with Beyer DT-150 headphones (Beyerdynamic GmbH, Heilbronn, Germany) and instructed to use them when performing sessions. Volume levels were set to provide stimuli at 65–70 dB.

A schematic representation of experimental tasks is shown in Fig. 3. After a visual identification pre-test, audiovisual training progressed through five phases, in which stimulus distributions were gradually broadened. Daily training sessions within each phase alternated between explicit and incidental training tasks. The progression of training phases depended on the attainment of performance criteria by each participant. Specifically, when a participant reached a signal-to-noise ratio of −10 dB at any point in the explicit training task, the next training phase was initiated in the subsequent session.[3] Thus, the number of sessions varied across participants (10–22 sessions; the median number of sessions was 12). The rationale for this design was to obtain a similar level of expertise in identifying artificial visual stimuli for all participants by the end of training (a similar pilot experiment that used a standard number of sessions for all participants found substantial variability across participants in the ease with which they learned the artificial visual cues).

Factorial audiovisual identification tasks were performed at the beginning of the third, fourth, and fifth phases, to examine the use of the artificial visual cues in bimodal speech identification over the course of training. After the performance criterion was reached in the fifth phase, a visual identification post-test was immediately performed. The next and final session of the experiment then consisted of (in this order): a factorial audiovisual identification task, an audiovisual mismatch identification in noise task with animated-robot visual stimuli, and a second audiovisual mismatch task with visual stimuli of a human face.

2. Audiovisual training tasks

a. "Explicit" training task. The purpose of the explicit task was to provide direct instruction to participants on how to interpret the artificial visual stimuli. During the task, participants were instructed to watch and listen to the robot and indicate which consonant the robot spoke on each trial, using the laptop's J, K, and L keys (re-labeled "B," "D," and "G"). The task was presented as a video game in which learning about the robot's movements would enable participants to perform well (as indicated by displays of point score and difficulty level and occasional sound effects). The Space bar could be used to replay an audiovisual stimulus combination. After correct responses, the word "Correct!" was displayed in green text; after incorrect responses, the correct response was displayed in red text. Each incorrect trial was immediately repeated once to provide an opportunity for correct identification.

Background noise was added to auditory stimuli so that participants were encouraged to use visual information. Noise amplitude varied according to performance. Thus, the signal-to-noise ratio was set to +10 dB at the beginning of each explicit training session, and then adjusted by −1 dB after any six consecutive correct responses and by +1 dB after any two consecutive incorrect responses. Noise samples were randomly selected at the time of presentation from a 60 s recording of multi-speaker babble with overall RMS power matched to the auditory stimuli.

Each training session consisted of 30 trial blocks in which one audiovisual combination was presented from each of the 12 VCV distributions for the current phase of training, for a total of 360 base trials (as described above, repetition of some trials occurred based on participants' responses). The audiovisual combinations used in each block were selected randomly from the current stimulus distributions. The ordering of base trials was random within each block.

b. "Incidental learning" task. The purpose of the incidental task was to give participants additional exposure to audiovisual stimulus combinations in the absence of noise and without the requirement of overt identification responses. In this task, participants were asked to watch and listen to the speech robot and simply indicate whether it produced a malfunction. The malfunctions were relatively infrequent events in which an anomalous auditory or visual stimulus was presented. Five auditory and five visual stimuli were created to represent the malfunctions: the auditory stimuli consisted of edited samples of white noise and tones, and the visual stimuli were animated videos of the robot in which the robot's parts moved erratically.

On each trial, a participant indicated whether the robot had produced normal output, an auditory malfunction, or a visual malfunction, using keys labeled "Normal," "Auditory," and "Visual" on the keyboard (the A, S, and D keys). There were 30 trial blocks, each of which consisted of 12 normal audiovisual combinations which were randomly selected from the VCV distributions for the current phase of the experiment, and two randomly selected malfunctions, one containing anomalous auditory information and one containing anomalous visual information. Each type of auditory and visual malfunction occurred with equal frequency (six times) over the course of the task. The ordering of trials within each trial block was random.

3. Visual identification (pre-test and post-test)

The unimodal visual identification task tested participants' labeling of consonants based solely on the robot's movements, with no accompanying acoustic speech stimulus. Twelve visual stimuli (those containing /i/, /ae/, /a/, and /u/) were part of the stimulus set used in audiovisual training. Six additional videos (containing /e/ and /o/) were included to test whether participants could generalize their knowledge of visual consonants to vowel contexts that had not been trained. There were 12 blocks in which the 18 unimodal visual stimuli were presented in random order. After each presentation, the participant used the keyboard to identify which consonant had been produced by the robot.

4. Factorial audiovisual identification

The factorial audiovisual identification task tested for effects of visual information on identification of auditory consonants that ranged perceptually from /b/ to /d/. Participants' identification responses were recorded for unimodal presentations of stimuli from auditory and visual series and for bimodal combinations of auditory and visual stimuli (i.e., 6 auditory stimuli, 5 visual stimuli, 30 combinations). On unimodal auditory trials, the robot remained on the screen in its neutral configuration. The task consisted of 10 trial blocks, in which each of these 41 stimulus combinations was presented in random order. After each presentation, the participant used the keyboard to indicate whether the robot had produced /b/ or /d/. Participants were instructed to do their best to identify which consonant was *heard* on each trial, except for unimodal visual trials. All repetitions of the task between phases of training were identical to each other, and the task was identical for participants trained on the original and rearranged audiovisual distributions (responses were merely coded differently depending on which visual stimulus was trained as /b/ and which was trained as /d/).

5. Audiovisual mismatch identification in noise

The audiovisual mismatch identification task evaluated the extent to which newly-learned visual cues affected speech perception in noise. Auditory and visual cues for /aba/, /ada/, and /aga/ were presented in combinations that were either consistent or inconsistent with participants' audiovisual training. This task was given twice, first with the animated robot and second with visual stimuli of a speaker's face. Identification responses were recorded for each of the nine possible combinations of auditory and visual stimuli and for each of the three auditory stimuli alone, at three signal-to-noise ratios: +20 dB (noise inaudible), −4 dB, and −8 dB.[4] As in the audiovisual training task, a noise segment of appropriate length was randomly sampled from a 60 s multi-speaker babble recording on each trial. The task consisted of 10 trial blocks, in each of which the 36 different combinations (12 auditory/audiovisual combinations ×3 noiselevels) were presented in random order. Participants were instructed to do their best to identify which consonant was *heard* on each trial. The task was identical for participants trained on the original and rearranged audiovisual distributions (again, data were coded such that visual stimuli were classified as congruent or incongruent depending on which stimuli were trained as /b/ and /d/).

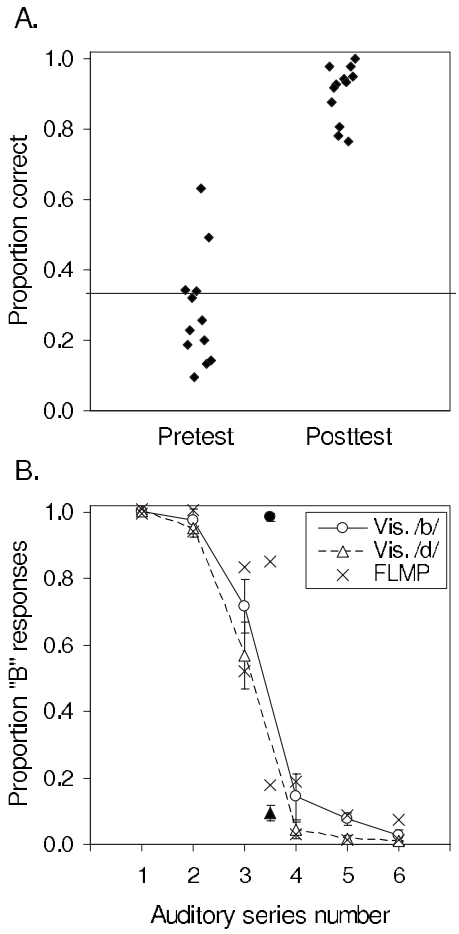

FIG. 4. (a) Learning of novel visual speech cues as indicated by performance of individual participants in the visual identification task. The horizontal line represents chance performance. (b) Mean "B" responses to combinations of auditory stimuli with artificial visual /b/ and /d/ in the final factorial audiovisual identification task (open symbols). Also plotted are "B" responses to unimodal visual /b/ and /d/ (closed symbols) and FLMP predictions for the data (Xs). FLMP predictions are based on model fits to the entire data set (see text), which includes conditions not represented in the figure. Error bars represent standard error of the mean.

III. RESULTS

A. Visual identification

Figure 4(a) displays proportion correct visual identification responses for each of the 12 participants, at pre-test and post-test. All participants exhibited good visual identification performance at post-test. Improvement in proportion correct unimodal visual identification between pre-test and post-test occurred across all three consonants, for both arrangements of audiovisual categories. A 2(pre-test vs. post-test) ×3(visual consonant)×2(category arrangement) repeated-measures ANOVA found a highly significant effect of test, $F(1,10)=219.8$, $p<0.001$, $\eta_p^2=0.96$, and no reliable effect of consonant $F(2,20)=3.141$, $p=0.065$, nor of category arrangement, $F(1,10)=0.106$, $p=0.75$. No test by consonant interaction was found, nor was there a consonant by category arrangement interaction (both $F<1$). However, there was a three-way interaction of test, consonant, and category ar-

rangement, $F(2,20)=5.37$, $p=0.014$, $\eta_p^2=0.349$. Inspection of the data suggested that this interaction was due to different patterns of accuracy across consonants at post-test for the two groups of participants trained on different category distributions. This was to be expected given that the most distinctive visual stimulus represented different consonants in the two category distributions (i.e., /b/ in the original distributions and /d/ in the rearranged distributions).

B. Factorial audiovisual identification

1. Visual influence

In the interest of brevity, data are only presented from the final factorial audiovisual identification task given after the completion of training. Figure 4(b) displays average proportion "B" responses across the auditory series for combinations containing end point visual stimuli. Data in the figure are combined from both groups of participants trained on original and rearranged distributions (note that the particular stimuli used to represent visual /b/ and /d/ were reversed for the two groups). A 6(auditory stimulus) $\times$2(visual stimulus) $\times$2(original vs. rearranged distributions) repeated-measures ANOVA on these data revealed significant effects of auditory stimulus, $F(5,50)=144.7$, $p<0.001$, $\eta_p^2=0.94$, and visual stimulus, $F(1,10)=22.7$, $p=0.001$, $\eta_p^2=0.69$. The auditory $\times$ visual interaction was also significant, $F(5,50)=3.77$, $p=0.006$, $\eta_p^2=0.27$. Thus, the visual stimuli influenced consonant identification when paired with auditory stimuli ranging from /b/ to /d/. The interaction reflects the tendency for this influence to be greater in the middle of the auditory series. The effect of original vs. rearranged distributions was not significant, $F(1,10)=3.52$, $p=0.09$, $\eta_p^2=0.26$; however, there was a trend toward fewer "B" responses overall in the group trained on rearranged distributions. This trend was greatest in the middle of the auditory series, and was reflected in a significant interaction of auditory stimulus $\times$ distribution arrangement, $F(5,50)=2.68$, $p=0.032$, $\eta_p^2=0.211$. More importantly, however, the interaction of distribution arrangement with visual stimulus was not significant, and neither was the three-way interaction (both $F<1$). Thus, the degree to which newly-learned visual stimuli influenced identification of /b/ and /d/ did not differ depending upon which visual endpoint stimulus was trained as /b/ and which was trained as /d/. Figure 4(b) also displays responses given to the endpoint visual stimuli when presented unimodally. It can be seen from the figure that the differences in "B" responses across unimodal visual conditions were considerably greater than the differences caused by visual stimuli in bimodal conditions, even when auditory information was ambiguous. This suggests that participants were not making optimal use of the visual cues to resolve auditory ambiguity (cf., Massaro, 1998, ch. 4).

2. Model comparisons

The question of how participants combined the two information sources in this task was addressed formally by comparing the fits of three models to the full data sets from the four factorial audiovisual identification tasks. The procedure used here followed that of Massaro (1998, ch. 2), by comparing the Fuzzy Logical Model of Perception (FLMP) to a Single Channel Model (SCM).

TABLE I. RMSD for model fits to data from the four factorial audiovisual identification tasks.

	FLMP		SCM	
Task	*M*	*SD*	*M*	*SD*
1	0.068	0.013	0.062	0.015
2	0.067	0.015	0.064	0.014
3	0.073	0.025	0.058	0.018
4	0.061	0.012	0.052	0.015

a. FLMP. The FLMP assumes that information sources are optimally integrated in perception. In the case of a two-alternative task such as the one used here, the form of the FLMP is equivalent to Bayes' rule (Massaro, 1998, ch. 4). For example, the proportion of "B" responses in the factorial audiovisual identification task would be represented by:

$$P(\text{“}B\text{”}\,|A_i,V_j)=\frac{a_i v_j}{a_i v_j+(1-a_i)(1-v_j)}, \tag{1}$$

where A_i and V_j represent the auditory and visual sources of information, and a_i and v_j are "fuzzy truth values," which are parameters representing the degree of support for the /b/ category from the auditory and visual sources, respectively. When one source is absent, it is represented by a fuzzy truth value of 0.5. In the present model-fitting exercise, 11 parameters were estimated for each expanded-factorial task, representing the degree of support for the /b/ category from each of the six auditory and five visual stimuli used in the task.

b. SCM. The SCM assumes that on any given trial, the perceiver only uses information from one source. In the current task the SCM predicts "B" responses thus:

$$P(\text{“}B\text{”}\,|A_i,V_j)=p_a a_i+(1-p_a)v_j, \tag{2}$$

where a_i and v_j represent the probabilities of choosing a "B" response based on auditory and visual sources, and p_a represents the probability of responding based on the auditory source in any given trial. For this analysis, 12 parameters were estimated: one for each of the auditory and visual stimuli, and one parameter representing the probability of an auditory response.

c. Model fits. Each model was fit to the data by initially setting all parameters to random values between 0 and 1 and iteratively adjusting each parameter so as to minimize the root mean squared deviation (RMSD) between the predicted and observed results. To avoid the possibility of settling into local minima, the fitting procedure was carried out 20 times for each model with different initial parameters. Average RMSD values for each model were computed across the 12 participants and are presented in Table I. Although the models performed similarly, the SCM provided better fits to the data than the FLMP. A 2(model) $\times$4(task number) repeated-measures ANOVA comparing the SCM to the FLMP found a significant main effect of model on RMSD, $F(1,11)=10.5$, $p=0.008$, $\eta_p^2=0.49$. The effect of task

number was not significant, $F(3,33)=1.63$, $p=0.20$, $\eta_p^2=0.13$, nor was the model × task number interaction, $F(3,33)=1.15$, $p=0.35$, $\eta_p^2=0.09$.

These results are in stark contrast to typical findings for natural audiovisual speech, for which the FLMP performs much better than the SCM (e.g., Massaro, 1998, ch. 2). Some of the FLMP predictions are displayed alongside observed data in Fig. 4(b), which illustrates that the FLMP has difficulty reconciling the small size of the visual influence on audiovisual identification with participants' excellent ability to identify visual stimuli that are presented unimodally. Recall that the FLMP represents completely ambiguous information in the same way as if that information were absent (i.e., with fuzzy truth values of 0.5). Thus when auditory information is ambiguous, the FLMP predicts that response patterns to audiovisual stimuli should approach the patterns observed for unimodal visual stimuli. The model fits confirm that the current data do not conform to this prediction. It should also be noted that the SCM did not perform much better than the FLMP. In particular, the SCM is not capable of predicting an auditory × visual interaction like the one observed in the data of Fig. 4(b). Thus, it is likely that participants used the two sources of information in a way that somewhat reflected the ambiguity of each source. However, there is no evidence that participants optimally combined the two sources.

C. Audiovisual mismatch identification in noise

1. Intelligibility

Figure 5 displays average proportion of responses corresponding to the auditory stimulus for congruent and incongruent audiovisual combinations and for unimodal auditory stimuli, at each noise level and in each version of the task (artificial vs. natural visual speech stimuli). The data in the figure are divided between the two groups that were trained on different distributions. It can be seen from the figure that, under noisy conditions, visual stimuli had a substantial influence on participants' identification of auditory consonants. These intelligibility effects of natural versus artificial visual cues were compared in separate 2(stimulus type; robot vs. face) × 3(visual condition) repeated-measures ANOVAs for each group of participants (original and rearranged training distributions) on the data from the high noise condition (−8 dB). For participants trained on the original audiovisual category distributions, there was an effect of visual condition, $F(2,10)=24.2$, $p<0.001$, $\eta_p^2=0.83$. There was no significant main effect of stimulus type, $F(1,5)<1$. There was a significant interaction of visual condition and stimulus type, $F(2,10)=4.81$, $p=0.034$, $\eta_p^2=0.49$, corresponding to the greater magnitude of effects of congruent and incongruent visual cues for natural visual stimuli than for artificial visual stimuli. For participants trained on rearranged audiovisual category distributions, the effect of visual condition was significant, $F(2,10)=70.3$, $p<0.001$, $\eta_p^2=0.93$, and the main effect of stimulus type was not significant, $F(1,5)=1.24$, $p=0.32$, $\eta_p^2=0.20$. In contrast to the participants trained on the original arrangement of stimulus distributions, the interaction of visual condition and stimulus type was not significant, $F(2,10)<1$. Thus, the magnitude of effects of artificial visual stimuli (in the high noise condition) did not differ from those of natural visual stimuli for participants trained on rearranged distributions.

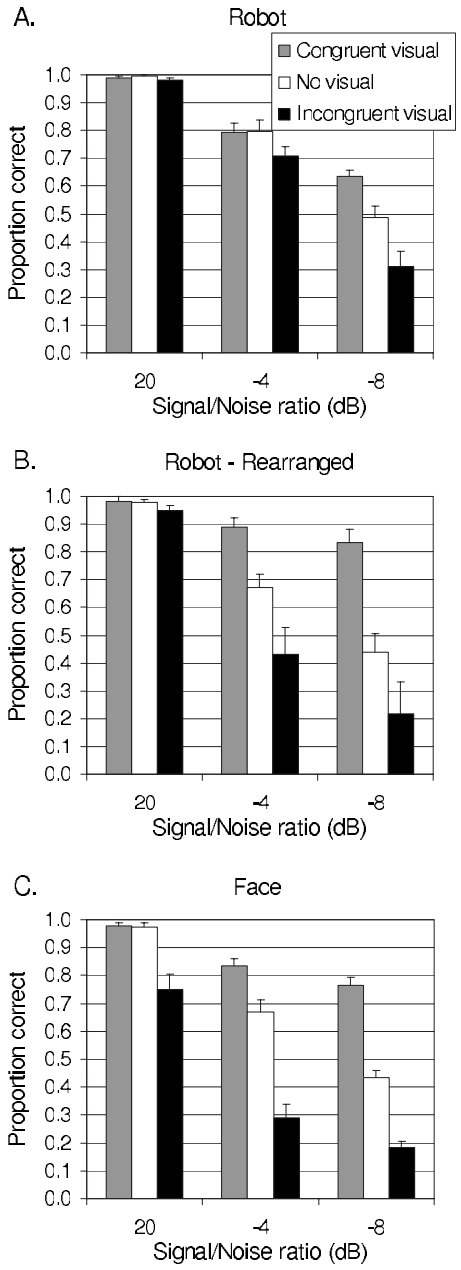

FIG. 5. Proportion correct identification of auditory consonants in the audiovisual mismatch in noise task. (a) Data for artificial visual stimuli with original distributions. For the participant who received different noise levels than other participants (+20, −8, −12 dB), only data from the +20 and −8 dB conditions are included in the figure. (b) Data for artificial visual stimuli with rearranged distributions. (c) Data for natural visual stimuli. Data for natural visual stimuli represent averages of all participants trained on both sets of stimulus distributions. Error bars represent standard error of the mean.

The effects of artificial visual stimuli on intelligibility in high noise were as great as the effects of natural visual stimuli among participants trained on the rearranged distributions, but not among participants trained on the original distributions. This finding suggests that the structure of the stimulus space did have effects on the use of visual information in speech perception.

TABLE II. Response proportions for congruent visual information versus no visual information, in the audiovisual mismatch in noise task, at −8 dB S/N ratio.

	Response					
	B		D		G	
Stimulus	*M*	*SD*	*M*	*SD*	*M*	*SD*
/b/+no visual[a]	0.41	0.19	0.28	0.13	0.31	0.19
/b/+face	0.89	0.21	0.08	0.12	0.03	0.12
/b/+robot-orig.	0.85	0.12	0.03	0.05	0.12	0.10
/b/+robot-rearr.	0.77	0.19	0.12	0.12	0.12	0.12
/d/+no visual	0.27	0.16	0.42	0.16	0.30	0.13
/d/+face	0.02	0.04	0.82	0.17	0.17	0.14
/d/+robot-orig.	0.13	0.15	0.65	0.15	0.22	0.18
/d/+robot-rearr.	0.05	0.12	0.90	0.15	0.05	0.05
/g/+no visual	0.18	0.14	0.32	0.13	0.51	0.15
/g/+face	0.00	0.00	0.42	0.19	0.58	0.19
/g/+robot-orig.	0.13	0.14	0.47	0.23	0.40	0.15
/g/+robot-rearr.	0.12	0.12	0.05	0.05	0.83	0.14

[a]Data in the "no visual" conditions are averaged across both tasks (robot and face).

Why did the two distributions lead to different degrees of visual influence? Table II presents a confusion matrix for participants' responses in congruent audiovisual versus unimodal auditory conditions at the high noise level (−8 dB). For participants trained on the original audiovisual distributions, congruent visual information disambiguated /b/ from the other consonants but was less successful in helping participants to correctly identify /d/ and /g/. Natural visual stimuli showed a similar pattern in that identification of /g/ improved little with congruent visual information (although /d/ identification did benefit considerably). In contrast, for participants trained on rearranged audiovisual distributions, artificial visual speech improved identification of all three consonants, including /g/. From the response proportions in the unimodal auditory conditions, it can be seen that auditory /b/ and /g/ were least likely to be confused with each other, whereas auditory /d/ was highly confusable with the other consonants in noise. Thus, making /b/ the most distinctive visual stimulus (as in the original audiovisual distributions) did little to help distinguish between the most confusable consonants. On the other hand, making /d/ the most distinctive visual stimulus (as in the rearranged distributions) provided visual information that was complementary to the available auditory cues and maximized the visual benefit.

2. McGurk effect

As with the data from the factorial audiovisual identification task, the question arises whether artificial visual cues were integrated with auditory information in a similar manner to natural visual speech cues. For example, how do participants respond when presented with combinations of auditory /b/ and visual /g/? A "D" response in this condition constitutes the classic McGurk effect, in which auditory and visual information are apparently perceptually "fused." However, "D" responses to auditory /b/ plus visual /g/ are also possible without perceptual fusion. For instance, a perceiver might produce a "D" response while selectively attending only to auditory /b/ or visual /g/, as in the Single Channel Model. Starting from Eq. (2) above, if p_a is assumed to remain constant throughout the experiment, the SCM predicts that

$$P(\text{“}D\text{”}\,|A_bV_g) = P(\text{“}D\text{”}\,|A_bV_b) + P(\text{“}D\text{”}\,|A_gV_g) - P(\text{“}D\text{”}\,|A_gV_b). \quad (3)$$

As a result, it was possible to test the predictions of the SCM within the data sets from the audiovisual mismatch task, by computing both sides of Eq. (3) from each participant's response proportions in the relevant conditions. Figures 6(a) and 6(c) show "D" responses in the McGurk condition for original distributions of artificial visual stimuli and for natural visual stimuli compared to the control condition represented by the right side of Eq. (3). For rearranged distributions [Fig. 6(b)], an analogous analysis was performed that examined "B" responses to auditory /d/ and visual /g/, since /b/ was the intermediate audiovisual category in these distributions. It can be seen from the figure that only natural visual stimuli produced McGurk-style fusions at a greater rate than would be expected from selective attention to a single modality on each trial. To confirm this observation, separate 2(fusion vs. control)×3(noise level) ANOVAs were conducted on the three data sets presented in Fig. 6 (excluding the participant with different noise levels). The data from the task with natural visual stimuli revealed significant main effects of fusion vs. control, $F(1,11)=23.5$, $p=0.001$, $\eta_p^2=0.68$, and noise level, $F(2,22)=4.67$, $p=0.020$, $\eta_p^2=0.30$, and a significant interaction, $F(2,22)=7.17$, $p=0.004$, $\eta_p^2=0.39$. For the task with robot stimuli from original distributions, there was a main effect of noise level $F(2,8)=6.00$, $p=0.026$, $\eta_p^2=0.60$, but there was no main effect of fusion vs. control, $F<1$, nor was the interaction significant, $F(2,8)=3.92$, $p=0.065$, $\eta_p^2=0.50$. For the task with robot stimuli from rearranged distributions, there was no effect of noise level, $F(2,10)=1.52$, $p=0.26$, $\eta_p^2=0.23$, no main effect of fusion vs. control, $F<1$, and no interaction, $F<1$. Thus, although participants made a measurable number of McGurk-like responses with artificial visual stimuli, there is

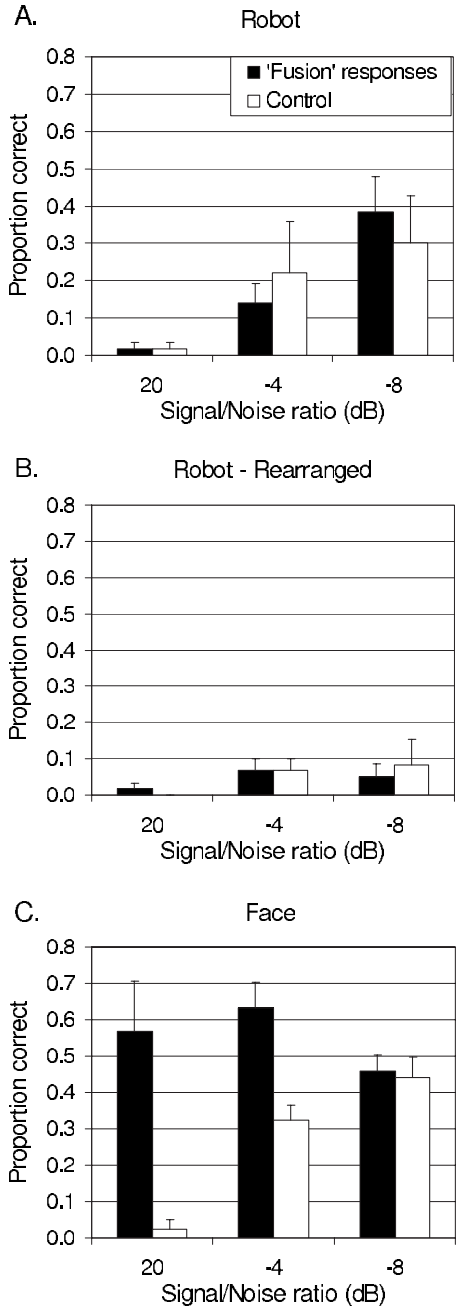

FIG. 6. Proportion "fusion" responses characteristic of the McGurk effect, compared to control predictions based on the assumption of selective attention to a single modality on each trial (see text). (a) Data for artificial visual stimuli with original distributions. For the participant who received different noise levels than other participants (+20, −8, −12 dB), only data from the +20 and −8 dB conditions are included in the figure. (b) Data for artificial visual stimuli with rearranged distributions. (c) Data for natural visual stimuli. Data for natural visual stimuli represent averages of all participants trained on both sets of stimulus distributions. Error bars represent standard error of the mean.

no evidence from this task that these responses resulted from integration of the auditory and visual cues.

IV. DISCUSSION

A. Summary of current findings

In the present study, participants learned a new and arbitrary source of visual information for phonetic categories. The visual cues were dynamic and were synchronized with auditory speech. After training, participants correctly applied phonetic category labels to these artificial visual speech cues, and made use of the visual cues in phonetically labeling acoustically ambiguous speech stimuli. Although the visual stimuli shifted participants' identification responses in the identification tasks, the effects of artificial visual cues differed from those typically seen with natural visual cues from a speaker's face.

In the factorial audiovisual identification task, the effect of artificial visual information was limited mainly to audiovisual combinations in which the auditory stimulus was highly ambiguous. With ambiguous auditory stimuli, the size of the visual effect was not commensurate with participants' high accuracy in identification of unimodal visual stimuli. As a result, the Fuzzy Logical Model of Perception (Oden and Massaro, 1978) did not fit the data more accurately than an alternative model that does not assume information integration across modalities.

In the audiovisual mismatch in noise task, artificial visual cues influenced participants' ability to correctly identify auditory consonants presented in noise. Under certain conditions, the effects of artificial visual cues on accuracy were as great as those observed with natural visual stimuli. Additionally, the specific visual effects on individual consonants reflected the structure of the audiovisual categories on which participants were trained. As in the factorial audiovisual identification task, there was no evidence that artificial visual cues were integrated with auditory information in a manner similar to natural visual cues, and an analysis of "fusion responses" characteristic of the McGurk effect (McGurk and MacDonald, 1976) indicated that the effects of artificial visual cues were consistent with the predictions of a model that does not assume integration of information across modalities.

B. Theoretical implications

The use of information from separate sensory modalities has figured prominently in theoretical approaches to speech perception. Computational models of speech perception (e.g., FLMP: Oden and Massaro, 1978; TRACE: McClelland and Elman, 1986; and Merge: Norris *et al.*, 2000) posit that information is perceptually integrated based on associations among relevant information sources and perceptual categories. Further, research on such models has increasingly emphasized the *optimality* (in the Bayesian sense) of information integration in speech perception (e.g., Massaro, 1998, ch. 4; Movellan and McClelland, 2001; Norris and McQueen, 2008). That is, the probability of identifying a phonetic category reflects an ideal combination of the conditional probabilities of that category given the information from each available source (e.g., auditory, visual, lexical). Based on this literature, it might be expected that newly-learned artificial visual cues for consonants would be optimally exploited by the perceptual system in this way. This was not the case in the current study, demonstrated by the comparison between current data and the predictions of an optimal model (FLMP). Rather, the effects of newly-learned

visual information in the current study are consistent with some form of attention-switching between modalities rather than information integration (perhaps reflecting task demands learned in training, in which use of the visual cues was encouraged; see also Fowler and Dekle, 1991; Massaro, 1998, ch. 2).

There is no ready explanation for why well-learned visual cues from the speech robot would not be combined with auditory speech. Thus, the central theoretical question raised by this study is: what conditions are sufficient to allow for optimal use of an information source in speech perception? One possible answer is that the newly learned visual cues were in fact used optimally in some sense, if it is taken into consideration that the perceiver has relatively little experience with them (compared to natural cues) and should treat them as unstable or unreliable. Computational models of information integration might therefore be able to accommodate the current findings by incorporating some extra mechanism by which new information sources compete with (or are otherwise disadvantaged by) more established information sources in perception. Predictions could then be made regarding the degree of experience and other conditions necessary for novel information sources to be perceptually integrated.

An alternative possibility is that information sources such as the speech robot can never be perceptually integrated with auditory speech, no matter how much experience is provided. According to the direct realist account of audiovisual speech perception (e.g., Fowler, 1996), perceptual integration only occurs when both information sources are linked to a shared environmental cause (i.e., the gestures of a speaker's vocal tract). The current results are thus consistent with direct realism, although they are equally consistent with the interpretation that the amount of experience provided was simply not sufficient for perceptual integration.

C. Implications for speech comprehension in adverse conditions

Aside from theoretical considerations, the benefit of visual stimuli from rearranged category distributions on speech identification in noise is interesting from a purely practical standpoint. Even though it is unlikely that participants integrated the visual and auditory stimuli as in natural audiovisual speech perception, they nonetheless were able to use the artificial visual stimuli to achieve similar levels of accuracy in a noisy environment. A particularly interesting aspect of the data was that the rearranged categories apparently disambiguated auditory cues more effectively because they provided the most distinctive visual cues for the consonant (/d/) that was most confusable with other consonants in noise. A possible implication of these findings is that artificial visual cues may have some value as an aid to speech perception, since they could be constructed specifically to disambiguate confusable phonetic categories and thus maximize the information available to perceivers when auditory information is degraded.

D. Conclusion

The current study establishes a novel empirical paradigm for training perceivers on novel visual cues for consonant categories. The results raise important theoretical questions about the nature of information integration in speech perception and how it may be brought about through experience. The findings also imply that supplementing auditory speech with dynamic visual information may be beneficial to speech comprehension, even if the underlying perceptual mechanisms are not equivalent to those of natural speechreading.

ACKNOWLEDGMENTS

This research was supported by NIH award 1 F31 DC007284-01 to JDWS, by National Science Foundation award BCS-0345773 to LLH, by a grant from the Bank of Sweden Tercentenary Foundation to LLH, and by the Center for the Neural Basis of Cognition. The authors thank Christi Gomez for help with data collection and manuscript preparation.

[1]Example audiovisual stimuli may be viewed at the following URL: http://www.psy.cmu.edu/~lholt/php/gallery_audiovisual.php.

[2]The entire set of 160 tokens may also be downloaded from: http://www.u.arizona.edu/~alotto/ACNS/StimuLibrary.htm.

[3]For the three participants who completed the experiment on a desktop computer in the laboratory, an unforeseen programming bug affected the adjustment of noise in the explicit training task and prevented these participants from attaining the criterion in some sessions (two sessions for two participants and one session for the other participant). The only consequence of this error was that it may have slowed the progress of these participants through the phases of the experiment.

[4]Due to exploration of the best noise levels for use in the task, noise levels in the version of the task that featured the robot were +20 dB, −8 dB, and −12 dB for one participant (trained on original distributions).

Atal, B. S., and Hanauer, S. L. (**1971**). "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am. **50**, 637–655.

Bernstein, L. E., Auer, E. T., Jr., and Takayanagi, S. (**2004**). "Auditory speech detection in noise enhanced by lipreading," Speech Commun. **44**, 5–18.

Bertelson, P., Vroomen, J., and de Gelder, B. (**2003**). "Visual recalibration of auditory speech identification: A McGurk after effect," Psychol. Sci. **14**, 592–597.

Boersma, P. (**2001**). "PRAAT, a system for doing phonetics by computer," Glot International **5**, 341–345.

Fowler, C. A. (**1996**). "Listeners do hear sounds, not tongues," J. Acoust. Soc. Am. **99**, 1730–1741.

Fowler, C. A., and Dekle, D. J. (**1991**). "Listening with eye and hand: Crossmodal contributions to speech perception," J. Exp. Psychol. Hum. Percept. Perform. **17**, 816–828.

Ganong, W. F. (**1980**). "Phonetic categorization in auditory word perception," J. Exp. Psychol. Hum. Percept. Perform. **6**, 110–125.

Grant, K. W., and Seitz, P. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am. **108**, 1197–1208.

Lachs, L., Pisoni, D. B., and Kirk, K. I. (**2001**). "Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: A first report," Ear Hear. **22**, 236–251.

Markel, J. D., and Gray, A. H., Jr. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, New York), pp. 1–288.

Massaro, D. W. (**1998**). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle* (MIT, Cambridge, MA), pp. 35–79, 95–127, and 415–443.

Massaro, D. W. (**1999**). "Speechreading: Illusion or window into pattern recognition," Trends Cogn. Sci. **3**, 310–317.

Massaro, D. W., Carreira-Perpinan, M. A., and Merrill, D. J. (**2009**). "Opti-

mizing visual perception for an automatic wearable speech supplement in face-to-face communication and classroom situations," in Proceedings of the 42nd Hawaii International Conference on System Sciences, Waikoloa, HI, January 5–8.
Massaro, D. W., and Chen, T. H. (**2008**). "The motor theory of speech perception revisited," Psychon. Bull. Rev. **15**, 453–457.
Massaro, D. W., and Cohen, M. M. (**1999**). "Speech perception in hearing-impaired perceivers: Synergy of multiple modalities," J. Speech Lang. Hear. Res. **42**, 21–41.
Massaro, D. W., Cohen, M. M., and Gesi, A. T. (**1993**). "Long-term training, transfer, and retention in learning to lipread," Percept. Psychophys. **53**, 549–562.
McClelland, J. L., and Elman, J. L. (**1986**). "The TRACE model of speech perception," Cognit Psychol. **18**, 1–86.
McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature (London) **264**, 746–748.
Mirman, D., McClelland, J. L., and Holt, L. L. (**2006**). "Interactive activation and Hebbian learning produce lexically guided tuning of speech perception," Psychon. Bull. Rev. **13**, 958–965.
Montgomery, A., and Jackson, P. (**1983**). "Physical characteristics of the lips underlying vowel lipreading performance," J. Acoust. Soc. Am. **73**, 2134–2144.
Movellan, J. R., and McClelland, J. L. (**2001**). "The Morton-Massaro law of information integration: Implications for models of perception," Psychol. Rev. **108**, 113–148.
Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (**1996**). "Temporal constraints on the McGurk effect," Percept. Psychophys. **58**, 351–362.
Norris, D., and McQueen, J. M. (**2008**). "Shortlist B: A Bayesian model of continuous speech recognition," Psychol. Rev. **115**, 357–395.
Norris, D., McQueen, J. M., and Cutler, A. (**2000**). "Merging information in speech recognition: Feedback is never necessary," Behav. Brain Sci. **23**, 299–325.
Norris, D., McQueen, J. M., and Cutler, A. (**2003**). "Perceptual learning in speech," Cognit Psychol. **47**, 204–238.
Oden, G. C., and Massaro, D. W. (**1978**). "Integration of featural information in speech perception," Psychol. Rev. **85**, 172–191.
Samuel, A. G., and Kraljic, T. (**2009**). "Perceptual learning for speech," Atten. Percept. Psycho. **71**, 1207–1218.
Stephens, J. D. W. (**2006**). "The role of learning in audiovisual speech perception," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, pp. 97–119.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
Tyler, R. S., Lowder, M. W., Parkinson, A. J., Woodworth, G. G., and Gantz, B. J. (**1995**). "Performance of adult ineraid and nucleus cochlear implant patients after 3.5 years of use," Audiology **34**, 135–144.
van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (**2004**). "Integration of letters and speech sounds in the human brain," Neuron **43**, 271–282.
Walden, B. E., Busacco, D. A., and Montgomery, A. A. (**1993**). "Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons," J. Speech Hear. Res. **36**, 431–436.