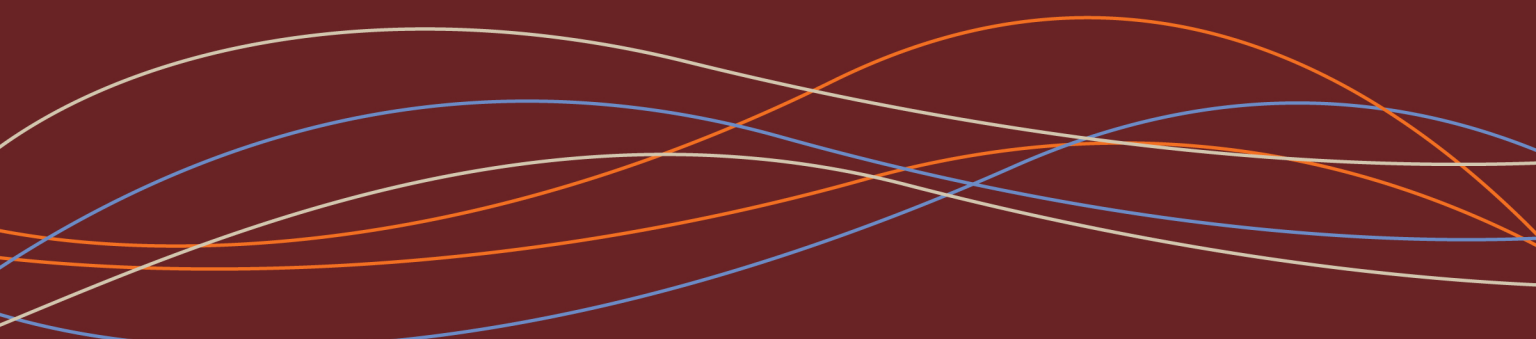


AJP

Volume 124 • Number 3 • Fall 2011

The American Journal of Psychology



Edited by Robert W. Proctor *Purdue University*

Book Reviews Dominic W. Massaro *University of California, Santa Cruz*

Obituaries and History of Psychology Alfred H. Fuchs *Bowdoin College*

Integration of Facial and Newly Learned Visual Cues in Speech Perception

DOM MASSARO, MICHAEL M. COHEN, HEIDI MEYER, TRACY STRIBLING, CASS STERLING, and SAM VANDERHYDEN
University of California, Santa Cruz

We are developing technology to translate acoustic characteristics of speech into visual cues that can be used to supplement speechreading when hearing is limited. Research and theory have established that perceivers are influenced by multiple sources of sensory and contextual information in spoken language processing. Previous research has also shown that additional sources of information can be learned and used to supplement those that are normally available but have been degraded by sensory impairment or difficult environments. We tested whether people can combine or integrate information from the face and information from newly learned cues in an optimal manner. Subjects first learned the visual cues and then were tested under three conditions. Words were presented with just the face, just the visual cues, or both together. Performance was much better with both cues than with either one alone. Similar to the description of previous results with audible and visible speech, the present results were well described by the Fuzzy Logical Model of Perception (Massaro, 1998), which predicts optimal or maximally efficient integration.

The need for language aids is pervasive in today's world; for example, 36 million people in the United States live with hearing deficits and have extraordinary difficulty participating in spoken interaction (Kochkin, 2005; NIDCD, 2008). Worldwide, in 2015 there will be about 700 million people with a hearing loss of 25 dB or greater (Hear It, n.d.). There has also been an increase in hearing loss in U.S. teens: In the last 15 years the rate of slight loss increased by 30% and mild or worse hearing loss increased by 77% (Shargorodsky, Curhan, Curhan, & Eavey, 2010).

Many people rely on lipreading, cued speech, cochlear implants, or hearing aids to help them perceive spoken language. Our goal is to develop another tech-

nology to enhance common face-to-face conversation that offers some advantages over existing aids. For example, Cued Speech is a recent deliberate solution in which one uses hand gestures while speaking to provide the perceiver with information that potentially disambiguates what is seen on the face (Hage & Leybaert, 2006; LaSasso, Crain, & Leybaert, 2010). However, very few people know Cued Speech or have the motivation to learn it, and therefore people with hearing impairments receive insufficient input in many face-to-face and classroom-like environments.

Our current research involves developing and testing embellished eyeglasses (iGlasses™), which will perform two simultaneous functions (Massaro,

Carreira-Perpinan, & Merrill, 2009, 2010). First, real-time acoustic analysis of an interlocutor's speech will track several speech-relevant acoustic features. In this case, the features are voicing, frication, and nasality. Second, these acoustic features will be transformed into continuous visual cues displayed on the eyeglasses. By integrating these visual cues with lipreading (called speechreading because it involves more than just the lips), the user will gain more complete perceptual access to the conversation.

The iGlasses are being designed as a regular pair of eyeglasses with two small microphones and three colored light-emitting diodes (LEDs). The wearer looks at the interlocutor, and the microphones deliver the interlocutor's speech to a processing device such as an iPhone, which processes the acoustic input. The input is analyzed for low-frequency voicing information, high-frequency frication energy, and nasal resonance that are associated with the acoustic and phonetic properties of voicing, frication, and nasality. The three properties are then transformed in real time into simple visual cues displayed on the three vertically mounted LEDs, as shown in Figure 1. These particular phonetic properties were chosen because they are fairly easy to track in the speech signal and, more importantly, because they distinguish instances within a viseme category (i.e., subsets of phonemes that are highly confusable in speechreading). Seeing the upper teeth over the lower lip could indicate /f/ or /v/, but the adding the voicing feature would uniquely indicate /v/. These cues also require no literacy, which allows preliterate children and other nonreaders to use the iGlasses.

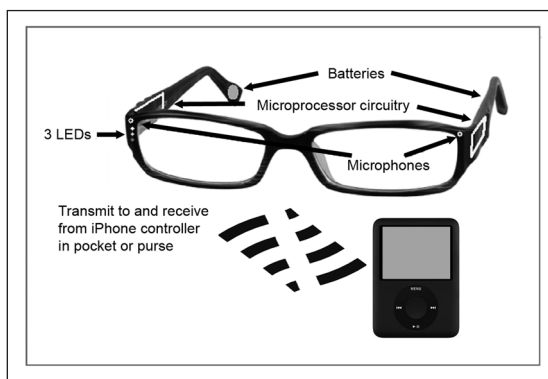


FIGURE 1. Early prototype of the iGlasses device that transforms properties of the speech signal in real time into simple visual cues displayed on 3 vertically mounted LEDs

Previous research showed that the LED cues can be learned to assist in speechreading comprehension and help the user better understand his or her conversational partner (Massaro et al., 2009, 2010). The primary question for the current research is whether the visual cues can be integrated with facial information in the same manner as the integration of audible and visible speech (Massaro, 1998; Massaro & Cohen, 2000). In exploring this question, we begin with a short discussion of relevant empirical and theoretical research, which addresses the issue of information integration and its efficiency in auditory-visual speech perception.

Processing Spoken Language

Speech perception during a face-to-face conversation is a particular form of pattern recognition that exploits many different possible sources of bottom-up and top-down information (Massaro, 1998). Bottom-up sources include audible cues from the voice and visible cues from the face. Top-down cues include situational and linguistic constraints within the conversation. Combining or integrating these multiple cues makes speech perception robust in a variety of challenging situations, such as in a noisy cafe. Experimental studies have shown that the joint presentation of the face and the voice increases comprehension over either visual or audio cues presented individually (Sumbly & Pollack, 1954). This result is found for syllables in isolation, as well as for words and sentences (Massaro, 1998).

These results and other research in a variety of domains and tasks support the predictions of the Fuzzy Logical Model of Perception (FLMP) (Oden & Massaro, 1978; for a summary, see Massaro, 1998), which are that perceivers have continuous rather than categorical information from each of these sources; each source is evaluated with respect to the degree of support for each meaningful alternative; each source is treated independently of other sources; the sources are integrated to give an overall degree of support for each alternative; decisions are made with respect to the relative goodness of match among the viable alternatives; evaluation, integration, and decision are necessarily successive but overlapping stages of processing; and crosstalk between the sources of information is minimal.

The FLMP has also been able to account for a range of contextual constraints in both written and

spoken language processing. Normally we think of context effects occurring at a conscious and deliberate cognitive level, but context effects have been found at even the earliest stages of linguistic processing (Ganong, 1980). The FLMP has also been shown to be mathematically equivalent to Bayes's theorem, an optimal method for combining several sources of evidence to infer a particular outcome (Massaro, 1987). The fact that the model appears to be optimal based on support from a plethora of empirical and theoretical studies provides an encouraging framework for the study of verbal and nonverbal communication (Movellan & McClelland, 2001).

Testing for Integration of Two Sensory Modalities

Unfortunately, the term *integration* tends to be used very loosely in cognitive science and neuroscience, even though one goal of science is to be very precise in its terminology. Dictionary definitions are naturally very broad to account for the fundamental fuzziness of language use. *Integration* might mean "the act of combining into an integral whole," "forming into a whole or introduced into another entity," or "becoming one." In behavioral science, however, a term should be formalized within a precise and testable model. Gick and Derrick (2009) concluded that "perceivers integrate naturalistic tactile information during auditory speech perception without previous training" (p. 502). This conclusion was based on their findings that inaudible air puffs on the skin increased the accuracy of distinguishing between speech sounds that differed in aspiration. It is well documented that more accurate performance does not necessarily mean that integration occurred (Massaro, 1987).

In our research (Massaro, 1987, 1998) we define integration explicitly and make an important distinction between integration and nonintegration processes. Multisensory integration involves combining continuous information from two sensory modalities so that the perceptual experience to a given multisensory event reflects the contribution of both modalities. In auditory-visual speech perception, for example, the pairing of two different syllables often produces a unique identification that is not observed when either of the modalities is presented alone. An auditory /ba/ paired with a visual /da/ often produces /va/ or /ɔ̃a/ identifications, whereas these identifications seldom

occur for the corresponding single-modality presentations (Massaro, 1987, 1998).

A nonintegration process involves identification of a multisensory event that is based on just one of several sensory inputs. Distinguishing between these integration and nonintegration alternatives is not easy and usually requires more elaborate experimentation and model testing than researchers typically carry out (Massaro & Chen, 2008). Although there are several paradigms to study auditory-visual integration in speech perception, we will limit ourselves to a simple task in which test alternatives are presented auditorily with just the voice, visually with just the face, and both together (called a bimodal condition).

Determining whether integration actually occurred usually entails formalizing and distinguishing between integration and nonintegration models. A single-channel nonintegration model (SCM) that has been repeatedly tested assumes that only a single modality or source of information is used even though both are present. Even though audible and visible speech might be presented, for example, the audible speech is used on some proportion p of the trials, and the visible speech is used on the remaining $1 - p$ of the trials. Given many trials, the predicted proportion of a response such as /da/ to a bimodal speech event $A_i V_j$ would be

$$P(/da/) = pa_i + (1 - p)v_j, \quad (1)$$

where a_i is probability of identifying A_i as /da/ and v_j is probability of identifying V_j as /da/. These two probabilities are assumed to be equal to the probabilities of a /da/ identification on the unimodal auditory and visual trials, respectively.

As just described, the FLMP is an integration model. The degree of auditory support for /da/ can be represented by a_i , where the subscript i represents the i th auditory stimulus. In a two-alternative task given auditory and visual speech with /ba/ and /da/ alternatives, the support for /ba/ would be $(1 - a_i)$. Similarly, the degree of visual support for /da/ can be represented by v_j and the support for /ba/ by $(1 - v_j)$. The probability of a response to the unimodal stimulus is simply equal to the appropriate feature value. For bimodal trials, the degree of support for /da/ and /ba/ is $a_i v_j$ and $(1 - a_i)(1 - v_j)$, respectively. The predicted probability of a response, $P(/da/)$, is equal to the support for /da/ divided by the sum of the support for /da/ and /ba/ (a relative goodness rule [RGR]).

$$P(/da/) = \frac{a_i v_j}{a_i v_j + (1 - a_i)(1 - v_j)}. \quad (2)$$

As can be seen in Equation 2, the auditory and visual sources of support are multiplied to give an overall degree of support for each response alternative. The value a_i representing the degree of auditory support is assumed to be the same on both unimodal auditory and bimodal trials. This same property holds for the visual support. The multiplicative integration and the RGR entail the process to be optimal and thus maximally efficient (see Massaro, 1998, pp. 115–117; Massaro & Cohen, 2000; Massaro & Stork, 1998). We now describe the test of the SCM and FLMP against an experiment carried out by Grant and Seitz (1998). We describe this experiment in detail because it is exactly analogous to our current experiment testing whether the face and visual cues are integrated in speech perception. In addition, it sets a standard from auditory–visual experiments to compare to our experiment on visible speech and newly learned visual cues.

Grant and Seitz (1998) Research

Grant and Seitz (1998) tested 40 subjects in a vowel–consonant–vowel consonant identification task with 18 test alternatives. The vowel–consonant–vowel stimuli consisted of the consonants /p/, /b/, /t/, /d/, /k/, /g/, /m/, /n/, /s/, /z/, /f/, /v/, /th/, /dh/, /sh/, /zh/, /jh/, and /ch/ interposed between two /a/ vowels. The 18 alternatives were presented under auditory alone degraded by noise, visual alone, and bimodal conditions. Subjects identified the test items, which were randomly and repeatedly presented.

The results are best analyzed in a confusion matrix for each of the three A , V , and AV conditions in which the test alternatives are presented along the row and the response alternatives along the columns. The entries correspond to the proportion of times each response occurs given each stimulus. The results reflected the psychophysical properties of the test stimuli. Similar stimuli tended to be confused for one another, there was a different similarity relationship for the auditory than the visual stimuli, and the bimodal condition produced a synergistic outcome, which gave much better performance relative to either of the unimodal conditions. For our purposes, the test of the SCM and FLMP provide a strong test of whether the auditory and visual speech were integrated.

There are two ways to represent the auditory and visual information in this type of experiment. Each

modality can be represented by a single source of information, or it can be represented by a set of characteristics called features. Both of these representations, called modality analysis and feature analysis, are informative. They allow independent tests of the models and also indicate the extent to which the features used to describe each modality are perceptually real (Massaro & Cohen, 1999, 2000).

Modality Analysis Implementation

This analysis assumes that each modality contributes an overall measure of support for each alternative. For the SCM, it is necessary to know the probability p of basing the perceptual judgment on the auditory modality and the probability that each unimodal speech item is identified as a given alternative. The predicted probability of a /b/ response given audible /b/ and a visible /b/ would be

$$P(/b/) = p a_{bb} + (1 - p) v_{bb}, \quad (3)$$

where p is the probability of basing the perceptual judgment on the auditory modality, a_{bb} is the probability of identifying an auditory /b/ as /b/, and v_{bb} is the probability of identifying a visual /b/ as /b/. Analogous predictions are made for all other possible pairs on bimodal trials. For example, the predicted probability of a /b/ response given audible /b/ and a visible /d/ would be

$$P(/b/) = p a_{bb} + (1 - p) v_{bd}, \quad (4)$$

where p is the probability of basing the perceptual judgment on the auditory modality, a_{bb} is the probability of identifying an auditory /b/ as /b/, and v_{bd} is the probability of identifying a visual /b/ as /d/.

More generally, the probability of response i given auditory item A_j and visual item V_k is equal to

$$P(/i|A_j V_k) = p a_{ij} + (1 - p) v_{ik}, \quad (5)$$

where a_{ij} is equal to the probability of identifying A_j as alternative i and v_{ik} is equal to the probability of identifying V_k as alternative i .

It is necessary to estimate free parameters to test the SCM. The probability p is a free parameter because we do not know the likelihood that the perceiver will use one modality or the other. With 18 auditory stimuli and 18 visual stimuli, we must estimate the probability of each of the 18 responses to each unimodal test stimulus. In addition to p , 324 free parameters are needed for the auditory modality and 324 for the visual modality. We

are able to test the model by predicting $3 \times 324 = 972$ data points with 649 free parameters. The SCM gave a poor description of the individual subject results, with an average root mean square deviation (RMSD) for each subject of .120.

For the FLMP, it is necessary to estimate a unique parameter to represent the degree to which each source of information supports each alternative. We use a_{ij} to represent the degree to which the audible speech A_j supports the alternative i . The term v_{ik} would represent the degree to which the visible speech V_k supports the alternative i , and so on for the other response alternatives. Given both audible and visible speech, the total support for the alternative /b/, $s(b/)$, would be

$$s(b/|A_j V_k) = a_{ij} v_{ik}, \quad (6)$$

and so on for the other test conditions and the other alternatives.

After the degree of support for each response alternative is calculated, the stimulus is categorized according to the RGR, which states that the relative probability of choosing an alternative is the support for that alternative divided by the sum of the support for all alternatives. With 18 stimulus–response alternatives, each test stimulus provides different degrees of support for each alternative. It is necessary to estimate 18 free parameters for each of the 18 test stimuli in each modality. Thus, 324 free parameters are needed for the auditory modality and 324 for the visual modality. We are able to test the model by predicting 3×324 data points with 2×324 free parameters. The FLMP gave a good description of the individual subject results, with an average RMSD of .011. This goodness-of-fit is more than 10 times better than that given by the SCM.

Feature Analysis Implementation

The modality analysis model makes no assumptions about the psychophysical relationship between the different test items. A unique parameter is estimated for each possible stimulus–response pair. For example, a unique parameter is estimated to represent the amount of support a visual /b/ provides for the response alternative /d/. To test the psychological reality of various speech features and to reduce the number of free parameters, we have articulated the FLMP in terms of audible and visible support for these features (Massaro & Cohen, 1999). This formulation has the potential

to save a large number of free parameters because it is assumed that a given feature in a given modality has the same impact regardless of what segment it is in. Following the tradition begun with Miller and Nicely (1955), we can define the 18 consonants in terms of five features: voicing, nasality, place, duration, and frication (see Massaro & Cohen, 1999).

Table 1 gives the feature values for the 18 consonants used in the Grant and Seitz (1998) study. We assume that features for the 18 consonants are simply sensory primitives that distinguish speech categories. Although the features used in the following tests are chosen to be equivalent to the linguistic features, they should be thought of as simply handy labels for the underlying sensory features. Because the features are sensory, the auditory feature value for place would not necessarily be equivalent to the visual feature value for place. In fact, the auditory feature values appear to be unrelated to or negatively correlated with the visual feature values. Voicing and nasality have informative feature values for auditory speech and neutral feature values for visible speech. The place feature, on the other hand, gives informative values for visible speech. The features at the evaluation stage are not linguistic but perceptual.

The fundamental difference between the SCM and the FLMP remains for the feature analysis implementation. Only a single modality is used on bimodal trials for the SCM, whereas both are used for the FLMP. We determined the influence of each modality for both models by assuming that the feature integration operates identically. Each of the 18 syllables is described by the conjunction of 5 features for unimodal speech and the conjunction of 10 features for bimodal speech. Even though each feature is defined as a specific value or its complement (e.g., voiced or voiceless), its influence in the perception of visible speech is represented by a value between 0 and 1. The parameter value for the feature indicates the amount of influence that feature has. Therefore, if the /ma/ and /na/ prototypes are each expected to have a nasal feature and the calculated parameter value for this feature is .90, then the nasal feature is highly functional in the expected direction. Alternatively, if the calculated parameter value for the nasal feature is .50, then the nasal feature is not functional at all. Because of the definition of negation as one minus the feature value, a feature value of .5 would give the same degree of support

TABLE 1. Features based on Miller and Nicely (1955) Face Cues for the English Consonants That Were in the Grant and Seitz (1998) Study and the Current Experiments

Phoneme	Word	Voicing	Nasality	Place	Duration	Frication
p	Pay	-	-	1	+	-
b	Bee	+	-	1	+	-
m	Me	-	+	1	+	-
t	Too	-	-	3	+	-
d	Day	+	-	3	+	-
n	No	-	+	3	+	-
k	Key	-	-	4	+	-
g	Go	+	-	4	+	-
ng	Sing	-	+	4	+	-
th	Thin	-	-	3	+	+
dh	Then	+	-	3	+	-
f	For	-	-	2	+	+
v	Vote	+	-	2	+	-
s	See	-	-	3	+	+
z	Zoo	+	-	3	+	-
sh	She	-	-	2	-	+
zh	Azure	+	-	2	-	-
j	You	+	-	1	-	-
jh	Judge	+	-	2	-	-
ch	Church	-	-	2	-	+
hh	He	-	-	4	-	+
l	Let	+	-	2	-	-
r	Read	+	-	2	-	-
w	We	+	-	4	-	-

for an alternative that has the feature as it should for an alternative that does not have the feature. If the calculated parameter value is .20, then the nasal feature is functional but the opposite of the expected direction. Finally, it should be noted that the features are not marked in this formulation: Absence of nasality is as informative as presence of nasality. Thus, if a nasal stimulus supports a nasal response alternative to degree .9, then a nonnasal stimulus also supports a nonnasal alternative to degree .9.

The overall match of a test stimulus to each syllable prototype was calculated by combining the feature matches according to the assumptions of the FLMP. These constraints dictate that the features are the sources of information that are evaluated independently of one another, and the features are integrated multiplicatively to give the overall degree of support

for a syllable alternative. It follows that the overall degree of support for /ba/, $s(/ba/)$, given the presentation of a /ba/ syllable, is

$$s(/ba//ba/) = a_v a_n a_p a_d a_f v_n v_p v_d v_f, \quad (7)$$

where each feature value indexes a match between the feature in the stimulus and the corresponding feature in the /ba/ prototype. The feature a_v corresponds to auditory voicing, v_n to visual nasality, and so on. A mismatch between the feature in the stimulus and the corresponding feature in the prototype would be indexed by $(1 - f)$, where f corresponds to the modality's feature value. Thus, the support for the /ka/ prototype, given presentation of a /ba/ syllable, is

$$s(/ka//ba/) = (1 - a_v) a_n (1 - a_p) a_d (1 - a_f) (1 - v_v) v_n (1 - v_p) v_d (1 - v_f), \quad (8)$$

where $(1 - f_i)$ indexes a mismatch between the feature in the /ba/ stimulus and the corresponding feature in the /ka/ prototype. Even though the place feature has six levels rather than just the two levels for the other four features, we assume a mismatch if the two syllables differ in place regardless of the degree of mismatch. Finally, it should be noted that the model predicts a symmetric confusion matrix because it is assumed that $s(/ka//ba/) = s(/ba//ka/)$.

The SCM was tested by Equation 5 except that the probabilities of a response given each unimodal stimulus were assumed to be equal to the overall degrees of support, as illustrated in Equations 7 and 8. Thus, in addition to the free parameter for p , five parameters are necessary to describe the auditory information and five to describe the visual.

For the FLMP, after the overall degree of support for each syllable is calculated, the stimulus is categorized according to the RGR, which states that the relative probability of choosing an alternative is the goodness of match of that alternative divided by the sum of the goodness of match of all alternatives. This model implementation parallels the previous one in all aspects except in terms of the featural description of the stimulus and response alternatives. The FLMP can thus be tested against the confusion matrix by estimating the amount of information in each feature and the featural correspondence between the stimulus and response prototypes. Thus, five parameters are necessary to describe the auditory information and the same number to describe the visual.

The FLMP was fit to the results in the same manner, with the neutral feature value of .5. The average RMSD of the standard FLMP fit is .1001. Although this goodness of fit is much worse than the fit of the modality implementation (.0111), it also requires only 10 parameters, compared with the 648 parameters of the modality implementation. Thus, the fit can be considered at least as good because a roughly 65 to 1 reduction in free parameters resulted in only a 10 to 1 decrease in goodness of fit.

In summary, our model tests revealed that the FLMP gave a much better description of the confusion matrices than the SCM.

Supplementary Visual Cues in Speech Perception

As mentioned previously, our iGlasses device is designed to analyze the speech input for the acoustic-phonetic properties of voicing, frication, and nasality. These properties would then be transformed in real time into simple visual cues displayed on the three vertically mounted LEDs. These particular phonetic properties were chosen because they are fairly easy to track in the speech signal and, more importantly, because they distinguish instances within a viseme category (i.e., subsets of phonemes that are highly confusable in speechreading). Wearers of the iGlasses must learn to associate these visual cues with the appropriate speech properties, use them in speech recognition, and integrate them with information from the face. The following experiment provides evidence that such learning, use, and integration is possible.

EXPERIMENT

METHOD

Subjects

There were five subjects. All were young adults (approximately early twenties). One subject wore hearing aids that had partially compensated for a hearing loss since early childhood. The exact hearing level of the subjects was not measured because the experiment involved speechreading without sound. All subjects received monetary compensation.

All subjects were familiarized with the iGlasses project and received sufficient background training on the cues and speechreading. The preliminary time on task for familiarization with the cues ranged from 30 to 50 hr. The approximate distribution of hours is as follows: Subject E, 40 hr; Subjects H, K, and S,

50 hr; Subject T, 30 hr on this background training plus about 20 hr on another version of the cues with another device. All of the learning and testing was carried out on an iPhone, iPod Touch, or iPad, using an internal BaldiExp application. Rather than laboriously record video of real talkers, we used Baldi as the interlocutor. Baldi is a computer-generated animated talking head shown to be almost as visually accurate as a real speaker (Cohen, Massaro, & Clark, 2002; Massaro, 1998).

Colored disc cues were used to simulate the LEDs on the iGlasses, as shown in Figure 2. Table 2 lists the phonemes and their corresponding visual cues for the English consonants that help identify the consonant when only the visible speech from the face is present. There are three groups of segments: nasality (red), frication (white), and voicing (blue). (It should be noted that, for facilitating learning and use, the cues were defined ideally to distinguish phonemes within a viseme class. For example, /f/ was defined as fricative and /v/ was defined as voiced even though in real speech /v/ has some frication.) The cues are not meant to distinguish between the vowels, which all are assigned to the voicing group. The colored discs light up when a phoneme from the associated group is presented. In the present case of the BaldiExp application, the cues are small discs located near Baldi's mouth, whereas in the final device they will be presented on iGlasses.

The sequence of cues helps distinguish between words because they can differ from one another. This sequence of cues combined with speechreading should be enough to distinguish a word from other words that may have the same color sequence or the same visual appearance. An example of this would be distinguishing *bat* from *mat*, which look mostly identical on the face but have different visual cues. Both words will end with a blue cue for the vowel /ae/ followed by a white cue for the fricative /t/. However, the word *bat* will begin with a blue voicing cue for /b/, whereas the word *mat* will begin with a red nasality cue. Similarly, *vad* and *bag* will have the same visual cues but will differ from one another in the face.

Learning Procedure

Subjects became familiar with the cues by performing three different types of tasks. The test items were pronounceable syllables, words, word sequences, and short phrases. Examples are the words shown in Table 3, sequences of these words and others, and phrases such as "A date book" and "Find the dead mouse."

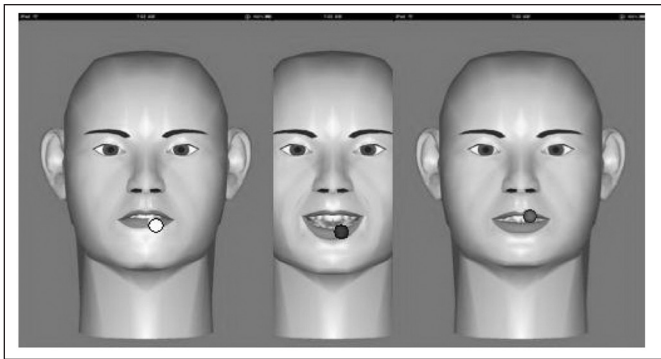


FIGURE 2. Three frames of Baldi's mouthing of the word *fan*. The white disc occurs during /f/, the blue disc during /æ/, and the red disc during /n/. The position of the red, white, and blue cues is fixed at top, middle, and bottom, respectively

TABLE 2. Definition of the Visual Cue Values for the English Consonants That Help Identify the Consonant When Only the Visible Speech From the Face Is Present

Phoneme	Word	Nasality	Frication	Voicing
p	Pay	-	+	-
b	Bee	-	-	+
m	Me	+	-	-
t	Too	-	+	-
d	Day	-	-	+
n	No	+	-	-
k	Key	-	+	-
g	Go	-	-	+
ng	Sing	+	-	-
th	Thin	-	+	-
dh	Then	-	-	+
f	For	-	+	-
v	Vote	-	-	+
s	See	-	+	-
z	Zoo	-	-	+
sh	She	-	+	-
zh	Azure	-	-	+
j	You	-	-	+
jh	Judge	-	-	+
ch	Church	-	+	-
hh	He	-	+	-
l	Let	-	-	+
r	Read	-	-	+
w	We	-	-	+

Note. A + means that the cues occur for that phoneme. Nasality is signaled by the top red disk, frication is signaled by the middle white disk, and voicing is signaled by the bottom blue disk.

TABLE 3. Test Items in the Two 18-Alternative Tests

Test A	Test B
Path	Pan
Pang	Pat
Pad	Pad
Bath	Ban
Bang	Bat
Bad	Bad
San	Man
Sad	Mat
Sat	Mad
Zan	Tap
Zad	Tab
Zat	Tam
Fan	Dap
Fad	Dab
Fat	Dam
Van	Nap
Vad	Nab
Vat	Nam

The different types of exercises used during the learning phase are described in Table 4. A learn exercise consists of Baldi presenting each word in a test set by speaking aloud simultaneously with cues, followed by Baldi mouthing the word with no sound simultaneously with cues. The idea behind a learn exercise is to allow the user to become familiar with the word set and how each word will look spoken by Baldi with the cues. A learn-speak exercise times the second repetition of the word so that the user can speak the word simultaneously with Baldi's mouthing while seeing the corresponding cues. The idea is that this will connect the cues with the mouth movements in a more integrative way, which will increase proficiency in using cues.

A test exercise presents each word with Baldi mouthing simultaneously with cues, followed by a response request. There are two types of response options: a two-alternative forced choice (2AFC) and an open-ended (OE) response. For 2AFC there will be two options to choose from: the correct answer and an incorrect one. In many trials the incorrect answer is a word that looks very similar to the correct answer when no cues are presented. The user simply touches

TABLE 4. Description of Exercises Used During the Learning Phase of the Experiment

Exercise	Stimulus presentation	Subject response	Feedback
Learn	Voiced presentation with cues	None	Silent presentation with cues
Learn–speak	Voiced presentation with cues	Imitates articulation of	
	Silent presentation with cues	silent presentation with cues	
Test	Silent presentation with cues	2AFC, OE, OE from list	Voiced presentation with cues
Evaluation	Silent presentation with or without cues	2AFC, OE, OE from list	Voiced presentation with cues

Note. 2AFC = 2-alternative forced choice; OE = open-ended response; OE from list = choose alternative from list.

his or her choice on the device screen to make a selection. The OE response option has two variations. One involves typing in the actual word, and the other involves typing in the number corresponding to that word. In both cases the subject is given a list of all words in the test set, with associated numbers for the second response case. Word lists are usually short, between 10 and 30 words, so as not to mar results by losing time searching for the word in the list. After the answer is made, Baldi says the correct word aloud simultaneously with cues. There is also the option to program the test to repeat the word once more with Baldi mouthing with cues as a reinforcement, in the case of an incorrect answer. When a test is completed, a result screen displays the score.

An evaluation exercise randomly presents all words in a test set twice, once as a given condition and once as a no condition. There are two types of evaluation exercises. One is a vowel evaluation, because we were interested in whether the presence of voicing during the vowels influences performance on the consonants. The vowel evaluation evaluates the effect of the voicing cue being on during vowels (the given condition) compared with the voicing cue being off during vowels (the no condition). The cue evaluation assesses the influence of the cues compared with a control condition with no cues. Response options are the same as for a test.

Test Procedure

The test procedure followed the typical procedure used in auditory–visual speech perception experiments. Three test conditions were randomly intermixed in a given experimental session: face alone, cues alone, and both face and cues. Two different

tests with 18 words and pronounceable nonwords each were used (Table 5). These items were chosen so that different subsets of the items looked very similar to one another but differed on the visual cues. (The two lists had unique words except that the words *pad* and *bad* occurred in both lists.) For example, *ban*, *bat*, and *bad* look very similar, but each ends with a different visual cue. Subjects were given a list of the 18 items in alphabetical order to refer to during each test. Each of the 18 items occurred once under each of the three conditions. Table 5 gives the number of test sessions for each of the five subjects.

TABLE 5. Number of Sessions and Proportion Correct Words for the Cue, the Face, and Both the Cue and the Face for Each of the 5 Subjects, Tests A and B

Subject	Sessions	Cue	Face	Both
Test A				
E	13	0.33	0.23	0.91
H	11	0.39	0.23	0.97
K	14	0.52	0.43	0.77
S	4	0.29	0.25	0.96
T	10	0.39	0.19	0.97
Test B				
E	12	0.48	0.09	0.86
H	11	0.52	0.14	0.89
K	11	0.59	0.31	0.80
S	5	0.47	0.10	1.00
T	10	0.56	0.14	0.96

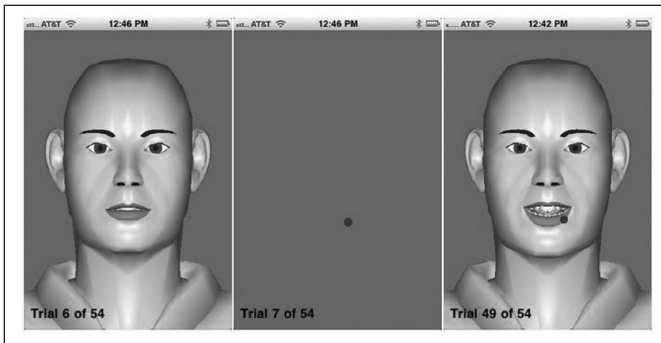


FIGURE 3. Screen shot of the three types of trials (left to right): face alone, cues alone, and both together

The research was carried out on an iPod or an iPad, depending on availability of the device during the testing session. Figure 3 shows single-frame screen shots of the test item on the three types of trials: face alone, cues alone, and both together. The visible speech was presented by Baldi, a computer-animated face developed to produce accurate visible speech. As can be seen in the figure, the cues were presented as colored dots superimposed on Baldi's face. The cues were generated on the basis of the phonemic segments in the test item. (In the actual application of the iGlasses, the cues would be generated by the acoustic analyses of the interlocuter's speech.)

RESULTS

Table 5 lists the proportion correct for each of the five subjects. As can be seen in the table, performance was much better when both sources of information were available compared with just one. Four of the five subjects were nearly perfect in the both condition. This result provides strong evidence against the SCM because it cannot predict better performance in the both condition than the best performance in the face and cues condition.

Integration of the Face and Newly Learned Cues

The extension of the analysis of audible and visible speech to the new paradigm of the face and visual cues is straightforward. The modality analysis is equivalent to that given to the Grant and Seitz data. The predictions of the SCM and FLMP are given by Equations 1 and 2, respectively, except that the second source of information is the newly learned visual cues rather than auditory information. Table 6 gives the outcomes of these descriptions. As can be seen in the

table, the FLMP provides a much better description of the results than that given by the SCM.

In our 18-alternative task, there were two sets of 18 consonant–vowel–consonant alternatives. As can be seen in Table 3, the test items differ in both the initial and final consonants. Although the modality analysis is equivalent to the audible and visible speech conditions of Grant and Seitz, the feature analysis is somewhat more complicated because the features are defined differently for the face and the newly learned visual cues and because both the initial and final consonant must be considered. For example, for the visual cues, the feature cue fricative would be positive for the segment /s/ and negative for the segment /z/. For the visible speech on the face, both of these segments are defined as fricative. Similarly, /n/ would be nasal and not voiced for the visual cue but nasal and voiced for the face. Table 2 defines the features for the visual cues.

The overall degree of support for /bad/, s(/bad/), given the presentation of a /bad/ syllable, is

$$s(/bad/|/bad/) = (ic_v ic_n ic_f if_v if_n if_p if_d if_f) (fc_v fc_n fc_f ff_v ff_n ff_p ff_d ff_f), \quad (9)$$

TABLE 6. Root Mean Square Deviation Values for the Single-Channel Nonintegration Model (SCM) and Fuzzy Logical Model of Perception (FLMP) for Each of the 5 Subjects, Tests A and B

Subject	SCM	FLMP
Test A		
E	.125	.006
H	.134	.004
K	.111	.022
S	.132	.013
T	.133	.004
Test B		
E	.121	.008
H	.127	.010
K	.113	.023
S	.136	.007
T	.131	.008
Mean	.126	.011

where each feature value indexes a match between the feature in the stimulus and the corresponding feature in the /bad/ prototype. The feature ic_v corresponds to cue voicing, if_n to face nasality, and so on of the *initial* consonant, and feature fc_v corresponds to cue voicing, ff_n to face nasality, and so on of the *final* consonant. The goodness of match of the stimulus with the word is simply a multiplicative function of the goodness of match of the initial and final consonants, as shown by Equation 9. Analogous to the single-consonant condition, a mismatch between the feature in the stimulus and the corresponding feature in the prototype would be indexed by $(1 - ic_i)$ and $(1 - ff_i)$ for the initial and final consonants, respectively. The support for the /vat/ prototype, given presentation of a /bad/ syllable, is

$$s(/vat//bad/) = [ic_v ic_n ic_f] [if_v if_n (1 - if_p) (1 - if_d) (1 - if_f)] [(1 - fc_v) fc_n (1 - fc_f)] [(1 - ff_v) ff_n ff_p (1 - ff_d) (1 - ff_f)], \quad (10)$$

where one minus the feature value indexes a mismatch between the feature in the /bad/ stimulus and the corresponding feature in the /vat/ prototype. The cue features are identical for the initial consonant, whereas the face features for the initial consonant differ on place, duration, and frication. The cue features differ for voicing and frication for the final consonant, whereas the face features for the final consonant differ on voicing, duration, and frication.

The FLMP was fit to the result of the 5 subjects for both the modality analysis and feature analysis implementations. Table 7 gives the outcomes of these descriptions. As can be seen in the table, the FLMP provides a good description of the results.

Integration Efficiency

Given that the FLMP is efficient, the good description of the results allows us to conclude that integration performance is efficient. The model can also be formalized to provide a quantitative measure of integration efficiency. As can be seen in Equation 1, the auditory and visual sources of support are multiplied to give an overall degree of support for each response alternative. A feature value representing the degree of auditory and visual speech or visual cue support is assumed to be the same on both unimodal auditory and bimodal trials. This property and the

TABLE 7. RMSD Values for the Fuzzy Logical Model of Perception Feature and Weighted Feature Models for Each of the 5 Subjects, Tests A and B

Subject	Feature	Weighted feature
Test A		
E	.0851	.0851
H	.0782	.0782
K	.0930	.0921
S	.1080	.1080
T	.0844	.0844
Test B		
E	.1071	.1071
H	.1161	.1161
K	.1044	.1044
S	.1271	.1271
T	.1080	.1080
Mean	.101	.101

multiplicative integration rule, followed by the RGR, cause the process to be optimal and thus maximally efficient (see Massaro, 1998, pp. 115–117; Massaro & Stork, 1998).

A direct way to measure integration efficiency in the FLMP is to determine whether there is any loss of information on bimodal trials relative to unimodal trials. A potentially inefficient integration model can be formalized within the FLMP framework. One simply assumes that less information from the available sources of information can be present on bimodal trials relative to unimodal trials. In this case, the degree of support from the face on bimodal trials, v_{jB} , is compromised by the function

$$v_{jB} = w_f v_{jU} + (1 - w_f) .5, \quad (11)$$

where v_{jU} is the degree of support from the face on unimodal trials. Equation 11 shows that the actual support from the face on bimodal trials can be between v_{jU} and .5. A value of .5 means completely neutral support. An analogous function describes the visual cue information

$$v_{jB} = w_c c_{jU} + (1 - w_c) .5, \quad (12)$$

where $(1 - w_c)$ and w_c correspond to the weights given the face and cue feature values, respectively.

For tasks with two response alternatives or for models with features that lie between 0 and 1 (as in the present case), the feature values represent more support for an alternative to the extent that the weight value is greater than .5. Because the weights can lie between 0 and 1, the smaller the weight value, the less the support is controlled by the unimodal feature value and the more it is controlled by .5 (complete ambiguity). A weight value of 1 makes the same prediction as the original FLMP.

Tests of efficiency therefore simply involve testing this new model and determining the weight values and the extent to which this model gives a better description of performance compared with the standard FLMP. We applied this model to the dataset described by Grant and Seitz (1998) and the results of our current iGlasses integration experiments. The fit was essentially equal to the fit of the standard FLMP for the Grant and Seitz results (Massaro & Cohen, 2000). In the present case, the two-weight model and one-weight models gave RMSDs that were essentially equivalent to the standard FLMP, except for subject K, who had a weight for the visual cue of .698 for Test A only. However, this weight did not lead to a meaningful improvement in the fit of the FLMP. Thus, we can conclude that subjects were able to integrate the newly learned cues with the face in an efficient (optimal) manner.

DISCUSSION

The question motivating the present study was whether people could integrate newly learned cues with visible speech in speechreading. Five subjects learned visual cues to phoneme identity and then were tested under three conditions. Words were presented with just the face, just the visual cues, or both together. As in the previous results with audible and visible speech, performance was much better with both cues than with either one alone. These results were well described by the FLMP (Massaro, 1998), which predicts optimal or maximally efficient integration. This result reinforces our goal of developing technology to translate acoustic characteristics of speech into visual cues that can be used to supplement speechreading when hearing is limited.

Presenting auditory speech characteristics visually on LEDs is a form of sensory substitution (Bach-y-Rita & Kerckel, 2003), although in our case it is supplementary because the presence of auditory speech is not precluded. Several studies have shown a benefit of sensory substitution (e.g., Auvray, Hanneton, & O'Regan, 2007). Brain plasticity allows adaptations in the central nervous system that result in highly comparable perceptual experiences coming from one sensory modality being substituted for another (Rosenblum, 2010). Sensory substitution can result in inputs from one sensory modality reaching brain structures usually devoted to another sensory modality. An example is that sign language can activate the auditory cortex in deaf people (Finney, Fine, & Dobkins, 2001).

Perceptual learning has a long and varied history, even though the literature is somewhat sparse. Not surprisingly, there are several types of perceptual learning. Perceptual learning in vision usually involves the discrimination of fundamental visual characteristics, such as contrast, orientation, and positional offset (Xiao et al., 2008). An important characteristic of this type of perceptual learning that distinguishes it from the current one is whether an association must be learned. In the Xiao et al. task, for example, observers practiced contrast discrimination (i.e., "Which interval contained a higher contrast stimulus in a two-interval trial?"). In our task, subjects were required to determine which of several words was presented given information from the cue, the face, or both together. They had to not only perceive the relevant information but also associate it correctly with one of the test items.

Given the additional complexity of learning and using the information in our task, we can partition it into two separate but important processes. First there is the visual information processing of the cues (and the face). Thus, we can ask how the perceptual learning of the cues depends on retinal location, contrast, color, and other factors. In addition, however, it is necessary to address how the association between the cues and the test items is learned. It appears that accounting for both perceptual learning and association learning is necessary to find an optimal paradigm for learning. Most importantly, we might be able to determine the extent to which the integration of the

cue and face occurs during the perceptual process or the associative process in the task.

Research on the learning of visual cues reflecting speech information can also benefit from research on learning more generally. Research and applications have consistently revealed the value of time on task so much so that it has become legend. From the initial stages of learning to the attainment of expertise, time spent on focused deliberate practice is essential, and the more time spent, the more learning (Ericsson, Charness, Feltovich, & Hoffman, 2006). A second effective principle of learning is the value of distributed or spaced practice relative to massed practice. Given an equivalent amount of time for learning, it is better to space this practice across multiple learning sessions rather than having fewer sessions of learning (Rohrer & Pashler, 2007).

It is also important to determine whether it will be necessary to use different learning paradigms for different people. It is difficult to anticipate what individual differences might necessitate unique learning procedures. Although learning style has been often touted as important, there appears to be no substantial evidence for this influence in learning (Pashler, McDaniel, Rohrer, & Bjork, 2008). Furthermore, we have found that the ability to integrate auditory and visual speech in multisensory speech perception occurs across a broad range of subject differences (Massaro, 1998; Massaro & Cohen, 2000). Consistent with this finding, all subjects in the present experiment revealed the ability to learn the visual cues and to integrate them with the face.

As far as we know, this is the first demonstration of integration of newly learned cues with already-learned cues in pattern recognition. Although Cued Speech certainly facilitates speech processing, a benefit could occur even though the cues are not actually integrated (see the analysis of the Sumbly & Pollack, 1954, results in Massaro, 1987, pp. 85–86). As discussed earlier, the concept of integration must be formalized and tested within a specific model. The present experiment and model tests have only scratched the surface of this potentially important question. They offer a proof of concept but do not measure the relationship between the degree of learning cues and the ability to integrate them. Although the answer to this question will not come

easily, it is important that it be addressed in future research.

NOTES

This research was supported by grants from the Center for Information Technology Research in the Interest of Society and the National Science Foundation (Small Business Innovation Research grant to Animated Speech Corporation, now TeachTown).

Address correspondence about this article to Dr. Dominic W. Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064 (e-mail: massaro@ucsc.edu).

REFERENCES

- Auvray, M., Hanetton, S., & O'Regan, J. K. (2007). Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition with "The vOICe." *Perception, 36*(3), 416–430.
- Bach-y-Rita, P., & Kercel, S. W. (2003). Sensory substitution and the human-machine interface. *Trends in Cognitive Neuroscience, 7*(12), 541–546.
- Cohen, M. M., Massaro, D. W., & Clark, R. (2002). Training a talking head. In D. C. Martin (Ed.), *Proceedings of the IEEE Fourth International Conference on Multimodal Interfaces, (ICMI'02)* (pp. 499–510). Pittsburgh, PA: Pebbles.
- Ericsson, A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge, England: Cambridge University Press.
- Finney, E. M., Fine, I., & Dobkins, K. R. (2001). Visual stimuli activate auditory cortex in the deaf. *Nature Neuroscience, 4*, 1171–1173.
- Ganong, W. F., III. (1980). Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110–125.
- Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature, 462*, 502–504.
- Grant, K. W., & Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104*, 2438–2450.
- Hage, C., & Leybaert, J. (2006). The effect of cued speech on the development of spoken language. In P.-E. Spencer & M. Marschark (Eds.), *Advances in the spoken language development of deaf and hard of hearing children* (pp. 193–211). Oxford, England: Oxford University Press.
- Hear It. (n.d.). *More and more hearing impaired people*. Retrieved from <http://www.hear-it.org/page.dsp?area=134>
- Kochkin, S. (2005). MarkeTrak VII: Hearing loss population tops 31 million people. *Hearing Review, 12*(7), 16–29.
- LaSasso, C. J., Crain, K. L., & Leybaert, J. (Eds.). (2010). *Cued speech and cued language for deaf and hard of*

- hearing children (pp. 503–530). San Diego, CA: Plural Publishing.
- Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.
- Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
- Massaro, D. W., Carreira-Perpinan, M. A., & Merrill, D. J. (2009). Optimizing 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*. Washington, DC: IEEE Computer Society Press.
- Massaro, D. W., Carreira-Perpinan, M. A., & Merrill, D. J. (2010). iGlasses: An automatic wearable speech supplement in for individuals' speech comprehension in face-to-face and classroom situations. In C. J. LaSasso, K. L. Crain, & J. Leybaert (Eds.), *Cued speech and cued language for deaf and hard of hearing children* (pp. 503–530). San Diego, CA: Plural Publishing.
- Massaro, D. W., & Chen, T. H. (2008). The motor theory of speech perception revisited. *Psychonomic Bulletin & Review*, 15(2), 453–457.
- Massaro, D. W., & Cohen, M. M. (1999). Speech perception in hearing-impaired perceivers: Synergy of multiple modalities. *Journal of Speech, Language, and Hearing Science*, 42, 21–41.
- Massaro, D. W., & Cohen, M. M. (2000). Tests of auditory–visual integration efficiency within the framework of the fuzzy logical model of perception. *Journal of the Acoustical Society of America*, 108, 784–789.
- Massaro, D. W., & Stork, D. G. (1998). Speech recognition and sensory integration. *American Scientist*, 86, 236–244.
- Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27, 338–352.
- Movellan, J., & McClelland, J. L. (2001). The Morton–Massaro law of information integration: Implications for models of perception. *Psychological Review*, 108, 113–148.
- NIDCD, National Institute on Deafness and Other Communication Disorders. (2008). *Health information: Statistics on voice, speech, and language*. Bethesda, MD: Author. Retrieved from <http://www.nidcd.nih.gov/health/statistics/vsl.asp>
- Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, 85, 172–191.
- Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*, 9, 106–119.
- Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, 16(4), 183–186.
- Rosenblum, L. D. (2010). *See what I'm saying: The extraordinary power of our five senses*. New York, NY: W.W. Norton.
- Shargorodsky, J., Curhan, S. G., Curhan, G. C., & Eavey, R. (2010). Change in prevalence of hearing loss in US adolescents. *Journal of the American Medical Association*, 304, 772–778.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212–215.
- Xiao, L. Q., Zhang, J. Y., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology*, 18, 1922–1926.

Library Recommendation Form

THE
American Journal
OF
Psychology

To: Librarian / Library Acquisition Committee

From: _____

Position: _____ **Dept:** _____

Email: _____

Phone: _____

I recommend that the library subscribe to *The American Journal of Psychology*. Please include this journal in your next serials review meeting with my recommendation to subscribe.

Signature: _____

Date: _____

For subscription information please contact the University of Illinois Press online at <http://www.press.uillinois.edu/journals/ajp.html>