51 Massaro, D.W. (1988) Ambiguity in perception and experimentation *J. Exp. Psychol. Gen.* 117, 417–421

52 Clark, J.J. and Yuille, A.L. (1990) *Data Fusion for Sensory Information Processing Systems,* Kluwer Academic Publishers

53 Landy, M.S. *et al.* (1995) Measurement and modeling of depth cue combination: in defense of weak fusion *Vis. Res.* 35, 389–412

54 Bruno, N. and Cutting, J.E. (1988) Minimodularity and the perception of layout *J. Exp. Psychol. Gen.* 117, 161–170

55 Law, D.J. *et al.* (1993) Perceptual and cognitive factors governing performance in comparative arrival time judgments *J. Exp. Psychol. Hum. Percept. Perform.* 19, 1183–1199

56 Fischer, S.C. *et al.* (1994) Strategic processing in dynamic spatial acuity tasks *Learn. Indiv. Diff.* 6, 65–105

57 McLeod, P., McGlaughlin, C. and Nimmo-Smith, I. (1985) Information encapsulation and automaticity: evidence from the visual control of finely timed actions, in *Attention and Performance Vol. XI* (Posner, M. and Marin, O., eds), pp. 391–406, Erlbaum

58 Cutting, J.E. *et al.* (1992) Wayfinding on foot from information in retinal, not optical, flow *J. Exp. Psychol. Gen.* 121, 41–72

59 Wann, J.P., Edgar, P. and Blair, D. (1993) Time to contact judgment in the locomotion of adults and preschool children *J. Exp. Psychol. Hum. Percept. Perform.* 19, 1053–1065

60 Alderson, G.K., Sully, H. and Sully, D. (1974) An operational analysis of a one-handed catching task using high speed photography *J. Motor Behav.* 6, 217–226

61 Bridgeman, B. *et al.* (1979) The relationship between cognitive and motor oriented systems of visual position perception *J. Exp. Psychol. Hum. Percept. Perform.* 5, 692–700

62 Milner, A.D. and Goodale, M.A. (1992) *The Visual Brain in Action*, Oxford University Press

63 Ungerleider, L.G. and Mishkin, M. (1982) Two cortical visual systems, in *Analysis of Visual Behavior* (Ingle, D.J., Goodale, M.A. and Mansfield, R.J., eds), pp. 549–586, MIT Press

64 Goodale, M.A., Pellison, D. and Prablanc, C. (1986) Large adjustments in visually guided reaching do not depend upon vision of the hand or perception of target displacement *Nature* 320, 748–750

65 Pylyshyn, Z. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception *Behav. Brain Sci.* (in press)

66 Tyldesley, D.A. and Whiting, H.T.A. (1975) Operational timing *J. Hum. Mov. Stud.* 1, 172–177

67 Bootsma, R.J. *et al.* (1997) On the information-based regulation of movement: what Wann (1996) may want to consider *J. Exp. Psychol. Hum. Percept. Perform.* 23, 1282–1289

Speechreading: illusion or window into pattern recognition

Dominic W. Massaro

In the Fuzzy Logical Model of Perception (FLMP) perceivers are conceptualized as forming perceptual judgments by evaluating and integrating multiple ambiguous sources of information, in an optimal manner based on relative goodness of match. This model has been tested favorably against a variety of competing theories and models. Recent extensions of the FLMP are described in this article along with empirical applications and verification, and progress in the study of speech perception by ear and eye is reviewed within this general theoretical framework. The model illuminates the differences that are observed across different languages in terms of information as opposed to information-processing. Pattern recognition of bimodal speech is representative of pattern recognition in a variety of other domains, such as emotion perception, and there are several domain-dependent reasons why multimodal presentation of audible and visible speech is particularly conducive to accurate pattern recognition. We believe that the positive outcome of this research provides a framework for the development of computer-animated agents, which can serve as language tutors and as conversational characters in other domains, easing the interaction of humans and machines.

D.W. Massaro is at the University of California, Santa Cruz, CA 95060, USA.

tel: +1 831 459 2330
fax: +1 831 459 3519
e-mail: massaro@fuzzy.ucsc.edu

It has been well over two decades since the publication of an article entitled 'Hearing lips and seeing voices' by the late Harry McGurk and his colleague John McDonald[1]. The so-called McGurk effect has obtained widespread attention in many circles of psychological inquiry and cognitive science. The classic McGurk effect involves the situation in which an auditory /aba/ is paired with a visible /aga/ and the perceiver reports hearing /ada/. This outcome is dubbed a so-called fusion response because two different segments are fused into a third. The reverse pairing, an auditory /aga/ and visual /aba/, tends to produce a perceptual judgment of /abga/, a so-called combination response. The McGurk effect had such an

1364-6613/99/$ – see front matter PII: S1364-6613(99)01360-1

impact on research because the visual input actually changes the perceiver's auditory experience.

One question is whether this illusion reveals something essential about speech perception, or about multimodal perception more generally. If one accepted speech as a Fodorian input module[2], then clearly the McGurk effect provides a potential window into the functioning of this module. From a broader perspective, however, we should not be all that surprised by the finding that our auditory experience of speech is influenced by the visual input. For example, ventriloquism, inner voices while reading, and localizing voices in film are additional cases of crosstalk between modalities implying that the McGurk effect might not be unique.

We should be encouraged that the McGurk effect resembles other avenues of experience, such as localizing sound in space[3–5]. Its similarity to other domains offers the expectation of a more general account of sensor fusion and modality-specific experience rather than one unique to speech perception by ear and eye. Research from several laboratories has documented that bimodal speech perception and bimodal localization are highly analogous processes[6,7]. These situations reflect cases of pattern recognition in which several sources of information from different modalities contribute to the perceptual outcome. There are also amodal influences on perceptual experience, however. Without this prior knowledge of the words, the listener cannot make heads or tails of the message. An experimental demonstration of this type of illusion is the so-called phonemic restoration effect in which we claim to hear the /s/ in the word legislatures even when it is replaced by a cough, a buzz or even a pure tone[8,9].

When a spoken word is masked by noise having the same amplitude envelope, subjects report that they hear the word much more clearly when they see the word in print at the same time[10]. This result supports the idea that written text can influence our auditory experience. To show effects of written information on auditory judgment at the perceptual level, we compared the contribution of lip-read information to written information[11]. Although there was a large effect of visible speech, there was only a small (but significant) effect of the written segments BA or DA on the judgments. To better test for the possible influence of text on speech perception, we aimed to obtain a larger effect of written text[12]. Given that letters of the alphabet have a strict spelling-to-sound mapping and are pronounced automatically and effortlessly, the letters B and D were used. This is convenient because the letter sequences BA and DA are not necessarily pronounced /ba/ and /da/, but the letters B and D are pronounced only as they are named in the alphabet, i.e. /bi/ and /di/.

Subjects were instructed to watch a monitor and listen to speech sounds. As can be seen in Fig. 1, there were substantial effects of both visible speech and written letters on perceptual judgments. The effects of written information on auditory judgment can also be seen in Fig. 1. Clearly, we can conclude that written text, as well as visible speech, can influence our auditory experience and that the FLMP accounts for both types of influence.

One important issue is whether integration in the speech domain follows different rules from those describing integration in other domains such as spatial localization. As described by Calvert[13], the fact that sensor fusion occurs in both speech perception and spatial localization in no way implies that they share common neural mechanisms or neural sites. In our view[14,15], these two domains involve different sources of information but they might follow the same algorithm of information-processing. Thus we might expect that these two situations be influenced by different variables but still might follow the same algorithm of combination. This stance, of course, is a testable hypothesis, which we have explored in a variety of domains. As an example, spatial proximity of the two modalities is critical in localization but less important in bimodal speech perception[13]. However, in systematic tests of the FLMP and competing models, the FLMP best described both syllable identification and spatial localization judgments[6]. The two sources of information appear to be combined in similar ways in both localization and speech perception[7].

The fact that perceptual experience is primarily in one modality might not be reflective of the processing that led to the experience. Speech information from the visual and auditory modalities provides a situation in which the brain combines both sources of information to create an interpretation that is easily mistaken for an auditory one. We believe we hear the speech perhaps because spoken language is usually heard. Crosstalk between modalities might simply mean that we couldn't trust modality-specific experience as a direct index of processing[16].

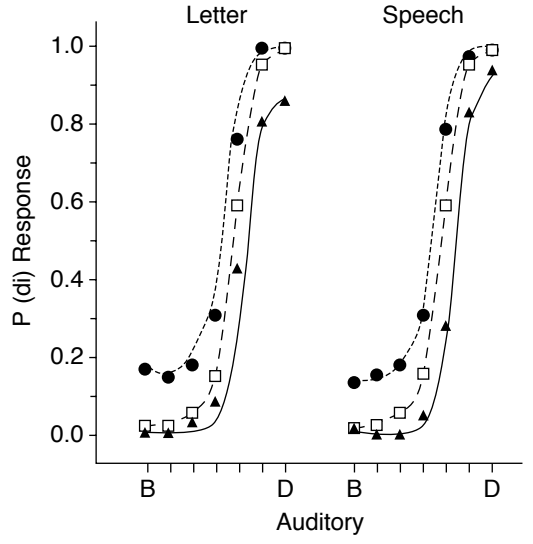

Fig. 1. Average results for the letter and speech conditions in experiments by the author[11,12]. Subjects were instructed to watch a monitor and listen to speech sounds. The proportion of /di/ responses as a function of the seven auditory levels is shown with the visual B (filled triangles) or D (filled circles) stimulus or no visual information (open squares) as the curve parameter. The average proportion of /di/ responses increased significantly as the auditory syllable went from the most /bi/-like to the most /di/-like level. There was also a significant effect on the proportion of /di/ responses as a function of the visual stimulus, with fewer /di/ responses for visual B than for a visual D. The interaction of these two variables was also significant given that the effect of the visual variable was smaller at the less ambiguous regions of the auditory continuum. The main result of interest is that the visual effect was substantial and of similar size for the letter and for the speech conditions.

Implications for theories and models of speech perception

How do the impressive findings of bimodal speech perception impact on extant theories? According to psychoacoustic accounts, speech perception can be understood in terms of the processing of complex auditory signals. A theory of acoustic invariance makes two claims: (1) each phonetic feature contains an invariant acoustic pattern that specifies the value of that feature, and (2) the perceptual system uses this information for speech perception[17,18]. A more modest claim is that it is a conglomeration of auditory dimensions that provide a direct relationship between the acoustic signal and the appropriate percept[19]. The dramatic influence of visible speech seems to falsify these proposals in that we would expect them to predict the putatively direct relationship between sound and percept to be impermeable to the influence of another modality.

Of course, these scientists have recognized the influence of visible speech (see for example, Ref. 19), but they have not specified exactly how visible speech makes its contribution. It would appear that visible speech would somehow have to be secondary to audible speech as, for example, in an auditory dominance model. In this formulation, an effect of visible speech occurs only when the auditory speech is not completely intelligible[20,21]. Because speech is primarily auditory, it might seem reasonable to assume that visible speech plays a secondary role, influencing perception only when the auditory information is not intelligible. The McGurk effect would seem to disqualify this model because the auditory speech is usually identified easily when it is presented alone. More reasonably, it could be proposed that the perceiver uses just a single modality for identification, sometimes the visual and sometimes the auditory. This more general single-channel model has also been systematically falsified in a series of experimental tests[14,15].

Integration models

The models we have described to this point can be classified as non-integration models. For any perceptual experience, there is only a single influence. Integration models, on the other hand, assume that perceptual experience is jointly influenced by both auditory and visible speech. The simplest type of integration model is the Additive Model of Perception (AMP). Additive models have been proposed and tested to explain perception and pattern recognition in several domains[22–24]. In the AMP, it is assumed that the sources of information are simply added together at the integration stage. For generality, it can also be assumed that one modality of information has more influence than the other modality. To implement this assumption, the influence given to each modality has an additional weight parameter. The AMP has been shown to give a very poor description of speech perception in a broad range of experimental conditions[14,15].

The motor theory assumes that listeners analyse the acoustic signal to generate hypotheses about the articulatory gestures that were responsible for it. The perceiver uses the sensory input to best determine the set of articulatory gestures that produced this input[25–28]. The inadequate auditory input is assessed in terms of the articulation, and it is only natural that visible speech could contribute to this process. To postulate a motor explanation of integration in speech seems to violate parsimony, in my opinion, because integration occurs in many other domains[15] that involve no analogous motor medium. Although motor theory was originally developed to account for acoustic or phonetic perception, it has difficulty accounting for the influence of higher-order linguistic context. For example, if the ambiguous auditory sentence '*My bab pop me poo brive*', is paired with the visible sentence '*My gag kok me koo grive*', the perceiver is likely to hear '*My dad taught me to drive*'. Two meaningless sources of information are combined to create a meaningful interpretation[29]. Even if some representation is necessary to account for the joint influence of audible and visible speech, there is as yet no compelling reason why this representation should be a motor one.

The direct perception theory states that persons directly perceive the distal causes of sensory input[30]. In spoken language, the distal cause is the vocal tract activity of the talker, and it is reasonable that visible speech should also influence speech perception because it also reveals the vocal-tract activity of the talker. Speech perceivers therefore obtain direct information from integrated perceptual systems responding to the flow of stimulation provided by the talker[31]. This theory has trouble, however, with the finding that written language can influence speech perception (Fig. 1).

Fuzzy Logical Model of Perception (FLMP)

Pattern recognition is viewed as central to cognition, and the perception of speech by eye and ear is deemed as a prototypical case of pattern recognition. Within the FLMP, perceivers are assumed to utilize multiple sources of information supporting the identification and interpretation of the language input. As illustrated in Fig. 2, the model specifies a set of rules or an algorithm to describe how pattern recognition occurs. There are four successful but overlapping stages of processing. At the evaluation stage, each source of information is evaluated to give the continuous degree to which that source specifies various alternatives. The auditory and visual sources are evaluated independently of one another[32,33].

One of the central assumptions of the FLMP is the independence of the auditory and visual information at the evaluation stage. Mesulam[34] and Calvert[13] indicate that heteromodal cortex might ensure the binding of the modality-specific information, but it might not retain this combination. As an alternative, the sensory-specific information might be represented in the unimodal sensory cortices responsible for their initial processing. If evaluation occurred at these sites, then processing of one modality would be achieved without influence from concurrent processing at other sensory-specific sites.

At integration, the sources are combined multiplicatively to provide an overall degree of support for each alternative. The decision process, which has received less attention than the other processes up to now, has been decomposed to include two component operations. Assessment takes into account all of the viable response alternatives. Response selection follows a probability-matching rule in which the likelihood of a given response is equal to its relative goodness-of-match to the input. We proved that a criterion rule (as assumed in signal-detection theory, for example, Refs 15,35)

could be employed in the FLMP to make the same predictions as probability matching[36].

Within the FLMP, why does auditory /ba/ paired with a visible /ga/ produce a perceptual report of hearing /da/ (i.e. the McGurk effect)? These two sources of information are integrated and the outcome can be explained by the psychophysical properties of the audible and visible sources of information. Visual /ga/ is very similar to visual /da/ and auditory /ba/ is somewhat more similar to an auditory /da/ than to an auditory /ga/. Thus, the alternative /da/ is the best solution given both sources of information[29]. There might also be other sources of information (or constraints) contributing to performance. Higher-order context might be influential in that the segment /d/ appears to be more frequent in initial position in English than the segment /g/.

One inherent attribute of this theoretical model is the important distinction between information and information-processing[15,37]. The sources of information from the auditory and visual channels make contact with the perceiver at the evaluation stage of processing. The reduction in uncertainty effected by each source is defined as information. In the fit of the FLMP, for example, the parameter values indicating the degree of support for each alternative from each modality correspond to information. These parameter values represent how informative each source of information is. Information-processing refers to how the sources of information are processed. In the FLMP, the evaluation, integration, assessment, and response selection stages describe information-processing.

Some of our recent research has also attempted to make progress on the question of the information contained in visible speech. In one experiment, visible speechreading was studied to determine which features are functional and to test several models of pattern recognition[38]. Nine test syllables were presented in intact form or under various levels of spatial quantization. Performance decreased in increasing quantization but remained relatively good at moderate levels of degradation. Six features were identified as functional in distinguishing among the nine consonant-vowel syllables.

The features that appeared to have psychological validity were duration, tongue-tip movement, lip rounding, mouth narrowing, dental adduction, and lower-lip tuck. These features were used as sources of information in the FLMP and an additive model (AMP). The FLMP provided a significantly better description of the confusion matrices, showing that speechreading is analogous to other domains of pattern recognition such as face recognition and facial affect perception.

Selecting among theories and models

Our goal is to broaden the domain of the techniques of model selection in our tests of extant models of speech perception and pattern recognition. Various theories of speech perception have been implemented in quantitative form in order to allow them to be tested against empirical results. A categorical perception (CMP) model grounded in the categorical perception of the auditory and visual speech provides a poor description of performance[14,15]. The CMP predicts that the curves in Fig. 1 would be parallel to one another, even though the distance between the curves is several times larger in the middle

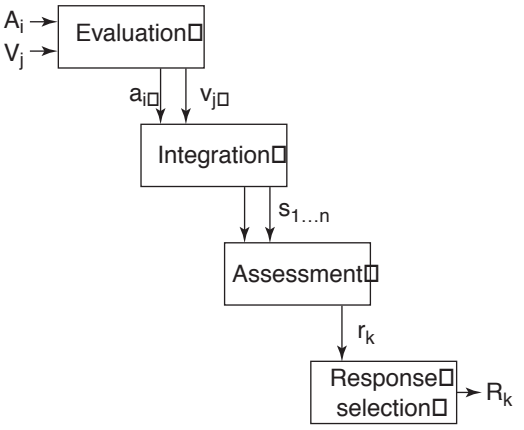

Fig. 2. A schematic representation of the four processes involved in perceptual recognition. The four processes are shown in sequence, left to right, to illustrate their necessarily successive, but overlapping operations. These processes make use of prototypes stored in long-term memory. Uppercase letters represent the sources of information. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j). These sources are then integrated to give an overall degree of support (s_k) for each speech alternative, k. The assessment operation uses these to arrive at some overall measure of support for a given alternative. The response-selection process maps that value into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely.

range of the *x*-axis than at the extremes. The limitation with this type of model is that the categorical outcomes of processing each source cannot be used in an informative manner. If the categorization of the visible speech agrees with the categorization of the audible speech, no new information is obtained. If the two categorizations disagree, then there is nothing to guide the perceiver to choose one or the other outcome.

In the FLMP, the two sources of information are treated as independent of one another at the initial evaluation stage. A contrasting candidate that has been considered is the TRACE model of speech perception[18]. Several researchers have proposed that this model can account for the McGurk effect[21,39]. TRACE is an interactive activation model in which information-processing occurs through excitatory and inhibitory interactions among a large number of simple processing units. Three levels or sizes of units are used in TRACE: feature, phoneme, and word. Features activate phonemes that activate words, and activation of units at a particular level inhibits other units at the same level. In addition, activation of higher-order units activates their lower-order units; for example, activation of a given word unit would activate the phonemes that make up that word.

O'Reilly lists bidirectional activation (interactivity) as one of six principles for biologically based computational models of cortical cognition[40]. He cites evidence for the well-known bidirectional connectivity in cortex[41]. The critical issue, however, is what these bidirectional connections imply about neurological processing in pattern recognition. One interpretation is the exchange of activation during perceptual processing. However, it is equally possible that the second

Box 1. Neurological mechanisms

We consider three possible neurological mechanisms to account for the integration of auditory and visual speech, as assumed by the FLMP (Ref. a). In sensory penetration (Fig. IA), the processing of one modality activates the location that receives activation from the other modality. As illustrated in the figure, the activation from the visible speech is sent to a location that receives activation from the auditory modality. This possibility appears to be inconsistent with the many findings that the processing of audible and visible speech is described by the FLMP law in which the two modalities are represented independently of one another.

In feedforward convergence, the activation from the two modalities is sent to a third location that combines their inputs. As illustrated in Fig. IB, the neural activation from the auditory and visible speech activates a third location that is sensitive to the inputs from both modalities. An important set of observations from single cell recordings in the cat could be interpreted in terms of convergent integration (Ref. b). Convergent integration offers a potential implementation of the FLMP.

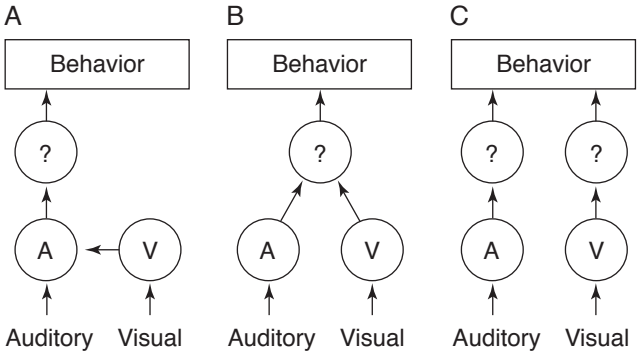

Fig. I. Representations of three possible mechanisms for the integration of auditory and visual speech. (A) sensory penetration; (B) simple feedforward convergent integration; and (C) non-convergent temporal integration. (See text for details.)

In non-convergent temporal integration (Fig. IC), integration involves the combination of information from two or more remote regions of the brain. Corticocortical pathways (pathways that connect regions of the cortex) synchronize the outputs of these regions and enable them to feed forward, independently, but synchronously, to other areas (Refs c,d). This type of brain processing appears to be most consistent with the findings that an integrated percept can exist simultaneously with and independently of representations of the separate sources of information.

One limitation in distinguishing among these neurological alternatives by localizing specific sites for integration is that the auditory and visual sites are intertwined in the cortex. Neuroimaging techniques revealed that speechreading without auditory speech activated superior temporal sulcus (STS). Calvert proposed that the observed contribution of speechreading to the enhancement of activity in the auditory cortex could be subsequent to the integration of these two sensory streams in heteromodal regions proximal to the STS (Ref. e). Sams found a delay of processing visual speech relative to auditory speech (Ref. f). This raises the possibility that the auditory and visual signals are integrated first in the association cortex (close to and including the STS), and only then is this information fed back to the auditory speech areas. Perhaps it is at this point the phenomenal speech as being heard is created.

References

a Massaro, D.W. (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press

b Stein, B.E. and Meredith, M.A. (1993) *The Merging of the Senses*, MIT Press

c Liederman, J. (1995) A reinterpretation of the split-brain syndrome: implications for the function of corticocoertical fibers, in *Brain Asymmetry* (Davidson, R.J. and Hugdahl, K., eds), pp. 451–490, MIT Press

d Zeki, S. (1993) *A Vision of the Brain*, Blackwell Science

e Calvert, G.A., Brammer, M.J. and Iverson, S.D. (1998) Crossmodal identification *Trends Cognit. Sci.* 2, 247–253

f Sams, M. *et al.* (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex *Neurosci. Lett.* 127, 141–145

direction of activation only occurs after recognition is complete, and the activation is used in learning or updating memory.

A simple addition to TRACE is to include visual features as well as auditory ones[21,39]. This elaboration is consistent with the general scheme of integration models in which there is separate feature evaluation of the audible and visible sources of information. The important difference is that the TRACE model involves feedback. Activation of the phoneme level would in turn activate the feature level. Featural information in one modality would be sufficient to activate features in the other modality. One of the central differences between TRACE and the FLMP is the independence of the auditory and visual information at the evaluation stage. Although this is a major difference between TRACE and the FLMP, a stochastic version of the model has been demonstrated to make equivalent asymptotic predictions to the FLMP. More elaborate experimental tasks and manipulations involving the dynamics of information-processing are necessary to distinguish between the models[42,43].

A broad set of model tests confirms the robustness of the FLMP account of pattern recognition. We face the challenge that experimental psychology is plagued with variability in a variety of guises. Behavior is not deterministic in the sense that the same stimulus situation does not always lead to the same behavior. In our experiments, we give repeated tests of the same stimulus to estimate the likelihood of responses to the stimulus. This estimation has sampling variability, which makes exact prediction more difficult. For this reason, we have developed an absolute benchmark for goodness-of-fit, which provides a standard for determining a model's accuracy. We have also used different methods of fitting models to observed results. The FLMP also maintains its superiority when it and competing models are challenged with a set of broader tests and model-fitting procedures (see Ref. 15, Chap. 10).

Broadening the domain of inquiry

We have carried out experiments to broaden our domain of inquiry in several directions. These new results test a framework for understanding individual differences, allow a distinction between information and information-processing[15,44], and illuminate cross-linguistic differences. This research analyses the results of individual subjects because it is possible that average results of an experiment do not reflect the results of any individual making up that average. We have explored a broad variety of dimensions of individual variability in terms of the distinction between information and information-processing. These include (1) life-span variability, (2) language variability, (3) sensory impairment, (4) brain trauma, (5) personality, (6) sex differences, and (7) experience and learning. The results of experiments with native English, Spanish,

Japanese, and Dutch talkers showed substantial differences in performance across the different languages[15,45,46]. The application of the FLMP indicated that these differences could be completely accounted for by information differences with no differences in information-processing. The differences that are observed are primarily the different response categories used by the different linguistic groups, which can be attributed to differences in the phonemic repertoires, phonetic realizations of the syllables, and phonotactic constraints in these different languages. In addition, talkers of different languages are similarly influenced by visible speech, with its contribution largest to the extent the auditory source is ambiguous. The details of these judgments are predicted by the FLMP, but not by competing models such as a single-channel model, auditory dominance, or categorical perception.

A second direction of our research concerns ecological variability, which refers to different perceptual and cognitive situations involving pattern recognition and to variations in the task itself[15]. Generally, we need to know to what extent the processes uncovered in the task of interest generalize across (1) sensory modalities, (2) environmental domains, (3) test items, (4) behavioral measures, (5) instructions, and (6) tasks. The processes involved in bimodal language processing, for example, might be revealed more readily by addressing these variables in addition to those traditionally manipulated. The belief is that the interactions with these variables will inform and constrain the kinds of processing mechanisms used to explain the basic observations (see Box 1).

Pursuing the question of whether our model of pattern recognition is valid across different domains, we examined how emotion is perceived by manipulating facial and vocal cues of a speaker[47]. The results shown in Fig. 3 indicate that participants use both the face and the voice to perceive emotion and the influence of one modality is greater when the other is ambiguous (see also Refs 48,49). Given that the FLMP fit the judgments significantly better than several alternative models, the perception of emotion appears to be well described by our theoretical framework[50]. Analogous to speech perception, we find a synergistic relationship between the face and the voice. A message communicated by both of the modalities is more informative than either one alone[15].

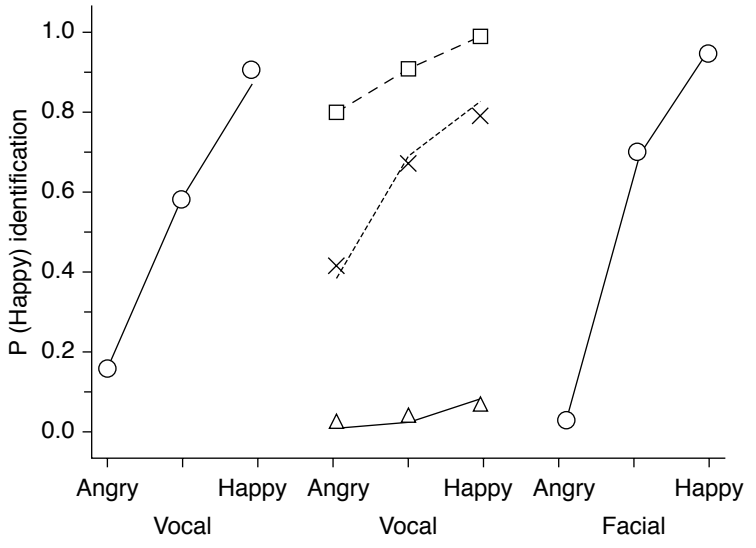

Fig. 3. Results of a study on the perception of emotion[47]. The observed proportion (data points) and FLMP predicted proportion (lines) of happy judgments as a function of the facial and vocal affect are shown for the two conditions and for their combination. The left plot shows performance for just the face and the right plot for just the voice. The middle plots show performance for the factorial combination of the two modalities. The influence of one modality was greater to the extent that the other was ambiguous. The FLMP fit the judgments significantly better than an additive model, which weakens theories based on an additive combination of modalities, categorical perception, and influence from only a single modality. (Modified from Ref. 47.)

The value of auditory–visual speech

There are several reasons why the use of auditory and visual information together is so successful, and why they hold so much promise for educational applications such as language tutoring (see Box 2). These include: (1) robustness of visual speech; (2) integration of the two modalities even though they are slightly asynchronous in time; (3) complementarity of auditory and visual speech; and (4) optimal integration of these two sources of information.

Empirical findings show that speechreading, or the ability to obtain speech information from the face, is robust. Research has shown that perceivers are fairly good at speechreading even when they are not looking directly at the talker's lips[51]. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when portions of the face are missing[52], when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer[15,53]. These findings indicate that speechreading is highly functional in a variety of nonoptimal situations.

Another example of the robustness of the influence of visible speech is that people naturally integrate visible speech with audible speech even when the temporal occurrence of the two sources is displaced by about a fifth of a second. Given that light and sound travel at different speeds and that the dynamics of their corresponding sensory systems also differ, a multimodal integration must be relatively immune to small temporal asynchronies. In several experiments, the relative onset time of the audible and visible sources was systematically varied[54,55]. The tests of formal models made it possible to determine when integration of audible and visible speech did occur. The FLMP gave the best description of the results, but only when the temporal arrival of the two sources of information was within 250 ms. This finding supports the conclusion that integration of auditory and visual speech is a robust process and is not easily precluded by offsetting the temporal occurrence of the two sources of information.

Complementarity of auditory and visual information simply means that one of the sources is most informative in those cases in which the other is weakest[15,52]. Two segments that are robustly conveyed in one modality tend to be relatively ambiguous in the other modality. For example, the difference between /ba/ and /da/ is easy to see on the face but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. The fact that two sources of information are complementary makes their combined use much more informative than would be the case if the two sources were non-complementary or redundant (Ref. 15).

Box 2. Applications in language training

Given our theoretical framework, there are potential applications for the development of computer-animated agents, who can serve as language tutors and as conversational characters in a variety of educational and human–machine domains. With respect to the ecological validity of our research findings, the analysis of individuals with hearing loss has confirmed many of the principles derived from studies of individuals with normal hearing (Refs a–c; H.W. Campbell, PhD thesis, University of Nijmegen, The Netherlands, 1974). The good description given by the FLMP to these data sets indicates that persons with hearing loss benefit greatly from visible speech.

Given the powerful contribution of speechreading, it follows that there is value in the technology and science of creating talking faces. With our completely animated, synthetic, talking head Baldi (Fig. I) we can control the parameters of visible speech and study its informative aspects. His speech is controlled by 33 parameters including: jaw rotation and thrust, horizontal mouth width, lip corner and protrusion controls, lower lip tuck, vertical lip position, horizontal and vertical teeth offset, tongue angle, width and length. Realism of the visible speech is measured in terms of its intelligibility to speechreaders. Experiments have shown that visible speech produced by the synthetic head, even in its adumbrated form, is almost comparable to that of a real human (Ref. d, Chapter 13).

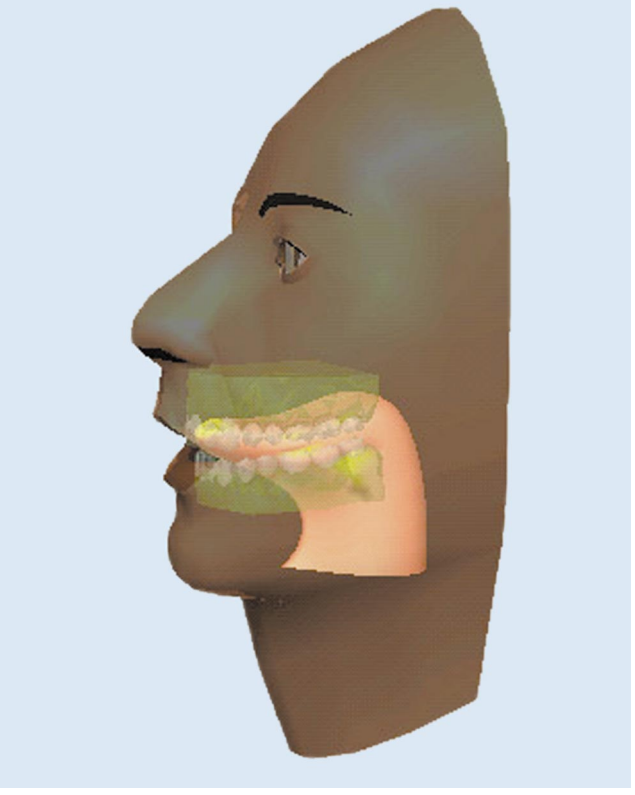

Fig. I. Illustration of computer-animated talking head, Baldi, with transparent skin to reveal the tongue, hard palate and teeth.

The pursuit of visible speech technology could be of great practical value in many spheres of communication (Ref. e). Children with hearing impairment, for example, require guided instruction in speech perception and production. However, as in other domains of learning, a great deal of time on task is necessary. Furthermore, many of the subtle distinctions among speech segments are not visible on the outside of the face, making it difficult for the human speech therapist. The animated talker solves both of these problems. Because it is freely available on a modest PC, it is essentially always available (Ref. f). In addition, the synthetic talker's skin can be made transparent in order to show the internal articulators that are normally hidden (Ref. g). Language training more generally could utilize this technology, as in the learning of non-native languages and in remedial instruction with language-disabled children. Speech therapy during the recovery from brain trauma could also benefit. Finally, we expect that children with reading disabilities could profit from interactions with our talking head.

References

a Walden, B. *et al.* (1990) Visual biasing of normal and impaired auditory speech perception *J. Speech Hear. Res.* 33, 163–173

b Massaro, D.W. and Cohen, M.M. (1999) Speech perception in perceivers with hearing loss: synergy of multiple modalities *J. Speech Lang. Hear. Res.* 42, 21–41

c Erber, N.P. (1971) Auditory, visual, and auditory–visual recognition of consonants by children with normal and impaired hearing *J. Speech Hear. Res.* 15, 413–422

d Massaro, D.W. (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press

e Massaro, D.W. *et al*. Developing and evaluating conversational agents, in *Human Performance and Ergonomics: Handbook of Perception and Cognition* (Vol. 17) (Hancock, P.A., ed.), pp. 173–194, Academic Press (in press)

f Cole, R. *et al*. (1999) New tools for interactive speech and language training: using animated conversational agents in the classrooms of profoundly deaf children, in *Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*, pp. 45–52 (London, UK, April 1999), University College London

g Cohen, M.M., Beskow, J. and Massaro, D.W. (1998) Recent developments in facial animation: an inside view, in *Proc. Audit. Visual Speech Percept.* (Terrigal, Australia, 1998), pp. 201–206, Causal Productions, Rundle Mall

The final characteristic of auditory–visual speech perception is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner[15]. There are many possible ways to treat two sources of information: we could use only the most informative source, average the two sources together, or integrate them in such a fashion in which both sources are used but that the least ambiguous source has the most influence. Perceivers in fact integrate the information available from each modality to perform as efficiently as possible. A wide variety of empirical results have been described by the FLMP, which describes an optimally efficient process of combination.

Conclusion

Understanding multimodal speech perception has created a productive interdisciplinary endeavor that should be considered ideal for cognitive science. The empirical findings and tests of quantitative models provide a plethora of constraints on theoretical explanations and possible neurological mechanisms. We have learned a great deal, specifically about speech perception and language processing, and more generally about how we cope with many different sources of information in pattern recognition. I look forward to future innovations that will facilitate the explanation of our impressive linguistic behavior and its incumbent phenomenology.

References

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

2 Fodor, J.A. (1983) *Modularity of Mind*, Bradford Books/MIT Press

3 Bertelson, P. and Radeau, M. (1981) Cross-modal bias and perceptual fusion with auditory–visual spatial discordance *Percept. Psychophys.* 29, 578–584

4 Bertelson, P. (1994) The cognitive architecture behind auditory–visual

interaction in scene analysis and speech identification *Cahiers de Psychologie Cognitive* 13, 69–75

5 Radeau, M. (1994) Auditory–visual spatial interaction and modularity. *Cahiers de Psychologie Cognitive* 13, 3–51

6 Fisher, B. (1992) Integration of visual and auditory information in the perception of speech events *Dissert. Abstr. Int*. 52, 3324B

7 Massaro, D.W. (1992) Broadening the domain of the fuzzy logical model of perception, in *Cognition: Conceptual and Methodological Issues* (Pick, H.L., Jr, Van den Broek, P. and Knill, D.C., eds), pp. 51–84, American Psychological Association

8 Warren, R.M. (1970) Perceptual restoration of missing speech sounds *Science* 167, 392–363

9 Samuel, A.G. (1981) Phonemic restoration: insights from a new methodology *J. Exp. Psychol. Gen*. 110, 474–494

10 Frost, R., Repp, B.H. and Katz, L. (1988) Can speech perception be influenced by simultaneous presentation of print? *J. Mem. Lang*. 27, 741–755

11 Massaro, D.W., Cohen, M.M. and Thompson, L.A. (1988) Visible language in speech perception: lipreading and reading *Visible Lang*. 1, 8–31

12 Massaro, D.W. (1998) Illusions and issues in bimodal speech perception, in *Proc. Audit. Visual Speech Percept*. (Terrigal, Australia), pp. 21–26, Causal Productions, Rundle Mall

13 Calvert, G.A., Brammer, M.J. and Iverson, S.D. (1998) Crossmodal identification *Trends Cognit. Sci*. 2, 247–253

14 Massaro, D.W. (1987) *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*, Erlbaum

15 Massaro, D.W. (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press

16 Dennett, D.C. (1991) *Consciousness Explained*, Little, Brown & Co.

17 Blumstein, S.E. (1986) On acoustic invariance in speech, in *Invariance and Variability in Speech Processes* (Perkell, J.S. and Klatt, D.H., eds), pp. 178–197, Erlbaum

18 Blumstein, S.E. and Stevens, K.N. (1979) Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants *J. Acoust. Soc. Am*. 66, 1001–1017

19 Diehl, R.L. and Kluender, K.R. (1989) On the objects of speech perception *Ecol. Psychol*. 1, 121–144

20 Sekiyama, K. and Tohkura, Y. (1991) McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility *J. Acoust. Soc. Am*. 90, 1797–1805

21 Sekiyama, K. and Tohkura, Y. (1993) Inter-language differences in the influence of visual cues in speech perception *J. Phonet.* 21, 427–444

22 Cutting, J. E. *et al*. (1992) Selectivity, scope, and simplicity of models: a lesson from fitting judgments of perceived depth *J. Exp. Psychol. Gen*. 121, 364–381

23 Massaro, D.W. (1988) Ambiguity in perception and experimentation *J. Exp. Psychol. Gen*. 117, 417–421

24 Massaro, D.W. and Cohen, M.M. (1993) The paradigm and the fuzzy logical model of perception are alive and well *J. Exp. Psychol. Gen*. 122, 115–124

25 Liberman, A.M. (1996) *Speech: A Special Code*, MIT Press

26 Mattingly. I.G. and Studdert-Kennedy, M., eds (1991) *Modularity and the Motor Theory of Speech Perception*, Erlbaum

27 Robert-Ribes, J. *et al*. (1995) Expliting sensor fusion architectures and stimuli complementarity in AV speech recognition, in *Speechreading by Humans and Machines* (Stork, D.G. and Hennecke, M.E., eds), pp. 193–210, Springer-Verlag

28 Robert-Ribes, J., Schwartz, J-L. and Escudier, P. (1995) A comparison of models for fusion of the auditory and visual sensors in speech perception *Artif. Intell. Rev*. 9, 323–346

29 Massaro, D.W. and Stork, D.G. (1998) Speech recognition and sensory integration *Am. Sci*. 86, 236–244

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

31 Best, C.T. (1995) A direct realist view of cross-language speech perception, in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (Strange, W., ed.), pp. 171–204, York Press, Baltimore

32 Oden, G.C. and Massaro, D.W. (1978) Integration of featural information in speech perception *Psychol. Rev*. 85, 172–191

33 Massaro, D.W. and Friedman, D. (1990) Models of integration given multiple sources of information *Psychol. Rev*. 97, 225–252

34 Mesulam, M-M. (1994) Neurocognitive networks and selectively distributed processing *Rev. Neurol. (Paris)* 150, 564–569

35 Green, D.M. and Swets, J.A. (1966) *Signal Detection Theory and Psychophysics*, John Wiley & Sons

36 Massaro, D.W. (1996) Bimodal speech perception: a progress report, in *Speechreading by Humans and Machines* (Stork, D.G. and Hennecke, M.E., eds), pp. 79–101, Springer-Verlag

37 Campbell, C.S. and Massaro, D.W. (1997) Perception of visible speech: influence of spatial quantization *Perception* 26, 627–644

38 McClelland, J.L. and Elman, J.L. (1986) The TRACE model of speech perception *Cognit. Psychol*. 18, 1–86

39 Campbell, R. (1988) Tracing lip movements: making speech visible *Visible Lang*. 22, 32–57

40 O'Reilly, R.C. (1998) Six principles for biologically based computational models of cortical cognition *Trends Cognit. Sci*. 2, 455–462

41 White, E.L. (1989) *Cortical Circuits: Synaptic Organization of the Cerebral Cortex: Structure, Function, and Theory*, Birkhauser

42 McClelland, J.L. (1991) Stochastic interactive processes and the effect of context on perception *Cognit. Psychol*. 23, 1–44

43 Massaro, D.W. and Cohen, M.M. (1991) Integration versus interactive activation: the joint influence of stimulus and context in perception *Cognit. Psychol*. 23, 558–614

44 Grant, K.W., Walden, B.E. and Seitz, P.F. (1998) Auditory–visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory–visual integration *J. Acoust. Soc. Am*. 103, 2677–2690

45 Massaro, D.W. *et al*. (1993) Bimodal speech perception: an examination across languages *J. Phonet*. 21, 445–478

46 Massaro, D.W., Cohen, M.M. and Smeele, P.M.T. (1995) Cross-linguistic comparisons in the integration of visual and auditory speech *Mem. Cognit*. 23, 113–131

47 Massaro, D.W. and Egan, P.B. (1996) Perceiving affect from the voice and the face *Psychonomic Bull. Rev*. 3, 215–221

48 Hess, U., Kappas, A. and Scherer, K.R. (1988) Multichannel communication of emotion: synthetic signal production, in *Facets of Emotion: Recent Research* (Scherer, K.R., ed.), pp. 161–182, Erlbaum

49 De Gelder, B. *et al*. The combined perception of emotion from voice and face: early interaction revealed by electric brain responses *Neurosci. Lett*. (in press)

50 Campbell, C.S., Schwartzer, G. and Massaro, D.W. Face perception: an information-processing framework, in *Computational, Geometric, and Process Perspectives on Facial Cognition* (Wenger, M.J. and Townsend, J.T., eds), Erlbaum (in press)

51 Smeele, P.M.T. *et al*. (1998) Laterality in visual speech perception *J. Exp. Psychol. Hum. Percept. Perform*. 24, 1232–1242

52 Summerfield, Q. (1987) Some preliminaries to a comprehensive account of audio–visual speech perception, in *Hearing by Eye: The Psychology of Lip-Reading* (Dodd, B. and Campbell, R., eds), pp. 3–51, Erlbaum

53 Jordan, T.R. and Sergeant, P.C. (1998) Effects of facial image size on visual and audio–visual speech recognition, in *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory–Visual Speech* (Dodd, B., Campbell, R. and Burnham, D., eds), pp. 155–176, Psychology Press/Erlbaum

54 Massaro, D.W. and Cohen, M.M. (1993) Perceiving aysnchronous bimodal speech in consonant–vowel and vowel syllables *Speech Commun*. 13, 127–134

55 Massaro, D.W., Cohen, M.M. and Smeele, P.M.T. (1996) Perception of asynchronous and conflicting visual and auditory speech *J. Acoust. Soc. Am*. 100, 1777–1786