Speech Recognition and Sensory Integration

A 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech

Dominic W. Massaro and David G. Stork

Consider what goes through your mind when you attempt to identify an apple's variety—Granny Smith, Golden Delicious, McIntosh or Fuji, for example. You see the apple's shape and color, feel its weight and the smoothness of the skin, hear the snap as you take a bite, feel the firmness and texture in your mouth, and, of course, savor the taste and smell. Any one of these clues taken alone might prove insufficient for classifying the apple. But together, matched with your past experience of each variety, they enable you to accurately classify the apple.

In this decision you may deliberately combine the input from several senses, but in many cases, the brain automatically combines sensory information. How can the brain bring together information from such diverse senses to permit accurate categorization? For that matter, how do we bring together information sources within a *single* sensory modality, such as the visual perception of surface texture and of color? What strategies have evolved in the human brain to solve this "sensory integration" problem, what can we learn from studying this solution, and how can we apply such lessons to designing and programming machines that categorize from perceptual input?

One of us (Massaro) has worked for many years on understanding how people integrate sensory information in speech perception, especially when lipreading is a source of information; the other (Stork) has been building lipreading machines that can decode speech. In such machines, the designer must specify the method for integrating visual and acoustic information. Despite the difference in our disciplines, we have found a satisfying convergence of results that has led us to the identification of a law of human information processing as well as to the creation of a new class of accurate automatic speech-recognition systems. Both results rely on the same 240-year-old theorem in statistics used widely for prediction in science.

Human beings are social animals, and despite the onslaught of advances in telecommunications—from the telegraph to the Internet—we prefer our messages embodied with a view of our correspondent, as in the recent technology called videoconferencing. The image of a talker's face provides not only a window into her emotions and motivations, but also important information used in understanding speech. People with normal hearing and sight use both modalities to understand speech, although we are generally not aware of the visual component. For example, when you have difficulty understanding speech in a very noisy room, you will do better by watching the speaker's face more closely. Some people claim they can understand the dialogue on TV better with their glasses on. Research suggests that, whenever possible, people use and integrate both sight and sound for speech recognition.

The contributions of sight and sound become more obvious when people must use only one modality. Children blind from birth seem to have greater difficulty learning certain subtle distinctions in acoustic speech than sighted children, and highly skilled lipreaders can understand much of the speech in a movie—even in a silent movie. The deaf watch the whole face—mouth, tongue, teeth, jaw, even eyebrows—to aid understanding, hence the more comprehensive term "speechreading."

The McGurk Effect

The most striking demonstration of the combined (bimodal) nature of speech understanding appeared by accident. Harry McGurk, a senior developmental psychologist at the University of Surrey in England, and his research assistant John MacDonald were studying how infants perceive speech during different periods of development. For example, they placed a videotape of a mother talking in one location while the sound of her voice played in another. For some reason, they asked their recording technician to create a videotape with the audio syllable "ba" dubbed onto a visual "ga." When they played the tape, McGurk and MacDonald perceived "da." Confusion reigned until they realized that "da" resulted from a quirk in human perception, not an error on the technician's part. After testing children and adults with the dubbed tape, the psychologists reported this phenomenon in a 1976 paper humorously titled "Hearing Lips and Seeing Voices," a

Dominic W. Massaro is chair and professor in the Department of Psychology at the University of California, Santa Cruz. He holds a Ph.D. in psychology from the University of Massachusetts and previously taught at the University of Wisconsin at Madison. Past president of the Society for Computers in Psychology, he is currently book-review editor of the American Journal of Psychology *and co-editor of* Interpreting. *His research interests include perception, memory, cognition, learning and decision-making. A book on his current work,* Perceiving Talking Faces, *has just been published by MIT Press. David G. Stork, who earned his Ph.D. in physics at the University of Maryland, is chief scientist and head of the Machine Learning and Perception Group at Ricoh Silicon Valley and also consulting associate professor and visiting scholar in psychology at Stanford University. He has worked extensively in human and machine pattern recognition and led the first international workshop on speechreading in humans and machines. He edited* HAL's Legacy: 2001's Computer as Dream and Reality *(MIT Press), which was recently translated into Japanese. Address for Massaro: Department of Psychology, UC-Santa Cruz, Santa Cruz, CA 95064. Internet: massaro@fuzzy.ucsc.edu.*

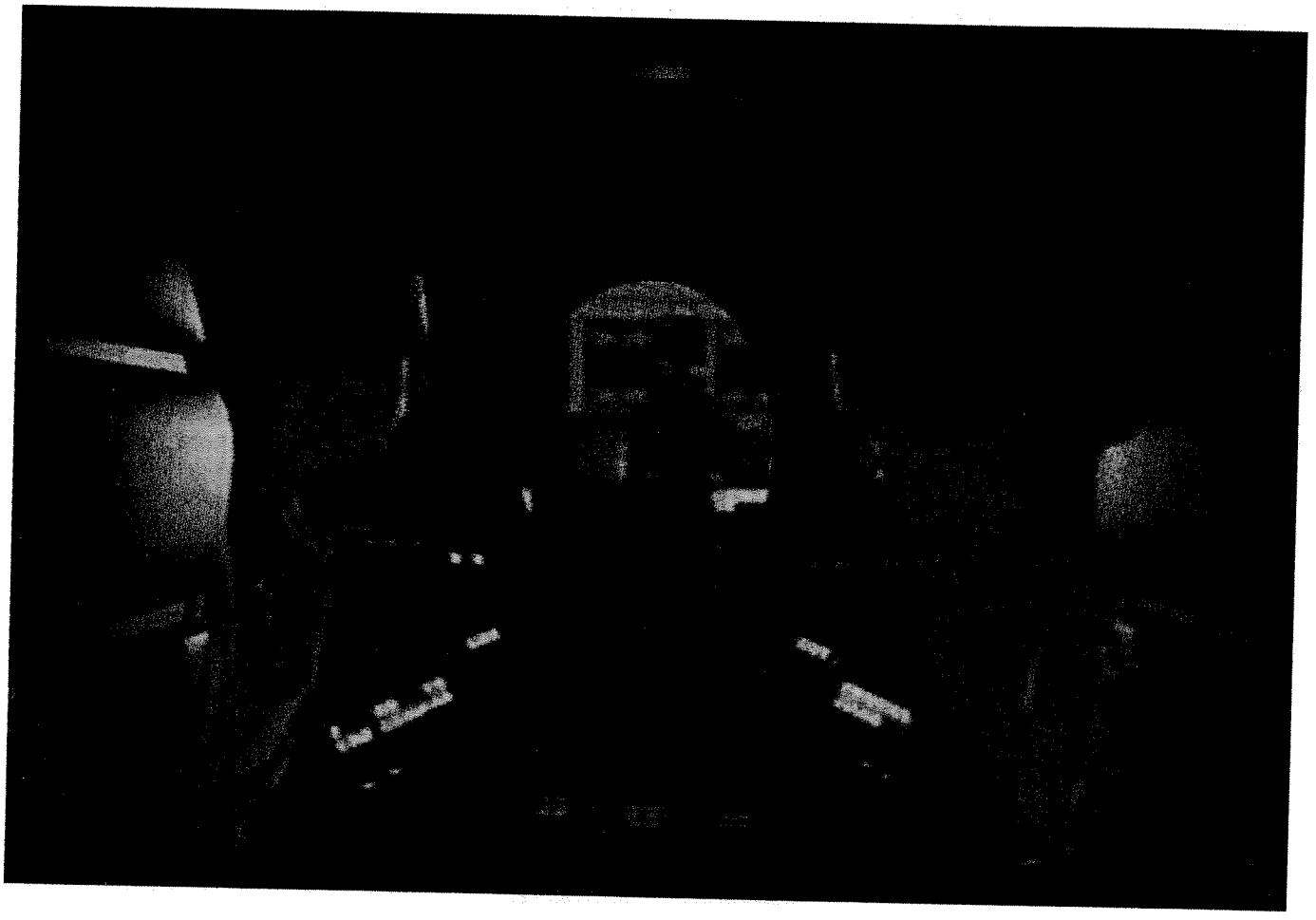


Figure 1. In the 1968 Stanley Kubrick film *2001: A Space Odyssey*, the spaceship's computer not only controls the ship and converses with the crew, but in a crucial plot twist also reads lips. Here (in a clip from the popular home-video version) HAL has begun to act strangely, and astronauts Dave and Frank plot to turn the computer off while ensconced in an escape pod. But HAL reads their conversation through the glass. Thirty years after the movie, computerized speech recognition is on the verge of becoming a reality. The authors find that similar algorithms explain how visual and auditory information may be integrated by the senses and how computers can use visual information in speechreading.

landmark in the field of human sensory integration. This audio-visual illusion has become known as the McGurk effect or McGurk illusion.

In the archetypal example of the McGurk effect, the brain seems to simply combine sound and sight. The auditory perception of the *front* consonant in "ba" (vocal tract closed at the lips) and the visual perception of the *back* consonant in "ga" (closure at the back of the throat) yield an integrated perception of the *middle* consonant: "da" (closure behind the alveolar ridge). But the McGurk effect shows up in more than just single syllables. If you make a videotape with the audio nonsense sentence "My bab pop me poo brive," dubbed onto the video nonsense sentence "My gag kok me koo grive," most viewers perceive "My dad taught me to drive." Experimental subjects generally cannot decode the video by itself. Given only the audio, they clearly hear "My bab pop me poo brive." Any model of sensory integration in human beings must explain such rich and robust McGurk phenomena.

Although these demonstrations seem surprising at first, a simple principle explains the McGurk effect. Perceivers tend to interpret an event in a way that is most consistent with all the sensory information—in speechreading, both sight and sound. The syllable "ba" sounds somewhat similar to "da." Analogously, visual "ga" looks quite like visual "da." Visual "ga," on the other hand, does not fit with auditory "ba," so "ba" would not be a good interpretation of the speech event. Visual "da" matches both the auditory and visual inputs reasonably and therefore wins the competition for the best interpretation. Similar processes operate in the sentence example, with the added dimension that subjects tend to think about the meaning of a sentence.

Speech segments—such as the *b* in the syllable "bi"—consist of several features that assist us in distinguishing one segment from the other. The voicing of some segments proves essentially impossible to see but fairly easy to hear. For example, "bi" and "pi" are visually indistinguishable. (They belong to the same "viseme" class, so-called by analogy to acoustical "phonemes.") Nevertheless, "bi" and "pi" are easily distinguished acoustically using a feature called *voice onset time*—the delay between the initial burst sound and the onset of vibration of vocal folds. The voice onset time in "pi" is noticeably longer than in "bi."

The segments "mi" and "ni" sound quite similar; listeners frequently confuse them, especially in noisy locations. "Mi" and "ni" have the same voicing and nasality (extra resonances from the nasal cavity), differing only in where the speaker closes the focal tract with the lips or tongue or teeth, the place of articulation. Nevertheless, those utterances are particularly easy to distinguish by

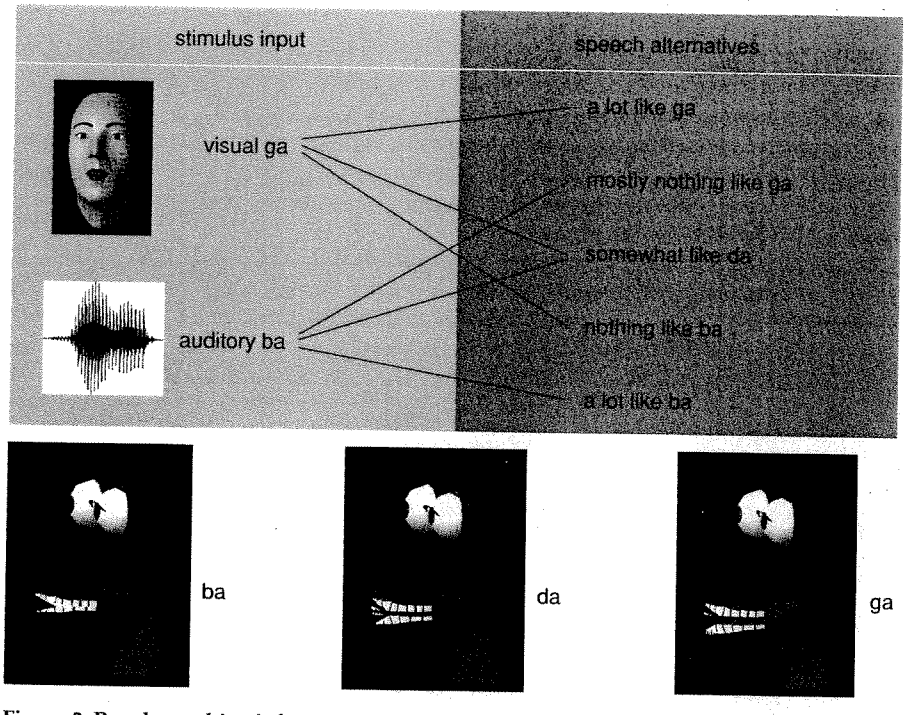


Figure 2. People combine information from different modalities to decode speech. When visual and auditory inputs conflict, the brain still comes up with a best-fit solution. Developmental psychologist Harry McGurk and research assistant John MacDonald discovered this phenomenon while studying how infants perceive speech. When a technician made a videotape in which a voice saying "ba" was dubbed onto an image of a person saying "ga," McGurk and MacDonald heard "da." Although they first thought the tape mistakenly contained "da" in both modalities, they soon learned the truth—and that they could replicate the illusion, now called the McGurk effect, at will. One of the authors (Massaro) has used variations of the McGurk effect to study how the brain processes information to recognize speech. His recent work uses versions of "Baldi," a computer simulation, developed in collaboration with Michael M. Cohen, that can combine audio and video syllables in any combination. Stripping away Baldi's skin *(bottom)* reveals how the formation of a syllable might lead to the overlapping perceptions shown at upper right. The auditory perception of the front consonant in "ba" (vocal tract closed at the lips) seems to be integrated with the visual perception of the back consonant in "ga" (closure at the back of the throat) to yield a perception of the middle consonant, or the syllable "da" (closure behind the alveolar ridge).

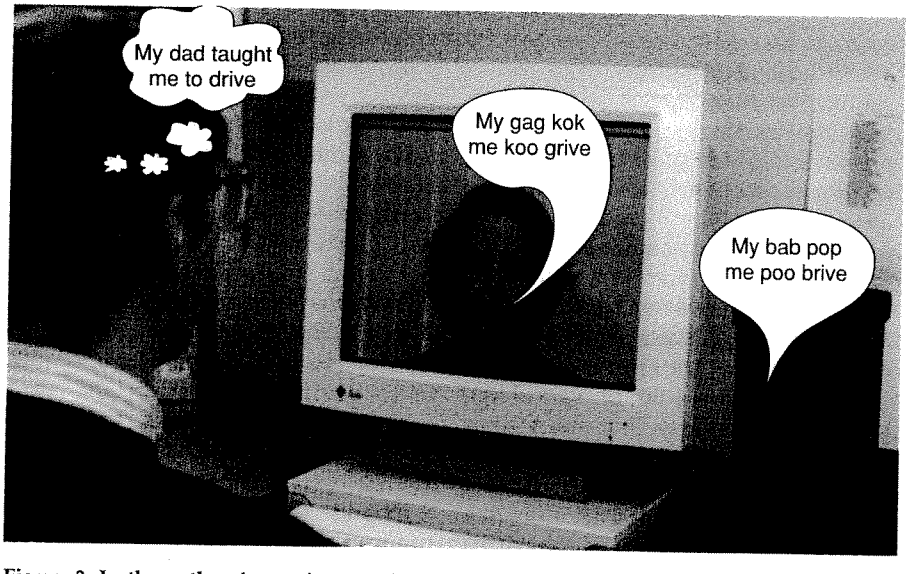


Figure 3. In the authors' experiments, the McGurk effect works just as well for whole sentences as for isolated syllables. Presented with a voice saying "My bab pop me poo brive" dubbed to a face mouthing "My gag kok me koo grive," most subjects think they hear "My dad taught me to drive." Eliminate the video and subjects hear nonsense. (Photograph courtesy of the authors.)

sight—in "mi" the lips close at onset, whereas in "ni" they do not. One of the key properties of bimodal speech to emerge from such analysis is that of complementarity: Features that are *hardest* to distinguish acoustically are the *easiest* to distinguish visually, and vice versa.The sensory integration of auditory and visual information in speech perception and the complementarity between these modalities shows clearly in experiments that independently vary auditory and visual information.

Varying Two Inputs

To study bimodal speech perception we usually vary independently two sources of information. We decorrelate the two modalities with experiments using a so-called expanded factorial design.

The grid in Figure 4 illustrates the design used in one study in which we presented four auditory syllables in combinations with each of four visible syllables. In addition, we presented each of the syllables unimodally. Each participant (human or machine) identified separately just the auditory syllables, just the visible syllables, and all combinations of the auditory and visual syllables. We dubbed the auditory syllable onto the visual syllable to maintain the temporal alignment found in the natural syllable. This type of study can determine how the separate sources of information combine to achieve speech perception. This experimental design provides a strong test of sensory integration because it examines both unimodal and bimodal conditions. Any explanation of how the brain integrates sensory information must describe the relationship between unimodal and bimodal performance.

The four syllables differ primarily in place of articulation and nasality. The syllables "bi" and "mi" are articulated in the same place in the front of the mouth, making any difference difficult to see. Similarly, "di" and "ni" are both said with an open mouth. Differences in place of articulation are easy to see but difficult to hear, for example, compare "mi" and "ni." Nasality, on the other hand, is easy to hear but hard to see, an example of *complementarity*. The nasality associated with "mi" makes this syllable easily distinguished from "bi." Similarly, nasality makes it easy to distinguish "ni" and "di," both of which require the tongue to touch the palate. Diagrams that show the probability of each category response for a particular input utterance, called per-

ceptual confusion matrices, best reveal the complementary nature of bimodal speech recognition.

The confusion matrix in Figure 5 shows that subjects sometimes confused "bi" and "di," which sound similar, when they received only auditory input. But adding visual information made the syllables easy to distinguish because of the difference in lip closure. The opposite holds as well: Subjects confuse the consonants "bi" and "mi" when they are presented solely visually, but rarely confuse these consonants in a bimodal presentation. In unimodal presentations, acoustic recognition is more accurate than visual among hearing subjects who have not been trained in lipreading.

Complementarity

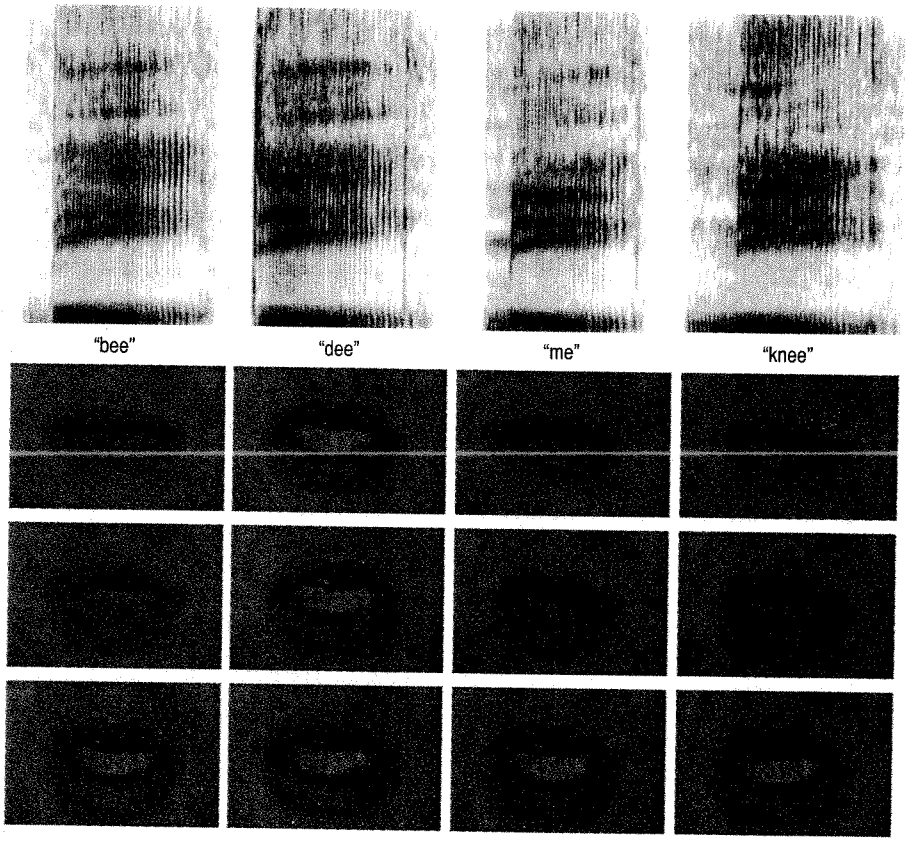
These perceptual-confusion matrices reveal complementarity of the auditory and visual components of speech. Not only do audible and visible speech provide two independent sources of information, but each also provides strong information where the other is weak. No fundamental theory describes the evolution of complementarity in bimodal speech. We can say only that informative acoustic information arises in visually inaccessible regions of the vocal tract (such as the vocal folds and glottis) whereas acoustically challenging information (such as place of articulation) correlates with visually obvious features such as lip closure or tongue placement.

Complementarity confers two important research benefits. First, it increases speechreading's value as a system in which to probe the general properties of sensory integration in human beings. The bimodal effects are most pronounced when both modes are fallible for some features, but one mode is accurate while the other is not. Second, complementarity is especially welcome in automated speechreading systems because the visual information improves recognition of those utterances most difficult for current acoustic recognition systems to discriminate. Furthermore, strong formal similarities appear between bimodal speech and electronic communication systems in which signals in one channel help to eliminate the effects of noise or transmission errors in the other channel (paired error-correcting codes).

The results of experiments producing confusion matrices help determine exactly how sensory integration takes place, or whether it takes place at all. The two sources may *not* be integrated for distinguishing a particular feature. Instead one source may serve to recognize one feature and the other source to recognize the other feature. McGurk and MacDonald (1976) took this stand in their first publication of the illusion, claiming that visual input informed the perception of place of articulation, and auditory information dominated the perception of voicing.

But this conclusion is amenable to testing. If the visual modality dominates the perception of place of articulation, then inconsistent auditory information should not confuse subjects when they attempt to recognize syllables that differ visually but sound similar. Visual "di" paired with auditory "bi" should be no more difficult to distinguish than visual "di" paired with auditory "di." However, experiments show that the dominant modality for a given feature does not entirely dominate the judgments. For example, the likelihood of a subject perceiving "di" given visual "di" paired with auditory "di" was significantly larger in experiments than the likelihood of a "di" judgment given visual "di" paired with auditory "bi." Other results in Figure 6 follow the same pattern.

Such evidence demonstrates convincingly that human brains do integrate information, but how? And can we program speechreading machines to do the same?



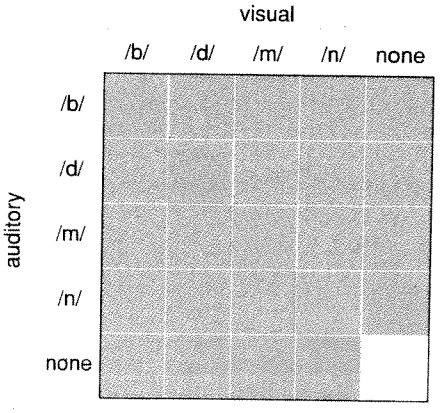


Figure 4. Four syllables—written phonetically /bi/, /di/, /mi/ and /ni/—differ in articulation and nasality, as evidenced by variations in the spectrograms above. Differences in the position of the lips at voice onset, in the middle of articulation and in the pronunciation of the vowel *(top, middle and bottom mouth images)* help distinguish vocally similar pairs. Author Massaro uses an expanded factorial design *(right)* in which human subjects respond to every combination of sound and sight for four syllables. The responses are mapped onto a grid called a confusion matrix *(see Figure 5)*. (Images courtesy of The Johns Hopkins University.)

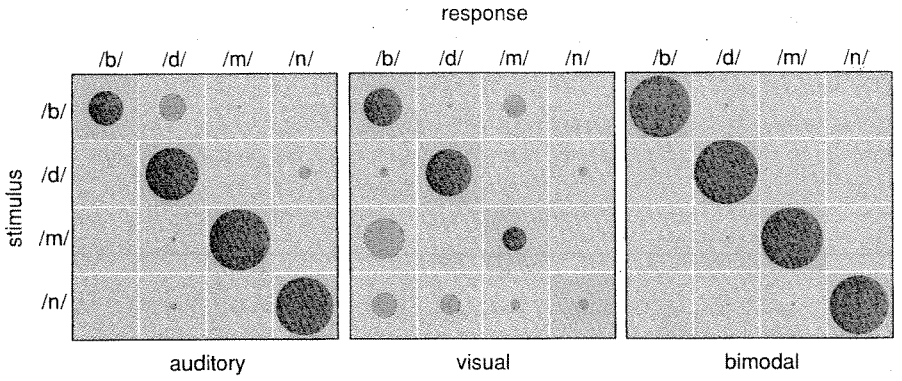


Figure 5. Auditory and visual speech information appear to complement each other in speech recognition. Where auditory information fails, visual information comes to the fore. The confusion matrix shown here reveals the complementary nature of bimodal speech. The size of each circle represents, from human experimental data, the probability of a response given a particular stimulus. For example, the consonant "d" is commonly perceived when "b" is presented solely acoustically, but visual information corrects this perception. In a bimodal presentation, syllables are seldom confused. Confusion matrices for human subjects and for the speechreading machines discussed in the text appear quite similar, although people outperform machines.

How People Do It

The experiments described above provide an impressive demonstration of the brain's integration of auditory and visual data in recognizing speech patterns. To fully understand this integration and to use that understanding to build speechreading machines requires a model of how the integration takes place. Investigators in several fields have tested many models against experimental results like these, and, perhaps surprisingly, a clear victor emerges. Even more impressive, the winning model is mathematically identical to a proposal more than 200 years old. The English amateur mathematician Reverend Thomas Bayes wrote his theorem in 1761, the year of his death. Published posthumously, the Bayes theorem could fairly be considered the foundation of much of the predictive ability of the sciences, finding use in a wide range of topics from medical diagnosis to predicting weather to putting a spacecraft in orbit.

Simply put, Bayes's theorem offers a way to statistically quantify the probability of one hypothesis among several being true, and to update that probability as new data come in. Simple statistical problems—coin flips or dice rolling, for example—yield to probability analysis based on counting. Bayes's theorem applies to problems in which scientists cannot count frequencies or repeat trials.

For example, this winter the authors—and many other central Californians—suffered from rains, floods and power outages caused by El Niño. Earlier in the summer of 1997, forecasters warned of the high likelihood of an extremely wet winter. They based this prediction on several observations, such as the warmer-than-usual ocean currents and air-temperature patterns in various parts of the world. Forecasters combined all of these observations using the simple recipe proposed by Bayes to arrive at their prediction.

In speechreading, each time a brain decodes visual and auditory information to make a decision, it must choose between several competing hypotheses to answer the question "What is the syllable I heard and saw?" We choose a syllable with a high likelihood of being the right one based on the available visual and auditory data.

When we recognize speech, we evaluate and then integrate sounds and sights, which provides a psychological value indicating the degree to which available data support a particular hypothesis—for example, should the brain categorize the perceived syllable as "bi" or "di?" Next, a decision-making process called a relative-goodness-of-match rule applies. This compares the "score" for one syllable against the combined scores for all others. This rule predicts that the probability of a particular syllable being the correct one, given certain auditory and visual input, equals the total support for a particular syllable divided by the sum of the goodness-of-match values of all alternative syllables.

We can apply the Bayes theorem to speech data integration with a bit of simple notation. The probability that a perceived syllable falls into a speech category (c) given the acoustic evidence (A) is denoted $P(c \mid A)$. We can state this probability in terms of the acoustic evidence given the category, the probability $P(A \mid c)$, the probability of the category c, and the sum of the probabilities of observing all possible categories—in this case the total probability of finding the acoustic evidence A:

$$P(c \mid A) = \frac{P(A \mid c)P(c)}{\text{sum}_A}$$

Exactly the same logic holds for the probability of a category c given the visual evidence V:

$$P(c \mid V) = \frac{P(V \mid c)P(c)}{\text{sum}_V}$$

The desired probability given evidence from both modalities, $P(c \mid A \,\&\, V)$, also arises from Bayes's theorem. If A and V are conditionally independent—that is, if $P(A \,\&\, V \mid c) = P(A \mid c)\,P(V \mid c)$—Bayes's theorem can yield the optimal sensory-integration scheme:

$$\frac{P(c \mid A)P(c \mid V)P(c)}{\text{sum}_{AV}}$$

Assigning a perceived syllable to a category can be seen as a pattern-recognition process. Most pattern-classification techniques employ a statistical method called *discriminant functions*, which take an input pattern and assign to each candidate category a numerical value or score. The perceived pattern can then be assigned to the category with the highest score. A Bayes discriminant function insures the minimum classification error. For the most accurate classification overall, such discriminant functions should relate simply to the probability that the input pattern fits the category in question. The problem of sensory integration comes down to specifying the appropriate bimodal discriminant function from the component ones, that is, computing $P(c \mid A \,\&\, V)$ based on $P(c \mid A)$ and $P(c \mid V)$. This is precisely what Bayes's theorem accomplishes.

More than Math

Mathematics alone cannot describe behavior. We find it valuable to couch the mathematical description as a set of operations taking the perceiver from the test stimulus to the interpretive response. The model, called a fuzzy-logical model of perception, relies on the assumptions that perceiving speech is fundamentally a pattern-recognition problem and that signals

correspond to probabilities. Within this framework, speech-pattern processing occurs in three stages: feature evaluation, feature integration and decision. These stages take place in succession, but also overlap in time.

These processes make use of prototypes stored in long-term memory. Source prototypes and input are then integrated to give an overall degree of support, *s*, for a given decision. (Did I perceive "bi" or "di?") Finally, the decision operation maps the outputs of integration into some response alternative, *R*. The response can take the form of a discrete decision or a rating of the likelihood of the alternative.

The inner ear transduces spoken language, making available to the brain a set of primitive characteristics, called sensory *cues* or *features*. As members of a linguistic community, we store in memory knowledge about what segments of speech occur in our language. Memory stores each segment as a prototype defined in terms of its ideal cues. When we receive some spoken language, we can compare its features to each prototype stored in memory.

In contrast to most models of speech perception, we assume the features provide continuous rather than discrete information. In this case, we can say that a particular feature fits a particular prototype to some degree on a continuous scale. The fit can also be interpreted as a subjective probability. The integration process combines the information from each feature to give an overall degree of fit to each prototype. Finally, the decision process makes some judgment based on the relative fit with the relevant prototypes.

The integration stage uses mathematics similar to Bayesian analysis. For example, consider the speech category "di." Physical input is transformed to a psychological value indicating the degree to which available speech data support the hypothesis that the correct category is "di." If one labels the auditory and visual values for support of prototype category "di" a and v, respectively, then the integrated total support, s("di"), for the alternative "di" would be given by the product of a and v: $s(\text{"di"}) = a \times v$.

Finally, as in Bayesian analysis, a decision operation follows that requires that, in this case, $P(\text{"di"} \mid A \ \& \ V)$ be determined. The fuzzy-logical model of perception follows a decision-making process called a *relative-goodness rule*. According to this rule, the probability of a particular categorization is assumed to be equal to the relative goodness-of-match. Mathematically, $P(\text{"di"} \mid A \ \& \ V)$ is equal to the total support (s) for "di" divided by the sum of the goodness-of-match values of all alternatives—just as in Bayesian analysis the multiplied probabilities are divided by the sum of all of the alternative probabilities.

How Machines Can Do It

It is easy to imagine the eyes and ears of a speechreading machine: a video camera and a microphone. Programming a brain, however, poses a larger problem. As in other kinds of computerized pattern recognition, speechread-

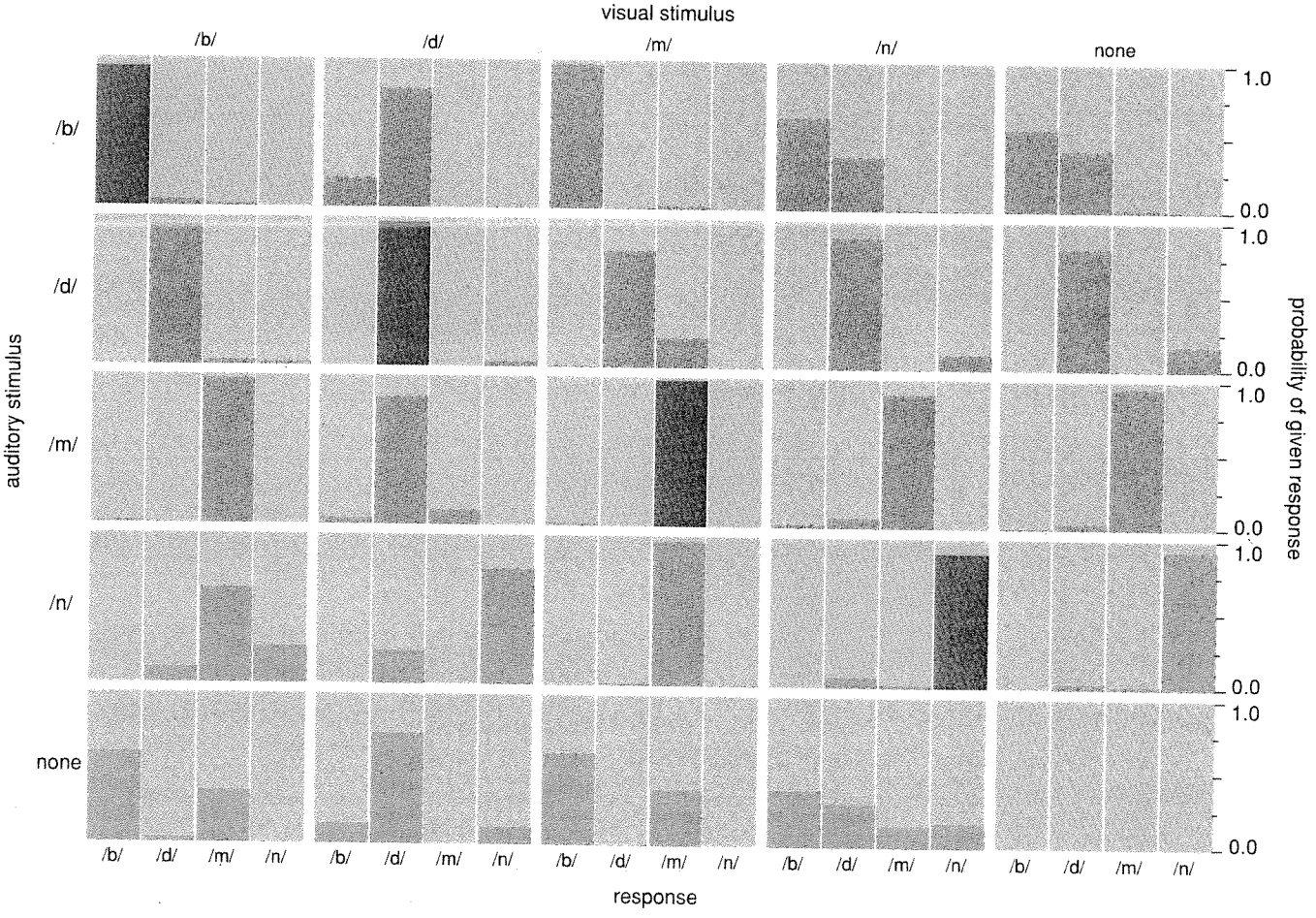


Figure 6. Not all "crossed signals" produce the same probability of a correct response. Full results from both unimodal and bimodal trials show that the brain appears to give greater weight to auditory information in some cases, but in others it relies heavily on visual cues. The dominant modality for a given feature does not entirely dominate the judgments.

ing machines must preprocess the raw sensor signals to extract features, and then analyze them to yield the syllable category. For audio, the computer must electronically filter the signal from the microphone, and break it down into component frequency bands. Other acoustic preprocessing might involve finding the overall loudness at each moment, detecting the presence of relatively repetitive tones (associated with voicing) and so forth. Because the *visual* components of speechreading systems are most novel, we shall not dwell on acoustic processing.

Video signals must first be broken into picture elements, or pixels. The first stages in processing the video signal include locating the face, mouth and chin. Our system locates the face by comparing the image of a talker to a "background" image taken when the talker is not present. Any changed pixels become candidates for further analysis. Next, the computer compares the color of each of these changed pixels to a sort of "universal skin color." It turns out that if you ignore the overall lightness, skin color is remarkably similar throughout the world—from Nigeria to China to Sweden. Pixels whose color closely matches this universal skin color remain as possible face pixels.

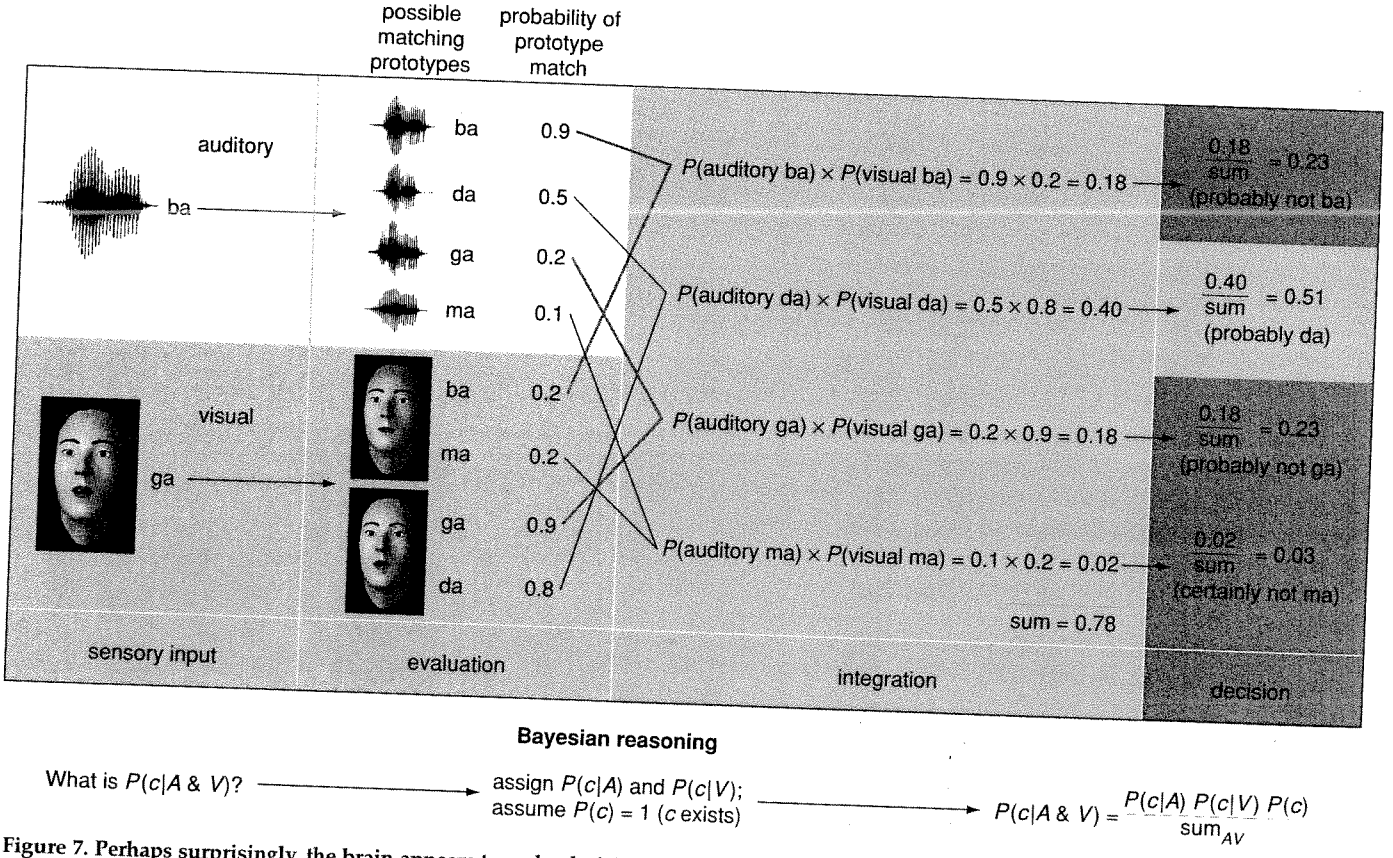
Because we are looking for a head, the program fits an oval surrounding the remaining candidate pixels, and considers the darkest pixels toward the top of the oval to be eyes. Some simple triangulation from these eye positions as well as the detection of motion (by standard "optic flow" techniques from computer vision) gives an excellent estimate of the position of the mouth. Once the computer finds the mouth, it fits a computer model called a "deformable lip template" to the image in each successive video frame *(Figure 9)*. The four arcs making up the template arise from several parameters, whose numerical values are adjusted until the template matches the lip shape in the video image. The final values of the parameters, determined for each video frame, make up the features used for visual classification.

The parameters governing the shape of the template, in particular the curvatures and separations of the component arcs, provide features used by the computer to determine the visual features of syllables. Some lip features, such as the tilt and overall position in the frame, carry no linguistic information and are used only for tracking lip location.

Other image-processing techniques give the position of the jaw and the visibility of the tongue, both of which are also used for recognition.

Crosstalk, or Not?

The schematic in Figure 8 shows audio and video clearly separated until the signals reach the computer speech integrator. But in people, might these systems interact before reaching the in-



Bayesian reasoning

What is $P(c|A \& V)$? → assign $P(c|A)$ and $P(c|V)$; assume $P(c) = 1$ (c exists) → $P(c|A \& V) = \frac{P(c|A)\,P(c|V)\,P(c)}{\text{sum}_{AV}}$

Figure 7. Perhaps surprisingly, the brain appears to make decisions during speech recognition following the 240-year-old statistical theorem scientists often use to evaluate the predictive power of hypotheses. Thomas Bayes posited a way to determine the probability that one hypothesis among many is true, given certain conditions. When experimental subjects receive auditory and visual input, they must effectively choose among several competing hypotheses, that is, they must posit an interpretation of their data. In the classic McGurk-effect example, a subject is presented with conflicting information—contrasting stimuli A (auditory) and V (visual). Each piece of information is evaluated with reference to a stored prototype to determine the degree to which the data support a given category (c)—for instance, $P(c|A)$, the probability the syllable fits category c given auditory information A. Auditory and visual probabilities are then integrated, and finally the relative-goodness-of-match rule normalizes the results by comparing the support for each syllable to the combined scores for all. This process shows how a subject might consider "da" the best fit in the classic problem described in Figure 2.

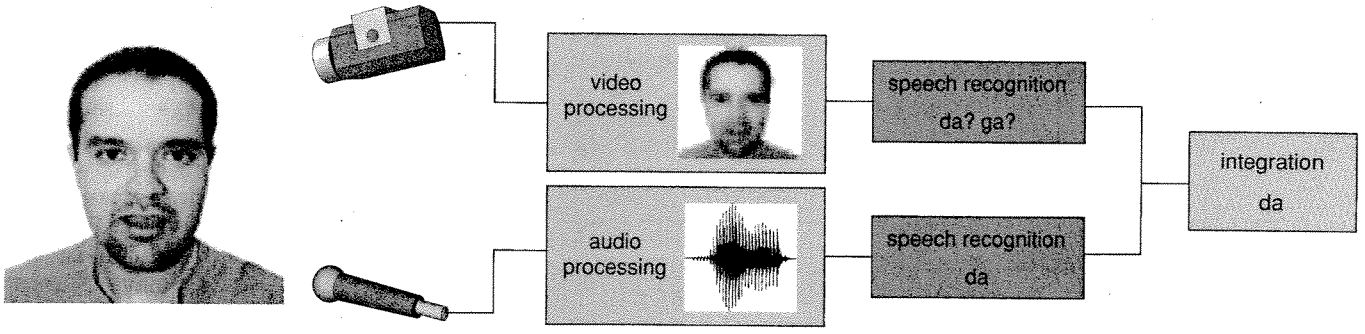


Figure 8. Machine speechreading uses input from a video camera and microphone. A system developed by one of the authors (Stork) processes the two signals separately, assuming no crosstalk, before integrating them for categorization of speech. Audio signals are filtered and broken into component frequency bands as video images are preprocessed to extract features such as the location of the face, mouth and chin. Each system computes a discriminant function for each category—in this case two possible syllables. The final, bimodal category corresponds to the maximum of some discriminant function. The process parallels the Bayesian integration scheme thought to be used in human speech recognition.

tegration areas of the brain? Many investigators have argued that crosstalk occurs between visual and auditory modalities, and the most popular neural-network model, called interactive activation, assumes crosstalk. However, ample evidence suggests very little crosstalk takes place in human brains. Signals from the eye and ear pass through several independent stages of neural processing before they come together in the parietal region of the brain. Furthermore, the fuzzy-logical model of perception, which assumes no crosstalk, accurately describes human performance. Thus, we will concentrate here on a particular form of sensory integration in machines that is easy to implement without crosstalk. Without crosstalk, the overall system functions as two component recognizers, each of which takes the input and computes a discriminant function, a single numerical probability score, for each candidate syllable category. The winning category is the one with the largest discriminant function.

Since our primary concern is with sensory integration, we shall not describe the details of the calculation of the discriminant functions in each subsystem except to say that we use statistical models used extensively in speech and speechreading research (such as hidden Markov models, named after Russian mathematician Andrei A. Markov) and neural networks, which have figured prominently in acoustic speech recognition for many years. Some fascinating challenges remain, however, related to the difference in overall speed of signal variation in audio and visual information.

Although we have discussed Bayes's theorem above, are there other candidate models of integration that can explain our human data? One simple way sensory integration might take place is through competition between the channels, the chosen category depending solely on which channel has the highest probability, that is, $P(c \mid A \ \& \ V) = \text{Max}[P(c \mid A), P(c \mid V)]$. This scheme, however, would fail to show an advantage of performance given the test item visual "di" paired with auditory "di" over the test item visual "di" paired with auditory "bi." Using the experimental results described above, we can immediately dismiss this as a model of human performance. This naive method, furthermore, leads to poor recognition in artificial speechreading systems, especially in noisy conditions where the acoustic information is unreliable.

Another alternative method for sensory integration, used in early speechreading machines, had the acoustic system present its top two candidate categories. Then the computer chose between these two based on their probabilities computed from the visual data. This method, however, cannot adequately explain the McGurk illusion and other experiments in people, and it too leads to poor performance in speechreading machines.

A Bayesian integration that assumes independence within each syllable category, but dependence between video and audio *(class-conditional independence)*, most accurately predicts human performance *(see Figure 10)*. If the audio and video representations possess class-conditional independence—$P(A \ \& \ V \mid c) = P(A \mid c)\ P(V \mid c)$—then Bayes's theorem suggests that the integration method leading to optimal recognition is the discriminant function $P(c) = P(c \mid A)\ P(c \mid V)$.

A Bayes Law?

Indeed, amid the infinite wealth of sensory integration methods, it is a bit surprising that one as simple as the Bayesian method (analogous to the fuzzy-logical model of perception in people) leads to such excellent performance in machines. Furthermore, Bayesian analysis can explain a wide range of human

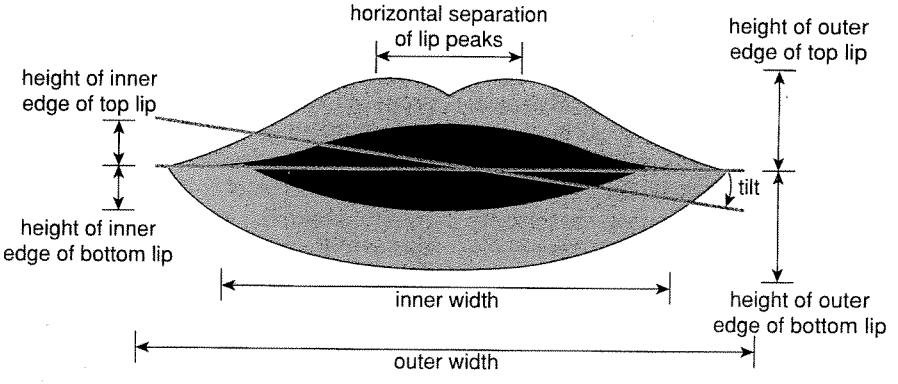


Figure 9. "Cartoon lips" serve as a template for analyzing features for speech recognition from videotape by computers. Features of such a deformable lip template include the curvatures and separations of four component arcs of the mouth. Once it has located a moving mouth, a speech-recognition system fits the template to a succession of video frames to extract the features necessary for categorizing speech.

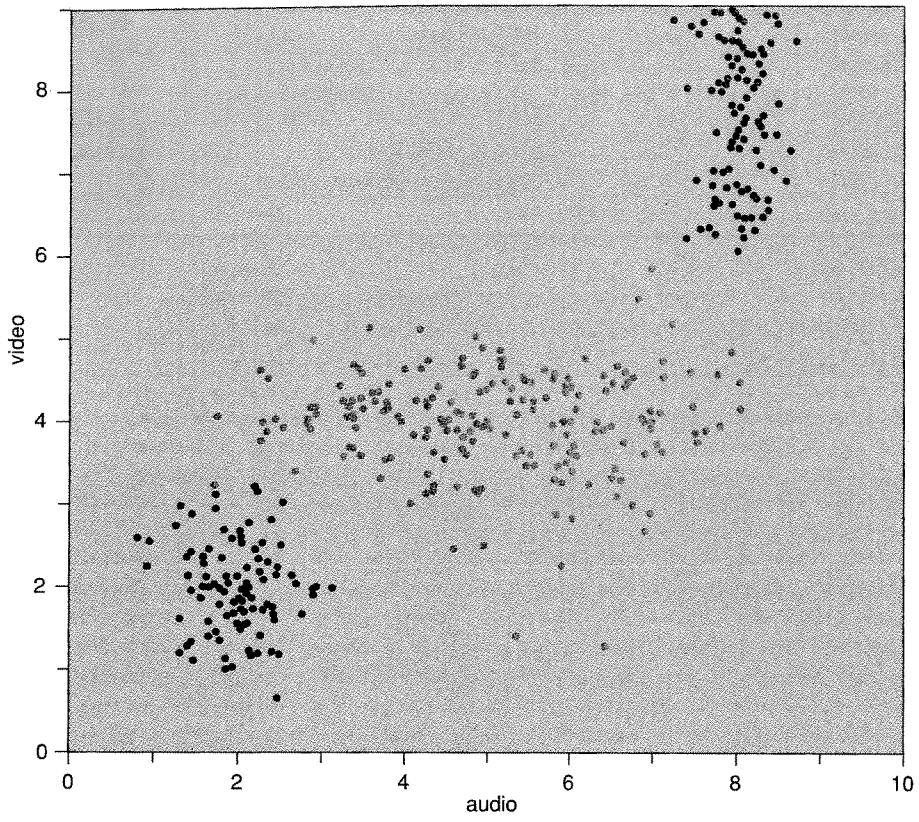


Figure 10. Each point in this scatter graph represents an audio-visual signal belonging to one of four syllable categories, each shown in a different color and graphed in arbitrary units. Overall, a strong correlation appears between the acoustic and the visual signals. The qualities of the video and audio signals increase together. But within a single syllable category, dots appear to be scattered randomly. Statisticians call such a relationship class-conditional independence. The Bayesian sensory integration described in the text yields the highest predicted accuracy for speech recognition, given the class-conditional independence illustrated here.

data—in speechreading and elsewhere. Of course, as in all psychological research, this and other empirical laws will be subject to fine (empirical) corrections, and may not apply in every foreseeable condition. This method applies to integration of three or more sources of information, and fits well experiments in a wide range of other domains, from the perception of emotional state based on the face and voice, to the judgment of linguistic interpretation based on grammatical and semantic information. It seems that this law of sensory integration might eventually gain the same status as some of the canonical laws in other realms of human information processing. Perhaps it will join the Weber-Fechner law, named for the discoveries of German scientists Ernst H. Weber and Gustav T. Fechner, who discovered the quantitative relation between stimulus and sensation nearly two centuries ago, or Ivan Pavlov's classical conditioning, which showed how pairing neutral and non-neutral stimuli can induce a trained response to the neutral stimulus.

We are naturally drawn to speculate on the fundamental reasons for the success of this form of Bayesian sensory integration and its ultimate roots in conditional independence. In cases where widely divergent kinds of sensory information exist—the smell of an apple and the smoothness of its skin—we can expect class-conditional independence; we would not expect strong correlations between such representations. In other cases, where such correlations exist, the human nervous system may learn to re-represent the signals as independent, before sensory integration. Because learning is fundamental to achieving expert pattern recognition, the most valuable aspect of independent representation of the modality-specific signals may be that it makes it easier for a system to learn the appropriate contingencies. This theory leads to predictions in underlying neural processing, including learning, that may one day find support in neurophysiology.

Automatic speech recognition has proved notoriously difficult, and any method to improve its accuracy would be a most welcome step in the development of new technologies. The incorporation of video signal processing into speechreading systems yields significant improvement under noisy conditions and modest improvement in noise-free environments. Thus their greatest contribution may be to broaden the range of applicability of automatic speech recognition. Noise or multiple distracting talkers affect automated teller machines, noisy offices, airports or bus terminals and the cockpits of fighter planes, making these ideal places to apply automated speechreading. Automatic transcription of television news reports could be aided by speechreading, using the image and sound of the newscaster in the broadcast signal. In many cases video input is already available: security cameras at automatic teller machines and video cameras atop computer consoles for video mail and videoconferencing. With continued development of speechreading technology, we can hope to improve the accuracy of such important devices.

Acknowledgment

The authors are grateful to Michael M. Cohen and Jonas Beskow for their help on the illustrations.

Bibliography

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian Theory.* New York: Wiley.

Massaro, D. W. 1987. *Speech Perception By Eye And By Ear: A Paradigm for Psychological Inquiry.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

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

Massaro, D. W., and D. Friedman. 1990. Models of integration given multiple sources of information. *Psychological Review* 97:225–252.

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

Stork, D. G., and M. E. Hennecke, eds. 1996. *Speechreading by Humans and Machines,* New York: Springer-Verlag.

Stork, D. G., and M. E. Hennecke. 1996. Speechreading: An overview of image processing, sensory integration and pattern recognition techniques. *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition.* Killington, Vt., pp. xvi–xxvi.

Links to Internet resources for further exploration of "Speech Recognition and Sensory Integration" are available on the *American Scientist* Web site:

http://www.amsci.org/amsci/articles/98articles/massaro.html