Perception, 1997, volume 26, pages 627–644

Perception of visible speech: influence of spatial quantization

Christopher S Campbell, Dominic W Massaro
Department of Psychology, University of California at Santa Cruz, Santa Cruz, CA 95064, USA
Received 15 October 1996, in revised form 17 April 1997

Abstract. Visible speech reading was studied to determine which features are functional and to test several models of pattern recognition. Nine test syllables differing in their initial consonant 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. Different models were tested against the confusion matrices. Six features were identified as functional in distinguishing among the nine consonant–vowel syllables. These features were used as sources of information in a fuzzy-logical model of perception and an additive model. The fuzzy-logical model provided a significantly better description of the confusion matrices, showing that speech reading is analogous to other domains of pattern recognition such as face recognition and facial-affect perception.

1 Introduction

The talker's face as well as his or her voice convey linguistic information in face-to-face communication. This contribution of visible speech has been shown in various speech-perception contexts; namely, with auditory speech that is conflicting (McGurk and McDonald 1976), ambiguous (Massaro and Cohen 1990), or degraded (Sumby and Pollack 1954). In the absence of audible speech subjects with normal hearing have been shown to speech read reliably without any systematic training (Heider and Heider 1940; Massaro and Cohen 1995). Research has also demonstrated that subjects with normal hearing (Massaro et al 1993) and hearing-impaired subjects (Walden et al 1977) can be trained to discriminate nine classes of visible consonant phonemes (visemes) without the aid of audible speech. In addition, eight vowel-viseme categories can also be speech read reliably (Montgomery and Jackson 1983).

In further inquiry into visible-speech perception it has been asked what features of the face convey the perceptually functional information for speech reading. Not surprisingly, studies have shown that functional features reside in the lower half of the face. For example, experiments by Erber (1974) with $\frac{3}{4}$-facial views indicate that jaw rotation and cheek movement are important. Functional features have also been shown to reside in the lips. Summerfield (1979) presented subjects with different lip representations in four conditions to determine their contribution to perception when the acoustic speech was distorted with interfering prose: (a) a control in which the entire face was shown; (b) lips isolated by painting with phosphorescent paint and filming in ultraviolet lighting; (c) a point-light display of the lips similar to that of Johansson (1974); (d) a moving-ring representation of the lips. In all conditions, the visual display was shown in synchronization with auditory speech. Correct responding was increased by 42.6% when the entire face was shown relative to hearing the distorted auditory speech alone. With only the painted lips, correct responding was improved by 31.3%. The third and fourth conditions, however, did not significantly improve performance. Summerfield concluded that untrained subjects use the functional features of the lips to perceive speech. It was speculated that these features may include lip occlusion, horizontal lip extension, and oral area. Benoit et al (1996) found that a view of the jawbone enhanced speech reading relative to just a view of the lips. Including the skin around the jaw improved speech reading even more.

In a seminal study H W Campbell (1974) evaluated the contribution of phonetic features by using the hierarchical clustering analysis of confusion matrices (for a published example of this analysis, see Walden et al 1977). In this type of analysis, phonemes are organized in steps into a hierarchy of groups on the basis of their similarity. The similarity of two phonemes is given by the degree of their mutual confusions: two phonemes are similar to the extent that they are confused with one another. Subjects identified consonant–vowel syllables under conditions of auditory, visible, and bimodal speech presentation. The auditory syllables had one of three levels of noise distortion added. The hierarchical analysis of various phonetic features showed that the visible features from most salient to least salient were coronal, continuant, and voicing. Coronal is a feature that distinguishes place of articulation and accounts for our ability to speech read the difference between labial (/b/, /p/) and nonlabial (/d/, /t/) consonants. Continuant describes manner of articulation and accounts for our ability to speech read the difference between stops (/d/) and fricatives (/s/). Voicing describes the difference between voiced (/b/) and voiceless (/p/) cognates that are otherwise identical. The salience of the auditory features had just the opposite ordering: voicing, continuant, and coronal. This hierarchical analysis helped identify potentially functional features of visible and audible speech and simultaneously revealed their complementary relationship. A salient feature in one modality is less salient in the other.

Other classification schemes have been proposed for visible speech. Binnie et al (1974) suggested five articulation sets based on phonetic theory: bilabials, labiodentals, interdentals, rounded labials, and linguals. The Jena method provides a simpler classification scheme in which three visible structures of the face define the functional features. These facial structures are the lips, tongue, and tongue–palate (Bunger 1952). Most descriptions of speech reading, however, group phonemes into categories called visemes (Berger 1972). Viseme classes are constructed to maximize the visual similarity of phonemes within the class while minimizing the visual similarity between viseme classes. Several viseme classifications have been developed; some with as few as four visemes and others with as many as twelve (Berger 1972). This lack of agreement arises from the noncrisp or fuzzy nature of visible speech. Phonemes, for example, do not fit into viseme classes in an all-or-none fashion but, rather, they fit to various degrees. Thus, any number of viseme classes can be defined based on the criterion level of performance for class membership and the quality of the test stimulus. In our studies, we test nine consonant phonemes corresponding to the nine viseme categories of Walden et al (1977). Given our concern with the features used in visible-speech perception, we now look to the literature in face recognition to determine if research in that domain can inform our study.

1.1 *Face recognition*

Perceptual information has also been a topic of study in the face-recognition and general-visual-perception domains. Unlike the speech reading domain, however, face-recognition research is focused more directly on the nature of perceptual information and the methodology used in exploring perceptual information. Current research in face recognition gives the impression that faces are perceived as both a set of individual features and the structural relations among features. Faces in general are thought to be perceived in a more configural manner than nonface patterns (Tanaka and Farah 1993). Diamond and Carey (1986) distinguish between two types of structural relations in object-perception tasks, first order and second order. Second-order structural relations are those that remain constant across a class of stimuli whereas first-order structural relations do not. For faces as a class of stimuli, the relations of the nose, mouth, and eyes are fixed and therefore are second-order relations. For landscapes, however,

the relations of the trees, mountains, and river are first order because they can vary considerably across specific instances. Configural processing is used not only in face-recognition tasks but also in any task in which one has acquired a certain level of expertise (Diamond and Carey 1986). Given that configural processing develops through experience, then the amount of configural processing involved in the perception of a face is thought to increase as experience with the face increases (Carey et al 1980). Highly learned patterns like familiar faces contain greater perceptual information about the structural relations among features than unfamiliar faces.

Although some investigators have advocated that faces are processed differently from other domains of visual perception (eg Farah 1995), we operate on the assumption of a continuity across these different domains. Thus, the methodologies and theoretical distinctions of face-recognition research are useful in understanding perceptual information in speech-reading tasks.

In research on general visual perception roughly the same view of feature organization is held as in face-recognition research with the primary difference being terminology. Features are generally classified as local and global instead of featural and configural (Bruce 1988). Although clear definitions are scarce in the literature, local and global features refer to the size of elements in relation to a larger image. The feature/configural distinction is used to focus more on features as points of reference for the description of structural relations in an image. In most cases, local features are isolated features and global features are the structural relations among features. However, this is not always the case. In face recognition, for example, the eye is considered a local feature, yet the eye has many potential configural properties of its own. Some of these include the relation of the upper lid to the lower, the right corner to the left, and the roundness of the eye opening. Likewise, global properties need not necessarily depend on the relations among elements (Kimchi 1992). Some global aspects of the face, hairstyle for example, are not made up of the spatial arrangements of components. Rather, hairstyle is more like a global feature or form that occupies a large proportion of the face.

1.2 *Spatial-frequency information*

Research in visual perception has demonstrated that global and local features are processed somewhat independently in the visual system (Navon 1977, 1981; Lagasse 1993). It has been proposed that the visual system performs a spatial-frequency analysis whereby global features are processed by low-spatial-frequency channels and local features are processed by high-spatial-frequency channels (DeValois and DeValois 1988). Support for multiple spatial-frequency channels comes from early psychophysical experiments. For example, adaptation to specific spatial frequencies was demonstrated by using square-wave (Pantle and Sekuler 1968) and sine-wave (Blakemore and Campbell 1969) gratings. These experiments showed that the sensitivity to specific spatial frequencies decreased after adaptation to gratings of the same frequency. Also, adaptation to a specific spatial frequency has been found to affect the subsequent viewing of gratings of different spatial frequency (Blakemore and Sutton 1969). Generally, a contrast effect is observed in which the gratings at lower spatial frequencies than the adaptor are judged as even lower than the adapting frequency and gratings at higher spatial frequencies are judged higher. A more recent spatial-frequency-adaptation experiment supports the idea that global and local information is processed independently through channels of low and high spatial frequency (Shulman et al 1986). Subjects in this experiment were required to judge the orientation of small or large 'C' character(s) after exposure to an adapting sine-wave grating. Results showed that response time and errors increased for local-orientation judgments as the frequency of the adapting sine-wave grating increased. Likewise, response time and errors increased for global-orientation judgments as the frequency of the adapting sine-wave grating decreased.

Riedl and Sperling (1988) in an impressive experiment demonstrated that multiple spatial-frequency bands provide redundant perceptual information for American Sign Language (ASL) stimuli. The ASL stimuli used were initially 512×512 pixels per frame samples at 30 frames s^{-1}. One hundred of these ASL signs were randomly assigned to one of four frequency bands and then band-pass filtered. The range of the adjusted frequency bands in cycles per frame width were (a) 0–4.2, (b) 4.8–6.5, (c) 9.3–17.6, (d) 21.5 and above. The results showed that intelligibility was roughly comparable across the four frequency bands, with somewhat better performance on the higher frequencies. The mean percentage correct from the first band to the fourth was 66, 68, 88, and 80. In a follow-up experiment, Riedl and Sperling were able to selectively mask each frequency band with dynamic Gaussian noise of the same frequency. The results of Riedl and Sperling (1988) provide an explanatory basis for results from other experiments in ASL perception demonstrating high levels of intelligibility at low information rates (Sperling 1980; Tartter and Knowlton 1981; Pearson 1986).

The claim that different spatial-frequency bands convey redundant perceptual information has also been suggested by researchers in face recognition (Bruce 1988). Current research suggests that because faces are overlearned stimuli then subjects are sensitive to structural relations among features which are transmitted through low-spatial-frequency bands. According to this view, high-spatial-frequency information can be considered to be redundant (Ginsburg 1978, 1980). It has been demonstrated, however, that high spatial frequencies are themselves sufficient for face recognition (Fiorentini et al 1983). Thus, one could also conclude that low-spatial-frequency information is redundant. Given the complexity of these results, it is clear that spatial-frequency analysis provides only a partial account of performance on perceptual tasks (eg Uttal et al 1995a, 1995b). It is also necessary to understand pattern recognition as a set of processes that occur after the spatial-frequency analysis of the visual system. Spatial-frequency analysis is a biological constraint that determines what type of information or features are available to the pattern-recognition system. This fits with our general findings that there are multiple sources of information supporting perception and pattern recognition (Massaro, in press). Further, how these sources of information are evaluated and integrated is one of the main concerns of the present paper.

The stimulus-distortion method chosen for this experiment is the spatial-quantization method first used by Harmon (1973) and later elaborated by Costen et al (1994, 1996) and Uttal et al (1995a, 1995b). Spatial quantization is the process of reducing the resolution of an image through the local averaging of pixels into larger blocks. This process is sometimes referred to as pixelization. These blocks act to eliminate spatial-frequency information depending on block size. These blocks eliminate the relatively high-spatial-frequency information in the image while preserving the relatively low-spatial-frequency information. Also, the edges of these blocks introduce high-spatial-frequency components into the image that are not representative of the original image. Owing to the fact that spatial quantization introduces nonrepresentative high-frequency components, it tends to be more devastating to performance than other image-degradation methods such as low-pass Fourier transform or Gaussian blurring. However, it has been shown that spatial quantization has a comparable response function to either of these two image-degradation methods in face-recognition tasks (Costen et al 1994, 1996). This view implies that spatial quantization is reducible to low-pass Fourier transform and Gaussian blurring. However, it has been argued that spatial quantization may be more complex than originally thought (Uttal et al 1995b).

Spatial quantization was first used to study recognition processes in face-recognition research. Harmon (1973), using this method, was able to show that natural faces without high-spatial-frequency information were still easy to recognize. Faces were recognizable with only 16×16 pixels or about 8 cycles per face. The same results were obtained in

a more recent experiment by Bachmann (1991), who distorted natural faces by using the spatial-quantization method and measured accuracy of face identification. Identification performance remained highly accurate with increased quantization across conditions up to 18 horizontal pixels per face, which corresponds to 9 cycles per face, but fell off sharply at 15 pixels per face.

Another experiment by Costen et al (1994) showed that increases in image information above 21 horizontal pixels (10.5 cycles per face) does not improve accuracy or decrease response time in an identification task. Performance deteriorated only when information in lower–mid and lower spatial frequencies was not available. All three experiments taken together indicate that spatial frequencies below roughly 10 cycles per face convey adequate information for face recognition.

Perception of facial affect has also been shown to be highly robust to spatial quantization. For example, Wallbott (1992) measured facial-emotion identification for natural faces with fourteen basic emotional states: euphoria, happiness, fear, terror, contempt, shame, sadness, despair, cold anger, hot anger, interest, boredom, disgust, and pride. The four conditions included one control and three levels of quantization: 150×150, 75×75, and 38×38 pixels per face. These three levels conform to about 75, 37.5, and 19 horizontal cycles per face. Mean identifications from control to the third level were 66%, 65%, 47%, and 35% correct. These results show a high level of resistance to distortion but not as high as that seen in face-recognition experiments.

The purpose of the present investigation was to extend the analysis of spatial-frequency information used in face-perception research to the domain of visible-speech perception. It was hoped that this method would clarify the perceptual importance of information from the local to the global level in the general task of visible-speech recognition and in the task of disambiguating individual viseme classes. Spatial information was manipulated by varying the block size of the display in the presentation of the visible speech. As block size is increased, relatively lower spatial frequencies are eliminated. This effect is appropriate for addressing the ecological concerns of speech readers. In natural speech reading situations, high-spatial-frequency information is usually lost first, leaving only low-spatial-frequency information for recognition. Such situations occur when the speaker is at a distance (Small and Infante 1988), the speaker is in the peripheral viewing area, the speaker's face is obscured as when viewed through a screen, or the speech-reader suffers from reduces contrast sensitivity due to old age (Thorn and Thorn 1989).

1.3 *Models of pattern recognition*

Another purpose of this investigation was to use the pattern of viseme confusions (confusion matrices) to test the fuzzy-logical model of perception (FLMP) (Massaro and Friedman 1990). The FLMP has been previously shown to describe performance better than alternative models in several domains of perception, including facial-affect, audible-speech, and bimodal-speech experiments (Massaro and Cohen 1990; Ellison and Massaro 1997). In the case of bimodal speech, for example, Massaro and Cohen (1990) tested the perception of audible and visible speech in an expanded factorial design using a /ba/ to /da/ continuum. The subjects' identification responses were used to test the FLMP and additive model of perception (AMP). The primary difference between these two models is that the FLMP predicts multiplicative integration and the AMP predicts additive integration of features. The FLMP produced a much better fit to the responses than the AMP.

In the present study, the FLMP was again tested against the AMP. The AMP was chosen because it is currently a proposed mechanism of information integration in perception (Cutting et al 1992; Massaro and Cohen 1993). The AMP was chosen also because it is mathematically equivalent to several theories of psychological processes including the single-channel model, the categorical model, and the weighted-averaging

model of perception. The single-channel model predicts that the subject attends to only one source of information (one feature) on each trial. In the categorical model it is assumed that each feature is categorized and then an alternative is chosen by taking into account a bias factor for each alternative. The weighted-averaging model holds that the value of each feature is weighted according to a bias factor and the result is averaged to support a given alternative.

2 General procedure

2.1 *Subjects*

Eighteen undergraduates served as subjects as an option to fulfill a class requirement. All were native speakers of English and showed normal or corrected vision.

2.2 *Apparatus and materials*

A computer-generated synthetic head with face was used to produce the visible-speech stimuli in all conditions. The synthetic head was generated with a polygon topology and animation was controlled through a set of parameters (Cohen and Massaro 1990). The polygon topology was made up of approximately 900 surfaces connected at the edges to create the three-dimensional head with eyes, pupil, iris, sclera, eyebrows, nose, skin, lips, tongue, teeth, and neck (see figure 1). In the present study, the synthetic head did not have ears or hair. Animation of the synthetic head was generated in real time (approximately 30 frames s^{-1}). The real-time animation was generated by a Silicon Graphics Crimson Reality Engine with 96 megabytes of RAM and an R400 100 MHz microprocessor. Speech was controlled by eighteen parameters, such as jaw rotation and thrust, horizontal mouth width, moving the corners of the lips, protruding the lips, lower-lip 'f' tuck, raising the upper and lower lips, horizontal and vertical teeth offset, and tongue angle, width, and length. Prior experiments by Cohen et al (1996) and Massaro (in press) have shown that visible speech produced by the synthetic head is almost comparable to that of a real human. The spatial-quantization filtering was performed by using the Mosaic effect of a Panasonic MX-50 mixing/special-effects board. Examples of the stimuli are given in figure 1.

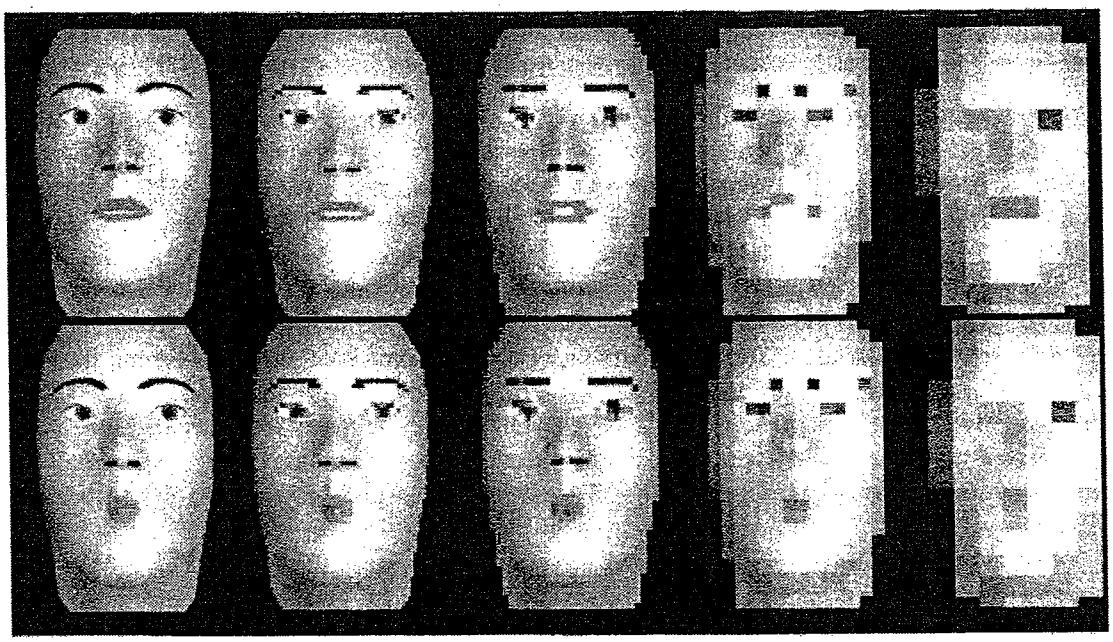

Figure 1. The synthetic head in the control condition (323 cycles per face) at far left through the highest level of distortion (4 cycles per face) at the far right. Extreme articulations are shown for two visemes, /va/ (top) and /wa/ (bottom).

2.3 *Design and procedure*
Spatial quantization was varied at five steps from level 0 (normal display) to level 16 of the Mosaic effect on the MX-50 mixer. These five levels were equivalent to the following cycles per face: level 0, 323 cycles per face; level 2, 31 cycles per face; level 4, 16 cycles per face; level 8, 8 cycles per face; level 16, 4 cycles per face. Cycles per face were calculated by dividing the number of quantized blocks that run horizontally at the face's eye level by two. Cycles per face or cycles per image is a common measure in spatial-frequency research and is claimed to be independent of viewing distance (Riley and Costall 1980). Figure 1 shows the synthetic head from the control condition (323 cycles per face) to the highest level of distortion (4 cycles per face) for the two visemes /va/ (top) and /wa/ (bottom). The five levels were crossed with nine visemes: /ba/, /va/, /tha/, /da/, /ra/, /la/, /za/, /ja/, /wa/, for a total of forty-five conditions. A random sequence of trials was determined for each test by sampling the forty-five conditions randomly without replacement within a block of 45 trials. There were fifteen test blocks for a total of 675 trials. There were three sessions of 225 trials each with a 5 min break between sessions.

Subjects were instructed to watch a video monitor (NEC 12 inch C12-202A), choose the one out of the nine visemes that was articulated by the synthetic head, respond as quickly as possible on each trial with the first answer that came to mind, and use the feedback given after each trial to improve their performance. Each subject was tested in small (5 ft × 6 ft) soundproofed room with a door at one end and sound-absorption material covering part of the walls. Subjects were seated about 1 m from a video monitor which was at eye level. At this distance, the synthetic head was presented at a constant size of 16.95 vertical and 13.42 horizontal deg of visual angle. The lower half of the face was measured from the tip of the nose to the bottom of the chin. Spatial quantization from level 1 (control) to level 5 comprised the following cycles deg^{-1} for the lower half of the face: 48.7, 5.2, 2.7, 1.3, 0.7. The spatial resolution of the lower half of the synthetic face for the quantization conditions fell largely within an optimal range of contrast sensitivity which is highest between about 0.5 and 10 cycles deg^{-1} of visual angle (cited in Wandell 1995).

Responses were input on a TV1950 keyboard positioned between the subject and the video monitor. There were nine keys on the bottom row (from 'x'; to '/') marked with labels for each of the nine visemes. These labels were explained to all the subjects before the experiment and could be considered intuitively appropriate. The viseme labels were BA, VA, THA, DA, RA, LA, ZA, JA, and WA. On each trial, subjects were required to respond within 5000 ms of the stimulus presentation or they were prompted by a 100 Hz beep sound. A total of four subjects could take part at the same time and advancement from one trial to another could not occur until all subjects responded. After all subjects responded, they received feedback in the form of the written label without the face in the bottom left corner of the video monitor.

3 Results
The mean proportion of correct identification was defined for each subject for each quantization condition as the total number of correct viseme classifications divided by the total number of trials. Figure 2 gives the mean proportion of correct identification across subjects as a function of spatial quantization. A positively decelerated curve best represents the functional relationship between identification performance and spatial quantization. This curve spans the entire stimulus range from asymptote at the control condition to chance performance with nine response alternatives (11% correct) at the highest level of distortion (4 cycles per face). A two-factor within-subject analysis of variance indicated that spatial quantization reduced correct identification from 68% to 12.5% ($F_{4,68} = 113.52$, $p < 0.001$). Multiple comparisons by means of the Tukey HSD

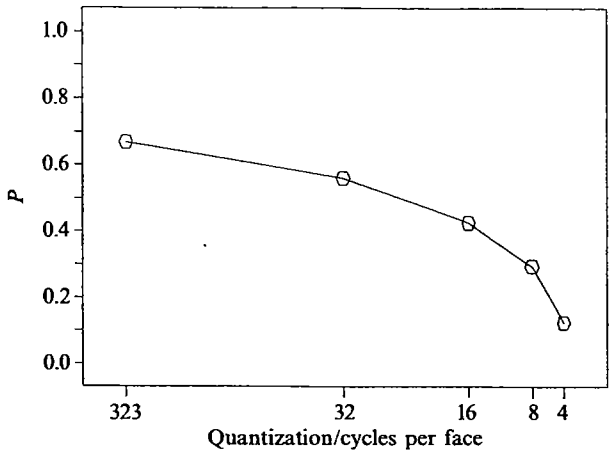

Figure 2. The mean proportion of correct identification (P) across subject and viseme for each level of spatial quantization. The relationship between the dependent measure, proportion of correct identification, and spatial quantization is a positively decelerated function.

method with $\alpha = 0.05$ and df $= 765$ showed that each change in spatial quantization produced a significant change in performance.

Figure 3 gives the proportion of correct performance for each of the eighteen subjects, averaged over all visemes as a function of the five levels of spatial quantization. Inspection of individual performance shows that the function given in figure 2 was not merely the result of averaging but that spatial quantization had a similar effect on the performance of each subject. Although subjects differed in overall performance, the performance of each subject is best described by a monotonically decreasing function of spatial quantization.

Figure 4 gives the mean proportion of correct identification for each viseme and each level of quantization. Speech-reading performance differed significantly for the different visemes ($F_{8,136} = 30.14$, $p < 0.001$). Consistent with other speech-reading experiments (eg Walden et al 1977), correct identification was higher for /ba/, /va/, /tha/, /la/, /ja/, and /wa/ than for /da/, /za/, and /ra/. Spatial quantization differentially influenced

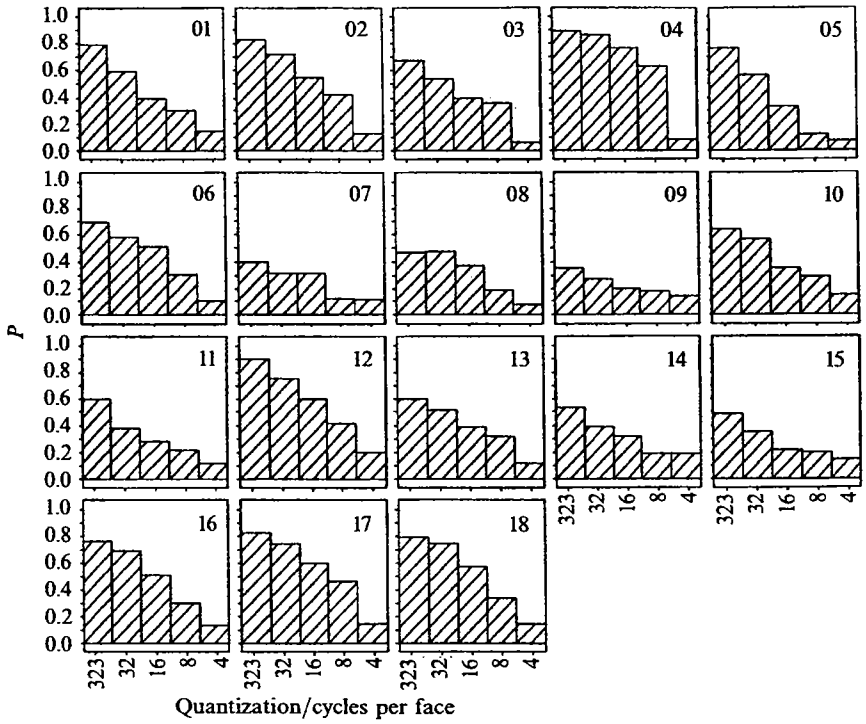

Figure 3. The proportion of correct performance (P) for each of eighteen subjects averaged over all visemes at each of the five levels of spatial quantization (control at far left to most degraded at far right). Spatial quantization is on a logarithmic scale.

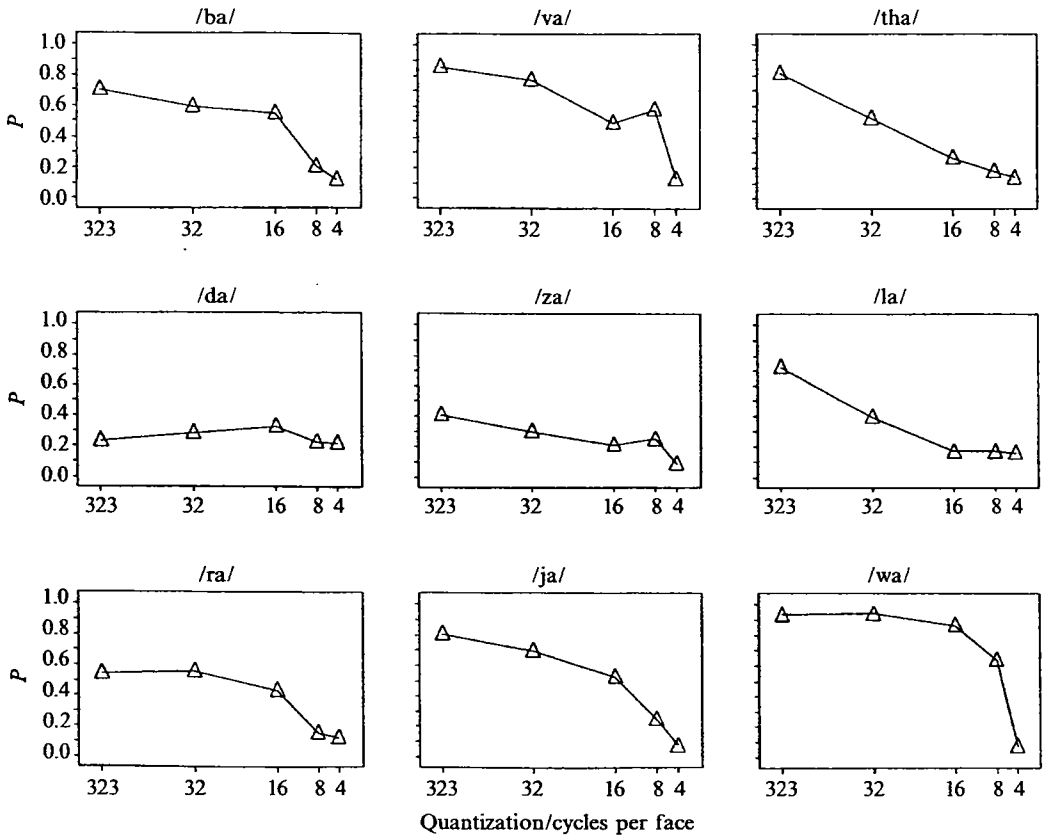

Figure 4. The mean proportion of correct identifications (P) across subjects for each viseme and each level of distortion. Spatial quantization is on a linear scale.

speech reading of the different visemes, as indicated by an interaction between viseme and distortion ($F_{32,544} = 15.77$, $p < 0.001$). Most of the visemes were influenced similarly and it can be seen that performance is largely resistant to spatial quantization up to and including 16 cycles per face. However, the visemes /tha/ and /la/ are exceptions because performance degraded quickly with increasing spatial quantization. Performance for /da/ and /za/, on the other hand, was fairly poor at the control level and therefore could not be further degraded by spatial quantization.

Confusion matrices for each level of spatial quantization are given in figures 5a–5e. The diagonal line of circles with crosses show the proportion of correct responses whereas the off-diagonal circles show confusions. Circle size depicts the proportion of responses in a given cell; the greater the proportion of responses, the larger the circle. Without any distortion (figure 5a), the proportion of responses on the diagonal is very large, showing a high degree of correct responding along with a small rate of confusions. The visemes /da/ and /za/ violate this trend, however. The test viseme /da/ was often identified as /la/ and to a lesser extent /za/, /tha/, and /ra/. The viseme /za/ was most highly confused with /da/ and somewhat less with /ja/, /tha/, /la/, and /ra/. The confusions do not appear completely symmetric. For example, the visible syllable /da/ was called /la/ much more often than the visible syllable /la/ was called /da/.

Across levels of increasing quantization, the correct identifications on the diagonal decrease to chance performance at 4 cycles per face and proportions of confusions increase. This decrement in performance is not the same for all visemes. At the first level of distortion (32 cycles per face) only /tha/, /la/, and /za/ showed any substantial decrease in mean proportion of correct identifications across subject. The visemes /va/ and /ja/ began to break down at the second level (16 cycles per face), followed by /ra/, /wa/, and /ba/ as distortion was further increased.

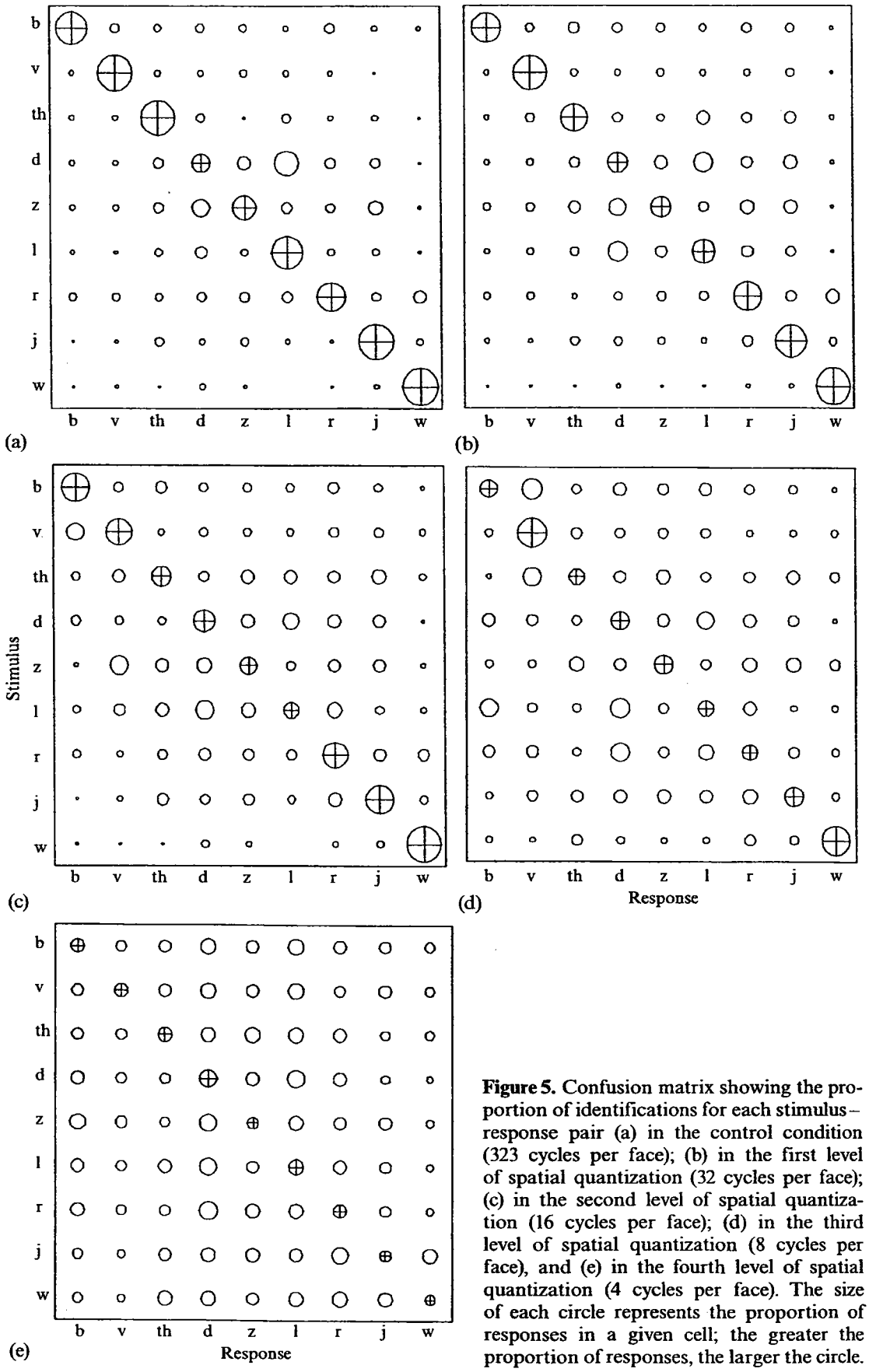

Figure 5. Confusion matrix showing the proportion of identifications for each stimulus–response pair (a) in the control condition (323 cycles per face); (b) in the first level of spatial quantization (32 cycles per face); (c) in the second level of spatial quantization (16 cycles per face); (d) in the third level of spatial quantization (8 cycles per face), and (e) in the fourth level of spatial quantization (4 cycles per face). The size of each circle represents the proportion of responses in a given cell; the greater the proportion of responses, the larger the circle.

A more formal examination of the confusion matrices was performed with an information-transmission analysis, similar to that used by Miller and Nicely (1955). This analysis gives the degree of covariance between the input and output in a stimulus–response system. Information transmission (T) is defined as the maximum uncertainty in a stimulus–response system minus the actual uncertainty (H) of the responses:

$$T_{(x;y)} = H_{\max} - H, \tag{1}$$

where

$$H_{\max} = -\sum p_i \log_2 p_i + -\sum p_j \log_2 p_j \tag{2}$$

and

$$H = -\sum p_i p_{ij} \log_2 p_i p_{ij} . \tag{3}$$

The variable p_i is the probability of the stimulus i, p_j is the probability of the response j, and p_{ij} is the probability of the response j given the stimulus i. Logarithms are taken to base 2 so that information is scaled in bits per stimulus or the number of binary decisions needed to discriminate the input alternatives. The proportion of information transmitted from x to y in bits per stimulus is given by

$$T_{\mathrm{rel}(x;y)} = \frac{T}{H_{(x)}}. \tag{4}$$

Figure 6 shows the normalized covariance measure or proportion of information transmitted at each level of distortion. The relationship between normalized covariance and distortion is given by a positively decelerated function similar to the overall accuracy performance shown in figure 2. This functional similarity indicates that the observed function was not an artifact of the accuracy scale. It can be seen from figure 6 that at the control level a relatively moderate amount of information was transmitted (53%) by the synthetic face for the visemes overall. This result suggests that faces in general and the synthetic face specifically are reasonably informative for nonexpert speech readers. If the reduction in the quality of the test syllable is assumed to be a linear function of spatial quantization, the positively decelerated function of figure 6 indicates that information transmission is fairly resistant to degradation and speech-reading is robust across a fairly substantial amount of degradation. (The term 'data' is defined as a physical dimension such as photons or frequency while 'information' is a psychological dimension that depends on the importance of data for a given task. Thus, if a segment of data is used to decide between two possible alternatives in a particular task, then that segment of data is information.)

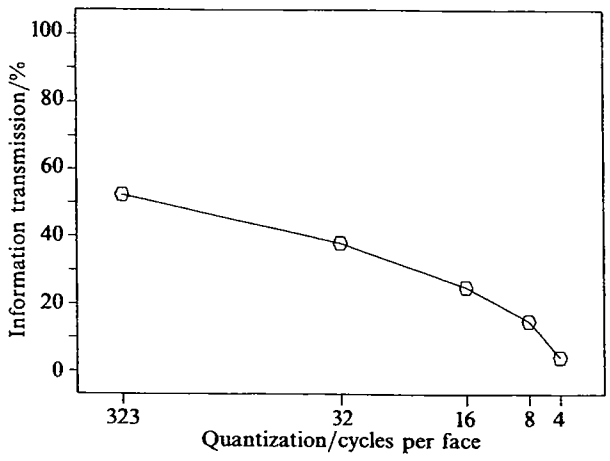

Figure 6. The proportion of information (or normalized covariance measure) transmitted across subject and viseme at each level of distortion. The relationship between normalized covariance and distortion is given by a positively decelerated function.

4 Tests of models

Given our experience with the analysis and synthesis of visible speech, it was possible to construct a set of visible features that should distinguish among nine consonant visemes in English. Six features were deemed to be necessary information and our goal is to test the psychological validity of these proposed features. Table 1 summarizes the feature assignments for the nine visemes used in our experiment. The duration feature differentiates visible phonemes that have a long versus a short duration: /za/ is long and /ba/ is short. The visible-tongue-tip-movement feature distinguishes whether or not movement of the tip of the tongue is visible during articulation. For example, the tip of the tongue can be seen in /tha/ and /la/ but not in /ba/ or /ja/ (as in judge). Lip rounding distinguishes the 'rounded' visible consonants /ra/, /wa/, and /ja/ from others. Within the rounded consonants, horizontal mouth narrowing distinguishes /wa/ in which there is a small distance between the corners of the mouth from /ra/ and /ja/ with a large distance. Dental adduction distinguishes consonants in which the teeth can be seen to come together from those in which the teeth are either not visible or seem to be apart; thus, /tha/, /za/, and /ja/ versus the others. Last, lower-lip tuck refers to the lower lip elevating and tucking under the upper teeth. This is a highly distinctive feature of /va/.

Table 1. Feature set describing the nine consonant visemes in English.

Feature	Viseme								
	/ba/	/va/	/tha/	/da/	/za/	/la/	/ra/	/ja/	/wa/
Duration	−	−	−	−	+	+	+	+	+
Tongue-tip movement	−	−	+	+	+	+	−	−	−
Lip rounding	−	−	−	−	−	−	+	+	+
Mouth narrowing	−	−	−	−	−	−	−	−	+
Dental adduction	−	−	+	−	+	−	−	+	−
Lower-lip tuck	−	+	−	−	−	−	−	−	−

It should be noted that this specification does not imply an all-or-none (categorical) definition of features. The binary specification of features in table 1 only represents the direction of the feature property for each viseme. A plus means that the feature is present whereas a minus means the absence of the feature. The feature should not be interpreted as being present or absent but rather as the direction of the continuous feature value it assumes with the corresponding viseme. For example, the value for the feature 'rounded' might be 0.8. In this case, presentation of a viseme with the rounded feature supports all rounded visemes to degree 0.8. Nonrounded visemes would receive support $(1 - 0.8)$ or 0.2. Thus, the lips are described as rounded and the lower lip not occluded for /ja/ whereas the lips are described as unrounded and the lower lip occluded for /va/.

A feature analysis of the confusion matrices was performed to predict the pattern of responses given in the confusion matrices. The feature analysis was conducted by first specifying, a priori, one set of features used by subjects in speech reading. This feature set is a priori because it is specified prior to analysis, similar to independent variables chosen prior to a discriminant analysis. If the feature set was a posteriori, then it would be the result of analysis as is the case in multidimensional-scaling modeling or principle-components analysis, or would be tailored to the correct results by using information from the model fits with various feature sets. The feature set defines the expected direction of the prototype representation for each viseme and thus is binary specifying either a feature or its complement. Referring to the feature set shown in table 1, an example of the prototype for /va/ would be:

/va/: short duration, no tongue-tip movement, no lip rounding, no mouth narrowing, no dental adduction, lower-lip tuck.
Similarly, the features for the /wa/ prototype would be:
/wa/: long duration, no tongue-tip movement, lip rounding, mouth narrowing, no dental adduction, no lower-lip tuck.

Even though each feature is defined as a specific value or its complement, its influence in the perception of visible speech is represented by a value between 0 and 1. In the simple model there is one parameter for each feature and the parameter values range from 0 to 1. The parameter value for the feature indicates the amount of influence that feature has. Therefore, if the /va/ prototype is expected to have a short-duration feature and the calculated parameter value for this feature is 0.90 then the duration feature is highly functional in the expected direction. Alternatively, if the calculated parameter value for the duration feature is 0.50, then the conclusion would be that the duration feature is not functional at all. Because of the definition of negation as one minus the feature value, a feature value of 0.5 would give the same degree of support for a viseme that has the feature as it would for a viseme that does not have the feature. Last, if the calculated parameter value is 0.20 then the duration feature is functional but in the direction opposite from the expected. In this case the /va/ prototype should have been defined as having a long duration instead of a short duration.

The overall match of the feature set to the prototype was calculated by combining the features according to the constraints of the FLMP. These constraints dictate that features are sources of information that are evaluated independently of one another and the features are integrated multiplicatively (conjoined) to give the overall degree of support for a viseme alternative. Thus, the overall degree of support for /va/, $S(/va/)$, given the presentation of a /va/ syllable, is

$$S(/va/|/va/) = f_1 f_2 f_3 f_4 f_5 f_6 , \tag{5}$$

where f_i indexes a match between the feature in the stimulus and the corresponding feature in the /va/ prototype. A mismatch between the feature in the stimulus and the corresponding feature in the prototype would be indexed by $(1 - f_i)$. Thus, the support for the /wa/ prototype, given presentation of a /va/ syllable, is

$$S(/wa/|/va/) = (1 - f_1) f_2 (1 - f_3)(1 - f_4) f_5 (1 - f_6) , \tag{6}$$

where $(1 - f_i)$ indexes a mismatch between the feature in the stimulus and the corresponding feature in the /va/ prototype.

After the overall degree of support for each viseme is calculated, the stimulus is categorized according to the relative-goodness rule, which states that the relative probability of choosing an alternative is the goodness of match of that alternative divided by the sum of the goodness of match of all alternatives. The probability of responding /va/ given a /va/ stimulus is

$$p(/va/|/va/) = \frac{f_1 f_2 f_3 f_4 f_5 f_6}{\sum s_k} , \tag{7}$$

where f_i indexes a match between the feature in the stimulus and the denominator is the sum of the overall degree of support, s_k, for each of the nine alternatives. The probability of responding /wa/ given the same /va/ stimulus is

$$p(/wa/|/va/) = \frac{(1 - f_1) f_2 (1 - f_3)(1 - f_4) f_5 (1 - f_6)}{\sum s_k} . \tag{8}$$

The confusion matrix observations of this experiment are shown in figures 5a–5e. There are five confusion matrices, one for each level of spatial quantization, with 81 cells each for a total of 405 observed points to be predicted. Our set of six features