Cogn Process (2004) 5: 106–112
DOI 10.1007/s10339-004-0014-2

RESEARCH REPORT

Steven K. de la Vaux · Dominic W. Massaro

Audiovisual speech gating: examining information and information processing

Received: 3 November 2003 / Revised: 7 February 2004 / Accepted: 17 February 2004 / Published online: 23 April 2004

Abstract We investigated the perceptual processing of facial information and vocal information to test the nature of the multisensory integration process. Single-syllable words were presented in background noise at one of eight stimulus durations ranging from 45 to 80% of the total word duration. Performance was evaluated relative to a fuzzy logical model of perception (FLMP) and an additive model (ADD). These two models differ in the nature of the integration process: optimal multiplicative integration in the FLMP and additive in the ADD. The FLMP provided a significantly better description of the word identifications relative to the ADD. Consistent with the outcomes of several other tasks in audiovisual speech perception, the FLMP also accurately describes the continuous uptake of information during word perception.

Keywords Speech perception · Pattern recognition · Multisensory processing · Quantitative models

Introduction

How we process spoken language is a complex question. Because the speech signal varies over time, an important empirical question is the temporal nature of the recognition process (Massaro 1975) and the continuous mapping of sensory information onto lexical forms (Marslen-Wilson and Warren 1994). This question has been addressed productively in the area of auditory word recognition using a "gating task," in which part of the word presented is eliminated or gated out (Grosjean 1980; Cotton and Grosjean 1984; Tyler 1985; Salasoo and Pisoni 1985; Walley et al. 1995; Warren and Marslen-Wilson 1987, 1988; Grosjean 1997). Fragments of words are presented by eliminating some variable portion of the end of the word, and participants are asked to identify what was said on the basis of the presented information. This procedure allows one to vary systematically the amount of information presented to an experimental participant to study the process of word recognition.

Warren and Marslen-Wilson (1987, 1988) utilized the gating task to address the continuous uptake of acoustic cues in spoken word recognition. Listeners heard successively larger fragments of words such as *scoot* or *scoop* that differed in place of articulation of the final consonant, and words such as *crown* or *crowd* in which the manner of articulation of the final consonant differed. Correct word recognition occurred well before the complete word was presented. These results reveal that spoken language is processed on-line, some information is made available before other information, and recognition is achieved when there is sufficient information to indicate a given word unambiguously. The results also indicated that the minimum amount of sensory information necessary for recognition differed depending on the final vowel–consonant pair in the test word and the number of lexical competitors—words that are *phonetically* similar to the test word. These researchers argued that earlier arriving coarticulatory information can be used to reduce uncertainty about a speech segment and allow the perceiver to discriminate a word from competitors in its lexical neighborhood. Our pronunciation of a speech segment is changed depending on its neighboring segments, and this dependence can provide coarticulatory information in an earlier segment about a following segment. For example, the lip protrusion of the vowel in the word *stew* can begin during the pronunciation of the earlier consonants in the word. The acoustic consequences of this coarticulation can provide

Edited by: Marie-Hélène Giard and Mark Wallace

S. K. de la Vaux · D. W. Massaro (✉)
Department of Psychology, University of California,
Santa Cruz, CA 95064, USA
Tel.: +1-831-4592330
Fax: +1-831-4593519
E-mail: massaro@fuzzy.ucsc.edu
http://mambo.ucsc.edu/psl/dwm/

information about the identity of the vowel. More recent research has replicated and extended these findings (e.g., McQueen et al. 1999; Smits et al. 2003).

However, the auditory modality is not the only source of information used in speech perception. It is well known that visual information from the face contributes to speech perception (Sumby and Pollack 1954; Summerfield 1979; Massaro 1987, 1998). In face-to-face communication, visible aspects of speech production provide information in addition to that provided in the acoustic signal. The classic demonstration of the influence of visual information on speech perception is the "McGurk effect" in which the simultaneous pairing of an auditory /ba/ with the visual presentation of /ga/ produces a perception of /da/, /va/, or /θa/ (Massaro 1998; McGurk and McDonald 1976).

Most importantly, the presence of the face and the voice together can be more informative than either of the single modalities presented alone. In one study, single-syllable English words were presented at either a normal rate or at three times the normal rate (a compressed triple rate) in auditory only, visual only, and bimodal conditions (Massaro 1998, Chap. 1). At the normal rate, identification of the auditory word was nearly perfect (0.951), and bimodal performance did not differ much from the auditory alone condition (0.962). However, in the compressed condition, where auditory and visual identification was found to be 0.55 and 0.04, respectively, performance in the bimodal trials was 0.725. This outcome was interpreted as an example of the multiplicative nature of the integration algorithm, which also correctly predicts that visual information has a greater impact on performance when the auditory information is ambiguous.

Since speech perception involves the use of multiple sources of information, it is important to understand the articulatory dynamics of *acoustic and visual* aspects of speech, and how these temporally covarying sources of information are evaluated and integrated in the word recognition process. Given the success of the gating paradigm in addressing how acoustic information is evaluated and mapped on lexical representations, an extension of this task to audiovisual speech can prove fruitful in examining the nature of the integration algorithm across a range of information during the recognition process (see also Smeele 1994). In the current study, two opposing models of information integration are formulated and tested against the observed word recognition results.

Fuzzy logical model of perception (FLMP)

According to the fuzzy logical model of perception (FLMP; Massaro 1987, 1993, 1998, 2002; Oden and Massaro 1978), speech perception can be described by three temporally overlapping stages of information processing: *evaluation*, *integration*, and *decision*. During evaluation, multiple sources of information are continuously matched against stored prototypes in memory. The representational structure of a prototype is conceptualized as a conjunction of auditory and visual features made available by the auditory and visual sources of information. During evaluation, the degree of support of a particular feature for each relevant alternative is computed. The degree of support is expressed in terms of a fuzzy truth value that lies between zero (false) and one (true). A fuzzy truth value represents the degree to which a feature in a given stimulus matches a particular response alternative. As such, fuzzy truth values provide a common metric for representing this degree of match of different features to response alternatives (Massaro 1987).

The integration stage consists of a multiplicative combination of all of the feature values supporting a given alternative. The output of the integration stage gives the overall degree of support for each relevant response alternative. The final stage, decision, computes the total support for a given word alternative divided by the sum of the support for all relevant alternatives. This decision stage is formally equivalent to Luce's (1959) choice rule and has been termed a relative goodness rule (Massaro and Friedman 1990). The assumptions lead to the prediction that the least ambiguous source of information will have the greatest influence on performance. It has been shown that the FLMP is mathematically equivalent to Bayes' theorem, an optimal algorithm for using sources of evidence to test a hypothesis (Massaro 1987, 1998).

Given the FLMP framework, we are able to make an important distinction between "information" and "information processing." 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 provided by each source is defined as information and represents the degree of support for each alternative from each modality. In a test of the FLMP, for example, the predicted response probability in the unimodal condition is predicted to be the information given by that stimulus. These values represent how informative each source of information is. Information processing refers to how the sources of information are processed. In the FLMP, this processing is described by the evaluation, integration, and decision stages.

Most of the experiments in audiovisual speech that the FLMP has been successful in describing involve syllable recognition. It is important to extend tests of the model to word recognition, and we chose the gating task for such a test. The gating task involves an open-ended set of possible responses whereas the FLMP has traditionally predicted performance of a finite number of response alternatives (Massaro 1987). The model has been extended to accuracy of word recall from memory (Massaro et al. 1991; Weldon and Massaro 1996), however, and the current study uses this formulation to make quantitative predictions.

Extending the FLMP to the gating task, there are three possible sources of information contributing to

correct word identification: auditory information, visual information, and a constant background source of information. The background source of information corresponds to all of the information that is not in the stimulus presentation and is necessary to assume when extending the model to directly predict accuracy of performance. In a two-alternative task, such as the responses /ba/ and /da/, the model predicts the proportion of each relevant response. In the absence of any information, the degree of support for each response alternative will be 0.5.

However, in a open-ended task with numerous response alternatives, the absence of information will be more likely to support incorrect responses. As such, it is necessary to assume a background degree of support that is less than 0.5 for a correct response and greater than 0.5 for an incorrect response. Therefore, the overall degree of support for a correct word identification in the auditory condition, $S(\text{Correct} \mid A_i)$, is equal to

$$S(\text{Correct}|A_i) = ba_i \tag{1}$$

where b represents the background source of information and i indexes the levels of the auditory source of information. The overall degree of support for an incorrect word identification, $S(\text{Incorrect}|A_i)$, can be represented as the additive complements of the parameters supporting the correct response:

$$S(\text{Incorrect}|A_i) = (1-b)(1-a_i) \tag{2}$$

Inserting these degrees of support for the correct and incorrect response alternatives into the relative goodness rule for proportion correct word identification, $P(\text{Correct}|A_i)$, gives

$$P(\text{Correct}|A_i) = \frac{ba_i}{ba_i + (1-b)(1-a_i)} \tag{3}$$

Exactly analogous predictions exist for performance given just the visual information, with support v_j rather than a_i.

For the bimodal condition, the overall degree of support for a correct word identification, $S(\text{Correct} \mid A_iV_j)$, is equal to

$$S(\text{Correct}|A_iV_j) = ba_iv_j \tag{4}$$

where b represents the background source of information and i and j index the levels of the auditory and visual sources of information, respectively. The parameters representing the overall degree of support for an incorrect word identification can be represented as the additive complements of the parameters for the correct response:

$$S(\text{Incorrect}|A_iV_j) = (1-b)(1-a_i)(1-v_j) \tag{5}$$

Inserting these degrees of support for the correct and incorrect response alternatives into the relative goodness rule for proportion correct word identification, $P(\text{Correct} \mid A_iV_j)$, gives

$$P(\text{Correct}|A_iV_j) = \frac{ba_iv_j}{ba_iv_j + (1-b)(1-a_i)(1-v_j)} \tag{6}$$

It is assumed that each gating interval gives unique information for each modality. Fitting the FLMP to the current gating task with eight gating intervals, therefore, requires 17 free parameters: eight visual, eight auditory, and one for the background source of information. To test a competing prediction that the impact of a source of information remains constant, we now develop an additive model of perception (ADD).

Additive model (ADD)

A second model that will be tested against the results is an ADD (Massaro 1998; Cutting et al. 1992). The ADD is similar to the FLMP at the evaluation stage in that multiple sources of information are independently evaluated, but the information sources are added at the integration stage. The amount of informational support for the correct response alternative is considered to be the simple addition of the background, auditory, and visual sources of information. Given just the visual information, the support for a correct response, $S(\text{Correct} \mid V_j)$, is predicted to be

$$S(\text{Correct}|V_j) = b + v_j \tag{7}$$

where b is the parameter representing a background source of information, and j indexes the levels of visual information. The support for an incorrect identification is

$$S(\text{Incorrect}|V_j) = (1-b) + (1-v_j) \tag{8}$$

Additive integration with a relative goodness rule at decision reduces to a simple averaging model (Massaro 1987). Inserting these degrees of support for the correct and incorrect response alternatives into the decision rule predicts the proportion correct word identification

$$P(\text{Correct}|V_j) = \frac{b + v_j}{b + v_j + (1-b) + (1-v_j)} = \frac{b + v_j}{2} \tag{9}$$

Exactly analogous predictions exist for performance given just the auditory information, with support a_i rather than v_j.

For the bimodal condition, the overall degree of support for a correct word identification, $S(\text{Correct} \mid A_iV_j)$, is equal to

$$S(\text{Correct}|A_iV_j) = b + a_i + v_j \tag{10}$$

where b is the parameter representing a background source of information, and i and j index the levels of auditory and visual information, respectively. The support for an incorrect identification is

$$S(\text{Incorrect}|A_iV_j) = (1-b) + (1-a_i) + (1-v_j) \tag{11}$$

Additive integration with a relative goodness rule at decision reduces to a simple averaging model (Massaro 1987). Inserting these degrees of support for the correct and incorrect response alternatives into the decision rule predicts the proportion correct word identification

$$P(\text{Correct}|A_iV_j) = \frac{b + a_i + v_j}{b + a_i + v_j + (1-b) + (1-a_i) + (1-v_j)} = \frac{b + a_i + v_j}{3} \quad (12)$$

Similar to the FLMP, fitting the ADD to the current gating task with eight gating intervals requires 17 free parameters: eight visual, eight auditory, and one for the background source of information. The FLMP and ADD make different predictions, as indicated by Eqs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12. As noted earlier, the FLMP predicts that one source of information has a greater effect on perceptual identification when the other source is ambiguous. The ADD, on the other hand, predicts that the contribution of one source of information is independent of the ambiguity of the other source. Including a decision stage in the ADD reduces it to an averaging model (Massaro 1987, 1998). For this reason, the ADD predicts that the bimodal condition can never be larger than in either of the unimodal conditions. In contrast, the FLMP predicts an advantage of the bimodal condition over either unimodal one.

Methods

Participants

Twelve and 16 undergraduate psychology students participated in the two experiments for course credit. All participants were native English speakers and reported normal hearing ability and normal or corrected-to-normal vision. All participants were tested individually in sound-attenuated rooms.

Materials/apparatus

The test stimuli consisted of 120 one-syllable English words taken from the Bernstein and Eberhardt (1986) lipreading corpus. These words were selected from this corpus based on two criteria. First, all words were consonant-vowel-consonant (CVC) words. Second, to keep the dynamics and duration of the vowels relatively constant, words were selected as much as possible not to contain a diphthong—a vowel combination usually involving a quick but smooth movement from one vowel to another (only four words in the stimulus set contained a diphthong). To allow a direct comparison across the three presentation conditions, each unique word was presented in auditory, visual, and bimodal forms at the same gating duration. Thus, there were 15 different words at each of the eight gating durations. The gating durations were the proportion of the entire stimulus word measured from the beginning of the stimulus word. These gating durations were chosen after initial pilot work determined accuracy of unimodal visual identification to be near 0 at gate durations up to 0.40. Consequently, the gating durations used were 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, and 0.80 of the stimulus word.

Previous gating studies have demonstrated that the presentation of only roughly half of a CVC stimulus word will result in high accuracy in identifying the stimulus word due to the presence of earlier coarticulatory information about the final consonant (Warren and Marslen-Wilson 1987, 1988). We added auditory noise to keep performance in the auditory conditions relatively inaccurate at these gating durations to optimize the potential impact of the visual information. The words were presented in experiment 1 at a signal/noise ratio of 64.4/60.0 dB SPL (weight B, slow), which decreased the amount of auditory information available while allowing for maximal visual information across the allotted duration range. Past research has shown that presenting auditory speech in noise lowers correct recognition with the largest interference on resolving the place of articulation of the speech segments (Miller and Nicely 1955; Green and Kuhl 1989).

Each word was randomly selected to be at one of the eight duration levels at each of the three presentation conditions. Each word at each modality was presented in a separate block of 120 trials. The experiment consisted of three blocks (sessions) of 120 test words preceded by five practice words. The order of stimulus presentation was the same for all participants. The stimuli were presented via a computer-controlled SONY LDP-1500 laser disk player on NEC C12-202A 12-inch color terminals.

The second experiment was identical to the first experiment except that the speech signal was bandpass filtered between 200 to 3,600 kHz (KH Krohn—Hite Model 3500) and it was presented in a background of speech-based noise. The purpose of bandpass filtering the speech was to decrease the accuracy on the auditory trials somewhat more than in the first experiment to provide a second test at a different level of overall performance. Bandpass filtering the speech lowered the signal/noise ratio to 60.0/58.9 dB SPL (weight B, slow).

Procedure

There were three sessions of 120 test trials. Each session began with five practice trials with different words but with the same characteristics as the test words. A short 5-min break was given at the end of each session. All participants were instructed that on each trial they would hear, see, or hear and see single-syllable English

words presented at different levels of completeness. They were instructed to type their best guess as to the identity of the stimulus word on the basis of the partial auditory, visual, or auditory–visual information. All responses were typed on Televideo TVI-950 computer terminals. Each participant had 8 s to type in a word response; if the response was not completed in this time interval, the computer would beep, signaling the participant to enter their response.

Results

Figures 1 and 2 show the average proportion of correct word identifications as a function of condition in the first and second experiments, respectively. As can be seen in these figures, performance improved with increases in the gating duration and performance in the auditory condition was significantly better than in the visual condition. Importantly, performance in the bimodal condition significantly exceeded that of the unimodal conditions. Given that this result can be predicted by the FLMP but not by the ADD, it provides strong qualitative support for the FLMP over the ADD. An analysis of variance (ANOVA) was carried out on the proportion correct identification with duration and modality as independent variables. For both experiments, the main effects for duration and modality, and the interaction, were found to be significant, $P < 0.0001$. Analyses were also carried out on the proportion correct identification of the initial consonant, vowel, and final consonant. Again, the main effects for duration and modality, and the interaction of duration with modality, were all found to be significant ($P < 0.001$) for each of the three segments.

Model fits to proportion correct word identification

The central purpose of this study was to examine the nature of the integration algorithm in audiovisual speech perception as information increased with increases in the gating duration. The FMLP and ADD were fit to the results of each of the individual participants. The models are represented in a parameter estimation program (STEPIT, Chandler 1969) as a set of prediction equations that iteratively minimize the difference between the observed and predicted results (Massaro 1998). The root mean square deviation (RMSD) provides a measure of the overall goodness of fit of a particular model. A smaller RMSD value indicates that a model explains the data better relative to a model with a larger RMSD value.

As can be seen by the close correspondence between the observed points and predicted lines in Figs. 1 and 2, the FLMP fit the observed data extremely well with an average RMSD across the individual fits of 0.0319 for experiment 1, and 0.0422 for experiment 2. In contrast, the ADD provided a very poor description of the results with an average RMSD of 0.1706 for experiment 1, and 0.1558 in experiment 2. A one-way ANOVA carried out on the RMSD values for each participant for each model showed that the FLMP provided a significantly better fit than the ADD for both the first experiment: $F(1,11) = 184$, $P < 0.001$, and the second experiment: $F(1,15) = 268$, $P < 0.001$.

Given the poor fit of the ADD, we modified this model to eliminate the decision stage and consequently to make it a pure additive model. In this formulation, the predicted performance on visual alone and bimodal trials would be given by the numerators of Eqs. 9 and 12, respectively. This revised model gave a significant improvement in its description of the results with

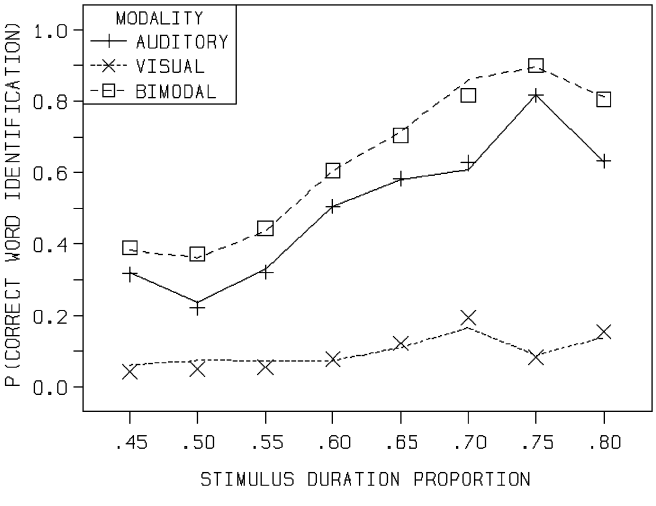

Fig. 1 Mean proportion correct word identification as a function of stimulus duration and modality for experiment 1. The *points* represent the observed results and the *lines* are the predictions of the FLMP. The good fit of the FLMP is indicated by the close correspondence between the observed points and predicted lines

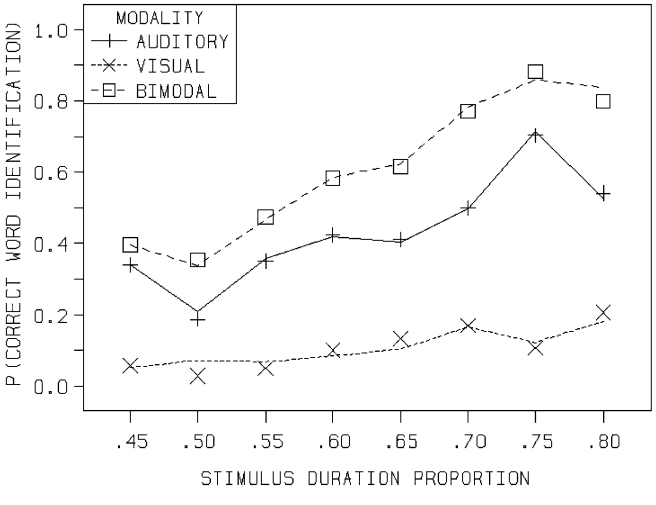

Fig. 2 Mean proportion correct word identification as a function of stimulus duration and modality for experiment 2. The *points* represent the observed results and the *lines* are the predictions of the FLMP. The good fit of the FLMP is indicated by the close correspondence between the observed points and predicted lines

average RMSDs of 0.0454 and 0.0568 for the two experiments, but it still fell significantly below the good fit of the FLMP. A one-way ANOVA carried out on the RMSD values for each participant for each model showed that the FLMP provided a significantly better fit than the modified ADD for both the first experiment: $F(1,11) = 13.7$, $P < 0.005$, and the second experiment: $F(1,15) = 8.56$, $P < 0.01$.

It should be noted that, contrary to expectation, there were a few nonmonotonic points in the performance functions across the gating intervals in Figs. 1 and 2. These probably occurred because different sets of words were used at each gating interval and some sets were slightly more difficult than others. The models are able to predict these nonmonotonicities because unique parameter values are estimated for each gating interval. The data therefore allow a strong test of nature of the processing of auditory and visual speech.

Discussion

The goal of this study was to investigate the nature of the integration algorithm underlying audiovisual speech perception by formally testing opposing models of the integration process as auditory and visual information was systematically varied. The gating paradigm was extended from the area of auditory word recognition to audiovisual speech perception and provided a means whereby information could be manipulated while observing information processing. In principle, neither information nor information processing is directly observable. Formal models must be tested to provide measures of these. An FLMP better accounted for the observed data relative to an ADD.

The FLMP specifies an optimal integration of the auditory and visual information in speech perception. It has been shown that the FLMP is mathematically equivalent to Bayes' theorem (Massaro 1987; Massaro and Friedman 1990), which is an optimal algorithm for combining several sources of evidence to determine the likelihood of one of many hypotheses. In the integration algorithm, the less ambiguous sources of evidence have a larger influence on the outcome relative to more ambiguous sources. More recently, Anastasio et al. (2000) have successfully employed Bayesian statistics to describe multisensory neural information processing.

There is also evidence that the auditory and visual sources in speech perception are complementary. Complementarity refers to the two sources of information as being non-redundant. Auditory and visual contributions to speech are complementary to the extent that a speech distinction is differentially conveyed by these two sources of information. For example, the distinction between /ba/ and /da/ is easy to perceive visually but becomes difficult to distinguish auditorily when the speech is noisy or when the auditory signal is attenuated (e.g., hearing loss). In contrast, the auditory voicing distinction between /ba/ and /pa/ is perceived more robustly but is almost impossible to see (Massaro 1998). Grant and Walden (1996) found that visual enhancement of bandpass-filtered speech was highest when each modality provided complimentary information. Other analyses have shown that the advantage of bimodal speech is due to both the complementarity of auditory and visual information and the optimal integration of these two sources of information (Massaro 1998, pp 424–427).

Visual word recognition performance improves as the duration of the test word is increased but it is also important to note that performance can improve if the duration of a static image is increased (Massaro 1975). Teigland and Wilson (1982), for example, examined the discrimination of certain visemes by tachistoscopically presenting static images of pictures of a female talker saying the phonemes /θ/, /i/, /u/, /p/, and /a/ at ten exposure durations ranging from 12 to 44 ms. The visemes /u/, /p/, and /i/ showed smooth discrimination growth as a function of exposure duration, while /a/ and /θ/ plateaued over several durations. These results show that performance can benefit from increased duration even when the information is static and that the time course of evaluation can differ for different features and segments.

In summary, our results indicate that the gating methodology can be profitably applied to audiovisual speech. By systematically varying the amount of stimulus information, the nature of the integration algorithm operating on this information can be examined. The results supported the FLMP in accounting for correct word identification. These results provide additional evidence regarding the generality of the FLMP in accounting for behavior in a wide variety of tasks and situations (Massaro 1998).

Acknowledgements This work was supported in part by grants from the National Science Foundation (NSF CHALLENGE Grant CDA-9726363 and NSF Grant BCS-9905176), a grant from the Public Health Service (Grant PHS R01 DC00236), cooperative grants from the Intel Corporation and the University of California Digital Media Program (D97-04), and grants from the University of California, Santa Cruz. The authors thank Alexandra Jesse for a critical reading of the manuscript.

References

Anastasio TJ, Patton PE, Belkacem-Boussaid K (2000) Using Bayes' rule to model multisensory enhancement in the superior colliculus. Neural Comp 12:1165–1187

Bernstein LE, Eberhardt SP (1986) Johns Hopkins lipreading corpus video-disk set. The Johns Hopkins University Press, Baltimore

Chandler JP (1969) Finds local minima of a smooth function of several parameters. Behav Sci 14:81–82

Cotton S, Grosjean F (1984) The gating paradigm: a comparison of successive and individual presentation formats. Percept Psychophys 35:41–48

Cutting JE, Bruno N, Brady NP, Moore C (1992) Selectivity, scope, and simplicity of models: a lesson from fitting judgments of perceived depth. J Exp Psychol 121:362–381

Grant KW, Walden BE (1996) Evaluating the articulation index for auditory–visual consonant recognition. J Acoust Soc Am 100:2415–2424
Green KP, Kuhl PK (1989) The role of visual information in the processing of place and manner features in speech perception. Percept Psychophys 45:34–42
Grosjean F (1980) Spoken word recognition processes and the gating paradigm. Percept Psychophys 28:267–283
Grosjean F (1997) Gating. In: Grosjean F, Frauenfelder UH (eds) A guide to spoken word recognition paradigms. Psychology, Sussex
Luce RD (1959) Individual choice behavior. Wiley, New York
Marslen-Wilson WD, Warren P (1994) Levels of perceptual representation and process in lexical access: words, phonemes, and features. Psychol Rev 101:653–675
Massaro DW (1975) Understanding language: an information processing analysis of speech perception, reading, and psycholinguistics. Academic, New York
Massaro DW (1987) Speech perception by ear and eye: a paradigm for psychological inquiry. Lawrence Erlbaum Associates, Hillsdale
Massaro DW (1993) Broadening the domain of the fuzzy logical model of perception. In: Pick HL Jr, Van Den Broek P, Knill DC (eds) Cognition: conceptual and methodological issues. APA, Washington, D.C., pp 51–84
Massaro DW (1998) Perceiving talking faces: from speech perception to a behavioral principle. MIT Press, Cambridge
Massaro DW (2002) Multimodal speech perception: a paradigm for speech science. In: Granstrom B, House D, Karlsson I (eds) Multilmodality in language and speech systems. Kluwer Academic, Dordrecht, The Netherlands, pp 45–71
Massaro DW, Friedman D (1990) Models of integration given multiple sources of information. Psychol Rev 97:225–252
Massaro DW, Weldon MS, Kitzis SN (1991) Integration of orthographic and semantic information in memory retrieval. J Exp Psychol Learn Mem Cogn 17:277–287
McQueen JM, Norris D, Cutler A (1999) Lexical influence in phonetic decision making: evidence from subcategorical mismatches. J Exp Psychol Hum Percept Perform 25:1363–1389
McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:744–748
Miller GA, Nicely P (1955) An analysis of perception confusions among some English consonants. J Acoust Soc Am 27:338–352
Oden GC, Massaro DW (1978) Integration of featural information in speech perception. Psychol Rev 85:172–191
Salasoo A, Pisoni DB (1985) Interaction of information in spoken word identification. J Mem Lang 24:210–231
Smeele PT (1994) Perceiving speech: integrating auditory and visual speech. Unpublished doctoral dissertation, Delft University of Technology
Smits R, Warner N, McQueen JM, Cutler A (2003) Unfolding of phonetic information over time: a database of Dutch diphone perception. J Acoust Soc Am 113:563–574
Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215
Summerfield AQ (1979) Use of visual information in phonetic perception. Phonetica 36:314–331
Teigland AD, Wilson WR (1982) Visual backward masking of selected visemes. J Speech Hear Res 25:269–274
Tyler LK (1985) The structure of the initial cohort: evidence from gating. Percept Psychophys 36:417–427
Walley A, Michela V, Wood D (1995) The gating paradigm: effects of presentation format on spoken word recognition by children and adults. Percept Psychophys 57:343–351
Warren P, Marslen-Wilson WD (1987) Continuous uptake of acoustic cues in spoken word recognition. Percept Psychophys 41:262–275
Warren P, Marslen-Wilson WD (1988) Cues to lexical choice: discriminating place and voice. Percept Psychophys 43:21–30
Weldon MS, Massaro DW (1996) Integration of orthographic, conceptual, and episodic information on implicit and explicit tests. Can J Exp Psychol 50:72–85