Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables *

Dominic W. Massaro and Michael M. Cohen

Program in Experimental Psychology, University of California, Santa Cruz, CA 95064, USA

Abstract. Subjects naturally integrate auditory and visual information in bimodal speech perception. To assess the robustness of the integration process, the relative onset time of the audible and visible sources was systematically varied. In the first experiment, bimodal syllables composed of the auditory and visible syllables /ba/ and /da/ were present at five different onset asynchronies. The second experiment replicated the same procedure but with the vowels /i/ and /u/. The results indicated that perceivers integrated the two sources of information at all asynchronies. Cluster responses (for example, /bda/ given visual /ba/ and auditory /da/) occurred primarily for the consonants but not for the vowels. In addition, cluster responses require that both the visual and the auditory information be reasonable compatible with the physical properties of a cluster articulation. For both vowels and consonant-vowel syllables, information from the auditory and visual sources is continuous, independent and combined in a three-stage process of feature evaluation, integration and decision.

Zusammenfassung. Menschen integrieren automatisch die optischen und akustischen Informationen in einer bimodalen Erkennung. Um die Beständigkeit des Integrationsprozesses zu bewerten, wurden die Zeitpunkte für die Auslösung der akustischen und optischen Quellen systematisch verändert. Im ersten Test zeigten bimodale Silben, die aus den auditiven und optischen Versionen der Silben (ba) and (da) bestanden, fünf verschiedene asynchrone Werte. Im zweiten Test wurde der erste kopiert, aber die Vokale i und u verwendet. Die Ergebnisse zeigen, daß die Personen bei ihrer Aufnahme die zwei Informationsquellen für alle asynchronen Werte integrieren. Die gruppierten Antworten, z.b. /bda/ für eine optische Quelle, /ba/ und eine auditive Quelle /da/) treten hauptsächlich bei Konsonanten und nicht bei Vokalen auf. Außerdem erfordern diese gruppierten Antworten, daß die optischen und auditiven Informationen mit den physischen Eigenschaften der Artikulierung einer Konsonantengruppe kompatibel sind. Sowohl für die Vokale als auch für die Silben mit Konsonanten und Vokalen wird die Information der auditiven und optischen Quellen kontinuierlich und unabhängig in einem Verarbeitungsverfahren in drei Ebenen kombiniert: Bewertung, Integration und Entscheidung.

Résumé. Les sujets intègrent naturellement les informations auditives et visuelles en perception bimodale. Pour évaluer la robustesse de ce processus d'intégration, on a fait varier systématiquement les instants de déclenchement de sources auditives et visuelles. Dans la première expérience, des syllabes bimodales composées des versions auditives et visuelles des syllabes /ba/ et /da/ on été présentées avec cinq valeurs différentes d'asynchronie. La deuxième expérience dupliquait la première en utilisant les voyelles /i/ et /u/. Les résultats montrent que les sujets intègrent dans leur perception les deux

---

* Hiroya Fujisaki has been a source of penetrating insight and understanding of the important issues in spoken language understanding. Although his discipline is engineering, his contributions have advanced the linguistic and psychological underpinnnings of how we process speech. His work on memory contributions to the perception and discrimination of consonants and vowels remains as one of the milestones in the development of our field. Before this research, students of speech science did not seriously consider the variety of psychological processes involved in the commonly accepted tasks used to study speech perception. When a group of students and I set the goal of understanding language from an information-processing perspective, Fujisaki's research and theory were an important source of inspiration (Massaro, 1975). In addition to our respect for Hiroya-san as a scholar, he is a wonderful and kind gentleman and a model representative of his country and culture. We dedicate this paper to him and anticipate his continued leadership in the field.

sources d'information pour toutes les valeurs d'asynchronie. Les réponses groupées (par exemple /bda/ pour une source visuelle /ba/ et une source auditive /da/) apparaissent essentiellement pour les consonnes et non pour les voyelles. De plus, ces réponses groupées nécessitent que les informations visuelle et auditive soient, chacunes, raisonnablement compatibles avec les propriétés physiques d'une articulation de groupe consonantique. Tant pour les voyelles que pour les syllabes consonnes-voyelles, l'information provenant des sources auditive et visuelle est continue, indépendante et combinée dans un processus de traitement des traits à trois niveaux: évaluation, intégration et décision.

Keywords. Bimodal speech perception; lipreading; auditory-visual temporal asynchrony; phonetic classification.

1. Introduction

Fujisaki recognized that spoken language is much more than auditory speech. Consistent with his view, we know that visible speech from the talker's face is a potentially valuable source of information in speech perception (Dodd and Campbell, 1987; Summerfield, 1991). When this visible speech is added to auditory speech, people find it naturally to perceive bimodally – that is, to use both the audible and visible information. The influence of visible speech might be considered surprising because only a subset of speech distinctions is carried by visible speech. Information about place of articulation and duration are visible to some extent, whereas voicing and manner tend not to be visible. Thus, for example, differences between /ba/ and /da/ are visible but not those between /ba/, /ma/ and /pa/. However, perceivers have no qualms about exploiting a source of information even though it is not a sure thing.

Recent research has clarified the processes involved in bimodal speech perception (Massaro, 1987, 1989; Massaro and Cohen, 1990). In (Massaro and Cohen, 1983) subjects identified speech events consisting of synthetic auditory syllables varying along a continuum from /ba/ to /da/ combined with a videotape of a person articulating a /ba/ or /da/ syllable or with no visible articulation. Although subjects were instructed specifically to report what they heard, the visible information also influenced identification. The points in Figure 1 give the observed proportion of identifications as a function of the auditory syllable; the visual condition is the curve parameter. As can be seen in the figure, subjects tended to identify the syllables as one of five alternatives: /ba/, /da/, /bda/, /ð/ and /va/. These five alternatives accounted for 95.7% of the judg-

ments. It is clear from the results in Figure 1 that both modalities contributed to the judgments. For example, a visible /ba/ articulation increases the likelihood of identifying a /ba/-like auditory syllable as /ba/.

The design and analysis of the Massaro and Cohen (1983) experiment allowed the test of competing quantitative models of the results. The fuzzy logical model of perception (FLMP) is based on three operations in perceptual recognition: feature evaluation, feature integration and decision (Massaro, 1987). It is assumed that independent, continuously-valued features are evaluated, integrated and matched against prototypical descriptions of syllables in memory. The integration process combines the auditory and visual modalities as independent sources of evidence for the occurrence of syllable prototypes. An identification decision is made on the basis of the relative goodness of match of the stimulus information with relevant prototype descriptions. The predic-
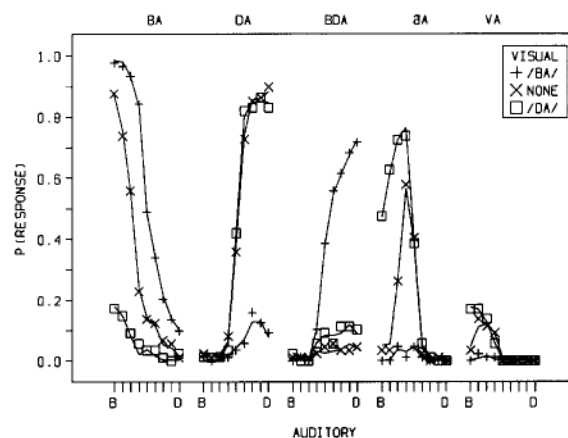

Fig. 1. Proportion of observed (points) and predicted (lines) identifications as a function of the auditory and visual sources of information (results of 8 subjects (Massaro and Cohen, 1983)).

tions of the FLMP, given by the lines in Figure 1, provide a good description of the observed results (Massaro, 1987, pp. 128–129). The FLMP gave a significantly better fit than a model in which the auditory and visual sources were perceived categorically.

The nature of the judgements as a function of the two sources of information reveals the sophistication of bimodal speech perception. Although a visible /ba/ articulation increases the likelihood of identifying a /ba/-like auditory syllable as /ba/, a visible /da/ decreases /ba/ judgments for these same auditory syllables (see Figure 1). In these conflicting cases, subjects tend to identify the syllables as /va/ or /ð/. These latter judgments are reasonable because they make a relatively good match with *both* the auditory and the visual information. Judgments /ba/ and /da/ are not good solutions to the conflicting sources of information because of the complete mismatch with one of the two sources. Figure 1 also shows that a visible /ba/ decreases the likelihood of a /da/ judgment given an auditory syllable from the /da/ end of the continuum. Subjects tend to respond with /bda/ in this situation.

The research on bimodal speech perception illustrates the ease with which perceivers integrate auditory and visual speech. There is a growing body of literature supporting the robustness of this integration process. Integration appears to occur when the two sources of auditory and visual information come from synthetic and/or natural speech (Massaro and Cohen, 1990, unpublished results), different spatial locations (Fisher, 1991) and different genders (Green et al., 1991). In the present experiments, the temporal occurrence of the visual and auditory syllables will be varied. One question of interest is to what extent integration occurs when the two modalities are presented asynchronously. We expect the two modalities to be integrated even if their temporal onsets are not perfectly aligned.

McGrath and Summerfield (1985) added a soundtrack of rectangular pulses synchronized to the closing of the talker's vocal track. This soundtrack facilitated identification of the visible sentences. If the soundtrack was delayed 160 msec, however, performance was equivalent to the vi-

sual-alone condition. Delays of 80 msec or less did not diminish the contribution of the soundtrack. This result shows that subjects benefit from having two sources of information relative to just one, and this advantage can be eliminated if the asynchrony between the sources is too large. In a second study, subjects had to judge whether a complex tone started before or after the opening of a pair of liplike Lissajous figures. Subjects were able to judge asynchrony when the auditory stimulus preceded the visual by about 80 msec and when the visual preceded the auditory by about 140 ms.

2. Experiment 1

In the first experiment, the auditory and visual syllables /ba/ and /da/ were presented either in agreement or in conflict. The two modalities were either presented simultaneously or with a temporal offset between the two. Subjects were asked to judge what they heard on each trial. This manipulation should provide some information on the robustness of integrating auditory and visual speech across small temporal asynchronies. In addition, this experiment addresses the relative importance of two contributions to cluster responses. The first is the relative compatibility of each of the sources of information with each possible response. The second is the temporal asynchrony between the auditory and visual sources of information. Figure 1 shows that a visual /ba/ paired with an auditory /da/ often produces the response /bda/. Both sources are reasonably compatible with the properties of /bda/. The visual information resulting from a /ba/ articulation is compatible with a /bda/ articulation under ordinary conditions of illumination. An auditory /da/ is also somewhat compatible with the auditory characteristics of the alternative /bda/. Thus, a /bda/ perceptual judgment occurs fairly frequently given these two sources of information. There is also some evidence that visual /ba/ is identified somewhat sooner than auditory /da/ (Cohen, 1984; Massaro, 1987, p. 160). If differences in the time of arrival are sufficient for a cluster identification, then it should be possible to produce /dba/

judgments by presenting a visual /da/ before an auditory /ba/. The visual characteristics of a visual /da/, however, differ greatly from those for /dba/. If compatibility between the stimulus sources and the response alternative is critical, /dba/ judgments should not occur even if visual /da/ occurs somewhat earlier than auditory /ba/.

2.1. Method

2.1.1. Subjects

Twelve students from introductory psychology and psychology statistics classes served as subjects. Half of the group served in this experiment before serving in another study on bimodal speech perception. For the other half of the group, the order was reversed. Some of the subjects participated for class credit and the remainder were paid 5 dollars for their time.

2.1.2. Stimuli

Prior to the experiment, a video-audio master tape was recorded. A speaker (the senior author) was seated in front of a wood panel background, illuminated with ordinary fluorescent fixtures in the ceiling. The speaker's head was centered in the video field and filled about 2/3 of the frame in the vertical direction. Both the video and the audio were recorded on a Sony AV 2800 3/4″ U-matic color cartridge video tape recorder, using a Panasonic PK-802 color camera with a f1.4 12–72 mm 6:1 zoom lens from a distance of about 1.5 meters. On each trial the speaker said either /ba/ or /da/ as cued by a video terminal under control of a DEC PDP-11/34a computer. Each recording trial started with a 400 msec warning tone presented by the emitter in the terminal keyboard base. One hundred (100) ms after the offset of the warning tone, the computer cued the speaker by displaying either a large lower case b or d on the screen for 500 msec. Following the offset of the cue there was a 2000 ms period before the next warning tone. The cues were presented in 13 blocks of 20 stimuli for a total of 260 trials. The first block of 20 was recorded to provide practice trials. Within each block of 20, randomization without replacement

was done for three variables: 2 levels of visual information, 2 levels of auditory information and 5 levels of (SOA). The tape was reviewed to ensure that the stimuli were created according to the cues.

An experimental tape was created from the 3/4″ master tape on a Panasonic NV-8200 1/2″ VHS video tape recorder according to the following design. The auditory test syllables were the synthetic endpoint stimuli from the ba/-/da/ continuum from (Massaro and Cohen, 1983). On each trial of the experimental tape, a visual /ba/ or /da/ was combined with an auditory /ba/ or /da/ with a SOA of −200, −100, 0, 100 or 200 ms between the auditory stimulus and the voice onset of the original audio. As noted in the Introduction, subjects in the (McGrath and Summerfield, 1985) study were able to judge asynchrony when the auditory stimulus preceded the visual by about 80 ms and when the visual preceded the auditory by about 140 ms. Thus, the subjects might have noticed an asynchrony between the auditory and visual stimuli on some of these trials, but we do not know how this would have influence the results.

The synthetic speech replaced the original audio track on the videotape. The original audio speech signal was monitored by a Schmitt trigger circuit. When the original audio channel on the videotape exceeded a preset threshold, one of the synthetic syllables was recorded onto the experimental tape. The timing of the SOA was accomplished by, prior to the dubbing, measuring the time from the onset of the warning tone to voice onset of the original syllable. This measurement was done by monitoring the audio portion of the master tape with a Schmitt trigger circuit connected to the programmable clock of the computer. During the creation of the experimental tape, the time for presentation of the auditory stimulus stimulus with reference to the warning tone onset could be calculated by adding the intended SOA to the previously measured SOA between the warning tone and original voice onset. Given that the original vowels triggered the circuit at their onset, the dubbing was accurate to within 1 ms. The synthetic speech was played at a rate of 10000 samples per second and filtered 20–4900 Hz (KHRON-HITE 3500R).

2.1.3. Procedure

During the experiment, the experimental tape was played to the subjects over individual NEC model C12-202A 12″ inch color monitors. Four subjects could be tested simultaneously in individual sound attenuated rooms. These rooms were each illuminated by two 60 watt incandescent bulbs in a frosted glass ceiling fixture.

The audio portion of the experimental tape was presented to the subjects over the built-in speakers of the monitors at a comfortable listening level of about 67 dB-A measured at the approximate position of the observers head using a B&K 2123 sound level meter with a 4134 microphone. The audio signal from the videotape was monitored by the Schmitt trigger of the programmable clock of the DEC PDP/11/34a computer. This trigger information was used by the computer to sense the beginning of a response interval. The subjects had 3000 msec to make their response by pressing buttons labeled /da/, /ba/, /va/, /ð/, /ga/, /dba/ and "other" on a detached keyboard of a Televideo TVI-950 video terminal.

Each subject was instructed:

"...you will watch a speaker and listen to what is spoken. Your task will be to identify what you heard. On each trial, you will look at the TV monitor and listen to a speech sound coming from the TV. Your task will be to identify what you heard."

A single session of 20 practice trials plus 240 recorded trials was run, giving 12 observations per subject for each of the 20 unique experimental conditions.

2. Results and discussion

Figure 2 gives the probability of each of the eight possible responses as a function of stimulus onset asynchrony (SOA) for each of the four possible combinations of the auditory and visual variables. An analysis of variance was carried out on these data with subjects, visual and auditory levels, SOA, and response as factors. This analysis revealed significant effects of response, response as a function of the auditory level, response as a function of SOA, and the three-way
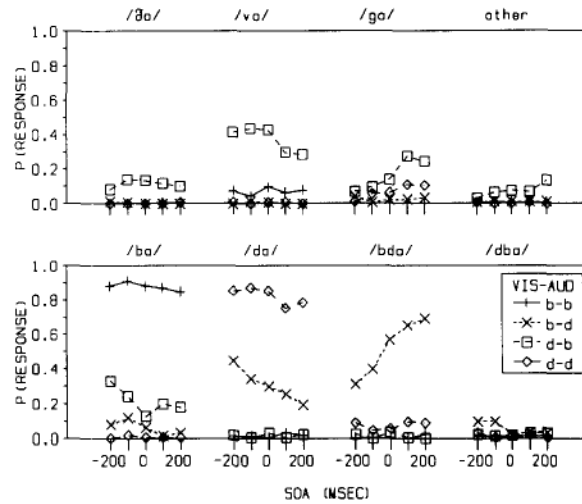

Fig. 2. Proportion of responses for Experiment 1 as a function of SOA: visual-auditory level is a curve parameter.

interaction of response, auditory level and visual level. Two other three way interactions, response by visual by SOA, and response by auditory by SOA, were significant, along with the four-way interaction of response by visual by auditory by SOA.

When the visual and auditory components had the same identity, the responses mostly agreed with that identity (/ba/ 87.8%, /da/ 82.3%) regardless of SOA. This result shows the robustness of bimodal speech perception across differences in SOA. Interesting results also occurred when the visual and auditory components disagreed. When visual /ba/ was combined with auditory /da/, the predominant responses were /bda/ (52.4%) and /da/ (30.8%). The large number of /bda/ judgments is consistent with the idea that visual /ba/ is highly similar to a visual /bda/ articulation. The /bda/ responses increased from 31.1% to 69.0% as the visual lead went from −200 to 200 msec, while the /da/ responses decreased from 44.7% to 19.3%. The increase in /bda/ judgments as the visual /ba/ was presented earlier than the auditory information is consistent with the hypothesis that cluster responses occur because one modality is processed sooner than the other.

The other conditions indicate that the temporal difference between the two modalities is not sufficient to produce cluster responses. When

visual /da/ was combined with auditory /ba/, /dba/ judgments did not occur. Furthermore, the number of /dba/ judgments did not increase when the visual /da/ was presented earlier than the auditory /ba/. The reason is that a visual /da/ is highly dissimilar to a /dba/ articulation. The most frequent responses given a visual /da/-auditory /ba/ were /va/ (37.1%), /ba/ (21.6%), /ga/ (16.5%) and /ða/ (11.6%). Given a visual /da/ paired with an auditory /ba/, /ba/ and /va/ responses decreased somewhat while /ga/ responses increased, to the extent that the visual /da/ preceded the auditory /ba/. The /ða/ responses were mostly flat with respect to SOA.

Given that predicted /dba/ responses rarely occurred, it is probably not the case that consonant clusters occur simply because of differences in arrival times of the two information sources during perceptual processing. Rather, the results lend support to the hypothesis that clusters occur because both the visual and auditory syllables are consistent with the articulation of a consonant cluster. However, when clusters are consistent with the visual evidence, the temporal relationship between the auditory and visual information can influence the degree to which clusters are perceived. Thus, the number of /bda/ judgments increased as the auditory /ba/ was delayed relative to the visual /ba/.

The cluster responses provide additional evidence for the integration of auditory and visual speech. If subjects did not integrate the two sources of information, but simply used one source or the other on a given trial, then no cluster responses would be expected. Not only do cluster responses occur, they change in a reasonable manner with changes in the SOA. Increasing the SOA between the onset of a visual /ba/ and an auditory /da/ increased the number of /bda/ identifications. This result indicates that both the visual and the auditory information contributes to the perceptual experience. On the other hand, increasing the SOA between the onset of a visual /da/ and an auditory /ba/ did *not* increase the number of /dba/ identifications. The reason for this is that a visual /da/ is very different from a visual /dba/. Thus, the degree of fit of the auditory and visual information is necessary for

cluster judgments, while the temporal occurrence of the information sources can modulate the number of cluster responses.

3. Experiment 2

In the next experiment, the same paradigm will be used to investigate the perception of vowel clusters. Cohen and Massaro (1991) carried out experiments aimed at discovering similarities and differences in the auditory-visual perception of consonants and vowels. The differences between stop consonants and vowels indicated that visible speech may not influence the identification of vowels in the same manner as stop consonants. Thus it was of interest to test whether the same model describing bimodal perception of consonants also applies to vowels. Subjects identified consonant-vowel or vowel syllables varying on both audible and visible speech information. Compared to a categorical model, a single channel model, and a weighted averaging model (all of which are mathematically equivalent), the FLMP gave a much better description of the identification of both vowels and consonants, for both two and eight response alternatives. In addition, the results were used to test the recent prelabeling model (PRLM) of Braida (1991). Although the FLMP and PRLM made almost identical predictions for binary responses, the FLMP provided a significantly better fit to both consonant and vowel experiments with eight response alternatives. For both vowels and consonant-vowel syllables, information from the auditory and visual sources is continuous, independent, and combined in a three stage process of feature evaluation, integration and decision.

Experiment 2 replicates the asynchrony experiment with the vowels /i/ and /u/. Is the integration of auditory and visual information as robust for vowels as it is for consonant vowel syllables? Although consonant clusters have clearly been observed with conflicting auditory and visual stimuli, the same cannot be said for vowels. Summerfield and McGrath (1984) reported a set of experiments on the auditory-visual perception of vowels. In one experiment three vowel continua were used. In separate blocks, 11 member

series between pairs of the three point vowels, /i/, /a/ and /u/ were presented in a /bVd/ environment approximately synchronized with a visual articulation from either end of the given continuum. Although the visible speech influenced the identification judgments, vowel clusters were not among the response set and were not reported.

Experiment 2 explicitly tests whether vowel clusters will occur analogous to those occurring with incongruous auditory-visual consonants. The cluster /ui/ is very similar to the common word "we" while /iu/ closely resembles "you". Thus we can aks whether a visual /u/ combined with a auditory /i/ will produce the judgment /ui/ analogous to /bda/? Both /u/ and /b/ are similar in having somewhat earlier and more salient visual articulations at their onsets. Also, will /iu/ judgments be rare analogous to the rarity of /dba/ judgments?

3.1. Method

3.1.1. Subjects

A group of 12 students from introductory psychology and psychology statistics classes participated. Half of the group served in another experiment after this task. For the other half of the group, the order was reversed. Some of the subjects participated for class credit and the remainder were paid 5 dollars for their time.

3.1.2. Stimuli and procedure

The same recording and testing procedure was used in this experiment as Experiment 1 except that the stimuli were /i/ and /u/ rather than /ba/ and /da/ and the response buttons were labeled as /i/, /u/, /iu/, /ui/, /a/, /I/, /o/ or "other". The latter two are close to /i/ and /u/ respectively in the F_1–F_2 vowel space and there is some indication that subjects made a number of responses in these directions in the (Summerfield and McGrath, 1984) study.

The two auditory test vowels were created using a software formant serial resonator speech synthesizer (Klatt, 1980). The vowel duration was 500 msec. The first three formants were 315, 2245 and 3135 Hz for /i/ and 355, 1187 and 2294 Hz for /u/. The voicing amplitude rose from 24 dB

at time 0 to 50 dB at 70 msec to 60 dB at 150 msec and then fell linearly to 48 db at 450 msec and finally to 0 dB at 500 msec. To keep the overall loudness equal for the two vowels, the voicing amplitudes were adjusted upwards for the stimuli at the /u/ end of the continuum. The pitch of the vowels rose from 150 Hz at time 0 to 172 Hz at 300 msec, remaining at that value until the last 100 msec during which it descended to 160 Hz. The amplitude and pitch contours closely approximated those of the natural vowels of the speaker as analyzed using linear prediction with cepstrally based pitch estimation. The F_4 frequency was fixed at 4100 Hz. The resulting stimuli, quantitized at a 10000 sample per second rate, were stored on the computer disk for later use.

The lips were closed at the beginning and end of each vowel syllable. The mean auditory durations of five tokens each of the /i/ and /u/ visible vowels were 483 and 437 msec, respectively.

3.2. Results and discussion

Figure 3 gives the probability of each of the eight possible responses as a function of SOA for each of the four possible combinations of the auditory and visual syllables. As can be seen in the figure, all of the response judgments were
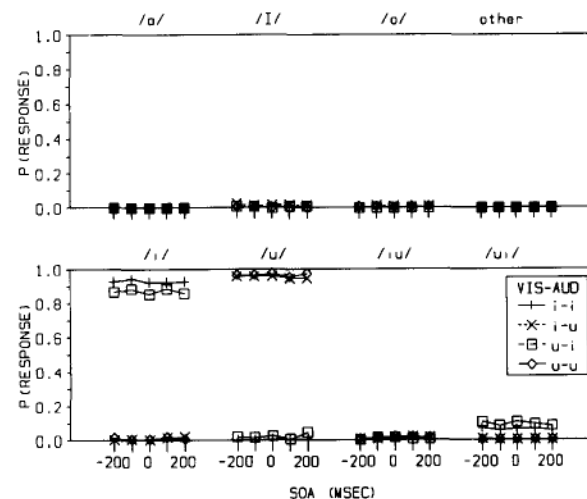

Fig. 3. Proportion of responses for Experiment 2 as a function of SOA: visual-auditory level is a curve parameter.

/i/, /u/ or /ui/. The latter occurred only 4% of the time. An analysis of variance was carried out on these data with subjects, visual and auditory levels, SOA and response as factors. This analysis revealed significant effects of response, response as a function of the auditory level, response as a function of the visual level, and the interaction of response, auditory level and visual level. As can be seen in Figure 3, response as a function of SOA was not significant.

The results showed an influence of auditory and visible speech regardless of the SOA. As can be seen in Figure 3, a visible /u/ increased the number of /u/ judgments at all SOAs. Similarly, a visible /i/ increased the number of /i/ judgments. In contrast to Experiment 1 with consonant vowel syllables, however, cluster judgments were rarely observed.

Although very few clusters were observed resulting from the auditory-visual combination of the vowels /i/ and /u/, cluster responses might occur under other conditions. Perhaps shorter-duration auditory and visual vowels would be more likely to produce the perception of vowel clusters. A faster rate of speaking would perhaps increase the compatibility of the visible vowel gesture and that of a cluster. Thus, a rapidly articulated visual /u/ paired with a short auditory /i/ might be perceived as /ui/ (we). Given the potent influence of other sources of information, the visual /u/ and auditory /i/ placed in an appropriate semantic/syntactic context should increase the perception of "we" even more.

Acknowledgments

References

L.D. Braida (1991), "Crossmodal integration in the identification of consonant segments", *Quart. J. Experimental Psychology*, in press.

M.M. Cohen (1984), Processing of visual and auditory information in speech perception, Dissertation, University of California, Santa Cruz.

M.M. Cohen and D.W. Massaro (1991), *Perceiving visual and auditory information in consonant-vowel and vowel syllables*.

B. Dodd and R. Campbell, eds. (1987), *Hearing by Eye: Experimental Studies in the Psychology of Lipreading* (Lawrence Erlbaum, Hillsdale, NJ).

B.D. Fisher (1991), Integration of visual and auditory information in speech events, Unpublished Doctoral Dissertation, University of California, Sant Cruz.

H. Fujisaki and T. Kawashima (1970), "Some experiments on speech perception and a model for the perceptual mechanism", *Annual Report of the Engineering Research Institute,*, Vol. 29, Faculty of Engineering, University of Tokyo, pp. 206–214.

K.P. Green, P.K. Kuhl, A.N. Meltzoff and E.B. Stevens (1991), "Integration speech information across talkers, gender and sensory modality: Female faces and male voices in the McGurk effect", *Perception & Psychophysics*, Vol. 50, pp. 524–536.

D.H. Klatt (1980), "Software for a cascade/parallel formant synthesizer", *J. Acoust. Soc. Amer.*, Vol. 67, pp. 971–995.

D.W. Massaro, ed. (1975), *Understanding Language: An Information Processing Analysis of Speech Perception, Reading, and Psycholinguistics* (Academic Press, New York).

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

D.W. Massaro (1989), "Multiple book review of Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Behavioral and Brain Sciences", Vol. 12, pp. 741–794.

D.W. Massaro and M.M. Cohen (1983), "Evoluation and integration of visual and auditory information in speech perception", *J. Experimental Psychology: Human Perception and Performance*, Vol. 9, pp. 753–771.

D.W. Massaro and M.M. Cohen (1990), "Perception of synthesized audible and visible speech", *Psychological Sci.*, Vol. 1, pp. 55–63.

M. McGrath and Q. Summerfield (1985), "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults", *J. Acoust. Soc. Amer.*, Vol. 77, pp. 678–685.

Q. Summerfield (1991), "Visual perception of phonetic gestures", in *Modularity and the Motor Theory of Speech Perception*, ed. by I.G. Mattingly and M. Studdert-Kennedy (Lawrence Erlbaum, Hillsdale, NJ), pp. 117–137.

Q. Summerfield and M. McGrath (1984), "Detection and resolution of audio-visual incompatibility in the perception of vowels", *Quart. J. Experimental Psychology*, Vol. 36A, pp. 51–74.