THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 1992, *45A* (1) 1–20

Lexical Mediation Between Sight and Sound in Speechreading

Bruno H. Repp

Haskins Laboratories, New Haven, Connecticut, U.S.A.

Ram Frost

Hebrew University of Jerusalem, Israel

and

Elizabeth Zsiga

Yale University and Haskins Laboratories, New Haven, Connecticut, U.S.A.

In two experiments, we investigated whether simultaneous speech reading can influence the detection of speech in envelope-matched noise. Subjects attempted to detect the presence of a disyllabic utterance in noise while watching a speaker articulate a matching or a non-matching utterance. Speech detection was not facilitated by an audio-visual match, which suggests that listeners relied on low-level auditory cues whose perception was immune to cross-modal top-down influences. However, when the stimuli were words (Experiment 1), there was a (predicted) relative shift in bias, suggesting that the masking noise itself was perceived as more speechlike when its envelope corresponded to the visual information. This bias shift was absent, however, with non-word materials (Experiment 2). These results, which resemble earlier findings obtained with orthographic visual input, indicate that the mapping from sight to sound is lexically mediated even when, as in the case of the articulatory–phonetic correspondence, the cross-modal relationship is non-arbitrary.

The interaction of the visual and auditory modalities in word perception is of interest to psychologists concerned with the nature of the representation of words in the mental lexicon. That such an interaction exists has been

Requests for reprints should be sent to Bruno H. Repp, Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695, U.S.A.

This research was supported by NICHD Grant HD01994 to Haskins Laboratories. A brief report of the results was presented at the 31st Annual Meeting of the Psychonomic Society in New Orleans, LA, November 1990.

demonstrated in many studies. For example, the popular cross-modal semantic priming paradigm (Swinney, Onifer, Prather, & Hirshkowitz, 1979) demonstrates facilitation of lexical access in one modality by the recent occurrence of a related word in the other modality. Visual articulatory information (i.e. a speaker's moving face) has long been known to aid the recognition of spoken words in noise (e.g. O'Neill, 1954; Erber, 1969), and, conversely, auditorily presented speech features that may not be intelligible by themselves can increase word recognition in speechreading (e.g. Breeuwer & Plomp, 1984, 1986). Cross-modal interactions can occur prior to word recognition: printed single letters or non-word letter strings can facilitate the response to a phoneme presented in the auditory modality (Dijkstra, Schreuder, & Frauenfelder, 1989; Layer, Pastore, & Rettberg, 1990). Prelexical cross-modal influences have also been demonstrated when the visual information consists of articulatory gestures (McGurk & MacDonald, 1976): simultaneous presentation of a spoken CV syllable and of a speaker's face uttering a different syllable can lead to the illusion of hearing the syllable suggested by the visual modality. This interaction even takes place prior to the categorization of the phonemes involved (Massaro & Cohen, 1990; Summerfield, 1987).

In a recent study, Frost, Repp, and Katz (1988) investigated whether influences from the visual modality can penetrate to earlier, precategorical levels of auditory perception by requiring their subjects to *detect* rather than recognize speech in noise. Auditory speech-plus-noise and noise-only trials were accompanied by a visual orthographic stimulus that either matched or did not match the masked speech. Frost et al. found that matching visual input did not improve subjects' speech detection performance, which suggested that the information subjects relied on (probably bursts of low-frequency spectral energy) was immune to cross-modal top-down influences. However, the visual input did have a strong effect on the bias parameter in this signal detection task: subjects claimed to hear speech more often when they saw the word to be detected than when they saw a different printed word or no word at all. This *bias shift*, which may represent a genuine perceptual effect (viz. an illusion of hearing speech in noise), was evidently due to the fact that, in that study, the amplitude envelopes of the masking noises had been matched to those of the words to be masked. This so-called signal-correlated noise has very desirable properties as a masking agent (it enables precise specification of the signal-to-noise ratio and keeps that ratio constant as the signal changes over time), but it does retain some speechlike features. Although these features are not sufficient to cause perception of the noise as speech, let alone to identify a specific utterance, they do convey considerable prosodic and phonetic information. More specifically, the amplitude envelope conveys information about the rate of speech (Gordon, 1988), number of syllables (Remez & Rubin, 1990), relative stress (Behne, 1990), and several major

classes of consonant manner (Van Tasell, Soli, Kirby, & Widin, 1987). (See also Smith, Cutler, Butterfield, & Nimmo-Smith, 1989, who employed speech heavily masked by unmodulated noise.) Apparently, the subjects in the Frost et al. (1988) study automatically detected the correspondence between a printed word and an auditory presented noise amplitude envelope. As a result, they perceived the masking noise as more speechlike and concluded that there was "speech in the noise". Frost et al. considered this an interesting and novel demonstration of rapid and automatic phonetic recoding in silent reading: as signal-correlated noise is too impoverished to suggest a definite orthographic representation, the cross-modal correspondence must be established by mapping the print into an internal speechlike representation specific enough to contain amplitude envelope features matching those of the noise and accessed rapidly enough to be linked to the transitory auditory stimulus.

According to many models of visual word recognition, the mapping from print to speech may be accomplished either via stored phonological codes attached to lexical entries or via prelexical spelling-to-sound conversion rules (see Patterson & Coltheart, 1987; Van Orden, Pennington, & Stone, 1990, for reviews). Hence it was especially interesting to find that the bias shift just described was reduced considerably when the materials were meaningless pseudowords (Frost et al., 1988: Exp. 2). Frost (1991) has replicated this finding in the Hebrew orthography, both with and without vowel symbols, using a within-subject design. His results suggest that the stronger bias shift for words than for non-words is independent of spelling-to-sound regularity, and of the speed of processing the printed stimuli. It seems, therefore, that subjects' ability to detect the orthographic–acoustic correspondence in the speech detection paradigm is, at least in part, lexically mediated—that is, when the visual input is a word, it activates a lexical entry and, with it, an internal speechlike representation containing considerable phonetic detail, including amplitude envelope features. In contrast, when the visual input is a non-word, its internal phonetic representation (if any) must be assembled via analogy with known lexical items (Glushko, 1979) or via spelling-to-sound translation rules, and because of this piecemeal construction it may be less coherent or less vivid than the phonetic representation of a familiar word; hence the match with an auditory amplitude envelope is less evident.

Our aim in the present study was to examine further the hypothesis that detailed phonetic information is stored with, or is part of, lexical representations. We conducted two experiments analogous to Experiments 1 and 2 of Frost et al. (1988), but instead of print we employed a video recording of a speaker's face.

Visual articulatory information differs from orthography in several important ways. On one hand, whereas the relations of graphemic forms to phonologic structures are a cultural artifact, the relations of articulatory

movements to phonological and phonetic structure are non-arbitrary. There is a natural isomorphism between visible articulatory movements and some acoustic properties of speech, particularly between the degree of mouth opening and overall amplitude. Therefore, lexical mediation may not be required for viewer–listeners to perceive a correspondence between the timing of opening/closing gestures and variations in signal amplitude.[1] On the other hand, visual articulatory information is less specific than print and generally conveys only distinctions among major consonant and vowel classes, the so-called visemes (see, e.g. Owens & Blazek, 1985). Visually observed speech gestures are often compatible with a number of lexical candidates. It may be hypothesized, therefore, that in order for a speech-read utterance to be associated with the sound of a particular word, lexical access may be necessary, after all. Finally, we must note that articulatory information unfolds over time, whereas print is static and presents all information at once (provided it can be viewed in a single fixation). Thus there is an added dimension of temporal synchrony in audio-visual speech perception, which may enhance the interaction of the two modalities.

These considerations led us to hypothesize that the original finding of lexical mediation in the access of speechlike representations from orthography (Frost et al., 1988) might be replicated when the visual information consists of articulatory gestures: subjects might be able to detect a correspondence between the speaker's gestures and auditory amplitude envelopes, but only when the stimuli are familiar words. In that case, the auditory envelope information would supplement the visual gestural information to constrain word identification.[2] A lexical representation would automatically link two types of information, and a significant increase in perceptual bias on "matching" trials would be the result. However, when the speech-read stimuli are clearly non-words, lexical mediation would not occur, and this might also eliminate the bias shift, if it indeed originates at the lexical level.

Although the bias shift (i.e. the influence of visual information on perception of the *masking noise*) was of primary interest in our study, we

---

[1]Kuhl and Meltzoff (1982) have shown that 18 to 20-week-old infants perceive the correspondence between visually presented /i/ and /a/ articulations and the corresponding speech sounds. However, the infants did not recognize any relationship when the amplitude envelopes of these vowels were imposed on a pure tone, so they probably relied on spectral rather than amplitude information when listening to speech.

[2]It is known from research on possible aids for the hearing-impaired that the auditory speech amplitude envelope, even when carried just on a single pure tone, constitutes an effective supplement to speech reading (Blamey, Martin, & Clark, 1985; Breeuwer & Plomp, 1984, 1986; Grant, Ardell, Kuhl, & Sparks, 1985). Note that in our experiments it does not matter whether or not the word recognized is "correct" (i.e. the one intended by the speaker), as long as it fits both the auditory and the visual information.

also examined whether the detectability of the masked speech signal was influenced by seeing matching articulatory information. Our earlier studies with orthographic stimuli revealed absolutely no change in subjects' sensitivity to masked speech. However, because of the close relationship between visible articulatory information and speech acoustics, and because of the added dimension of audio-visual synchrony, we considered it possible that the speech gestures would aid listeners in separating the speech signal from the accompanying noise.

EXPERIMENT 1

Experiment 1 employed words as stimuli. Because we suspected that low-frequency energy provided the major cues for speech detection, and that utilization of these cues may be insensitive to cross-modal top-down influences, we included in Experiment 1 two auditory conditions, the first employing natural phonated speech and the second using whispered speech, which contains little low-frequency energy. These conditions provide very different opportunities for speech-read information to exert an influence on auditory detection performance, as well as somewhat different amplitude envelopes for individual words to test the generality of the expected bias shift.

Method

Stimuli and Design. The stimuli were 48 disyllabic English words with stress on the first syllable (examples: "mountain", "baby", "canvas", etc.). A female speaker was recorded twice producing these words, once with normal phonation and list intonation, and once in a whisper, with the microphone much closer to her mouth. The first session was also videotaped, with the picture showing a frontal view of the speaker's face. Half the recorded words were used to generate the auditory stimuli. The same 24 words in each production mode (phonated and whispered) were digitized at 20 kHz and low-pass filtered at 9.6 kHz. Signal-correlated noise was generated from each word by a simple procedure that randomly reversed the polarity of half the digital sampling points (Schroeder, 1968). Such noise has exactly the same amplitude envelope as the original signal (obviously, as the envelope is derived from the rectified signal, i.e. regardless of the direction of the sound pressure change) but a flat spectrum, like white noise.[3] Speech-plus-noise stimuli were generated by

[3]Although the noise had a flat spectrum in its digital form, it was output through hardware designed to remove high-frequency pre-emphasis and thus had a sloping spectrum in its acoustic form. For the purpose of the present experiments, this was irrelevant.

adding the digital waveforms of each word and of its signal-correlated noise after multiplying them with weighting factors that added up to 1, so that the overall amplitude of the sum remained virtually the same. Two such weightings were used, which, on the basis of pilot results, were expected to yield detection performance of 70–80% correct. In the phonated condition they corresponded to signal-to-noise (S/N) ratios of −12 and −14 dB. In the whispered condition, which was much more difficult, the S/N ratios used were −4 and −6 dB. All these ratios were well below the speech recognition threshold.

Within each production type (i.e. phonated or whispered) and S/N ratio condition, each of the 24 words appeared 6 times: 3 times as signal-plus-noise and 3 times as signal-correlated noise only. Each of these two auditory presentations occurred in three visual conditions: in the matching condition, the subjects saw the speaker produce the word that had been used to generate the auditory stimulus; in the non-matching condition, they saw the speaker say a different disyllabic word, drawn from the 24 words not used as auditory stimuli; in the neutral condition, they saw the speaker's still face. The 6 audiovisual conditions for each of the original 24 words were distributed across 6 blocks of 24 trials according to a Latin square design. Thus each of the 24 words (in one of its two auditory incarnations) occurred exactly once in each block, and each of the 6 audiovisual conditions occurred 4 times per block (with different words). The 24 trials within each block were randomized. The more difficult condition with the lower S/N ratio always followed that with the higher S/N ratio, with the 144 trials of each following the same sequence. The phonated and whispered conditions also used the same stimulus sequences. The order of these production type conditions was counterbalanced across subjects.

The experimental video tapes were generated as follows: (1) Using professional video dubbing equipment, the video recordings from the phonated condition (with the original sound track on audio channel A) were copied one by one from the master tape onto the experimental tape, according to the randomized stimulus sequence for the video track. Each video segment started about 1 sec before, and ended about 1 sec after, the audible utterance. A view of the speaker's still face, of similar total duration, served as the neutral stimulus. About 3 sec of blank screen intervened between successive video segments. (2) The resulting audio track was digitized in portions, and the exact intervals between the onsets of the original spoken words were measured in displays of the digitized waveforms. (Most words began with stop consonants; for a few that began with nasals, the point of oral release following the nasal murmur was considered the onset.) (3) A computer output sequence was created containing the audio items to be substituted for the original utterances,

according to the experimental stimulus schedule for the audio track, with exactly the same onset-to-onset intervals as those measured on audio channel A. Audio trials for the neutral condition were timed to start about 1 sec after the onset of the still face video. (4) This auditory substitute sequence was output and recorded onto audio channel B, which was the one played back during the experiment.[4]

Subjects and Procedure. The subjects were 12 paid volunteers, all native speakers of American English and claiming to have normal hearing. They were tested singly in a quiet room. The subject sat in front of a colour monitor at a comfortable viewing distance and listened to the audio output over the video loudspeaker at a comfortable intensity. The task was described as one of speech detection in noise, and 24 practice trials using a higher signal-to-noise ratio (−10 dB in the phonated condition, 0 dB in the whispered condition) were provided without any accompanying video; these trials contained words not used as audio stimuli later on. Subjects were informed that a spoken word (either phonated or whispered, depending on the condition) was present in the noise on some of the trials. They were told to watch the video screen, but it was emphasized that what they saw had nothing to do with whether or not a word was hidden in the noise. The subjects wrote down their response (*S* for speech or *N* for noise only) on an answer form in the interval between trials. The whole experimental session (4 × 144 trials) lasted about 60 min.

Analysis. The data were analysed in terms of the detectability and bias indices proposed by Luce (1963), which we call *d* and *b* here for simplicity, and which are comparable to the *d'* and Beta indices of Signal Detection Theory. They are defined as

[4]As we did not have equipment available to trigger the output sequence precisely and thus to ensure exact audio-visual synchrony, we started and restarted the output sequence manually until it seemed in synchrony with the video channel. Subsequently, we measured the onset asynchrony between audio channels A and B on matching trials, using two-channel digitization and digital waveform displays. If any asynchrony exceeded ±100 msec, we re-recorded the output sequence. Asynchronies within this range are difficult to detect (Dixon & Spitz, 1980; McGrath & Summerfield, 1985) and seem to have only a negligible effect on audiovisual speech perception (Tillmann, Pompino-Marschall, & Porzig, 1984; McGrath & Summerfield, 1985). Although we believed at the time to have satisfied this criterion, postexperimental checks revealed some inaccuracies in the test sequence specifications that led to onset asynchronies in excess of 100 msec for some stimulus combinations. These asynchronies were always such that the sound lagged behind the visual stimulus, which is less detectable than the opposite (Dixon & Spitz, 1980), and they occurred only in the phonated condition. Although this aspect should have been under better control, we have no indication that audio-visual asynchrony had any effect whatsoever on our results; in particular, as will be seen, the phonated and whispered conditions yielded very similar bias shifts.

$$d = \frac{\ln[p(H)p(1 - FA)/p(1 - H)p(FA)]}{2}$$

and

$$b = \frac{\ln[p(H)p(FA)/p(1 - H)p(1 - FA)]}{2}$$

where p(H) and p(FA) are the proportions of hits and false alarms, respectively.[5] The d index (normally) assumes positive values similar to d', and a positive b index indicates a bias to respond "*S*" (i.e. "*speech present*"). The indices we report below were computed for each subject and then averaged; however, we also computed indices for each item and did statistical analyses both ways. Separate analyses of variance were conducted on the phonated and whispered conditions, with S/N ratio and visual condition as within-subject factors; the F ratios for the subject and item analyses will be reported as $F1$ and $F2$, respectively.

Results

Detectability. In the phonated condition, the average d indices for the two S/N ratios were 2.14 (−12 dB) and 1.88 (8b14 dB). This difference, which pitted the positive effect of practice against the negative effects of reducing the S/N ratio, was significant across items [$F1(1, 11) = 3.83$, $p < 0.08$; $F2(1, 23) = 18.63$, $p < 0.0004$] but is of little interest. The important result was that detection performance was unaffected by visual condition [$F1(2, 22) = 0.38$, $p > 0.5$; $F2(2, 46) = 2.19$, $p > 0.1$]; the average ratios in the three conditions were 1.95 (match), 2.02 (mismatch), and 2.07 (neutral). Thus, seeing a matching articulation did not aid speech detection. If anything, a match reduced sensitivity: in the item analysis, but not in the subject analysis, there was a significant S/N ratio × visual condition interaction [$F1(2, 22 = 1.00$, $p > 0.3$; $F2(2,46) = 5.48$, $p < 0.008$]: at the higher S/N ratio, performance was best in the neutral condition and worst in the matching condition; this difference disappeared at the lower S/N ratio.

The average d indices were lower in the whispered than in the phonated condition, despite the much higher S/N ratios: 1.44 and 1.12, respectively, at the −4 dB and −6 dB ratios. The decline in sensitivity as a function of S/N ratio was significant [$F1(1, 11) = 16.10$, $p < 0.003$; $F2(1, 23) = 23.51$, $p < 0.0002$]. The performance levels were ideal for observing effects of

[5]Values of ½n and 1-½n were substituted for proportions of 0 and 1, respectively. Due to the different frequencies of these substitutions and the non-linear nature of the indices, the average d and b indices were not identical when computed across subjects and across items.

visual condition. Still, there was no trace of a visual condition main effect [$F1$, $F2 < 1$]; the average values in the three conditions were 1.34 (match), 1.20 (mismatch, and 1.31 (neutral). Thus, even when low-frequency cues were eliminated, an audiovisual match did not facilitate detection performance. The S/N ratio × visual condition interaction was likewise non-significant [$F1$, $F2 < 1$].

Bias. We turn now to the results of primary interest. The *b* indices for both production mode conditions, averaged across the two S/N ratios, are shown in Figure 1 as a function of visual condition.

In the phonated condition, there was a strong bias to respond "*S*" in the matching condition, a lesser bias in the non-matching condition, and hardly any bias in the neutral condition. This pattern of results matches that obtained with orthographic stimuli (Frost et al., 1988; Frost, 1991). The main effect of visual condition was highly significant [$F1(2, 22) = 11.32$, $p = 0.005$; $F2(2, 46) = 29.79$, $p < 0.0001$]. Planned comparisons revealed reliable differences between the matching and non-matching conditions [$F1(1, 11) = 15.24$, $p < 0.003$; $F2(1, 23) = 5.94$, $p < 0.03$], and between the non-matching and neutral conditions in the item analysis [$F2(1, 23) =$

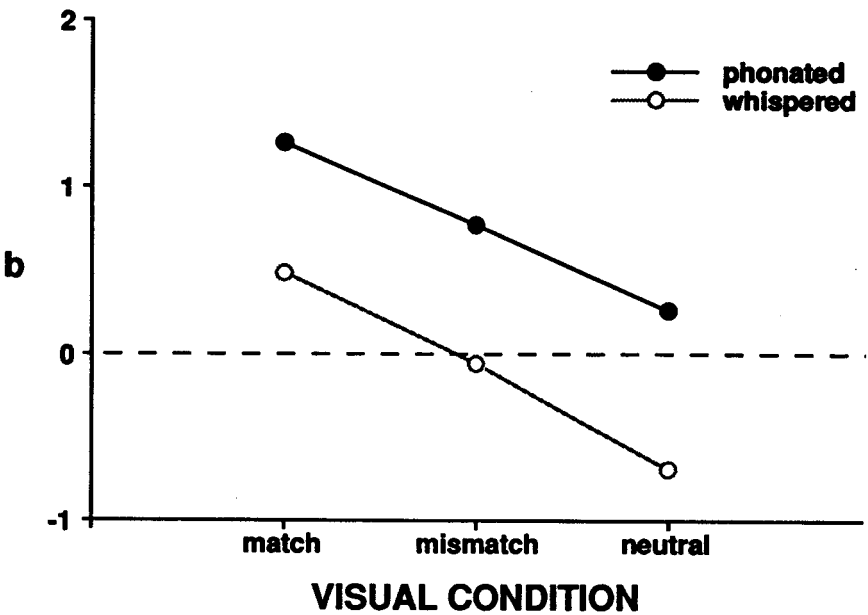

FIG. 1. Bias indices in the phonated and whispered conditions of Experiment 1 (word materials) as a function of visual condition.

26.84, $p < 0.0001$] but not in the subject analysis [$F1(1, 11) = 3.70$, $p < 0.09$]. There were no significant effects involving S/N ratio.

In the whispered condition, the absolute b indices were much lower, but a very similar main effect of visual condition emerged [$F1(2, 22) = 20.61$, $p < 0.0001$; $F2(2, 46) = 55.14$, $p < 0.0001$]. There was a small bias to say "*S*" in the matching condition, no bias in the non-matching condition, and a bias to say "*N*" in the neutral condition. Planned comparisons showed reliable differences between the matching and non-matching conditions [$F1(1, 11) = 14.42$, $p < 0.004$; $F2(1, 23) = 20.08$, $p < 0.0003$], and between the non-matching and neutral conditions [$F1(1, 11) = 12.81$, $p < 0.005$; $F2(1, 23) = 38.87$, $p < 0.0001$]. There were no significant effects involving S/N ratio.

In summary, the results of Experiment 1 replicate almost exactly the findings of Frost et al. (1988) with orthographic word stimuli. Clearly, subjects were able to perceive a correspondence between speech gestures presented visually and amplitude envelopes presented auditorily. Like matching printed information, matching articulatory information, too, seems to create an illusion of hearing speech in amplitude-modulated noise. The bias shifts in the phonated and whispered conditions were equivalent. The difference between these conditions in absolute bias values must have a different origin (see General Discussion); whatever its cause, it is orthogonal to the relative bias shift with which we are concerned.

In order to determine whether the detection of correspondence between the speaker's articulatory gestures and the noise amplitude envelopes is lexically mediated, we examined in Experiment 2 whether non-word materials would produce the same effect.

EXPERIMENT 2

As similar bias shifts were obtained in Experiment 1 regardless of production mode, only a phonated condition was employed in Experiment 2. Otherwise, except for the difference in materials, the experiment was an exact replication of Experiment 1. If there is a direct (i.e. prelexical) link between visible articulatory movements and the auditory amplitude envelope, then the results of Experiment 2 should replicate those of Experiment 1. If, on the other hand, this connection can only be established via the lexicon, then there should be no effect of audio-visual match on response bias. In particular, there should be no difference between the matching and non-matching conditions; as it is conceivable that the mere presence versus absence of articulatory movements has an independent effect on response bias (see discussion below), the comparison with the neutral condition is less crucial.

Methods

The stimuli were 48 disyllabic non-words stressed on the first syllable, produced by the same female speaker and videotaped. In Frost et al. (1988), orthographic non-words had been generated from words by changing one or two letters. This would not do for speech reading because of the phonological ambiguity of visemes. To ensure that our stimuli were not speech-read as English words, we used phonotactically atypical but easily pronounceable utterances containing the point vowels /a, i, u/ and visually distinctive consonants. (Examples: "vumuv", "kichaf", "fafiz", etc.). Of the non-words, 24 were used as auditory stimuli, the other 24 as non-matching visual stimuli. The generation of stimulus tapes, the test sequences, and the procedure were identical with those in Experiment 1. Because detectability scores in the phonated condition of Experiment 1 had been somewhat high, the S/N ratios were set slightly lower in Experiment 2: at −13 and −16 dB. The subjects were informed that the utterances were meaningless. The subjects were 12 volunteers from the same general population. Two of them had participated in Experiment 1.

Results

Detectability. The average *d* indices for the two S/N ratios were 1.62 and 1.11, respectively—significantly lower than the corresponding indices for phonated words in Experiment 1 [$F1(1, 22) = 5.84$, $p < 0.03$; $F2(1, 46) = 10.96$, $p < 0.002$, in a combined ANOVA], in part due to the somewhat lower S/N ratios used.[6] The main effect of S/N ratio was significant [$F1(1, 11) = 49.84$, $p < 0.0001$; $F2(1, 23) = 24.88$, $p < 0.0001$]. Surprisingly, there was also a significant main effect of visual condition here [$F1(2, 22) = 10.00$, $p < 0.0009$; $F2(2,46) = 6.50$, $p < 0.004$]. This effect was due to a lower *d* index in the non-matching condition (1.17) than in either the matching condition (1.51) or the neutral condition (1.42). In a combined ANOVA on the data of Experiment 1 (phonated condition) and of Experiment 2, with the added factor of lexical status (word/non-word), a significant interaction of visual condition and lexical status was obtained [$F1(2, 22) = 3.46$, $p < 0.05$; $F2(2, 46) = 4.14$, $p < 0.02$]. This suggests some inhibition or distraction caused by an audiovisual mismatch for

[6]Performance for non-words was somewhat lower than expected on these grounds alone. Of course, this could have reflected a random difference between subject samples. However, Frost et al. (1988), too, found lower detection performance for non-words than for words in different experiments, even though the words and non-words were equally detectable when presented randomly within the same experiment (Repp & Frost, 1988; Frost et al., 1988: Exp. 3). It is as if subjects listened less carefully when they are presented with nonsense.

non-words, but no facilitation due to a match. The S/N Ratio × visual condition interaction was non-significant.

Bias. The bias results are shown in Figure 2, averaged over the two S/N ratios. There was a significant effect of visual condition [$F1(2, 22) = 11.86$, $p < 0.0004$; $F2(2, 46) = 13.68$, $p < 0.0001$], but, as can be seen in the figure, it was entirely due to the matching and non-matching conditions versus the neutral condition. There was absolutely no difference between the former two conditions, both of which exhibited a small positive bias. The effect of visual condition did not interact with S/N ratio. There was a marginally significant main effect of S/N ratio [$F1(1, 11) = 4.62$, $p < 0.06$; $F2(1, 23) = 4.44$, $p < 0.05$], due to an absolute decrease in the bias to say "*S*" when the S/N ratio was lowered. In order to compare directly the differences between the matching and the non-matching conditions obtained for words and for non-words, we combined in one ANOVA the data of Experiments 1 (phonated) and 2 for these two visual conditions. The interaction of visual condition and lexical status was significant across subjects [$F1(1, 11) = 11.11$, $p < 0.004$], and nearly so across items [$F2(1, 23) = 3.53$, $p < 0.07$].

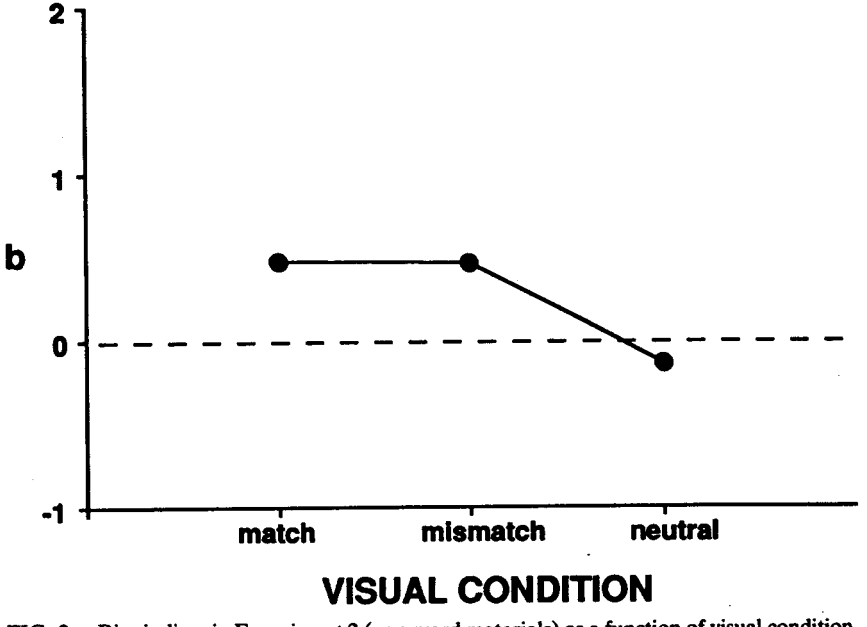

FIG. 2. Bias indices in Experiment 2 (non-word materials) as a function of visual condition.

GENERAL DISCUSSION

In the present study we examined the effect of a visual presentation of a speaker's face uttering words and non-words on the detection of these words and non-words in amplitude-modulated noise. Our experiments yielded three main results:

1. There was no facilitative effect of audio-visual match on speech detectability.
2. However, subjects recognized a correspondence between speech gestures and amplitude envelopes when the stimuli were words. Such an audio-visual match created an increased tendency to report the presence of speech in the detection task.
3. This bias shift was absent when the stimuli were non-words.

We will discuss these three results in turn.

Speech Detection in Noise and Speechreading

The absence of a systematic effect of visual conditions on speech detectability is not too surprising, in view of the fact that the task of detecting speech in noise requires only relatively low-level auditory processing. When the masking noise is coextensive with the speech and has the same amplitude envelope, as in our study, this means that the listeners must detect local spectral peaks that rise above the flat spectral level represented by the masking noise. When the speech is phonated, such peaks are most likely to occur in the lowest harmonics of voiced portions, and listeners therefore hear snippets of a human voice somewhere in the noise. As speechreading does not provide information about the presence or absence of voicing, it cannot guide the listener to any portions of the signal that are especially likely to yield spectral evidence of voicing.

When the speech is whispered, listeners probably detect spectral prominences in the region of the second formant, or at higher frequencies if the word contains fricatives with strong noise components, such as /s/. This task is difficult because the speech itself has a noise source, and the S/N ratio must be raised considerably to chieve above-chance accuracy. Speechreading can provide some limited information about the occurrence of fricatives, but the most visible consonant articulations (/bpm/, /vf/, /θð/ have weak acoustic correlates, and fricatives such as /s/ were rare in our stimuli. Thus there is not much to be gained from speechreading here either, and auditory detection strategies therefore seem to be uninfluenced by visual input.

There were two instances in which visual input did affect detectability scores, but the influence was negative rather than positive. In the phonated

condition of Experiment 1, there was a tendency for detection to be best in the neutral condition, but only at the higher S/N ratio. More strikingly, in Experiment 2 detection scores were depressed in the non-matching condition. Seeing articulatory movements may have had a slight distracting effect on listeners, especially when there was an obvious mismatch with the auditory input. Mismatches may have been more obvious in the non-word experiment, due to the different construction of the materials.

The Bias Shift for Words

The result of primary interest is the relative change in bias as a consequence of audiovisual match. Our findings suggest that the visual presentation of speech gestures matching the auditory amplitude envelope causes an auditory illusion of hearing speech that is similar to the illusion obtained by Frost et al. (1988) with printed stimuli. This may not seem surprising: if subjects can detect the correspondence between the auditory amplitude envelope and print, whose relationship to each other is merely conventional, then they certainly should also detect the correspondence between the envelope and articulatory movements, which are intrinsically linked. In particular, the visible time course of jaw opening is a direct optic correlate of the gross amplitude envelope. It is not necessary to invoke lexical access to explain the results for words. Lexical access probably did occur, however, due to the joint constraints effected by the auditory amplitude envelope and the visual articulatory information, and it probably happened more often in the matching than in the non-matching condition.

Two aspects of subjects' sensitivity to audio-visual matches deserve comment: (1) An effect of match was obtained even though the auditory and visual inputs were not in perfect synchrony; this suggests, in accordance with earlier findings (see Footnote 4), that temporal offsets smaller (and occasionally larger) than 100 msec do not interfere substantially with the detection of audio-visual correspondence, especially if the sound lags behind. (2) The bias shift was obtained for both phonated and whispered speech, even though the amplitude envelope of a given word was different in the two production modes. As the same video was used in both conditions and relative bias shifts of the same magnitude were obtained, this means that the audio-visual match was equally good for both kinds of amplitude envelopes. The amplitude envelopes thus must have retained crucial phonetic properties across the change in phonation type (cf. Tarter, 1989, on phonetic information in whispered speech). The extent to which the speech amplitude envelope conveys invariant phonetic features is a worthwhile topic for investigation, and it has received only very limited attention so far (e.g. Mack & Blumstein, 1983; Nittrouer & Studdert-Kennedy, 1985).

There was one difference, however, between the phonated and whispered conditions: the absolute bias indices were considerably lower in the whispered condition. As the masking noises were rather similar in the two conditions, the difference in bias must reflect differences in subjects' expectations of hearing speech. The greater difficulty of the whispered condition and the atypicality of whispered speech may have been sufficient reasons for subjects' relative conservatism, as reflected in the absolute bias indices.

So far, we have focused on the difference between the matching and non-matching conditions for words, which constitutes the predicted bias shift. However, there was also a reliable difference between the non-matching and neutral conditions, with the bias to say "*S*" being relatively greater in the non-matching condition. This difference was also obtained in the earlier study with print (Frost et al., 1988). There are two possible interpretations: (1) The effect may represent a different kind of response bias, caused by *any* structured visual input (print or articulation), regardless of match. According to this view, there are really two bias shifts: a less interesting one (postperceptual response bias), which accounts for the difference between neutral and non-neutral conditions, and a more interesting one (perceptual in origin), which accounts for the difference between the matching and non-matching conditions. (2) Alternatively, the difference between the neutral and non-matching conditions may represent an effect of partial match. After all, the non-matching stimuli had the same general prosodic pattern as the matching stimuli (i.e. two syllables, with stress on the first). This may have been sufficient to obtain a small bias shift. According to this view, there is a single bias shift effect that is present to varying degrees in the matching and non-matching conditions, and the "non-matching" condition really should have been called "partially matching" in this case.

The present data for word stimuli cannot decide between these two alternatives. However, a previous experiment that bears on the issue is Experiment 3 of Frost et al. (1988), which used orthographic visual stimuli. In that experiment, white noise without amplitude modulation was used as a masker. Thus, there was no auditory basis for either whole or partial matches. Yet, a difference in bias was obtained between the neutral condition and the other two conditions. This suggests that the first explanation given above is correct, at least for print.

The Absence of a Bias Shift for Non-words

This suggestion seems to be confirmed by the present results: the difference in the bias between the matching and non-matching conditions, obtained for word stimuli in Experiment 1, was absent for non-word stimuli in Experiment 2. There was, however, a reliable difference

between the neutral condition and the other two visual conditions even for non-words, and this difference was similar in magnitude to that between the neutral and non-matching conditions for words. If the relative bias increase in the non-matching condition represented an effect of partial match, then it would be difficult to explain why an additional effect of complete match was obtained for words only. Therefore, the difference between the neutral and non-matching conditions may well represent an "uninteresting" response bias, due to the occurrence of any verbal event in the visual modality.

However, the partial match explanation can still be upheld by noting that the partial match reflects only general prosodic characteristics (number of syllables, stress pattern), whereas the complete match reflects the added effect of matching segmental envelope characteristics as well as prosodic detail such as the exact timing pattern. To account for the effect of lexical status, one is then led to the interesting (but highly speculative) conclusion that the detection of segmental (and exact prosodic) cross-modal matches requires lexical access, whereas the detection of gross prosodic matches can occur without the involvement of the lexicon.

A similar conclusion was reached independently, and on the basis of quite different kinds of evidence, by Cutler (1986). In a cross-modal priming task, auditorily presented words drawn from semantically distinct pairs that differed only in stress pattern but not in segmental structure (quite rare in English; e.g. "FORbear"—"forBEAR") had equal priming effects on lexical decision for visual targets that were semantically related to one or the other member of the pair. In other words, the auditory stress pattern did not constrain lexical access and only postlexically disambiguated the two semantic alternatives. Our results are complementary to those of Cutler in that they suggest that global prosodic information, including stress pattern, is processed independently of lexical access. This result makes sense when we consider the fact that prosodic parameters are not specific to speech but also play an important role in music, in animal communication, and even in environmental sounds. Lexical access, perhaps necessarily, is governed by speech-specific (segmental and detailed prosodic) properties of the acoustic signal; global prosodic properties, on the other hand, feed into the nonverbal systems of auditory event perception and emotion. They may also be processed in different parts of the brain.

The above interpretation remains speculative because we do not know what would happen on trials on which there is a striking *prosodic mismatch* between the auditory and visual inputs. An experiment including such trials remains to be conducted. Our results show very clearly, however, that an audio-visual match of segmental (and detailed prosodic) characteristics leads to a bias shift only for words, not for non-words. This result

replicates earlier findings obtained with print (Frost et al., 1988; Frost, 1991) and, in fact, is more dramatic: Whereas a small difference between matching and non-matching conditions was consistently obtained with printed non-words, there was no difference at all with speechread non-words, perhaps because the latter were less similar to English words than were the printed non-word stimuli. In the case of print, the results suggested that lexical access through the visual modality results in a detailed phonetic representation that shares amplitude envelope features with a matching signal-correlated noise. The alternative process of letter-to-sound translation by rule or analogy, which—according to traditional dual-route models—must be employed for non-words, is either too slow to enable subjects to relate its product to the auditory stimulus, or, more probably, does not result in a detailed, complete, or coherent phonetic representation. The latter interpretation is favoured by Frost's (1991) recent results, which show that manipulations known to affect speed of word recognition (viz. word frequency and Hebrew vowel diacritics) have no effect on the magnitude of the bias shift for words; by implication, the absence of a bias shift for non-words is probably not due to a slower processing speed. Can the same arguments be made in the case of speechreading?

In the introduction, we pointed out three important differences between print and visual articulatory information. Two specific aspects of speechreading—the temporal nature of the information and its non-arbitrary relation to the sounds of speech (including the amplitude envelope)—led to the expectation that an effect of audio-visual match might be obtained regardless of lexical status. This was clearly not the case; thus, speechread information is not directly translated into a phonetic representation. The reason for this lies probably in the third aspect: the visual information is not specific enough. Inner speech consists of the sounds of words, not just of their amplitude envelopes, which are features of the complete sound patterns. Speechread information rarely specifies a unique word, however, and hence it does not (or only rarely) lead to lexical access in the case of isolated words, nor does it enable a viewer to construct a detailed sound pattern by direct translation, bypassing the lexicon. Normally, the incomplete information needs to be supplemented by additional information that constrains the possible lexical choices. The auditorily presented amplitude envelope probably functioned as such a source of supplementary information (see Footnote 2). In addition, its spectral masking power may have enabled the auditory illusion of hearing speech, as in the phonemic restoration effect (cf. Warren, 1984).

This role of the auditory amplitude envelope in conjunction with speechreading is somewhat different from the role Frost et al. (1988) attributed to it in their studies with print, where they saw it as *probing* into

the process of lexical access from (unambiguous) print. In the case of speechreading, the noise envelope is not so much a probe as an active ingredient in the processes leading to lexical access. (When print stimuli are made ambiguous, as in a recent, still unpublished study by Frost, the same is probably true.) The best way, then, to characterize what happened in our present experiments is that amplitude envelope information and speechread information often converged onto a lexical entry in the case of words, but failed to do so in the case of non-words. Whether the bias shift for words was a direct consequence of this lexical convergence, or whether a separate postlexical process detected the match between the phonetic representation stored in the lexicon and the noise envelope, is a moot and probably unresolvable question. It may be concluded, however, that it is the lexically mediated activation of an internal phonetic representation that accounts for the illusion of hearing speech in the noise, and hence for the bias shift.

REFERENCES

Behne, D.M. (1990). The position of the amplitude peak as an acoustic correlate of stress in English and French. *Journal of the Acoustical Society of America, 87*, S65–66 (abstract).

Blamey, P.J., Martin, L.F.A., & Clark, G.M. (1985). A comparison of three speech coding strategies using an acoustic model of a cochlear implant. *Journal of the Acoustical Society of America, 77*, 209–217.

Breeuwer, M., & Plomp, R. (1984). Speechreading supplemented with frequency-selective sound-pressure information. *Journal of the Acoustical Society of America, 76*, 686–691.

Breeuwer, M., & Plomp, R. (1986). Speechreading supplemented with auditorily presented speech parameters. *Journal of the Acoustical Society of America, 79*, 481–499.

Cutler, A. (1986). *Forbear* is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech, 29*, 201–220.

Dijkstra, T., Schreuder, R., & Frauenfelder, U.H. (1989). Grapheme context effects on phonemic processing. *Language and Speech, 32*, 89–108.

Dixon, N.F., & Spitz, L. (1980). The detection of audiovisual desynchrony. *Perception, 9*, 719–721.

Erber, N.P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12*, 423–425.

Frost, R. (1991). Phonetic recoding of print and its effect on the detection of concurrent speech in amplitude-modulated noise. *Cognition, 39*, 195–214.

Frost, R., Repp, B.H., & Katz, L. (1988). Can speech perception be influenced by simultaneous presentation of print? *Journal of Memory and Language, 27*, 741–755.

Glushko, R.J. (1979). The organization of activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 674–691.

Gordon, P.C. (1988). Induction of rate-dependent processing by coarse-grained aspects of speech. *Perception & Psychophysics, 43*, 137–146.

Grant, K.W., Ardell, L.H., Kuhl, P.K., & Sparks, D.W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *Journal of the Acoustical Society of America, 77*, 671–677.

Kuhl, P.K., & Meltzoff, A.N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138–1141.

Layer, J.K., Pastore, R.E., & Rettberg, E. (1990). The influence of orthographic information on the identification of an auditory speech event. *Journal of the Acoustical Society of America, 97*, Suppl. 1, S125 (abstract).

Luce, R.D. (1963). Detection and recognition. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. New York: Wiley.

Mack, M., & Blumstein, S.E. (1983). Further evidence of acoustic invariance in speech production: The stop-glide contrast. *Journal of the Acoustical Society of America, 73*, 1739–1750.

Massaro, D.W., & Cohen, M.M. (1990). Perception of synthesized audible and visible speech. *Psychological Science, 1*, 55–63.

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

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America, 77*, 676–685.

Nittrouer, S., & Studdert-Kennedy, M. (1985). The stop-glide distinction: Acoustic analysis and perceptual effect of variation in syllable amplitude envelope for initial /b/ and /w/. *Journal of the Acoustical Society of America, 80*, 1026–1029.

O'Neill, J.J. (1954). Contributions of the visual components of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders, 19*, 429–439.

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research, 28*, 381–393.

Patterson, K., & Coltheart, V. (1987). Phonological processes in reading: A tutorial review. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 421–447). Hove, UK: Lawrence Erlbaum Associates Ltd.

Remez, R.E., & Rubin, P.E. (1990). On the perception of speech from time-varying acoustic information: Contributions of amplitude variation. *Perception & Psychophysics, 48*, 313–325.

Repp, B.H., & Frost, R. (1988). Detectability of words and nonwords in two kinds of noise. *Journal of the Acoustical Society of America, 84*, 1929–1932.

Schroeder, M.R. (1968). Reference signal for signal quality studies. *Journal of the Acoustical Society of America, 43*, 1735–1736.

Smith, M.R., Cutler, A., Butterfield, S., & Nimmo-Smith, I. (1989). The perception of rhythm and word boundaries in noise-masked speech. *Journal of Speech and Hearing Research, 32*, 912–920.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Swinney, D.A., Onifer, W., Prather, P., & Hirshkowitz, M. (1979). Semantic facilitation across sensory modalities in the processing of individual words and sentences. *Memory & Cognition, 7*, 159–165.

Tartter, V.C. (1989). What's in a whisper? *Journal of the Acoustical Society of America, 86*, 1678–1683.

Tillmann, H.G., Pompino-Marschall, B., & Porzig, U. (1984). Zum Einfluss visuell dargebotener Sprechbewegungen auf die Wahrnehmung der akustisch kodierten Artikulation. *Forschungsberichte des Instituts für Phonetik und sprachliche Kommunikation*(University of Munich, FRG), *19*, 318–336.

Van Orden, G.C., Pennington, B.F., & Stone, G.O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review, 97*, 488–522.

Van Tasell, D.J., Soli, S.D., Kirby, V.M., Widin, G.P. (1987). Speech waveform envelope cues for consonant recognition. *Journal of the Acoustical Society of America*, *82*, 1152–1161.

Warren, R.M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin*, *96*, 371–383.

Revised manuscript received 10 September 1991