LANGUAGE AND SPEECH, 1992, 35(4), 391-417

AUDITORY AND VISUAL CUEING OF THE [±ROUNDED] FEATURE OF VOWELS*

LEIGH LISKER
University of Pennsylvania and Haskins Laboratories
and
MARIO ROSSI
Université de Provence

That lipreading plays a role in phoneme recognition, even when the acoustic signal alone is phonologically unambiguous, has been concluded from experiments in the perception of discrepant combinations of acoustic and visual speech signals. Little is known about the effect of visual information on explicitly phonetic judgments, the kind of judgments made by trained observers that are the basis for describing the phonological pattern of a language. In this study some isolated vowels, most of them similar to vowels in standard French, were produced in ten random orders by an experienced phonetician. The acoustic signals and frontal views of the lower half of the speaker's face were recorded on video tape. By computer editing, audiovisual stimuli were prepared in which pairs of vowels supposed to differ primarily in rounding were variously combined. Twenty French-speaking speech researchers carried out three tasks: to decide on the rounding of each vowel by sound alone, by sight alone, and by sound when accompanied by matching or discrepant images of the talker. Their summed responses indicate that, despite the instruction to base decisions on the auditory signal, visual evidence of speech activity significantly "perturbed" subjects' rounding judgments. However, the lipreading effect varied greatly across both subjects and vowels. Most subjects judged most vowels strictly on the basis of the auditory information, while for others lipreading exerted paramount influence. Only a small minority responded so as to indicate any integration of discrepant rounding information registered by ear and eye.

Key words: rounding, audiovisual speech perception, lipreading, French

---

* The first author wishes here to thank the University of Pennsylvania and the American Philosophical Society in Philadelphia, and the Haskins Laboratories of New Haven, CT (through NICHD Grant HD–01994) for financial support that made possible his stay at the University of Provence, Aix-en-Provence. The work reported here could not have been accomplished without the wholehearted cooperation of the staff and graduate students of the Institute of Phonetics of the University of Provence. We want also to thank Robert Espesser and Bernard Teston, members of the engineering staff of the Institut de Phonétique whose technical expertise was essential to the preparation of the test stimuli. In addition, we should like to express our gratitude to Dominique Abry, who served as a "neutral" phonetician in making auditory judgments of our stimuli in terms of the French vowel categories, as well as to Aline Tabourin, who did the same as a native speaker of French without any background in phonetics. Patient instruction in the vagaries of analysis of variance by Leonard Katz, of Haskins Laboratories and the University of Connecticut is also acknowledged with thanks. Finally we are grateful for very helpful review comments by Alan Montgomery, Bruno Repp, and Quentin Summerfield.

Communications to the authors may be addressed to Prof. Leigh Lisker, Department of Linguistics, University of Pennsylvania, Philadelphia PA 19104, USA; or to Prof. Mario Rossi, Institut de Phonétique, L.A. CNRS 261, Université de Provence, 29 Avenue Robert Schuman, 13621 Aix-en-Provence, France.

INTRODUCTION

Under modern conditions of communication, speech perception much of the time is accomplished largely, or even exclusively, by ear. Perception by eye is in general much less "accurate" than is the linguistic interpretation of the acoustic effects of articulatory activity of the vocal tract. This is in part explained by the fact that even under optimum viewing conditions a significant amount of what goes on in the speaker's vocal tract is invisible, so that even for the experienced lipreader the number of visemes is smaller than the number of phonemes perceived under reasonably normal listening conditions. However, we may plausibly suppose that during primary language acquisition most infants gain their implicit knowledge of the connection between a speech sound and a particular position or movement of certain body parts from the audiovisual experience of speech (Kuhl and Meltzoff, 1984). Some of a phonetician's wider and more explicit knowledge of relations between sound and articulation is similarly based. In both cases there is no question but that both the acoustic and the visual information originated in one and the same phonetic event. Researchers looking into the contribution of lipreading to speech intelligibility generally also suppose that speech is most reliably perceived when the talker can be both heard and seen. When only visual information about a speech event is available, message intelligibility is invariably reduced; when only auditory information is accessible, it *may* be reduced because of channel noise of some kind. Recent evidence suggests that lipreading, despite its limitations, is not simply a perceptual stratagem that is adopted when the acoustic channel is unsatisfactory. It also plays a role when another kind of noise factor is introduced – the "ecological invalidity" of a sound-image combination reflecting two different speech events. In this unnatural situation an observer asked to attend to just one of the signals, the auditory, seems unable to screen out ostensibly discrepant information from the other.

Evidence of a role for lipreading in speech perception under the condition of auditory adequacy comes largely from lexical or pseudolexical identifications elicited from observers selected as typical speakers of a particular language, i.e., observers without special phonetic awareness, but capable of making consistent phonological judgments. The data have come, for the most part, from speakers of English[1] tested in a situation where the visual information provided was not derived from the true source of a synchronously presented acoustic signal. When the observer was asked to report what was **heard** (so that the purely auditory judgment was by definition correct), then the most striking evidence of the significance of lipreading was the spoiler effect of discrepant visual information on the "auditory" perception of stop consonant place of articulation. In some part that effect can be accounted for by supposing that observers categorize audiovisual stimuli according to a simple binary rule based on the most readily

---

[1] That the role of lipreading may be language or culture dependent is suggested by data indicating that the perception of acoustically unambiguous consonants by Japanese speakers is not perturbed by discrepant visual "information" (Sekiyama and Tohkura, 1991). Possibly relevant is the observation that "many Japanese feel it is rude to look a person directly in the face" (Martin, 1959, p. 4).

accessible visual information: Whatever the accompanying acoustic signal might be, if the lips are seen to close, then /p/, /b/, or /m/ will be heard; if they are seen not to close, then none of these consonants characterized by lip closure will be reported (McGurk and Macdonald, 1976; Macdonald and McGurk, 1978). In short, what the phonemes look like dictates what they sound like. Thus, while generally there is little question of the primacy of the acoustic signal in speech communication (Easton and Basala, 1982), it might be said that in this situation the eye not only influences perception, but takes precedence over the ear, at least so far as perceiving a stop as [+labial] or [–labial].

There is also convincing evidence that lipreading affects the perception of vowel sounds, but it is not as easy to account for much of its effect on vowel labeling behavior by appealing to any simple two-valued measure of lip configuration, for the size and shape of lip aperture varies with respect to a number of dimensions (Fromkin, 1964; Lindau, 1978; Linker, 1982; Montgomery and Jackson, 1983; Nearey, 1980), and conveys information that influences perceived location along more than one dimension of the continuous "vowel space" (Summerfield and McGrath, 1984). The effect of a visual vowel different from the auditory one is to shift to some extent the perceived quality of the latter toward another contiguous to it within the vowel space (Summerfield and McGrath, 1984). For consonants, a tentative conclusion that the lipreading effect is in some measure referrable to a [±labial closure] dimension is derived from an analysis of the phoneme labeling behavior of phonetically untrained observers. In the case of vowels, too, lip positions are traditionally described by linguists in a way analogous to a binary division of consonants into [+labial] and [–labial] types: Vowels are categorized as being either unrounded ("[–rnd]") or rounded ("[+rnd]"). This kind of labeling, however, is not one of phoneme identification, but of phonetic classification, and is therefore properly assigned only to observers with some phonetic awareness. Moreover, it can be argued that it is properly done **only** on the basis of lipreading, no matter how certain the observer without visual information might be in deciding on the rounding status of a vowel. In other words, decisions about this feature should "rationally" be made primarily on the basis of any available visual evidence, even when the task imposed on an observer requires that responses be based on auditory evaluation. Rounding judgments from suitable subjects can provide answers to several questions: (1) How closely do subjects agree in their auditory, visual and audiovisual rounding judgments? (2) Given combinations of audio and video signals with different classifications by ear and by eye, and alerted to the possibility of such combinations, (a) can they disregard discrepant visual evidence in reporting what they heard? (b) are their responses guided primarily by the nature of what was seen? or (c) do their responses reflect some sort of compromise resolution of the information presented to the two senses?

The vowel space: Articulatory

The "phonetic quality" of a vowel is traditionally specified by locating it within a three-dimensional space. Two dimensions are coordinates of putative tongue position (height and backness), while the third characterizes lip posture (rounding).[2] While most

[2] For a brief overview of the history of vowel classification see Catford (1981).

graphical representations of the vowel space allow us to suppose that these dimensions are mutually orthogonal and of equal rank, certain asymmetries of the space are suggested by the ways in which it is exploited in natural languages. Thus phoneme inventories tend to include fewer central than either front or back vowels, high and mid vowels are often more numerous than low vowels, and distinctive rounding is unevenly distributed over the height-backness plane, often being confined to the higher vowels (Lindau, 1978; Maddieson, 1984). Furthermore, in phonological discussion the dimensions of height and backness have sometimes been set off against rounding, either by being judged to be primary (e.g., Joos, 1948; Abercrombie, 1967; Nearey, 1980), or essentially acoustic/auditory rather than articulatory in nature (Ladefoged, 1982; Ladefoged and Maddieson, 1990). Moreover, while at least four "cardinal" positions along the height dimension and three degrees of backness are generally recognized (i.e., by the International Phonetic Association), for the rounding dimension, only two values are commonly taken into account, so that vowels are either unrounded or rounded. Of course, upon closer examination, rounding is not all that simple a property, differing continuously in degree and in kind, depending for the most part on position with respect to the other coordinates of the space.[3] Thus, for example, in high unrounded vowels the lips are likely to be spread, and in high rounded vowels they are also often protruded (that is, if "rounding" is not taken as synonymous with "protrusion"); in low vowels neither spreading nor protrusion of the lips is usually observed. However, in the analysis of most vowel inventories a simple binary classification is apparently considered sufficient for phonological description and "for most practical purposes" (Abercrombie, 1967, p. 57), whatever the cost might be in phonetic precision. It may be noted, moreover, that of the three dimensions of the vowel space, the one of rounding seems to provide the best visual information as to the identity of a vowel, even if it is alone not entirely sufficient (Tseva, 1989).

The vowel space: Acoustic

A direct relation is often said to hold between tongue position and the frequencies of the lowest resonances of the vocal tract (or the formants of its acoustic output), so that for a given talker the phonetic quality of a vowel, insofar as it is interpreted as tongue height and backness, can be inferred from a sound spectrogram. To a first approximation tongue position "maps" quite well into first and second formant frequencies, at least for a vowel inventory like that of English (Ladefoged, Harshman, Goldstein, and Rice, 1978). The height dimension is correlated with the frequency of the first formant, while the second formant frequency (or perhaps the frequency

[3] Linguists usually find it enough to classify a vowel of any specified tongue position as either rounded or unrounded, the most commonly cited exception being the two high front Swedish vowels distinguished by "inner" *vs.* "outer" rounding (following Henry Sweet, as per Henderson, 1971). Furthermore, while linguists classify vowels dichotomously into rounded and unrounded types, both types may be further subdivided (Heffner, 1969; Catford, 1977; Ladefoged and Maddieson, 1990; Zerling, 1992).

difference between the first and second formants) reflects the degree of tongue backing. One of several complications[4] that make this picture less than entirely adequate across languages, however, is the fact that lip rounding also leaves its trace in the resonance pattern of a vowel, its effect on the second formant being rather like that of tongue backing (Delattre, 1968; Fant, 1960; Lindblom and Sundberg, 1971).[5] Thus, the two cardinal vowels [i] and [y] might just as well be perceived to differ in degree of backness as in rounding, although they are described as being different only with respect to rounding. This is a problem of no great moment so far as the English vowels are concerned, since in English, lip rounding can be regarded as a maneuver without a phonologically contrastive function distinct from backing (Bloomfield, 1933; Ladefoged, 1982), serving only to "enhance" the acoustic effect of tongue backing.[6] Moreover, it may well be true that a speaker of English can produce satisfactory tokens of words such as *cool cook call,* containing the "rounded" vowels /u/ /ʊ/ /ɔ/, but with little or no visible lip rounding, although just what *is* done in such a case to produce the appropriate quality is uncertain, since possibly more than one "compensatory" maneuver is available (Joos, 1948; Riordan, 1977; Tuller and Fitch, 1980).[7] Nor can we be certain that even phonetically rather sophisticated listeners would not hear a visibly unrounded /u/, for example, as "rounded". This would be consistent with the belief of some field linguists/phoneticians that lip position is not reliably diagnosed from the acoustic signal alone (Abercrombie, 1985). If this view is well founded, and there are data consistent with it (Ladefoged, 1967; Lisker, 1989), then visual information, even when redundant for lexical and phoneme identification by listeners with normal hearing, is essential for the phonetic (i.e., articulatory) description of a vowel. In other words, there may be no auditory dimension of "rounding color", distinct from a "backing color", that reliably signals lip posture,[8] from which it would follow that a purely auditory judgment of

---

[4] Variations in vowel duration and nasalization have been linked with differences in perceived vowel height (Mermelstein, 1978; Wright, 1975).

[5] A striking regularity in the relation between F2, backing, and rounding is found in two-formant synthesized vowels, but matters are a bit muddled when F1 is high (Delattre, Liberman, Cooper, and Gerstman, 1952).

[6] This is not universally accepted. Thus for Roach (1983) English /u/ is distinctively [+rounded], with no phonological role for its [+back] property. Moreover, phonetic studies of coarticulation at least implicitly accept the rounding of English /u/ as distinctive (e.g., Boyce, Krakow, Bell-Berti, and Gelfer, 1990). It would seem, too, that English speakers perceive French /y/ more on the basis of its rounded than its fronted quality, classifying it with English /u/ rather than with English /i/ (Rochet, 1991).

[7] Conversely, if we can follow Stevens and House (1955) and Summerfield (1983), unrounded vowel qualities, such as [i], can be managed with rounded lips.

[8] Joos (1948, p. 42) seems to suggest this when he invokes the vowel quality of "muffled voice", defined acoustically as weakening of the higher harmonics of the glottal signal, as a property common to both rounded and backed vowels.

tongue position is also compromised. Nor can we ignore evidence of a further complication, which is that liprounding is regularly accompanied by lowering of the larynx (Perkell, 1969; Ewan and Krones, 1974; Riordan, 1977), a maneuver with acoustic consequences like those of backing and rounding.[9] Of course we may not draw from all this the patently false conclusion that a vowel's phonological identifiability must thereby be reduced: Thus, for example, a failure to classify the vowel in *cool* as high, back and rounded does not preclude a listener identifying it as /u/, i.e., a vowel identical *phonologically* with one certifiably high, back, and rounded. (Such a vowel would, no doubt, be termed "phonologically rounded" by some linguists.)

If lip position is not always reliably conveyed by the acoustic signal, this does not mean that auditorily based decisions about lip posture are invariably inaccurate, even for speakers of English, over the full range of tongue positions and shapes assumed in vowel production across languages. There are no doubt some vowel qualities, at least the familiar ones, which elicit articulatory interpretations, whether in the form of mimicry responses or explicitly phonetic judgments, that are both consistent and correct. Furthermore, the behavior of English speaking observers, even those with some phonetic sophistication, may not provide the best test of the null proposition that the acoustic reflexes of backness and rounding are not auditorily distinct; the behavior of speakers in whose native language backing and rounding are more nearly independent might be a better basis for its rejection. French, for example, is a language for which rounding is commonly recognized to be distinctive over the part of the vowel space occupied by the non-low front vowels, while elsewhere it is either nondistinctively absent, in the case of the lowest vowels, or nondistinctively present, in the non-low back ones. The question may be asked whether phonetically trained speakers of this language are able to separate the acoustic effects of backing and rounding, and if they are, whether this ability extends to the entire vowel space or is confined to that part of the space within which the [±rounded] feature is distinctive in French. In other words, we may ask whether native competence in French confers a general ability to distinguish auditorily between variations along the dimensions of tongue backing and lip rounding, or whether this ability is restricted to the part of the vowel space in which the second is distinctive in French, i.e., the front non-low vowels.[10]

---

[9] Mattingly (1990) proposes to lump the larynx lowering together with lip protrusion as an integral vocal-tract lengthening gesture.

[10] Speakers of languages with a rounding difference only in non-front vowels (e.g., Russian, Thai) would serve just as well in elucidating this question of generalizability. Turkish, however, with distinctive rounding of both front and back vowels, will not do, since acquisition of that language presumably entails learning how to manage this feature over the entire vowel space.

METHODS

Stimuli

Eighteen vowel symbols, i e ɛ a ɑ ɔ o u y ɸ œ ʌ ɯ ɨ ɛ̃ œ̃ ɑ̃ ɔ̃,[11] were presented serially in ten different random orders to one of the present authors (MR) a trained phonetician and native speaker of French. The vocal tract shapes and auditory qualities that are by international convention associated with most of these symbols (Principles of the International Phonetic Association, 1949) are virtually the same as those ascribed to vowels of standard French; the three exceptions, ʌ ɯ ɨ, represent vowels not found in that language. The 18 items were read aloud as isolated monophthongal vowel qualities, perhaps more or less approximating their commonly accepted IPA values, but quite possibly sometimes departing from them in the direction of the norms of standard French.[12] Whatever the deviations from established IPA definitions of the vowel symbols, it was expected that in reading them aloud the speaker would produce them with the prescribed rounding value: [i e ɛ a ɑ ʌ ɯ ɨ ɛ̃ ɑ̃] as unrounded, [ɔ o u y ɸ œ œ̃ ɔ̃] as rounded. The duration of each recorded vowel was about 350 msec, and successive vowels were separated by silent intervals of about three seconds. A video camera was set to record a frontal view of the lower half of the speaker's face, while his acoustic output was being recorded on one of the audio channels of the video tape.[13] Representative tokens of the fourteen oral vowels[14] of the recorded set are shown in Figure 1, plotted in standard fashion on the F1–F2 plane, their frequency values determined by LPC analysis.

---

[11] In citing the vowels of this study we follow the tradition of using square brackets ([ ]), but with no implication that the vowel tokens were subjected to any close comparison with the cardinal vowel qualities of IPA transcription practice.

[12] While none of the vowel sounds was intentionally French, two native French speakers and one American linguist with a good command of the language identified [ɯ ɨ] as non-French, and the others as the French vowels we should expect. One of the French speakers in addition volunteered the opinion that [ɯ ɨ] both resembled somewhat the Russian vowel "bl".

[13] This assumes that the French rounded vowels are produced with lip maneuvers readily seen in a frontal view of the lower face. This is probably true at least for the high vowels, judging from the measurement data of Linker (1982). Data presented by Wozniak and Jackson (1979) indicate that lip-reading of English vowels is as accurate from a frontal as from a lateral view of the talker's face. And while this view may not be best for judging lip protrusion, a frequent but not invariant accompaniment of "rounding", it is optimal for judging the distance between the corners of the lips, which is, according to Ladefoged (1982, p. 263), the physical correlate of degree of rounding.

[14] Given the notorious acoustic effects of opening the velopharyngeal port, which are not readily separated from those that are correlated with tongue height, we omit the nasalized vowels in this display.

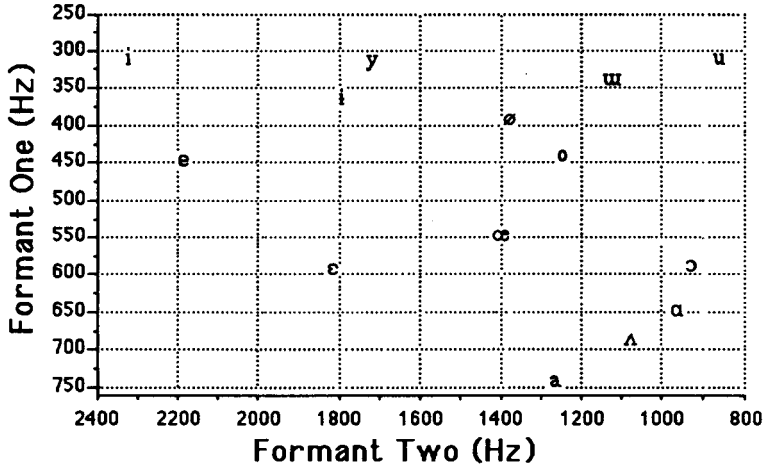

Fig. 1. Locations of the 14 oral vowels in the F1 X F2 plane.

A computer program was designed so that the acoustic onset of each vowel triggered a search for some specified one of the first eighteen recorded vowel tokens, which had been stored after digitization at a 10 kHz sampling rate. The vowel stored was then dubbed to the audio channel of a second video tape, in close synchrony with a copy of the visually recorded facial movements accompanying the original vowel articulation. Five randomly selected tokens of a given "visual vowel" were synchronized with a token of the auditory vowel quality appropriate to it in nature, while the remaining five tokens of the same vowel were combined with a vowel quality produced by a closely similar tongue position but different lip posture. Thus, for example, the [ɛ] image was combined five times with an [ɛ] sound quality, and five times with an [œ] sound quality, since these vowels are defined as being produced with identical tongue position and different degrees of lip rounding (International Phonetic Association, 1949). The following discrepant sound-image combinations were prepared: [i/y e/ɸ ɛ/œ ɛ̃/œ̃ ʌ/ɔ ɑ̃/ɔ̃ ɯ/u ɨ/o y/i ɸ/e œ/ɛ œ̃/ɛ̃ ɔ/ʌ ɔ̃/ɑ̃ u/ɯ o/ɨ ɸ/a o/ɑ]. (The inadvertent absence of the combinations a/ɸ and ɑ/o should be noted, so that [a/ɑ] were not combined in symmetrical fashion with any "rounded images".) It was expected that for many, if not all, of the first eight pairs the first vowel would elicit significantly fewer "rounded" judgments than the second, either by ear or by eye, and that the converse would be true for the remaining pairs, and that any visual within-pairs differences associated with differences in vowel height and/or backness would have little or no effect on rounding perception. Moreover, while tongue positions were unlikely to be identical in the [ɨ/o ɑ̃/ɔ̃ ɸ/a o/ɑ] combinations, and we cannot be entirely certain that they were in the others,[15] it is not far-

[15] For possibly relevant Dutch and German data see Raphael, Bell-Berti, Collier, and Baer (1979), Hoole and Tillmann (1991).

fetched to believe that a salient articulatory difference in all of them was that of rounding (Delattre, 1946).

Subjects

Test subjects were 20 native or near-native[16] speakers of French. They were closely associated with the Institute of Phonetics of the University of Provence in Aix-en-Provence, either as research staff or as graduate students pursuing research in phonetics, and were thus thoroughly familiar, both practically and theoretically, with the notion of lip rounding as an articulatory feature having linguistic significance.

Procedure

Testing was carried out over a period of about a month, each participant being tested on three different occasions. Each session involved the presentation of 180 stimuli to a single subject; its duration was just slightly more than ten minutes. The subject was seated in an anechoic chamber facing a loudspeaker and 25-inch television screen at a distance of between five and six feet. Sound and light levels were set for "comfortable" hearing and viewing.[17] At all times subjects were under visual observation afforded by a large sound-proofed window set into one wall of the chamber. Prior to each test session subjects were given verbal instructions (in French) concerning the nature of the specific task, and provided with an answer sheet on which to enter responses. In the initial sessions half the participants were exposed to the video signals alone, and were asked to respond to each stimulus vowel by writing a plus (+R) if they thought the lips were rounded and a minus (−R) if they thought not. The other ten participants were presented with the audio signals and asked to decide, in the absence of any visual information, whether they had been produced with or without lip rounding. In the second series of sessions each participant then performed the alternate task. The final sessions involved the presentation of combined auditory and visual stimuli. Since from a preliminary session it appeared likely that mismatches between what was heard and what was seen would sometimes be detected by at least some observers, prior to testing all participants were informed of the nature of the stimuli and the purpose of the experiment. They were told that they should base their judgments on auditory impression alone, in the event that they thought the audio and video signals were mismatched. Because, given that instruction, a subject might be tempted simply to avoid looking at the monitor screen, participants were tested individually and closely watched to make

[16] One subject is of North African background, and one other is of English origin; both have resided in France for many years, and speak French fluently and without detectable foreign accent. Moreover, their test responses differed in no way from those of the native speakers of French.

[17] Although none of our French-speaking subjects voiced any criticism of the tests, the complaints and difficulties of several American linguists tested later suggest that interstimulus intervals were too brief for comfort. In view of this, the high level of response consistency shown by our 20 subjects is all the more impressive.

sure that they were in fact giving their attention to the visual image as the acoustic signal sounded.

RESULTS

Taking it for granted that, strictly speaking, the "true" rounding status of a vowel can only be decided by phonetically trained observers making judgments based on visual information, we begin the description of our experimental results with an account of the responses to the video signals presented without accompanying sound.

Judgments by eye

According to their IPA definitions the 18 items of the stimulus set should fall into two groups: ten unrounded and eight rounded vowels. The visually based judgments are generally pretty much in agreement with these definitions, to the extent that all the vowels for which $+R_{mean}$ is less than 50% are unrounded per IPA prescription, and all those having $+R_{mean}$ greater than 50% are by the same token rounded (Figure 2). At the same time, however, despite the binary nature of both the instruction to the subjects and the assumed articulatory intentions of the speaker, we see no very sharp binary division into categories that we might call [–rnd] and [+rnd]. Thus, the two putatively unrounded vowels [ɑ ɑ̃] were judged rounded about 25%, i.e., more often than were, e.g., [i e ɛ], while the "rounded" vowels [ɔ œ̃ œ] elicited a percentage of +R responses only slightly better than chance.[18] If we decide, rather arbitrarily to be sure, to designate as [–rnd] those vowels for which $+R_{mean}$ judgments were less than 25%, and as [+rnd] those for which $+R_{mean}$ judgments exceeded 75%, then [ɑ̃ ɔ œ̃ œ] hardly qualify as either, and it seems prudent to accommodate them within a third category of vowels, "[?rnd]", of ambiguous rounding status.

It may be asked whether the deviation from the traditionally binary rounding classification of vowels reflects predominantly within-subject or across-subject uncertainty. When the responses of individual subjects to each vowel are subjected to tripartite classification (Figure 3), two observations seem warranted: (1) Responses were generally quite consistent, in that for the vowels classified as either [–rnd] or [+rnd] on the basis of mean percentage of +R responses only about 5% of the 280 10-item response sets (20 Ss × 14 vowels) fall within the 25–75% range. (2) For the four [?rnd] vowels, inconsistency of judgment ($25\% < +R_{mean} < 75\%$) characterizes just slightly more than half of the 80 10-item response sets. The remaining sets include both subjects predisposed to –R judgments and those with a contrary bias toward +R. The [?rnd]

---

[18] Although these vowels may not be identical with the vowels of standard French, it is possibly relevant that at least one group of phoneticians has described French /ɑ/ as rounded (Abri, Boë, Corsi, Descout, Gentil, and Graillot, 1980), and that many French speakers do not distinguish between the vowels commonly represented as [œ̃] and [ɛ̃]. Another phonetician, also a native speaker of French, omits any mention of /ɑ/ in his rounding classification of the vowels of the language (Delattre, 1946).

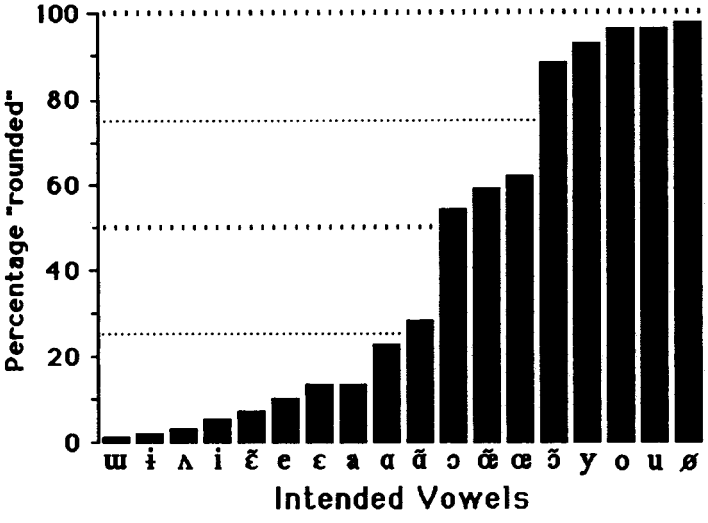

Fig. 2. Percent +R judgments of isolated vowels by eye.

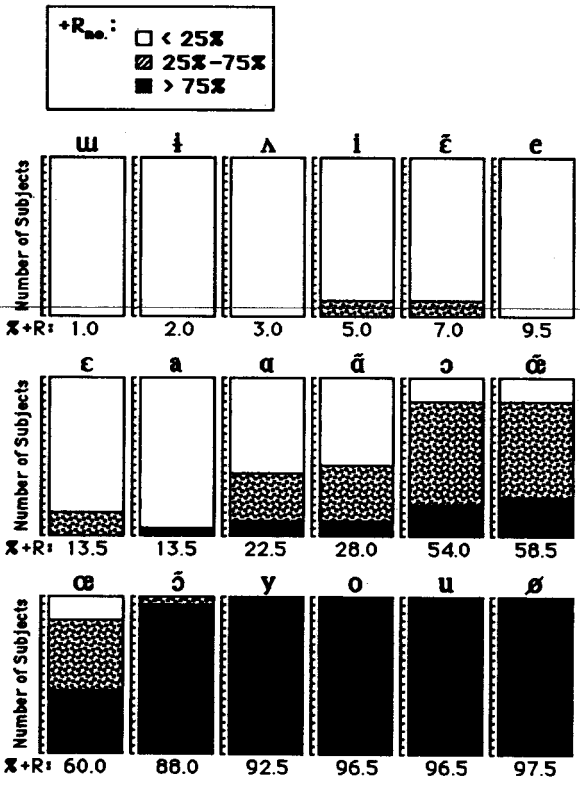

Fig. 3. Consistency of visual judgments by 20 subjects.

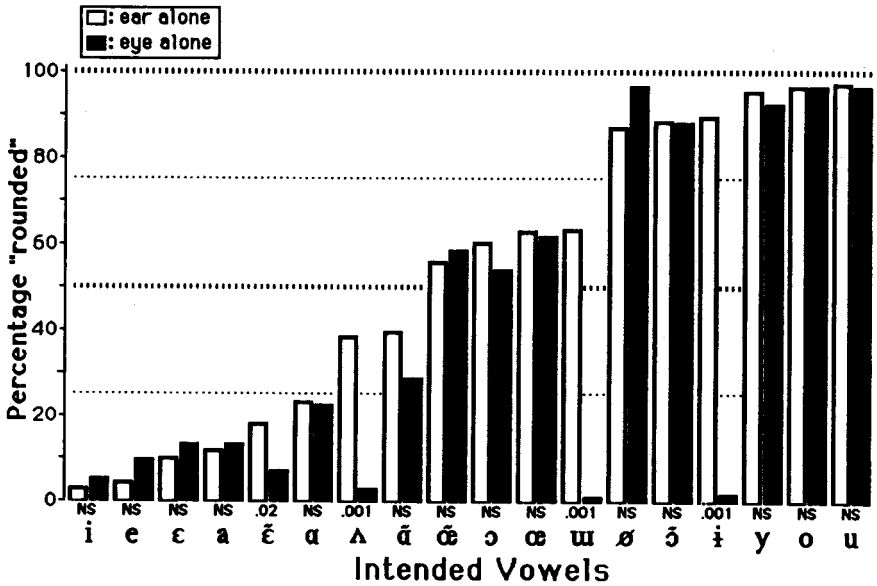

Fig. 4. Percent +R judgments of isolated vowels by ear and by eye. Levels of significance of Newman-Keuls tests are indicated.

category, then, reflects in about equal measure both within-subject and across-subject uncertainty.[19] While it may be true that "lip posture is perfectly obvious to inspection" (Abercrombie, 1967, p. 157), the binary classification of these vowels by visual inspection is clearly not all that simple a task.

Judgments by ear

Initial examination of the auditory rounding judgments in comparison with the visually based responses (Figure 4) suggests that, with a few exceptions, there is little overall difference between the two. However, an analysis of variance, with stimulus mode and phonetic identity as factors, reveals both to be important determinants of rounding classification: mode – $F(1, 19) = 33.365$, $p < 0.0001$; phonetic identity – $F(17, 323) = 86.757$, $p < 0.0001$. Interaction between the two factors is also significant [$F(17, 323) = 18.262$, $p < 0.0001$], so that the very large effect of stimulus mode turns out, per vowel-by-vowel *post-hoc* Newman-Keuls testing, to be limited to the four vowels [ɛ̃ ʌ ɯ ɨ]. Apart from these four vowels, which were all more often **heard** as rounded than

[19] Examination of responses token by token does not indicate variability in production to be a factor explaining response uncertainty.

seen to be so,[20] differences between mean ear and eye judgments were not significant ($p > 0.1$), i.e., the rounding information provided to the ear was in the main correctly interpreted. Thus [i e ɛ a ɑ] can be said to be unrounded both by eye and by ear ($+R_{mean} < 25\%$), and both modes of presentation indicate rounding for [ɸ ɔ̃ y o u] ($+R_{mean} > 75\%$). Moreover, the visually ambiguous vowels [ɑ̃ ɔ̃ œ̃ œ] ($25\% < +R_{mean} < 75\%$) were just as uncertainly judged by ear. For three of the four vowels judged differently by ear and eye, differences were large enough to justify assigning them to different rounding categories in the two modes: [ʌ] and [ɯ] [?rnd]$_{ear}$ and [i] [+rnd]$_{ear}$, but all three [−rnd]$_{eye}$.[21] The striking difference between auditory and visual classifications in the case of [i] suggests the possibility that, under appropriate testing, our observers would accept a combination of the [i]-sound with the [o]-image as ecologically more valid than the same sound accompanied by its matching [i]-image.

The failure of our subjects, as a group, to judge all 18 vowels as being clearly either [−rnd]$_{ear}$ or [+rnd]$_{ear}$ suggests the same question that was raised about the visually derived data: Does the [?rnd]$_{ear}$ category reflect vacillation by individual subjects or differences of opinion between subjects? As with the judgments by eye, we find a considerable degree of consistency in auditory judgment (Figure 5): All subjects heard most of the vowels as either [−rnd]$_{ear}$ or [+rnd]$_{ear}$, with only 15% of the 360 10-item response sets 20 Ss X 18 vowels) reflecting individual uncertainty. However, cross-subject variability of auditory judgments was significantly greater, as determined from a comparison of the mean standard deviations of ear and eye judgments by paired two-tailed *t*-test ($t = 5.827$, $p \leqslant 0.0005$). A closer look at the data for the vowels [ɑ̃ œ̃ ɔ œ], which both visually and auditorily are [?rnd], indicated that the lesser variability of the visual judgments masks the fact that many more subjects showed greater consistency in their auditory judgments. Thus, for example, the vowel [œ̃], for which auditory judgments were most evenly divided between −R and +R ($+R_{mean} = 56\%$), was inconsistently judged (i.e., $25\% < +R_{mean} < 75\%$) by only six of the 20 listeners; six regularly reported it to be −R, while eight others just as decisively labeled it +R. (This same vowel, on the basis of visual information, was inconsistently classified by a majority of the subjects.) The uncertainty of the six listeners who "split their vote" more or less evenly between −R and +R responses is possibly explained by the presumed "open rounding" of this vowel, more rounded than the "spread" vowels [i e], but less rounded than the "closely rounded" vowels [y ɸ u] (International Phonetic Association, 1949, p. 6).

---

[20] The rounded quality attributed to [ʌ ɯ i] is expecially striking in view of the fact that all tokens of these vowels were produced with extreme retraction of the lip commissures, although this is not explicitly called for by their IPA specifications.

[21] These data, then, seem to bear out Riordan's (1977) conclusion that "the position of the lips is not crucial for the production of rounded vowels", and/or that observers are not in substantial agreement as to just how much rounding, if any, is necessary for a vowel to count as "rounded".

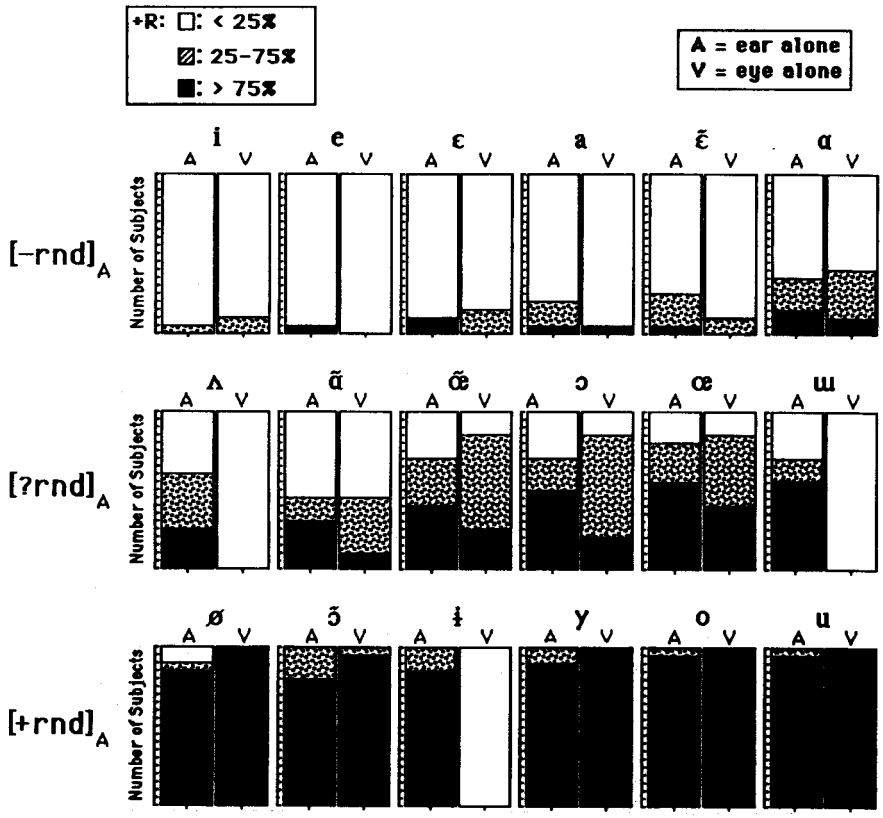

Fig. 5. Comparison of subjects' consistency in judging vowels by ear and by eye.

Audiovisual judgments

In preparing our audiovisual stimuli we had made three assumptions about the nature of their acoustic and visual components: (1) that matching components, presented separately, would signal more or less the same lip position; (2) that in each of the 16 discrepant combinations the two vowels, one auditory and the other visual, would be products of distinctly different degrees of rounding, and that this difference would be reflected in both auditory and visual judgments; (3) that the audiovisual stimuli composed of matching acoustic and visual signals would elicit responses consistent with those evoked by their components when separately presented. In other words, the perceived rounding of a vowel should not depend on which of the three ecologically valid modes of presentation was used. We have already seen, however, that for the

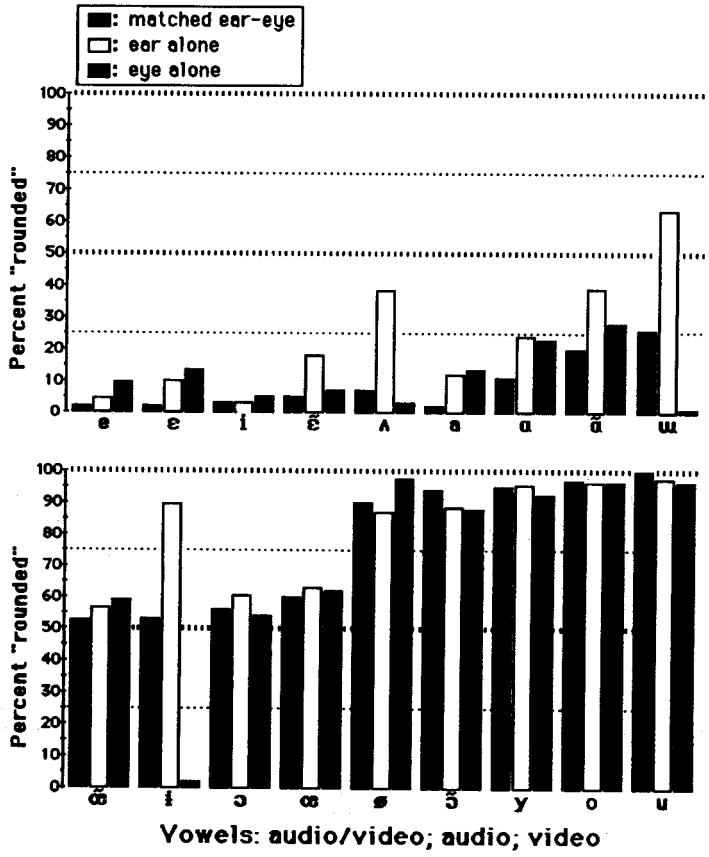

Fig. 6. Rounding judgments of matched audiovisual stimuli, compared with ear-alone and eye-alone responses.

unfamiliar vowels [ʌ ɯ i] the rounding classification based on our subjects' responses differs, depending on their mode of presentation. Since the matched combination of an acoustic and optical signal provides more phonetic information than does either by itself, it would seem reasonable to regard the former as the most reliable basis for vowel classification.

As the data of Figure 6 show, partitioning of the 18 vowels of our matched audiovisual stimuli on the basis of the quartile criterion yields nine audiovisually unrounded vowels, [e ɛ i ɛ̃ ʌ a ɑ ɑ̃ ɯ], and five rounded ones, [ø ɔ̃ y o u], with the four remaining vowels of ambiguous status, [œ̃ i ɔ œ].[22] An overall comparison of these responses with

[22] Strictly speaking, the vowel [ɯ] should be grouped with the ambiguous set, since it elicited 26% rounded judgments, but the pattern of responses was strongly unimodal, with half the observers unanimously calling it unrounded, so that it is neither clearly [–rnd] or [?rnd].

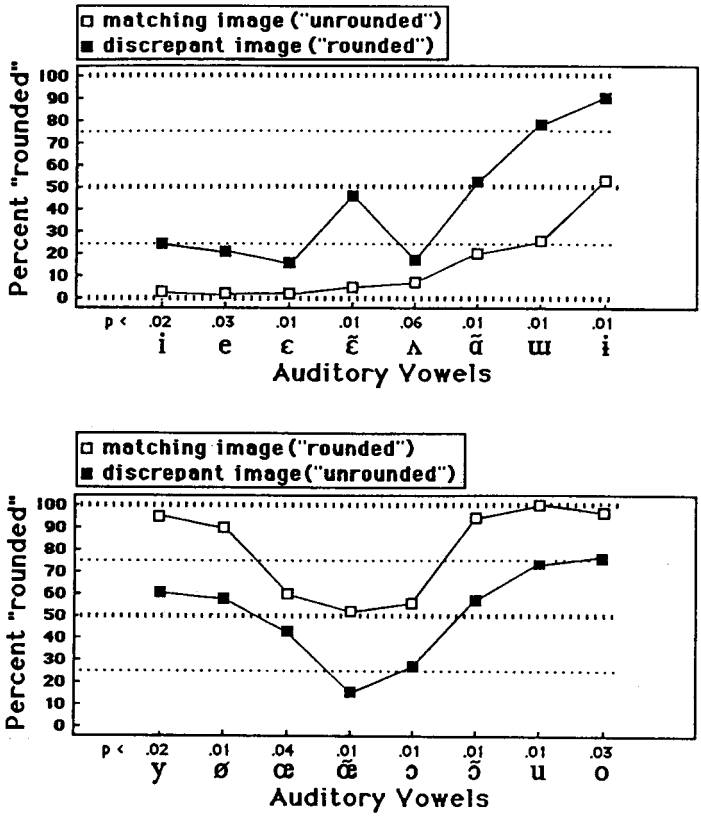

Fig. 7. Comparison of rounding judgments of individual vowel sounds with matched *vs.* discrepant images, with level of significance for each difference.

the auditory and visual judgments reveals a picture similar to the one found for the relation between auditorily based and visually based responses: While stimulus mode is overall highly significant as a factor affecting rounding judgment, its effect is again confined to a small subset of the vowels, this time [ɛ] as well as [ɛ̃ ʌ ɯ ɨ]. For the first two of these vowels, which were most often judged unrounded both by eye and ear separately, the combination of auditory and visual information elicited even fewer +R responses. (Although the vowels [e ɑ̃] were also audiovisually more unequivocally [–rnd] than by eye or ear alone, intermodal differences fall short of being significant.) On the other hand, the audiovisual presentations of the three vowels [ʌ ɯ ɨ], whose mean visual and auditory evaluations put them into different categories, elicited $+R_{mean}$ values somewhere intermediate between the two.

Since for the vowels other than [ɛ ɛ̃ ʌ ɯ ɨ] the information provided by a matching image had no significant effect on mean rounding judgments, we must turn to the

responses elicited by the audiovisual stimuli made up of sounds and discrepant images in order to uncover any measurable lipreading effect on their phonetic evaluation. An analysis of variance was applied to compare responses to matching and discrepant combination of the 16 vowels that were used in our symmetrically paired audiovisual stimuli. From the analysis a difference between mean responses to the two kinds of audiovisual stimuli turns out to be highly significant [$F(15, 285) = 4.020$; $p = 0.001$]. As a group, then, our subjects were not successful in basing their judgments of rounding solely on the auditory components of the complex stimuli.

When the responses to matching and discrepant audiovisual stimuli are examined vowel by vowel (Figure 7), *post hoc* testing indicates that in all but one case ([ʌ]: $p < 0.06$) there is a significant difference in rounding perception, depending on whether the auditory vowel is accompanied by a matching or a discrepant image. Thus every one of the "rounded" vowels (as per IPA definition) elicited fewer +R responses when accompanied by a discrepant image. With the "unrounded" vowels (by IPA definition), while the effect of a matching image is rather less consistent relative to purely auditory judgment, the relation between judgments of matching and discrepant audiovisual stimuli is just the same: discrepant stimuli were more often judged rounded.

Since the presence of discrepant visual information posed the severest test of our subjects' ability to make purely auditory rounding judgments, it is of interest to compare more closely the responses to discrepant combinations with the rounding evaluations of their constituent signals (Figure 8). First of all, we see that virtually all these stimuli qualify as perceptually discrepant, the single exception being the combination of auditory [i] and visual [o]. For three audiovisual combinations, [ɛ/œ e/ɸ ɑ̃/ɔ̃], mean +R responses are not significantly different from judgments of their auditory components, i.e., the subjects as a group here apparently succeeded in responding purely on the basis of the auditory signal. On the other hand, in the judgments of [œ̃/ɛ̃] and [ɛ̃/œ̃] the dominant influence seems to have been the visual rather than the auditory component. For the remaining ten audiovisual stimuli judgments were significantly different from both auditory and visual judgments, with mean number of +R responses lying somewhere between those elicited separately by their acoustic and optical components.[23]

With fully half the discrepant audiovisual combinations eliciting $+R_{mean}$ values falling within the 25%–75% range, it would seem that subjects' responses indicate a significant amount of integration of auditory and visual information. However, an examination of the responses of individual subjects reveals that the large number of ambivalent $+R_{mean}$ values in many cases does not reflect perceptual integration of auditory and visual information. This is especially the case for the audiovisual combinations [i/y y/i e/ɸ ɸ/e], made up of audio and video components maximally different in degree of perceived rounding (Figure 9). Thus, for example, the [i/y] combination was ambiguously judged by only three subjects – 14 reported it to be unrounded ($+R_{mean}$

[23] The sole exception is the [ʌ/ɔ] combination, which elicited fewer rounded judgments than did either component individually, and thus suggests the intuitively unacceptable conclusion that the effect of the relatively rounded visual [ɔ] was to decrease the rounded quality of the auditory [ʌ].

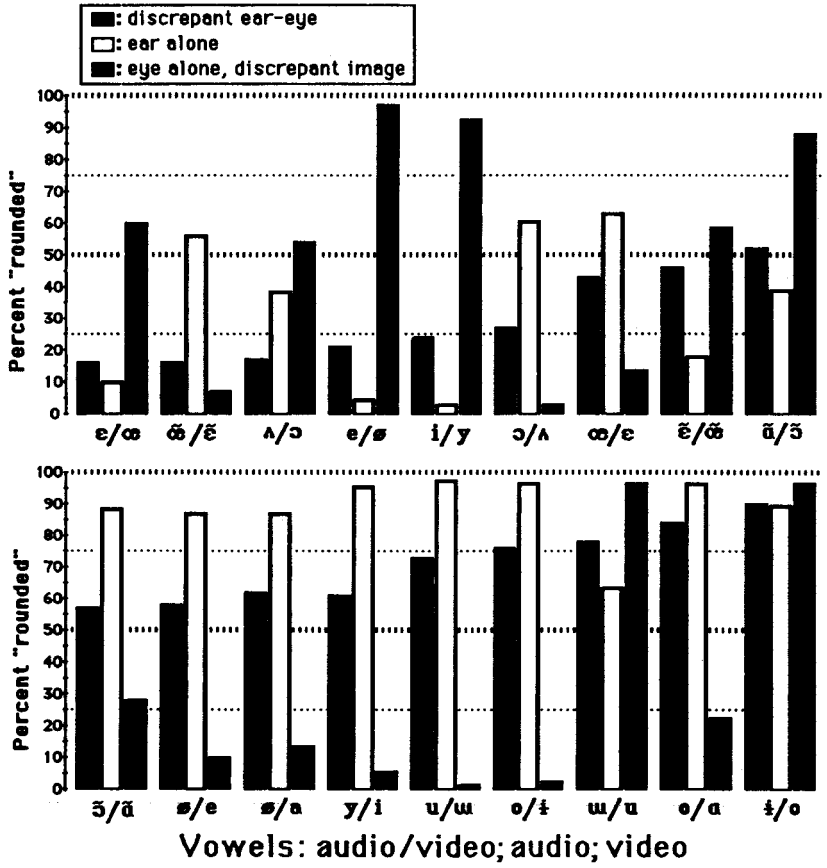

Fig. 8. Rounding judgments of individual vowel sounds in discrepant audiovisual stimuli compared with perceived rounding of their components when judged alone.

< 25%), while three heard it as rounded (+R_{mean} > 75%). For the [e/ϕ] combination the responses even more clearly indicate that subjects chose just one input mode as the source of information – 16 heard this stimulus as unrounded, while the remaining four decided that it was rounded. For these two discrepant combinations, then, subjects made consistent judgments, and most of them, moreover, **succeeded** in basing their judgments solely on the auditory signals, notwithstanding the obvious rounding of their visual accompaniments. The picture is somewhat similar for the [y/i ϕ/e] combinations, in that very few subjects gave other than consistent rounding judgments. However, for these stimuli with [+rnd] auditory components, subjects were not quite as successful

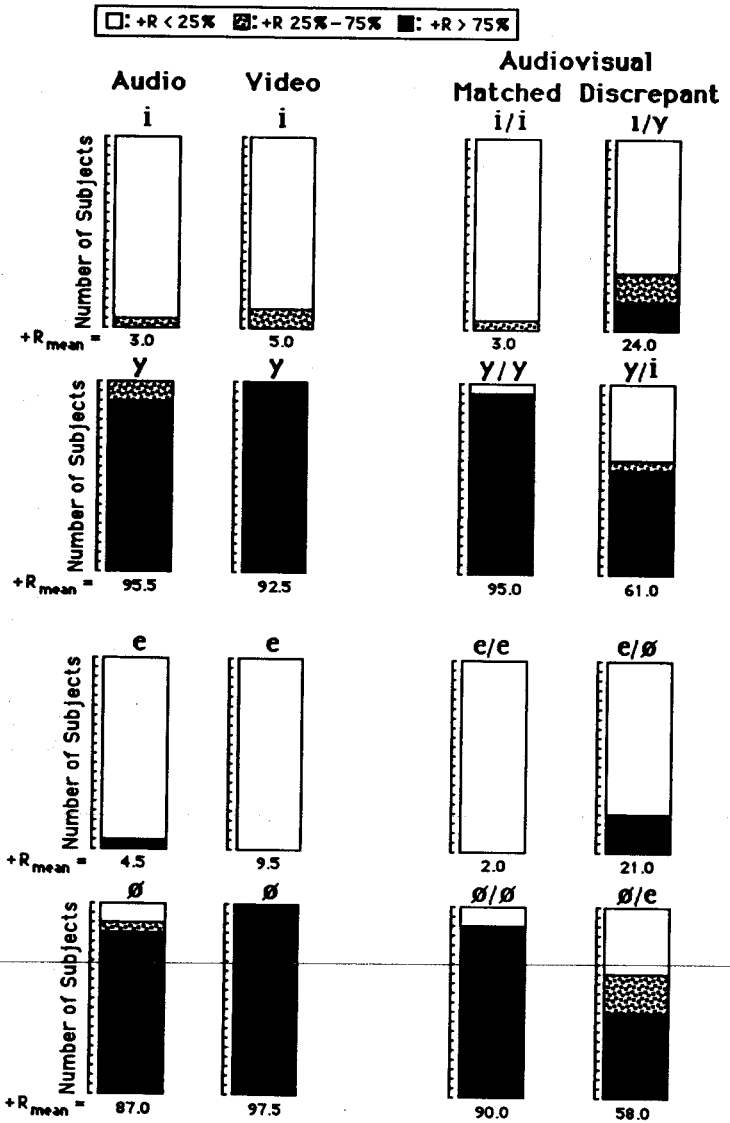

Fig. 9. Individual consistency of 20 subjects in judging matching and discrepant combinations of auditory and visual vowels minimally ambiguous with respect to rounding.

in discounting contrary visual information as they were when assessing the [i/y] and [e/ø] combinations.

In view of the instruction given, it is perhaps not very surprising that in the judgment of discrepant stimuli auditory "rounding" exercised a somewhat stronger overall effect than visual evidence of rounding.[24] Thus, a discrepant combination of an unrounded auditory signal with a rounded visual image was less often reported as rounded than the combination of a rounded auditory signal with an unrounded visual signal. This is most clearly evident in the responses to the strongly [–rnd] vowels [i e] in combination with the unambiguously [+rnd] vowels [y ø]. Thus, while audiovisual [i/y] and [e/ø] combinations elicited somewhat less than 25% +R responses, for the [y/i] and [ø/e] combinations the number of such responses was about 60%. Three combinations for which $[-\text{rnd}]_{\text{ear}}$ + $[+\text{rnd}]_{\text{eye}}$ combinations elicited more +R responses, [ɛ̃/œ̃ ɯ/u i/o], are cases where one member of the pair was often not "correctly" judged even when auditory and visual signals were recordings of the same event. (That is to say, [ɯ i] were not judged auditorily as per their purely visual assessments, while [œ̃] was not judged very often as rounded, as per the phonetic prescription of "œ̃", under any ecologically valid mode of presentation.[25]) While an analysis of variance, with difference in the manner of combination of [–rnd] and [+rnd] information as one factor and "tongue position" as the other, shows that manner of combination of discrepant auditory and visual information is overall not significant, a closer examination of individual combinations reveals significant differences for the front vowel combinations [i/y e/ø ɛ/œ ɛ̃/œ̃]. For the first three of these auditory information was significantly more important ($p < 0.05$) in determining rounding responses, but in judgments of the [ɛ̃/œ̃] combination it was the visual signal that even more strongly determined rounding judgments ($p < 0.01$). Since [ɛ̃] was judged +R only 18% by ear alone and 7% by eye alone, while for [œ̃] the corresponding values are 56% and 58.5%, this latter finding seems anomalous and must at this point remain unexplained (though it is tempting to look for some explanation based on the phonologically ambiguous status of the [ɛ̃]–[œ̃] relationship in French).

DISCUSSION

In studies of lipreading it is generally assumed that visual indications of vocal tract activity play no detectible role in speech perception unless the acoustic signal alone is

---

[24] An instruction to base rounding judgments on the visual evidence, which after all is how phoneticians might be expected actually to decide a vowel's status in this respect, would tell us how much auditory quality can perturb the perception of rounding.

[25] The relationship between [ɛ̃] and [œ̃] is not readily explained by the reported tendency toward the neutralization of the contrast between the phonetically similar phonemes in contemporary French, since both by ear and by eye [œ̃] elicited significantly more +R judgments than did [ɛ̃].

not fully intelligible. Otherwise, by definition, it can do no more than ratify the percepts cued by the acoustic signal. (The reverse is not the case: Lipreading rarely if ever conveys a linguistic message so unambiguously that its audible accompaniment is redundant.) By the perception of a speech signal is usually meant its identification as a particular sequence of linguistic elements (words, phonemes) "intended" by its producer. This requires no more of the perceiver than the linguistic competence that virtually all members of a speech community share, together with knowledge of an alphabetic writing system. But speech perception may also be understood to include the identification of the phonetic properties that characterize words and phonemes. Awareness of these properties, also in some sense intended by the speaker, usually comes only with explicit linguistic/phonetic training. Moreover, the ability to perceive phonetic properties is not necessarily better by ear than by eye. Thus, while lipreading contributes nothing to the perception of voicing or nasalization, the rounding status of a vowel or other phonetic element is by definition determined by how phonetically trained observers see it, not how they hear it. Unlike the situation in the case of word and phoneme perception, it may be anticipated that auditory and visual judgments of vowel rounding will sometimes differ, and if they do, then the latter are to be preferred, that is, unless "rounding" is **defined** as an auditory quality.

The rounding decisions reported by our group of trained observers indicate that while auditory rounding judgments of many vowels may closely match those based on visual inspection, this need not be so for all vowels. Despite phonetic awareness and native command of a vowel system in which rounding is distinctive, the ability of our subjects to make auditory judgments consistent with visual judgments does not appear to extend beyond the vowels familiar to them as native speakers of French. In fact, their auditory assessments of the unfamiliar vowels [ʌ ɯ ɨ] agree to a startling degree with auditory judgments of the Daniel Jones recordings of "the same" cardinal vowels by an otherwise similar group of English-speaking subjects (Lisker, 1989). The ability to diagnose the auditory effect of unrounded *vs.* rounded lip positions on high front vowels was somewhat greater for the French speakers, but they were as ready as the English speakers to report hearing the high non-front vowels [ɯ] and [ɨ] (the second in particular) as rounded.[26] In other words, both groups tended to confound the acoustic properties marking a vowel as non-front with the effects of lip rounding. Thus any greater sensitivity of the French group to the [±rnd] dimension, plausibly attributable to its distinctive role in French, is limited to the region within the vowel space in which rounding is utilized distinctively. Elsewhere rounding is no more distinctive in French than it is in English, and auditory judgments of rounding are no more reliable for speakers of one language than the other.

---

[26] Given the fact that F2 of [ɯ] is much lower than that of [ɨ] (Figure 1), one might wonder why the second was more often judged rounded. Perhaps the proximity of [ɨ] to [y] in the F1 × F2 plane led subjects to judge it similarly in relation to the front unrounded vowels, though none of the three persons who made auditory identifications of the vowels assigned any token of [ɨ] to French /y/. Perhaps [ɯ] was more often heard as a distorted [u], but with considerable uncertainty as to whether the distortion was a matter of being less backed or less rounded than [u].

Whatever the basis on which such phonetic judgments are usually made,[27] if the rounding status of a vowel is ultimately ascertainable only by eye, we should expect visual judgments to be more consistent than judgments by ear. This we found to be the case when relative consistency was determined by comparing the means of standard deviations of $+R_{mean}$ judgments by ear and by eye over the 20 subjects (Figure 6). For those vowels perceived largely as [−rnd] or [+rnd] in both modes, mean standard deviations of auditory judgments are roughly twice those of visual judgments. But for the vowels of uncertain rounding status (our [?rnd] category) the greater consistency of the visual judgments is not a reflection of greater agreement as to the rounding status of those vowels: Subjects were individually less consistent by eye than by ear, and the smaller mean standard deviation of their visual judgments reflects the greater number of subjects who failed to judge the visual vowels consistently. If phonological considerations force us to make a binary choice, then the basis we choose for deciding the rounding status of the [ɑ̃ œ̃ ɔ œ] vowels cannot be the responses of a jury of trained subjects with native competence in French, at least as these were elicited under the conditions of our experiments. (Possibly a test asking for selection of the more rounded member of each of the pairs [ɑ̃/ɔ̃ œ̃/ɛ̃ ɔ/ʌ œ/ɛ] would yield perceptual data in closer agreement with our expectations.)

As for the role of lipreading in deciding whether or not a vowel quality was produced with rounding, our findings at first glance are consistent with those that first demonstrated the "McGurk effect" on obstruent consonant identification (McGurk and Macdonald, 1976; Macdonald and McGurk, 1978), as well as with those showing similar effects on vowel perception (Summerfield and McGrath, 1984). Thus an audio-visual stimulus with components derived from two phonetically different speech events of maximum simplicity elicits judgments that in sum reflect both components, even from a group of phonetically trained listeners alerted to the possibly misleading nature of the visual signal, and despite the fact that its auditory component is similar to, or perhaps even identical with, a vowel of their native language having a generally recognized status with respect to distinctive rounding. However, there were significant differences in perception across both subjects and audiovisual combinations, and they lead us to a conclusion quite different from those based on experiments in phoneme identification. Audiovisual stimuli involving components of uncertain rounding status led an appreciable minority of subjects (never a majority) to give responses reflecting uncertainty of judgment, so that we may suppose there was perceptual integration of the two kinds of information. On the other hand, most of the response sets elicited by stimuli conveying very different auditory and visual information show no significant lipreading effect, while, contrariwise, many of the remaining response sets reflect no influence of the auditory information. The perception of such stimuli, we conclude therefore, does not as a rule involve integration of the information to ear and eye on the part of individual subjects; most often subjects simply differ in which kind of information they

[27] According to Ladefoged (1960), "a precise statement about the degree of lip-rounding is usually a statement about what the vowel sounds like, and not necessarily about what it looks like".

attend to as the basis for judgment. For a small minority of subjects lipreading may be said to exercise a "perturbing" effect on the auditory judgment of rounding, but by and large the effect is only apparent when we sum the responses of all subjects.

The summed data, though not the responses of individual subjects, also provide a further striking demonstration of the importance of visual information for speech perception, in that a measurable lipreading effect is not limited to ecologically invalid combinations of auditory and visual signals. It is seen when auditory and visual responses to unfamiliar vowels, particularly [ɯ] and [ɨ], are compared with those elicited by their presentation in **matched** audiovisual combinations. It may be noted that in the case of the vowel [ɨ] we may even speak of a phonetic "McGurk effect", for while this vowel falls auditorily into the [+rnd] category and visually is [−rnd], audiovisually it is neither, i.e., it is [?rnd]. But this McGurk effect is spurious, since no more than a small minority of subjects failed to judge audiovisual [ɨ] consistently either as [−rnd] or [+rnd]. For those who called it unrounded, we cannot decide whether they did so despite the "rounded" quality of the auditory signal, or because they recognized the source of the unfamiliar vowel quality to be a vocal tract articulating the un-French combination of backed tongue position with unrounded lips. The greater number of subjects who judged [ɨ] in the presence of its matched and strongly unrounded visual image most certainly were responding only to the auditory signal when they attributed its quality, erroneously, to rounding of the lips.

Since matched audiovisual stimuli provide more information than either acoustic or visual signals alone, we might expect them to be judged more consistently, Our data, however, do not bear out such an expectation. (Quite possibly they would have, if subjects had had no reason to be uncertain of their ecological validity.) While judgments of the matched audiovisual combinations are generally more consistent than those based solely on the acoustic signal, it is the visually based judgments that show the smallest degree of variability. We may suppose that awareness of the possibility of audiovisual discrepancy puts the subject squarely between the horns of a dilemma – whether to follow the instruction to discount the visual signal or to take the more "reasonable" course of using it as the primary basis for the rounding decision, and treating the acoustic signal as effectively the "distractor", despite the instruction to report on the basis of auditory impression. This instruction, in demanding that the auditory signal alone be treated as relevant, is tantamount to asking the subject to disregard that component of the audiovisual stimulus which most unambiguously conveys the information needed to make the required judgment. Thus, in the transmission of at least one particular kind of phonetic information, it seems that the visual signal is not only less ambiguous than the auditory (although its binary classification is less sure), it can be even "better" when no auditory information is available, perhaps especially for subjects whose linguistic experience has led them to have particular expectations as to the relationship between lip position and auditory quality over much, but not all, of the vowel space.

Finally, it might be objected that the examination of phonetic judgments of the kind we have been considering is at most only marginally relevant so far as advancing our understanding of speech perception as narrowly defined, in which the explicit recognition of phonetic properties plays no role. Such an objection might remind us of

the many studies of speech and language behavior in which subjects are chosen for their phonetic/linguistic naiveté, with the implication that a study of non-naïfs could very possibly yield results of doubtful (in some sense) validity. According to such a view, the proper object of interest is the behavior of typical members of a language community, and not the linguist's metalinguistic activity. The linguistic awareness that informs the behavior of the typical language user is no doubt of different degree from that of the linguist-phonetician, and certainly must be a matter of central interest to students of language behavior. At the same time, however, the role played in speech research by the phonetic judgments of linguists should not be minimized. After all, to the extent that phoneticians and linguists are able, perhaps by virtue of their professional qualification, to transcend perceptual biases imposed by the phonetic/phonological patterns of their native language, their behavior should be more indicative of general limitations on our ability to relate speech sound to speech sound production. The discrepancies we have found between auditory and visual judgments of rounding are hardly a basis for arguing the irrelevance of phoneticians' opinions as compared to the direct examination of articulatory behavior, for rounding as a dimension of phonetic classification, like other phonetic features, does not emerge of itself from the raw data of physical examination. The glaring discrepancy between the auditory and visual assessments of a very few of our vowels means that the phonetic dimension of rounding is not always directly accounted for by lip posture. Of course, it is conceivable (though in our view most unlikely) that the same phonetician who judged a vowel of [i]-quality to be rounded not only could, but **would** mimic that quality without any apparent lip activity. But even from such a hypothetical situation we would argue that the behavior of phoneticians *is* a legitimate object of research attention, since to demonstrate the mismatch between judgment and attempted mimicry would require corroboration by another phonetician (or phoneticians), who would be obliged to express two opinions: (1) that the auditory quality of the "repetition" was acceptably close to the target, and (2) that it was produced without visible rounding. An opinion, whatever its source, as to whether or not two speech events are the same in quality is at best of uncertain scientific status, but at least on intuitive grounds it seems reasonable, *faute de mieux,* to accept the dictum that "two vowels can be equated if, and only if, a trained phonetician regards them as being the same" (Ladefoged, 1960).[28]

In short, in speech research there is ultimately no way to evade the obligation to make phonetic judgments, including both interpretations of the "hard" data of laboratory measurements and opinions that may be difficult to formulate in terms allowing for disconfirmation on the basis of "objective" observational data. This is certainly the case when we deal with the elusive property or properties of "vowel quality" – a notion which ultimately has no meaning other than auditory – even if linguists are by training inclined to label it with ostensibly articulatory terms. Speech sounds do, of course, tell us a good deal about what a human vocal tract is doing, but so far as the perception of vowel quality is concerned, they do not always inform us precisely about the individual

---

[28] The only improvement on the opinion of a "trained phonetician" would be, according to Hurford (1969), the opinions of two such persons.

contributions of tongue, lips, and other associated parts (such as the larynx). When the ear tells the phonetician that the lips are rounded but the eye sees otherwise, it is not enough simply to dismiss the phonetician as "not competent", for whatever proof of competence we might choose, it cannot be performance of the required task in accordance with our expectation. Instead we must ask what acoustic properties give rise to the rounded quality, how they resemble the properties produced by visually attested rounding, and what other articulatory maneuvres might plausibly be held responsible for the illusion of lip rounding. The eye, we know, cannot capture directly all the linguistically motivated gestures of the vocal tract, and even has difficulty detecting the lip rounding of low vowels, but the ear may not be any better when it comes to distinguishing between the effects of lip rounding and other maneuvers, such as tongue backing, that have similar effects on the acoustic output of the vocal tract.

(Received August 11, 1989; accepted August 14, 1992)

REFERENCES

ABERCROMBIE, D. (1967). *Introduction to General Phonetics.* Edinburgh: Edinburgh University Press.

ABERCROMBIE, D. (1985). Daniel Jones's teaching. In V.A. Fromkin (ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged* (pp. 155–224). Orlando, FL: Academic Press.

ABRI, C., BOË, I.-J., CORSI, P., DESCOUT, R., GENTIL, M., and GRAILLOT, P. (1980). *Labiabilité et Phonétique. Données fondamentales et études expérimentales sur la géométrie et la motricité labiales.* Grenoble: Université des Langues et Lettres de Grenoble.

BLOOMFIELD, L. (1933). *Language.* New York: Henry Holt.

BOYCE, S.E., KRAKOW, R.A., BELL-BERTI, F., and GELFER, C.E. (1990). Converging sources of evidence for dissecting articulatory movements into core gestures. *Journal of Phonetics,* **18,** 173–188.

CATFORD, J.C. (1977). *Fundamental Problems in Phonetics.* Edinburgh: Edinburgh University Press.

CATFORD, J.C. (1981). Observations on the recent history of vowel classification. In R.E. Asher and E. Henderson (eds.), *Towards a History of Phonetics* (pp. 19–32). Edinburgh: Edinburgh University Press.

DELATTRE, P. (1946). *Principes de Phonétique Française à l'Usage des Étudiants Anglo-Saxons.* Middlebury, VT: The College Store.

DELATTRE, P. (1968). La radiographie des voyelles françaises et sa correlation acoustique. *French Review,* **XLII,** 48–65.

DELATTRE, P., LIBERMAN, A.M., COOPER, F.S., and GERSTMAN, L.J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word,* **8,** 195–210.

EASTON, R.D., and BASALA, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics,* **32,** 562–570.

EWAN, W.G., and KRONES, R. (1974). Measuring larynx movement using the thyroumbrometer. *Journal of Phonetics,* **2,** 327–335.

FANT, G. (1960). *Acoustic Theory of Speech Production.* The Hague: Mouton.

FROMKIN, V. (1964). Lip positions in American English vowels. *Language and Speech,* **7,** 215–225.

HEFFNER, R-M.S. (1969). *General Phonetics.* Madison: University of Wisconsin Press.

HENDERSON, E.J.A. (1971). *The Indispensable Foundation: A Selection of the Writings of Henry Sweet.* London: Oxford University Press.

HOOLE, P., and TILLMANN, H.G. (1991). An articulatory investigation of front rounded and unrounded vowels. In *Actes du XII^ème Congres International des Sciences Phonétiques*, **Vol. 2** (pp. 362–365). Aix-en-Provence: Université de Provence.

HURFORD, J.R. (1969). The judgment of vowel quality. *Language and Speech*, **12**, 220–237.

International Phonetic Association (1949). *The Principles of the International Phonetic Association* (2nd ed.). London: University College.

JOOS, M. (1948). *Acoustic Phonetics*. Language Monograph No. 23. Baltimore MD: Waverly Press.

KUHL, P.K., and MELTZOFF, A.N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, **7**, 361–381.

LADEFOGED, P. (1960). The value of phonetic statements. *Language*, **36**, 387–396.

LADEFOGED, P. (1967). The nature of vowel quality. In *Three Areas of Experimental Phonetics* (pp. 50–142). London: Oxford University Press.

LADEFOGED, P. (1982). *A Course in Phonetics* (2nd ed.). New York: Harcourt Brace Jovanovich.

LADEFOGED, P., HARSHMAN, R., GOLDSTEIN, L., and RICE, L. (1978). Generating vocal tract shapes from formant frequencies. *Journal of the Acoustical Society of America*, **64**, 1027–1035.

LADEFOGED, P., and MADDIESON, I. (1990). Vowels of the world's languages. *Journal of Phonetics*, **18**, 93–122.

LINDAU, M. (1978). Vowel features. *Language*, **54**, 541–563.

LINDBLOM, B.E.F., and SUNDBERG, J.E.F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, **50**, 1166–1179.

LINKER, W. (1982). Articulatory and acoustic correlates of labial activity in vowels: A cross-linguistic study. *UCLA Working Papers in Phonetics*, **56**, 1–134.

LISKER, L. (1989). On the interpretation of vowel "quality": The dimension of rounding. *Journal of the International Phonetic Association*, **19**, 24–30.

MACDONALD, J., and MCGURK, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, **24**, 253–257.

MADDIESON, I. (1984). *Patterns of Sounds*. Cambridge U.K.: Cambridge University Press.

MARTIN, S.E. (1959). *Easy Japanese*. Rutland, VT: Charles E. Tuttle Co.

MATTINGLY, I.G. (1990). The global character of phonetic gestures. *Journal of Phonetics*, **18**, 445–452.

MCGURK, H., and MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746–748.

MERMELSTEIN, P. (1978). On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perception & Psychophysics*, **23**, 331–336.

MONTGOMERY, A.A., and JACKSON, P.L. (1983). Physical characteristics of the lips underlying lipreading performance. *Journal of the Acoustical Society of America*, **73**, 2134–2144.

NEAREY, T.M. (1980). On the physical interpretation of vowel quality: cinefluorographic and acoustic evidence. *Journal of Phonetics*, **8**, 213–241.

PERKELL, J.S. (1969). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Cambridge, MA: MIT Press.

RAPHAEL, L.J., BELL-BERTI, F., COLLIER, R., and BAER, T. (1979). Tongue position in rounded and unrounded front vowel pairs. *Language and Speech*, **22**, 37–48.

RIORDAN, C.J. (1977). Control of vocal-tract length in speech. *Journal of the Acoustical Society of America*, **62**, 998–1002.

ROACH, P. (1983). *English Phonetics and Phonology: A Practical Course*. Cambridge, U.K.: Cambridge University Press.

ROCHET, B.L. (1991). Perception of the high vowel continuum: A crosslanguage study. In *Actes du XII^ème Congres International des Sciences Phonétiques*, **Vol. 4** (pp. 94–97). Aix-en-Provence: Université de Provence.

SEKIYAMA, K., and TOHKURA, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, **90**, 1797–1805.

STEVENS, K.N., and HOUSE, A.S. (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, **27**, 484–493.

SUMMERFIELD, Q. (1983). Audio-visual speech perception, lipreading and artificial stimulation. In M.E. Lutman and M.P. Haggard (eds.), *Hearing Science and Hearing Disorders* (pp. 131–182). London: Academic Press.

SUMMERFIELD, Q., and MCGRATH, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, **36A**, 51–74.

TSEVA, R. (1989). L'arrondissement dans l'identification visuelle des voyelles du français. *Bulletin du Laboratoire de la Communication Parlée*, **3**, 149–186. Grenoble: Institut National Polytechnique de Grenoble.

TULLER, B., and FITCH, H.L. (1980). Preservation of vocal tract length in speech: A negative finding. *Journal of the Acoustical Society of America*, **67**, 1068–1071.

WOZNIAK, V.D., and JACKSON, P.L. (1979). Visual vowel and diphthong perception from two horizontal viewing angles. *Journal of Speech and Hearing Research*, **22**, 354–365.

WRIGHT, J.T. (1975). Effects of vowel nasalization on the perception of vowel height. In C.A. Ferguson, L.M. Hyman, and J.J. Ohala (eds.), *Nasálfest: Papers from a Symposium on Nasals and Nasalization*. Stanford, CA: Language Universals Project, Stanford University.

ZERLING, J-P. (1992). Frontal lip shape for French and English vowels. *Journal of Phonetics*, **20**, 3–14.