American
Speech-Language-
Hearing
Association

Journal of Speech and Hearing Research, Volume 37, 1195–1203, October 1994

Effects of Phonetic Context on Audio-Visual Intelligibility of French

Christian Benoît
Tayeb Mohamadi
Sonia Kandel
Institut de la Communication Parlée
URA CNRS n° 368
INPG-ENSERG/Université Stendhal
Grenoble, France

Bimodal perception leads to better speech understanding than auditory perception alone. We evaluated the overall benefit of lip-reading on natural utterances of French produced by a single speaker. Eighteen French subjects with good audition and vision were administered a closed set identification test of VCVCV nonsense words consisting of three vowels [i, a, y] and six consonants [b, v, z, ʒ, ʀ, l]. Stimuli were presented under both auditory and audio-visual conditions with white noise added at various signal-to-noise ratios. Identification scores were higher in the bimodal condition than in the auditory-alone condition, especially in situations where acoustic information was reduced. The auditory and audio-visual intelligibility of the three vowels [i, a, y] averaged over the six consonantal contexts was evaluated as well. Two different hierarchies of intelligibility were found. Auditorily, [a] was most intelligible, followed by [i] and then by [y]; whereas visually [y] was most intelligible, followed by [a] and [i]. We also quantified the contextual effects of the three vowels on the auditory and audio-visual intelligibility of the consonants. Both the auditory and the audio-visual intelligibility of surrounding consonants was highest in the [a] context, followed by the [i] context and lastly the [y] context.

The information provided by the speaker's face increases message comprehension, particularly in a noisy background. In the past, several studies have aimed at quantifying the intelligibility gain that is due to visual information in situations where acoustic information was degraded (Binnie, Montgomery, & Jackson, 1974; Erber, 1969, 1975; Grant & Braida, 1991; MacLeod & Summerfield, 1987; Miller & Nicely, 1955; Neely, 1956; Sumby & Pollack, 1954; Summerfield, 1979). These studies show that facial information greatly increases the intelligibility of speech in noise, even for normal listeners with no special lip-reading training. Under degraded acoustic conditions, visual and auditory modalities complement each other in the perception of speech. What has been masked by noise in the speech spectrum can be partly recovered by the visual perception of the most salient aspects of the lips, teeth, and tongue shapes that determine the articulation place of several consonants (McGrath, Summerfield, & Brooke, 1984).

The positive influence of the visibility of the speaker's face on auditory perception was also observed when acoustic conditions are not degraded (Reisberg, McLean, & Goldfield, 1987). Furthermore, visual information may distort auditory perception if the acoustic and the optic sources of information are not coherent. In the well-known McGurk effect (McGurk & MacDonald, 1976), an acoustic /ba/ stimulus dubbed onto a visual /ga/ stimulus is perceived as /da/. Nevertheless, Sekiyama and Tohkura (1991) suggested that the influence of vision on audition may be subject to cross-linguistic variability. These authors reported that Japanese subjects were not as sensitive to the McGurk effect in clear acoustic conditions. They had to add noise to the acoustic stimuli in order to reproduce the McGurk effect with Japanese subjects. The latter results were reinterpreted by Massaro, Tsuzaki, Cohen, Gesi, and Heredia (1994), who demonstrated that there is a strong influence of visible speech in Japanese as well. They found similar results for Spanish and Dutch. More cross-linguistic comparisons are needed for a better understanding of how auditory and

0022-4685/94/3705-1195

visual information are integrated by subjects from different linguistic groups. As far as we know, all experiments on global audio-visual intelligibility of speech in noise have been conducted in English. This study in French aimed at providing auditory and audio-visual intelligibility scores as a function of acoustic degradation in French.

Benguerel and Pichora-Fuller (1982) compared the visual intelligibility of three English vowels and nine consonants in V1CV2 sequences uttered by a speaker who was particularly easy to lip-read. Their results showed that for normal-hearing and hearing-impaired subjects initial vowels were all easily lip-read, but that the final [u] was visually more intelligible than [i], which was in turn more intelligible than [æ]. They also showed that there were vocalic effects on the visual intelligibility of consonants. The averaged consonantal intelligibility in VCV stimuli was much higher in an [æ] or an [i] context than in an [u] context. In the current investigation, we compared vowel intelligibility of natural French for auditory and audio-visual presentations and quantified phonetic contextual effects of vowels on consonant intelligibility.

Benoît, Lallouache, Mohamadi, and Abry (1992) provided a descriptive analysis of the labial geometry of French, taking into account the coarticulatory effects of vowels on consonants and vice versa. They generated an extended corpus consisting of nine repetitions of the 14 French vowels and of various V1CV2CV1 productions. The films were analyzed using the special software/hardware developed by Lallouache (1991) according to relevant factors unveiled in studies on labiality (intero-labial area, height, width, the upper and lower lip protrusion, and chin lowering). Multidimensional analyses performed on measures on labiality have shown the wide extent to which vowels and consonants are subject to geometrical changes according to the context in which they are produced (Benoît, Boë, & Abry, 1991). This led to the definition of a set of French "visemes" that account for coarticulation in French (Benoît et al., 1992). The latter observed that at the production level, the labial shape of [y] is quite independent from the surrounding consonants and highly different from the shapes of [a] or [i], which not only vary according to the consonantal context in which they are inserted but also overlap. Furthermore, the shapes of six French consonants were very similar to each other in an [y] context, quite different in an [i] context, and in turn less different in an [a] context.

In agreement with perceptual data reported by Benguerel and Pichora-Fuller (1982) for English and based on a geometric analysis of labial shapes, Mohamadi (1993) suggested that in French [y] would be easier to identify visually than [a] and [i], and that consonants would be globally more intelligible in an [a] context than in an [i] context and, then, than in an [y] context. The study reported in this article aimed at providing perceptual confirmation of these predictions for natural French. We thus selected a portion of the corpus analyzed by Benoît et al. (1992) so that the most relevant coarticulated lip gestures could be perceptually tested. We considered only the three vowels [i, a, y] and the six consonants [b, v, z, ʒ, ʀ, l].

Hence, we quantified the contribution of visual information in bimodal speech perception as a function of various masking noise levels for French. Then, vowel intelligibility scores for auditory and audio-visual presentations were compared. Finally, the effects of vocalic context were calculated to determine a hierarchy of contextual effects in the auditory and audio-visual modalities.

Method

Corpus

From the reference corpus described by Benoît et al. (1992), we considered only the items of the form [VCVCVz] (e.g., [iviviz], [uʒuʒuz], [ababaz], etc.). The three chosen vowels [i, a, y] corresponded to the extreme positions of the labial movement in French vowels (Abry & Boë, 1986). The labial shape of the first four consonants [b, v, z, ʒ] has specific characteristics in French, whereas the last two consonants [ʀ, l] are apparently neutral and their influence regarding labiality is not well known. For acoustic reasons, the selected consonants were representative voiced phonemes of labial prototypes. By presenting three identical vowels and two identical consonants in a single stimulus, we expected phonetic redundancy to emphasize the effect of context on each unit. The corpus consisted of 18 different items, embedded in a carrier interrogative sentence: *C'est pas* "VCVCVz"? (i.e., *Is it* "VCVCVz"?).

Stimuli and Preparation of the Test

The audio-visual recording was made in a sound-treated room. The apparatus used for recording was the one developed by Lallouache (1991). Color video movies of a French speaker with high dynamics and symmetry in lip displacements were made using two cameras to obtain front and profile views. The speaker's face was presented in a black-and-white front view. Because the speaker wore goggles, "eye prosody" was not part of the visual information.

Speech material was stored in a video tape recorder (Sony BVU VP 9000P). The speech signal resided on one channel, whereas the second channel was used to transmit the masking noise. The masking noise was produced by a white noise generator (Nakamichi T-100, flat spectrum between 20 Hz and 20 kHz). After 8 kHz low-pass filtering, the noise was digitized (16 bits, Fs = 16 kHz) through a signal processing board (OROS AU20) on to a lab computer so that its duration exceeded the entire duration of each utterance. The D/A converter was programmed to vary by 6 dB steps, so that the signal-to-noise ratio (S/N) varied between −24 dB and 0 dB. Because the sentences were excised from a recording of about one thousand sentences, small differences in the overall intensity of the speaker's voice were observed. To eliminate these small natural fluctuations between two selected utterances, the speech level was measured during the realization of the phoneme [a] in the carrier sentence "*c'est pas*" and was used as the common reference for all stimuli. This adjustment never exceeded ±1 dB. Hence, the intrinsic intensity of the speech material was preserved without artificial adjustments. The noise and speech signals were mixed together on the audio output of the VTR.

Each stimulus, its presentation mode (audio alone [A] or audio-visual [AV]), the starting and ending frames of the whole carrier sentence, and the noise output attenuation were stored in an ASCII descriptor file that served as script for the experiment. Because of pseudo-random order of presentation, each subject was assigned an individual descriptor file. Finally, the whole process was fully automatic.

Subjects

Each subject received an audiological evaluation and a test of visual acuity before the study began. We selected 18 French subjects (11 females and 7 males) between the ages of 19 and 26 years (mean = 21.5 years). None of them had any particular background in speech sciences or familiarity with people with hearing impairment. One potential subject was rejected because of hearing loss. Subjects were paid for participation.

Procedure

The test was divided into two subtests: the audio only (A) and audio-visual (AV) presentation modes. The presentation order of the two subtests was counterbalanced across subjects. Subjects were tested individually in a sound-treated chamber. A video monitor (Sony-Trinitron PVM 1442 QM) and a loudspeaker (CHORALE III SP 3021) were situated 1.5 m in front of the table at which the subject was seated. Subjects were first administered instructions concerning the experiment. Before each subtest, they were trained with a five-stimuli presentation and then given five answer sheets, each with 18 identical lines of the six consonants and three vowels. The subjects were asked to circle the perceived consonant and vowel for each of the 18 stimuli. They were asked to guess as much as possible in cases of high degradation. If no speech was perceived because of the masking noise, they could cross out the whole line corresponding to that stimulus. A recorded voice indicated that the answer sheet should be changed after every 18 stimuli. Each stimulus was preceded by a beep and followed by a 2-sec silence. A response time of 15 sec allowed subjects to fill in the answer after each stimulus. This brief time also permitted an automatic search for the first frame of the next stimulus on the magnetic tape. The order of presentation of the stimuli differed from one subject to another. It was pseudo-randomized by avoiding repetition of the degradation conditions and of the stimuli. Each session lasted 24 min.

Results

We first analyzed the global intelligibility of speech in noise in both audio only (A) and audio-visual (AV) conditions. Then, the mean intelligibility of the three tested French vowels was quantified. Finally, we compared the contextual effects of each vowel on the mean auditory and visual intelligibility of the six consonants.

Global Auditory and Audio-Visual Intelligibility

We quantified the global intelligibility scores of all items in the two presentation modes at the various masking noise levels. Responses were considered correct only if both the consonant and the vowel were correctly identified. The percentages of correct responses averaged across the 18 subjects and their standard deviations are shown in Figure 1. An Analysis of Variance (ANOVA) with presentation mode (A and AV) and S/N level (−24, −18, −12, −6, and 0 dB) as within-subjects factors was carried out. A presentation mode main effect indicated that differences between A and AV scores were globally significant [*F* (1,17) = 707.37; *p* < .0001]. S/N level was significant as well [*F* (4,17) = 641.22; *p* < .0001]. A significant interaction between the two factors [*F* (4,68) = 145.45; *p* < .0001] was observed. Differences between A and AV conditions were significant at all S/N levels: *F* (1,17) = 415.06, *p* < .0001 at −24 dB; *F* (1,17) = 641.78, *p* < .0001 at −18 dB; *F* (1, 17) = 229.21, *p* < .0001 at −12 dB; *F* (1,17) = 164.36, *p* < .0001 at −6 dB; *F* (1,17) = 28.71, *p* < .0001 at 0 dB.

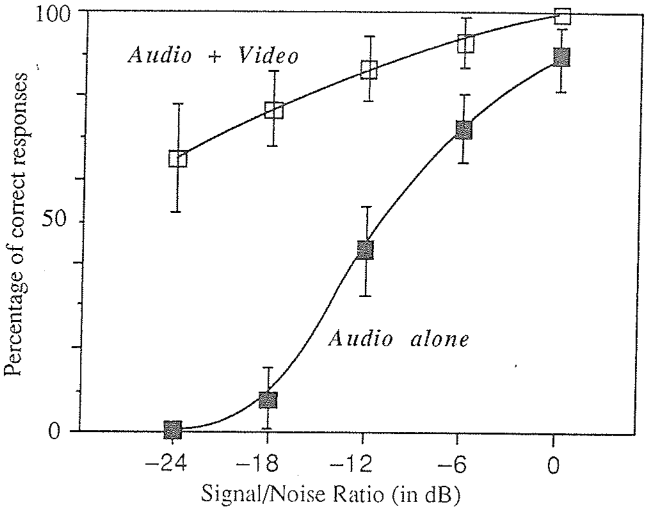

FIGURE 1. Global A and AV intelligibility scores for 18 nonsense words by 18 subjects as a function of the S/N ratio.

Under the A condition the mean identification score decreased from 72% to 8% within a 12 dB interval (between S/N = −6 dB and S/N = −18 dB). In the audio-visual presentation, and within the same range of acoustic interference, the intelligibility score decreased from 93% to 77%. At −24 dB, the A score approached zero, whereas in AV, it was still 65%. Subjects reported that they could not detect the speech signal at −24 dB. The AV condition at S/N = −24 dB appears to be comparable to a purely visual condition. The intelligibility score obtained under this *visual-only* condition may thus be considered as a measure of lip-reading performance and will hereafter be referred to as the V condition.

Intelligibility of Vowels [a, i, y]

We quantified the intelligibility of the three tested vowels [i, a, y] in the auditory and audio-visual modes of presentation. Intelligibility was defined as the percentage of correctly identified vowels, averaged over the six consonantal contexts.

Auditory (A). An ANOVA for the A condition, with S/N level (−24, −18, −12, −6, and 0 dB) and vowel [i, a, y] as within-subject factors was carried out. A main S/N level effect was observed [F (4,17) = 529.35, $p < .0001$]. A main vowel effect [F (2,34) = 20.92, $p < .0001$] was observed as well, indicating that [i], [a], and [y] globally differ in intelligibility (I) according to a I_A[a] > I_A[i] > I_A[y] hierarchy.[1] The interaction between the two factors was significant [F (8,34) = 10.05, $p < .0001$]. Vowel intelligibility was extremely high at 0 and −6 dB (greater than 97.2%), but no differences between [i], [a], and [y] were observed [F (2,34) = 1, $p = .378$ at 0 dB, and F (2,34) = 1.55, $p = .228$ at −6 dB]. In contrast, at −12 dB differences in vowel intelligibility were significant [F (2, 34) = 32.4, $p < .0001$]. Indeed, at −12 dB I_A[a] (99.1%) was higher than I_A[i] (84.3%) (Newman-Keuls $p < .01$), and was in turn higher than I_A[y] (62.0%) (Newman-Keuls $p < .01$). At −18 dB a simple effect was observed [F (2,34) = 7.36, $p = .002$] as well: I_A[a] (40.7%) remained higher than I_A[i] (19.4%) and I_A[y] (14.8%), but the difference between I_A[i] and I_A[y] was not statistically significant (Newman-Keuls $p > .01$). At −24 dB the three vowel intelligibility scores were extremely low (I_A[a] = 5.6%; I_A[i] = I_A[y] = 1.9%) and did not yield significant differences [F (2,34) = 1.36, $p = .27$]. Since differences in vowel intelligibility were most apparent at −12 dB, when referring to auditory intelligibility, we will consider only the −12 dB S/N level.

Audio-visual (AV). An ANOVA for the AV condition, with S/N level (−24, −18, −12, −6, and 0 dB) and vowels [i, a, y] as within-subject factors was carried out. A main S/N level effect was observed [F (4,17) = 34.28, $p < .0001$]. Analysis revealed a main vowel effect [F (2,34) = 3.17, $p = .05$] as well as a significant interaction between the two factors [F (8,34) = 3.18, $p = .0025$]. Ceiling effects occurred at S/N ≥ −18 dB [F (2,34) = .106, $p = .9$], above which all vowel intelligibility scores were greater than 96% [F (2,34) = 1.55, $p = .228$ at −12 dB; F(2,34) = .000, $p = 1$ at −6 dB; F(2,34) = .000, $p = 1$ at −0 dB]. The vowel main effect is thus due to differences in vowel intelligibility observed at S/N −24 dB [F(2,34) = 3.61, $p = .038$], that is, in purely visual conditions. At −24 dB, the hierarchy I_v[y] = 94.4% > I_v[a] = 85.2% > I_v[i] = 76.9% was observed. Pairwise comparisons indicate that all differences were significant: Newman-Keuls $p < .01$.

***Auditory versus visual intelligibility of** [a, i, y].* Table 1 shows confusion matrices of the three vowels [a, i, y], irrespective of errors on consonant identification, for 18 subjects in both the auditory and the visual modes. Table 1 shows results observed in A at S/N = −12 dB and results observed in the V condition (i.e., the AV condition at −24 dB).

In the A condition, there was no confusion between [a] and the other two vowels, whereas some confusion arose between [i] and [y]. Vowel [i] was apparently "response attractive" (125 [i] percepts vs. 108 [i] stimuli, out of a total of 324 stimuli) in comparison to [a] (107 percepts) and [y] (only 79 percepts). Conversely, in the visual condition, intelligibility for vowel [y] was the highest and was hardly ever confused with the other two vowels. Vowels [i] and [a] were subject to some confusion between each other, and [i] was the least attractive response. Therefore, auditory confusions appeared between [i] and [y], whereas visual confusions arose between [a] and [i].

TABLE 1. Confusion matrices of vowels [a, i, y], **averaged over six consonantal contexts, for 18 subjects, for 108 stimuli. Question mark stands for nonresponse.**

	Responses			
Stimuli	a	i	y	?
A −12 dB				
a	107	–	–	1
i	–	91	11	6
y	–	34	68	6
Total	107	125	79	13
AV −24 dB				
a	92	12	1	3
i	15	83	–	10
y	1	–	103	4
Total	108	95	104	17

In order to display the intelligibility scores of the three tested French vowels [a, i, y] in an audio-visual space, Figure 2 plots the percentages of correctly identified vowels (irrespective of errors on consonants) in a double-axis presentation. On the X-axis are plotted the mean intelligibility scores in A at −12 dB. On the Y-axis are plotted the mean intelligibility scores in AV at −24 dB.

In summary, the obtained auditory intelligibility hierarchy for the three French vowels in six consonantal contexts was I_A[a] > I_A[i] > I_A[y]. When the auditory information was absent (i.e., at S/N = −24 dB), the intelligibility hierarchy was I_v[y] > I_v[a] > I_v[i]. This reinforces the idea that vision and audition complement each other, at least in the discrimination of these three extreme vowels.

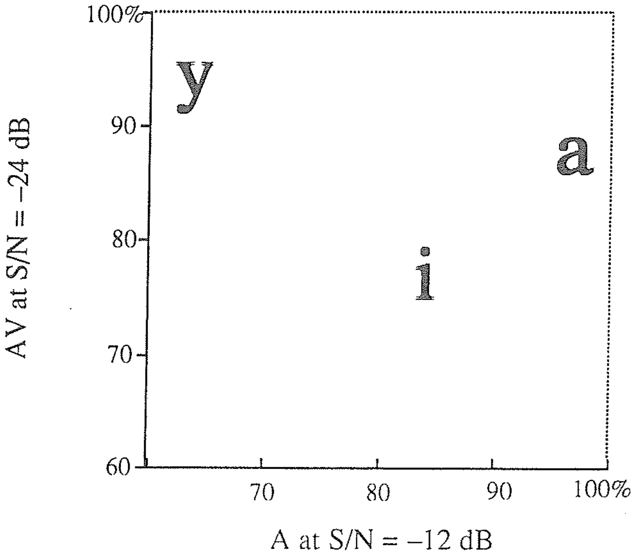

FIGURE 2. Percentages of correctly identified vowels (irrespective of errors on consonants). On the X-axis are plotted the averaged results in A for −12 dB (out of 108 responses). On the Y-axis are plotted the results in the V condition (out of 108 responses).

[1] We will use hereafter the notation I_A[X] and I_v[X] for the intelligibility of vowel X in the auditory (A) and the visual mode (v).

Effect of Vocalic Context on Consonant Intelligibility

We examined how the phonetic context naturally contributed to auditory and visual phoneme identification. The perceptual influence of the three tested vowels were globally evaluated on the averaged intelligibility of our set of six consonants.

Effect of [a], [i], and [y] on the auditory intelligibility of consonants. An ANOVA for the A condition with S/N levels (−24, −18, −12, −6, and 0 dB), consonants [b, v, z, ʒ, ʀ, l], and vocalic context [i, a, y] as within-subject factors was carried out. The dependent measure concerned correct consonant identification. The three factors yielded significant main effects: S/N level [F (4,17) = 476.58, p < .0001], consonant [F (5,68) = 8.05, p < .0001], and vocalic context [F (2,34) = 117.12, p < .0001]. Consonant intelligibility was affected by vocalic context at almost all S/N levels except at −24 dB: F (2,34) = .000, p = 1 at −24 dB; F (2,34) = 10.52, p < .0001 at −18 dB; F (2,34) = 52.73, p < .0001 at −12 dB; F (2,34) = 51.92, p < .0001 at −6 dB; F (2,34) = 8.93, p = .001 at 0 dB. The intelligibility of the six consonants was extremely poor at −24 dB (less than 1%). In all other S/N conditions, the six consonants were more intelligible in the [a] context than in the [i] context and, then, than in the [y] context. Results yielded a C_A[a] > C_A[i] > C_A[y] hierarchy.[2] As for vowel intelligibility, the −12 dB S/N condition was the one that best emphasized differences between contextual effects, avoiding any ceiling or floor effect: C_A[a] = 81.5% > C_A[i] = 33.3% > C_A[y] = 26.9% (Newman-Keuls p < .01 for pairwise comparisons between the three values).

Effect of [a], [i], and [y] on the audio-visual intelligibility of consonants. An ANOVA for the AV condition with S/N level (−24, −18, −12, −6, and 0 dB), consonants [b, v, z, ʒ, ʀ, l], and vocalic context [i, a, y] as within-subject factors was carried out. The dependent measure concerned correct consonant identification. The three factors yielded significant main effects: S/N level [F (4,17) = 54.85, p < .0001], consonant [F (5,68) = 24.41, p < .0001], and vocalic context [F (2,34) = 139.35, p < .0001]. Vocalic context effects were observed at most S/N levels except at 0 dB: F (2,34) = 35.0, p < .0001 at −24 dB; F (2,34) = 44.33, p < .0001 at −18 dB; F (2,34) = 42.65, p < .0001 at −12 dB; F (2,34) = 11.14, p < .0001 at −6 dB; F (2,34) = 2.13, p = .135 at 0 dB. Vocalic contextual effects followed a C_V[a] > C_V[i] > C_V[y]. Contextual effects on consonant intelligibility were best emphasized in the V condition [F (2,34) = 35.0, p < .001]: C_V[a] = 89.8% > C_V[i] = 75.9% > C_V[y] = 48.2% (Newman-Keuls p < .01 for pairwise comparisons between the three values).

Auditory vs. visual effects of [a], [i], and [y] on the intelligibility of consonants. The effects of the three vowels [a], [i], and [y] on consonant intelligibility in an audio-visual space are shown in Figure 3. The percentages of correctly identified consonants [b, v, z, ʒ, ʀ, l] (irrespective of errors on vowels) in the three vocalic contexts are shown in a double-axis presentation. On the X-axis are plotted the averaged intelligibility scores in A at −12 dB. On the Y-axis are plotted the averaged intelligibility scores in the visual condition (AV at −24 dB).

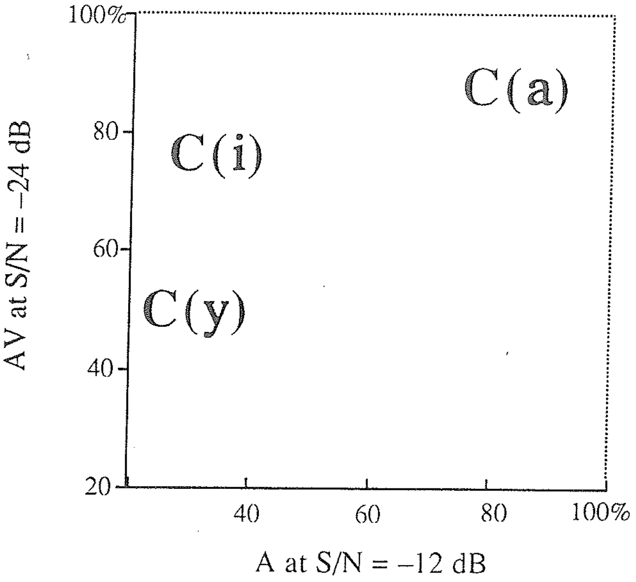

FIGURE 3. Effects of vowels [a], [i], and [y] on consonant intelligibility, in an audio-visual space: mean percentages of correctly identified consonants [b, v, z, ʒ, ʀ, l] (irrespective of errors on vowels) in the three vocalic contexts. On the X-axis are plotted the averaged auditory intelligibility scores (A at −12 dB). On the Y-axis are plotted the averaged visual intelligibility scores (AV at −24 dB).

Discussion

Global results indicated that the visibility of the speaker's face enhanced speech perception, especially in situations where acoustic information was reduced. Our results for French are thus similar to those of Sumby and Pollack (1954) and Erber (1969) for English. Data analysis yielded a I_A[a] > I_A[i] > I_A[y] auditory-alone intelligibility hierarchy. When speech was presented in the AV mode at −24 dB S/N, the resulting hierarchy was I_V[y] > I_V[a] > I_V[i]. This reinforces the idea that vision and audition complement each other, at least in the discrimination of these three extreme vowels. Contextual vocalic effects on consonant intelligibility were analyzed as well, revealing a C[a] > C[i] > C[y] hierarchy for both the audio-alone and visual-alone conditions.

These perceptual hierarchies seem to be due to articulatory constraints at the production level. In the audio-alone condition, the I_A[a] > I_A[i] > I_A[y] hierarchy is consistent with previously reported data on the *intrinsic intensity* of the three tested vowels. Indeed, Rossi (1971) reported that the specific intensity of [a] is 4 dB higher than that of [i], which is 2 dB higher than the one for [y]. Guérin and Boë (1978) observed a similar difference between [a] and [i], with the mean intensity of [y] comparable to that of [i]. These results were based on sophisticated acoustic analyses of vowels uttered in isolation. Hence, they do not provide a complete psychophysical explanation of differences in intelligibility observed

[2] We will use hereafter the notation C_A[X] and C_V[X] for the average intelligibility of consonants under the contextual effect of vowel X in the auditory (A) and the visual (v) mode.

with vowels repeated in a consonantal context. Intrinsic duration and formant structure should probably be considered, as well as the robustness of these features in a consonantal environment, but these factors cannot by themselves explain why the auditory *vocalic intelligibility* was somewhat transmitted to surrounding consonants. Therefore, the differences in intelligibility of vowels and consonants in vocalic contexts could be partly explained by differences in the intrinsic intensity of the corresponding phonemes uttered in isolation.

In the visual-alone condition, results yielded a I_V[y] > I_V[a] > I_V[i] intelligibility hierarchy. These results are in agreement with the predictions made by Mohamadi (1993) on the basis of a strictly geometrical analysis. The geometrical constraints for the production of the three tested vowels, the specific role of labial gestures, as well as differential resistivities to modifications in consonantal environments at the articulatory level, may explain the superiority of the vowel [y] in the visual domain. Lip gesture is highly constrained in the production of [y] because the lip internal area cannot exceed 100 mm^2 (Abry & Boë, 1986). We may thus assume that subjects exploit this highly visible articulatory constraint in the visual identification of the vowel [y]. Conversely, the labial shape of vowel [i] is highly dependent on the consonantal context in which it is produced. Therefore, it is not surprising that subjects have difficulties in identifying [i] when it is uttered in a consonantal context. We may consider [a] as a vowel with an intermediate articulatory behavior between [i] and [y] as far as visual modifications of its lip/jaw shapes are concerned.

To our knowledge, the only studies of visual identification of French vowels were done by Mourand-Dornier (1980) with normal hearing subjects and by Gentil (1981) with hearing-impaired subjects. The two studies tested 13 French vowels. Globally, within the full set of French vowels the I_V[i] > I_V[a] > I_V[y] intelligibility hierarchy was observed. This may be due to the high degree of confusion between [i] and [e], [a] and [ɛ] (or [ɑ] only included in Gentil's test), or [u] and [y], among others. Despite the discrepancies between the confusion matrices provided by the two authors (partly due to differences in the type of stimuli, the populations, and the vowels they tested) and the fact that [i], [a], and [y] are not the easiest vowels to discriminate within a set of 13, a multidimensional analysis (INDSCAL) performed on both results revealed that [i], [a], and [y] were systematically situated at the extremes of a roughly triangular projection on the first two factors (Tseva, 1989).

Great care must be taken when comparing the visual hierarchy we observed with previous ones. First, the intelligibility of a vowel largely depends on the number of tested vowels. In French, confusions arise between visual realizations of phonemes that *look alike* on the lips, such as [u] and [y]. Second, most of the perceptual studies on the visual intelligibility of vowels were done in English, for which lip rounding and vowel (back) position are redundant. Although comparing our results for French with the ones observed in experiments in English would be rather hazardous, it must be pointed out that the most protruded vowels ([y] or [u] in French, [u] in English) are undoubtedly the easiest to identify visually when they are tested against spread vowels such as [a] or [i]. The most salient results from earlier measurements of the intelligibility of vowels uttered in a consonantal context are summarized with our own results in Table 2.

The data reported in this paper indicate that contextual effects in the audio-alone condition follow the hierarchy C_A[a] > C_A[i] > C_A[y]. Barth and Chulliat (1980) compared the contextual effects of the three vowels [a, i, u] on the identification scores of six French fricatives [f, v, s, z, ʃ, ʒ] auditorily presented to moderately, severely, and profoundly hearing-impaired subjects. They observed a C_A[a] > C_A[i] > C_A[u] hierarchy for the three groups. Although this experiment was different from ours on several aspects, the results are globally in agreement with the ones obtained in the present study.

In the visual condition the same hierarchy as in the audio-alone condition was observed: C_V[a] > C_V[i] > C_V[y]. Once again, our perceptual results were in agreement with the predictions of Mohamadi (1993) based on a geometrical analysis of similar stimuli. The contextual vocalic effect on the intelligibility of surrounding consonants depends on the way the labial gesture needed to produce the vowel can be combined with the labial movement needed for the production of the consonant. Whereas the hierarchy for vocalic intelligibility was I_V[y] > I_V[a] > I_V[i] and was established according to the most salient labial shapes, the intelligibility of consonants was enhanced when surrounded by vowels whose labial shape was less salient. [y] was the most easily identified vowel but, because of this, it was the vocalic context that most distorted the intelligibility of surrounded consonants. As stated above, the production of [y] is subjected to articulatory constraints that restrict the vocal tract output area. In addition, lip protrusion is largely anticipated across most of the preceding consonants in order for the lips to accurately reach the specific pattern of [y] in time (Abry & Lallouache, 1991; Benguerel & Cowan, 1974).

Earlier studies have shown that the degree of alteration of a consonant's labial shape by the vocalic context determines its visual intelligibility score (Benguerel & Pichora-Fuller, 1982; Erber, 1971; Owens and Blazek, 1985; Massaro, Cohen, & Gesi, 1993 for English; Barth & Chulliat, 1980 for French). Table 3 matches our data with results obtained in these studies (many of the reported scores have been recalculated on the basis of figures and/or tables). In both languages and for all the tested consonants, spread vowels enhance the intelligibility of surrounded consonants, whereas rounded vowels decrease the intelligibility of surrounded consonants. Despite the differences in the phonological distribution and phonotactical combinations of vowels and consonants between English and French, all the studies on the influence of the vocalic context on consonant visual intelligibility are systematically consistent with the following hierarchy: C_V[a, æ, or ɑ] > C_V[i] > C_V[u or y].

Finally, the fact that in natural French vocalic context affects the intelligibility of surrounding consonants in the same manner (C[a] > C[i] > C[y]) for both modalities is rather striking, given the complementarity between A and V observed for vowel intelligibility.

Conclusion

The complementarity between auditory and visual information provided by the vocal tract gestures has been widely

TABLE 2. Percentage of correct lip-read vowels as reported in the literature (i.e., vowel intelligibility irrespective of the consonantal context). Some data have been obtained from graphic representation or by calculating partial—or averaged—results. (NH = normal-hearing subjects and HI = hearing-impaired subjects)

Authors	Language	Number of vowels	Number of subjects	Stimulus and consonantal context	Vowel intelligibility in % (desc. order)
Erber (1971)	American English	10	6 HI	CVC C = [b]	u 91.0 ɝ 89.0 i 85.0 ɑ 85.0 ɔ 83.0 æ 63.0
Wozniak & Jackson (1979)	American English	16	10 NH	C_1VC_2 C_1 = [h] C2 = [g]	u 80.0 i 78.0 æ 51.0 ɑ 50.0 ɪ 36.0 ʌ 18.0
Mourand-Dornier (1980)	French	13	30 NH	C_1VC_2 words	a 58.7 ɛ 52.5 i 51.7 u 49.3 e 46.7 y 19.1
Gentil (1981)	French	13	51 HI	C_1V or C_2VC_1 existing words C_1 = [l]	i 57.9 a 41.3 u 38.8 ɛ 31.3 y 26.3 ɑ 18.8 e 15.0
Benguerel & Pichora-Fuller (1982)	Canadian English	3	5 NH 5 HI (mixed)	VCv C ɛ [p t k tʃ f θ s ʃ w]	u 99.5 i 89.6 æ 77.7
Montgomery et al. (1987)	American English (2 talkers)	5	30 HI	C_1VC_1 and C_2VC_3 C_1 ɛ [p b f v t d ʃ] C_2 ɛ [h w r] C_3 = [g]	ɑ 56.7 i 53.8 u 49.8 ɪ 41.5 ʌ 29.8
Benoît et al. (1994) (this paper)	French	3	18 NH	$VC_1VC_1VC_2$ C_1 ɛ [b v z ʒ ʀ l] C_2 = [z]	y 94.4 a 85.2 i 76.9

emphasized in the literature. This phenomenon is observed in our experiment by the modality-dependent hierarchies of vowel intelligibility I_A[a] > I_A[y] versus I_V[y] > I_V[a]. However, we see from the various hierarchies reported in the literature that such a ranking is highly dependent on the size of the set of tested vowels, on their position within the stimulus, and on the consonantal environment. Because of this, it is practically impossible to correlate any of the reported vowel intelligibility hierarchies with hierarchies of vowel distribution across vocalic systems, or with hierarchies of vocalic frequency of occurrence in a given language. However, it is implicitly considered that phonological systems have emerged only on an auditory-based communication between humans of the same linguistic community. Prediction models of vocalic systems (Liljencrants & Lindblom, 1972; Lindblom, 1986; Schwartz, Boë, Perrier, Guérin, & Escudier, 1989; Vallée, Boë, & Schwartz, 1991), for instance, typically rely on acoustic or articulatory-acoustic constraints and integrate acoustic principles such as the Dispersion Theory (Liljencrants & Lindblom, 1972), Quantal Theory (Stevens, 1989), and the perceptual formant hypothesis (Bladon & Fant, 1978; Carlson, Fant, & Grandstrom, 1970). None of the latter take directly into consideration any principle of visual discrimination.

Some support for the importance of vision in speech communication may come from the Motor Theory of Speech Perception (Liberman & Mattingly, 1985). Once it is assumed that listeners/viewers are not primarily interested in the

TABLE 3. Global lip-read consonant intelligibility as a function of vocalic context, as reported in the literature. (NH = Normal-hearing subjects and HI = hearing-impaired subjects)

Authors	Language	Number of vowels	Number of subjects	Tested consonant	Consonant intelligibility in % (desc. order)
Erber (1971)	American English	3	6 HI	b d g h l r ð v s z	a 89.5 i 71.9 u 71.2
Barth & Chulliat (1980)	French	3	30 HI	f s ʃ v z ʒ	a 65.3 i 63.7 u 60.3
Benguerel & Pichora-Fuller (1982)	Canadian English	3	5 NH 5 HI (mixed)	p t k ʃ f θ s tʃ w	æ 78.0 i 77.0 u 58.0
Owens & Blazek (1985)	American English	4	5 NH 5 HI (mixed)	23 consonants	ɑ 43.0 ʌ 39.5 i 32.5 u 21.5
Massaro et al. (1993)	American English	3	6 NH	22 consonants	a 34.4 i 30.5 u 29.2
Benoît et al. (1994) (this paper)	French	3	18 NH	b v z ʒ ʀ l	a 89.8 i 75.9 y 48.1

patterns of sound that talkers produce, but rather in the articulatory gestures that generate the sounds, it may follow that articulatory gestures may be perceived through the ears and/or eyes. Summerfield (1991) suggested that the evolutionary pressure on humans to develop refined hearing has been stronger than to develop lip-reading. This is not a reason, however, to neglect the importance of the visual perception of articulatory gestures in the explanation of any phonological system. Vowel intelligibility yields different hierarchies in the two modalities (I_A[a] > I_A[y] versus I_v[y] > I_v[a]). In contrast, vocalic context affects the intelligibility of surrounding consonants in the same manner (C[a] > C[i] > C[y]) in both the auditory and visual modalities. Because consonants are more frequent than vowels in French and they cannot be uttered without a vocalic context, the evolutionary pressure on humans to develop vocalic systems could have somehow exploited the synergy of audition and vision in speech perception.

Acknowledgments

We gratefully acknowledge Christian Abry, Marie-Agnès Cathiard, Tahar Lallouache, Tom Sawallis, Shelley Peery for friendly, technical, scientific, and/or linguistic support, and Rebecca Eilers and an anonymous reviewer for enlightening comments on an earlier version of this manuscript and for the many improvements they kindly brought to our article throughout the reviewing process.

References

Abry, C., & Boë, L.-J. (1986). Laws for lips. *Speech Communication, 5,* 97–104.

Abry, C., & Lallouache, M.-T. (1991). Audibility and stability of articulatory movements. Deciphering two experiments on anticipatory rounding in French. *Proceedings of the 12th International Congress of Phonetic Sciences, 1,* 220–225. Aix-en-Provence, France.

Barth, S., & Chulliat, R. (1980). Perception auditive des fricatives par les déficients auditifs. *Actes des 11èmes Journées d'Etude sur la Parole,* Groupe Communication Parlée de la Société Française d'Acoustique, Strasbourg, France, 18–24.

Benguerel, A. P., & Cowan, H. A. (1974). Coarticulation of upper lip protrusion in French. *Phonetica, 30,* 41–55.

Benguerel, A.-P., & Pichora-Fuller, M. K. (1982). Coarticulation effects in lipreading. *Journal of Speech and Hearing Research, 25,* 600–607.

Benoît, C., Boë, L.-J., & Abry, C. (1991). The effect of context on labiality in French. *Proceedings of the 2nd Eurospeech Conference,* Vol. 1, 153–156, Genoa, Italy.

Benoît, C., Lallouache, T., Mohamadi, T., & Abry, C. (1992). A set of French visemes for visual speech synthesis. In G. Bailly & C. Benoit (Eds.), *Talking machines: Theories, models and applications* (pp. 485–504). North Holland: Elsevier Science Publishers.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research, 17,* 619–630.

Bladon, R. A. W., & Fant, G. (1978). A two-formant model and the cardinal vowels, *STL-QPSR 1,* 1–8.

Carlson, R., Fant, G., & Granström, B. (1970). Some studies concerning perception of isolated vowels. *STL-QPSR 2–3,* 19–35.

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

Erber, N. P. (1971). Effects of distance on the visual reception of speech. *Journal of Speech and Hearing Research, 14,* 848–857.

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders, 40,* 481–492.

Gentil, M. (1981). *Etude de la perception de la parole: Lecture labiale et sosies labiaux.* IBM France.

Grant, K. W., & Braida, L. D. (1991). Evaluating the articulation index for auditory-visual input. *Journal of the Acoustical Society of America, 89,* 2952–2960.

Guérin, B., & Boë, L. J. (1978). Étude d'un indice acoustique des voyelles: La puissance intrinsèque. *Actes des 9èmes Journées d'Étude sur la Parole, 1,* 167–176. Lannion: Société Française d'Acoustique.

Lallouache, T. (1991). *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres,* Doctoral thesis, Institut National Polytechnique, Grenoble, France.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perception contrast. *Language, 48,* 839–862.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. J. Ohala (Ed.), *Experimental phonology.* (pp. 13–44). Orlando, FL: Academic Press.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21,* 131–141.

Massaro, D. W., Cohen, M. M., & Gesi, A. (1993). Long-term training, transfer, and retention in learning to lip-read. *Perception & Psychophysics, 53,* 549–562.

Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., & Heredia, R. (1994). Bimodal speech perception: An examination across languages. *Journal of Phonetics, 21,* 445–478.

McGrath, M., Summerfield, Q., & Brooke, M. (1984). Roles of lips and teeth in lipreading vowels. *Proceedings of the Institute of Acoustics, 6,* 401–408.

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

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

Mohamadi, T. (1993). *Synthèse à partir du texte de visages parlants: Réalisation d'un prototype et mesures d'intelligibilité bimodale.* Doctoral thesis, Institut National Polytechnique, Grenoble, France.

Montgomery, A. A., Walden, B. E., & Prosek, R. A. (1987). Effects of consonantal context on vowel lip reading. *Journal of Speech and Hearing Research, 30,* 50–59.

Mourand-Dornier, L. (1980). *Le rôle de la lecture labiale dans la reconnaissance de la parole.* Doctoral thesis, Faculty of Medecine, Besançon.

Neely, K. K. (1956). Effect of visual factors on the intelligibility of speech. *Journal of the Acoustical Society of America, 28,* 1275–1277.

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

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–114). Hillsdale NJ: Lawrence Erlbaum Associates.

Rossi, M. (1971). L'intensité spécifique des voyelles. *Phonetica, 24,* 129–161.

Schwartz, J.-L., Boë, L.-J., Perrier, P., Guérin, B., & Escudier, P. (1989). Perceptual contrast and stability in vowel systems: A 3D simulation. *Proceedings of the 1st Eurospeech Conference,* Paris, 1/2, 63–66.

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

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3–45.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36,* 314–331.

Summerfield, Q. (1991). Visual perception of phonetic gestures. In I.G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 117–137). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tseva, A. (1989). L'arrondissement dans l'identification visuelle des voyelles du français. *Bulletin du Laboratoire de la Communication Parlée, 3,* 149–186.

Vallée, N., Boë, L. J., & Schwartz, J. L. (1991). Tendances universelles et stabilité des systèmes vocaliques. *Proceedings of the 12th International Congress of Phonetic Sciences, 3,* 142–145. Aix-en-Provence, France.

Wosniak, V. D., & Jackson, P. L. (1979). Visual vowel and diphthong perception from two horizontal viewing angles. *Journal of Speech and Hearing Research, 22,* 355–365.

Received August 16, 1993
Accepted May 3, 1994

Contact author: Christian Benoît, PhD, Institut de la Communication Parlée, URA CNRS n° 368, INPG-ENSERG/Université Stendhal, BP 25X-38040. Grenoble Cedex 9, France.