Published in Talking Machines: Theories, Models & Designs
G. Bailly & C. Benoît, Eds., Elsevier Science Publishers, North-Holland

A set of French visemes for visual speech synthesis

Christian Benoît, Tahar Lallouache, Tayeb Mohamadi, and Christian Abry

Institut de la Communication Parlée, INPG/ENSERG - Université Stendhal, BP 25X, 38040 Grenoble, France

Abstract

An extensive audio-visual corpus of transitions between French phonemes was recorded (signal, plus front and profile views) from a speaker with high dynamics and symmetry in lip displacements. Nine repetitions of 14 vowels and 6 consonants in various combinations were analysed, and 3300 video fields were labeled on the centers of the acoustic wave-forms. The vermilion zone of the speaker's lips was carefully made up in an intense blue, and a Chroma-Key transformed the lips area to a saturated black color in real-time. Each selected video field was digitized on PC-based visual speech workstation. The internal and external contours of the vermilion zone were automatically extracted and stored. A program measured various areas from these patterns (upper, lower, and inter-lip areas) and various distances between front and profile reference lines and anatomical points. From 786 selected video fields, 17 parameters were ultimately measured and computed. Clustering and discriminant analysis allowed us to isolate 21 lip-jaw shapes that constitute the "labial space" of our French speaker. These 21 visemes are here discussed and described in the perspective of their integration into visual speech synthesisers, following a general introduction to the visual intelligibility of speech.

1. INTRODUCTION

"Speech is rather a set of movements made audible than a set of sounds produced by movements." (Stetson, 1928)

1.1. Visual intelligibility of speech

1.1.1. Hearing impaired subjects

Most of the psychological studies on the visual perception of speech were dedicated to the hearing impaired. The early investigations went from l'abbé de l'Epée (18th century) to Kitson (1915). Lip-reading is necessary for hearing impaired persons in order to (partially) understand speech, specifically, by using what is recoverable from visual speech. For a tutorial review of this problem, see Risberg (1982).

1.1.2. Normal-hearing subjects

In 1935, Cotton stated that “there is an important element of visual hearing in all normal individuals.” Even if the auditory modality is most important for speech perception by normal hearers, the visual modality may allow subjects to understand some speech. One must notice that visual information, provided by movements of the lips, the jaw, the teeth, etc., cannot, in itself, provide normal speech intelligibility (Dodd, 1977). It has been demonstrated however that vision may almost entirely replace spectral information if the laryngeal timing is provided (Rosen et al., 1979). Even tactile speech communication, such as the Tadoma method, where a hand is placed on the face of the talker, allows deaf-blind people to achieve almost normal communication (Norton et al., 1977). Vision of a talker’s face may also enhance spectral information that is distorted by added noise.

The phonetic interpretation of speech intelligibility improvement by visual information has been studied by Summerfield (1979). Several studies have also been devoted to factors influencing the comprehension of visual speech: physical factors, such as the optimal angle at which to view the face, the distance between speaker and listener, the intensity of the light, etc.; human factors, such as the familiarity with the speaker, the speech rate, the listener’s auditory and visual acuity, etc.; and psychological factors, such as motivation, early learning of lipreading, etc. (For a recent tutorial on face processing, see Young & Ellis, 1989, or Cathiard, 1988/1989, in French.)

1.1.3. Bimodal perception of speech in noise

A number of investigators have studied the effect of distortion by noise on speech intelligibility according to whether the message is heard only, or heard with the speaker’s face also provided (Sumby and Pollack, 1954; Neely, 1956; Binnie et al., 1974; and Erber, 1969 and 1975, among others.). Mohamadi and Benoît (to appear) recently adapted the Erber’s experiment to French: Figure 1 shows the articulation score on 18 nonsense words by 18 normal hearers in two test conditions, audition only and audition plus vision.

We observe that vision does not add much in optimal acoustic conditions (S/N >> 0 dB), whereas seeing the speaker’s face allows the listener to understand around two third of the 18 items under highly degraded acoustic conditions (S/N < -20 dB). We can also note that the standard deviation of errors is greater under visual presentation only as there are noticeable individual differences in the processing of visual information.

The same effect was simulated by Yuhas et al. (1989) using a feedforward neural network model to associate mouth shapes and short time spectral amplitude envelopes of the corresponding acoustic signal spectra. This experiment demonstrated that vowel recognition in a noisy acoustic environment is significantly enhanced by by the addition of visual speech signals.

A mean 11 dB “benefit of lip-reading” has been found by Summerfield et al. (1989). This corresponds to the average difference between the lowest S/N ratios at which test sentences are understood given the presence or absence of visual information.

Visual cues are obviously used by listeners to improve speech intelligibility in highly degraded conditions. Pols (1979) however reported that listeners who could

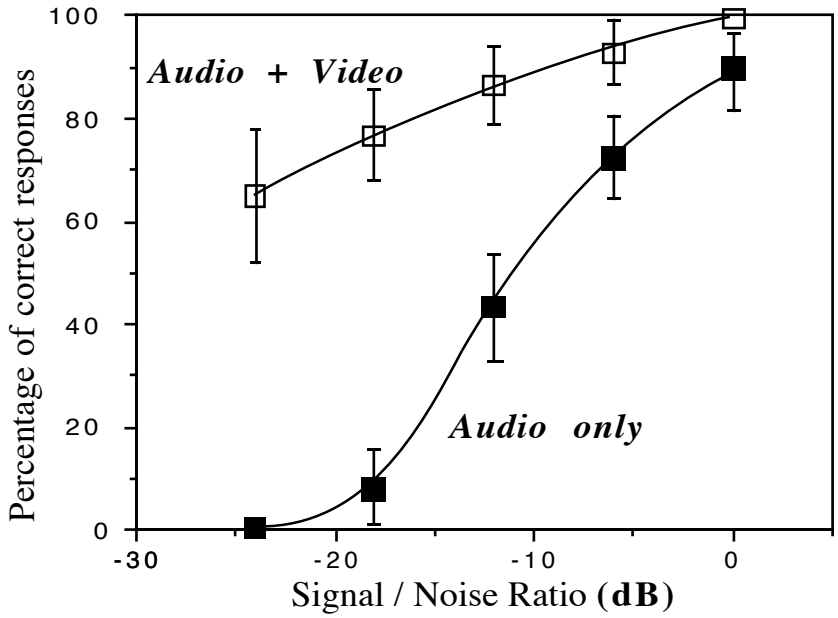

Figure 1. Means and SDs of 18 non-sense words identification in 2 conditions by 18 subjects.

control the speech level and were requested to set it up to a preferred listening level depending on different background noise conditions (quiet, 30, 40, 50 and 60 dB) did not modify their listening level according to whether the video picture of the speaker's face was provided or not. The author concluded that "although controlling a level is not the same as measuring intelligibility, this result nevertheless gives some indication of the apparently minor importance of lipreading on intelligibility for normal listeners under not too severe noise conditions." However, this result may simply show that listening "comfort" is independent of the presence of visual information.

1.1.4. Visual similarities between phonemes

Several experiments have tried to estimate visual similarities between phonemes. Confusion matrices suggest the existence of classes of lip/jaw shapes (visemes) that are similar to each other. Walden et al. (1977) measured the percentages of correct responses and of confusions between visually perceived American consonants by 31 normal-hearing subjects. Summerfield (1987) developped an ordered tree of confusions shown in Figure 2 (right panel). The tree presents (on the Y-axis) the order of confusion of consonants, and may thereby give a first classification of visemes. Figure 2 also shows that visual confusions are mainly independent from acoustic confusions (left panel, after Summerfield, 1987; from similar data obtained by Miller & Nicely,1955).

It has long been known that consonants differing in place of articulation are relatively difficult to distinguish under degraded auditory conditions, whereas consonants differing in voicing or nasality are relatively easy to distinguish (Miller & Nicely, 1955). The opposite is observed for visual presentation (Binnie et al., 1974;

Owens & Blasek, 1985). Vowel similarities are not so clear as consonant similarities: it seems impossible to determine reliable classes of vowels (vocalic visemes). The most robust feature is rounding (Jackson et al., 1976; Tseva, 1989).

In the end, the partial complementarity between the auditory and the visual mode is basic in the integration of visual and acoustic cues by listeners. We may hence hypothesize that TtS synthesis will be more intelligible with an added visual component (leading to TtAVS synthesis: Text-to-Audio-Visual-Speech), especially given the current quality of synthesizers.

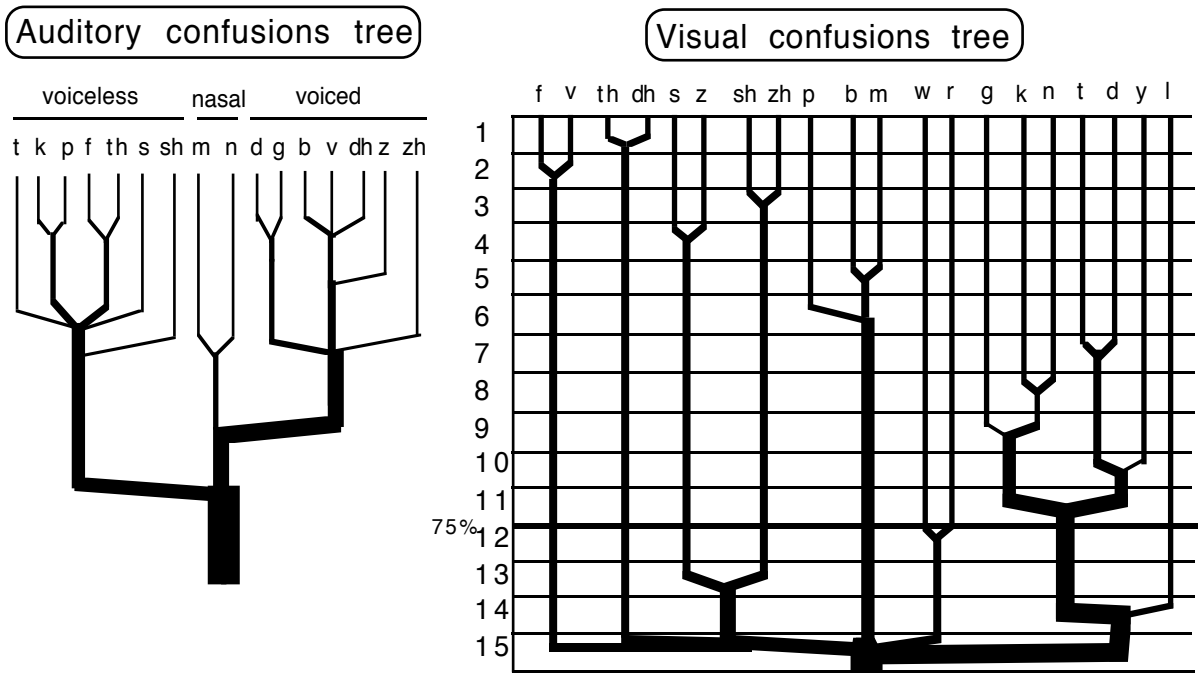

Figure 2. Auditory (left) and visual (right) confusions of consonants in CV syllables with the vowel “ah”, after Summerfield (1987, p. 13,15). Left panel: auditory data from Miller & Nicely (1955). Right panel: visual data from Walden et al. (1977).

1.2. Audio-visual speech timing

Few studies have investigated the temporal aspects of acoustic and visual perception of speech. Dixon and Spitz (1980) have shown that subjects are sensitive to the desynchronization between the noise of a hammer and its visual impact when the acoustic signal is presented more than 75 ms before, or more than 190 ms after the visual signal . For continuous speech, the delay sensitivity reported by these authors is respectively 130 ms and 260 ms, whereas selective perception of a single feature, such as voicing, is closer to the “hammer” effect: 80 ms and 140 ms (McGrath & Summerfield, 1985).

In a pilot experiment, Escudier et al. (1990) determined the auditory identification functions for the final target vowel [i] or [y] from progressively shortened /zizi/, /zizy/, /zyzi/, and /zyzy/, and compared it with the visual identification functions for front

photos of the speaker's face, coordinated with the acoustic shortening. They reported that visual identification of rounding was approximately 80 ms earlier than auditory identification in the /zizy/ transition, while no desynchronisation was observed in the /zyzi/ transition. Cathiard et al. (1991) showed that crosspausal rounding anticipation is clearly perceived by viewers up to 180 ms before the sound is produced. These studies suggest that a TtAVS would have to incorporate some (even asymmetrical) anticipation of visemes on the corresponding phonemes.

1.3. The concept of viseme

1.3.1. What does a viseme look like?

Given that many phonemes differ in non visible acoustic cues (such as voicing or nasality), there is a natural trend for phonemes to be visually grouped according to place similarities (e.g., [p,b,m]). Such classes were first called visemes by Fisher (1968) who derived the word from visual phonemes. As long as a viseme is considered on a "visuophonological" basis, the concept seems to be highly relevant, but coarticulation occurs at least as much in visual as in acoustic speech. For both sources, context distorts the pattern that would ideally been expected from observations of isolated vowels for instance. This is why, following Montgomery et al. (1987), we prefer to use the concept of viseme as a construct that is more experimental than theoretical and that is closer to the similar (acoustic) concept of allophone than to phoneme. In this perspective, the viseme could be a visual reference pattern answering the basic question: "How many visual shapes of the mouth and the jaw are necessary to describe all possible speech realizations in a given language?". The answer is obviously related to the production and to the perception of speech, since the visual description of speech depends on geometric aspects of the vocal tract which affect acoustics on one hand, and visual identification on the other. Variability between different speakers' mouth shapes, and therefore in the patterns of individual visemes, has been reported (Plant, 1980; Montgomery & Jackson, 1983; Abry & Boë, 1986 among others). Differences are even frequently seen between the two sides of a talkers' mouth (Graves et al., 1982; Wolf & Goodale, 1987).

Which anatomical components of the talker's face are most relevant to speech perception? Summerfield (1979) compared the percentages of correctly understood main words in sentences presented to subjects under degraded acoustic conditions and in various visual conditions: audio alone (A); full video of the speaker's face (B); the speaker's lips only (C); only four dots painted on the speaker's lips (D); and circles whose diameter varied with the envelope of the audio signal (E). The minimum, maximum and mean percentages of correct identification are replotted on Figure 3: it is clear that lips play a major role in visual speech intelligibility. The lack of information from jaw movement (obtained either from the jaw skin or from the teeth) leads obviously to the conclusion that lips ([C]) may not fully explain the increase in intelligibility, from [A] to [B]. These results can be related to a further study by McGrath (1985). He tested identification of 11 English vowels in a [bVb] context by 12 normal-hearing subjects in three visual conditions with the following results: full face = 77.8%; lips and teeth = 56.1%; lips only = 50.3%.

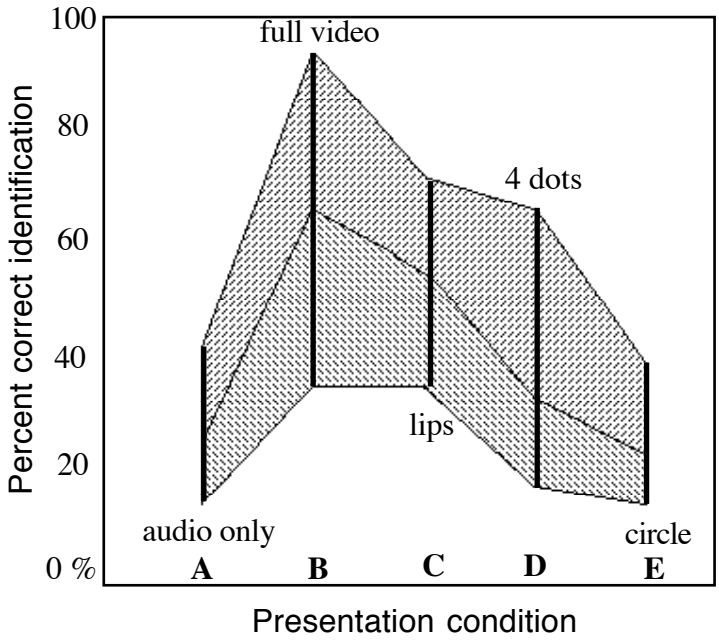

Figure 3. Mean percentages and ranges of main words correctly transcribed by 10 subjects in 5 conditions (data replotted from Summerfield, 1979)

1.3.2. Visemes from analysis of natural utterances

Based on speech production, classes of mouth/jaw patterns may be analyzed from video-taped corpora of natural speech. Several investigators have studied various shapes of speakers' faces (Fromkin, 1964; Abry & Boë, 1986; Zerling, 1990; for American and French speakers). Some authors have even related geometric measurements to visual confusions of lip shapes by subjects for American English, Australian and French (Jackson et al., 1976; Plant, 1980; Montgomery & Jackson, 1983; Cathiard, 1988).

Studies have seldom considered the effect of transitions between phonemes that take into account coarticulatory effects: Benguerel and Pichora-Fuller (1982) showed that variability in articulatory parameters resulting from coarticulatory effects increased overall lipreading difficulty; and Montgomery et al. (1987) showed that vowel intelligibility was significantly poorer in most contexts involving highly visible consonants. Most of the experiments that aimed at defining sets of visemes were based upon perceptual confusions by subjects rather than on geometric analyses of the shapes: Owens and Blazek (1985) proposed 21 visemes for General American consonants in various vocalic contexts (see their article for a compilation of such studies based upon perceptual analysis). However, analysis of geometric parameters extracted from face patterns has allowed investigators to define sets of visemes that are intended to be valid for any description, or synthesis, of the labial space in a given language: Fukuda and Hiki (1982) proposed 13 visemes for Japanese.

We present below the methodology that we used to analyze geometric measurements of a French speaker's face in order to define a set of French visemes and to discriminate among them.

1.4. Visual speech synthesis

Audio-visual synthesis provides a very promising tool for multi-media communication. Personal computers are now more and more frequently equiped with video facilities and digital telephone networks may fruitfully use this technology.

The reader is encouraged to look at Brooke's article (1991) in this book for an introduction on how to synthesize talking faces. Briefly, talking faces may be synthesized by various techniques, such as synthesis by rule either of vector-driven face shapes (Brooke & Summerfield, 1983), parametrized 3D facial models (Parke, 1982; Cohen & Massaro, 1990), concatenation either of video sequences of real faces (Girard, 1987), or of synthetic images (Saintourens et al., 1990; Magnenat-Thalmann, 1989). Whatever the technique that may use our results, we here focus on the analysis of a limited set of geometric parameters that can be used to calculate the shape of synthetic lips. Our objective is to define an optimal set of visemes to serve as a basis for any French TtAVS synthesizer. We describe in the following sections the corpus and the methodology that we used to measure, classify and discriminate target shapes, and we detail their corresponding phonetic classes.

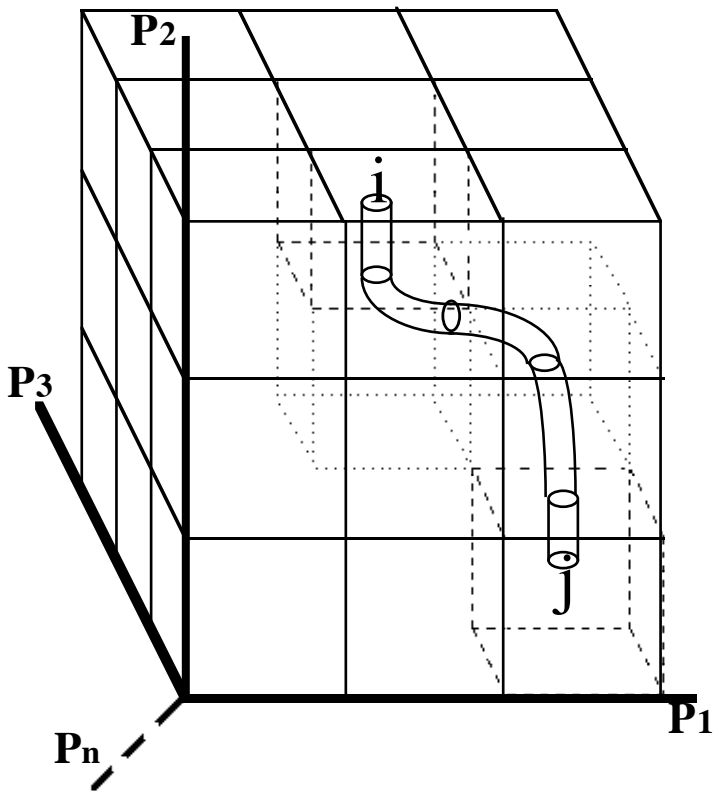

Figure 4. Theoretical structure of the "jaw-lip space": the multidimensional space (defined by the geometric parameters P_i, ..., Pn) is supposed to be composed of non-overlapping visemes (i, j, etc., that are here represented by adjacent cubes of equal size, instead of dispersed multinomial clouds, for simplicity purposes) whose trajectories between each other must be defined by a set of appropriate rules for interpolation between target faces in a TtAVS synthesiser.

To do so, we first supposed a hypothetical “lip-jaw space” (see Figure 4) composed of non-overlapping multidimensional areas. These areas (small cubes in Figure 4) correspond to natural variations of one speaker’s utterances of similar allophones. In order to analyze the whole space as accurately as possible, we limited our corpus to contextual realizations of vowels and of consonants that most strongly influence lip and jaw gestures. A critical observation of results from a dynamic clustering analysis is hereby able to provide us with phonetically labeled clusters of geometric patterns (i, j, etc.) whose distinctiveness from one another leads to the above mentioned definition of a set of visemes. The suggested independence can be assessed by means of discriminant analysis. A systematic study of trajectories between the visemes (i → j) that can be phonetically produced in serial order should ultimately provide us with a set of rules to control a lexicon-based TtAVS synthesizer.

2. CORPUS

Three vowels ($V_i \in$ [i, a, y]) and six voiced consonants ($C_k \in$ [b, v, z, ʒ, l, r] U {Ø}) were chosen, based on lip and jaw displacements, as labial types, and were used to create ”Indian names” of the form [$V_iC_kV_j$]. These were followed by a French third person present (existing or nonsense) verb, [C_kV_iz], to create 63 sentences of the type [$V_iC_kV_j$]+[C_kV_iz]. The remaining French vowels ($V_i \in$ [e, ɛ, ø, œ, u, o, ɔ, ɛ̃, œ̃, õ, ã]) were also included in a similar pattern ([V_iV_i]+[V_iz]). The 74 total items were embedded in the carrier sentence: “C’est pas [$V_iC_kV_j$] [C_kV_iz] ?”. The interrogative form stressed the last vowel and the following [z] lengthened it. Each sentence was repeated ten times in a pseudo-random order, modified to avoid consecutive repetitions of the same stimulus.

A male French native-speaker was chosen on the basis of sufficiently high dynamics and good symmetry in his lip displacements. He pronounced the 740 stimuli in a single session.

3. RECORDING

The audio-visual recording of the speaker was done in a sound-proof anechoic cubicle where an experimenter read aloud the sentences to the speaker, who was temporarily blinded with black goggles because of the 1,000 W halogen lamp necessary to illuminate his face (see Figure 5).

Two video cameras filmed the speaker’s face from front and profile (90°). A microphone was located 20 cm from the speaker’s mouth. The speaker’s lips were carefully made up in a blue color that was absent in the background and on the speaker’s face. A Chroma Key was connected to the output of the front camera so that the blue color was transformed to saturated black in real time. Both images (front and profile) were mixed on the same U-Matic magnetic tape. A tape without Chroma Key effects was archived. The reader-experimenter pressed a button to initiate a synchronized 1 kHz beep on the audio track and LED flash in the video field, for later verification of audio-visual coordination. A time-coder inserted the frame numbers to facilitate the labelling and subsequent research on each token.

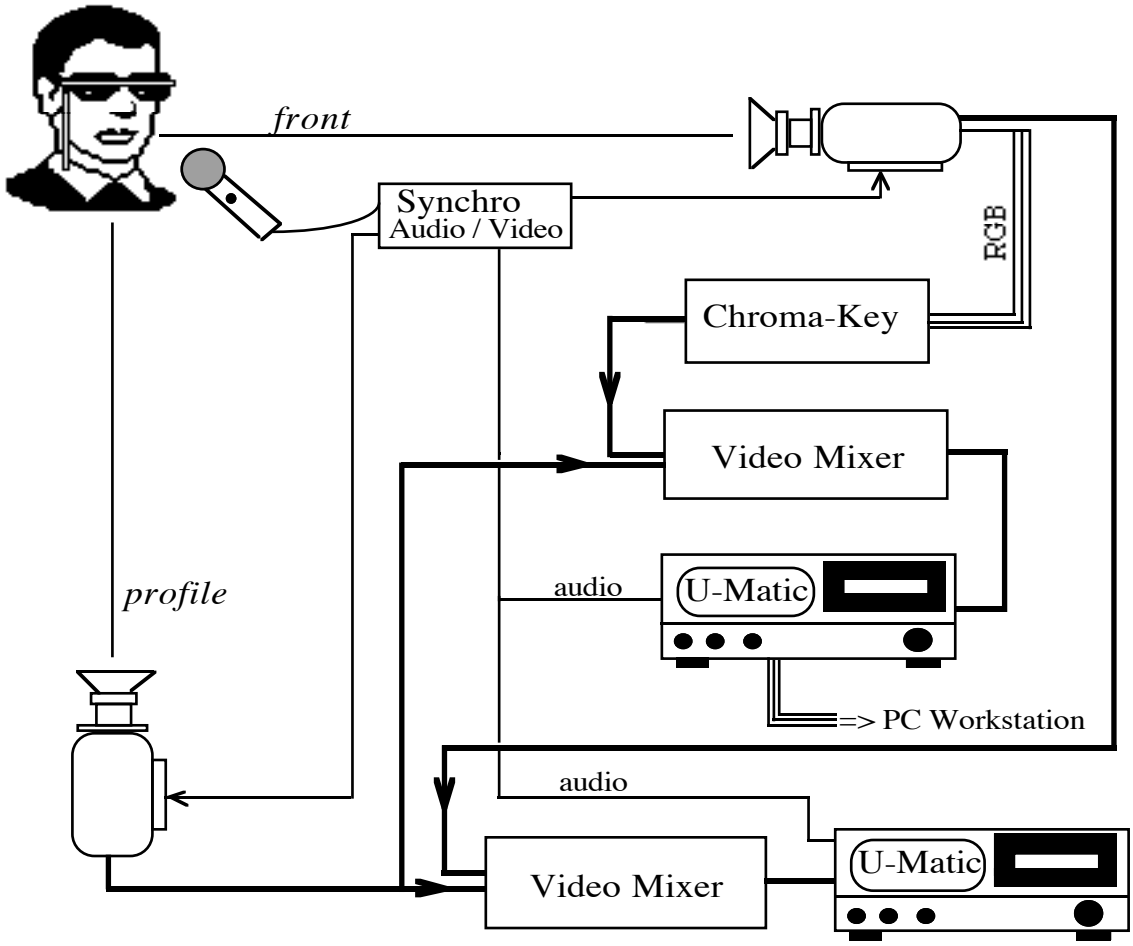

Figure 5. Schematic of the audio-visual recording setup (see text above).

4. LABELLING AND SELECTION OF THE STIMULI

Firstly, all sentences were labeled according to their LED flash address. The audio part of the video tape was then digitized on the "Video-Speech" workstation (Lallouache, 1990) so that the experimenters could hand-label the acoustic waveform with a specially designed signal editor by means of mouse-based dialog. 3300 video fields were labeled at the centre of each acoustic segment. We limited our analysis to nine repetitions of each of the following utterances:

- the middle vowel of each $[V_i \mathbf{V_i} V_i z]$; $\mathbf{V_i} \in \{i, a, y, e, \varepsilon, \emptyset, œ, u, o, ɔ, \tilde{\varepsilon}, \tilde{œ}, \tilde{o}, \tilde{a}\}$;
- the middle vowel of each $[V_i C_k \mathbf{V_i} C_k V_i z]$; $\mathbf{V_i} \in \{i, a, y\}$; $C_k \in \{b, v, z, ʒ, l, r\}$;
- the second consonant of each $[V_j C_k V_i \mathbf{C_k} V_j z]$; $(V_i, V_j) \in \{i, a, y\}$; $\mathbf{C_k} \in \{b, v, z, ʒ, l, r\}$.

A few views of the lips in a prephonatory position, and in a quiet closed position, were also analysed .

We therefore could analyze the effect of a dynamic self-context on 14 vowels, the effect of 6 consonants on 3 vowels and the effect of 3 vowels on 6 consonants.

5. AUTOMATIC EXTRACTION OF FACE PARAMETERS

The programs developed by Lallouache (1990) and adapted to this corpus computed various geometric parameters from the desired video fields. Data were then analysed by multidimensional analyses: dynamic clustering (Diday, 1973), discriminant analysis (Romeder, 1973) and correspondence analysis (Benzecri, 1973). For a tutorial presentation of these techniques, see Benoît (1983).

Seven distances and 4 areas were selected from the front view of the speaker and 3 others from his profile, so that his mouth's main characteristics could be determined (see Figure 6): upper lip (S1), lower lip (S2), lip opening (S) and total lip (St = S1 + S2 + S) areas; horizontal width (A) and vertical height (B) of the internal lip opening; horizontal width (A') and vertical height (B') of the external vermilion margins; Cupid's Bow width (L) and height (h); chin distance to goggles (M); protrusion distances between the vertical profile reference and the upper lip (P1), the lower lip (P2) and the lip contact point (C). Three derived parameters were also analysed: A·B, A'·B' and A'/B'. This was done to verify that A·B was strongly correlated to S (as previously shown by Abry and Boë in 1986 on French vowels), to test St and A'·B' correlation, and to evaluate the discriminative power of A'/B'.

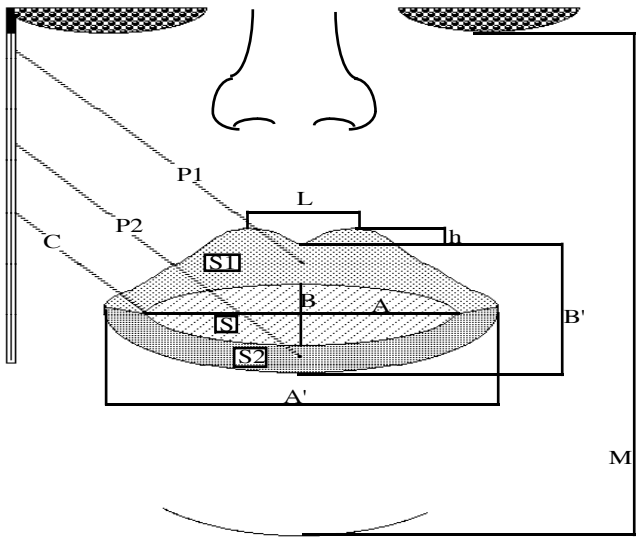

Figure 6. Parameters automatically measured on the speaker's face (see text)

6. DATA ANALYSIS

Dynamic clustering and correspondence analysis suggested several classes. Generally, all 14 vowels and all contextual allophones of [i, a, y] and of the 6 consonants were rather easily classified into 19 main classes. Two exceptions, [r] and [l] allophones adjacent to a single [y], are widely dispersed and overlap with other classes. Their barycentres are however located in a vacant place within the multidimensional space of our 19 visemes.

Overall, 21 main classes of allophones can be discerned (see Annex).

To separate these 21 classes, the most discriminant parameters are the inside lip width (A) and height (B) (or the area S, which is correlated to A·B, $r > .999$), the

external “pattern factor” (A’/B’) and the lip contact protrusion (C). When applied to the 774 vectors used by a discriminant analysis as a learning set to separate the 21 classes, a Mahalanobis distance (Romeder, 1973) correctly classified 80 % of these 774 vectors with only three parameters: A, S and A’/B’. Figure 7 displays the projection of the 21 visemes (barycentres of all data regrouped within one class of lip/jaw shapes) and of 6 of the most discriminant parameters on the two first factors obtained by a correspondence analysis, accounting for 80% of the 14 parameters’ total variance.

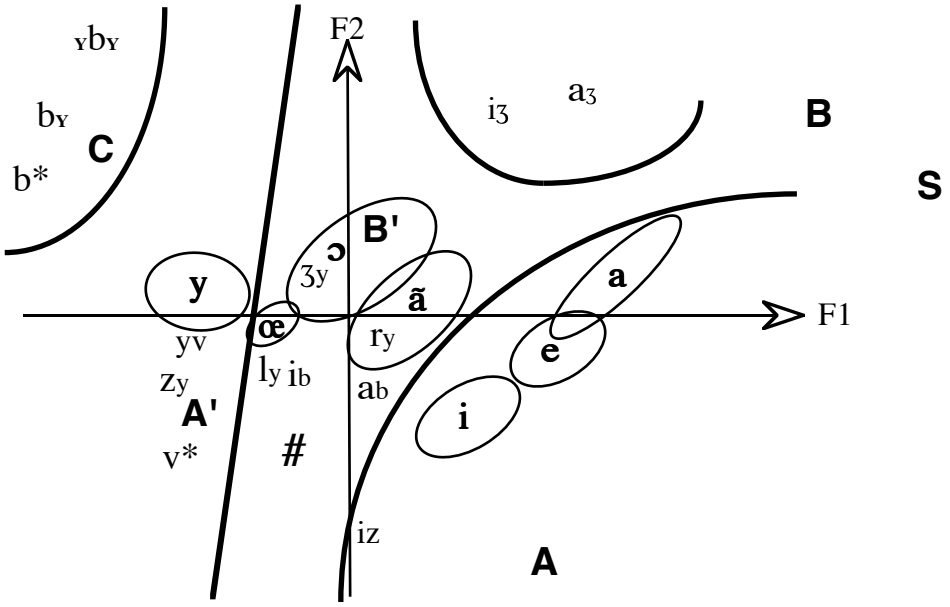

Figure 7. Simultaneous projection of 6 main parameters and of 21 visemes on the two first factors obtained by a correspondence analysis (see caption next page).
A, B, A’, B’, C, S: six of the most discriminant parameters
a, e, i, ã, ɔ, œ, y: the 7 visemes consisting of vowels in a self-context; barycentres are plotted in dispersion ellipses of the corresponding data (90% confidence level)
vc : visemes with just vowels (V) in a consonantal context (C)
cv : visemes with just consonants (C) in a vocalic context (V)
vc : visemes combining vowels (V) in a consonantal context (C) and vice-versa
: prephonatory shape of the lips
(See the Annex for the list of allophones corresponding to these visemes.)

Correspondence analysis allows graphics to simultaneously display projections of data and of their describing variables, as it mainly compares the profiles of data or of variables. The χ^2 distance that is used is symmetrical between data and variables so that: i) the closer the visemes are plotted, the more similar they look according to the parameters that were measured; ii) the closer the parameters are plotted, the more similar they behave according to the set of allophones that were analysed; and finally, iii) the closer a parameter and an allophone are plotted, the more specifically this allophone behaves according to this parameter when compared to other parameters and to other allophones.

Five main groups of visemes may be discerned depending on the values of A (internal width), B (internal height) and C (lip contact protrusion). Separated in Figure 7 by thick curves, they can be phonetically described as follows, using the notation ÿ for "rather small", ⁄ for "rather large", and ÷ for "medium":

- top left: A = B = 0, C⁄; closed lips (various contextual realizations of [b]);
- centre left: Aÿ, Bÿ, C⁄; rounded-protruded lips (mainly [y, u, ø, o, õ} or rather closed lip slit ([v]);
- centre: A÷, B÷, C÷; medium shapes (mainly [œ, ã, ɔ] + prephonatory [#]);
- top right: A÷, B⁄, C÷; open-protruded lips (mainly [ʒ]);
- bottom right: A⁄, B÷, Cÿ; spread lips (mainly [i, e, a]).

7. BI-DIMENSIONAL REPRESENTATIONS

7.1. The highest correlations between selected parameters

Table 1 shows the highest values (r > .95) that were observed when correlating all parameters between each other. The highest correlation (r = .999) is obtained with S (internal lip area) and A·B (area of the rectangle in which the internal area could be included). This result is coherent with previous observations (Fromkin, 1964; Lindblom & Sundberg, 1971; Abry & Boë, 1986). A slightly smaller tendancy is observed with external lips (vermillion contour), as r = .99 between St and A'·B'.

Table 1
Highest correlation coefficients between lip shape parameters (786 utterances)

	B	B'	S	St	A·B
B'	.96				
S	.97	.92			
St	.98	.95	.99		
A·B	.97	.92	.999	.99	
A'·B'	.98	.97	.98	.99	.98

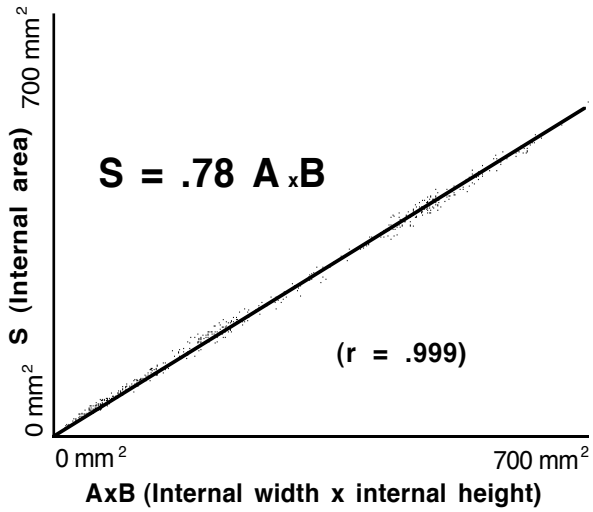

Figure 8. Correlation between the internal lip area (S) and the product A·B (rectangle constituted of the internal width and the internal height).

Figure 8 plots parameter S vs. the product of parameters A and B. The

regression line follows the “rule” S = .78 A·B, determined by the lips of our speaker. Abry and Boë (1986) reported S = .75 A·B for the lips of a female French speaker. It is therefore apparent that the internal lip area (the visual parameter most closely related to the acoustics of the vocal tract) may be adequately calculated from the measured distances A and B, whatever the shape of the mouth.

7.2. Internal lip patterns

Figure 9 shows the projection of 786 utterances on the {A,B} plane. Two tendencies can be observed from the physiological patterns of the lips in their internal width (A) and height (B) displacements. In one sub-area (right part) are plotted the internal lip patterns of spread allophones whose trajectories between each other are followed with a relatively constant lip width. On the other hand (left part), a sub-area consists of all other patterns whose trajectories between each other roughly follow a rule of the form $B = k A^n$ ($B \# .00144 A^{2.5}$, here). This second category even includes spread vowels that are surrounded by protruding consonants (namely [ʒ]). It could hence be possible to approximately predict one parameter from the other, once given the surrounding context of a phoneme: this result may be of major importance for rule-based synthesis of faces. It is noticeable that the prephonatory pattern of internal lip ([#]) stands right between the two categories, just as if the speaker stands ready to produce one category or the other.

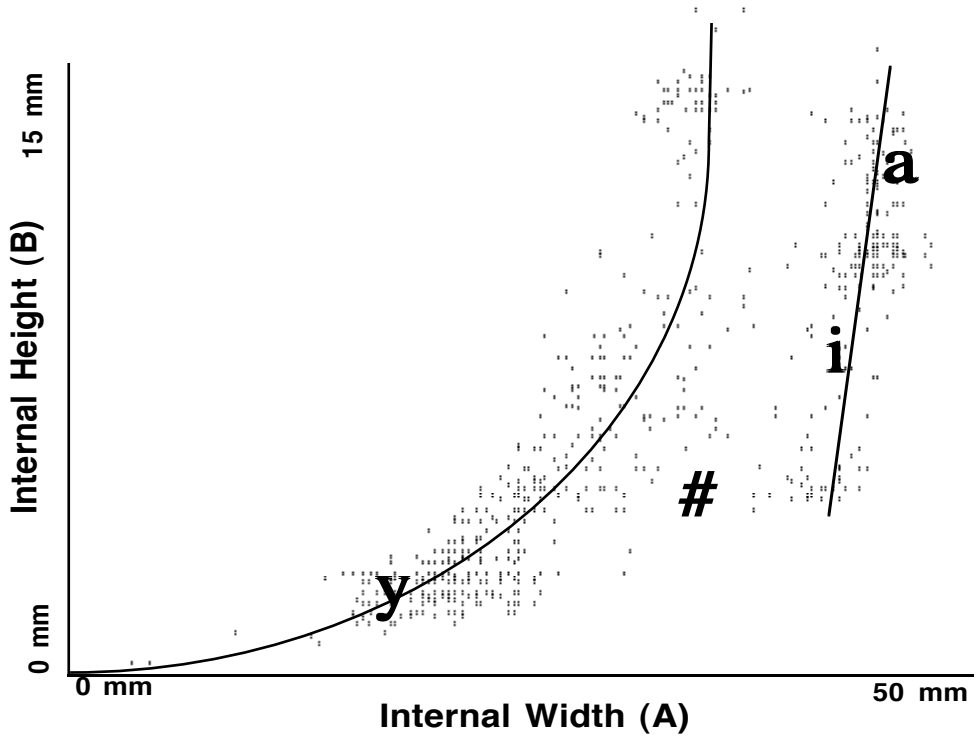

Figure 9. Projection of B vs. A. Two main tendencies can be observed: protrusion gesture (left side: $B \# .00144 A^{2.5}$), and spreading gesture (right side: A # Cst.). The figure shows the barycentres of the three extreme vowels in a self-context [i, a, y], as well as the prephonatory shape of the lips [#].

7.3. External lip patterns

Figure 10 shows the projection of 786 utterances on the {A',B'} plane. The overall correlation between A' and B' surprisingly equals zero. There are few studies related to the external lips. Plant (1980) however reported the visual importance of the external lips. The visemes are not displayed on Figure 10 for clarity purposes, but the three visemes representing [b] in various allophonic contexts could be distinguished on this figure whereas they could not on Figure 9, as A = B = 0.

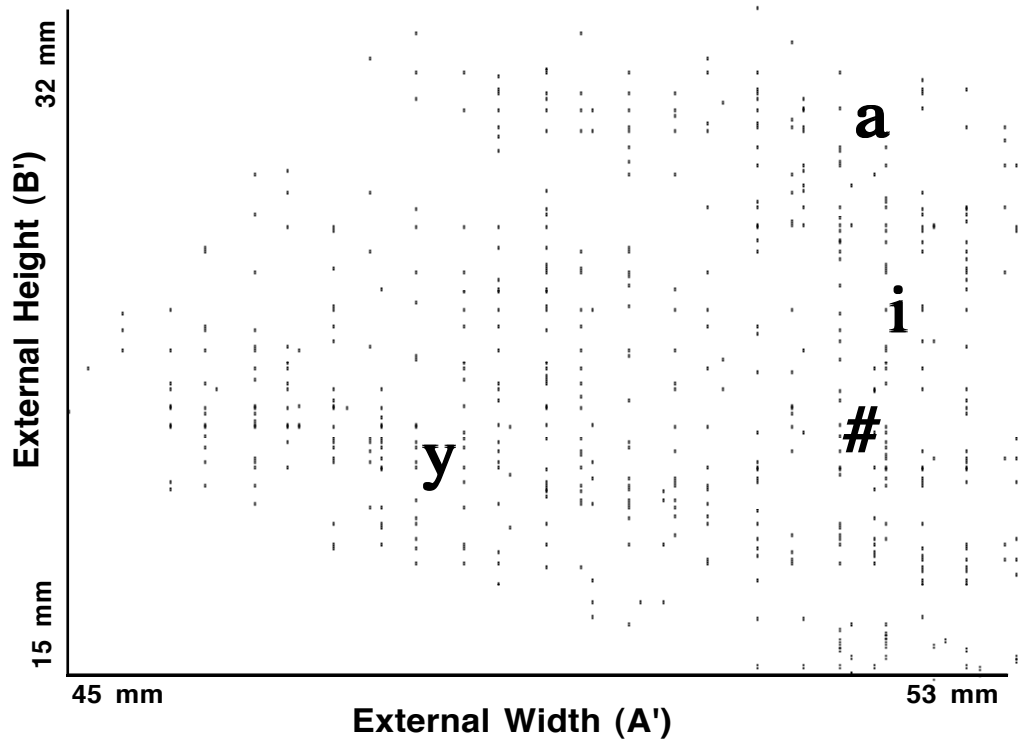

Figure 10. Projection of B' vs. A': external lip contour representation. The figure shows the barycentres of the three extreme vowels in a self-context [i, a, y], as well as the prephonatory shape of the lips [#].

7.4. Lip contact and "chin patterns"

Figure 11 plots the 786 measured distances of the chin and of the lip contact protrusion from reference points on the speaker's face. This projection is mainly related to the profile view of the speaker. It may be seen that, once again, all patterns, except for [#], stand in the "triangle" whose corners are defined by the three extreme vowels: [a] with the maximum lowering of the chin, [i] with the highest chin and the smaller lip contact protrusion, and [y] with the longest lip contact protrusion. It is noticeable that the prephonatory position [#] stands out of the "speech space". This observation confirms that a specific viseme must be dedicated to this special position. Measurements of the chin cannot be strictly extended to the jaw, as protrusion of the lower lip implies a raised chin.

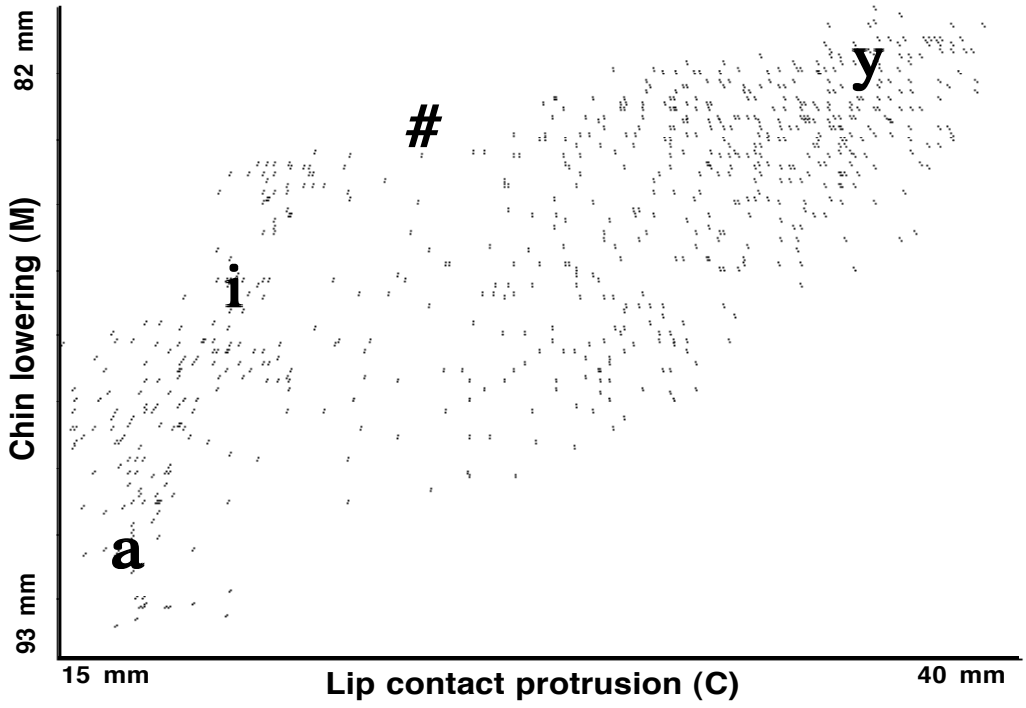

Figure 11. Projection of M vs. C: profile representation. The figure shows the barycentres of the three extreme vowels in a self-context [i, a, y], as well as the prephonatory shape of the lips [#].

8. DISCUSSION

We classified the 14 French vowels into 7 classes (see Annex: visemes N° 1 to 7) that are coherent with previous classifications based upon internal lip measurements (Abry & Boë, 1983; Zerling, 1990).

We employed parameters that were presumed to play a major role in speech, both in visual perception and in face movement synthesis. It is hardly a surprise that the parameters best discriminating our set of 21 visemes are related to internal lip geometry, and thus to the acoustics of the vocal tract output.

The two visemes [y**ʁ**V, V**ʁ**y] and [y**l**V, V**l**y] (where V is [i] or [a]) look at least as polymorphic as their corresponding phonemes: their dispersions are very large, almost ranging from the viseme [y] to the accompanying vocalic visemes, [i] or [a]. Nevertheless, their barycentres occupy two places that are vacant inside the multidimensional space of visemes, and no other viseme can replace them, even if they overlap a great deal with their neighbours. We must hence take them into account, even if they are to be used as “glide visemes”, i.e., as transit visemes rather than as target visemes.

It must be noticed that the 21 visemes were obtained from multidimensional analysis, and that second order variations allow a more or less accurate classification of these shapes. For instance, class 5 and class 20 are composed of two classes of phonetically very different constituents. While these two classes are close to each other according to main variations, they differ somewhat in lip

protrusion, as class 20 (ʒ adjacent to a y) present higher lip protrusion than class 5 (ɔ) A similar observation may be made when considering jaw displacement of the vowels that are collapsed within class 3: [ø, o, õ] present less jaw lowering than [y, u]. Such detailed differences within the classes would require perceptual tests to validate their hypothetical significance.

9. CONCLUSION

Our analysis showed that in our French reference speaker, the lip configurations can be adequately described by twenty-one visemes. Two other "pseudo-visemes" must be added to take into account non-linguistic shapes: the "quiet closed lips" and the "prephonatory position". A viseme-based synthesiser for French should take this finding into account.

Still, many studies remain to be done before synthesis of an intelligible "talking face" will be possible. First of all, we must extrapolate contextual realizations to other vowels than the [a, i, y] classes of our observations, either using the general effects of consonants on vowels that were here observed, or by research based on an extended corpus. Second, we must interpolate appropriate shapes for the transition between any two selected visemes. Third, we must define a transducer to convert a string of phonemes into a string of existing visemes. Finally, we must elaborate a set of rules to specify the timing of viseme appearances from the durations of phonemes and of their transitions.

ACKNOWLEDGEMENTS

This research was supported by the C.N.R.S. and by the Agence de Coopération Culturelle et Technique (ACCT). We want to thank Marie-Agnès Cathiard for her profond contribution to this introduction, Argyro Tseva for editing the corpus and labelling the sentences, and Jean-Luc Schwartz for speaking, blindly but clearly, in front of 1,000 W lamps. Special thanks to the Assistant-Editor of this whole book, Tom Sawallis, for interest, editing, and patience!

ANNEX: the set of French visemes

Notation: * = [i] or [a];
V (c) = {[c**V**c] / V ∈ [i, a, y]};
C (v) = {[v**C**v] / v ∈ [i, a, y]};
C (v*) = {[v**C**i], [v**C**a] / v ∈ [i, a, y]};
C (*v) = {[i**C**v], [a**C**v] / v ∈ [i, a, y]}.

N°	Symbol	$[v_i \underline{V}_i v_i z]$	$[v_j c_k \underline{V}_i c_k v_j z]$	$[v_i c_k v_j \underline{C}_k v_i z]$
1	[a]	a	a (r, l)	l (a), r (a*, *a)
2	[i]	i	a (z)	
3	[y]	y, u, ø, o, õ	y (b, z, ʒ, r, l)	[z, ʒ, r, l] (y)
4	[e]	e, ɛ, ɛ̃	i (r, l)	r (i), l (i*, *i)
5	[ɔ]	ɔ		
6	[œ]	œ, œ̃		
7	[ã]	ã		
8	[ab]		a (b, v)	
9	[aʒ]		a (ʒ)	
10	[ib]		i (b, v)	
11	[iz]		i (z)	z (**)
12	[iʒ]		i (ʒ)	ʒ (**)
13	[yv]		y (v)	v (y)
14	[b*]			b (**)
15	[by]			b (y*, *y)
16	[v*]			v (**)
17	[zy]			[v, z] (y*, *y)
18	[ry]			r (y*, *y)
19	[ly]			l (y*, *y)
20	[ʒy]			ʒ (y*, *y)
21	[yby]			b (y)
22	[#]	prephonatory shape of the lips		
23	[=]	quiet closed lips		

REFERENCES

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

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

Benoît C. (1983). Vers une herméneutique de lanalyse des données applicable au locuteur, Bulletin de l'Institut de Phonétique, Grenoble, 12, 155-183.

Benzecri J.P. (1973). L'Analyse de Données, Tome 2: L'Analyse des Correspondances, Dunod, Paris, 619 p.

Binnie C.A. Montgomery A.A. & Jackson P.L. (1974). Auditory and visual contributions to the perception of consonants, Journal of Speech & Hearing Research, 17, 608-618.

Brooke M. (1991). Computer graphics synthesis of talking faces", this volume.

Brooke M. & Summerfield Q. (1983). Analysis, synthesis and perception of visible articulatory movements, Journal of Phonetics, 11, 63-76.

Cathiard M.A. (1988). Identification visuelle des voyelles et des consonnes dans le jeu de la protrusion et de la rétraction des lèvres en Français, Mémoire de Maîtrise, Université Grenoble II, 106 p.

Cathiard M.A. (1988/1989). La perception visuelle de la parole : Aperçu des connaissances, Bulletin de l'Institut de Phonétique de Grenoble, 17/18, 109-193.

Cathiard M.A., Tiberghien G., Cirot-Tseva A., Lallouache M.T. & Escudier P. (1991). Visual perception of anticipatory rounding during acoustic pauses: a cross-language study, Proceedings of the XIIth International Congress of Phonetic Sciences, 4:50-53, Aix-en-Provence, France.

Cohen M.M. & Massaro D.W. (1990). Synthesis of visible speech, Behavior Research Methods, Instruments, & Computers, 22, 260-263.

Cotton J. (1935). Normal "visual-hearing" , Science, 82, 592-593.

Diday E. (1973). The dynamic clusters method and sequentialization in non-hierarchical clustering, in Proceedings of the International Joint Conference on Pattern Recognition, Washington DC, 14 p.

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

Dodd B. (1977). The role of vision in the perception of speech, Perception, 6, 31-40.

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

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

Escudier P., Benoit C. & Lallouache T. (1990). Identification visuelle de stimuli associés à l'opposition /i/-/y/ : étude statique, First French Conference on Acoustics , Lyon, France, 541-544.

Fisher C.G. (1968). Confusions among visually perceived consonants, Journal of Speech & Hearing Research, 15, 474-482.

Fromkin V. (1964). Lip position in American English vowels, Language and Speeh, 7, 215-225.

Fukuda Y. & Hiki S. (1982). Characteristics of the mouth shape in the production of Japanese: Stroboscopic observation, Journal of the Acoustical Society of Japan, 3, 75-90.

Girard P. (1987). DAVOS: a computerized videodisc animation with synchronized speech, Section Informatique du Département. des Sciences Economiques. et Administratives, Université du Québec à Chicoutimi, Canada, 20p.
Graves R., Goodglass H. & Landis T. (1982). Mouth asymmetry during spontaneous speech, Neuropsychologia, 20, 371-381.
Jackson P.L., Montgomery A.A. & Binnie C.A. (1976). Perceptual dimensions underlying vowel lipreading performance, Journal of Speech & Hearing Research, 19, 796-812.
Kitson H.D. (1915). Psychological tests for lipreading ability, Volta Review, 17, 471-476.
Lallouache T. (1990). Un poste visage-parole. Acquisition et traitement des contours labiaux, Journées d'Etude sur la Parole, Société Française d'Acoustique, Montréal, 282-286.
Lindblom B. & Sundberg, J. (1971). Acoustical consequences of lip, tongue, jaw and larynx movement, Journal of the Acoustical Society of America, 50, 1166-1179.
Magnenat-Thalmann N. (1989). The problematics of facial animation, pp. 47-55 in State-of-the-art in computer animation, Magnenat-Thalmann N. & Thalmann D., Eds., Springer-Verlag.
McGrath M. (1985). An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces, Ph.D. Thesis, University of Nottingham, UK.
McGrath M. & Summerfield Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults, Journal of the Acoustical Society of America, 77, 678-685.
Miller G.A. & Nicely P.A. (1955). An analysis of perceptual confusions among some English consonants, Journal of the Acoustical Society of America, 27, 338-352.
Mohamadi T. & Benoît C. (to appear). Apport de la vision du locuteur à lintelligibilité de la parole bruitée, Bulletin de la Communication Parlée, 2, Cahiers de l'ICP, INPG, Grenoble, France.
Montgomery A.A. & Jackson P.L. (1983). Physical characteristics of the lips underlying vowel lipreading performance, Journal of the Acoustical Society of America, 73, 2134-2144.
Montgomery A.A., Walden B.E. & Prosek R.A. (1987). Effects of consonantal context on vowel lipreading, Journal of Speech & Hearing Research, 30, 50-59.
Neely K.K. (1956). Effect of visual factors on the intelligibility of speech, Journal of the Acoustical Society of America, 28, 1275-1277.
Norton S.J., Schult M.C., Reed C.M., Braida L.D., Durlach N.I., Rabinowitz W.M. & Chomsky, C. (1977). Analytic study of the Tadoma method: Background and preliminary results, Journal of Speech & Hearing Research, 20, 574-595.
Owens E. & Blasek B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers, Journal of Speech & Hearing Research, 28, 381-393.
Parke F.I. (1982). Parametrized models for facial animation, I.E.E.E. Computer Graphics, 2, 61-68.
Plant G.L. (1980). Visual identification of Australian vowels and diphtongs, Australian Journal of Audiology, 2, 83-91.

Pols L.C.W. (1979). Intelligibility of intervocalic consonants in noise, Procee-dings of the Fifth I.E.E.E. International Conference on Acoustics, Speech & Signal Processing, 460-463.
Risberg A. (1982). Speech coding in aids for the deaf: An overview of research from 1924-1982, Speech Transmission Laboratory Quarterly Progress & Status Report, 4/1982, KTH, Stockholm, 65-98.
Romeder J.M. (1973). Méthodes et programmes d'analyse discriminante, Dunod, Paris .
Rosen S.M., Moore B.C.J. & Fourcin A.J. (1979). Lipreading with fundamental frequency information, Proceedings of the Institute of Acoustics (UK), 1A2, 5-8.
Saintourens M., Tramus M.H., Huitric H. & Nahas M. (1990). Creation of a synthetic face speaking in real-time with a synthetic voice, ESCA Workshop on Speech Synthesis, Autrans, France, 249-252.
Stetson R.H. (1928). Motor Phonetics, Archives Néerlandaises de Phonétique Expérimentale, 3, 1-216 (2nd ed., 1951, Amsterdam; re-ed. 1988, by J.A.S. Kelso & K.G. Munhall, Boston).
Sumb 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. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception, in Hearing by eye: The psychology of lipreading, B. Dodd & R. Campbell, Eds.
Summerfield Q., MacLeod A., McGrath M. & Brooke M. (1989). Lips teeth and the benefits of lipreading, in Young & Ellis, Eds, 1989.
Tseva A. (1989). L'arrondissement dans l'identification visuelle des voyelles du français. Premiers acquis, Bulletin du Laboratoire de la Communication Parlée, INPG-ENSERG, Grenoble. 149-186.
Walden B.E., Prosek R.A., Montgomery A.A., Scherr C.K. & Jones C.J. (1977). Effects of training on the visual recognition of consonants, Journal of Speech & Hearing Research, 20, 130-145.
Wolf M.E. & Goodale M.A. (1987). Oral asymmetries during verbal and non-verbal movements of the mouth, Neuropsychologia, 25, 375-396.
Young A.W. & Ellis H.D. (1989), Eds. Handbook of research on face processing, Elsevier Science Publishers B.V., North-Holland, Amsterdam.
Yuhas B.P., Goldstein M.H. & Sejnowski T.J. (1989). Integration of acoustic and visual speech signals using neural networks, I.E.E.E. Communications Magazine, Nov. 1989, 65-71.
Zerling J.P. (1990). Aspects articulatoires de la labialité vocalique en français, Thèse de Doctorat d'Etat, Institut de Phonétique, Strasbourg, 545 p.