Lips and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications

Piero Cosi and Emanuela Magno Caldognetto
CSRF, Centro di Studio per le Ricerche di Fonetica - CNR, Padova, Italy.

Abstract. This research focuses on the spatio-temporal characteristics of lips and jaw movements and on their relevance for lip-reading, bimodal communication theory and bimodal recognition applications. 3D visible articulatory targets for vowels and consonants are proposed. Relevant modifications on the spatio-temporal consonant targets due to coarticulatory phenomena are exemplified. When visual parameters are added to acoustic ones as inputs to a Recurrent Neural Network system, high recognition results in plosive classification experiments are obtained.

Keywords. Lips, jaw, visible articulatory movements, lip-reading, bimodal recognition, auditory models, neural networks.

1 Introduction

Lip-reading and bimodal recognition research is following the same trend which occurred to the studies on the transmission of the linguistic information by the auditory channel. The first stage was focused on visual intellegibility tests, i.e on the quantification of the information trasmitted by the visual channel. In the second stage the research proceeds with the identification of the characteristics of the signal which trasmit the information. To that purpose, various devices capturing and recording *distal* and *proximal* signals have to be designed, built and tuned up, and various techniques for the creation of synthetic stimuli for experimental tests have to be developed. Only relying on a great amount of experimental data, sufficient to capture the complexity of the whole phenomenum, and, possibly, characterized by a cross-linguistic nature in order to separate the fundamental mechanisms from more liguo-specific characteristics, the elaboration of adequate theories of visual perception of articulatory movements and of bimodal perception of speech will be possible. The experimental data presented in the following are intended to contibute to this second stage of the research (Magno Caldognetto et al., 1995). In fact they constitute the natural development of previous studies executed at CSRF focused on auditory (Magno Caldognetto and Vagges, 1990a, 1990b) and visual (Magno Caldognetto, Vagges and Ferrero, 1980) intelligibility tests which enabled us to quantify and verify the characteristics of the phonological information transmitted separately by both channels. As illustrated in

Figure 1, the intelligibility of visible articulatory movements, as it was expected and parallely to other languages, is high only for bilabial (/p/, /b/) and labiodental (/f/, /v/) consonants, while the correct identifications gradually reduce from anterior to posterior loci of articulation. As for the manner of articulation, the visible identification of nasals and of all voiced consonants is particularly difficult due to the fact that neither the movements of the velum neither that of the vocal folds are visible. On the contrary, considering the auditory intelligibility tests in various noise masking conditions (Magno Caldognetto, Ferrero and Vagges, 1982), (Magno Caldognetto, Vagges and Ferrero, 1980), (Magno Caldognetto, Vagges and Ferrero, 1988), nasals, liquids (laterals and trills) and all sonorant consonants are well identified. The analysis of identification errors enabled us to build the two dendrograms illustrated in Figure 2. The groups of consonants visually confused (*visemes*) tend to share the locus of articulation, while the clusters obtained in the auditory identification tests correspond to unvoiced, voiced and sonorant consonants, which are characterized by different spectral patterns implying different manners of articulation and different activity of the vocal folds. These results are similar to those obtained for other languages, by Summerfield (1987) or Mohamadi and Benoit (1992), and support the idea of the *auditory and visual bimodal synergism* relevant for the development of the theories on language acquisition by normal and pathological infants, on speech communication and for various technological applications, such as audio-visual speech synthesis (Benoit et al., 1992), (Cohen and Massaro, 1990) or audio-visual speech recognition systems (Petajan, 1984).

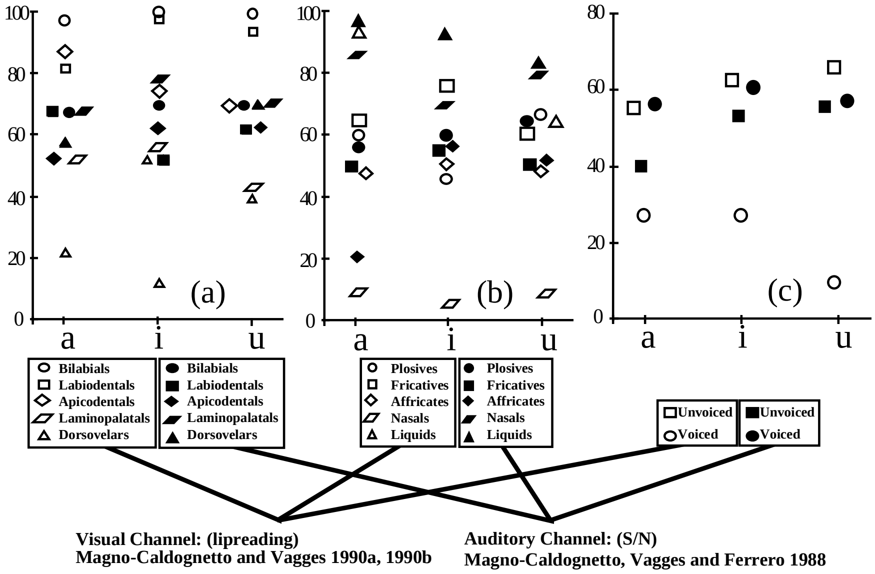


Figure 1. Results of the intelligibility tests for Italian consonants: (a) loci of articulation, (b) manners of articulation, (c) voiced/unvoiced opposition.

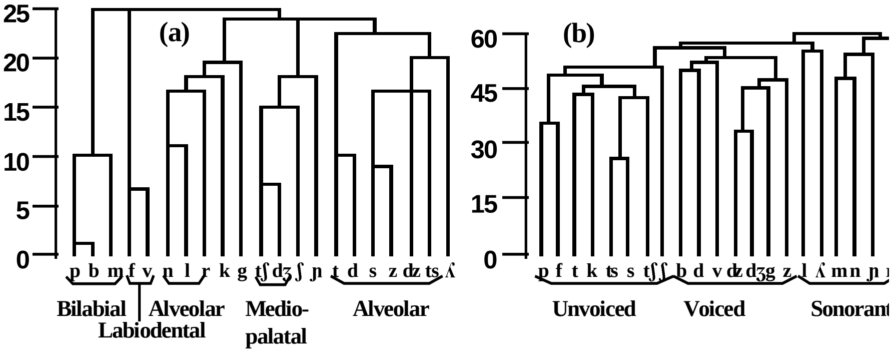


Figure 2. Cluster analysis of (a) visual (Magno Caldognetto, Vagges and Ferrero, 1980) and (b) auditory (Magno Caldognetto and Vagges, 1990a, 1990b) confusions among Italian consonants.

2 Spatio-temporal characteristics of lips and jaw movements

The definition of the spatio-temporal characteristics of visible articulatory movements is actually important because it provides the basic experimental material whith which various relevant theoretical problems could be tackled, such as, for example:

- the quantification of the available visible information relative to each phonological unit;
- the definition of the perceptive role of various articulatory parameters and of their relation and possible cooccurrence with linguistic (distinctive) features;
- the identification of rules able to capture the variability induced by phonetic context, prosodic variations, speech rate;
- the determination of the iso- or aniso-morphism between articulatory movements and their correspondent acoustic product in order to formulate adequate rules for the integration of visual and auditory information useful for the synthesis of *visible speech* (talking heads).
- the implementation of new audio-visual technological applications such as new speech recognition systems (see Section 3);

At present, the spatial characteristics of vowels and of a significative subset of consonants have been rather exaustively explored. As for the dynamic characteristics, only few illustrative analyses will be proposed being conscious, however, that future researches will concentrate on this subject: in fact, the results of the bimodal recognition experiments which will be described in the third section are already supporting this statement.

For that purpose, lip and jaw articulatory movements were recorded and analyzed with ELITE (Magno Caldognetto et al., 1989), a fully automatic real-time movement analyzer for 3D kinematics data acquisition (Ferrigno and Pedotti, 1985). This system ensures a high accuracy and minimum discomfort to the subject. In fact, only small, non obtrusive, passive markers of 2mm of diameter, realized by reflective paper, are attached onto the speaking subject's face, as illustrated in Figure 3.

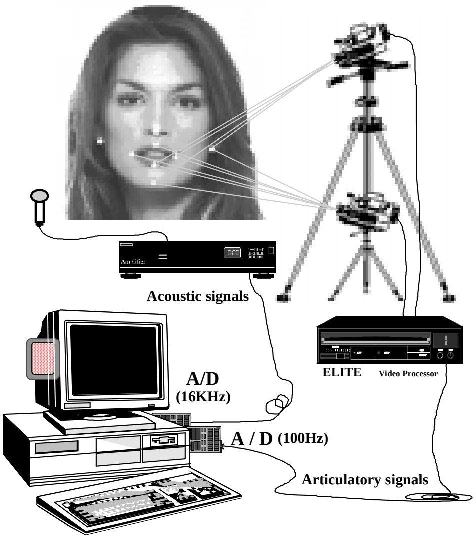


Figure 3. The ELITE System.

The subjects are placed in the field of view of two CCD TV cameras at 1.5 meters from them. These cameras light up the markers by an infrared stroboscope, not visible in order to avoid any disturbance to the subject. ELITE is characterized by a two level architecture. The first level includes an interface to the environment and a fast processor for shape recognition (FPSR). The outputs of the TV cameras are sent at a frame rate of 100 Hz to the FPSR which provides for markers recognition based on a cross-correlation algorithm implemented in real-time by a pipe-lined parallel hardware. This algorithm allows the use of the system also in adverse lighting conditions, being able to discriminate between markers and reflexes of different shapes although brighter. Furthermore, since for each marker several pixels are recognized, the cross-correlation algorithm allows the computation of the weighted center of mass increasing the accuracy of the system up to 0.1mm on 28cm of field of view. The coordinates of the recognized markers are sent to the second level which is constituted by a general purpose personal computer. This level provides for 3D coordinate reconstruction, starting from the 2D perspective projections, by means of a stereophotogrammetric procedure which allows a free positioning of the TV cameras. The 3D data coordinates are then used to evaluate the parameters described hereinafter. Simultaneously to the articulatory signals, ELITE records also the coproduced acoustic signal. In this study the markers were placed, as illustrated in Figure 4, on the central points of the vermilion border of the upper lip, and of the lower lip, at the corners of the lips, and at the center of the chin. The markers placed on the tip of the nose and on the lobes of the ears served as reference points to eliminate the effects of the head movement. In Table 1 are indicated all the most relevant articulatory movements and parameters which can be analyzed using ELITE. The axes are obviously related to phonological features, x being joined to *lip-rounding,* y to *lip-protrusion* and z to *lip-opening*.

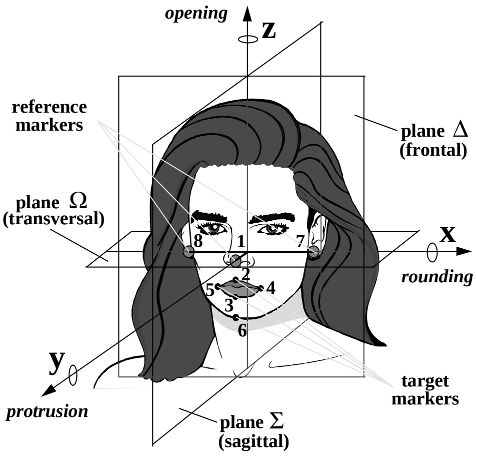


Figure 4. Position of the reflecting markers and of the reference plane.

abbreviations	meaning	definition
ULz	upper lip vertical movement	d(m2,Ω)
LLz	lower lip vertical movement	d(m3,Ω)
ULy	upper lip frontal movement	d(m2,Δ)
LLy	lower lip frontal movement	d(m3,Δ)
RCx	right corner horizontal movement	d(m4,Σ)
LCx	left corner horizontal movement	d(m5,Σ)
Jz	jaw vertical movement	d(m6,Ω)
LOH	lip opening height (ULz-LLz)	d(m2,m3)
LOW	lip opening width (RCx-LCx)	d(m4,m5)
	velocities	$\partial p/\partial t$
	accelerations	$\partial^2 p/\partial t$

Table 1. Definition of ELITE articulatory measurements (see Fig. 4).

2.1 Vowels

2.1.1 Spatial characteristics

Among the various possible measurements provided by ELITE, the following visible articulatory parameters corresponding to phonologically significant features were analyzed for the vowels:

- *lip opening height* (LOH*),* calculated as the distance between the markers placed on the central points of the vermilion border of upper and lower lips. T his parameter may be correlated with the feature *high/low.*

- *lip opening width* (LOW), corresponding to the distance between the markers placed at the corners of the lips, this parameter correlates with the feature *rounded / unrounded.*

- *jaw opening* (Jz), corresponding to the distance between the markers placed at the center of the chin and the plane Ω (see Fig. 4) containing the line passing from the markers placed on the lobes of the ears and the tip of the nose. This distance is primarily due to the jaw opening, but it is also influenced by the movement of the skin of the chin. This parameter is correlated with the feature *high / low.*

- *anterior-posterior movement* of the *upper lip* (ULy) and *lower lip* (LLy), calculated as the distance between the markers placed on the central points of the vermilion border of either the upper or lower lip and the plane Δ (see Fig. 4) containing the line passing from the markers placed on the lobes of the ears and perpendicular to plane Ω (see Fig. 4). This parameter correlates with the feature *protruded / retracted.*

The lip and jaw movements of 6 subjects (4 females and 2 males), talkers of northern Italian, were recorded and analyzed. All the subjects were university students, aged between 19 and 22 and were paid volunteers. They repeated five times, in random order, each of the 7 stressed /a/, /ɛ/, /e/, /i/, /ø /, /o/, /u/ and the 5 unstressed, /a/, /e/, /i/, /o/, /u/, Italian vowels. The vowels were in the first syllable of disyllabic (/'tasti/, /'tɛsi/, /'tesi/, /'tisi/, /'tɔsko/, /'tosko/, /'gusto/) or trisyllabic (/ta'stare/, /te'stare/, /ti'sane/, /to'skane/, /gu'stare/) words, and were preceded by a /t/ and followed by /s/ (with two exceptions, i.e. /'gusto/ and /gu'stare/). They occurred within the carrier phrase “dico _____ chiaramente” (I say _____ clearly). Portions of the articulatory signal corresponding to the vowel to be analyzed were segmented on the basis of the acoustic speech signal synchronous recorded. Since the dynamic aspects of vowel production were not taken into consideration in this study, a single point characterizing the vowel, as illustrated in Figure 5, was individuated for each articulatory parameter. The data were normalized subtracting the values related to the position of the lips and jaw at rest, from each parameter obtained, for each vowel and each subject. This elaboration assured the comparability of the results independently of the subjects variability in the shape and size of the articulators. The data obtained with such normalization correspond to the real extension of the lip and jaw movements and may also be connected to data relating to the internal borders of the lips, cf. Abry and Boë (1986) and Benoit et al. (1992).

Following Magno Caldognetto, Vagges and Zmarich (1995), the analysis of the normalized data showed some interesting relations between all the parameters examined: a clear correlation between the *upper* and the *lower lip protrusion* for both stressed and unstressed vowels, (r=.82 and r=.83 respectively), and a negative correlation between *lip width* and *upper lip protrusion*,(r=-.81 and r=-.83), as well as between *lip width* and *lower lip protrusion* (r=-.73 and r=-.80). Stressed vowels showed also a correlation between *lip height* and *jaw opening* (r=.85). For both stressed and unstressed vowels there was no significant correlation between *lip height* and *lip width* and between *lip height* and *upper* or *lower lip protrusion*.

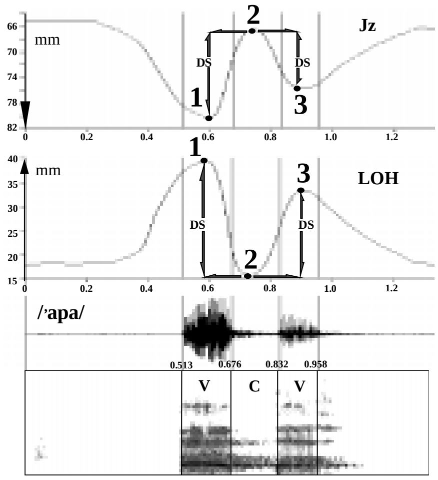


Figure 5. Definition of target points which characterize spatial characteristics of vowels and consonants for two particular articulatory parameters such as Jz and LOH. Points 1 and 3 refer to vowels while point 2 refers to consonants.

The presence or absence of correlation observed in this study for Italian is congruous with previous results reported for English by Fromkin (1964) and Linker (1982), and for French, by Abry and Boë (1986) and Benoit et al. (1992), independently of the instrumentation or reperee points used in defining the parameters. The normalized mean values, pooled over the 6 subjects and the 5 repetitions, for each parameter and each vowel, are reported in Table 2. The values may be either positive or negative depending on the parameter taken into consideration. For example, LOW values are negative when the distance between the corners of the lips decreases with respect to their distance at rest, as is evident for both stressed and unstressed /u/, while positive LOW values correspond to an increased distance with respect to the values at rest, as is the case of the unrounded vowel /i/ in both stressed and unstressed position. ULy and LLy may also show both positive and negative normalized mean values, while LOH and Jz are always positive.

		/i/	/e/	/ɛ/	/a/	/ɔ/	/o/	/u/
LOH	stressed	8.6	9.6	13.6	15.0	15.1	8.7	7.7
	unstressed	7.6	8.7		10.6		8.1	7.1
Jz	stressed	6.6	7.8	12.6	14.1	11.2	3.3	1.8
	unstressed	5.3	6.7		9.0		2.6	2.0
LOW	stressed	0.1	0.5	1.1	0.9	-5.1	-5.1	-6.1
	unstressed	1.2	0.8		1.2		-3.3	-5.3
ULy	stressed	-1.1	-0.3	-2.0	-2.1	2.6	3.9	4.4
	unstressed	-1.2	-1.0		-1.5		3.2	3.9
LLy	stressed	-1.4	-1.2	-3.3	-2.9	0.9	2.8	3.6
	unstressed	0.6	-1.1		-1.9		2.2	3.4

Table 2. Normalized mean values (mm) pooled over subjects (6) and repetitions (5), for each articulatory parameter and each vowel.

The results of two-way ANOVAs (7 or 5 vowels respectively and 6 subjects as a between factor) showed that *jaw opening* is the articulatory parameter that better distinguishes both stressed and unstressed vowels since it defines 4 degrees of jaw opening. In fact, Fig. 6 shows that stressed vowels are clustered in /u,o/, /i,e/, /ɛ,ɔ/, and /a/, and the unstressed are clustered in /a/, /e/, /i/ and /u,o/. LOH, which is traditionally considered to be parallel to *jaw opening*, does not identify all the degrees of opening defined by Jz (see Fig. 6).

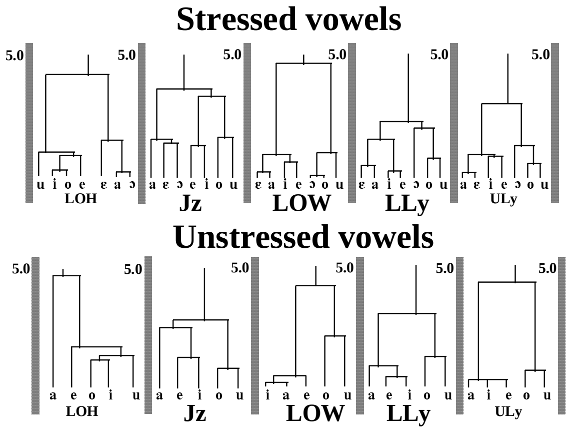
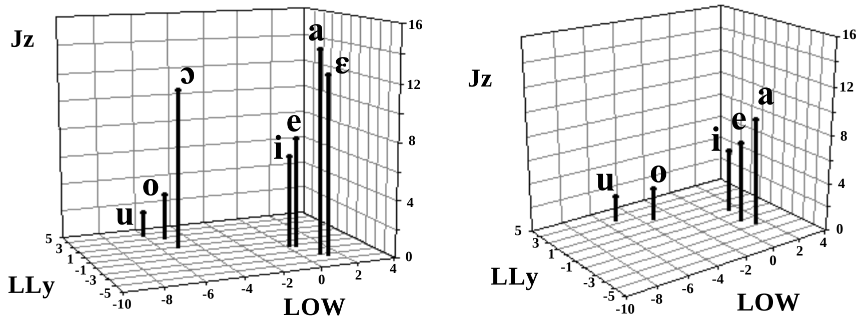


Figure 6. Hierarchical clustering of the stressed and unstressed vowels with respect to five articulatory parameters.

Moreover, the extension of its movement always shows greater values than Jz, cf. Table 2. It is clear that lips not only move in synergy with the jaw, but also in an independent specific manner for rounding and protrusion movements. LOW divides both stressed and unstressed vowels in two groups: rounded vowels and unrounded vowels (see Fig. 6). As for the two protrusions, LLy is the parameter that best distinguishes both stressed and unstressed vowels. As shown in Fig. 6, stressed and unstressed vowels are divided into 4 groups, i.e., two degrees of protrusion and two degrees of retraction. In particular, for stressed vowels, a higher degree of protrusion characterizes /u/ and /o/ with respect to /ɔ/, while /a/ and /ɛ/ are more retracted than /i/ and /e/, see Table 2. Based on the parameters, (*jaw opening, lower lip protrusion* and *lip width),* that distinguished the vowels in the most significant way, a three-dimensional representation of the stressed and unstressed vowel space was plotted in Figure 7. The data confirm the cooccurrence of rounding and protrusion for the vowels. In fact, all the vowels with positive values of *lip width*, /i,e,ɛ,a/, also have negative values for both *upper* and *lower lip protrusion*. That is, unrounded vowels are always also non protruded. Similarly, vowels with negative *lip width* values, i.e. the rounded vowels /ɔ,o,u/, are characterized by positive values of *upper* and *lower lip protrusion*, that is, they are also protruded. j*aw opening* and *lower lip protrusion* are the parameters that better distinguish the vowels. It should be noted that differences in *jaw opening* with respect to *lip height* may be due to the marker placed on the chin: the position of this marker was influenced not only by the jaw opening but also by the movement of the skin especially during the lip protrusion.

Based on the values of the parameters analyzed, the reduction of the unstressed with respect to the stressed vowels was confirmed. Moreover, the unstressed mid

vowels are more similar to the stressed mid-high /e/ and /o/ rather than to the mid-low stressed /ɛ/ and /ɔ/.



(a) **(b)**

Figure 7. 3D representation of the (a) stressed and (b) unstressed vowel space**.**

Based on the values of the parameters analyzed, the reduction of the unstressed with respect to the stressed vowels was confirmed. Moreover, the unstressed mid vowels are more similar to the stressed mid-high /e/ and /o/ rather than to the mid-low stressed /ɛ/ and /ɔ/.

2.2 Consonants

2.2.1 Spatial characteristics

The three-dimensional spatial targets were also elaborated for a significative subset of consonants: C = /p/, /b/, /f/, /t/, /d/, /s/, /ʃ/. To this purpose, the visible articulatory movements of 4 subjects were recorded. They pronounced 5 randomly ordered sets of the non-sense /'aCa/ symmetric sequences. The lips are involved directly as a primary articulator by some of the consonants like the bilabial /p/, /b/, or the labiodental /f/, while are not involved directly in the realization of some others like the alveolar /t/, /d/, /s/, or the mediopalatal /ʃ/, in which the lips move due to coarticulation or to coproduction. In particular, we considered the possible differences in movements due to the voiced/unvoiced opposition. Moreover the contrast /s/-/ʃ/ will be studied in order to verify the labialization of the palatal fricative (see for French, Abry and Boë, 1986). In order to characterize the 3D consonantal space, the LOH parameter, instead of Jz, was chosen together with LOW and LLy. This new parameter, highly correlated with Jz (r = 0.85) with respect to vowels, enables us to measure the possible lip compression. According to Figure 8 and Table 3, where the data of the three reference Italian isolated cardinal vowels are also illustrated, negative values of LOH compared to the rest position, characterize Italian bilabials /p/ and /b/. Similarly to the vowels, a full statistical analysis is in progress. Table 3 shows the mean values for the consonants, used to build Figure 8, computed in one target point for each consonants, such as point 2 illustrated in Figure 5.

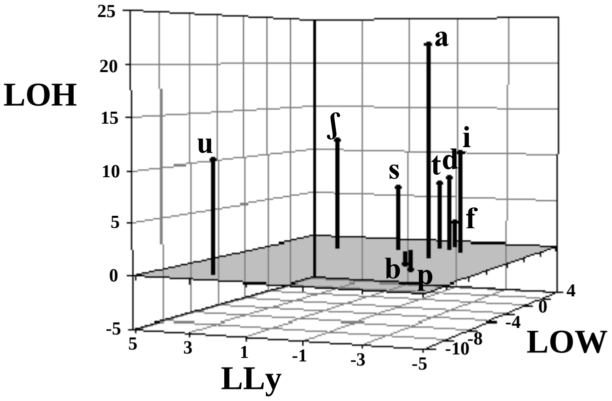


Figure 8. 3D representation of the consonantal space.

	/i/	/a/	/u/	/p/	/b/	/f/	/t/	/d/	/s/	/ʃ/
LOH	10.4	22.0	11.0	-2.2	-1.5	2.8	7.1	7.8	6.7	11.8
s.d.	(3.9)	(4.3)	(2.7)							
LOW	1.6	-1.4	-8.5	0.8	0.3	2.2	1.8	1.8	0.7	0.5
s.d.	(3.1)	(1.3)	(2.9)							
LLy	-2.0	-1.9	2.9	-0.4	-0.4	-1.6	-1.2	-1.6	0.1	2.4
s.d.	(2.0)	(2.5)	(1.1)							

Table 3. Normalized mean values (mm) pooled over subjects (4) and repetitions (5), for each articulatory parameter and each consonant. Isolated cardinal vowels parameters are also drawn for reference.

As for *opening* (LOH), the Table shows that the labial opening values increase from bilabial to mediopalatal consonants, that is following the degree of tongue retraction. It should be underlined that the degree of constriction of the vowels is always greater than that of the consonants and that there are negative values for the bilabials /p/ and /b/. In fact, when a bilabial plosive is generated, the lips get closer to themselves and also produce a certain degree of compression. Moreover, the pairs of omorganic voiced/unvoiced consonants are not distinguished by different values of LOH. As for *rounding* (LOW), all the examined consonants were spreaded compared to the rest position, in particular /f/, /t/, and /d/, whose values result similar to those obtained, for the same parameter, for the vowel /i/. As for *protrusion* (LLy), spreaded consonants are also retracted, while only /ʃ/ is characterized by a certain degree of protrusion and consequently it can be considered as a labialized palatal fricative. It is worth noticing that only a 3D description of consonantal targets enables to distinguish vowels from consonants and consonants among themselves. In the production of vowels, *rounding*, i.e. negative values of LOW, and *protrusion*, i.e. positive values of LLy, are concurrent, whereas in the production of consonants they can be independent. In fact, /ʃ/ is protruded but not retracted. Even for consonants similar for one feature, such as, for example, LOH for the alveolar /t/, /d/, and /s/, the differentiation is ensured by considering the other two remaining features LOW and LLy.

2.2.2 Spatial coarticulatory effects

Consonantal spatial targets are subject to possible variations depending on different contextual flanking vowels. In order to evaluate the relevance of these variations the effects of symmetrical vocalic contexts, constituted by the three cardinal vowels /a/, /i/ and /u/, on the labial movements of two different plosive consonants were examined. In particular the bilabial unvoiced plosive /p/, in which the lips constitute the primary articulator, and the apicodental unvoiced plosive /t/ where, on the contrary, the primary articulator is constituted by the tip of the tongue and the lips move due to coarticulatory effects, were considered. As illustrated in Figure 9 for /’VpV/ non-sense stimuli, flanking vowels determine different visual shapes for /p/. The context /u/ distinguishes itself, considering all three parameters, from the contexts /i/ and /a/, which instead result more similar. In particular a reduced compression (LOH) is evident for /p/ in the context /u/ due probably to the presence of the protrusion. Being /p/ the consonant showing the greater degree of labial constriction, a more significant difference between the data relating to the three contextual vowels and the consonantal target can be observed. Considering all three labial dimensions, relevant coarticulatory variations are noticeable also for /t/ in /’VtV/ non-sense stimuli (see Fig. 9). In the contexts /i/ and /a/ the degree of labial constriction for the consonant /t/ is evident and always lower than that of the corresponding stressed and unstressed flanking vowels. As regards LOW parameter, values for /t/ show a spreading effect in relation to each flanking vowel. Contexts /i/ and /a/ determine however greater LOW values than those pertaining to the context /u/. Also for protrusion, in the context /i/ and /a/ there is a clear effect of lip retraction for the consonantal targets, while context /u/ determine protrusion values similar to those of the vowels. Future research should obviously be devoted to study the coarticulatory effect of consonants on vowels and of asymmetric vowel context on consonants (for a review, see Farnetani, 1995).

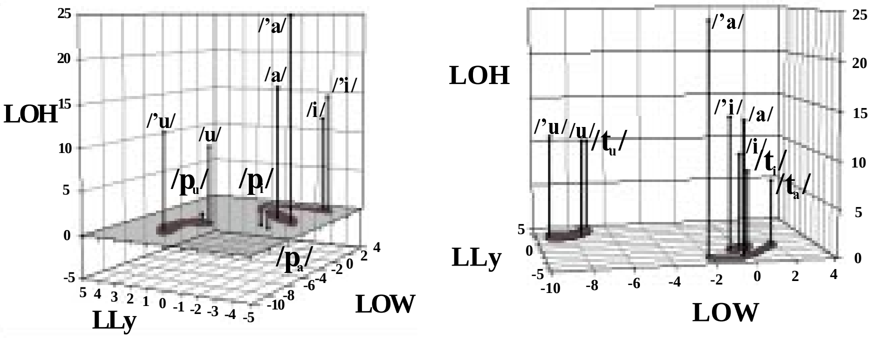


Figure 9. Coarticulation effect of flanking vowels on the target consonants (/’VpV/, /’VtV/ stimuli).

2.2.3 Temporal characteristics

The definition of dynamic characteristics of lips and jaw movements involved in the production of consonants is quite more complex and less clearly described than that of vocalic spatial targets, especially from a cross-linguistic point of view, even if articulatory phonetics produced a great amount of data and various interesting theories regarding this subject. In Figure 10 and Table 4 there is a description of some of the measurements which have been utilized in order to analyze the dinamics of the parameters and of the single articulators which could be relevant, not only for theoretical purposes, but also for the development of new *visual speech* synthesis and recognition applications.

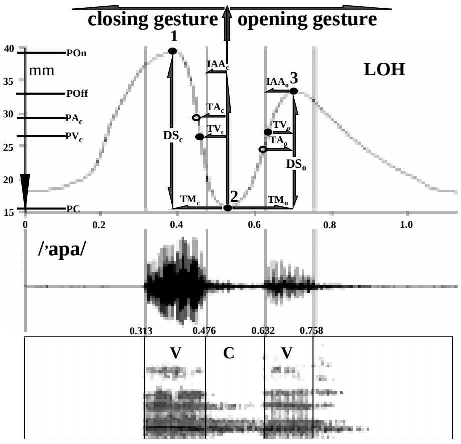


Figure 10. Target points and temporal definition for analyzing dynamic articulatory parameters.

POn	position of the marker at the onset of the closing movement
POff	position of the marker at the offset of the movement
PC	position of the marker at the peak closure
PV_c /PV_o	closure/opening peak velocity value
PA_c /PA_o	closure/opening peak acceleration value
DS_c /DS_o	closure/opening displacement (distance between onset/offset position and peak closing position)
TM_c /TM_o	closure/opening duration (time interval between the onset/offset of the movement and the peak closing position)
TV_c /TV_o	closure/opening velocity time (time interval between peak closing/opening velocity and peak closing position)
TA_c /TA_o	closure/opening acceleration (time interval between peak closing/opening acceleration and peak closing position)
IAA_c /IAA_o	closing/opening acoustic-articulatory delay (anisocrony) (time interval between the closing/opening acoustic segmentation landmark and the peak closing position)

Table 4. Some possible target points and some of the possible temporal definitions characterizing the dynamic characteristics of visible articulatory novements in the production of consonants.

These analyses are in progress for all the consonants. As an example, in Table 5 are illustrated the values obtained for the duration of the vowel-to-consonant closure (TM_c) and of the consonant-to-vowel opening (TM_o) movements relative to the LOH, LLy, LOW parameters for all the consonants previously utilized looking at the spatial characteristics (C = /p/, /b/, /f/, /t/, /d/, /s/, /ʃ/).

	/p/	/b/	/f/	/t/	/d/	/s/	/ʃ/		/p/	/b/	/f/	/t/	/d/	/s/	/ʃ/
LOH	**257**	**247**	**303**	**297**	**313**	**335**	**219**	**LOH**	**180**	**154**	**203**	**186**	**158**	**187**	**244**
LLy	**235**	**326**	**281**	**284**	**325**	**325**	**344**	**LLy**	**190**	**266**	**229**	**216**	**279**	**334**	**289**
LOW	**172**	**169**	**400**	**429**	**459**	**434**	**410**	**LOW**	**179**	**116**	**233**	**279**	**222**	**348**	**240**
				TM_c								**TM_o**			

Table 5. Mean values of the duration of the closure and opening movements.

Closing movements tend to be always longer than the opening ones. This effect can be related to the different prosodic characteristics of the initial (stressed) and final (unstressed) vowel which present different LOH values (cf. § 2). Reported values evidentiate also a different behaviour of the three parameters for each consonant.

In what follows, only an example of the presently available data will be presented, and precisely those relative to the vowel-to-consonant closure movements of the upper (UL) and lower lip (LL) for bilabial voiced and unvoiced stop consonants /p/ and /b/, produced 5 times by 4 talkers, within non-sense stimuli characterized by symmetric contexts (‘VCV: V= /i,a,u/) (Magno Caldognetto et. al. 1989). The data presented in Tables 6 and 7 refer to the mean values and standard deviations of some of the previously described spatio-temporal parameters, pooled over all the subjects and all the repetitions. The role of the voiced/unvoiced opposition and of the different vocalic contexts was investigated by the use of a series of three-way ANOVAs (3 vowels, 2 consonants and 4 subjects as a between factor) for each of the spatio-temporal measurements.

		/’apa/	/’ipi/	/’upu/			/’aba/	/’ibi/	/’ubu/
UL	POn mm	24.8	24.2	25.5	UL	POn mm	24.9	24.2	24.9
	s.d.	4.8	4.9	4.9		s.d.	4.7	4.9	4.5
	PC mm	27.7	27.7	27.9		PC mm	27.1	26.9	27.1
	s.d.	4.5	4.2	5.3		s.d.	4.4	4.4	4.9
	DS_C mm	2.9	3.1	2.4		DS_C mm	2.2	2.8	2.1
	s.d.	1.9	1.3	0.9		s.d.	1.4	1.1	0.7
	PV_C mm/sec	38.0	47.0	32.6		PV_C mm/sec	26.1	31.8	28.3
	s.d.	26.9	47.3	13.8		s.d.	13.3	13.9	10.5
LL	POn mm	58.8	50.3	48.6	LL	POn mm	59.1	50.1	47.8
	s.d.	5.0	6.3	4.1		s.d.	5.6	5.9	3.6
	PC mm	36.5	37.2	41.0		PC mm	37.9	38.1	41.9
	s.d.	4.5	3.7	4.5		s.d.	2.6	2.4	3.6
	DS_C mm	22.3	13.1	7.5		DS_C mm	21.2	12.0	5.8
	s.d.	4.2	4.8	2.3		s.d.	4.0	4.9	2.0
	PV_C mm/sec	273.1	153.0	80.4		PV_C mm/sec	264.0	150.1	70.0
	s.d.	41.8	61.8	26.1		s.d.	53.0	62.9	29.7

Table 6. Mean values and standard deviations, pooled over all subjects and repetitions, of the spatial characteristics of the upper and lower lips in non-sense ‘VCV stimuli (C=/p, b/, V=/a, i, u/).

			/’apa/	/’ipi/	/’upu/				/’aba/	/’ibi/	/’ubu/
UL	TM_c	msec	153	182	157	UL	TM_c	msec	169	187	143
		s.d.	71	72	47			s.d.	62	46	31
	TV_C	msec	56	67	76		TV_C	msec	60	77	65
		s.d.	26	46	33			s.d.	20	32	21
LL	TM_c	msec	207	185	199	LL	TM_c	msec	234	174	167
		s.d.	39	33	30			s.d.	61	32	32
	TV_C	msec	95	102	98		TV_C	msec	68	76	104
		s.d.	24	36	15			s.d.	12	13	118

Table 7. Mean values and standard deviations, pooled over all subjects and repetitions, of the temporal characteristics of the upper and lower lips in non-sense ‘VCV stimuli (C=/p, b/, V=/a, i, u/).

Both for the upper and lower lips, the temporal characteristics of the closing movement are not affected by the voiced/unvoiced contrast. In fact, neither the duration of the closing movement nor the time interval between peak velocity and peak closure are affected by the type of consonant. Our data on the duration of the closing movement are not in agreement with those reported by Summers (1987), who noted a longer duration for /b/ than /p/. As for the spatial characteristics, the voiced/unvoiced contrast does not affect the onset of the closing movement, but affects the peak closing position as well as the displacement. The data show that there is a greater degree of lip compression during the bilabial closure for the voiceless stop /p/ than for the voiced stop /b/. As it is shown in Table 6, the peak closure (PC) values show that the distance of the marker on the upper lip from the reference plane is always greater for /p/ than /b/. On the other hand, the peak closure values for the lower lip are always smaller for /p/ than /b/. This different degree of compression can be compared to the greater pressure of bilabial contact for word-final /p/ than /b/, as observed by Lubker and Parris (1970). Moreover, it can be compared to the reduced degree of linguopalatal contact noted for the voiced stop /d/ compared to the voiceless /t/ discovered by Farnetani (1989). The peak velocity values are also higher for /p/ than /b/, although this trend did not reach significance for the lower lip. Our data on the lip closure velocity are in agreement with the data obtained for the bilabial voiced and voiceless stops in word-final position, by Sussman, MacNeilage and Hanson (1973), Smith and McLean-Muse (1987), Summers (1987) and Flege (1988). As for the vowel context, the data show that it affects the duration of the closing movement, but not the time interval between peak velocity and peak closure position. With respect to the spatial characteristics, the onset of the closing movement depends on the vowel quality, for both the upper and lower lips. The peak closure position for the lower lip is affected by the flanking vowel, in particular, /u/ shows different peak closure values from /a/ and /i/. This difference may be due to the characteristic of lip protrusion which cooccurs with the rounding feature in Italian. The displacement values are, as expected, affected by the vowel quality, while the peak velocity values depend on the vowel context only with respect to the lower lip. In summary, the main effects of the voiced/unvoiced contrast on the bilabial closure movement are a greater degree of lip compression and a greater closure velocity for /p/ than /b/. The effect of the flanking vowels is evident in the different closure duration of the upper and lower lips and in the different closure velocity of the lower lip. Moreover, the time interval between the peak velocity and the peak closure is not affected by neither the voiced/unvoiced contrast nor by the vowel context. The asymmetric behaviour of most relevant parameters for upper (UL) and lower (LL) lip is visualized in Figure 11.

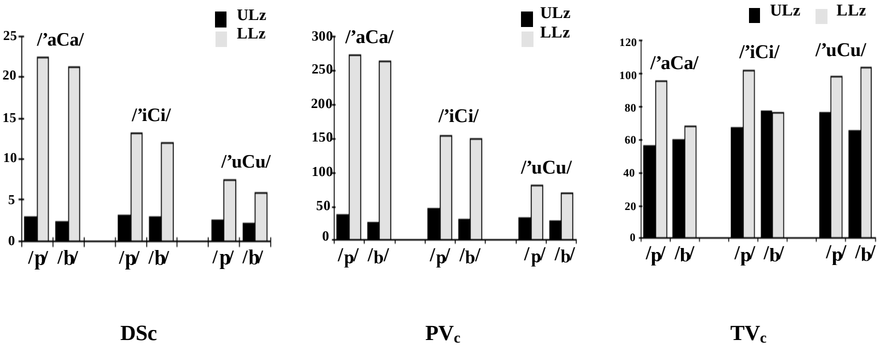


Figure .11 Upper and lower lip asymmetry in vertical movements.

It is evident, from previously presented data, the important role of the context in the definition of the variability of segmental targets and of the movements of single articulators. The design of a coarticulation model and the discovery of coarticulatory rules will be quite inportant for the future *visible speech* synthesis and recognition applications. In a previous research, reported in Magno Caldognetto et al. (1992), focused on the analysis of lip-rounding in bysillabic sequences of the type /ti'Cu/, /tiC'Cu/, and /ti'CCCu/, very strong anticipatory coarticulation effects were discovered, as exemplified in Figure 12.

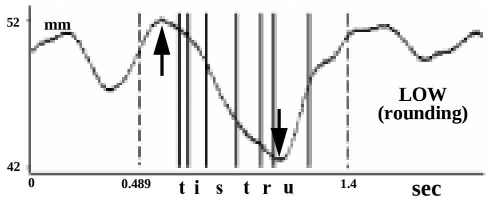


Figure 12. Anticipatory coarticulation effect for lip-rounding. Example for /ti'stru/ and LOW parameter.

The durations of the rounding movements differ depending on the type and number of intervocalic consonants, while the rounding movement appears to be independent of the presence of a syllabic boundary. Our data seems to be in agreement with the "look-ahead" (Perkell, 1980) rather than the "time-locked" (Fowler et al. 1980) theory because they do not appear to confirm the predictions of the intrinsic-timing control models, which predict an equal duration of the rounding movement before the beginning of the vowel /u/. In our data, rounding movements present different durations and they always begin during the occlusion of the initial consonant /t/ or during the front unrounded vowel /i/, and this rather supports a model of extrinsic-timing control. Future research should examine not only the duration of the rounding movement and the spatial characteristics of the maximum and minimum of the rounding target, but also the kinematic characteristics of the movement. In fact, maximum and minimum values can be accomplished using different strategies. Assessing the velocity and acceleration of the rounding parameter can provide evidence for the "hybrid" or "two-stage anticipatory coarticulation" model (Perkell, 1990). The between subjects variational behaviour should be underlined. In fact, irrespective of the model of speech production the data may point to, the specificity of individual strategies of motor control should not be ignored.

3 Bimodal Recognition of Italian Plosives

The articulatory data studied in the first section were used in conjunction with acoustic data in some Italian talker dependent and talker independent plosive classification experiments in order to verify a possible improvement of recognition performance, mostly in noisy conditions.

3.1 Introduction

Audio-visual automatic speech recognition (ASR) systems can be conceived with the aim of improving speech recognition performance, mostly in noisy conditions (Silsbee and Bovik, 1993). Various studies of human speech perception have demonstrated that visual information plays an important role in the process of speech understanding (Massaro, 1987), and, in particular, "lip-reading" seems to be one of the most important secondary information sources (Dodd and Campbell, 1987). Moreover, even if the auditory modality definitely represents the most important flow of information for speech perception, the visual channel allows subjects to better understand speech when background noise strongly corrupts the audio channel (MacLeod and Summerfield, 1987). Mohamadi and Benoit (1992) reported that vision is almost unnecessary in rather clean acoustic conditions (S/N > 0 dB), while it becomes essential when the noise highly degrades acoustic conditions (S/N <= 0 dB).

3.2 Method

The system being described takes advantage of jaw and lip reading capability, making use of ELITE (Magno Caldognetto et al., 1989) in conjunction with an auditory model of speech processing (Seneff, 1988) which have shown great robustness in noisy condition (Cosi, 1992). The speech signal, acquired in synchrony with the articulatory data, is prefiltered and sampled at 16 KHz, and a joint synchrony/mean-rate auditory model of speech processing (Seneff, 1988) is applied producing 80 spectral-like parameters at 500 Hz frame rate. In the experiments being described, spectral-like parameters and frame rate have been reduced to 40 and 250Hz respectively in order to speeding up the system training time. Input stimuli are segmented by SLAM, a recently developed semi-automatic segmentation and labeling tool (Cosi, 1993) working on auditory model parameters. Both audio and visual parameters, in a single or joint fashion, are used to train, by means of the Back Propagation for Sequences (BPS) (Gori, Bengio and

De Mori, 1989) algorithm, an artificial Recurrent Neural Network (RNN) to recognize the input stimuli. A block diagram of the overall system is described in Figure 13 where both the audio and the visual channel are shown together with the RNN utilized in the recognition phase.

3.3 Experiments

The results obtained in two phonetic classification experiments will be illustrated, the first one dealing with a Talker Dependent (TD) (Cosi, 1994), while the second with a Talker Independent (TI) environment (Cosi, 1995). For both experiments, the input data consist of disyllabic symmetric /’VCV/ non-sense words, where C=/p,t,k,b,d,g/ and V=/a,i,u/, uttered by 4 talkers (2 male and 2 female) in the TD case and by 10 male talkers in the TI condition. All the subjects were northern Italian university students, aged between 19 and 22, and were paid volunteers. They repeated five times, in random order, each of the selected non-sense words. The talker comfortably sits on a chair, with a microphone in front of him, and utters the experimental paradigm words, under request of the operator. In this study, the movements of the markers placed on the central points of the vermilion border of the upper lip, and lower lip, together with the movements of the marker placed on the edges of the mouth (markers 2, 3, 4, 5 of Fig 4. A total of 14 parameters, 7 movements plus their instantaneous velocity, constitute the articulatory vector which has been used together with the acoustic vector in order to represent the target stimuli. The chosen articulatory parameters were (see Fig. 4 and Table 1): ULz, LLz, Uly, Lly, LOH, LOW, Jz and velocities.

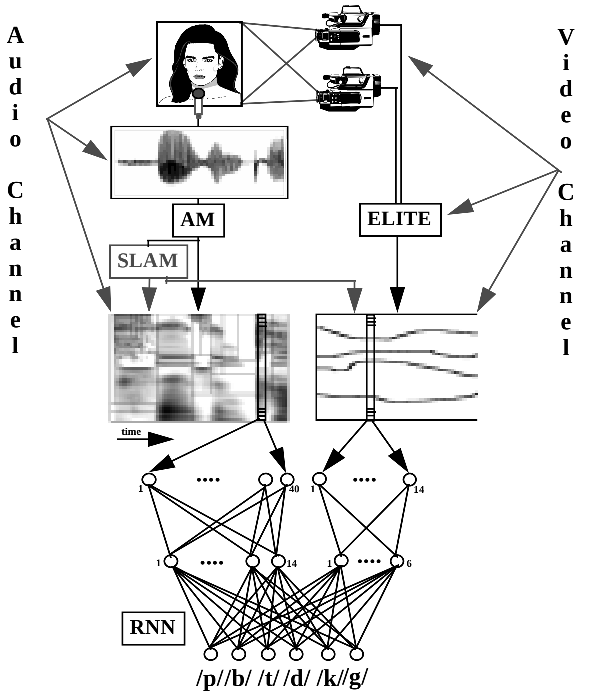


Figure 13. Structure of the bimodal recognition system.

As for the TD case two signal to noise condition were considered. A clean and a noisy condition with 0db S/N ratio. Also three situation were examined:

a) only the audio channel is active;
b) only the visual channel is active;
c) both audio and visual channel are simultaneously active

As for the Talker Independent environment, only the clean signal condition was explored and only the first and third situations in which only the audio channel or both audio and visual channels are simultaneously active were considered. In both TD and TI experiments, the network architecture which has been considered for the recognition was a fully or “two-part” (e.g. acoustic and articulatory input field of action were maintained disjoint) fully connected recurrent feed-forward BP network with dynamic nodes positioned only in the hidden layer. The structure of the RNN involved in the experiments is graphically illustrated in Fig 14.

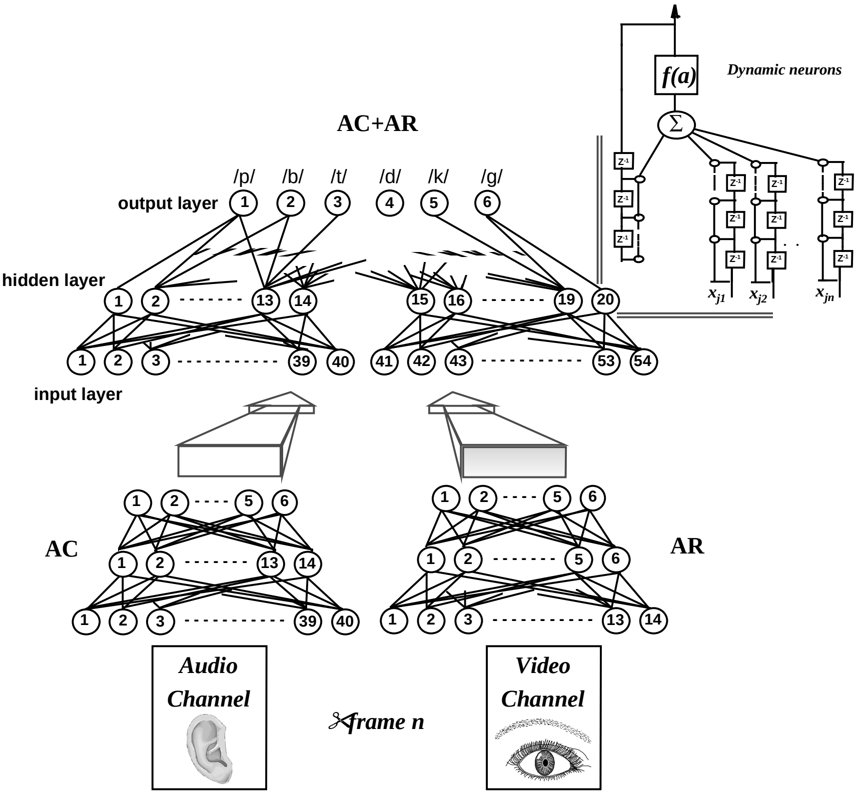


Figure 14. Network structures in the three different experimental settings: ACoustic, ARticulatory, ACoustic+ARticulatory (see text).

In order to use a learning algorithm which can be "local" in space and in time, thus reducing the computational complexity, (in other words, an algorithm which can operate on each neuron using only information relative to its connected afferent neurons, and using only the present input frame, utilizing information not explicitly related to previous frames) dynamic nodes were concentrated only in the hidden layer. In fact with this constraint the requested "local" conditions for the learning algorithm can be satisfied. The learning strategy was based on BPS algorithm (Gori, Bengio and De Mori, 1989), and only two supervision frames were chosen in order to speeding up the training time, as illustrated in Figure 15. The first one, focused on articulatory parameters, was positioned in the middle frame of the target plosive, the ‘closure’ zone, while the second, focused on acoustic parameters, was positioned in the penultimate frame, the ‘burst’ zone. A 20 ms delay, corresponding to 5 frames, was used for the hidden layer dynamic neurons. A 54(40+14)input * 20(14+6)hidden * (6)output RNN, as illustrated in Figure 14, was considered. Not all the connections were allowed from the input and the hidden layer, but only those concerning the two different modalities, which were thus maintained disjoint. Various parameter reduction schemes and various network structure alternatives were exploited but those described above represent the best choice in terms of learning speed and recognition performance.

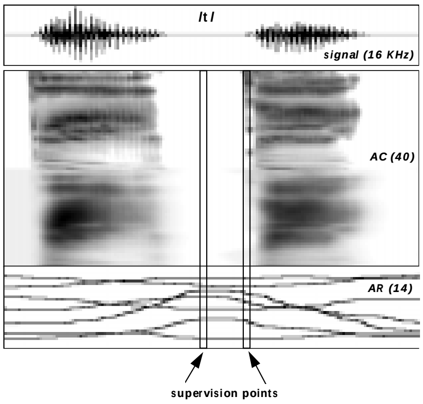


Figure 15. Two points supervision strategy.

3.4 Results

3.4.1 Talker dependent case

Table 8 refers to the results obtained in the “clean” speech experiment. "AC" refers to the experiment using only acoustic (auditory modeling) parameters, "AR" refers to the experiments using only articulatory (lips and jaw kinematics) parameters, while "AC+AR" refers to the simultaneous acoustic and articulatory case. The ar(pla) column refers to the grouping of similar phonemes given their "Place of Articulation". In other words /p,b/, /t,d/ and /k,g/ are considered as single classes.

talker	AC	AR	AR(PLA)	AC+AR
MA(m)	83	67	100	100
LI(m)	78	61	97	97
PA(f)	78	67	98	96
AN(f)	72	72	100	98
mean	78	67	99	98

Table 8. TD, “clean” condition, correct classification results (%)

It is immediately evident, analyzing the AR and AR(PLA) columns, that the articulatory net is quite poor in discriminate among all the plosives while on the contrary is quite good in classify the "PLA" classes. For all the four talker, the joint use of AC and AR parameters always improved the performance given by the AC parameters only. The 98% mean recognition rate is quite satisfactory given the difficult task. In Table 9 are illustrated the results obtained in the noisy speech experiment. The AR and AR(PLA) columns are obviously identical to the previous clean case, because acoustic noise do not corrupt articulatory data.

talker	AC	AC+AR
MA(m)	83	100
LI(m)	78	94
PA(f)	67	95
AN(f)	67	94
mean	74	96

Table 9. TD, “noisy” condition, correct classification results (%)

“AC” column shows a clear degradation of recognition performance on the respect of the clean case even if a 74% mean recognition rate cannot be considered terribly bad given the very critical noisy conditions (we "believe" that the auditory speech processing alone works better than classical techniques in noisy conditions). Like for the clean case, for all the four talker, the joint use of AC and AR parameters always improved the performance given by the AC parameters only, and the 96% mean recognition rate is quite similar to that of the clean speech case.

In order to qualitatively explain the network ability to well discriminate among the different classes of articulation (99 %), the influence of input articulatory parameters, in terms of activation of input nodes, on the correct outputs was measured, and various plots, as those illustrated in Fig. 16 and in Table 10 were built.

Looking at the data in the Table, referring to talker AN, obtained considering the learning set as the test set in order to have a big number of correct classifications, interesting qualitative deductions can be drawn. In particular it can be observed that a similar activation pattern represents the same plosive class. For example, high LOHv (lip opening height velocity) or LOWv (Lip Opening Width velocity) and low ULy (upper lip protrusion) well differentiate bilabial consonants /p/ and /b/ from the other two classes, as also Jzv (Jaw opening) well distinguishes apicodental consonants /t/ and /d/ from the other two classes. Similar plots, obtained for other talkers, show a similar tendency thus justifying the results obtained in the TD case. These observations obviously need further investigation which will be completed in the future. In particular, a more complete statistical description of articulatory data will be computed, so as to justify the hypothesized ability of the chosen RNN to identify the most valuable and reliable parameters for the PLA class discrimination.

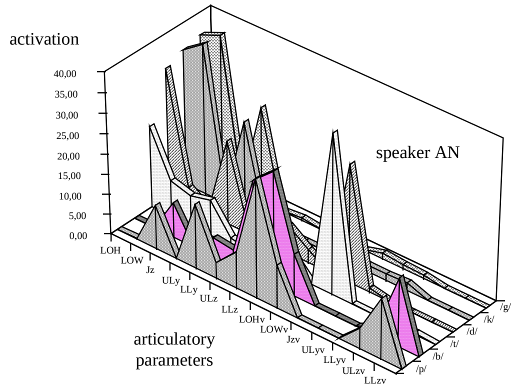


Figure 16. Input articulatory parameter influence for the plosives relatively to the talker AN (see text).

par/out	***/p/***	***/b/***	***/t/***	***/d/***	***/k/***	***/g/***
LOH	0,00	0,00	20,18	31,74	33,58	34,56
LOW	0,13	0,00	9,02	4,46	37,12	36,66
Jz	10,95	8,68	7,61	1,01	1,43	1,18
ULy	0,00	0,00	8,86	20,73	22,72	23,54
LLy	16,00	4,77	1,78	1,49	0,13	0,09
ULz	3,85	4,14	0,81	0,05	0,05	0,65
LLz	9,03	23,92	10,04	6,97	0,00	0,00
LOHv	29,44	28,48	5,35	3,26	0,00	0,00
LOWv	10,73	10,49	0,00	0,00	0,00	0,00
Jzv	0,00	0,00	39,59	29,25	1,56	1,49
ULyv	0,00	0,00	0,06	1,03	1,56	1,00
LLyv	0,00	0,00	0,00	0,00	1,85	0,82
ULzv	5,06	2,44	0,00	0,00	0,00	0,00
LLzv	15,05	17,08	0,33	0,00	0,00	0,00

Table 10. Input articulatory parameter influence for the plosives relatively to the talker AN.

3.4.2 Talker independent case

Two different experimental setting were considered in which, among the 10 talkers, 8 talkers were randomly picked up in order to form the learning set while the remaining two were considered as the test set. After having observed that a particular talker had a vary bad acquired audio signal, a third experiment was organized thus considering only 7 talkers for the learning set and two for the test set. The results for these three cases are illustrated in Table 11.

	E1	E2	E3
Talker 1	95.6	87.8	95.6
Talker 2	72.2	66.7	84.4
Mean	83.9	77.8	83.9

Table 11. TI correct recognition rate in three experimental settings with 8 talker for learning and 2 for testing in E1 and E2, and 7 talkers for learning and 2 for testing in E3.

In Table 12 the Talker-Pooled (TP) mean correct recognition performance for all the three experimental settings is illustrated. In this case each talker forming the learning set was also individually tested.

	E1	E2	E3
Mean	78.5	74.8	83.3

Table 12. TI mean correct recognition rate for the Talker-Pooled (TP) case.

In order to test the power of the bimodal approach all the three experiments were repeated eliminating visual information thus retaining only the audio channel input. The 40 input * 14 hidden * 6 output RNN utilized in this case is exactly the audio subnet of the global net utilized in the bimodal environment as indicated in Fig. 13. Results for this case are illustrated in Table 13.

	E1	E2	E3
Mean	68.9	58.3	65.0

Table 13. TI mean correct recognition rate with only Audio information.

2.4 Discussion

As indicated by a direct inspection of Tables 11-13, recognition performance significantly improves when both audio and visual channels are active. Looking at Table 12 referring to the talker-pooled results a good generalization power can be associated with the chosen RNN given that TI results were surprisingly better than TP results.

REFERENCES

Magno Caldognetto E., Vagges K., Ferrero F.E. and Cosi P. (1995) La lettura labiale: dati sperimentali e problemi teorici, *Proc. IV Convegno Nazionale Informatica Didattica e Disabilità*, Napoli, 9-11 Nov., 1995, (too be published).

Magno Caldognetto E., Vagges K. (1990a), Il riconoscimento visivo dei movimenti articolatori da parte di soggetti normali e ipoacusici. In *Scritti in onore di Lucio Croatto, Padova*, 1990, 153-166.

Magno Caldognetto E.,. and Vagges K. (1990b), Il riconoscimento delle consonanti in un test di lettura labiale, Atti del Congresso Nazionale della Società Italiana di Acustica, l'Aquila, 94-99.

Magno Caldognetto E., Vagges K. and Ferrero F.E. (1980), Un test di confusione fra le consonanti dell'italiano: primi risultati, *Atti del Seminario "La percezione* *del linguaggio"* (Firenze, 17-20 dicembre 1980), Accademia della Crusca 123-179.

Magno Caldognetto E., Ferrero F.E. and Vagges K (1982), Intelligibilità delle consonanti dell'italiano in condizioni di mascheramento (S/R), di filtraggio passa-alto (PA) e passa-basso (PB*), Bollettino Italiano di Audiologia e Foniatria*, vol. 5, 163-172.

Magno Caldognetto E., Vagges K. and Ferrero F.E. (1988), Intelligibilità e confusioni consonantiche in italiano, *Rivista Italiana di Acustica*, Vol. 12, 121-134.

Summerfield Q., (1987) Some Preliminaries to a Comprehensive Account of Audio-Visual Speech Perception, in Dodd B. and Campbell R. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading,* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 3-51.

Mohamadi T. and Benoit C., (1992) Apport de la vision du locuteur à l'intelligibilité de la parole bruitée en français, *Bulletin de la Communication Parlée*, n. 2, 1992, 32-41.

Benoit C., Lallouache T., Mohamadi T., and Abry C., (1992) A Set of French Visemes for Visual Speech Synthesis, in Bailly G., Benoit C., and Sawallis T.R. (Eds.), *Talking machines: Theories, Models, and Designs*, North-Holland, Amsterdam, 485-504.

Cohen M.M. and Massaro D., (1990) *Behavior Research Methods, Instruments and Computers*, Vol. 22 (2), 260-263.

Petajan E.D., (1984) Automatic Lipreading to Enhance Speech Recognition, *PhD Thesis*, Univ of Illinois at Urbana-Champaign.

Magno Caldognetto E., Vagges K., Borghese N.A., and Ferrigno G., (1989) Automatic Analysis of Lips and Jaw Kinematics in VCV Sequences, *Proc. of Eurospeech 1989*, Vol. 2:453-456.

G. Ferrigno and A. Pedotti, (1985) ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", *IEEE Transactions on Biomed. Eng.*, BME-32, pp. 943-950.

Abry C., and Boe L.J., (1986) "Laws" for Lips, *Speech Communication*, 5, 97-104.

Magno Caldognetto E., Vagges K. and Zmarich C., (1995) Visible Articulatory Characteristics of the Italian Stressed and Unstressed Vowels, Proc. of ICPhS 95, Stochkolm, 14-19 August, 1995, Vol. 1, 366-369.

Fromkin V., (1964) Lip Positions in American English Vowels, *Language and Speech*, 7, 217-225.

Linker W., (1982) Articulatory and Acoustic Correlates of Labial Activity in Vowels: A Cross-Linguistic Survey, *UCLA, Working Papers in Phonetics*, 56, 1-134.

Farnetani E., (1995) Labial Coarticulation, in *Quaderni del Centro di Studio per le Ricerche di Fonetica*, Vol. 13, 57-81.

Lubker J. and Parris P., (1970) Simultaneous measurements of intraoral pressure, force of labial contact, and labial electromyographic activity during production of the stop consonant cognates /p/ and /b/, *J. Acoust. Soc. Am*. 47, 1970: 625-633.

Farnetani E., (1989) V-C-V Lingual Coarticulation and its Spatiotemporal Domain, in Hardcastle W. J. and Marchal A. (Eds), *Speech Production and Speech Modelling*, 1989, Kluwer Academic Publishers, Dordrecht: 93-130.

Sussman H.M., MacNeilage P.F., and Hanson R.J., (1973) Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations, *J. of Speech and Hearing Research*, 16, 1973: 397-420.

Smith B.L., McLean-Muse A..(1987) Kinematic characteristics of post-vocalic labial stop consonants produced by children and adults, *Phonetica*, 44, 1987: 227-237.

Summers W.V., (1987) Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analysis, *J. Acoust. Soc. Am.*, 82 (3) 1987: 847-863.

Flege J., (1988), The development of skill in producing word-final English stops: Kinematic parameters, *J. Acoust. Soc. Am*., 84 (5) 1988: 1639-1652.

Magno Caldognetto E., Vagges K., Ferrigno G., and Busà G. (1992) Lip Rounding Coarticulation in Italian, *Proc. of International Conference on Spoken Language Processing*, Banff 1992, Vol. 1: 61-64.

Perkell, J.S., (1980) Phonetic Features and the Physiology of Speech Production in B. Butterworth (ed.), Language Production, Academic Press, London, Vol. 1, 337-372.

Fowler, C.A., Rubin P., Remez R.E. and Turvey M.T. (1980) Implications for Speech Production of a General Theory of Action, in B. Butterworth (ed.), *Language Production*, Academic Press, London, Vol. 1, 373-420.

Perkell, J.S., (1990) Testing Theories of Speech Production: Implication of Some Analyses of Variable Articulation Data, *Proc. NATO ASI, Speech Production and Modelling*, pp. 263-288.

Silsbee P.L. and Bovik A.C. (1993), Medium-Vocabulary Audio-Visual Speech Recognition, *Proc. NATO ASI, New Advances and Trends in Speech Recognition and Coding,* pp. 13-16.

Massaro D.W. (1987), Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry, Lawrence Erlbaum Associates, Hillsdale, New Jersey.
Dodd B. and Campbell R., Eds., (1987), Hearing by Eye: The Psychology of Lip-Reading, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

MacLeod A. and Summerfield Q. (1987), "Quantifying the Contribution of Vision to Speech Perception in Noise", *British Journal of Audiology*, 21 pp. 131-141.

Seneff S. (1988), "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing", *Journal of Phonetics*, 16, 1988, pp. 55 76.

Cosi P. (1992), Auditory Modelling for Speech Analysis and Recognition. In M. Cooke, S. Beet and M. Crawford (Eds.), *Visual Representation of Speech Signals*. John Wiley and Sons, pp.205-212.

Cosi P. (1993), "SLAM: Segmentation and Labelling Automatic Module", Proc. Eurospeech-93, Berlin, 21-23 September, 1993, pp. 665-668.

Gori M., Bengio Y. and De Mori R. (1989), "BPS: A Learning Algorithm for Capturing the Dynamical Nature of Speech", Proc. IEEE IJCNN89, Washington, June 18 22, 1989, Vol. II, pp. 417 432.

Cosi P., Magno Caldognetto E., Vagges K., Mian G.A. and Contolini M. (1994), “Bimodal Recognition Experiments with Recurrent Neural Networks”, Proceedings of IEEE ICASSP-94, Adelaide, Australia, 19-22 April, 1994, Vol. 2, Session 20.8, pp. 553-556.

Cosi P., Dugatto M., Ferrero F., Magno Caldognetto E., and Vagges K. (1995), Bimodal Recognition of Italian Plosives, Proc. 13th International Congress of Phonetic Sciences, ICPhS95, Stochkolm, Sweden, 1995.