STATISTICAL DEFINITION OF VISUAL INFORMATION FOR ITALIAN VOWELS AND CONSONANTS

Emanuela Magno Caldognetto, Claudio Zmarich, Piero Cosi

Istituto di Fonetica e Dialettologia, C.N.R., Padova

ABSTRACT

The aim of this research is to identify visual cues for Italian vocalic and consonantal visemes on a articulatory basis and to verify the results of visual intelligibility tests. These data will be useful when performing cross-linguistic comparisons, in defining some relevant visual parameters and variation ranges which could be used in the development of bimodal automatic speech synthesis and recognition systems.

1. INTRODUCTION

The identification of minimal units conveying visual linguistic information is essential in developing theories on audio/visual production and perception of speech [1, 2, 3] and also in supporting various technological applications [4] in telecommunications, in man-machine interaction or in language teaching and rehabilitation, such as bimodal audio/visual speech synthesis and recognition systems. With reference to the Italian language, a number of visual intelligibility perception tests for natural CV stimuli have identified the quality and quantity of the phonological information transmitted [5]. Assessment of the confusion errors recorded by the same test has enabled the identification of *homophenous* consonant groups whose visible articulatory movements are considered as being similar and therefore transmit the same phonological information. These groups have been given the name *visemes*. This kind of information is partially specific to the single languages owing to the fact that, despite the parallelism due to the high or reduced visibility of the different articulatory positions (i.e. front *vs.* back), there are also a number of idiosyncratic characteristics deriving from the different structure and dimension of the phonological inventories and from the phonotactic constraints, as demonstrated by the contrastive analysis for English [1, 3, 6], French [7] and Japanese [8]. At present, after having performed intelligibility tests of the natural visual stimuli, researchers aim to identify the optical cues used in the recognition processes. This has been the rationale for our analysis of a vast series of data regarding the visible lip and jaw articulatory movements for all the Italian vowels and consonants, and the definition of the relationships between these articulatory movements and the co-produced acoustic signal [9]. In this paper we present:

- the quantification of the labial orifice for all vocalic and consonantal targets on the basis of three parameters: lip height (LH), lip width (LW) and lower lip protrusion (LP);
- the identification, through statistical analysis (Cluster Analysis: Average Linkage Method, ANOVA and Linear Correlation) for each parameter, of groups of consonants and vowels characterized by non significantly different values, and therefore candidates for being perceived as homophenous;
- the definition of the phonetic variation for selected consonantal targets due to contextual vowels /a, i, u/.

These data, combined with the results of previous analyses on the dynamic characteristics and the articulatory-acoustic relationships [9], will be useful when performing cross-linguistic comparisons, in defining the most relevant visual parameters and variation ranges which could be used in the development of bimodal automatic speech synthesis and recognition systems.

2. METHOD

Instrumentation. In order to collect articulatory data, the ELITE system [9, 10, 11], a fully automatic real-time movement analyzer for 3D kinematics data acquisition was used. A synchronous recording of the acoustic signal was also obtained. On the basis of the analytical data referring to the upper lip (UL), lower lip (LL) and jaw (J), the following parameters were computed owing to their relevance in the definition of the area of the labial orifice [7] and thus for their connections with the corresponding acoustic signal:

- lip height (LH), i.e. the distance between the markers placed on the central points of

the UL and LL; this parameter may be correlated with the feature *high/low*;

- lip width (LW), corresponding to the distance between the markers placed at the corners of the lips, and correlated with the feature *rounded/unrounded* (or *spread*);
- anterior/posterior movement (protrusion) of UL (UP) and LL (LP), calculated as the distance between the marker placed on the central points of either the upper and lower lip and the frontal plane containing the line crossing the markers placed on the lobes of the ears. This parameter correlates with the feature *protruded/retracted.*

Data related to the Jaw parameter (J), though representing an important and stable index for both the degree of the oral cavity opening/closing and the syllabic vowel-to-vowel cycle [10], will not be dealt with in this paper.

Subjects and linguistic materials. For *vowels*, the visible articulatory movements of 6 subjects (4 females and 2 males), speakers of northern Italian, university students, aged between 19 and 22, were recorded and analyzed. These subjects repeated for a total of 5 times, in random order, each of the 7 stressed /a, ɛ, e, i, ɔ , o, u/ and the 5 unstressed, /a, e, i, o, u/, Italian vowels. The stressed vowels were in the first syllable of disyllabic words, the unstressed vowels in the first syllable of trisyllabic words and both types of words were embedded in carrier phrases [10]. The same subjects produced each of the isolated cardinal vowels /a,i,u/ 5 times. For *consonants*, all the 21 Italian consonants /p, b, m, f, v, t, d, n, s, z, ts, dz, ʃ, ʧ, ʤ, l, ʎ, ɲ, k, g, r/ were pronounced in the vocalic symmetric context /'aCa/, and repeated 5 times by 4 subjects, 2 female and 2 male university students and talkers of northern Italian. The same subjects repeated a selected group of consonants in the /'iCi/ and /'uCu/ context 5 times each.

Portions of the articulatory signal corresponding to the vowels and consonants to be analyzed were segmented on the basis of the acoustic speech signal. The single point characterizing vocalic and consonantal targets (i.e. the minimum or maximum value, depending on the parameter) was identified for each articulatory parameter. It is worthwhile noticing that, with reference to each parameter, these target values were normalized by subtracting the values related to the position of the lips at rest. This assured the comparability of the results independent of the subject variability in the shape and size of the articulators. The data thus obtained correspond to the real extension of the lip movements correlated to the data relating to the internal borders of the lips [7]. The values may be either positive or negative: for example, LW values are negative when the distance between the corners of the lips decreases with respect to their distance at rest, i.e. lips are rounded and *viceversa*. UP, LP, and LH may also show both positive and negative normalized mean values.

3. RESULTS

Vowels. The normalized mean values (mm) of stressed, unstressed and isolated vowels are reported in the 3D representation (Fig. 1) based on LH, LW, LP parameters.

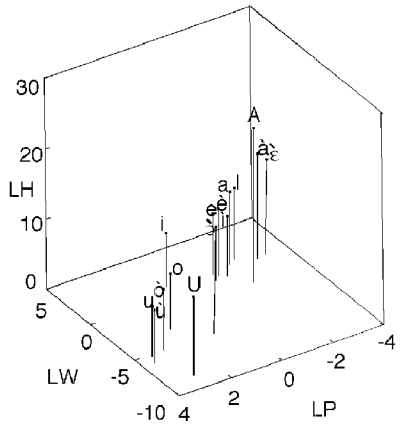

Figure 1. Spatial configurations of the labial orifice for the 7 stressed vowels /'i,'e,'ɛ,'a,'ɔ,'o,'u/, the 5 unstressed vowels /i,e,a,o,u/ and 3 isolated cardinal vowels (I, A, U in the fig.), based on LH, LW and UP values (mm).

LH presents only positive values, (with a 14.9 mm variability range), from 7.1 mm for the unstressed /u/ to 22 mm for isolated /a/. LW ranges from 1.6 mm for the most spread vowel, the isolated /i/ to –8.5 mm for the most rounded vowel, the isolated /u/, with a variability range of 10.1 mm. LP ranges from 3.6 mm for the most protruded vowel, the stressed /u/, to –3.3 mm for the most retracted vowel, the stressed /ɛ/ (a 6.9 mm variability range). The choice of these parameters is justified by the following statistical correlation analysis: for both stressed and unstressed vowels, the two parameters UP and LP show a clearly positive correlation ($p < .001$, $r = .82$ and $r = .83$ respectively), while there is a negative correlation between LW and UP, ($r = -.81$ and $r = -.83$), as well as between LW and LP ($r = -.73$ and $r = -.80$). It is worthwhile noticing that this negative correlation, in terms of phonetic features, corresponds to positive correlation between rounding and protrusion. Neither stressed nor unstressed vowels show any significant correlation between LH and LW and between LH and UP or LP. As to isolated vowels, LP correlates significantly with UP ($r = .85$), while there are two negative correlations between LW and UP ($r = -.90$) and between LW and LP ($r = -.81$) and

no correlation between LH and the other parameters. Therefore the resulting correlations are verified in every condition. In the following analysis, preference has been given to LP rather than UP, as its values are always greater, and more prominent, than UP values.

Stressed and unstressed vowels were analyzed with two-way ANOVAs (7 or 5 vowels respectively and 6 subjects as a between factor) to assess their effect on each of the articulatory parameters examined. Multiple *post hoc* Tukey comparisons were carried out when the vowel effects were significant. Only the data significant at $p<.01$ will be discussed.

As for the stressed vowels, LH identifies 3 groups of vowels: /i, e, u, o/: close and close-mid vowels (front and back); /a, ɔ/: open central and open-mid back vowels; /ɛ /: open-mid front vowel. LW divides stressed vowels in 2 groups: /i, e, ɛ, a/ unrounded vowels and /u, o, ɔ/, rounded vowels. LP divides the stressed vowels into 4 groups, characterized by two degrees of retraction and two degrees of protrusion: /i,e/ close and close-mid front vowels; /ɛ,a/ open-mid front and open central vowels; /o,u/ close and close-mid back vowels; /ɔ/ open-mid back vowel.

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 [10]). As for the unstressed vowels, LH distinguishes /a/, from all the close and close-mid /o, u, i, e/ vowels, whereas LW divides them in two groups: /i, e, a/ unrounded vowels, and /u, o/ rounded vowels, and LP identifies the 4 groups (as in the stressed condition): /i, e/, /a/, /o/, /u/.

When the results for the three different conditions affecting the cardinal vowels are compared, LH shows a systematic tendency to reduce the values from isolated to stressed and finally to unstressed conditions. Even for LW the tendency is that of producing the maximal contrast in the isolated condition, while for LP, the maximal value of protrusion is achieved by the stressed rather than the isolated condition.

Consonants. The analysis of visible articulatory characteristics of consonantal targets begins with the /'aCa/ context, because this context produced the greatest number of correct recognitions in the visual intelligibility perception tests [5]. Fig. 2 represents the three-dimensional co-ordinates (LH, LW, LP) defining the labial orifice for all the 21 Italian consonants and averaged along all the subjects' productions. As with vowels, the choice of LP was dictated by the results of the correlation analysis and the magnitude of values.

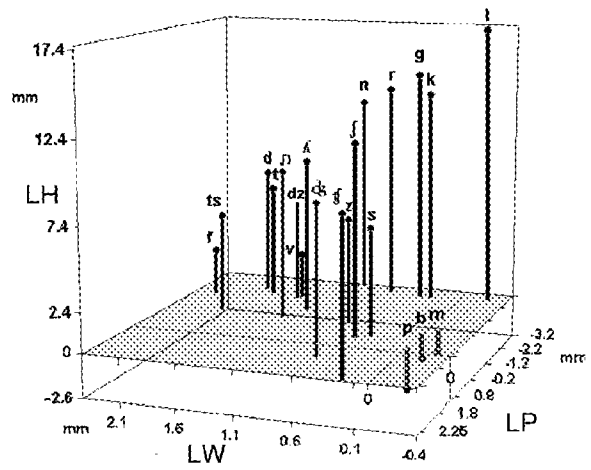

Figure 2. Spatial configurations of the labial orifice for the 21 Italian consonants in the context /'aCa/ based on LH, LW, and UP values (mm) .

- LH: only three consonants, /p, b, m/, produced negative values determined by the compression of the lips performing the bilabial closure. Minimum positive values were recorded for /f, v/, and maximum positive values for /l/;
- LW: the highest value (maximum spreading) characterizes /f/. The lowest value (maximum rounding) was achieved in the production of /ʧ/;
- LP: only /ʧ, ʤ/ were protruded, whereas the most retracted consonant was /n/.

A correlation test (r-ratio, Pearson, $p = .001$) applied to these data revealed the following results:

- for 10 consonants (i.e. /m, d, ts, dz, z, l, n, r, ʧ, k/) there was no significant correlation between the three parameters;
- a positive correlation was found between LH and LW for the consonants /t, ʃ, g/; between LH and LP for /ʤ/ and between LW and LP for the consonant /s/;
- a negative correlation was determined between LH and LW for /p, b, f, v, ʤ/ and between LW and LP for /ɲ, ʎ/.

As the three parameters were found to be correlated two-by-two only for /ʤ/ (i.e. LH and LP by positive correlation, LH and LW by negative correlation), these results appear to prove that the three parameters LH, LW and LP contribute independently in the

definition of the characteristics of the labial orifice for Italian consonants.

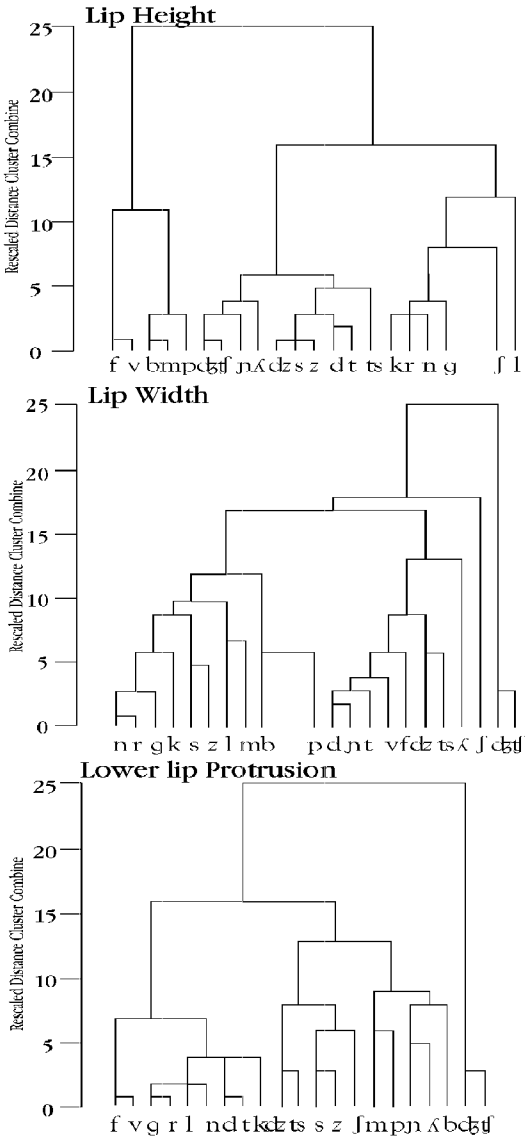

Figure. 3: Consonant groups calculated by using the Average Linkage Method with reference to the articulatory parameters LH, LW and LP.

In subsequent analyses, it was therefore essential to take into consideration all three parameters separately. Relative clusters were obtained (Figure 3) from all the data referring to all the consonants and for each of the three parameters, by applying the Averaged Linkage Method between Groups, which converts the numerical similarity between the mean consonant values for each parameter into spatial or geometrical distance proximity, so that consonants that are most similar to each other in terms of real values are grouped closer together. The scale 0-25 then provides an indication of relative distance values between the groups.

An ANOVA test with consonant type as a within factor was performed, followed by planned comparisons between consonant groups, using Scheffè method ($p<.001$). The parameter which best distinguishes the consonants from each other is LH, which makes it possible to identify 6 groups, each characterized by an articulatory variability range that is considerably different from that of the others. Minimum and maximum values (expressed in mm) of LH that define the phonetic boundaries of classes thus identified are as follows:

- /p, b, m/: -2.535 / -1.273
- /v, f/: 2.652 / 2.844 ;
- /ts, dz, z, s, t, d/: 5.879 / 7.808;
- /ʧ, ɕ, ʎ, ɲ,/: 8.798 / 9.553;
- /ʃ, n, k, r, g/: 11.716 / 14.443;
- /l/: 17.385.

With regard to LW,3 groups gave significant results:

- /ʧ, ɕ,/: -0.280 / 0.158;
- /l, m, b, k, g, s, z, r, p, n/: -0.166 / 1.022;
- /ʃ, ʎ, ɲ, v, dz, t, d, ts, f/: 1.078 / 2.185.

Concerning LP, 4 groups gave significant results:

- /n, l, r, g, k, d, t/: -3.128 / -2.041;
- /v, f/: -1.862 / -1.638;
- /ʎ, ts, dz, ɲ, z, m, b, p, s, ʃ/: -1.050 / 0.465;
- /ʧ, ɕ/: 2.128 / 2.294.

The above data reveals the ranges of the 3 parameters LH, LW and LP to be quite different from each other: i.e. 19.910 mm, 2.465 mm, and 5.522 mm, respectively. Besides identifying the greatest number of groups, LH keeps them more distinct from each other, due to a 3 mm separation on the edges of each group. For the narrowest parameter (LW), with only 3 classes, the reduced difference between the adjacent groups needs to be perceptively validated as to its relevance. Comparing the clusters referring to LH, LW and LP with the clusters obtained by the visual intelligibility test for all consonants coarticulated with /a/ [5] (see Ca in Fig. 4), evidence is given that the 6 visemes identified by the perception test coincide with those determined for the parameter LH.

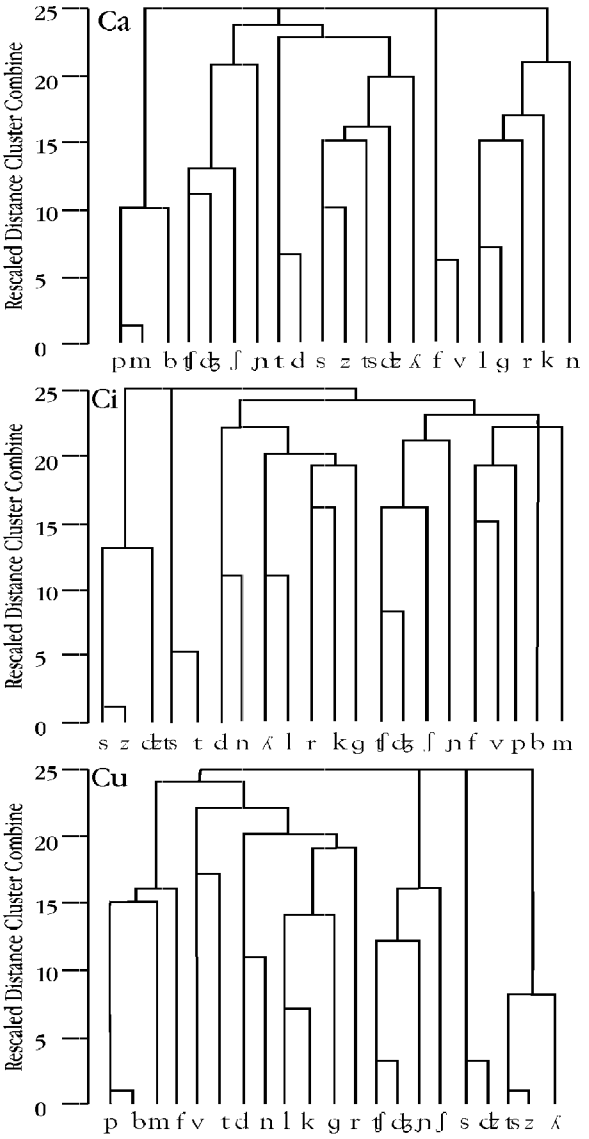

Figure 4. Consonant groups calculated by using Average Linkage Method with reference to intelligibility tests scores of natural visual stimuli for all 21 Italian consonants coarticulated with /a, i, u/ vowels.

The only differences regard the positioning of /ʃ, ʎ, l/ in differing visual clusters, the identification of which is most likely made with the contribution of the parameters LW and LP. For both the series, clusters are organized according to the place of articulation: bilabials, labiodentals, dentals (except for /n/), postalveolars and palatals. An exception is made for the viseme composed of /k, g, r, l/. These data would appear to justify the adoption of articulatory parameters for the definition of those homophenous consonant categories, or visemes, that have been primarily identified on the basis of perception tests.

Coarticulatory effects. The results of visual intelligibility tests for natural CV stimuli [5] demonstrated the tendency for correct recognition of the consonants to be reduced from /a/ (31%), to /i/ (28%) and finally to /u/ (25%) context, as well as for the visemes to change number and internal structure according to the subsequent vowel (see Ci and Cu in Fig. 4). In order to explain these results, the effects of the vocalic context on the LH, LW and LP values of a selection of consonants was measured. The consonants are those best identified within each visual cluster. For context /a/: /p, f, d, s, ʧ, l/; for /i/: /p, f, t, s, ʧ, l, ʎ/; and finally for /u/: /p, f, s, ʤ, l/.

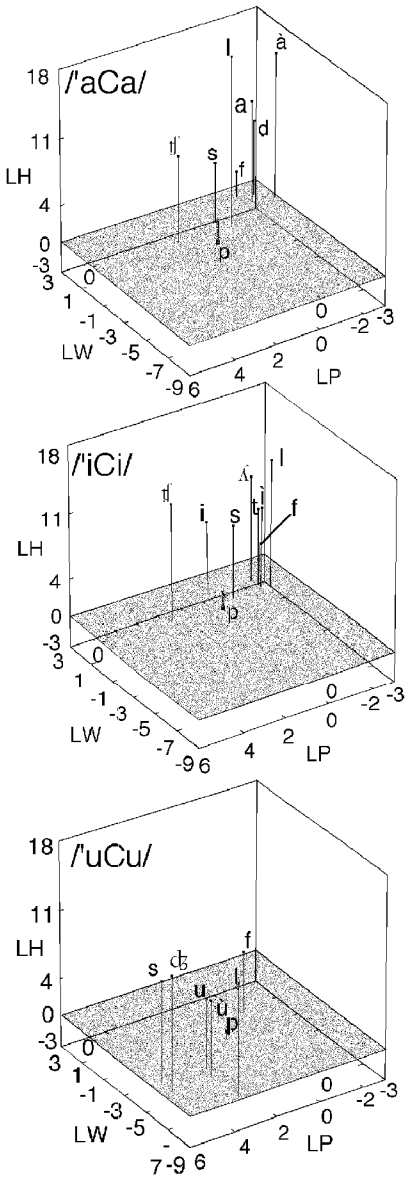

Figure 5. Spatial configurations based on LH, LW, and UP values (mm) of the labial orifice for the stressed and unstressed cardinal vowels and the Italian consonants best identified within each visual cluster in the /Ca/, /Ci/, and /Ca/ contexts.

Figure 5 represents the 3D positions (for LH, LW and LP) characterizing the labial orifice in the articulation of the selected consonant targets, together with the related stressed and unstressed vowel targets. As for LH, consonants are distinguished independently from the vocalic contexts. In the /u/ context, all the consonant are characterized by the presence of rounding and protrusion (the highest LP values are achieved for /s/ and /ʤ/, whereas the minimum value is achieved for /f/). In the /a/ and /i/ contexts all the consonant are unrounded and retracted, except for the protruded

/ʧ/. An Anova model, applied to /p, f, s, l/, i.e. the 4 consonants common to the 3 sets of best identified consonants of Fig. 5, and to the vocalic context /a, i, u/ as within factors, was performed for each parameter in order to verify the influence of the vocalic context on the consonant values. The results were highly significant ($p<.001$) for the main factors involved, and, more importantly, for the interactions between the factors (LH: $F(6,114)=40.133$); LW: $F(6,114)= 36.355$; LP: $F(6,114)=30.046$). Multiple comparisons were applied to each consonant to detect significant variations due to the vowel context (Scheffe' method, $p<.05$). For LH, the only significant comparison is /'ulu/ *vs.* /'ala/, while for LP there are significant differences for /f, l, s/ in the /u/ context *vs.* the /i/ and /a/ contexts. For LW, all the comparisons between the consonants in the /u/ and the /i,a/ contexts are significant. Looking at the numerical values, the LH parameter distinguishes the consonants from each other within the /a/ and /i/ contexts better than in that of /u/, while LW and LP distinguish the consonants from each other in the /u/ context better than in the /a/ and /i/ contexts.

4. CONCLUSIONS

The data presented in this paper reveal the fact that the definition of the labial orifice needs to be performed in terms of 3 parameters, as LH, LW and LP are independent, when referring to consonants, whereas for vowels LW and LP are correlated. The ranges of the 3 parameters are all different from each other (LH > LP > LW), and all 3 parameters present normalized positive and negative value variation with reference to the position at rest, with the exception of LH, which yields purely positive values for vowels. The identification of statistically significant value variation ranges for each single parameter is the first step in the identification of indexes that may be exploited in the recognition of visible articulatory movements. However, their role as indipendent characteristics or as indexes for holistic computation of the labial area still requires further discussion. Vowel-consonant co-production within a phonetic chain determines significantly different variations in the 3d consonant targets for which it will be necessary to devise a number of articulatory rules [12] in future researches. All these data may be used to specify the characteristics of visible articulatory movements for the Italian language, which will then need to be taken into account when summarizing and performing automatic bimodal audiovisual recognition.

5. REFERENCE

1. Summerfield, Q., "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 Ass., Hillsdale, NJ, 3-51, 1987.
2. Massaro, D.W., "Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry", in Dodd B. and Campbell R. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*, Lawrence Erlbaum Ass., Hillsdale, NJ, 53-83, 1987.
3. Massaro, D.W., "Bimodal Speech Perception: A Progress Report", in D.G. Stork and M.E. Hennecke (cit.), 79-102, 1996.
4. Stork, D.G. and Henneke, M.E. (Eds.), *Speechreading by Humans and Machine: Models, Systems and Applications*, NATO ASI Series F: Computer and Systems Sciences, Vol. 150, 1996.
5. Magno-Caldognetto, E. and Vagges, K., "Il riconoscimento delle consonanti in un test di lettura labiale", *Atti del XVIII Convegno Nazionale dell'Associazione Italiana di Acustica*, Lecce, 94-99, 1990.
6. Auer, E.T., Bernstein, L.E., Waldstein, R.S. and Tucker, P.E., "Effects of Phonetic Variation and the Structure of the lexicon on the Uniquess of words", in C. Benoit and R. Campbell (Eds.), *Proceedings of the Workshop on Audio-Visual Speech Processing*, Rhodes, 21-24, 1997.
7. Benoit, C., Lallouache, T., Mohamadi, T. and Abry, C., "A Set of French Visemes for Visual Speech Synthesis", in G. Bailly, C. Benoit, and T.R. Sawallis (Eds.), *Talking Machines: Theories, Models, and Designs*, North-Holland, Amsterdam, 485-504, 1992.
8. Hiki ,S. and Fukuda, Y., "Negative Effects of omophone on speechreading in Japanese", in C. Benoit and R. Campbell (Eds.), *Proceedings of the Workshop on Audio-Visual Speech Processing*, Rhodes, 9-12, 1997.
9. Magno-Caldognetto, E., Zmarich, C., Cosi, P. and Ferrero, F., "Italian Consonantal visemes: relationships between spatial/temporal Articulatory Characteristics and Coproduced Acoustic Signal", in C. Benoit and R. Campbell (Eds.), *Proceedings of the Workshop on Audio-Visual Speech Processing*, Rhodes, 5-8, 1997.
10. Magno-Caldognetto, E., Vagges, K. and Zmarich, C., "Visible Articulatory Characteristics of the Italian Stressed and Unstressed Vowels", *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, vol. 1, 366-369, 1995.
11. Cosi, P. and Magno-Caldognetto, E., "Lips and Jaw Movements for Vowels and Consonants:Spatio-Temporal Characteristics and Bimodal Recognition Applications", in D.G. Storke and M.E Henneke (cit.), 291-313, 1996.
12. Vatikiotis-Bateson, E., Munhall, K.G. and Hirayama, M., "The Dynamics of Audiovisual Behavior in Speech", in Stork D.G. and Hennecke M.E. (cit.), 221-232, 1996.