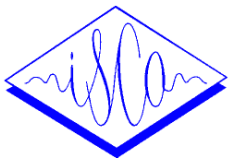

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

ITALIAN CONSONANTAL VISEMES: RELATIONSHIPS BETWEEN SPATIAL/ TEMPORAL ARTICULATORY CHARACTERISTICS AND COPRODUCED ACOUSTIC SIGNAL

E.Magno Caldognetto, C. Zmarich, P. Cosi and F. Ferrero
Centro di Studio per le Ricerche di Fonetica –C.N.R.
Consiglio Nazionale delle Ricerche
Via G. Anghinoni, 10 – 35121 Padova (ITALY)
Phone: +39 49 8274421, FAX: +39 49 8274416, E-mail: cosi@csrf00.csrf.pd.cnr.it

ABSTRACT

In order to identify the Italian consonantal visemes, to verify the results of perceptive tests and elaborate rules for bimodal synthesis and recognition, the 3D (lip height, lip width, lower lip protrusion) lip target shapes for all the 21 Italian consonants were determined. Moreover, the spatio-temporal characteristics of the closure/opening movements for the realisation of these consonantal targets were studied relative to the lip height (LH) parameter together with the temporal relationships between the characteristics of this articulatory movement and the co-produced acoustic signal.

1. INTRODUCTION

In order to develop theories on audio/visual production and perception of speech (Summerfield 1987, Massaro 1987,1996) and also to support various technological applications in telecommunications (Stork and Hennecke 1996), in man machine interaction or in language teaching and rehabilitation, such as bimodal audio/visual speech synthesis (Benoit et al. 1992, Cohen & Massaro 1990) and recognition systems (Petajan 1984, Stork et al. 1992, Silsbee & Allen 1993, Adjoudani & Benoit 1995), it is essential to:
- identify the minimal units conveying visual linguistic information (visemes), i.e. on the basis of articulatory parameters, specifying which are the consonants belonging to each of them;
- determine the relationships between the visible articulatory movements and the corresponding co-produced acoustic signal, i.e. determine the iso- or aniso-morphism between articulatory movements and their correspondent acoustic product.

This kind of information is partially language specific because, even though significant and expected cross-linguistic parallelisms are present due to the high versus low visibility of anterior versus posterior articulation loci, language specific characteristics arise due to the different size and structure of the phonological inventories (Walden et al. 1977, Cohen et al. 1996, for English visemes, and Benoit et al. 1992, 1996, for French visemes).

2. METHOD

In order to collect all these data, the ELITE system (Magno Caldognetto et al. 1989, 1995), a fully automatic real-time movement analyser for 3D kinematics data acquisition (Ferrigno and Pedotti 1985) was utilised. In particular we focused on the movements of markers attached to the upper lip (UL), lower lip (LL) and jaw (J). All the 21 Italian consonants /p, b, m, f, v, t, d, n, s, z, ts, dz, S, tS, dZ, l, L, J, k, g, r/ (SAMPA coding, see Fourcin et al. 1988) were pronounced in a vocalic symmetric context /'aCa/, and repeated 5 times by 4 subjects, 2 female and 2 male university students, talkers of northern Italian. On the basis of the analytical data referring to UL, LL and J, the following parameters were computed owing to their relevance in the definition of the area of the labial orifice (Benoit et al. 1992) and thus for their connections with the corresponding acoustic product:
- lip height (LH), calculated as the distance between the markers placed on the central points of the upper and lower lips; 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, a parameter which correlates with the feature rounded/unrounded;
- anterior/posterior movement (protrusion) of the upper lip (UP) and lower lip (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 (JL), even though they represent an important and stable index for both the degree of the oral cavity opening/closing and the syllabic cycle vowel-to-vowel, will not be presented in this paper but they will be discussed in next studies.
For the three parameters LH, LW and LP, the values corresponding to the consonantal targets were identified (see point '2' in Figure 1). It is worth noticing that, for each parameter, these target values were normalised by subtracting the values related to the position of the lips at rest. This assured the

comparability of the results independent of the 4 subject variability in the shape and size of the articulators. The so obtained data correspond to the real extension of the lip movements which are correlated to the data relating to the internal borders of the lips (Benoit et al. 1992). Moreover for LH parameter, the spatial and temporal characteristics, as illustrated in Figure 1, were analysed for all the 21 Italian consonants in terms of:

- displacement, i.e. the extension of the closing (initial stressed vowel-to-consonant)/opening (consonant-to-final unstressed vowel) movement (DSc/DSo);
- duration of the possible steady state (plateau), occurring when the velocity curve crossed the zero-axis more than once within a single gesture, and the max. value in the curve lying between the first and final cross-points was lower than one tenth of the maximal velocity along the same direction of movement;
- duration of the closing/opening movement, measured as the time interval between the onset of the movement and the peak closing/opening position (TMc/TMo);
- duration of the whole closing/opening cycle;
- consonantal acoustic duration;
- temporal relationships between the articulatory movements and the corresponding acoustic product, measured as the time interval between the acoustic onset of the consonant/second vowel and the peak closing/opening position (IAAc/IAAo);
- temporal relationships between the articulatory movements and the corresponding peak closing/opening velocity, measured as the time interval between onset of closing/opening movement and closing/opening velocity peak (ITVc/ITVo).

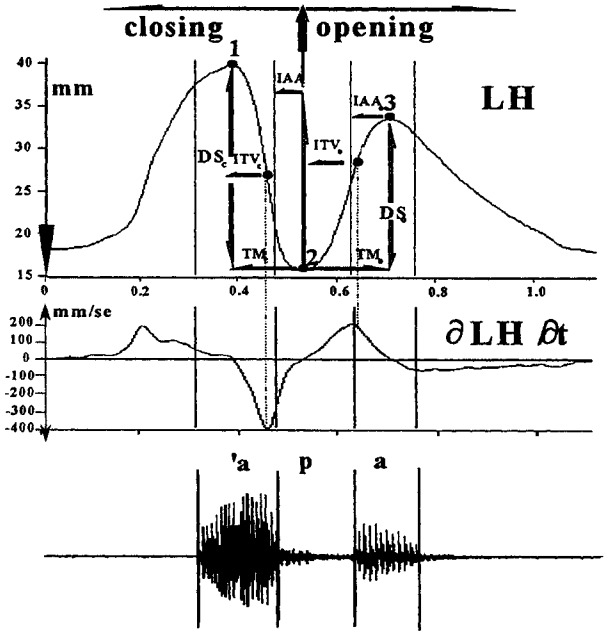

Figure 1. Target points and spatio/temporal definitions for analyzing LH parameter.

3. RESULTS

3.1. 3D spatial representation of consonantal targets

Figure 2 represents a 3D configuration of the lip orifice for all the 21 Italian consonants, based on the mean values of LH, LW, LP targets for all the speakers and productions. Between the two protrusion parameters, LP was preferred, as previous research studies (Magno Caldognetto et al. 1995, Cosi and Magno Caldognetto 1996) stated the high degree of correlation between them. Moreover, LP was always characterised by higher values with respect to UP. It is worth noticing that, for all the three parameters, both positive and negative values were found. It is evident that negative values indicate a reduction with respect to the values characterising the rest position. Italian consonants seem to be better distinguished by the values of the LH parameter. In fact, 5 classes were identified:

a) /p, b, m/: LH < 0mm;
b) /f, v/: 0mm < LH <3mm;
c) /t, d, s, z, ts, dz/: 6mm < LH < 8mm;
d) /N, L, S, tS, dZ/: 9mm < LH < 12mm;
e) /k, g, n, r, l/: 12mm < LH < 17mm.

These 5 groups of consonants should probably correspond to the 5 visemes of Italian, but a more robust statistical analysis will be executed in the future in order to justify this conclusion. As for LW, the most spread consonant was /f/ (LW=2.2mm greater than rest position), while, on the contrary, the most rounded was /l/ with a negative value of LW (-0.2). The other consonants could be divided into 2 separate groups: a more spread one, /v, t, d, ts, dz, N, L/, identified by 1.2mm < LW < 1.9mm, while the second put together all the other consonants located by 0mm < LW < 1mm.

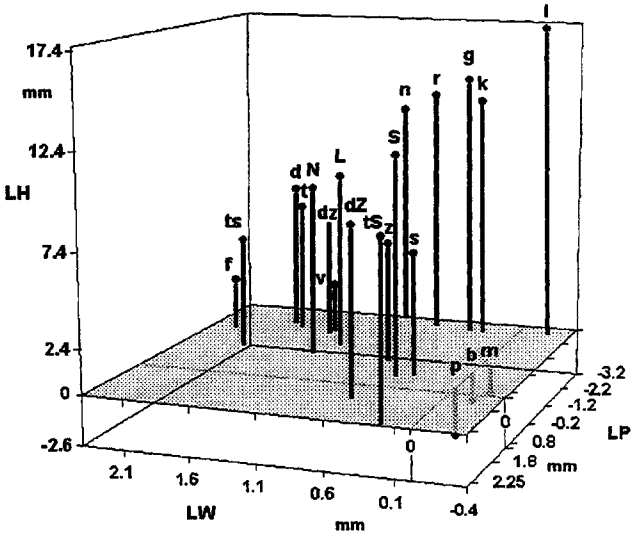

Figure 2. 3D spatial targets for 21 Italian consonants.

As for LP, almost all the consonants seem to be characterised by a reduction of the protrusion (LP < 0),

in particular /n,l, r/ resulted as the most retracted. Only medio-palatal affricates present positive values thus indicating labialisation. Along the 3 considered dimension the consonantal clusters are indeed different. Appropriate statistical analysis should be executed in order to quantify the relevance of these groupings and identify their relationships, thus defining 3D Italian visemes.

3.2. Spatio/temporal characteristics

Figure 3, shows the percentage of occurrence of the 4 identified articulatory strategies in the realisation of the consonantal targets. In 10 consonants the closing/opening (C+O) movement was the only chosen strategy. In 8 consonants C+O strategy is predominant; in 2 consonants (/S, ts/) the Plateau (C+P+O) strategy was observed in the majority of cases, while only in /r/ the Glide (G) movement
strategy was prevalent.

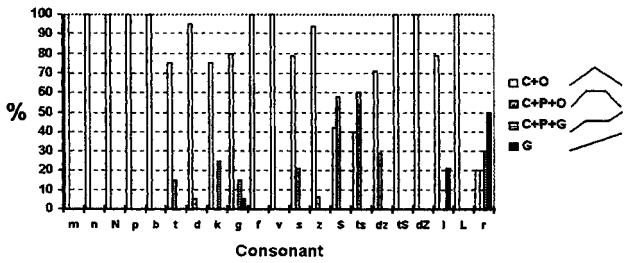

Figure 3. Percentage of occurrence of articulatory strategies.

The data presented in Figure 4 refer only to the predominant strategy for each consonant and correspond to the values of the displacement (DSc/DSo) and duration (TMc/TMo) of the closure/opening movement of the LH parameter.

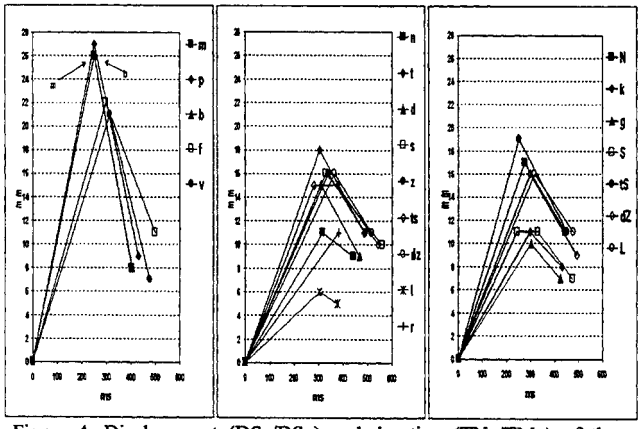

Figure 4. Displacement (DSc/DSo) and duration (TNc/TMo) of the closure/opening movement of the LH parameter

For all the consonants, which presented a closure-opening (C+O) movement (with the exception of /S/, /ts/, and /r/ characterised by different strategies) Figure 4 shows that the displacements of the closing movements are always greater than the opening ones, as the former start from the initial stressed vowel, which is the most open, while the latter end with the final unstressed vowel, which present a reduced degree of opening. As for the closure displacement (DSc), 5 major classes could be identified:

a) /p, b, m/: DSc>24mm;
b) /f, v/: 20mm<DSc<24mm;
c) /t, d, s, z, ts, dz, tS, dZ, N, L/: 14mm<DSc<20mm;
d) /n, r, S, k, g/: 6mm<DSc<14mm;
e) /l/: DSc<6mm.

As for the duration (TMc/TMo), the closure movement of all the consonants was always longer than the opening movement. As for the total duration of the articulatory movement of the consonant, bilabials /p, b, m/, and apicodentals /l, r/ showed the fastest movement, while /t, d, s, z, ts, dz/, were the longest ones.These groupings based on duration do not coincide with those based on displacement or target values. Appropriate statistical analysis shall have to demonstrate their different role in the construction of visemes.

As for the relationship between the articulatory movements and the corresponding acoustic production, Figure 5a, illustrates the duration of all the consonants together with that of the flanking vowels. For the affricates the duration of both the occlusive and fricative phase is highlighted. The acoustic duration of the consonants was always shorter than the corresponding global articulatory gesture duration, showing an anisomorphic behaviour between acoustic and articulatory data.

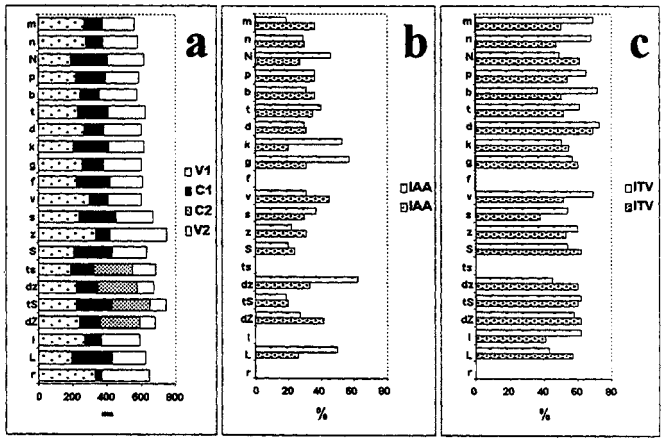

Figure 5. a) duration of all the consonants together with that of the flanking vowels; b) % of the total acoustic duration of the second/third segment taken by the time interval between the acoustic onset of the segment and the peak closing/opening position (IAAc/IAAo); % of the closing/opening movement taken by the time interval from onset of closing/opening movement to corresponding velocity peak (ITVc/ITVo).

Figure 5b shows the percentage value, related to the total acoustic duration of the consonant/second vowel, of the time interval between the acoustic onset of the segment and the peak closing/opening position (IAAc/IAAo). These data indicate that for almost all the consonants the values of these intervals range from 20% to 45% (except for /g/ and /dz/). Figure 5c illustrates the percentage of closing/opening movement

taken by the time interval between onset of closing/opening movement and closing/opening velocity peak (ITVc/ITVo).
The values for these parameters range between 40% and 60%. The parameters considered in Figure 5b-c will be quite useful in specifying rules for bimodal synthesis because they define the location of the consonantal target with respect to the correspondent acoustic landmarks (IAAc/IAAo) and the shape of the articulatory gesture (ITVc/ITVo).

4. CONCLUSIONS

All the data presented here demonstrate that in order to identify the consonantal visemes on the bases of articulatory data, the spatio-temporal characteristics of articulatory movements and their relationships with the co-produced acoustic signal should be taken into consideration, alongside the spatial target.

5. REFERENCES

Adjoudani A. and Benoit C. (1995), "Audio-Visual Speech Recognition Compared Across Two Architectures", *Proc. of Eurospeech-95*, 18-21 Sept. 1995, Madrid, Vol. 2, pp. 1563-1566.

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, pp. 485-504.

Benoit C., Guiard-Marigny T., Le Goff B. and Adjoudani A. (1996), "Which Components of the Face do Humans and Machines Best Speechread?", in Stork D.G. and Hennecke M.E. (Eds.), 1996, pp. 315-328.

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

Cohen M.M., Walker R.L. and Massaro D. (1996), "Perception of Synthetic Visual Speech", in Stork D.G. and Hennecke M.E. (Eds.), 1996, pp. 153-168.

Cosi P. and Magno Caldognetto E. (1996), "Lips and Jaw Movements for Vowels and Consonants:Spatio-Temporal Characteristics and Bimodal Recognition Applications", in Storke D.G. and Henneke M.E (Eds.), 1996, pp. 291-313.

Ferrigno G. and Pedotti A. (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.

Fourcin A.J., Harland G., Barry W. and Hazan W. (eds.) (1989), "Speech Input and Output Assessment, Multilingual Methods and Standards", Ellis Horwood Books in Information Technology.

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, pp. /453-456.

Massaro D.W. (1987), 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 Associates, Hillsdale, New Jersey, pp. 53-83.

Massaro D.W. (1996), "Bimodal Speech Perception: A Progress Report", in Stork D.G. and Hennecke M.E. (Eds.), 1996, pp. 79-102.

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

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.

Stork D.G. and Henneke M.E. (Eds.) (1996), 'Speechreading by Humans and Machine: Models, Systems and Applications', NATO ASI Series F: Computer and Systems Sciences, Vol. 150, 1996, Springer-Verlag.

Stork D.G., Wolff G. and Levine E. (1992), "Neural Network Lipreading System for Improved Speech Recognition", *Proc. of IEEE International Joint Conference on Neural Networks, IJCNN-92*, 285-295.

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.

Vatikiotis-Bateson E., Munhall K.G. and Hirayama M. (1996), "The Dynamics of Audiovisual Behavior in Speech", in Stork D.G. and Hennecke M.E. (eds.), 1996, 221-232.

Walden B.E., Prosek R.A., Montgomery A.A. Scherr C.K. and Jones C.J. (1977), "Effects of Training on the Visual Recognition of Consonants", *Journal of Speech and Hearing Research*, Vol. 20, 130-145.