Visual and acoustic modifications of phonetic labial targets in emotive speech: Effects of the co-production of speech and emotions

Emanuela Magno Caldognetto [1], Piero Cosi, Carlo Drioli, Graziano Tisato, Federica Cavicchio

Laboratory of Phonetics and Dialectology, ISTC-CNR, Institute of Cognitive Sciences and Technologies, Via Anghinoni 10, 35121 Padova, Italy

---

Abstract

This paper describes how the visual and acoustic characteristics of some Italian phones (/'a/, /b/, /v/) are modified in emotive speech by the expression of *joy*, *surprise*, *sadness*, *disgust*, *anger*, and *fear*. In this research we specifically analyze the interaction between labial configurations, peculiar to each emotion, and the articulatory lip movements of the Italian vowel /'a/ and consonants /b/ and /v/, defined by phonetic-phonological rules. This interaction was quantified examining the variations of the following parameters: lip opening, upper and lower lip vertical displacements, lip rounding, anterior/posterior movements (protrusion) of upper lip and lower lip, left and right lip corner horizontal displacements, left and right corner vertical displacements, and asymmetry parameters calculated as the difference between right and left corner position along the horizontal and the vertical axes. Moreover, we present the correlations between articulatory data and the spectral features of the co-produced acoustic signal.

Key words: Bimodal communication, Emotive speech, Audio-visual characteristics

of emotion, Co-production of speech and emotions, Phonetic modification induced by emotions.

1 Introduction

The studies on the bimodal audiovisual characteristics of emotive speech should specify the complex interactions among four categories of indices: a) the visual cues related to the lips, the jaw, and the tongue linguistic articulatory movements, carrying the information on visemes (Magno Caldognetto et al., 1997, 1998); b) the labial and facial visual cues related to the configurations of emotions; c) the acoustic cues conveying linguistic segmental, supra-segmental, prosodic, and intonational, characteristics of speech; d) the acoustic correlates of emotions: F0 parameters, voice quality, prosody, and intonation. Furthermore all of these indices superimpose the acoustic and visual extra-linguistic and idiosyncratic characteristics of the speaker. Traditionally, facial and speech acoustic cues (both segmental and supra-segmental) conveying emotions have been studied separately: from the first studies on facial configurations (Ekman and Friesen, 1978) and on acoustic characteristics (Scherer, 1986; Banse and Scherer, 1996) to more recent fundamental publications: Ekman et al. (2002); Camras et al. (1993); Scherer (2003); Scherer et al. (2003); Douglas-Cowie et al. (2003). The relevance of the interaction between audio and visual modalities in the transfer of emotions has been stressed particularly from the perceptual point of view (Hess et al., 1998; Massaro and Egan, 1998; deGelder et al., 1998; deGelder and Vroomen, 2000). We focused

[1] Corresponding author. Tel.: +39-049-8274409; fax.: +39-049-8274416.

E-mail address: magno@csrf.pd.cnr.it (E. Magno Caldognetto)

our research on the quantification of the labial articulatory parameters modifications induced by the different emotions, and on the definition of their acoustic correlates (see Magno Caldognetto et al. (2003); Nordstrand et al. (2003) and the recent papers on smile (Schroeder et al., 1998; Aubergé and Cathiard, 2003)). In particular, our study is aimed at identifying the effects on the parameters chosen to describe the labial configuration on the non-emotive speech production (LO, UL, LL, LR, ULP, LLP, LCX, RCX, ASYMX, LCY, RCY, ASYMY. See Section 3, Fig. 2). These effects are due to the compliance between both the phonetic-phonological constraints and the configurations required for the encoding of emotions. The obtained results will be useful for the formulation of cognitive theories on co-production and perception of linguistic and paralinguistic information, mixed in emotive speech, and for the working out of speech technologies, i.e. bimodal speech synthesis (Talking Heads) and recognition systems.

2 Method

In order to collect the articulatory and acoustic data, an automatic opto-tracking movement analyzer for 3D kinematics data acquisition (ELITE) was used, which also allows a synchronous recording of the acoustic signal (for previous applications of this data acquisition system to the definition of Italian visemes on an articulatory basis, see Magno Caldognetto et al. (1997, 1998)). This system ensures high accuracy (100 Hz sampling rate, maximum error of 0.1 mm for a 28x28x28 cm cube) and minimum discomfort to the subject because it tracks the infrared light reflected by small (2 mm diameter), passive markers glued on different points of the external lips contour and of the

[FIGURE 1 APPROX. HERE.]

Figure 1. Position of the 28 reflecting markers and of the reference planes for the articulatory movement data collection.

face, following the scheme in Fig. 1. Only the data relative to lip markers are presented in this work. A male University student, speaking a northern regional Italian and with recitation skills, pronounced two phonological structures 'VCV', corresponding to two feminine proper names: "Aba" /'aba/ and "Ava" /'ava/, simulating, on the basis of appropriate scenarios, six emotive states: *anger* (A), *joy* (J), *fear* (F), *sadness* (SA), *disgust* (D) and *surprise* (SU), apart from the neutral one (N), corresponding to a declarative sentence. This 14 words set was repeated many times in random order, leading to a total of 107 recordings [2]. The ELITE system allows the synchronous recording of both the lip movements and the co-produced acoustic signal. In this research we present, for all the emotions, the analysis of the mean values characterizing, for all the acoustic and articulatory parameters, the targets of the vowel /'a/ and the consonants /b/ and /v/, extracted with respect to the mid point of the corresponding acoustic signal [3]. Figure 2 proposes an example of nor-

[2] Due to the complexity of the movement tracking algorithm, not all the acquisitions resulted reliable. For this reason, 7 cases were used for A, 5 for D, 9 for N, 10 for J, 6 for F, 10 for SU, 7 for SA in the word /'aba/, and 7 cases were used for A, 6 for D, 11 for N, 8 for J, 6 for F, 9 for SU, 6 for SA in the word /'ava/. All the 107 recordings considered for the articulatory and acoustic analysis were judged by 3 listeners and scored between 4 and 5 on a 0-5 adequacy scale.

[3] The mid point data collected here are suitable for a rule-based bimodal text-to-speech synthesis.

[FIGURE 2 APPROX. HERE.]

Figure 2. Speech signal and spatio-temporal evolution of the articulatory parameters associated with the sequence /'aba/ expressing *disgust*.

malized values for the labial kinematic parameters. The mid points[4] of the vowel stationary signal, of the /b/ closure phase and of the /v/ friction noise, do not always correspond to the maximum or minimum values of the different parameters, already extracted in past researches on the Italian visemes (Magno Caldognetto et al., 1997, 1998). Future researches will describe the spatio-temporal evolution of the articulatory parameters.

We applied two different normalizations on the data obtained from the ELITE system: first, we subtracted the value of the resting position from each parameter to quantify the labial movements independently from the extra-linguistic characteristics of the lip shape. The second normalization was obtained subtracting the values of the neutral production from the first normalization data: in this way we highlight the specifical emotive variability for the articulatory parameters.

[4] A two-factor Anova verified that the mid point values of /'a/, /b/ and /v/ were not significantly different from the mean values calculated on each vowel and consonant acoustic segments.

3 Articulatory data analysis

3.1 First normalization procedure

For each acquisition session the articulatory data at resting position has been recorded as well, and the extracted parameters have been normalized with respect to this values (Magno Caldognetto et al., 2003). The parameters selected to quantify the labial configuration modifications are the following:

- *Lip Opening* (LO), calculated as the distance between the markers placed on the central points of the upper and lower lip vermilion borders; this parameter correlates with the HIGH-LOW phonetic dimension. Positive values correspond to the lip opening, negative values to lip closure.
- *Upper and Lower Lip vertical displacements* (UL and LL), calculated as a distance between the markers placed on the central point of either upper and lower lip and the transversal plane Ω (Fig. 1), passing through the tip of the nose and the markers on the ear lobes. Positive values correspond to a rising of the lips, i.e. a reduction of the distance of the markers from the Ω plane, while negative values stand for a lowering of the lips.
- *Lip rounding* (LR), corresponding to the distance between the left and right corners of the lips, which correlates with the ROUNDED-UNROUNDED phonetic dimension: negative values correspond to the lip spreading, positive values to lip rounding.
- *Anterior/posterior movements (Protrusion) of Upper Lip and Lower Lip* (ULP and LLP), calculated as the distance between the marker placed on the central points of either the upper and lower lip and the frontal plane Δ containing the line crossing the markers placed on the lobes of the ears

and perpendicular to Ω plane. These parameters correlate with the feature PROTRUDED-RETRACTED: negative values quantify the lip retraction, positive values correspond to lip protrusion.

- *Left and Right Corner horizontal displacements* (LCX and RCX), calculated as the distance between the markers placed on the left and the right lip corner and the sagittal Σ plane passing through the tip of the nose and perpendicular to the Ω plane. Positive values correspond to a stretching to the right of the right corner and to the left of the left corner with respect to the resting position, i.e. in both cases to an increasing of the distance from the Σ plane. Negative values correspond to a reduction of the distance from the Σ plane.
- *Left and Right Corner vertical displacements* (LCY and RCY), calculated as the distance between the markers placed on the left and right lip corner and the transversal plane Ω, containing the line crossing the markers placed on the lobes of the ears and on the nose. Positive values correspond to a corner lowering with respect to the resting position, i.e to an increment of the distance from the Ω plane, while negative values correspond to a corner rising and a decrease of the distance.
- The *Asymmetry parameter* ASYMX was calculated as the difference between right and left corner position along the X (RCX-LCX). Values different from zero indicate the presence of an asymmetry, so that positive values indicate right asymmetry and negative values indicate left asymmetry. For example, a positive ASYMX correspond to a larger distance of the right corner from the Σ plane, than the left corner. The asymmetry parameter ASYMY was calculated as the difference between right and left corner position along the Y (RCY-LCY) axes. A positive ASYMY correspond to a larger distance of the right corner from the Ω plane, than the left corner,

[FIGURE 3 APPROX. HERE.]

Figure 3. Mean values of LO, LR, and LLP mid points defining the /'a/, /b/, and /v/ targets for /'aba/ (panel a) and /'ava/ (panel b). Upper case labels indicate vowel /'a/, lower case labels indicate consonants /b/ and /v/.

Table 1

Results of the 2-factor ANOVA (emotion, stimulus, interaction) on articulatory parameters for vowels /'a/ in /'aba/ and /'ava/ and consonants /b/-/v/: F-ratio and associated p-value.

[TABLE 1 APPROX. HERE.]

i.e. a lower right corner than the left one.

Figure 2 proposes the labial kinematic parameters considered in this study, relative to /'aba/ expressing *disgust*: LO, LR, ULP, LLP, UL, LL, LCX, RCX, LCY, RCY, ASYMX and ASYMY.

3.2 First normalization results

In Fig. 3, the mean values of the LO, LR, and LLP in the stressed vowels (upper case labels) and consonants (lower case labels) of /'aba/ and /'ava/ are plotted in a 3D space. LLP is preferred over ULP since LL is the active articulator in the realization of the labiodental voiced fricative /v/ (Magno Caldognetto et al., 1997, 1998).

The different contribution of UL and LL to the realization of LO is presented in

[FIGURE 4 APPROX. HERE.]

Figure 4. Mean values of the UL and LL vertical displacement mid points for the /'a/, /b/, and /v/ in /'aba/ and /'ava/.

[FIGURE 5 APPROX. HERE.]

Figure 5. Mean values of horizontal and vertical asymmetries of the lip corners for the /'a/, /b/, and /v/ in /'aba/ and /'ava/.

Fig. 4. Figure 5 shows the data referring to the horizontal asymmetry ASYMX (solid line) and vertical asymmetry ASYMY (broken line) of the displacement of the lip corners.

A 2-factor ANOVA was performed to assess the effects of emotion (A, J, F, SA, D, SU, and also N) and stimulus (/'aba/, /'ava/) on each dependent articulatory variable for the vowels /'a/ and the consonants /b/ and /v/. F-ratio and associated p-values of the 2-factor ANOVA (emotion, stimulus, interaction) on articulatory parameters for vowels /'a/ and consonants /b/-/v/ are reported in Table 1. For all parameters, the proportion of variance accounted by emotions was always significant. For the vowels /'a/, emotion was the only significant effect, whereas for the two consonants the stimulus was significant too. For this reason, in the following the two vowels /'a/ of /'aba/ and /'ava/ are therefore considered together.

In order to understand the 2-factor ANOVA results, we now examine separately the mean values of the parameters for each emotion.

3.2.1 Lip Opening

As regards the vowels /'a/ (Fig. 3), with respect to *neutral* (N:15 mm), LO distinguishes significantly 3 groups of emotions: 1) the highest values are reached for *fear* (F: 22 mm) and *anger* (A: 19 mm); 2) *sadness* (SA) and *surprise* (SU) measure 14 mm and 13.9 mm respectively; 3) *disgust* and *joy* present the lowest LO displacement (D: 9 mm and J:10 mm). From the 2-factor ANOVA, we can establish that the consonants distinguish significantly by stimuli apart from emotions. Concerning /b/, *neutral* presents a characteristic negative LO value (n: -0.4 mm) determined by the lip compression related to the articulatory closure phase. The highest negative value is reached by *disgust* (d: -3.5 mm), followed by *joy* (j: -1.3 mm) and *anger* (a: -1.2 mm), while *fear* and *sadness* stand out for they exceed the resting position by respectively 3 mm and 1 mm. As regards /v/, LO presents a limited range of positive values, determined by the labiodental constriction, which is the relevant phonological feature. In fact, LO varies from the lowest value for *neutral* (n: 2 mm) to the highest value for *fear* (f: 5.5 mm), and middle values for *sadness* (sa: 4 mm) and *joy* (j: 3 mm).

3.2.2 Upper Lip Displacement

UL vertical displacement values in the two vowels /'a/ change significantly according to the emotions (Fig. 4). With respect to the neutral (N: -0.04 mm), *anger*, *joy* and *disgust* are characterized by positive values (A: 0.3 mm, J: 2.2 mm, D: 4.2 mm), corresponding to UL raising. In the other emotions, UL presents values between -1.2 mm for *fear* and 0.6 mm for *surprise*. The mean values for consonants vary significantly not only by emotions, but by stimuli too. For /b/, UL has always negative values, between -0.1 mm for *disgust* (d), -1.1 mm for *neutral* (n) and -3.4 mm for *fear* (f), corresponding to UL lowering. In /v/ consonant, UL is not involved in labiodental constriction while it is subjected to the coarticulatory effects of the contextual vowel. Therefore *disgust* (d: 4 mm) and *joy* (j: 3 mm) are characterized by positive displacements which correspond to UL raising movements. The other emotions present negative values for *anger* (-0.3 mm), *surprise* (-0.4 mm), *sadness* (-0.9 mm), and *fear* (-1.4 mm).

3.2.3 Lower Lip Displacement

In LL displacements the vowels and the consonants vary significantly according to the emotions (Fig. 4). It is possible to define 3 groups for the stressed vowels: the first includes the emotions characterized by the widest displacement: *fear* (F: -20 mm) and *anger* (A: -17.2 mm, which correspond to LL lowering and to lip opening). The second group contains *neutral* (N: -12.4 mm), *surprise* (SU: -12.3 mm) and *sadness* (SA: -12.9 mm). The third group consists of *joy* and *disgust* (J: -8.3 mm and D: -5.2 mm), which present the minimum LL lowering and lip opening. LL displacement mean values in /b/ and /v/ are similar and both narrower than the vowels, as LL realizes the lip closure and labiodental constriction. In /b/ consonant, we found in *fear* (f: -5.8 mm) the greater displacement value, comparing to *neutral* (n: -1.3 mm), followed by *sadness* (sa: -2.8 mm). Positive values, corresponding to a LL raising, characterize *disgust* (d: 2.8 mm) and *joy* (j: 0.6 mm). In /v/ consonant, *fear* (f: -4.8 mm) and *sadness* (sa: -4.3 mm) are characterized by the largest displacement mean values. Positive values are realized only for *disgust* (d: 2.2 mm).

3.2.4 Lip Rounding

Concerning the two vowels /'a/ (Fig. 3), with respect to *neutral* value (N: 1.7 mm), LR reaches the highest positive values for *fear* (F: 4.4 mm), *sadness* (SA: 3.4 mm), *anger* (A: 2.4 mm) and *surprise* (SU: 2.1 mm), whereas the lowest negative values, corresponding to the lip spreading, are found for *joy* (J: -6 mm) and *disgust* (D: -6.2 mm). For the consonant /b/, *anger* (a: 0.47 mm) and *sadness* (sa: 0.45 mm) determine higher rounding values than *neutral* (n: -0.3 mm), while *joy* presents intermediate values (j: -4.8 mm) and *disgust* is characterized by the maximum lip spreading (d: -8.8 mm). For /v/, with respect to neutral production (n: 0.08 mm), *anger* (a: 0.3 mm), *surprise* (su: 0.5 mm), *fear* (f: 1 mm) and *sadness* (sa: 1.4 mm) present positive lip rounding values, while *disgust* and *joy* are characterized by the maximum lip retraction values (d: -5.4 mm and j: -7.3 mm).

3.2.5 Lower Lip Protrusion[5]

As concerns /'a/ (Fig. 3), *disgust* and *joy* present LLP lowest values (D: -4.5 mm, J: -4.1 mm) corresponding to lip retraction, whereas the other emotions are characterized by mean values between -1.7 mm for *neutral* and 0.3 mm for *anger*. The consonants are significantly distinguished not only by emotions, but by stimuli too. For /b/, with respect to *neutral* mean value (n: 0.4 mm), *anger* (a: 1.8 mm) and *sadness* (sa: 0.9 mm) are characterized by the highest protrusion, and *disgust* (d: -3.2 mm) by the highest retraction. For /v/, all the emotions present negative values due to the labiodental constriction, with an intermediate value for *neutral* (n: -2.3 mm), and the highest lip retraction

[5] see Sec. 3.2 for the preference of LLP over ULP.

values for *disgust* and *joy* (d: -4.6 mm and j: -4.4 mm).

3.2.6 Right and Left Corner Horizontal Displacement

In LCX and RCX displacements the vowels and the consonants vary significantly according to the emotions and stimuli. For the stressed vowels /'a/ we can notice that the widest RCX displacement is shown for *joy* (J: 2.7 mm) and, above all, for *disgust* (D: 4.5 mm), which are the only two positive values, corresponding to a higher distance from the plane Ω, while all the other emotions present negative values, corresponding to a distance reduction, with the lowest negative values for *fear* (F:-1.5 mm) and *sadness* (SA: -1 mm). For LCX, similarly the widest displacement is in *joy* (J: 2.9 mm) and *disgust* (D: 3.2 mm), while the other emotions are characterized by negative values, corresponding to a reduction of the horizontal displacement, with the lowest negative mean values for *fear* (F: -2.3 mm), *sadness* (SA: -1.9 mm) and *anger* (A: -1.8 mm). RCX displacement mean values in /b/ are also characterized by positive values for *joy* (j: 2.1 mm) and *disgust* (d: 4.8 mm), while all the other emotions present a narrow range of positive values, close to the neutral production (n: 0.1 mm, a: 0.4 mm, su: 0.2 mm, sa: 0.1 mm); the only negative value is *fear* (f: -0.02 mm). In LCX, *joy* (j: 2.7 mm) and *disgust* (d: 3.7 mm) are still characterized by the widest horizontal displacements, while the other emotions have all negative values, from the lowest of *fear* (f: -1.1 mm) to *surprise* (su: -0.2 mm), while the neutral production has a little positive value (n: 0.2 mm). As regards RCX in /v/, *joy* (j: 3.3 mm) and *disgust* (d: 3.4 mm) too show the widest displacements, while the other emotions and neutral production are characterized by a narrow range of positive values between *surprise* (su: 0.05 mm) and *anger* (a: 0.5 mm). The only negative value is in *sadness* (sa: -0.4 mm). LCX is characterized by widest displacements in *joy* (j: 4 mm) and *disgust* (d: 2.1 mm), while the other emotions, apart from the neutral production (n: 0.2 mm), present a negative value range between *fear* (f: -1.1 mm) and *sadness* (sa: -0.7 mm).

3.2.7 Right and Left Corner Vertical Displacement

In RCY and LCY displacements, the vowels and the consonants vary significantly according to the emotions and stimuli. For the /'a/, RCY is characterized by positive values with the highest positive values for *fear* (F: 6.3 mm), corresponding to a higher distance from the plane Ω (i.e. a lowering of the corner with respect to the resting position), while the only negative value is in *joy* (J: -3 mm), due to a distance reduction from the plane (i.e. a corner rising with respect to the resting position). In LCY, the stressed vowels have positive values for all the emotions and the neutral production, except for *joy* (J: -4 mm), with the lowest positive value for *disgust* (D: 2.6 mm) and the highest for *fear* (F: 7 mm). The negative value in *joy* is due to a corner rising. In /b/ consonant RCY has positive values ranging from 2.2 mm for *anger* and 2.9 mm for the neutral production and the highest positive values for *fear* (f: 4.9 mm). The only negative value is in *joy* (j: -2 mm). LCY in /b/ has all positive values ranging from 2.1 mm for the neutral production to 4.4 mm for *fear*. The only negative value is, once again, in *joy* (j: -3.4 mm). In /v/, RCY determines three groups: in the first there are positive values for *anger* (a: 1 mm), *surprise* (su: 1.2 mm), *fear* and *sadness* (f, sa: 2.4 mm both); an intermediate group is composed by the neutral production (n: -0.3 mm) and *disgust* (d: -0.2 mm); in the third group, *joy* (j: -6.2 mm) presents the lowest negative value, corresponding to the corner rising. LCY has all positive values ranging from the lowest positive value for *disgust* (d: 1.1 mm) to the highest positive value for *fear* (f: 3.1 mm), while the only two negative values are for neutral production (n: -0.2 mm) and for *joy* (j: -7.9 mm), with the highest negative value.

3.2.8 Horizontal and Vertical Asymmetries[6]

The 2-factor ANOVA shows for the stressed vowels and the consonants that emotion is the only significant effect (Fig. 5). As expected, the neutral production is characterized by no asymmetries. Vowels and consonants in all emotions (except *joy*) present a right positive horizontal asymmetry (i.e. the right corner is displaced on the right side more than the left corner). Moreover, we found a vertical negative asymmetry ASYMY (i.e. the right corner is higher than the left one), except for *joy* for which the right corner is lower than the left one. Summarizing, *joy* is characterized, with respect to the other emotions, by a positive vertical asymmetry and by a left negative horizontal asymmetry.

3.3 Second normalization procedure

For each recording the extracted articulatory parameters, already normalized by the resting position, have been subsequently normalized with respect to the corresponding data of the neutral production. The meaning of the parameter values is the following:

- *Lip Opening* (LO): positive values correspond to an increase of lip opening if compared to the neutral production; negative values correspond to a lip

[6] see Palmer et al. (1986) for the problems relative to the asymmetry.

opening reduction.

- *Upper and Lower Lip vertical displacements* (UL and LL): positive values correspond to a lip rising with respect to neutral (i.e. a reduction of the distance from the Ω plane), while negative values stand for a lip lowering.
- *Lip Rounding* (LR): negative values correspond to higher lip spreading if compared to the neutral production; positive values correspond to higher lip rounding if compared to the neutral production.
- *Anterior/posterior movements (Protrusion) of Upper Lip and Lower Lip* (ULP and LLP): negative values quantify the growth of lip retraction with respect to the neutral production; positive values correspond to the growth of lip protrusion.
- *Left and Right Corner horizontal displacements* (LCX and RCX): positive values correspond to a higher distance from the neutral position; negative values correspond to a distance reduction.
- *Left and Right Corner vertical displacements* (LCY and RCY): positive values correspond to a lip corner lowering with respect to the neutral production; negative values correspond to a lip corner rising.
- The *Asymmetry parameter* (ASYMX and ASYMY): with the second normalization, positive values of ASYMX correspond to a larger distance of the right corner from the neutral position, than the left corner. For ASYMY, positive values correspond to a larger distance of the right corner from the Ω plane, than the left corner.

Table 2

Results of the T-test performed on the second normalization values of all parameters and of all emotions. Italic means that p value is higher than 0.1.

[TABLE 2 APPROX. HERE.]

3.4 Second normalization results

3.4.1 T-test

To verify the significative of the collected data a one sample T-test was performed (Table 2).

The analysis of the results of the T-test leads to interesting observations: 1) some emotions, such as *disgust* and *joy*, present, within each of the three acoustic segments, a large number of values significantly different from the other emotions; 2) the same emotion can give significantly different results according to the segments, e.g. *anger* presents a large number of data significantly different for the targets /'a/ and /b/ with respect to /v/; 3) the three phonetic targets are characterized by different variability of the parameters, since the articulatory linguistic constraints affect the modification induced by the emotive target. For example, in /v/ the LR parameter presents a narrow range of values, statistically not significant, due to the realization of the labiodental constriction.

3.4.2 Cluster analysis results

All the articulatory parameters were analyzed applying the Average Linkage Method between Groups, which converts the numerical similarity between the mean value of each parameter for each emotion into a special or a geometrical distance proximity. In this way, the emotions that are more similar to each other in terms of real values are grouped closed together (the parameters are grouped by phonetic coherence). All the clusters for the three phonetic targets and for each parameter considered, together with the mean values relative to each emotion, are reported in Appendix A.

Cluster analysis for the /'a/ vowel:

Disgust and *joy* constitute, for most of the parameters, a cluster that differentiates itself from all the other emotions. In fact *disgust* and *joy* are characterized by: maximum positive UL values (which correspond to the maximum upper lip rising with respect to *neutral*); maximum positive LL values (which correspond to the lower lip rising with respect to *neutral*); maximum negative LO values (which correspond to the maximum lip opening reduction with respect to *neutral*); lowest LR values (corresponding to maximum spreading), lowest ULP and LLP values (corresponding to maximum lip retraction); maximum RCX and LCX values (corresponding to the highest spreading). With regard to the other parameters (LCY, RCY, ASYMX, ASYMY), *joy* differentiates itself from all the other emotions: in particular the maximum negative values of LCY and RCY quantify the raising of the corners of the lips. ASYMX is distinguished by *joy* (-0.7 mm), which has the lowest negative value, corresponding to a larger relative left horizontal displacement of the left lip corner with respect to the right lip corner. ASYMY is also distinguished by *joy* (1.1 mm), that is the only positive value, corresponding to a relative lowering of right corner with respect to the left one. It should be noted that in the group of remaining emotions, *anger* and *fear* distinguish themselves by the maximum positive LO values and the maximum negative LL values. In particular, *fear* presents the maximum LO value (F: 9.2 mm), and also the maximum positive LR value. As a consequence, LO and LR result to be associated.

Cluster analysis for the /b/ consonant:

Disgust and *joy*, in the realization of the occlusive bilabial /b/, are grouped only by UL, LR, and LCX. Among these parameters, LR presents the maximum negative values (-8.5 mm for *disgust* and -4.6 mm for *joy*), and results to be the most important for it quantifies the labial stretching typical of these emotions. With regard to the other parameters, *disgust* and *joy* result independent from each other, and with regard to the other emotions. As regards ASYMX, *Joy* (-0.5 mm) is characterized by a negative value, corresponding to a relative horizontal displacement of the left lip corner with respect to the right one. The other emotions are all positive and restricted to a narrow range of values. As regards ASYMY, *joy* (0.6 mm) presents the only positive value, corresponding to a relative lowering of the right corner with respect to the left one.

In turn, *disgust* is distinguished by: the maximum negative LO values (i.e the maximum reduction of the opening); the maximum positive LL values (i.e. the maximum rising of the lower lip); the maximum negative ULP and LLP values (i.e. maximum lip retraction); the maximum ASYMY negative value (-3 mm), corresponding to a relative lowering of the left lip corner with respect to the right one.

Cluster analysis for the /v/ consonant:

Disgust and *joy* also constitute an unique cluster with respect to all the other emotions by the following five characteristics: maximum positive UL values (i.e. maximum upper lip rising with respect to *neutral*); maximum negative ULP and LLP values (i.e. maximum labial retraction); maximum positive LCX and RCX values (i.e. highest distances from the neutral position); ASYMX lowest negative value (-1.1 mm), corresponding to a relative horizontal displacement of the left lip corner with respect to the right one; positive value of ASYMY (1.7 mm), corresponding to a relative lowering of right corner with respect to the left one.

Disgust presents an unique negative LO value while *joy* distinguishes itself from all the other emotions by the maximum negative LCY and RCY values, corresponding to a lip corner raising.

The inspection of the LR cluster structure emphasizes in /v/ different aggregations with respect to those already described for /'a/ and /b/ in which *disgust* and *joy* result quite different from all the other emotions. For /v/, one of the cluster defined by *anger* and *fear* is characterized by the LR reduction with respect to *neutral*, even with minimum values (A: -1.3 mm and F: -1.2 mm). All the other emotions move within a very limited range around the *neutral*. These LR data point out the effect of the constraints due to the labiodental constriction on the emotive realization of this consonant.

4 Acoustic data analysis

4.1 Statistical results

In Fig. 6 we can see the 3D plots of the mean values of F1, F2, F3 spectral characteristics, extracted in the stressed vowel mid points. Table 3 presents the

Table 3

Mean values expressed in Hz of the F0 mid points for stressed vowels and consonants in /’aba/ and /’ava/.

[TABLE 3 APPROX. HERE.]

Table 4

Results of the 2-factor ANOVA (emotion, stimulus, interaction) on acoustic parameters for vowel /’a/ and consonant /b/-/v/, F-ratio and p-value associated.

[TABLE 4 APPROX. HERE.]

mean values of F0 mid points for /’a/, /b/ and /v/. A 2-factor ANOVA was performed to assess the effects of emotions and stimuli on the chosen acoustic parameters. F-ratio associated to the main effects of emotions, stimuli and interaction are reported in Table 4. For vowels and consonants, and for all parameters except F1, the proportion of variance accounted by emotions was significant. As previously shown in Fig. 6, the F1 values for /’a/ of /’aba/ vary between 794 Hz of *surprise* and 895 Hz of *joy*, while F1 values for /’a/ of /’ava/ vary between 780 Hz of *surprise* and 910 Hz of *fear*. Due to the limited range of variation, F1 is not distinguished significantly by emotions. On the contrary F2 values are significantly different by emotions. With respect to *neutral* (N: 1347 Hz), *sadness* presents the lowest value of F2 (SA: 1258 Hz), while we find the highest values for *joy* (J: 1522 Hz) and *fear* (F: 1786 Hz). As concerns F3, *neutral* is characterized by a mean value of 2585 Hz, while *anger* and *disgust* are distinguished by values ranging between 2490 Hz and 2600 Hz. *Joy*, *surprise* and *sadness* present F3 values ranging between 2700 Hz and 2800 Hz. *Fear* is characterized by the highest F3 value for /’a/

[FIGURE 6 APPROX. HERE.]

Figure 6. 3D representation of spectral characteristics F1, F2, F3 of the vocalic targets for all emotions and neutral production.

of /'aba/ (3163 Hz) while *sadness* presents the maximum for /'a/ in /'ava/ (3086 Hz).

4.2 Cluster analysis on normalized acoustic data

As for the articulatory parameters, acoustic data were normalized by the corresponding data of the neutral production. For the /'a/, F0mid values determine three groups. The first one, characterized by the highest F0mid values, aggregates *fear*, *joy* and *surprise*. The second one combines in a group *anger* and *sadness* and is characterized by F0mid intermediate values. In the last one, *disgust* presents the smallest difference from the *neutral*. The F1mid values emphasize a group of emotions (*anger*, *joy*, *fear*, *disgust*, and *sadness*), quite close to *neutral*, and an isolated *surprise* that is significantly different from the other emotions and from the neutral. With regard to F2mid, *sadness* is the only emotion that shows a frequency reduction with respect to the *neutral* and all other emotions, characterized by a frequency increase, with *joy* and *fear* presenting the highest variation. The F3mid values distinguish *anger* and *disgust* (characterized by values close to *neutral*) from the other emotions, which are characterized by an appreciable raising of F3mid, with *sadness* and *fear* that reach the highest differences.

The /b/ and /v/ consonants present different clusters, but we noted in both that *fear* is distinguished by greatest F0mid values. Moreover, *anger*, *disgust*, *surprise* and *sadness* values are greater in /b/ as in /v/ with the exception of *joy*.

5 Correlations between acoustic and articulatory data

In order to define the relationship between the acoustic and the articulatory parameters the correlations between the two series of data presented in Sec. 3 and Sec. 4 are calculated. In this paper we describe only the strong correlations, positive and negative, characterized by a score between 0.750 and 1.000. As concerns the *neutral*: LO, LR, ULP are negatively correlated to F0; LO and ULP are positively correlated to F1; LR is negatively correlated to F1 . For *anger* there are several strong correlations: LR and ULP are both strongly negatively correlated to F3; LO is strongly negatively correlated to F0; LR is strongly negatively correlated to F3 and ASYMX is strongly positively correlated to F0. As regards *joy*, we found only one negatively strong correlation between LO and F2. In *sadness* ULP is negatively correlated to F1; *disgust* has only one negative correlation between F1 and ASYMX while in *surprise* ASYMY is strongly positively correlated to F1. Let us first discuss the correlations between articulatory data and F0. It is recognized that the nonlinear acoustic interaction between the vocal tract and the voice source is responsible for acting in various ways on the shape of the glottal pulse (Fant and Lin, 1987; Stevens, 1998). However, no much trust is given to the hypothesis that the acoustics of the vocal tract can affect the glottal flow period length (1/F0) (Whalen and Levitt, 1995), at least for vowels. On the other hand, there are convincing arguments supporting the idea that the articulation, most likely the pull of the tongue on the larynx, can affect the fundamental frequency. This argument is used in the literature to discuss the phenomenon of intrinsic pitch (IF0), i.e. the fact that different vowels tend to have different fundamental frequency (Whalen and Levitt, 1995; Fischer-Jørgensen, 1990). Keeping this in mind, the negative correlations between F0 and LO found in *neutral* and *anger* can be explained with an increasing of the jaw opening and of the tongue action toward the back of the palate (supralaryngeal region), with a consequent reduction of the tension and an increase of the mass of the vocal folds, and thus a lowering of the pitch (Fischer-Jørgensen, 1990). The correlation between LR, ULP and F0, is not simple to explain by a direct relation. However, it has been observed that a positive correlation exists between these parameters and LO (Magno Caldognetto et al., 1995, 1998), and this could explain the statistical result. Finally, no clear relation can be drawn between ASYMX and F0 to explain the strongly positive correlation found in *anger*. Let us now discuss the correlations concerning F1, F2, F3, and the articulatory parameters. We should refer here to the well established rules that describe the acoustics of the vocal tract, given the position of articulators (e.g., Deller et al. (1993); Stevens (1998)). In the vowel /'a/, the jaw is lowered to open the oral cavity, the pharynx is constricted, and the tongue is pulled down in a central/back position. It is known that F1 is lowered by any constriction in the oral cavity and raised by pharyngeal constriction, to which correspond in general an open oral cavity. This explains the positive correlation between LO and F1: the lip opening and the jaw lowering increase the oral cavity (which produces the same effect of a shortening of the vocal tract), and reduce the pharyngeal tube length. Moreover, this also explains the negative correlation between LR and F1 found for *neutral* (LR increases the oral constriction and the vocal tract length). The positive correlation between ULP and F1 is less straightforward, however it can be hypothesized that the ULP increasing can be associated with an increment of the LO. In *sadness*, negative correlation between ULP and F1 is justified by the lengthening of the vocal tract due to the lip protrusion, and by a flattening of the lips (low LO values). Finally, the correlations ASYMX-F1 and ASYMY-F1 found in *disgust* and *surprise* have no straightforward explanation. The negative strong correlation between LO and F2 found in *joy* is in agreement with the fact that the frequency of F2 is lowered by a back tongue constriction (which increases in /'a/ for a LO increment). The strong negative correlations between LR and F3 found in *anger* is in agreement with the fact that the constriction of the front end of the oral cavity tends to lower all formants, the effect being more pronounced on the higher ones.

While we have much data and many acoustic models for the production of vowels and consonants of various languages, in emotive speech we are still far from having all the necessary experimental data (aerodynamic, glottic, articulatory and acoustic) essential for the modeling (with the exception of researches on the smile (Schroeder et al., 1998; Aubergé and Cathiard, 2003)). In the emotive speech, in fact, there are not recorded associations among paramenters (e. g. LR, ULP and LLP) detected in the vowel and consonant canonical targets.

6 Conclusions

Although the data presented in this paper are only a first contribution to the issue of visual and acoustic segmental characteristics of emotive speech, several results are interesting. In fact, the statistical analyses show that, in emotive speech, the articulatory targets /'a/,/b/ and /v/ vary significantly according to different emotions (first normalization results in Sec. 3.2). Moreover the T-test on the data of the second normalization (Sec. 3.4) shows that the parameter variability is significantly different according to the articulatory target taken into consideration. This happens because the lip configuration of emotions must respect the articulatory linguistic characteristic. The analysis of clusters, moreover, allows for the individuation of similarities and differences in emotions according to the parameters.

Some final considerations are discussed hereafter.

1- Our researches highlighs the modifications induced by emotions for all labial kinematic and acoustic chosen parameters. We verified that some labial parameters vary significantly not only by emotions, but by stimuli too: LO, UL and LLP values are distinguished by the consonants. On the basis of these quantifications it is possible to characterize visually and acoustically the emotions: i.e., *disgust* and *joy* are both characterized by minimum lip opening and maximum lip rising, by the highest spreading values (i.e. highest lip rounding), and highest retraction values (i.e. minimum lip protrusion). However, these two emotions are distinguished by the horizontal and vertical asymmetry values and also, as regards acoustic parameters, by F0, F2 and F3 values (always higher in *joy*, close to neutral values in *disgust*). *Fear* and *anger* is well characterized by the maximum lip opening (LO) and maximum LL displacement (i.e. by the maximum lip opening and maximum lower lip lowering), but are distinguished by F0, F2 and F3 values (which are always higher for *fear*).

2- Whenever different parameters mean values are analyzed contrastively in the realization of the three phonetic targets, it can be noticed that these values change for the same emotion according to the articulatory characteristics both of vowels and consonants. Some parallelism among targets, nevertheless, can be noticed even if the data differs. The LO parameter presents very different values, with regard to *neutral*, in the vowel /’a/ compared with /b/ and /v/. LO shows, reduced values due to the constraints of the labial closure and the labiodental constriction. LO and LL show a similar behavior because the movement of the lower lip affects mainly the LO values. As regards UL, similar values are shown in the realization of /’a/ and /v/ which present, for the different emotions, the minimum upper lip rising with respect to *neutral*. In /v/, in fact, the upper lip does not participate in the labiodental constriction realization and maintains the configuration of the earlier vowel. With respect to /b/, instead, the UL values present a lowering with respect to *neutral* in the different emotions, showing the role of the active articulator in the realization of the bilabial closure.

The LR parameter presents similar values, in different emotions, for the /’a/ and /b/ target against /v/, since the labiodental articulation characterizing this consonant reduces the variability of the labial configuration determining a restricted value range.

For ULP, as for UL, some similarities are found among the values of these parameters in /’a/ and /v/. In /v/ consonant there is the articulatory effect of the contextual vowel, while for /b/, ULP is constrained by the bilabial clousure.

For Left and Right Corner horizontal displacements (LCX and RCX), the data show a correspondence in all the three phonetic targets for each emotion, while for Left and Right Corner vertical displacements (LCY and RCY), similarities can be verified between /’a/ and /v/, where the displacements are wider than /b/, which presents reduced displacements. The left corner also shows a more significant displacement than the right one.

3-The clusters have shown a common tendency to the three articulatory targets by grouping together *joy* and *disgust*. With regard to *joy*, in /’a/ as well as in /b/ and /v/, the maximum negative values are in LCY and RCY corresponding to the lip corner vertical raising characterizing the labial configuration of *joy*. This vertical raising therefore is added to the articulatory linguistic behavior. So *Joy* presents negative LO values, negative LR values, UL and LL positive values and finally ULP and LLP negative values, determining a labial spreaded lip position whith lip corner rising. *Disgust* is characterized in /’a/, /b/ and /v/ by maximum negative LO values, UL and LL positive values, maximum negative LR values with a ULP and LLP reduction, (corresponding to maximum lip retraction) and maximum extension of LCX, RCX and ASYMY. As a result, *disgust* is characterized by a minimum lip opening, maximum spreading and retraction and also with negative right vertical asymmetry.

4-Finally, it is necessary to note that on the basis of the data presented and keeping in mind the data acquisition system adopted by us, three emotions, *surprise*, *sadness* and *fear* (the latter not systematically) result to be not very distinct among themselves. It will be necessary to verify if by comparing the lack of labial characterization these three emotions result to be better differentiated on the basis of the configuration of the upper part of the face (Ekman and Friesen, 1978; Ekman et al., 2002; Hess et al., 1998) and recently Nordstrand et al. (2003), or on the information contained in the acoustic signal, in particular the F0 values (Scherer, 1986; Banse and Scherer, 1996)(for Italian see Magno Caldognetto et al. (1997, 1998)). A future analysis of the relative data of the upper part of the face which was gathered together contemporarily with the lip movements here described, associated with the evaluation of the macro-prosodic characteristics in the emotive speech (e.g., F0 mean, intensity mean, speech rate), will permit the verification of this hypothesis and the presentation of the supporting data on the widely spread belief that for the transmission of the emotions it is certainly necessary to refer to bimodal (audio-visual) communication (Scherer, 2003; Scherer et al., 2003).

Acknowledgements

Part of this work has been sponsored by COMMEDIA (COMunicazione Multimodale di Emozioni e Discorso in Italiano con Agente animato virtuale, CNR Project C00AA71), PF-STAR (Preparing Future multiSensorial inTer-Action Research, European Project IST- 2001-37599, http://pfstar.itc.it) and TICCA (Tecnologie cognitive per l'Interazione e la Cooperazione Con Agenti artificiali, joint CNR - Provincia Autonoma Trentina Project).

References

Aubergé, V., and Cathiard, M. (**2003**). "Can we hear the prosody of smile?" Speech Communication, **40**, 87–98.

Banse, R., and Scherer, K. R. (**1996**). "Acoustic profiles in vocal emotion expression," Journal of Personality and Social Psycology, **70**, 614–636.

Camras, L. A., Holland, E. A., and Patterson, M. J. (**1993**). "Facial espression," In M. Lewis and J. Havilland (Eds.), Handbook of Emotions. The Guilford Press, pp. 199–208.

deGelder, B., and Vroomen, J. (**2000**). "The perception of emotions by ear and eye," Cognition and emotion, **14**, 289–311.

deGelder, B., Vroomen, J., and Bertelson, P. (**1998**). "Cross-modal bias of voice tone on facial expression: upper versus lower halves of a face," in *Proceedings of ICSLP'98* (Sydney, Australia), pp. 93–96.

Deller, J. R., Proakis, J. G., and Hansen, J. H. (**1993**). *Discrete Time Processing of Speech Signals* (Prentice Hall PTR), ISBN 0023283017.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (**2003**). "Emotional speech: towards a new generation of databases," Speech Communication, **40**, 33–60, ISSN 0167-6393.

Ekman, P., and Friesen, W. V. (**1978**). *Facial Action Coding System* (Consulting Psychologists Press Inc., Palo Alto, USA).

Ekman, P., Friesen, W. V., and Hager, J. C. (**2002**). *New version of the Facial Action Coding System (eBook)* (A Human Face Publ., Salt Lake City, USA).

Fant, G., and Lin, Q. (**1987**). "Glottal source - vocal tract acoustic interaction," STL-QPSR, pp. 13–27.

Fischer-Jørgensen, E. (**1990**). "Intrinsic F0 in tense and lax vowels with special reference to German," Phonetica, **47**, 99–140.

Hess, U., Kappas, H., and Scherer, K. R. (**1998**). "Multichannel communication of emotions: synthetic signal production," in K. R. Scherer Ed., Facets of emotion: recent research. Lawrence Erlbaum Ass. Publ., pp. 161–182.

Magno Caldognetto, E., Vagges, K., and Zmarich, C. (**1995**). "Visible articulatory characteristics of the italian stressed and unstressed vowels," in *Proceedings of ICPhS 95* (Stockholm, Sweden), pp. 366–369.

Magno Caldognetto, E., Zmarich, C., Cosi, P., and Ferrero, F. E. (**1997**).

"Italian consonantal visemes: relationship between spatial/temporal articulatory characteristics and coproduced acoustic signal," in *Proceedings of AVSP'97* (Rhodes, Greece), pp. 5–8.

Magno Caldognetto, E., Zmarich, C., and Cosi, P. (**1998**). "Statistical definition of visual information for italian vowels and consonants," in *Proceedings of AVSP'98* (Terrigal, Australia), pp. 135–140.

Magno Caldognetto, E., Cosi, P., Drioli, C., Tisato, G., and Cavicchio, F. (**2003**). "Coproduction of speech and emotion: bimodal audio-visual changes of consonant and vowel labial targets," in *Proceedings of AVSP 03* (S. Jorioz, France), pp. 209–214.

Massaro, D., and Egan, P. (**1998**). "Perceiving affect from the voice and the face," Psychological Bulletin, **3**, 1021–1032.

Nordstrand, M., Svanfeldt, G., Granstrom, B., and House, D. (**2003**). "Measurements of articulatory variations and communicative signals in expressive speech," in *Proceedings of AVSP 03* (S. Jorioz, France), pp. 233–237.

Palmer, A., Richard, D., and Strobeck, C. (**1986**). "Fluctuating asymmetry: measurements, analysis and patterns," Annual Review of Ecology and Systematics, **17**, 391–421.

Scherer, K. R., Johnstone, T., and Klasmeyer, G. (**2003**). "Vocal expression of emotion." In R. J. Davidson, H. Goldsmith, K. R. Scherer (Eds.). Handbook of the Affective Sciences. New York: Oxford University Press, pp. 433–456.

Scherer, K. R. (**1986**). "Vocal affect expression: a review and a model for future research," Psychological Bullettin, **99**, 143–165.

Scherer, K. R. (**2003**). "Vocal communication of emotions: a review of research paradigms," Speech Communication, **40**, 227–256.

Schroeder, M., Aubergé, V., and Cathiard, M. (**1998**). "Can we hear smile?" in *Proceedings of ICSLP'98* (Sydney, Australia), pp. 559–662.

Stevens, K. N. (**1998**). *Acoustic Phonetics*, Current studies in linguistics (The MIT Press, Cambridge, Massachussetts).

Whalen, D. H., and Levitt, A. G. (**1995**). “The universality of intrinsic F0 of vowels,” Journal of Phonetics, **23**, 349–366.

A Clusters

[FIGURE A.1 APPROX. HERE.]

Figure A.1. Clusters for the vowel /'a/

[FIGURE A.2 APPROX. HERE.]

Figure A.2. Clusters for the vowel /b/

[FIGURE A.3 APPROX. HERE.]

Figure A.3. Clusters for the vowel /v/

[FIGURE A.4 APPROX. HERE.]

Figure A.4. Clusters for acoustic parameters

List of Figures

List of Tables