Interrelation between Speech and Facial Gestures in Emotional Utterances: A single subject study

Carlos Busso, *Student Member, IEEE,* and Shrikanth S. Narayanan, *Senior Member, IEEE*

***Abstract*—The verbal and non-verbal channels of human communication are internally and intricately connected. As a result, gestures and speech present high levels of correlation and coordination. This relationship is greatly affected by the linguistic and emotional content of the message. The present paper investigates the influence of articulation and emotions on the interrelation between facial gestures and speech. The analyses are based on an audio-visual database recorded from an actress with markers attached to her face, who was asked to read semantically neutral sentences, expressing four emotion states (neutral, sadness, happiness and anger). A multilinear regression framework is used to estimate facial features from acoustic speech parameters. The levels of coupling between the communication channels are quantified by using Pearson's correlation between the recorded and estimated facial features. The results show that facial and acoustic features are strongly interrelated, showing levels of correlation higher than $r = 0.8$ when the mapping is computed at sentence-level using spectral envelope speech features. The results reveal that the lower face region provides the highest activeness and correlation levels. Furthermore, the correlation levels present significant inter-emotional differences, which suggest that emotional content affect the relationship between facial gestures and speech. Principal component analysis (PCA) shows that the audiovisual mapping parameters are grouped in a smaller subspace, which suggests that there is an emotion-dependent structure that is preserved from across sentences. The results suggest that this internal structure seems to be easy to model when prosodic-features are used to estimate the audiovisual mapping. The results also reveal that the correlation levels within a sentence vary according to broad phonetic properties presented in the sentence. Consonants, especially unvoiced and fricative sounds present the lowest correlation levels. Likewise, the results show that facial gestures are linked at different resolutions. While the orofacial area is locally connected with the speech, other facial gestures such as eyebrow motion are linked only at the sentence-level. The results presented here have important implications for applications such as facial animation and multimodal emotion recognition.**

***Index Terms*—Facial motion, articulatory movements, Affective state, Speech acoustic.**

I. Introduction

Manuscript received September 6, 2006; revised March 14, 2007. This work was supported in part by funds from the National Science Foundation (NSF) (through the Integrated Media Systems Center, an NSF Engineering Research Center, Cooperative Agreement No. EEC-9529152 and a CAREER award), the Department of the Army, and a MURI award from the Office of Naval Research. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. This work was performed when the authors were with Integrated Media Systems Center, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089.

C. Busso and S. Narayanan are with the Integrated Media Systems Center, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA (e-mail: busso@usc.edu, shri@sipi.usc.edu)

IN addition to speech, non-verbal communication plays an important role in day-to-day interpersonal human interaction. People simultaneously use their hand, change their facial expression and control the tone and their energy of the speech to consciously or unconsciously express or emphasize specific messages. The fact that all these communicative channels interact and cooperate to convey a desired message suggests that gestures and speech are controlled by the same internal control system [1], [2], [3], [4].

Human communication, manifested through a combination of verbal and nonverbal channels used in normal interaction, is a result of different communicative components that mutually modulate gestures and speech in a non-trivial manner. Notable among them are the linguistic, emotional and idiosyncratic aspects of human communication. The linguistic aspect defines the verbal content of what is expressed. In addition, the underlying articulation also affects the appearance of the face. Each phoneme is produced by the activation of a number of muscles that simultaneously shape the vocal tract and the face, in regions beyond the oral aperture [2]. Likewise, the emotional state of the speaker is directly expressed through both gestures and speech. Communicative channels such as facial expressions [5], [6], head motion [7], pitch [8], [9] and short time spectral envelope [10] all present specific patterns under emotional states. Similarly, the idiosyncratic aspects also influence the patterns of human speech and gestures, and are dependent on culture and social environment. These also include personal styles, such as the rate of the speech and intensity and manner of expressing emotions [5], [11].

Speech and gestures are directly connected and manipulated by one or more of these components of human communication. For example, the shape of the orofacial area is highly connected with the linguistic content [2], [12], hand gestures are related with idiosyncrasy [1] and facial expressions are most of the time triggered by emotions [5], [6]. Although, each communicative channel has been separately analyzed in the literature [1], [2], [5], [7], [8], [9], [13], [14], relatively few studies have addressed the interaction between gestures and speech as a function of the multiple aspects of human communication. Understanding the linguistic, emotional and idiosyncratic effects on the gesture-speech relationships is a crucial step toward improving a number of interesting applications such as emotion recognition [15], human-computer interaction [16], [17], human-like conversational agents [3] and realistic facial animation [18], [19], [20], [21]. It will also provide useful insights into human speech production and perception [22].

Toward understanding how to model expressive human

communication, the present paper focuses on the linguistic and emotional aspects of human communication and their influences on the relationships between gestures and speech. We investigate the interrelation between facial gestures, such as lip, eyebrow and head motion and acoustic features that represent the vocal tract shaping and prosody of speech. We analyze which gestures are more related with speech and how this relation changes in the presence of four different expressive emotion states: neutral, sadness, happiness and anger. The analysis presented here is based on a database recorded from a single actress with markers attached to her face, which provide detailed information about her facial expressions. We analyze how active facial gestures are as function of different emotions. The results indicate that the rate of facial gesture displacements for emotional utterances significantly differ from the ones observed in neutral speech. For example, the activeness, as quantified by the Euclidean distance between the facial features and their sentence-mean vector, for happiness and anger is more than 30% higher than for neutral speech. To quantify the relationship between facial gestures and speech, we compute Pearson's correlation between the original facial features and the speech-based estimated sequences generated with multilinear regression. This framework is implemented both at sentence-level, in which the mapping parameters are computed for each sentence assuming that the facial features are known, and at global-level, in which the mapping parameters are estimated using the entire database. At sentence-level, the audio-visual mapping presents significant inter-emotional differences, showing the influence of emotions in the interrelation between gesture and speech. When the mapping is estimated at global-level the correlation levels not surprisingly, decrease. However, the results indicate that there is a clear emotion-dependent structure that can be learned using more sophisticated techniques than linear mapping. Furthermore, the results reveal that the correlation levels also decrease for consonants, especially unvoiced and fricative sounds, and silence regions, when compared to other corresponding broad speech classes of voiced and sonorant sounds.

The main contribution of this paper is the analysis of the interplay between gestures and speech as a function of emotion and articulation. As far as the authors are concerned, this is the first attempt to quantize the linguistic and affective influence on the relationship between facial gestures and speech. As specific results based on the detailed framework presented here to analyze gestures and speech at different levels (global, sentence and phoneme level), important findings are presented:

- The existence of a strong relationship between facial gestures and speech, which greatly depends on the linguistic content of the sentence
- The existence of a low-dimension emotional-dependent structure in the gesture/speech mapping that is preserved across sentences
- An emotional-dependent structure in the mapping that seems to be easier to model when prosodic features are used to estimate the facial features
- Multiresolution nature of the relationship between facial gestures and speech, both in (feature) space and time

These results can guide the design of better models to capture the complex relationship between facial gestures and speech. Specifically, human-machine interfaces for applications such as games and educational and entertainment systems represent some of the wide range of applications that can be greatly enhanced by properly modeling and including human-like capabilities.

The rest of this paper is organized as follows. Section II describes related work about co-analysis of gestures and speech. Section III introduces the database, the features and the mathematical framework utilized in this paper. Section IV describes and discusses our results about the relationship between the various modalities. We analyze how active the facial features are during emotional speech and how much the audio-visual mapping is affected by articulation and emotions. Section V presents the audio-visual mapping at the phoneme level. We explore whether different broad phonetic classes have stronger or weaker relationships when the mapping is estimated at sentence level. Section VI discusses the implications of the results presented in this paper in the areas of facial animation and multimodal emotion recognition. Finally, Section VII gives the conclusion and future directions of this research.

II. Related Work

Based on high levels of correlation found between acoustic features and various human gestures, researchers have suggested that an internal control simultaneously triggers the production of both speech and gestures, sharing the same semantic meaning in different communication channels [1], [2], [3], [4]. Cassell *et al* mentioned that they are not only strongly connected, but also systematically synchronized in different scales (phonemes-words-phrases-sentences) [3]. This theory, referred to as the *excitatory* hypothesis [4] implies that a joint analysis of these modalities is needed to fully understand human communication.

The relation between gestures and speech has been studied under different perspectives. One line of research has focused on analyzing this relation in terms of conversational functions, also known as regulators and conversational signals [11]. Gestures and speech co-occur to fulfill functions during conversation, such as acknowledging agreement, changing turns and asking for clarification [16]. Valbonesi *et al* studied co-occurrence of *events* during speech and hand gestures. They showed that most of the acoustic events, defined as maximum and minimum in the pitch and the RMS energy, occur during hand gesture strokes [4]. Graf *et al* concluded that head and eyebrow motions are consistently correlated with the prosodic structure of the speech [19]. Granström *et al* conducted perceptual experiments showing that head and eyebrow motion help to stress prominence in speech [23]. They have also shown that smile is the strongest indicator of positive feedback. Such findings have been used in designing human-like virtual agents. Cassell *et al* proposed a rule-based system to generate facial expressions, hand gestures, head nod, eyebrow motion and spoken intonation, which were properly

synchronized according to rules [3]. In [16], they extended their work to create a human-like agent, called REA that was able to respond to discourse function using gestures.

Another line of research has analyzed the relation between speech and gestures, especially facial expressions, as results of articulatory processes. Vatikiotis-Bateson and Yehia showed that facial expressions are directly connected with the articulatory production [2]. They argued that the production of speech shapes the vocal tract and deforms the face, affecting regions quite further from the oral aperture. In [12], they continued their analysis presenting results about the relation between facial expressions, vocal-tract configuration and speech. They concluded that most of the facial motions can be predicted from acoustic features (*Line Spectral Pair* (LSP)) using linear estimations. In [13], they extended those results to non-linear mapping, showing higher accuracy. Jiang *et al* also studied the relation between facial, articulatory and acoustic features, using multilinear regression analysis [24]. The focus of their work was to quantify the audio-visual relationship for the *consonant-vowel* (CVs) syllables C/a/, C/i/ and C/u/. They concluded that the mapping was syllable-dependent, since they found better results for C/a/ than for C/i/ and C/u/. They also found differences in the correlation levels between the four speakers considered, suggesting that the mapping was speaker-dependent. Following a similar approach, Barker and Berthommier studied the mapping for isolated words with a fixed vowel-consonant structure [25]. They concluded that the correlation levels between facial (jaw and lips) and acoustic (LSP) features are higher during vowels than during consonants. One common aspect in all these works, however, is that the data used provides only sparse information of the facial area with relatively few markers on the subject's face (≤ 20).

With advances in data acquisition capabilities, probabilistic frameworks such as *Hidden Markov Models* (HMMs) and *Dynamic Bayesian Networks* (DBNs) have been successfully used to model both speech and facial expressions. Recently, a variety of these probabilistic frameworks have been proposed to jointly model facial expressions and speech for applications such as audio-visual emotion recognition [26], audio-visual speech recognition [27], user modeling [28], [29], and facial animation [7], [18], [30]. We believe that these statistic learning frameworks are attractive schemes to capture the temporal relationship between gestures and speech in the presence of emotions.

Another important aspect that influences the relation between speech and gestures is the emotion conveyed by the speakers. Since each communication channel can be greatly modulated under different emotions, it is expected that their relation will also be emotion-dependent. Many previous studies have shown that speech is colored by emotional effects [8], [9], [10]. Emotions influence not only supra-segmental characteristics of the speech (prosody and energy), but also short-time spectral envelope features such as *Mel-Frequency Cepstrum Coefficients* (MFCCs) [10]. The face is also highly affected by the affective state of the speaker. For instance, the group led by Ekman has extensively analyzed the relation between facial expressions and emotions. After studying apex poses of expressive faces, they concluded that specific facial patterns are displayed under certain family of emotions [5], [6]. The effect of emotions in the orofacial area has been especially analyzed for realistic facial animations. Nordstrand *et al* studied the effect of emotions in the shape of the lips for vowels. They concluded that there are significant differences in the patterns presented in neutral and emotional speech [31]. Similar inferences were also reported in [32]. As a direct consequence of these results, emotion-dependent models to synchronize lips movement with speech have been developed for human-like facial animations [20]. Other facial gestures such as head motion [7] and eyebrown motion [11] also show strong differences when the affective state of the speaker changes. Although these communication channels have been separately studied under different emotions, relatively few efforts have focused on the influence of emotion in the relation between these communication channels. This paper explores the influence of emotions on the relation between facial gestures and speech.

III. Methodology

The approach followed in this study to analyze the relation between speech and facial expression is to estimate the Pearson's correlation between the original facial expression signal, F_{Facial} and a predicted signal, $\widehat{F}_{Facial}$, estimated from speech acoustic features using a linear mapping. This approach is similar to the method presented by Yehia *et al* in [12]. Notice that it is also possible to estimate the acoustic features based on facial expression as presented in [12], [24], [25], which could be very useful for applications such as audio-visual speech recognition. However, for analysis simplicity we implemented only the unidirectional speech to gesture approach.

Voiced speech production is usually modeled as a quasi-periodic source signal that excites the vocal-tract transfer function; unvoiced sounds are modeled with noise source excitation. The excitation models both the air exhaled from the lungs through the trachea and pharynx, and the vocal cord, which adjusts its tension to create the oscillatory signal. The vocal-tract transfer function models the pharyngeal, oral and nasal cavities, which carries much of the phonetic information of the speech. Based on this broad description, two different sets of acoustic features can be defined: prosodic features, which provide the tonal and rhythmic aspect of the speech contained in the source; and, the vocal-tract acoustic features that model the time-varying vocal-tract transfer function [33]. In this paper, pitch and energy were used as prosodic features, and MFCCs were used as vocal-tract features. We chose MFCCs rather than other short-time spectral envelope, because our preliminary results showed that MFCCs have higher correlation with facial gestures. Furthermore, unlike LSP that are based on a parametric model using the all-pole assumption, MFCCs can handle zeros in the nasal sound spectrum, since they directly model the signal spectrum. Therefore, the use of MFCCs may improve the mapping during nasal phonemes, as discussed in Section V.

To analyze facial gestures, the face is divided in Voronoi cells centered on the facial markers. Each marker provides

(a) (b)

Fig. 1. Audio-visual database collection. (a)The figure shows the facial marker layout, and (b) the figure shows the motion capture system.

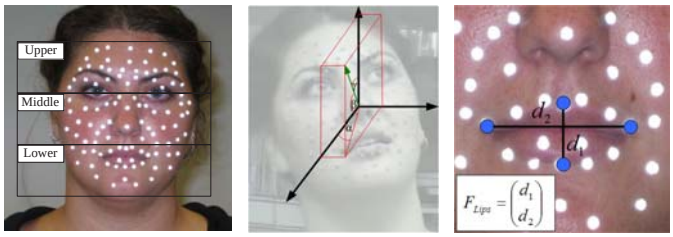
local information about her facial movements. Furthermore, rigid head motion was included as a part of the facial features. The shape of the lips and eyebrow, which are the most important distinguishable part of the face, were also considered in this study. These features were parameterized using specific facial markers, as explained in Section III-B. The complete set of the markers, head motion, eyebrow and lip features is referred together here on as facial features.

A. Audio-Visual Database

The audiovisual database used in this work was recorded from an actress, who was asked to read a custom-made, phoneme-balanced corpus four times, expressing different emotions (happiness, sadness, anger and neutral state). The use of a set of emotion categories simplifies the recognition and communication of emotional state for both people and computers [34]. Therefore, for the sake of simplicity, and for guiding interface design, the proposed analysis uses emotion categories, which is more appealing than the attribute characterization of emotion such as *arousal* and *valence*.

A detailed description of her facial expression and rigid head motion was acquired by using 102 markers attached to the face (Figure 1 (a)). A VICON motion capture system with three cameras was used to capture the 3D position of each marker (Figure 1 (b)). The sampling frequency was set to 120 Hz. The recording was made in a quiet room using a close talking SHURE microphone at a sampling rate of 48 kHz. The markers' motions and the aligned audio were simultaneously captured by the system. In total, 612 sentences were used in this work. Note that the actress did not receive any special instruction about how to express the target emotions, and was asked to be natural.

Even though acted facial expressions have some differences with genuine facial expression [5], databases based on actors have been widely used in the analysis of emotions. The main advantage of this setting is that a balanced corpus can be designed in advance, to include a wide range of phonetic and emotional variability. In addition, the proposed recording setting allows us to use markers that provide detailed facial information which could be very difficult to obtain in a more unconstrained production scenario. Such data are particularly useful for the types of analyses presented in this paper.



(a) (b) (c)

Fig. 2. Face parameterization. (a) the figure shows the facial markers subdivision (upper, middle and lower face regions), (b) the figure shows head motion features, and (3) the figure shows the lip features.

B. Feature extraction

The pitch (F0), the RMS energy and the 13 MFCC coefficients were extracted using the Praat speech processing software [35]. The analysis window was set to 25 milliseconds with an overlap of 8.3 milliseconds producing 60 frames per second. The smoothing and interpolation options of the Praat software were applied to remove any spurious spike in the pitch estimates and to avoid zeros in unvoiced/silence regions, respectively. In addition, the first and second derivatives of the pitch and energy were added to the prosodic feature vector to incorporate their temporal dynamics. As shown in [36], [25], these dynamic features improve the correlation levels of the audio-visual feature mapping. The first coefficient of the MFCCs was removed, since it provides information about the RMS energy rather than the vocal-tract configuration. The velocity and acceleration coefficients were also included as features. The dimension of this feature vector was reduced from 36 to 12 using *Principal Component Analysis* (PCA). This number was chosen to contain 95% of the variance of the MFCC features. This post-processed feature vector is what will be referred here on as MFCCs.

After the motion data were captured, all the markers were translated to make a nose marker at the local coordinate center of each frame, removing any translation effect. After that, the frames were multiplied by a rotational matrix, which compensates for rotational effects. This matrix was constructed for each frame as follows: A neutral pose of the face was chosen as a reference frame, which was used to create a 102×3 matrix, M_{ref}, in which the row of M_{ref} has the 3D position of the markers. For the frame t, a similar matrix M_t was created by following the same marker order as the reference. After that, the *Singular Value Decomposition* (SVD), UDV^T, of matrix $M_{ref}^T \cdot M_t$ was calculated. Finally, the product of VU^T gave the rotational matrix, R_t, for the frame t [37].

$$M_{ref}^T \cdot M_t = UDV^T \tag{1}$$

$$R_t = VU^T \tag{2}$$

After compensating for the translation and rotation effects, the remaining motion between frames corresponds to local displacements of the markers, which define the subject's facial expressions. To synchronize the frames of the acoustic features with the frames of facial features, the marker data was downsampled from 120 to 60 frames per second.

In the analyses, each of the facial markers, except the reference nose marker, was used as a facial feature. The markers were grouped into three main areas: upper, middle and lower face regions (Figure 2 (a)). The upper face region includes the markers above the eyes in the forehead and brow area. As Ekman *et al* observed, this facial area is strongly shaped by the emotion conveyed in the facial expression [6]. The lower face region groups the markers below the upper lip, including the mouth and jaw. As discussed in Section IV-B, this area is modulated not only by the emotional content, but also by the articulatory processes. Finally, the middle face region contains the markers in the facial area between the upper and lower face region (cheeks). This subdivision will be used to summarize the aggregated results in the tables presented in Sections IV and V.

In addition to the aggregated facial region features, parametric features describing the head, eyebrow and lip motion were also analyzed. Low dimensional features were selected that capture articulatory information (especially for the lips), and that are affected by emotional modulation. The matrix R_t defines the three Euler angles of the rigid head motion, which are added as visual features (Figure 2 (b)). Furthermore, the eyebrows were parameterized with a two-dimensional feature vector, computed by subtracting the position of two chosen markers in the right eyebrow from a neutral pose. After that, the vector was normalized in the range 0 to 1. Notice that right and left eyebrow motions are assumed symmetrical. Although Cavé *et al* suggested that there can be some differences between the magnitude of the two eyebrows' motions [14], this symmetry assumption is in general reasonable (especially for the subject analyzed). The lip features describe the width and the height of the opening area of the mouth. The lip features were computed by measuring the Euclidean distance between the markers shown on Figure 2 (c). This two-dimensional feature vector relates to three articulatory parameters that describe the shape of the lips: *upper lip height* (ULH), *lower lip height* (LLH) and *lip width* (LW).

C. Audio-visual mapping framework

In this study, our goal is to analyze the temporal relation between facial gestures and speech. We desire to measure the areas in the face that are shaped or modified by the articulatory and prosodic aspects of the speech. For this purpose, the Pearson's correlation was chosen as the measure to discern the relationship between the acoustic and facial features. Correlation provides a solid mathematical tool to infer and quantify how connected or disconnected different streams of data are. Its results are easy to interpret and no probability density functions need to be estimated, such as in mutual information calculation between feature streams.

In general, the speech and the facial features span different spaces which have dissimilar dimensions and scales. Therefore, preprocessing steps need to be implemented before computing Pearson's correlation. The complete framework proposed in this paper is depicted in Figure 3.

The first step after extracting the features is to map the acoustic features into the facial feature space. In this paper

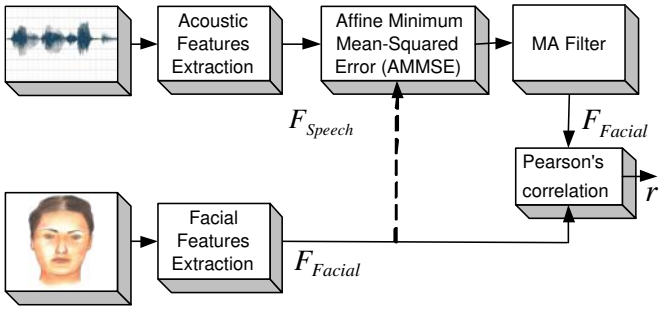


Fig. 3. Linear estimation framework to quantify the level of coupling between facial gestures and acoustic features. AMMSE is used to map the acoustic features into the facial feature space. A filtered version of this estimated signal, $\widehat{F}_{Facial}$, is used to measure the Pearson's correlation.

we used the *Affine Minimum Mean Square Error estimator* (AMMSE), which is defined by a transformation matrix, T, and a translation vector, M,

$$\widehat{F} = T \cdot F_{Speech} + M \tag{3}$$

$$T = K_{FS} \cdot K_S^{-1} \tag{4}$$

$$M = -K_{FS} \cdot K_S^{-1} \cdot U_S + U_F \tag{5}$$

where K_{FS} is the cross covariance matrix between the facial and the acoustic features, K_S and U_S are the covariance and mean vector of the speech features, and U_F is the mean vector of the facial features. This is an optimum linear estimator under mean square sense. We chose AMMSE because it is simple and has less overfitting problems than other non-linear techniques. In addition, linear estimation has shown better generalization than non-linear mappings in related studies [38]. Notice that in some applications such as facial animation, other mapping techniques such as *Artificial Neural Networks* (ANN) or HMMs could give better results than AMMSE [13], [25], [30], [38].

After the acoustic features are mapped onto the facial features space, a *moving average window* (MA) is applied to smooth the estimated signal, producing $\widehat{F}_{Facial}$. The final step is to compute the Pearson's correlation between the estimated and real facial features, in each dimension of the facial feature vector.

The code was implemented for off-line processing in Matlab. Since the models are linear, the computational requirements are not as high as other sophisticated time series modeling frameworks such as HMMs and DBNs.

Two implementations of this framework are presented. The first implementation is when the target facial features are known and the parameters of the mapping (T_u, M_u) are estimated for each sentence (u). This implementation is referred to as sentence-level mapping. The second implementation of this framework is when the parameters of the mapping are estimated using the entire database. In this case, the same parameters (T_{emo}, M_{emo}) for each emotional category are used to estimate each sentence. This implementation is referred to as global-level mapping. Both implementations provide

useful information in different applications, as we will discuss in the next section.

IV. Results of the Audio Visual Mapping

In this section, the facial and acoustic features are jointly analyzed in terms of the emotional categories considered in this work. Before studying the relation between facial expressions and speech, Section IV-A discusses the activeness of the facial features during speech. Knowing the motion rate of facial gestures during emotional speech is important to correctly interpret the results presented in this paper. Following the methodology described in Section III-C, Sections IV-B and IV-C analyze the results of the correlation between the original and estimated facial features at the sentence-level and global-level mapping, respectively. Section IV-D discusses the structure of the audio-visual mappings, by studying Eigen spaces of the mapping parameters (Equation 4).

A. Facial activeness during speech

During speech, some facial areas are naturally more active than others. This section explores the motion rate of facial gestures in terms of the articulatory and emotional aspects of human communication.

For each sentence, the *displacement coefficient*, Ψ, described in Equation 6, was calculated to measure the activeness of the facial features. This coefficient is computed as the average Euclidean distance between the facial features and their mean vector, at sentence-level:

$$\Psi_u = \frac{1}{N_u} \sum_{i=1}^{N_u} \mathrm{D_{eq}}(\vec{X}_i^u, \vec{\mu}^u) \tag{6}$$

where N_u is the number of frames in sentence u, $\vec{\mu^u}$ is the mean vector, and $\mathrm{D_{eq}}$ is the Euclidean distance, defined as:

$$\mathrm{D_{eq}}(\vec{X}, \vec{Y}) = \sqrt{\sum_{d=1}^{D} (x_d - y_d)^2} \tag{7}$$

where D is the dimension of the facial features. The *average displacement coefficient*,$\bar{\Psi}$, is obtained by computing the mean of the *displacement coefficients* across the N utterances, for each emotional class:

$$\bar{\Psi} = \frac{1}{N} \sum_{u=1}^{N} \Psi_u \tag{8}$$

Notice that in this analysis the acoustic features are not used, since Ψ is completely defined by the facial features. This coefficient provides global-level description of the activeness of facial areas and gestures.

The results for the *displacement coefficient* are presented in Tables I and II, and in Figure 4. Table I presents the *average displacement coefficient*, $\bar{\Psi}$, for the facial features related to head, eyebrow and lip motion, in terms of emotional categories. In addition, Table I summarizes the average activeness of the markers in the three facial areas described in Section III-B: upper, middle and lower face regions. To infer whether the emotional differences in the results presented in Table I are significant or not, one-way *Analysis of Variance* (ANOVA) tests were performed. For head (F[3,622], p=0.000), eyebrow (F[3,622], p=0.000) and lip motion (F[3,619], p=0.000), Table II provides the p-values for the multiple comparison tests. The same statistical test was individually applied to each marker. Instead of reporting the p-value results of the facial markers, Table II gives the percentage of the markers in the facial regions in which the differences were found significant, using a 95% confidence interval ($p \leq 0.05$). Figure 4 shows a visual representation of facial area activeness, per emotional category. This figure was created by computing the *average displacement coefficient* for each facial marker. Then, the coefficients were normalized across emotions in the range between 0 and 1. After that, gray-scale colors were assigned to each marker according to the palette shown in Figure 5. Finally, the Voronoi cells centered in the markers were colored according to the gray-scale assigned to the markers.

TABLE I
AVERAGE ACTIVENESS OF FACIAL FEATURES DURING EMOTIONAL SPEECH (*Neu*=NEUTRAL, *Sad*=SADNESS, *Hap*= HAPPINESS, *Ang*= ANGER)

Facial Area	Neu	Sad	Hap	Ang
Head Motion [deg]	2.30	4.52	5.05	5.00
Eyebrow	0.05	0.07	0.12	0.12
Lips	4.69	3.68	6.24	6.94
Upper region	0.72	0.85	1.51	1.37
Middle region	0.92	0.90	1.43	1.52
Lower region	3.24	2.49	4.20	4.47

TABLE II
STATISTICAL SIGNIFICANT OF INTER-EMOTION ACTIVATION DIFFERENCES (*Neu*=NEUTRAL, *Sad*=SADNESS, *Hap*= HAPPINESS, *Ang*= ANGER)

Facial Area	Neu-Sad	Neu-Hap	Neu-Ang	Sad-Hap	Sad-Ang	Hap-Ang
Head Motion	0.000	0.000	0.000	0.176	0.202	1.000
Eyebrow	0.000	0.000	0.000	0.000	0.000	1.000
Lips	0.000	0.000	0.000	0.001	0.000	0.000
Upper region	34.48%	100.00%	100.00%	100.00%	100.00%	58.62%
Middle region	25.64%	100.00%	100.00%	100.00%	100.00%	48.72%
Lower region	100.00%	100.00%	100.00%	100.00%	100.00%	41.38%

Figure 4 and Table I show that during speech, the lower face region, specifically the jaw, is the most active area in the face. This result confirms the important role of articulation in the dynamic motion of the facial expressions. However, when the values of the *displacement coefficients* associated with neutral speech are compared with those with emotional speech in each of the facial features, significant differences are found (Table II). Note that the lexical content and syntactic structure of the utterances used in each emotional category were identical, since the same sentences were used for each emotion. Therefore, the inter-emotional differences shown in Figure 4 and Table I can be attributed primarily to different emotional modulation of facial gestures.

In agreement with previous works [6], [31], [32], the results presented here show that the lower face region conveys important emotional clues. Table I and Figure 4 indicate that the inter-emotional differences in the *average displacement coefficient* of the markers are significant, with the exception of the pair happiness-anger, in which only 41.38% of the markers were found with significant differences (see Table II). Notice

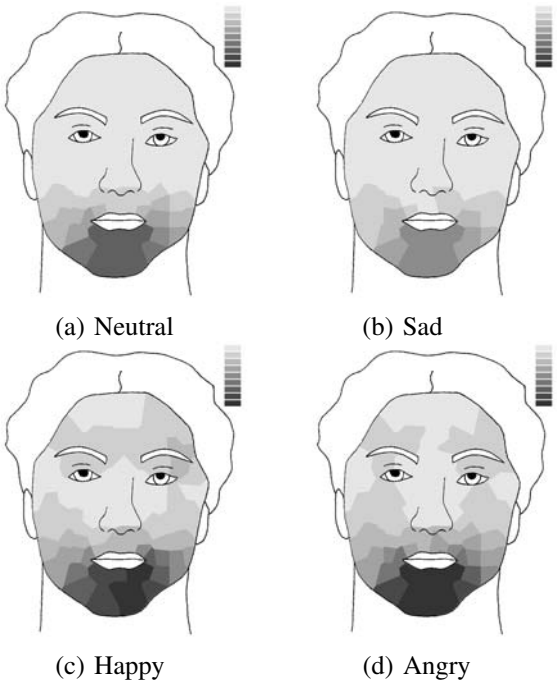


Fig. 4. Facial activeness during speech. The figures show that during speech, the lower face region is the most active area in the face. It also shows inter-emotional differences. During happiness and anger, the activeness of the face is higher than during neutral state. Conversely, during sadness the activeness of the face decreases.

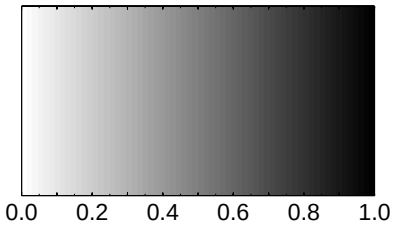


Fig. 5. Palette used in the plots. Darker shadings imply higher activeness (Figure 4) or higher correlation (Figure 6 and 7).

that $\bar{\Psi}$ for happy and angry sentences in the lower face region are more than 30% active than $\bar{\Psi}$ in neutral speech.

In the upper and lower face regions, the results in Tables I and II show that two emotional clusters are clearly grouped: happiness-anger and sadness-neutral. This result is also observed for the *displacement coefficient* in the head and eyebrow motion. As an aside, it is interesting to note that similar grouping trends were observed in the analysis of acoustic speech features presented by Yildirim *et al* [10], suggesting that these emotional categories share similarities across modalities.

The average activeness in sad and neutral sentences are very similar in the upper and middle face regions (see Figure 4), suggesting that those areas of the face are not modified during sad sentences. Notice that this result does not disagree with previous works, which have indicated that facial poses for sadness presents significant differences in those areas compared to neutral poses (inner corners of the eyebrows and cheeks are raised [5], [6]), since the *displacement coefficient* measures the average variance of facial gestures rather than the pose of the face itself (mean vector is removed in Equation 6). Conversely, in happy and angry sentences, the activeness of the upper and middle face region is over 60% more active than in neutral sentences, showing the differences in emotional modulation in those areas.

In general, similar trends were observed in the *displacement coefficient* for head, eyebrow, and lip motion. However, some important results are worth highlighting. The activeness of head motion for sad and neutral sentences are significantly different. Furthermore, the difference between the *average displacement coefficient* for angry and happy utterances for lips features is also significant. These results suggest that this coefficient could be used to discriminate between these pairs of emotional categories.

TABLE III
SUMMARY CORRELATION FOR SENTENCE-LEVEL MAPPING
(*N*=NEUTRAL, *S*=SADNESS, *H*= HAPPINESS, *A*= ANGER)

Facial Area	Prosodic				Mfcc			
	N	S	H	A	N	S	H	A
Head Motion	0.60	0.62	0.57	0.57	0.86	0.89	0.88	0.87
Eyebrow	0.52	0.54	0.53	0.53	0.76	0.80	0.85	0.85
Lips	0.57	0.60	0.57	0.57	0.88	0.89	0.90	0.89
Upper region	0.53	0.56	0.55	0.53	0.80	0.83	0.85	0.84
Middle region	0.51	0.56	0.54	0.53	0.82	0.84	0.86	0.85
Lower region	0.53	0.57	0.55	0.54	0.87	0.87	0.88	0.87

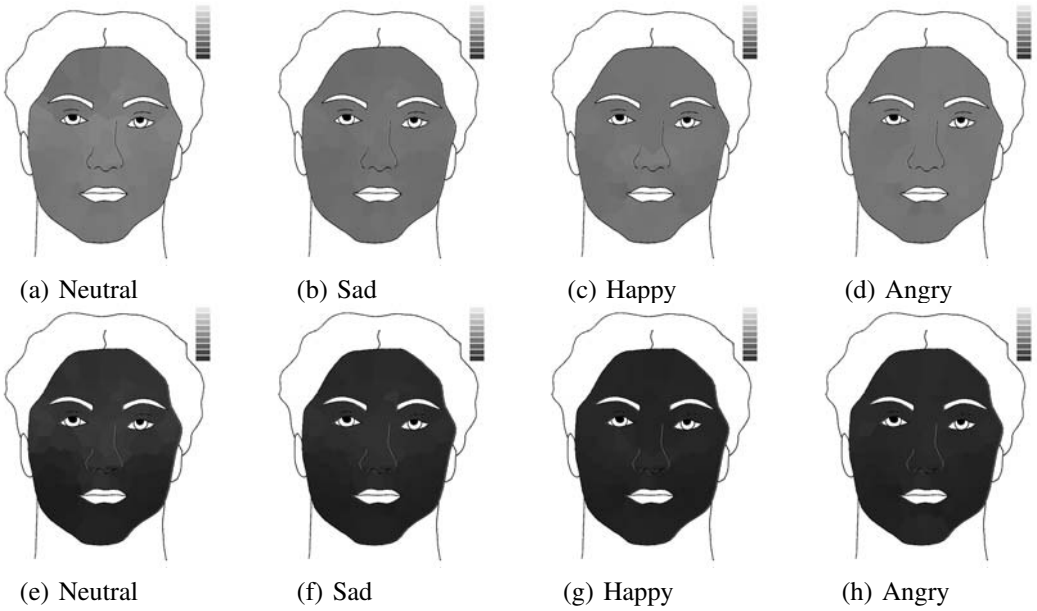
B. Sentence-Level mapping

In the previous section, the facial gestures were analyzed without considering acoustic features. Since gestures and speech are strongly interconnected, a deeper analysis of facial expressions needs to consider acoustic features. In this section, the audio-visual mapping framework presented in section III-C is used to shed light into the underlying relations between acoustic and facial features. The parameters of the linear transformation are calculated at the sentence-level, assuming that the acoustic and facial features are known. Although this is clearly not useful for an application such as animation, in which the facial features are unknown, the results presented here are important to understand better the coupling between speech and facial gestures. Also, areas such as multimodal emotion recognition can benefit from understanding the relation between the modalities.

In this section we are specifically interested in studying whether the correlation between the original and estimated facial features are affected by the emotional and linguistic content (lexical, syntactic) of the sentences.

Table III and Figure 6 show the results of the correlation at sentence-level. Table III presents the average correlation between the facial features and the signal separately estimated with prosodic and MFCC features, in terms of emotional categories. Figure 6 shows a graphical representation of these results. This figure was created following the same steps described in Section IV-A for Figure 4. Here, the correlation of each marker without normalization was used to assign the gray-scale color from Figure 5 to the Voronoi cells.

Table III and Figure 6 show high levels of correlation between the original and estimated facial features, which agree



(a) Neutral (b) Sad (c) Happy (d) Angry

(e) Neutral (f) Sad (g) Happy (h) Angry

Fig. 6. Correlation results for Sentence-Level mapping. (a-d) Prosodic features, (e-h) Vocal tract features. The figure shows high levels of correlation between the original and estimated facial features, especially when MFCCs are used to estimate the mapping. The figures also suggest that the link between acoustic and facial features is influenced by the emotional content in the utterance.

TABLE IV
STATISTICAL SIGNIFICANT OF INTER-EMOTION DIFFERENCES IN CORRELATION OF SENTENCE-LEVEL MAPPING (*Neu*=NEUTRAL, *Sad*=SADNESS, *Hap*= HAPPINESS, *Ang*= ANGER)

Facial Area	Neu-Sad	Neu-Hap	Neu-Ang	Sad-Hap	Sad-Ang	Hap-Ang
	Prosodic features					
Head Motion	0.444	0.332	0.061	0.005	0.000	1.000
Eyebrow	0.916	1.000	1.000	1.000	1.000	1.000
Lips	0.894	1.000	1.000	0.890	0.880	1.000
Upper region	51.72%	27.59%	10.34%	0.00%	55.17%	0.00%
Middle region	100.00%	35.90%	30.77%	17.95%	53.85%	2.56%
Lower region	89.66%	34.48%	3.45%	6.90%	37.93%	13.79%
	MFCC features					
Head Motion	0.012	0.098	1.000	1.000	0.221	0.743
Eyebrow	0.000	0.000	0.000	0.000	0.000	1.000
Lips	1.000	0.180	1.000	0.603	1.000	1.000
Upper region	96.55%	100.00%	100.00%	51.72%	6.90%	6.90%
Middle region	38.46%	53.85%	56.41%	30.77%	23.08%	5.13%
Lower region	0.00%	13.79%	10.34%	6.90%	6.90%	0.00%

with the hypothesis that the production of facial gestures and speech are internally connected [2], [3], [4]. The results also show that the correlation levels are higher when MFCC features are used to compute the mapping parameters. This result is observed not only near the orofacial area, in which the appearance of the face is directly modified by the configuration of the vocal tract, but also in areas far from the mouth such as cheeks and forehead, which agrees with the observations made in [2]. For example, head motion features, which are thought to be less dependent on the phonetic content of the speech, have higher correlation when the mapping is based on MFCCs rather than prosodic features. Notice that MFCCs carry the spectral envelope information and are closely related with the articulatory processes, while prosodic features are predominately related with the source of the speech. Therefore, these results indicate that articulatory events co-occur with the production of facial gestures.

When prosodic features are used to estimate the facial features, the observed correlation levels are similar along the facial regions, as shown in Figure 6 and Table III. However, when MFCCs are used to estimate the mapping parameters, the correlation of the lower face region is significantly higher than in any other facial region (see Table III). This result is also observed in lip features, which presents a correlation level higher than for eyebrow features. One explanation is that facial features with relatively low motion activity tend to have smaller correlation levels, as suggested in [12]. As mentioned in Section IV-A, the lower face region has higher motion activity than the upper and middle face region, and consequently, it presents higher correlation.

Up to this point, all the results presented in this subsection are observed across emotional categories. Another interesting question is whether those correlation levels between acoustic and facial features are emotion-type dependent. To answer this question, Table IV provides statistical significance measures of inter-emotion differences between the results presented in Table III. Similar to Table II, Table IV shows the statistical results for multiple comparison tests across emotional categories, using a 95% confidence interval. For head, eyebrow and lip motion, the p-values are given. For the aggregated facial regions, the percentage of the markers with significant differences at $\alpha = 0.05$ within each facial region is given.

Table IV shows that there are significant inter-emotional differences in the correlation levels between the original and estimated facial features presented in Table III. This result suggests that the link between acoustic and facial features is influenced by the emotional content in the utterance. As discussed in Section IV-A, the facial activeness changes under emotional speech. Likewise, acoustic features such as pitch, energy and spectral envelope vary as function of emotions [8],

[9]. These observations suggest that the audio-visual coupling presented here is also affected by this jointly emotional modulation.

When MFCC features are used to estimate the mapping parameters, Tables III and IV show that the upper face region presents significant inter-emotional difference in the correlation levels between the original and the estimated facial features. In this region, the results for neutral speech are statistically different from the results of any other emotional category (see Figure 6). However, in the middle and lower face region there are fewer markers with statistically significant inter-emotional differences. It is interesting to notice that facial areas that are less connected with the articulatory process, such as eyebrow and forehead (see Figure 4), present higher levels of inter-emotional differences. In those cases, the correlation levels for emotional utterances are higher than those with neutral speech.

When prosodic features are used to estimate the mapping parameters, the results show that the lower and especially the middle face region present stronger inter-emotional differences in the correlation levels, compared with the upper face region. The results also show that the correlation levels for emotional utterances are slightly higher than the ones for neutral speech.

In the case of lip features, the results indicate that there is no evidence to reject the null hypothesis that the correlation levels are similar across emotional categories. This result is observed when either MFCCs or prosodic features are used to estimate the mapping parameters. The fact that the emotional aspects do not significantly affect the coupling between speech and lip motion suggests that this link is mainly controlled by the articulation. This observation agrees with the results presented in Section IV-A, which show that the *relative* difference in the *displacement coefficient* for the lips between neutral and emotional speech is lower than in other facial areas. However, in other facial gestures the results indicate that emotional content do affect the relationship between facial gestures and speech.

C. Global-level Mapping

In the previous section, the mapping parameters between acoustic and facial features were computed at the sentence-level. This section extends those results when a single set of generic parameters (T_{emo}, M_{emo}) is computed across sentences (Equation 3). Since the level of correlation depends on the affective state of the speaker, as shown in Section IV-B, separate mapping parameters were calculated for each emotional category. Notice that speech is considerably easier and cheaper to collect, compared with any of the facial gestures considered here. Therefore, it is very convenient for applications such as facial animation to have reliable procedures to estimate facial features from speech.

Table V and Figure 7 show the average levels of correlation observed with global-level mapping. These results show that the correlation significantly decreases compared to the results with sentence-level mapping presented in the previous section (see Table III and Figure 6). This result is observed when either MFCCs or prosodic features are used to estimate the generic parameters T_{emo} and M_{emo}. Vatikiotis-Bateson and Yehia have also reported similar results [38].

TABLE V
SUMMARY CORRELATION FOR GLOBAL-LEVEL MAPPING (*N*=NEUTRAL, *S*=SADNESS, *H*= HAPPINESS, *A*= ANGER)

Facial Area	Prosodic				Mfcc			
	N	S	H	A	N	S	H	A
Head Motion	0.14	0.13	0.06	0.14	0.16	0.14	0.15	0.12
Eyebrow	0.18	0.05	0.01	-0.02	0.34	0.21	0.25	0.06
Lips	0.25	0.16	0.22	0.21	0.54	0.46	0.55	0.46
Upper region	0.15	0.07	0.10	0.02	0.28	0.15	0.32	0.13
Middle region	0.16	0.11	0.12	0.10	0.46	0.33	0.34	0.34
Lower region	0.21	0.18	0.20	0.17	0.58	0.49	0.51	0.51

TABLE VI
STATISTICAL SIGNIFICANT OF INTER-EMOTION DIFFERENCES IN CORRELATION OF GLOBAL-LEVEL MAPPING (*Neu*=NEUTRAL, *Sad*=SADNESS, *Hap*= HAPPINESS, *Ang*= ANGER)

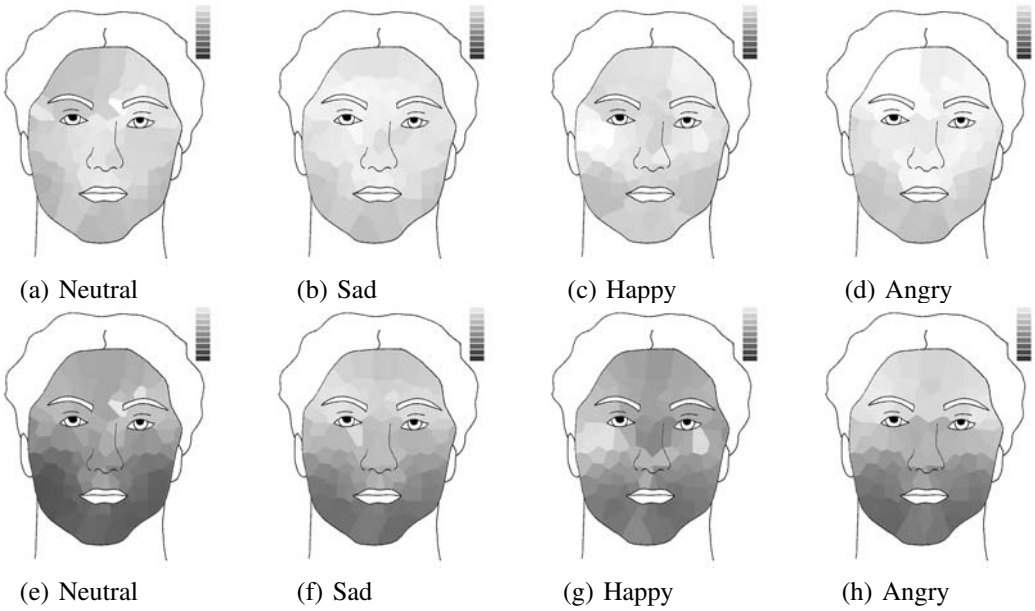
Facial Area	Neu-Sad	Neu-Hap	Neu-Ang	Sad-Hap	Sad-Ang	Hap-Ang
	Prosodic features					
Head Motion	1.000	0.030	1.000	0.078	1.000	0.071
Eyebrow	0.000	0.000	0.000	1.000	0.071	1.000
Lips	0.005	0.905	0.751	0.500	0.751	1.000
Upper region	58.62%	48.28%	86.21%	44.83%	20.69%	55.17%
Middle region	58.97%	43.59%	66.67%	33.33%	12.82%	41.03%
Lower region	37.93%	48.28%	44.83%	37.93%	0.00%	27.59%
	MFCC features					
Head Motion	1.000	1.000	0.543	1.000	1.000	1.000
Eyebrow	0.000	0.002	0.000	0.987	0.000	0.000
Lips	0.001	1.000	0.003	0.001	1.000	0.003
Upper region	96.55%	34.48%	96.55%	100.00%	13.79%	100.00%
Middle region	94.87%	84.62%	89.74%	56.41%	35.90%	56.41%
Lowe region	96.55%	75.86%	75.86%	34.48%	17.24%	48.28%

Similar to the sentence-level mapping section, Table V shows that the MFCC-based estimated facial features present higher levels of correlation than the prosody-based estimated facial features. In fact, in some cases of the latter, the correlation levels are close to zero (e.g. eyebrow with emotional utterances). This is probably because the link between some facial gestures and prosodic features, especially for emotional utterances, varies from sentence to sentence and it is not preserved after estimating the mapping parameters over the entire dataset.

The lower face region presents the highest correlation when either of the acoustic features is used to estimate the facial features. As shown in Table V and Figure 7, the levels of correlation decrease in the upper and middle face regions. These differences are also observed between lip and eyebrow features, in which the correlation levels in lip features are higher than in eyebrow features, across emotional category.

Similar to Table IV, Table VI presents details of statistical significance of inter-emotion differences in the correlation levels presented in Table V. The results indicate that there are strong emotional dependencies in those results.

In general, it is interesting to notice that the correlation levels for neutral speech are higher than any other emotional category. This result is in opposition with the results observed with sentence-level mapping, in which the correlation levels for neutral speech are equal or lower than other emotional categories. This result suggests that the coupling between facial gestures and emotional speech has a more complex



(a) Neutral (b) Sad (c) Happy (d) Angry

(e) Neutral (f) Sad (g) Happy (h) Angry

Fig. 7. Correlation results for Global-level mapping. (a-d) Prosodic features, (e-h) Vocal tract features. The figures show that the correlation levels significantly decreases compared with the results at sentence-level. MFCC-based estimated facial features also present higher correlation levels than when prosodic features are used. The lower face region presents the highest correlation levels in the face.

structure than for neutral speech, which is not preserved when a single set of parameters is used to estimate the linear mapping.

The fact that the levels of correlation decrease when global-level mapping is used indicates that the coupling between facial gestures and speech may change from sentence to sentence, depending on the underlying linguistic structure, as noted by Vatikiotis-Bateson and Yehia [38]. Since the parameters are averaged across sentences, the articulatory effects on the facial gestures are no longer reflected on the mapping. The high levels of correlation that were observed when MFCC features were used to estimate the parameter at sentence-level support this hypothesis. MFCC features model the configuration of the vocal tract, which shapes the appearance of the face [12]. Therefore, when the linguistic structure is blurred, the correlation levels significantly decrease. This hypothesis is addressed further in Section IV-D.

D. Analysis of Mapping Parameters

In Equation 3, the linear coupling between facial and acoustic features is mainly expressed through the parameter T, which is computed based on the cross-correlation between facial and acoustic features, K_{FS} (Equation 4). The structure of this parameter provides further insights about the relation between facial gestures and speech. This section presents a detailed analysis of the parameter T, when sentence-level mapping is used to estimate Equation 3.

The approach used to study the structure of T is based on PCA. In this technique, a reduced orthonormal subspace is selected such that it spans most of the variance of the multidimensional data. This subspace is formed from a subset of eigenvectors and eigenvalues of the covariance matrix of the data, associated with the highest eigenvalues. If the data is relatively clustered, only few eigenvectors will be needed to approximate the data. Therefore, by studying the eigenvectors of the covariance matrix of the parameter T, useful inferences can be drawn about the complexity of the assumed linear mapping between facial and acoustic features.

The parameter T is a $n \times m$ matrix whose dimensions depend on the acoustic and facial features used to compute the mapping (Equation 4). This matrix T is reshaped into a $nm \times 1$ vector, $\vec{t}$. The covariance matrix of this vector, $K_{\vec{t}}$, is approximated in Equation 9, in which $\vec{t}_u$ is the mean-removed vector associated with sentence u.

$$K_{\vec{t}} = [\vec{t}_1 \dots \vec{t}_N] \cdot [\vec{t}_1 \dots \vec{t}_N]^T \tag{9}$$

After computing the eigenvalues (λj) and eigenvectors (e_j) of $K_{\vec{t}}$, each parameter $\vec{t}_u$ can be expressed without errors as a linear combination of the eigenvectors (Equation 10, $P = n \cdot m$). Assuming that the eigenvalues are sorted in descending order, $\vec{t}_u$ can be approximated with only P eigenvectors ($P < n \cdot m$),

$$\vec{\hat{t}}_u(P) = \bar{t} + \sum_{j=1}^{P} < \vec{t}_u, \vec{e}_j > \vec{e}_j \tag{10}$$

where $\bar{t}$ is the vector mean of $\vec{t}$.

Table VII gives the fraction of eigenvectors (P) needed to span 90% or more of the variance of the parameters $\vec{t}$, for head, eyebrow and lip features. The same procedure was also applied for each marker. The average results within each facial area is presented in Table VII. The results are presented for each emotional category, in which only the parameters $\vec{t}_u$ of the corresponding emotional sentences were used to estimate $K_{\vec{t}}$ (Equation 9). In addition, the parameters $\vec{t}_u$ were concatenated together across emotional categories to estimate an emotion-independent covariance matrix. This procedure provides information for inferring whether the structure of

TABLE VII
FRACTION OF EIGENVECTORS USED TO SPAN 90% OR MORE OF THE VARIANCE OF THE PARAMETER T (N=NEUTRAL, S=SADNESS, H= HAPPINESS, A= ANGER, G= GLOBAL)

Facial	Prosodic					Mfcc				
Area	N	S	H	A	G	N	S	H	A	G
Head	0.22	0.33	0.28	0.28	0.28	0.36	0.47	0.44	0.53	0.56
Eyebrow	0.33	0.33	0.25	0.25	0.33	0.50	0.46	0.38	0.38	0.46
Lips	0.33	0.42	0.33	0.33	0.42	0.46	0.54	0.50	0.50	0.54
Upper	0.26	0.33	0.27	0.30	0.31	0.44	0.45	0.45	0.46	0.51
Middle	0.30	0.29	0.29	0.30	0.32	0.43	0.40	0.42	0.39	0.48
Lower	0.28	0.29	0.26	0.27	0.30	0.41	0.43	0.36	0.39	0.44

T depends on the emotional content of the utterance. These results are presented in Table VII under the letter G (global). Notice that the dimension of $\vec{t}$ varies when different facial and acoustic features are used to estimate the mapping. Therefore, the fraction of eigenvectors to cover 90% of the variance is easier to compare across facial features than just the dimension of this reduced subspace P.

Table VII indicates that when prosodic features are used to estimate the mapping parameters, between 22% and 42% of the eigenvectors of $K_{\vec{t}}$ are needed to cover 90% or more of the variance of T. When the mapping is based on MFCC features, between 36% and 54% of the eigenvectors are required to span this reduced subspace. These results suggest that the parameters $\vec{t}$ are clustered in a reduced subspace, showing a defined structure. Since the reduced subspaces for MFCC-based parameters have bigger dimensions than the ones for prosodic features, it can be infered that their structures are more difficult to model. Therefore, the mapping between prosodic features and facial gestures may be easier to generalize across sentences than the mapping between MFCCs and facial features.

As can be observed in Table VII, the percentage of eigenvectors of the emotion-independent covariance matrix needed to span 90% or more of the variance of $\vec{t}$ is generally higher than the percentage of eigenvectors of the emotion-dependent covariance matrices needed to cover the same reduced subspace. These results provide further evidence that the structures of the mapping parameters depend on the emotional content of the sentences.

The results presented in Table VII do not directly provide information about the correlation levels that will be observed when a reduced set of eigenvectors of $K_{\vec{t}}$ is used to approximate T. To analyze this question, the sentence-level mapping framework presented in section III-C was implemented to measure the correlation levels between facial and acoustic features when different numbers of eigenvectors (P) are used to approximate T (Equation 10). Figure 8 presents the results for the facial gestures considered in this paper. For the upper, middle and lower face regions the average results of the markers within each area is presented. The slopes of these curves indicate that the correlation levels slowly decreased as the number of eigenvectors used in the approximation of T decrease. Figure 8 also shows that when MFCC features are used to estimate the facial features, the slopes of the curves tend to be higher than the ones for prosodic features. These results support the hypothesis that there is a well-defined

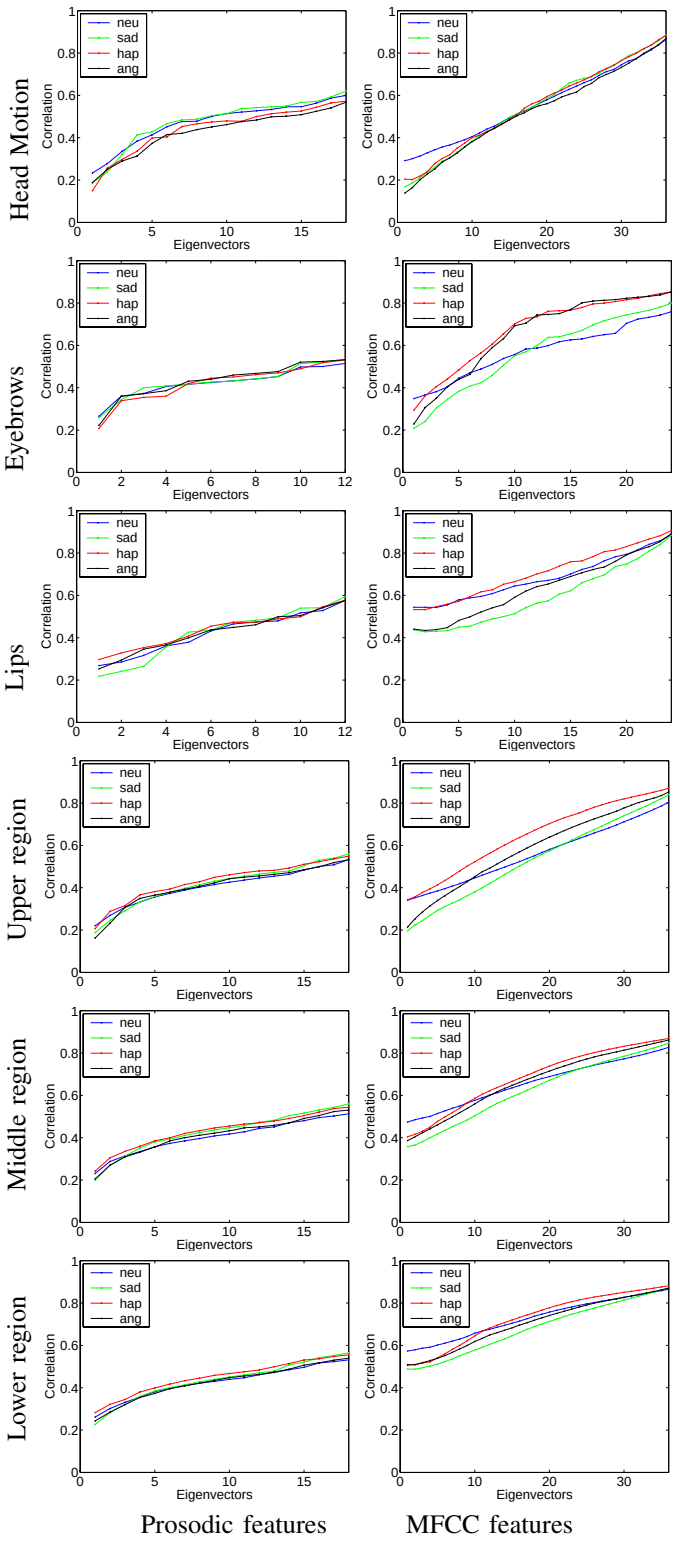


Fig. 8. Correlation levels as a function of the number of Eigenvectors (P) used to approximate $\vec{t}$ (Equation 10). The slopes of these curves indicate that the correlation levels slowly decreased as P decrease, supporting the hypothesis of an emotion-dependent structure in the audio-visual mappings.

structure in the parameter T, and that this structure seems to be simpler when prosodic features are used to estimate the facial gestures.

When MFCCs are used to estimate the mapping, the cor-

relation level for head motion is more affected than other facial gestures when low values of P are used in Equation 10 (see Figure 8). This result indicates that the coupling between head motion and speech varies from sentence to sentence, as suggested in [13]. The same result is observed for eyebrow motion. Conversely, the correlation level for lip motion is higher than 0.4, even when only one eigenvector is used to approximate T. This result indicates that not only there is a strong link between what is said and the appearance of the orofacial area, but also this map does not significantly change across sentences. This is not surprising given the tight coupling between lips and the segmental speech properties, so the articulatory effects dominate for all conditions. These observations suggest that the relation between speech and some facial gestures are easier to generalize than others.

Notice that the projections of $\vec{t}_u$ into the eigenvectors in Equation 10 is unknown when the facial features are not available. Therefore, if this approach is implemented for predicting facial gestures from speech, the terms $< \vec{t}_u, \vec{e}_j >$ will have to be estimated, which may be a non-trivial problem.

In sum, the results presented in this section suggest that even though the coupling between facial gestures and speech varies from sentence to sentence, there is an emotion-dependent structure in the mapping parameters across sentences that may be learned using more sophisticated non-linear estimation techniques such as HMMs or DBNs (e.g. see [7], [18], [30]).

V. Results of Phoneme Level Analysis

In the previous sections, the correlation levels were estimated for each utterance using either sentence-level, or global-level mapping parameters. This information provides general measures of the coupling between facial gestures and speech. This section analyzes whether this interrelations vary as function of broad phonetic categories. Instead of computing the Pearson's correlation over entire sentences, the speech is segmented into the constituent phones, over which the correlation levels are estimated. Notice that this approach differs from the framework followed in references [24], [25]. In those works, the authors estimated the mapping parameters for short time intervals corresponding to the target phonemes or syllables, which were spoken as nonsense words (e.g. /ma/, /cu/). Instead, we are interested in analyzing the phonetic dependency on the correlation levels when the mappings are estimated for the entire sentences. This procedure will indicate whether certain phonemes require specific attention for accurate facial feature estimation.

The broad phonetic categories considered in this paper are voiced/unvoiced, nasal, fricative and plosive phonemes. We also distinguish between vowels and consonant sounds. These phonetic categories share source characteristics and vocal-tract configurations that may influence the relation between gestures and speech. In addition, we also consider the correlation levels during periods of acoustic silence, which are separately analyzed. Note that silence periods can be linguistically meaningful (such as in phrase or sentence boundaries) and can be accompanied by specific gestures.

For each sentence, a single set of parameters (T,M) was estimated using the sentence-level mapping framework presented in Section III-C. Forced alignment, implemented with the HTK toolkit [39], was used to align the transcript with the speech. First, a small window centered on the target phoneme was selected in the original and estimated facial feature stream. The window was expanded in both directions as a function of the phoneme duration, to compensate for possible error in the automatic forced alignment procedure and the phase differences between gestures and speech [4], [40]. Finally, Pearson's correlation between the selected segments was computed. This procedure was repeated for each phoneme. The results were used to estimate average values for each broad phonetic category.

Table VIII gives the average correlation levels for the broad phonetic classes, for each emotional category, when prosodic and MFCC features were used to estimate the mapping between gestures and speech. To analyze whether the results presented in Table VIII are statistically significant, two-way ANOVA tests were implemented. The dependent factors are the emotional and phonetic categories. The results show that the p-values for the tests for each facial feature considered in this paper were less than 0.001, which indicate that there are significant differences in the emotional and phonetic levels. The p-values for the interaction terms were higher than 0.05, which suggest that the phonetic and emotional effects are additive.

To identify the phonetic classes that present statistical differences, a multiple comparison test with Bonferroni adjustment was applied to the data. A summary of the results is presented in Table IX. Similar to Tables II, IV and VI, Table IX gives the p-values for the head, eyebrow and lip motion test. It also gives the percentage of markers belonging to the upper, middle and lower face regions that present significant difference ($p < 0.05$).

In general, Table IX indicates that the average correlation levels for vowels is significantly higher than the levels for consonants. This result agrees with previous findings [25]. Similarly, voiced regions present higher correlation than unvoiced regions. Fricative phonemes present lower correlation levels than the average phoneme. Also, in contrast to the results presented by Yehia *et al*, our results suggest that the correlation levels for nasal phonemes do not have statistical differences compared to the average phonemes [13]. The different acoustic features used to estimate the mapping may explain this result. They used LSP coefficients, which, unlike MFCCs, do not adequately model the zeros in the nasal spectra. Like nasal phonemes, plosive phonemes presents similar correlation levels as the average phoneme.

Table VIII reveals that the lower face region and the lip features present the highest correlation levels. When compared to the case when the correlation was computed over the entire sentence (Table III), these facial features show the lowest reduction in the correlation levels. These results indicate that lips features are locally related to the segmental acoustic events, and that they have a different time resolution than eyebrow and head motion features. This is not surprising given the key role of the lips in the articulatory process.

An interesting result in Table VIII is that the correlation levels during acoustic silence is less than 50% lower than any

TABLE IX
STATISTICAL SIGNIFICANT IN DIFFERENCES BETWEEN BROAD PHONETIC CLASSES (*Al*=ALL, *Vw*=VOWELS,*Co*=CONSONANT, *Vo*=VOICED, *Uv*=UNVOICED, *Na*=NASAL, *Pl*=PLOSIVE, *Fr*=FRICATIVE)

Facial Area	Prosodic features									MFCC features								
	Al-Vw	*Al-Co*	*Al-Vo*	*Al-Uv*	*Al-Na*	*Al-Pl*	*Al-Fr*	*Vw-Co*	*Vo-Uv*	*Al-Vw*	*Al-Co*	*Al-Vo*	*Al-Uv*	*Al-Na*	*Al-Pl*	*Al-Fr*	*Vw-Co*	*Vo-Uv*
Head Motion	0.16	0.01	1.00	0.00	1.00	0.00	0.01	0.00	0.00	1.00	1.00	1.00	0.22	1.00	1.00	0.73	0.72	0.02
Eyebrow	1.00	1.00	1.00	0.06	1.00	1.00	0.40	1.00	0.00	1.00	1.00	1.00	0.05	1.00	1.00	0.00	0.29	0.00
Lips	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.06	1.00	0.06	1.00	0.00
Upper	51.7%	24.1%	0.0%	93.1%	0.0%	51.7%	100%	93.1%	100%	0.0%	0.0%	0.0%	100%	0.0%	0.0%	100%	24.1%	100%
Middle	53.9%	0.0%	0.0%	56.4%	2.6%	20.5%	53.9%	84.6%	79.5%	0.0%	0.0%	0.0%	84.6%	2.6%	0.0%	61.5%	15.4%	94.9%
Lower	31.0%	3.5%	0.0%	24.1%	20.7%	51.7%	27.6%	72.4%	72.4%	0.0%	0.0%	0.0%	79.3%	24.1%	0.0%	44.8%	27.6%	96.6%

other broad phonetic category. The main reason of this result is the behavior of acoustic features during silence. The RMS energy decreases in the entire frequency spectrum, affecting the MFCC coefficients. Also, in this paper the pitch (F0) is interpolated during unvoiced and silence regions (see Section III-B), which clearly affects the accuracy of the mapping. Notice that acoustic silence does not mean that there is no communication activity. In fact, while the verbal channel is passive, the face may continue to gesture which explains why the correlation levels are not zero. This result suggests that silence regions should be treated in a special way.

In sum, the results presented in this section suggest that special consideration need to be taken for consonants, especially unvoiced and fricative sounds as well as silence segments. In applications such as facial animation, the use of interpolation or pre-learned visemes during those segments may give better perceptive quality than just the use of the estimated facial features. Furthermore, these results support the observations presented in [3] regarding the use of different resolutions to integrate facial features for realistic avatars.

VI. DISCUSSION

While description of multimodal synthesis or recognition algorithms were not the goals of this paper, the results presented here aim to provide insights on how to jointly model facial gestures and speech for expressive interfaces. The results of the present paper indicate that linguistic and emotional aspects need to be jointly considered in the models. This applies not only to the orofacial area (e.g. with the use of visemes), but also in facial region such as forehead and even head motion. This section discusses and suggests new directions in the areas of facial animation and emotion recognition. It also presents observations with theoretical value.

Although the correlation levels decrease when the mapping parameters are estimated at global level from the entire database, the results presented here indicate that the relationship between facial gestures and speech has a structure that may be automatically learned with a more sophisticated mapping. This structure varies from emotion to emotion, indicating that specific emotional models are needed for expressive facial animations. As discussed in Section IV-A, when prosodic features are used to estimate the mapping parameters, the correlation levels are lower than when vocal-tract features are used. However, the internal structure of the mapping seems to be easier to generalize. Therefore, prosodic features may be better to use than vocal tract features to estimate facial features given the interplay between speech articulation and facial actions. As an example of these observations, in our related work, we have shown that an HMM-based framework can successfully capture the temporal relationship between head motion gesture and acoustic prosody features [18]. In [7], the approach was extended to include emotion-dependent models. The results showed that the generated sequences were perceived as natural as the original sequences, indicating that the relationship between head motion and prosody was successfully modeled. These works suggest that similar time series frameworks may be used to learn the structure of the audio-visual mapping for other facial gestures such as eyebrow motion.

As discussed in Section IV-D, some facial gestures showed more complex structure than others. Also, errors in some facial gestures are more perceptively detrimental for facial animation than others (e.g. lip motion). Therefore, the correlation levels between acoustic features may not be sufficient to generate some facial gestures, as suggested by Barker *et al* [25]. In those cases, extra constraints, specifications or pre-learned rules should be generated.

The results in Section V reveal that facial gestures and speech are connected at different resolutions. While the orofacial area is locally tightly connected with the spoken message, the forehead, eyebrow and head motion are coupled only at the sentence level. Therefore, a coarse-to-fine decomposition of acoustic and facial features may be beneficial to model the coupling between different facial areas and speech. This is an important observation, because believable models need to properly include the right timing and coupling resolution between different communication channels. Currently, this is one area that we are actively working to expand the results presented in this paper. Likewise, the results indicate that the correlation levels in the audio-visual mapping during consonant, unvoiced and especially silence regions tend to decrease. Therefore, special consideration is required during those regions.

The results presented in this paper indicate that linguistic and emotional aspects of communication jointly modulate human speech and gestures to communicate and express desired messages. Importantly, many of the same physical channels are involved actively or passively during the production of both speech (verbal) and the various face gestures (non-verbal). Since articulatory and affective goals appear to co-occur

TABLE VIII
PEARSON'S CORRELATION FOR THE AUDIO-VISUAL MAPPING AT PHONEME LEVELS (*N*=NEUTRAL, *S*=SADNESS, *H*=HAPPINESS, *A*=ANGER)

	Prosodic				MFCC			
	N	S	H	A	N	S	H	A
	Average all phonemes							
Head Motion	0.41	0.35	0.34	0.36	0.51	0.44	0.46	0.47
Eyebrow	0.39	0.37	0.34	0.35	0.47	0.45	0.50	0.50
Lips	0.54	0.52	0.49	0.51	0.74	0.70	0.71	0.75
Upper region	0.39	0.36	0.35	0.35	0.51	0.47	0.51	0.49
Middle region	0.44	0.40	0.39	0.40	0.62	0.56	0.57	0.59
Lower region	0.51	0.48	0.45	0.48	0.72	0.68	0.66	0.70
	Vowels							
Head Motion	0.42	0.37	0.37	0.39	0.52	0.47	0.46	0.49
Eyebrow	0.38	0.39	0.35	0.37	0.47	0.46	0.51	0.53
Lips	0.53	0.51	0.49	0.55	0.74	0.72	0.73	0.75
Upper region	0.40	0.38	0.37	0.38	0.52	0.48	0.52	0.51
Middle region	0.46	0.43	0.41	0.42	0.63	0.57	0.59	0.61
Lower region	0.53	0.50	0.48	0.49	0.72	0.69	0.69	0.71
	Consonant							
Head Motion	0.39	0.33	0.31	0.33	0.50	0.43	0.45	0.46
Eyebrow	0.40	0.37	0.32	0.34	0.47	0.45	0.48	0.48
Lips	0.54	0.54	0.48	0.50	0.75	0.69	0.70	0.75
Upper region	0.38	0.34	0.34	0.32	0.50	0.47	0.50	0.48
Middle region	0.43	0.39	0.38	0.38	0.62	0.55	0.56	0.59
Lower region	0.50	0.46	0.44	0.47	0.71	0.67	0.65	0.69
	Voiced phonemes							
Head Motion	0.42	0.36	0.36	0.37	0.51	0.45	0.47	0.48
Eyebrow	0.39	0.38	0.36	0.36	0.47	0.46	0.51	0.50
Lips	0.54	0.52	0.49	0.53	0.75	0.72	0.72	0.76
Upper region	0.39	0.37	0.36	0.36	0.52	0.48	0.53	0.50
Middle region	0.44	0.41	0.40	0.41	0.63	0.57	0.58	0.60
Lower region	0.51	0.49	0.46	0.49	0.72	0.69	0.68	0.71
	Unvoiced phonemes							
Head Motion	0.38	0.31	0.30	0.32	0.50	0.41	0.44	0.44
Eyebrow	0.39	0.36	0.27	0.33	0.46	0.44	0.44	0.48
Lips	0.55	0.53	0.45	0.46	0.72	0.64	0.65	0.73
Upper region	0.37	0.33	0.32	0.31	0.49	0.45	0.45	0.46
Middle region	0.43	0.37	0.36	0.37	0.61	0.53	0.51	0.56
Lower region	0.50	0.44	0.42	0.45	0.71	0.65	0.60	0.67
	Nasal phonemes							
Head Motion	0.41	0.36	0.34	0.34	0.49	0.43	0.47	0.45
Eyebrow	0.38	0.38	0.34	0.31	0.46	0.46	0.51	0.49
Lips	0.52	0.58	0.46	0.53	0.77	0.79	0.73	0.79
Upper region	0.36	0.36	0.31	0.34	0.52	0.50	0.53	0.50
Middle region	0.40	0.42	0.35	0.37	0.64	0.57	0.58	0.60
Lower region	0.47	0.50	0.40	0.46	0.73	0.69	0.68	0.71
	Plosive phonemes							
Head Motion	0.35	0.27	0.31	0.30	0.46	0.45	0.45	0.45
Eyebrow	0.41	0.34	0.28	0.32	0.48	0.46	0.51	0.47
Lips	0.59	0.53	0.50	0.56	0.75	0.71	0.75	0.78
Upper region	0.38	0.32	0.31	0.29	0.51	0.46	0.50	0.47
Middle region	0.46	0.39	0.41	0.42	0.64	0.56	0.57	0.56
Lower region	0.54	0.49	0.47	0.55	0.74	0.67	0.67	0.69
	Fricative phonemes							
Head Motion	0.36	0.32	0.27	0.34	0.51	0.40	0.42	0.44
Eyebrow	0.38	0.35	0.25	0.32	0.44	0.43	0.41	0.45
Lips	0.55	0.52	0.44	0.47	0.73	0.63	0.65	0.74
Upper region	0.34	0.33	0.30	0.30	0.48	0.45	0.44	0.45
Middle region	0.43	0.37	0.32	0.36	0.61	0.53	0.49	0.58
Lower region	0.50	0.45	0.40	0.45	0.70	0.66	0.58	0.68
	Silence							
Head Motion	0.22	0.19	0.18	0.21	0.29	0.21	0.19	0.25
Eyebrow	0.19	0.15	0.10	0.15	0.28	0.19	0.19	0.20
Lips	0.17	0.22	0.08	0.16	0.35	0.33	0.24	0.42
Upper region	0.17	0.15	0.12	0.15	0.28	0.19	0.18	0.18
Middle region	0.13	0.14	0.09	0.14	0.28	0.22	0.19	0.25
Lower region	0.16	0.16	0.12	0.16	0.35	0.31	0.26	0.35

during normal human interaction and share the same channels, some form of internal central control needs to buffer, prioritize and execute them in coherent manner. We hypothesize that linguistic and affective goals interplay interchangeably as primary and secondary controls [41]. During speech, for instance, articulatory processes targeting linguistic goals might have priority over expressive goals, which as a consequence are restricted to modulating the communicative channels under the given articulatory constraints to convey the desired emotions. During the silence period, in contrast, articulatory processes are passive and affective goals are dominant in the non-verbal communication channels. In our previous work, emotional modulation in acoustic and articulatory parameters were analyzed [10], [22]. The results showed evidence for this interplay between affective and linguistic goals in speech, where low vowels with less restrictive tongue position observed greater emotional coloring than high vowels such as /i/. This hypothesis is also supported by the results presented in Section IV-B. The results indicate that when MFCCs are used to estimate the mappings, the eyebrows and the upper face region, which are less constrained by the linguistic goals, present higher inter-emotional differences. In fact, the eyebrow pose may be enough to perceive the emotional message [11].

Another interesting research area is multimodal emotion recognition. In addition to the poses of facial expressions, the results presented in Section IV-A suggest that the activeness of facial gestures also provides discriminative information about emotions. For example, the *average displacement coefficients* for lips and head motion features may be used to discriminate between pairs of emotion that are usually confused in other modalities, such as happy-anger and sadness-neutral. It is interesting to note that in previous works, head motion has been usually compensated and removed to analyze facial expressions. In those works, the subjects were instructed to avoid moving their head. Since one of the channels is intentionally blocked, the resulting data may contain non-natural patterns in other facial gestures. The results presented here suggest that head motion sequences not only should be encouraged for natural interaction, but also can be used to discriminate between emotions, as proposed in [42], [43].

Although the use of facial markers is suitable for the kind of analysis presented here, facial expressions need to be automatically extracted for realistic applications. This challenging task can be done with automatic platforms such as the *CMU/Pitt Automated Facial Image Analysis* (AFA) system [43], which has been tested with head motion sequences that include pitch and yaw as large as 40°and 70°, respectively. While the orofacial area is clearly the target area for audio-visual speech recognition, it is not clear which area need to be considered for multimodal emotion recognition. The analysis presented here, especially in Section IV-A, shed light into the important facial areas used to display emotions. Figure 4 and Table I show that the *displacement coefficient* in the middle and upper regions in the face for happiness and anger is approximately 70% higher than during neutral speech. Conversely, the facial activeness in the lower face region increases only 30% for the same emotions. These results suggest that the forehead area seems to have more degree of freedom to display non-verbal

information such as emotion [41]. Therefore, this area needs to be especially considered for emotion recognition systems. These results agree with perceptual experiments, which have shown that the upper face region is perceptively the most important region to detect visual prominence [44], [45].

VII. Conclusions

This paper analyzed the influence of emotional content and articulatory processes in the relationship between facial gestures and speech. The results show that articulation and emotions jointly modulate the interaction between these communicative modalities. The articulatory process is strongly correlated with the appearance of the face. This result is observed not only in the lower face region but also in the upper and middle face regions. Likewise, we observe significant inter-emotion differences in correlation levels of the audio-visual mapping, which suggest that the emotional content also affect the relationship. Under emotional speech the activeness of facial gestures for anger and happiness increases by more than 30% than during neutral speech. This pattern directly affects the interrelation between facial gestures and speech.

The results presented here suggest that even though the relationship between facial gestures and speech can change from sentence to sentence, there is an emotion-dependent structure that may be learned using more sophisticated techniques than the commonly used multilinear regression. We are currently studying how to jointly model facial gestures and speech. Results from our recent work [7], [18] using time series models such as HMMs, are promising and hence seem to be suitable for learning the kinds of audio-visual mapping analyzed here.

We are currently analyzing the timing and resolution in the interrelation between facial and acoustic features. As shown in this paper, facial gestures and speech are coupled at different resolutions. Also, the timing between gestures and speech is intrinsically asynchronous. Even in the orofacial area there may be a phase difference of hundreds of milliseconds, because of the co-articulation process and articulator inertia [40]. An open question is how to learn and model this asynchronous behavior not only near the lips, but also in the entire face. Our goal is to appropriately integrate the facial gestures to generate believable human-like facial animations.

The main limitation of this work is that we analyzed the gestures and speech of a single actress. We are collecting similar data from more subjects, which will be useful to validate and expand the experiments presented here. This data will also be appropriate to study inter-subject variabilities in facial gestures. Our next step will be to control the emotional content and the personal styles of the facial animations. This will be possible if we understand how the emotional, idiosyncratic and linguistic aspects of human communication modulate each communicative modality.

Acknowledgment

The authors wish to thank the anonymous reviewers for their thoughtful and insightful comments. Thank also go to the colleagues in the emotion research group for their valuable comments.

References

[1] D. McNeill, *Hand and Mind: What gestures reveal about thought*. Chicago, IL, USA: The University of Chicago Press, 1992.

[2] E. Vatikiotis-Bateson, K. Munhall, Y. Kasahara, F. Garcia, and H. Yehia, "Characterizing audiovisual information during speech," in *Fourth International Conference on Spoken Language Processing (ICSLP 96)*, vol. 3, Philadelphia, PA, USA, October 1996, pp. 1485–1488.

[3] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone, "Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversational agents," in *Computer Graphics (Proc. of ACM SIGGRAPH'94)*, Orlando, FL,USA, 1994, pp. 413–420.

[4] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. McCullough, and R. Bryll, "Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures," in *European Signal Processing Conference (EUSIPCO 02)*, Tolouse, France, September 2002, pp. 75–78.

[5] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, April 1993.

[6] P. Ekman and E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. New York, NY, USA: Oxford University Press, 1997.

[7] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1075–1086, March 2007.

[8] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, April 2003.

[9] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[10] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, 2004.

[11] P. Ekman, "About brows: emotional and conversational signals," in *Human ethology: claims and limits of a new discipline*, M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, Eds. New York, NY, USA: Cambridge University Press, 1979, pp. 169–202.

[12] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[13] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, no. 3, pp. 555–568, Jul 2002.

[14] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and F0 variations," in *International Conference on Spoken Language (ICSLP)*, vol. 4, Philadelphia, PA, USA, October 1996, pp. 2175–2178.

[15] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Sixth International Conference on Multimodal Interfaces ICMI 2004*. State College, PA: ACM Press, 2004, pp. 205–211.

[16] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan, "Embodiment in conversational interfaces: Rea," in *International Conference on Human Factors in Computing Systems (CHI-99)*, Pittsburgh, PA, USA, May 1999, pp. 520–527.

[17] W. Wahlster, "Towards symmetric multi-modality: Fusion and fission of speech, gesture, and facial expression," in *Proceedings of the 26th German Conference on Artificial Intelligence*, A. Günter, R. Kruse, and B. Neumann, Eds. Berlin, Germany: Springer-Verlag Press, 2003, pp. 1–18.

[18] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, "Natural head motion synthesis driven by acoustic prosodic features," *Computer Animation and Virtual Worlds*, vol. 16, no. 3-4, pp. 283–290, July 2005.

[19] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Proc. of IEEE International Conference on Automatic Faces and Gesture Recognition*, Washington, D.C., USA, May 2002, pp. 396–401.

[20] E. Bevacqua and C. Pelachaud, "Expressive audio-visual speech," *Computer Animation and Virtual Worlds*, vol. 15, no. 3-4, pp. 297–304, July 2004.

[21] R. Gutierrez-Osuna, P. Kakumanu, A. Esposito, O. Garcia, A. Bojorquez, J. Castillo, and I. Rudomin, "Speech-driven facial animation with realistic dynamics," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 33–42, February 2005.

[22] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 497–500.

[23] B. Granström and D. House, "Audiovisual representation of prosody in expressive speech communication," *Speech Communication*, vol. 46, no. 3-4, pp. 473–484, July 2005.

[24] J. Jiang, A. Alwan, P. Keating, B. Chaney, E. A. Jr., and L. Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1174–1188, 2002.

[25] J. P. Barker and F. Berthommier, "Estimation of speech acoustics from visual speech features: A comparison of linear and nonlinear models," in *Conference Audio-Visual Speech Processing (AVSP 1999)*, Santa Cruz, CA, USA, August 1999, pp. 112–117.

[26] Z. Zeng, J. Tu, M. Liu, T. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 424–428, February 2007.

[27] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, September 2003.

[28] C. Conati and H. Mclaren, "Data-driven refinement of a probabilistic model of user affect," in *Proceedings of the Tenth International Conference on User Modeling (UM2005)*, L. Ardissono, P. Brna, and A. Mitrovic, Eds. Berlin, Germany: Springer-Verlag Press, 2005, pp. 40–49.

[29] S. Kshirsagar and N. Magnenat-Thalmann, "A multilayer personality model," in *Proceedings of the 2nd international symposium on Smart graphics (SMARTGRAPH 2002)*. Hawthorne, NY, USA: ACM Press, June 2002, pp. 107–115.

[30] Y. Li and H. Shum, "Learning dynamic audio-visual mapping with Input-Output Hidden Markov Models," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 542–549, June 2006.

[31] M. Nordstrand, G. Svanfeldt, B. Granström, and D. House, "Measurements of articulatory variations and communicative signals in expressive speech," in *Audio Visual Speech Processing (AVSP 03)*, S. Jorioz, France, September 2003, pp. 233–237.

[32] E. M. Caldognetto, P. Cosi, C. Drioli, G. Tisato, and F. Cavicchio, "Coproduction of speech and emotions: Visual and acoustic modifications of some phonetic labial targets," in *Audio Visual Speech Processing (AVSP 03)*, S. Jorioz, France, September 2003, pp. 209–214.

[33] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*. Piscataway, NJ, USA: IEEE Press, 2000.

[34] R. W. Picard, "Affective computing," MIT Media Laboratory Perceptual Computing Section, Cambridge, MA,USA, Technical Report 321, November 1995.

[35] P. Boersma and D. Weeninck, "Praat, a system for doing phonetics by computer," Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, Technical Report 132, 1996, http://www.praat.org.

[36] J. Jiang, A. Alwan, L. Bernstein, E. A. Jr., and P. Keating, "Predicting face movements from speech acoustics using spectral dynamics," in *IEEE International Conference on Multimedia and Expo (ICME 2002)*, vol. 1, Lausanne, Switzerland, August 2002, pp. 181–184.

[37] K. Arun, T. Huang, and S. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, September 1987.

[38] E. Vatikiotis-Bateson and H. C. Yehia, "Speaking mode variability in multimodal speech production," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 894–899, July 2002.

[39] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, England, December 2006.

[40] T. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1082–1089, May 2006.

[41] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 549–556.

[42] R. E. Kaliouby and P. Robinson, "Mind reading machines: automated inference of cognitive mental states from video," in *IEEE Conference on Systems, Man, and Cybernetic*, vol. 1, The Hague, the Netherlands, October 2004, pp. 682–688.

[43] J. Cohn, L. Reed, Z. Ambadar, J. Xiao, and T. Moriyama, "Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior," in *IEEE Conference on Systems, Man, and Cybernetic*, vol. 1, The Hague, the Netherlands, October 2004, pp. 610–616.

[44] C. Lansing and G. McConkie, "Attention to facial regions in segmental and prosodic visual speech perception tasks," *Journal of Speech, Language, and Hearing Research*, vol. 42, pp. 526–539, June 1999.

[45] M. Swerts and E. Krahmer, "The importance of different facial areas for signalling visual prominence," in *International Conference on Spoken Language (ICSLP)*, Pittsburgh, PA, USA, September 2006.

Carlos Busso Carlos Busso received his B.S (2000) and M.S (2003) degrees with high honors in electrical engineering from University of Chile, Santiago, Chile. He is currently a Ph.D. candidate in electrical engineering at the University of Southern California (USC), Los Angeles, USA. Since 2003, he has been a research assistant in the Speech Analysis and Interpretation Laboratory (SAIL) at USC. He was selected by the School of Engineering of Chile as the best Electrical Engineer graduated in Chile in 2003. At USC, he received a Provost Doctoral Fellowship from 2003 to 2005 and a Fellowship in Digital Scholarship from 2007 to 2008. His research interests are in digital signal processing, speech and video processing, and multimodal interfaces. His current research includes modeling and understanding human communication and interaction, with applications to automated recognition and synthesis to enhance human-machine interfaces. He has worked on audio-visual emotion recognition, analysis of emotional modulation in gestures and speech, designing realistic human-like virtual characters, speech source detection using microphone arrays, speaker localization and identification in an intelligent environment, and sensing human interaction in multi-person meetings.

Shrikanth Narayanan Shrikanth Narayanan (Ph.D'95, UCLA) is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC) where he holds appointments in Electrical Engineering and jointly in Computer Science, Linguistics and Psychology. Prior to USC he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal member, of its Technical Staff from 1995-2000. At USC he is a member of the Signal and Image Processing Institute and a research area director of the Integrated Media Systems Center, an NSF Engineering Research Center. Shri Narayanan is an Editor for the Computer Speech and Language Journal (2007-present) and an Associate Editor for the IEEE Signal Processing Magazine. He was also an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000-04). He serves on the Speech Processing and Multimedia Signal Processing technical committees of the IEEE Signal Processing Society and the Speech Communication committee of the Acoustical Society of America. Shri Narayanan is a Fellow of the Acoustical Society of America, a Senior member of IEEE, and a member of Tau-Beta-Pi, Phi Kappa Phi and Eta-Kappa-Nu. He is a recipient of an NSF CAREER award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost fellowship from the USC Center for Interdisciplinary research, a Mellon Award for Excellence in Mentoring, and a recipient of a 2005 best paper award from the IEEE Signal Processing society. Papers with his students have won best student paper awards at ICSLP'02, ICASSP'05, and MMSP'06. He has published over 235 papers and has fourteen granted/pending U.S. patents.