FACTORIZING SPEAKER, LEXICAL AND EMOTIONAL VARIABILITIES OBSERVED IN FACIAL EXPRESSIONS

Soroosh Mariooryad and Carlos Busso

Multimodal Signal Processing (MSP)
Department of Electrical Engineering, The University of Texas at Dallas
soroosh.ooryad@utdallas.edu, busso@utdallas.edu

ABSTRACT

An effective human computer interaction system should be equipped with mechanisms to recognize and respond to the affective state of the user. However, spoken message conveys different communicative aspects such as the verbal content, emotional state and idiosyncrasy of the speaker. Each of these aspects introduces variability that will affect the performance of an emotion recognition system. If the models used to capture the expressive behaviors are constrained by the lexical content and speaker identity, it is expected that the observed uncertainty in the channel will decrease, improving the accuracy of the system. Motivated by these observations, this study aims to quantify and localize the speaker, lexical and emotional variabilities observed in the face during human interaction. A metric inspired in mutual information theory is proposed to quantify the dependency of facial features on these factors. This metric uses the trace of the covariance matrix of facial motion trajectories to measure the uncertainty. The experimental results confirm the strong influence of the lexical information in the lower part of the face. For this facial region, the results demonstrate the benefit of constraining the emotional model on the lexical content. The ultimate goal of this research is to utilize this information to constrain the emotional models on the underlying lexical units to improve the accuracy of emotion recognition systems.

Index Terms— Facial expressions, emotion recognition, factor analysis, face analysis.

1. INTRODUCTION

Perceiving and expressing emotions are fundamental characteristics in human interaction. The emotional content in a conversation can clarify ambiguities in the message, influencing the expected reactions from the listeners. Recent findings suggest that emotions, at a cognitive level, impact our decision makings [1]. For all these reasons, it is important for a *human machine interface* (HMI) to recognize the affective state of the user. Among different modalities, facial expression is one of the most important channels used to externalize emotions. Facial emotion recognition systems rely on visual cues to deduce the affective state of the speaker. However, human communication is a sophisticated phenomenon which involves the interplay of cognitive, physiological, cultural and conversational processes. Therefore, the emotional state of the speaker is not the only factor that affects the facial appearance. The face is also manipulated by the verbal content of the spoken message, and the intrinsic cultural, physiological and idiosyncratic characteristics of the speaker (Fig. 1). This paper aims to quantify and localize the speaker, lexical and emotional variabilities observed in facial movements.

Conventional facial expression recognition systems have attempted to remove speaker dependency, by employing normalization techniques [2]. However, the verbal content of the message is generally disregarded [3], which may lead to erroneous assessments. In fact, our previous work showed that conditioning the emotional facial models on the speech-related unit (phoneme) improves the emotion recognition rate [4]. In the area of facial animation, Cao et al. explored these ideas to synthesize talking heads [5]. They used *Independent Component Analysis* (ICA) to identify emotion and lexical related subspaces for facial gestures. These spaces were used to manipulate the emotional content of the synthesized facial animations. A key open problem is to identify facial areas that are more affected by speaker and lexical variabilities. If these dependencies can be localized, then lexicon and speaker dependent models can be built for features extracted from these facial areas. This is the precise, ultimate goal of this study.

The proposed approach builds upon our previous study, which analyzed the interplay between linguistic and affective goals in the face [6]. This previous work was based on controlled experiments, in which the same sentences were uttered by a single actor expressing different emotions. In this paper, we extend the approach by considering the speaker, lexical and emotional variabilities in the face during spontaneous human interactions. The proposed factor analysis consists in estimating a metric similar to mutual information. The proposed metric measures the amount of decrease in variability of the facial movement models when the models are constrained by each of the three factors, compared to the case when there is no constraint on the factors. In this work, the trace of the covariance matrix of motion trajectories is used as the measure of variability. The results reveal that the lexical variability is localized on the lower region of the face. The speaker variability (after feature normalization) is evenly distributed across the face. The emotional variability is localized in the middle and upper facial regions. Moreover, conditioning the emotional facial models on the lexical information significantly reduces the variability in the orofacial area. These results support the benefit of lexical dependent models for emotion recognition. These findings can be leveraged to build robust lexical dependent models for facial expression recognition, which is our long term ultimate goal.

2. METHODOLOGY

This study focuses on the lexical information at syllable and word levels. In contrast with phoneme, these longer units offer better choice to capture motion trajectories and also to handle co-articulation effects. The face is represented with 53 markers that were placed on the subjects' face (Sec. 2.1). In this work, the emotion factor is represented by categorical labels (e.g., happy and sad). This section describes the database, the normalization scheme, the factors influencing facial movements and the proposed factor analysis technique.

2.1. IEMOCAP Database

This study uses the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database [7]. This corpus was collected to study multimodal expressive human communica-

978-1-4673-2533-2/12/$26.00 ©2012 IEEE

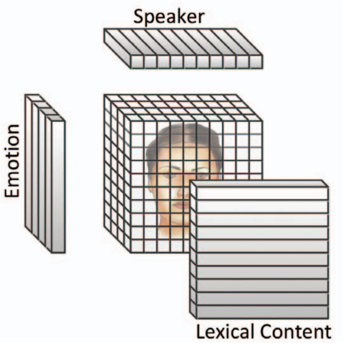
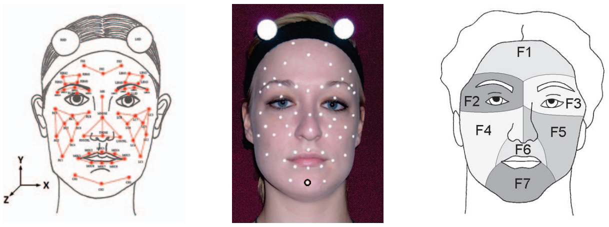


Fig. 1. Interplay between speaker, lexical and emotional content in the face, during expressive interaction.



(a) (b) (c)

Fig. 2. (a) Layout of 53 markers attached to the face of the subjects. (b) Subject with the markers. (c) Facial subdivision: (F1) Forehead, (F2) Left eye, (F3) Right eye, (F4) Left cheek, (F5) Right cheek, (F6) Nasolabial, and (F7) Chin.

tion. It contains five sessions of dyadic interactions. In each session, two actors (one male and one female) performed three carefully selected scripts. They were also asked to improvise hypothetical scenarios such as getting married. The scripts and scenarios were selected to elicit emotional manifestations. The recording includes speech and synchronized motion capture data, which provides detailed facial motion information. In total, 53 markers were placed on the face of the subjects. The markers' layout is depicted in Figures 2(a) and 2(b). For each marker, the motions in three directions (X, Y and Z) are extracted. The markers were translated and rotated so that the remaining motions correspond to the subjects' facial expressions [7]. The corpus was manually transcribed and segmented into turns. Forced alignment was used to estimate the phone and word boundaries. The syllable alignment was estimated using the syllabification software tsylb2 [8]. The emotional content of the utterances was assessed by subjective evaluations by three raters at turn level within the following categories: anger, sadness, happiness, disgust, fear, surprise, frustration, excited, neutral and other. The corpus is described in details by Busso et al. [7].

2.2. Speaker Normalization

The markers were manually attached to the subjects following the layout shown in Figure 2(a). Due to the differences in facial anatomy of the actors and the variation in the actual location of the markers, it is meaningless to directly compare the motion captured data across subjects. Therefore, the trajectory of the facial markers was normalized. Unfortunately, this normalization will affect the speaker variability measured by the proposed analysis.

The proposed speaker normalization method is motivated by the scheme developed by Zeng et al. [2]. The approach consists in matching the first and second order statistics of facial movements observed across all the speakers during neutral interaction (i.e., samples labeled as neutral). One speaker is selected as the reference (female speaker in the first session). The markers of other subjects are mapped into the markers' space of the reference subject. Equation 1 describes this speaker normalization technique. The i^{th} marker of the speaker s in the direction $d \in \{X, Y, Z\}$, ($m^{s}_{i,d}$), is transformed to match the reference speaker (ref), where μ and σ are the mean and standard deviation of the markers.

$$m^{\prime s}_{i,d} = (m^{s}_{i,d} - \mu^{s}_{i,d}) \times \frac{\sigma^{ref}_{i,d}}{\sigma^{s}_{i,d}} + \mu^{ref}_{i,d} \qquad (1)$$

Table 1. The 10 most frequent syllables and words in the IEMOCAP database. The values below the tokens are the corresponding number of repetitions in the corpus.

Syllables	AY	Y_UW	AX	N_OW	T_AX	AX_T	L_AY_K	DH_AX	G_OW	AX_N_D
	3232	2437	1748	1556	899	823	733	693	670	599
Words	I	YOU	KNOW	A	TO	THE	LIKE	AND	DO	ME
	3152	2402	955	952	739	690	566	541	523	503

2.3. Factors Influencing Facial Movements

Based on the size of the database, we limited the study by considering only the ten most frequent syllables and words across all speakers in the corpus. Therefore, we guarantee that the number of instances for each of these lexical tokens is large enough for the analysis. The selected lexical tokens with the number of repetitions are given in Table 1. The lower quartiles for durations of the syllables and words are 66 and 77 ms, respectively. The upper quartiles for both units are 150 ms. There are ten speakers in the database, which facilitates the analysis of the speaker factor for this study. Among the emotional categories in the IEMOCAP, the four most frequent classes are selected for studying the emotion factor (i.e., happy, angry, sad and neutral).

2.4. Facial Motion Trajectory Models

The goal of building facial models is to capture the movement pattern of each marker in each of the directions (X, Y and Z). The proposed approach consists in estimating the temporal shape of the markers' trajectory for syllables or words across many repetitions. To build the models, all the sequences are linearly aligned to have equal length N, which is empirically set to 25 frames (preliminary results showed that increasing N did not affect the reported results). This alignment is performed with an interpolation and resampling scheme. Then, for each marker m, the average pattern is modeled with the $N \times 1$ mean trajectory vector μ_m. The variation around this trajectory across time is captured with the $N \times N$ covariance matrix (Σ_m).

Figure 3 depicts an example of the models for the word *WELL*, with different emotions across all the speakers. The figure shows the trajectory of the marker on the middle of the chin (highlighted in Fig. 2(b)) in Y direction (i.e., up-down). The mean pattern μ_m is shown with solid line. The variance at each frame is shown with dashed lines. Figure 3 not only shows the existence of a lexical pattern, but also suggests characteristic emotional patterns that can be used for emotion recognition.

2.5. Measuring Uncertainty and Factor Analysis

If a variable depends on various factors, its overall variability reflects the influence of all the factors. If the variable is segregated according to one of these factors, it is expected that the within-class variability will be lower than the original overall variability. The difference can be considered as a measure of the dependency between the factor and the variable (the higher the variance reduction, the higher the dependency). If the feature is independent of the factor, then this segregation does not change the distribution and the variability of the features, leading to no variability reduction. The inner cube in Figure 1 represents the overall variability observed in the face. By building lexicon, speaker or emotion dependent models, we expect a reduction in the variability. For instance, breaking down the models with respect to the speaker factor will lead to ten speaker dependent models with lower average variability than the initial model (inner cube). This approach is used here to evaluate and quantify the dependencies of these three factors on localized facial regions during expressive speech.

The notion of uncertainty provides a suitable framework to measure variability. Mutual information, given in Equation 2, uses the Shannon's entropy $H(m)$ to estimate the uncertainty reduction achieved by the knowledge of factor F. Since the facial features are continuous variables, the entropy should be replaced by the differential entropy. However, differential entropy does not directly carry the uncertainty concept, because $i)$ it is not scale invariant, and $ii)$ it can get negative values. Also, in the case of multivariate Gaussian distributions, the entropy is reduced to the sum of the logarithms of eigenvalues of the covariance matrix plus a constant term, which is not robust and brings up computational issues when the covariance matrix has some small eigenvalues, as in this case.

$$I(m,F) = H(m) - \sum_{f \in F} P(f) H(m \mid f) \tag{2}$$

$$RM(m,F) = tr(\Sigma_m) - \sum_{f \in F} P(f) tr(\Sigma_m \mid f) \tag{3}$$

$$RM_n(m,F) = \frac{RM(m,F)}{tr(\Sigma_m)} \tag{4}$$

We propose a *relevance measure* (RM) to quantify the uncertainty reduction for continuous multivariate random variables (Equation 3). The uncertainty is measured following the approach used by Park et al. [9]. The authors used the trace of the covariance matrix (i.e., sum of the eigenvalues) as a rough measure of uncertainty. Notice that the eigenvalues of the covariance matrix give the variance in the principal components of the data. Therefore, the trace captures the overall variance. $RM(m,F)$ estimates the expected amount of decrease in uncertainty of marker m, when factor F (i.e., speaker, lexical or emotional content) is known. The probability distribution of each factor, $p(f)$, is estimated from their relative frequencies in the data (e.g., $p(Happy)$ for emotion). First, the uncertainty of the inner cube in Figure 1 is calculated $tr(\Sigma_m)$. Then, the model is broken down into separate models, which are conditioned on each factor F to calculate the conditional uncertainties (i.e., $tr(\Sigma_m|f)$). Then, the metric $RM(m,F)$ is estimated according to Equation 3 to quantify the relevance of factor F in the facial features m. The normalized version of the proposed metric is given in Equation 4 to compensate for different inherent unconditional uncertainties ($tr(\Sigma_m)$) of the facial points in a given direction. In Section 3, we use $RM_n(m,F)$ to factorize the three variabilities across different facial regions.

3. EXPERIMENT RESULTS

To compare the three factors, the face is divided into seven subdivisions which are depicted in Figure 2(c). Table 2 gives the average $RM_n(m,F)$ values for each factor. The results are aggregated for facial subdivisions (F1 to F7) and for marker directions (X, Y and Z). Two separate analyses are conducted for the selected data, by considering syllables and words as lexical units. For each region, the prominent factor with the highest $RM_n(m,F)$ value is highlighted. The results at word level are also shown in Figure 4 for visualization purpose. In this figure, darker regions represent higher $RM_n(m,F)$ values.

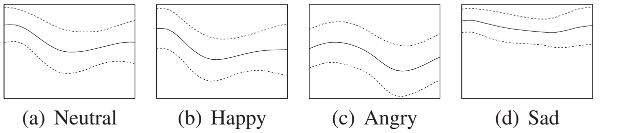

(a) Neutral (b) Happy (c) Angry (d) Sad

Fig. 3. Mean and variance of trajectories of the middle marker on the chin, in Y direction, for word *WELL*.

Table 2 and Figure 4 reveal that subdivision 7 (i.e., lower region in the face - see Fig. 2(c)) is strongly affected by the lexical information. We observe lower speaker and emotional variabilities in this region (both at the syllable and word levels). This result indicates that facial features extracted from this area should be constrained by the lexical information of the message. In other facial regions, the lexical variability is almost absent (especially for the upper facial area). Emotion is the dominant factor in other subdivisions (F1-F6). These results agree with our previous study [6], which confirms the presence of emotion modulation in the middle and upper facial regions. For these facial areas (F1-F6), lexicon-independent models should be enough to capture the emotional variability. Figure 4 also reveals the differences in the emotional variability across directions. In the X (left-right) and Y (up-down) directions, the emotional variability is mainly observed in the middle region of the face. In contrast, the emotional variability in the Z (in-out) direction is higher in the upper facial region. These results are useful to guide the design of facial emotion recognition systems and expressive talking head. The results show that the speaker variability is almost uniformly distributed and it is not the prominent factor in any of the subdivisions. This finding confirms the effectiveness of our speaker normalization.

4. LEXICON DEPENDENT EMOTIONAL MODELS

An important research goal of this paper is to assess the benefits of using lexicon-dependent models for emotion recognition. The proposed analysis in this section aims to estimate the decrease in uncertainty by conditioning on emotional factor, when the underlying lexical token is known. For this purpose, we compare the emotional variability across all lexical tokens (i.e., lexicon-independent condition) with the emotional variability when the models are also constrained by the lexical units (i.e., lexicon-dependent condition). The lexicon-independent condition corresponds to the experiments in Section 3 (*Emotion* columns in Table 2). For the lexicon-dependent condition, the data is segmented according to the lexical units and the emotions. A facial motion model is built for each combination (i.e., angry-WELL).

Important information can be derived by comparing the variation in the $RM_n(m,F)$ values for emotion factor between lexicon-dependent and lexicon-independent facial motion models. Table 3 gives the average $RM_n(m,F)$ values for emotion factor across all dimensions for the lexical dependent experiment. The columns $\Delta(\%)$ give the percentage of increase in $RM_n(m,F)$ compared to the corresponding values in the *Emotion* columns of Table 2. A positive value indicates the dependency of emotional variability on the verbal content. The table shows that constraining the emotional

Table 2. Average $RM_n(m, F)$ values for speaker, lexical and emotion factors in different regions of the face (see Fig. 2(c)).

Div#	Dir	Syllable Level			Word Level		
		Speaker	Syllable	Emotion	Speaker	Word	Emotion
F1	X	0.064	0.003	0.019	0.068	0.001	0.020
	Y	0.046	0.018	0.004	0.045	0.017	0.004
	Z	0.094	0.020	0.184	0.96	0.030	0.189
	Avg	0.068	0.014	**0.069**	0.070	0.016	**0.071**
F2	X	0.030	0.003	0.052	0.032	0.003	0.057
	Y	0.074	0.018	0.014	0.077	0.016	0.019
	Z	0.055	0.020	0.092	0.058	0.026	0.094
	Avg	**0.053**	0.014	**0.053**	0.056	0.015	**0.057**
F3	X	0.035	0.002	0.059	0.037	0.001	0.060
	Y	0.065	0.017	0.018	0.061	0.014	0.017
	Z	0.000	0.021	0.112	-0.008	0.029	0.114
	Avg	0.033	0.013	**0.063**	0.035	0.015	**0.064**
F4	X	0.077	0.026	0.099	0.085	0.026	0.098
	Y	0.081	0.016	0.163	0.077	0.023	0.153
	Z	0.066	0.050	0.060	0.070	0.066	0.055
	Avg	0.075	0.031	**0.107**	0.077	0.038	**0.102**
F5	X	0.067	0.022	0.125	0.067	0.022	0.124
	Y	0.085	0.020	0.164	0.084	0.026	0.155
	Z	0.088	0.054	0.051	0.091	0.071	0.047
	Avg	0.080	0.032	**0.113**	0.081	0.040	**0.109**
F6	X	0.068	0.033	0.123	0.71	0.036	0.122
	Y	0.089	0.018	0.139	0.087	0.023	0.133
	Z	0.029	0.168	0.088	0.031	0.208	0.085
	Avg	0.062	0.073	**0.117**	0.063	0.089	**0.114**
F7	X	0.052	0.045	0.075	0.053	0.053	0.073
	Y	0.029	0.188	0.055	0.031	0.228	0.045
	Z	0.033	0.226	0.013	0.035	0.272	0.012
	Avg	0.038	**0.153**	0.048	0.040	**0.184**	0.043

Table 3. The average $RM_n(m, F)$ values for emotion factor, in lexicon-dependent experiment. The $\Delta(\%)$ columns give the amount of increment in $RM_n(m, F)$ compared to the average values in the *Emotion* columns of Table 2.

Div#	Syllable Level		Word Level	
	Emotion(Syl Dep)	$\Delta(\%)$	Emotion(Syl Dep)	$\Delta(\%)$
F1	0.070	1.44	0.069	-2.28
F2	0.053	0.00	0.053	-7.01
F3	0.068	7.93	0.063	-1.58
F4	0.115	7.47	0.103	0.98
F5	0.122	7.56	0.111	1.83
F6	0.123	5.12	0.115	0.87
F7	0.067	**39.58**	0.063	**46.51**

models on the lexical information reduces the uncertainty in the orofacial area (F7). However, it has no significant effect on other facial subdivisions. These results confirm that only the orofacial area should be constrained by the lexical information of the message to build emotional models.

5. CONCLUSIONS

The paper introduced a metric to perform factor analysis on facial movements. The analysis shows that the lower facial region is mainly affected by the lexical information. The results show that conditioning the emotional models on the lexical units reduces the uncertainty only in the orofacial area. These results demonstrate the benefits of using lexicon-dependent models for features extracted from this facial area. While speaker variability is observed across the face, the results show that the emotional modulation is concentrated in the middle and upper facial regions.

Our next step is to build a facial emotion recognition system that leverages the insights provided by this study. We will combine lexicon-dependent models for the lower facial region with lexicon-independent models for the rest of the face. Another research direction is to define suitable lexical units. It is not practical to build separate model for each syllable, since there are more than 2000 syllables in English. We are exploring the use of clustering to find syllables with similar trajectories (i.e., visimes for syllables). These are some of the directions that we are working toward designing a robust emotion recognition system.

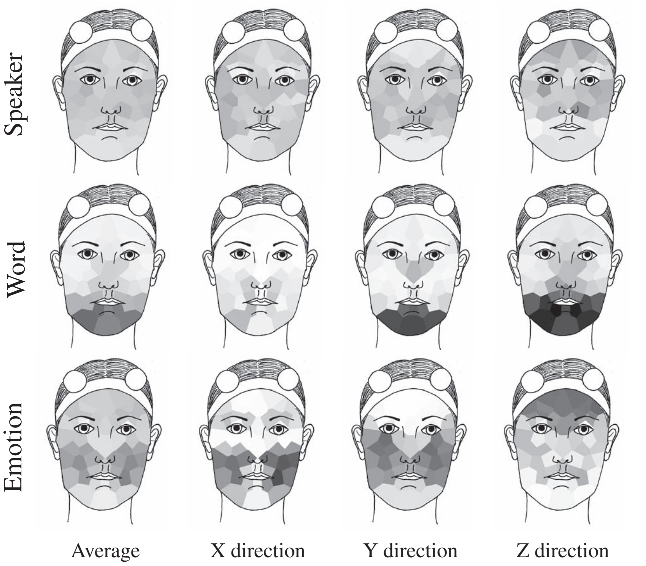


Fig. 4. $RM_n(m, F)$ for each factor (speaker, word and emotion) in the face. Darker regions represent higher dependency.

6. ACKNOWLEDGMENT

This work was supported by Samsung Telecommunications America, LLC.

7. REFERENCES

[1] R. Hastie and R. M. Dawes, *Rational choice in an uncertain world*, Thousand Oaks: Sage Publications, Jun 2001.

[2] Z. Zeng, J. Tu, B.M. Pianfetti, and T.S. Huang, "Audiovisual affective expression recognition through multistream fused HMM," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 570–577, June 2008.

[3] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, pp. 160–187, July 2003.

[4] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 2474–2477.

[5] Y. Cao, P. Faloutsos, and F. Pighin, "Unsupervised learning for speech motion editing," in *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, Aire-la-Ville, Switzerland, Switzerland, 2003, SCA '03, pp. 225–231, Eurographics Association.

[6] C. Busso and S.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.

[7] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[8] W. M. Fisher, "tsylb2-1.1 syllabification software," 1996.

[9] S. Park, H.L. Choi, N. Roy, and J.P. How, "Learning the covariance dynamics of a large-scale environment for informative path planning of unmanned aerial vehicle sensors," *International Journal of Aeronautical and Space Sciences*, vol. 11, no. 4, pp. 327–337, December 2010.