Feature and Model Level Compensation of Lexical Content for Facial Emotion Recognition

Soroosh Mariooryad and Carlos Busso

Abstract— **Along with emotions, modulation of the lexical content is an integral aspect of spontaneously produced facial expressions. Hence, the verbal content introduces an undesired variability for solving the facial emotion recognition problem, especially in continuous frame-by-frame analysis during spontaneous human interactions. This study proposes feature and model level compensation approaches to address this problem. The feature level compensation scheme builds upon a trajectory-based modeling of facial features and the whitening transformation of the trajectories. The approach aims to normalize the lexicon-dependent patterns observed in the trajectories. The model level compensation approach builds viseme-dependent emotional classifiers to incorporate the lexical variability. The emotion recognition experiments on the IEMOCAP corpus validate the effectiveness of the proposed techniques both at the viseme and utterance levels. The accuracies of viseme level and utterance level emotion recognitions increase by 2.73% (5.9% relative) and 5.82% (11% relative), respectively, over a lexicon-independent baseline. These performances represent statistically significant improvements.**

I. Introduction

The face is an expressive and rich channel to communicate feelings, intentions and desires. Emotions influence the externalized facial expressions [1]. Likewise, the articulation process modulates the facial appearance [2]. Therefore, facial expressions simultaneously convey both linguistic and affective messages [3], [4]. As a result, the underlying linguistic content imposes an undesired variability that affects the performance of facial emotion recognition systems (e.g., confusing a phoneme /ey/ with a smile). Decoding the emotional content requires the understanding and modeling of the interaction between the linguistic and affective goals. This work explores methods to compensate for the variability introduced by the articulation process (i.e., lexical content). These compensation approaches represent novel advances toward developing robust facial emotion recognition systems.

Given the strong influence of the lexical variability in the orofacial area, the facial features extracted from this area are usually ignored in video-based (i.e., dynamic) emotion recognition systems [5]. This approach is supported by our previous study that showed a strong effect of the lexical content on the orofacial area [6]. The lexical influence is considerably reduced in the middle and upper face areas. However, the lower face provides important information that an emotion recognition system should exploit. The challenge is to design schemes able to compensate or model the lexical variability observed in the orofacial area, especially in continuous frame-by-frame analysis of spontaneous interactions.

This study proposes strategies to enhance facial emotion recognition systems by modeling the underlying lexical content. The lexical content is represented in terms of visemes, which are the visual counterparts of the phonemes in facial expressions. The proposed strategies are implemented at the feature and model levels. The feature level approach consists in normalizing the facial features with respect to the given visemes. The normalization reduces the lexical variability increasing the performance of the emotion recognition system. The proposed normalization method builds upon trajectory-based modeling of the facial features. It consists in the whitening transformation of the feature trajectories. The model level approach is implemented by constraining the emotional models on the underlying visemes. The viseme level emotion recognition experiments give a significant improvement by utilizing the proposed methods, over a lexicon-independent baseline system. Then, several fusion approaches are explored to combine the viseme level classifiers to estimate the emotions at the utterance level. The utterance level emotion recognition results also show measurable improvements in performance, when the proposed compensation methods are incorporated.

The rest of the paper is organized as follows. Section II discusses related studies and our preliminary findings on factors influencing facial expressions. Section III describes the database and the required preprocessing steps. Section IV describes the proposed lexical variability compensation methods at the feature and model levels. Section V presents the experimental emotion recognition results that validate the proposed compensation methods. Section VI discusses the results, and summarizes potential extensions of this study.

II. Background

A. Related Works

While impressive advances have being been made in recognizing emotions from images with prototypical expressions, an important open challenge that remains open is the modeling of the temporal dynamics with respect to the lexical content of facial expression in spontaneous human interaction [7]. Ambadar et al. [8] used subjective evaluations to demonstrate the importance of perceiving the temporal dynamics to discriminate between subtle facial expressions as compared with the case in which the subject only evaluated static presentations. Pantic [9] highlighted the role of facial expression dynamics in different emotional and

This work was partially funded by Samsung Telecommunications America, LLC and NSF (IIS 1217183).

The authors are with the Multimodal Signal Processing (MSP) Laboratory, Electrical Engineering Department, The University of Texas Dallas, Richardson, TX 750 USA soroosh.ooryad@utdallas.edu, busso@utdallas.edu

cognitive states such as embarrassment, amusement, pain and mood. Although many approaches have been proposed to capture and analyze dynamic facial behaviors [10]–[12], the interaction between the underlying spoken and expressive messages has not being explored. Our previous study suggests a strong interplay between lexical and affective goals that influences the facial appearance in a localized manner [3]. The underlying emotion of an utterance manipulates the acoustic and prosodic parameters of the produced speech signal [13], [14]. Due to the interdependency of acoustic signal and articulatory configurations [2], it is expected to observe similar emotional variations in the orofacial area, as well as in the rest of the face [1]. Our previous studies also demonstrate the interplay between the articulatory process and expressive message in facial expressions. We proposed a controlled experiment, in which we contrasted neutral and emotional utterances with similar lexical content [3]. After aligning both sentences, we compared the shape of facial features using correlation and Euclidean distance. The evaluations showed that emotions have stronger influence on the upper face region. The comparisons also reveal that the articulatory process dominates the area around the mouth and lips. However, this facial region is still modulated to signal emotional cues that are easily perceived by other individuals (i.e., smile for happiness, jaw drop for surprise). While the upper and middle face areas are less dependent on the articulatory process, we have shown moderate and strong correlations between features from these facial areas and the lexical content [4].

B. Preliminarily Results

The interplay between lexical and emotional goals motivated a factor analysis framework proposed in our previous work [6]. The aim of that study was to quantify the contribution of lexical, speaker and affective content in the observed facial appearance. The study assumed that these are the three main factors modulating facial expressions. The ten most frequent words (or syllables) in the corpus were used to represent the lexical information. Four emotional classes (i.e., happiness, anger, sadness and neutral), and ten actors in the IEMOCAP corpus (see Section III) provide the emotional and speaker dependent variables. A factor analysis method was proposed to quantify the dependency of a given factor on facial features (e.g., 3D location of markers placed on the actors' faces). The dependency is measured in terms of the contribution of the factors on the overall variability observed in the temporal trajectory of facial features (details of the proposed trajectory model are described in Section IV-A). The approach consisted in measuring the reduction in variability observed when a given factor is known. Figure 1 depicts a graphical representation of the outcomes of this study. For each facial feature and direction (X – left/right; Y – up/down; Z – in/out), the figure displays the dependency on the given factors. Darker color indicates higher dependency. One interesting remark from this analysis is that the speaker dependency is distributed across the entire face and it is not the dominant factor. The lexical content is the dominant factor in the mouth, jaw and lips area. The variability of the rest of the face is mainly attributed to emotions. Further investigations showed that by constraining the lower regions of the face on the underlying words (or syllables), the emotion variability significantly increases. Hence, it is expected that compensating for the lexical variability in the orofacial area can increase the accuracy of a facial emotion recognition system. A key result from the analysis is that the lower face is the only facial region that benefits from this lexical dependent compensation – result that agrees with our previous findings [3]. These observations serve as the motivation of this work. This paper leverages these findings to propose compensation schemes for the orofacial area to recognize expressive behaviors.

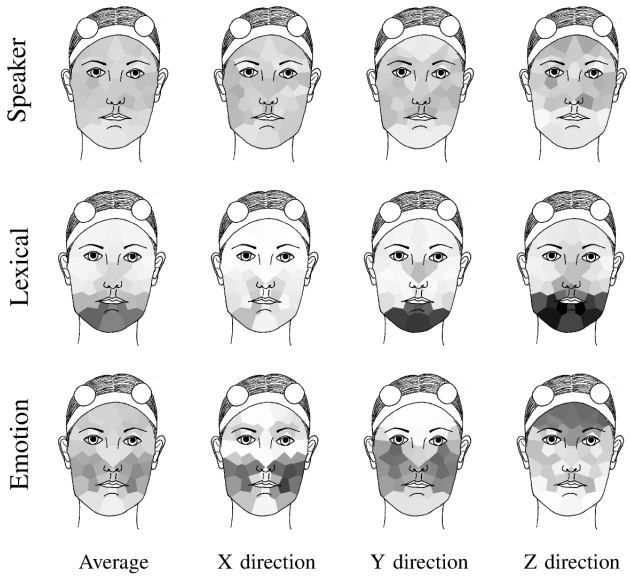
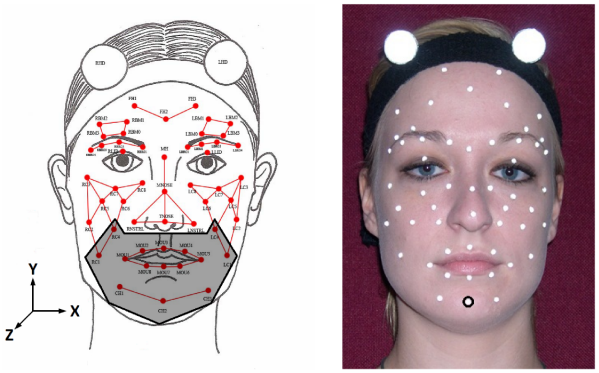


Fig. 1. The dependency of the factors (speaker, lexical and emotion) and different areas on the face according to the factor analysis technique introduced in [6]. Darker color represent higher dependency.



(a) (b)

Fig. 2. (a) The placement scheme of 53 motion capture markers in IEMOCAP. The markers specified around the mouth area are selected for the experiments (see Section V) (b) Actress with the motion capture markers on the face.

III. Database and Preprocessing Steps

This study considers the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database [15] to study different strategies for mitigating the lexical variability in facial expressions. This database was collected to analyze affective aspects of human communication through acoustic and visual channels (approximately 12 hours). The emotional manifestations are collected during dyadic conversation between two actors (one male and one female). In the first part of the recordings, the actors played three scripts. The second part consists of improvisations, in which the actors discussed hypothetical scenarios (e.g., losing a baggage at the airport or getting married). The scripts and scenarios are deliberately chosen to evoke spontaneous emotional reactions. In total, ten actors participated in five sessions of data collection. The corpus contains motion capture data and speech. The motion capture data comprises 53 markers placed on the actors' face to capture detailed facial movements (6 extra markers were also included to capture hand movements). Figure 2 depicts the markers placement on the face. The marker information is represented with its position in the three dimensional space (X, Y and Z). The markers are tracked with a VICON system at 120 Hz. The markers are translated and rotated to compensate for head motion (details of the approach are given in Busso et al. [15]). The spontaneous dialogs are segmented into turns, which are manually annotated. The precise phoneme and word boundaries are obtained with forced alignment. Three evaluators assessed the emotional state of the actors at the turn level. The evaluation included the affective states anger, sadness, happiness, disgust, fear, surprise, frustration, excited, neutral and other. The comprehensive description of the database is given in Busso et al. [15].

The inherent differences in the facial structure of the actors and actual placement of the markers introduce variabilities that requires proper normalization. A marker normalization scheme is proposed, in which the female speaker in the first session is arbitrarily selected as a reference. Using Z-normalization, the first and second order statistics of the markers are matched to the reference speaker. Equation 1 summarizes this process. Parameters μ and σ are the mean and standard deviation of the markers. For speaker s, the i^{th} marker in direction $d \in \{X, Y, Z\}$, $(m^s_{i,d})$, is transformed to have the same statistics as the reference speaker (ref),

$$m^{'s}_{i,d} = (m^s_{i,d} - \mu^s_{i,d}) \times \frac{\sigma^{ref}_{i,d}}{\sigma^s_{i,d}} + \mu^{ref}_{i,d}. \quad (1)$$

Building emotional models requires sufficient amount of samples (see Sec. IV-A). Hence, only the four most frequent emotions in the database are considered for the experiments. These emotional classes are happiness, anger, sadness and neutral. For consistency with other emotion recognition studies on this corpus [6], [16], the utterances labeled as excited and happiness are merged into a single class set. Table I reports the emotion distribution of the selected portion of the database.

TABLE I

Distribution of emotional states in the selected portion of the IEMOCAP database.

emotion	happiness	anger	sadness	neutral	all
number	838	574	630	585	2627

TABLE II

The phoneme to viseme mapping table adapted from Lucey et al. [17]. The column # contains the number of instances of the phonemes in the selected portion of IEMOCAP corpus.

#	Phoneme	Viseme	#	Phoneme	Viseme
1725	B	p	1354	EY	ey
1258	P		2312	EH	
3077	M		2610	AE	
1510	F	f	589	AW	
1470	V		2703	K	k
5454	T	t	1344	G	
1676	TD		6476	N	
713	TH		3714	L	
2921	D		1597	HH	
662	DD		2210	Y	
216	DX		1322	NG	
2414	DH		274	KD	
411	TS		3679	IY	iy
3952	S		3487	IH	
1930	Z		1475	AA	aa
2560	W	w	462	ER	er
3071	R		1059	AO	x
282	CH	ch	56	OY	
540	SH		874	IX	
13	ZH		2841	OW	
524	JH		596	UH	uh
2213	AH	ah	2254	UW	
3670	AY		1501	AXR	
7273	AX		-	SIL	sp

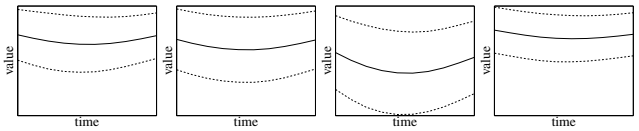
This study considers phoneme as the basic lexical unit. However, in facial expressions a group of phonemes may yield similar lip and mouth appearance. To avoid this redundancy, studies often consider visemes, which define the particular orofacial appearance characterizing a set of phonemes. This study adapts the phoneme-to-viseme mapping table proposed by Lucey et al. [17] to build the emotional models (Table II). Notice that only 13 non-silence visemes are considered for the evaluations. Table II also reports the number of instances of the corresponding phonemes in the selected portion of the corpus.

IV. Lexical Variability Compensation Methods

This section describes the proposed lexical compensation methods at the feature and model levels. The trajectory model for facial features is the basis of the feature level compensation scheme. The model level compensation scheme relies on separate classifiers for each viseme.

A. Feature Level Lexical Compensation

Our previous study considered the ten most frequent syllables and words as the lexical units for the analysis [6]. Given the large number of different syllables and words, this study implements the lexical normalization at the phoneme/viseme level. A drawback of using shorter units is that they are more susceptible to coarticulation effects (transition between



(a) Neutral (b) Happiness (c) Anger (d) Sadness

Fig. 3. The trajectory model of viseme /aa/, for the middle marker on the chin (see Figure 2(b)), in Y direction (i.e., up-down).

phonemes). From a practical perspective, however, the reduced number of visemes facilitates the training of robust viseme models.

The proposed trajectory model captures the distribution of the temporal shape of facial features. The models are created by temporally aligning facial features across many instances conveying the same lexical content (in this study, the same viseme). Due to the variable length of the visemes, an interpolation and resampling method is used to align all the instances. After this process, every sample is represented with an N-dimensional vector reflecting the temporal shape of the features. The average temporal shape of a marker m is given by an $N \times 1$ vector, referred to as the mean trajectory (μ_m). The deviations around this mean trajectory is modeled with an $N \times N$ covariance matrix (Σ_m). N is empirically set to ten, since our preliminary evaluations showed no significant improvement by increasing its value. Therefore, the trajectory model for feature m is characterized by μ_m and Σ_m. Figure 3 depicts the emotion dependent trajectory models, across all speakers, for the viseme /aa/, for the marker on the middle of the chin (highlighted in Figure 2(b)), in the Y direction (i.e., up-down). The solid lines are the mean trajectory vectors. The dashed lines give the standard deviation around the mean trajectory. This figure shows the existence of the lexical pattern. This figure also demonstrates the effect of emotions on the captured trajectory models.

The feature level compensation scheme leverages the viseme-dependent trajectory models to normalize the first and second order statistics of the trajectories across visemes. The normalization uses the whitening transformation [18] to achieve a unified zero mean random vector with the identity as the covariance matrix (i.e., whitened trajectory). For each viseme v_i with trajectory parameters μ_i and Σ_i, Equation 2 performs this normalization. $x_{N\times 1}$ is an instance of the viseme v_i. The matrix V_i contains the eigenvectors of Σ_i and D_i is a diagonal matrix with the corresponding eigenvalues of Σ_i on the diagonal. Following the normalization schemes proposed in previous studies [19], [20], only the neutral utterances in the database are used to estimate the transformation parameters (i.e., μ_i, V_i and D_i). Therefore, the normalization yields similar statistics across the neutral samples of all visemes. Deviation from these neutral statistics unfolds the presence of emotional variations. Therefore, given the normalized visemes, a single model is built to perform the emotion classifications.

$$x^w = D_i^{-\frac{1}{2}} V_i'(x - \mu_i) \tag{2}$$

B. Model Level Lexical Variability Compensation

The model level compensation method consists in using viseme labels to build the emotional classifiers. For each viseme, a separate model is built to classify the emotions. An important drawback of this approach is that the amount of training data per lexicon-dependent classifier is significantly reduced. The classification problem is split into 13 different classification problems in which each classifier is only trained with samples corresponding to the given viseme.

V. Experiment Results

This section evaluates the proposed methods with emotion recognition experiments. First, we present classification results at the viseme level. We approximate the emotional level of each viseme with the emotional label assigned to the entire sentence. Then, we combine the viseme level scores to derive utterance level labels.

Since the number of samples across emotional classes are unbalanced (see Table I), the results are presented in terms of *accuracy* (A), average *recall* (R), average *precision* (P) and F-score (F). The emotion recognition experiments are carried out using 10-fold leave-one-speaker-out cross validation scheme. While the middle and upper face regions provide emotional information, we are interested in the orofacial area, which is more affected by the lexical content. Therefore, following the findings from our previous work [6], summarized in Section II-B, the experiments only consider the markers around the mouth and lips (see highlighted region in Figure 2(a)). This set include 15 markers.

A. Viseme Level Emotion Recognition

For a given viseme, a set of statistics are extracted per each marker and its directions. These statistics include minimum, maximum, mean, standard deviation, median, lower quartile and upper quartile. Given the short duration of visemes, we do not consider other high order statistics that require many samples to achieve reliable estimations (e.g., kurtosis). For the classifiers we use linear kernel *support vector machine* (SVM), which has been commonly used in emotion recognition problems. The models are trained with the *sequential minimal optimization* (SMO) implementation of the WEKA machine learning package [21]. According to preliminary experiments, the *complexity* parameter of the SVM, c, yields the highest performance across all the settings when set to $c = 0.1$. Therefore, this value is used in all the classification experiments.

Table III gives the viseme level emotion recognition results. By incorporating the feature and model level lexical compensation methods, the accuracy increases 1.55% (relatively 3.34%) and 2.73% (relatively 5.9%), respectively, over the lexicon-independent emotion classifier (baseline). Similarly, the average precision, average recall and F-score also improve with the proposed compensation methods. According to the proportion hypothesis tests, the performance improvement for both approaches are statistically significant ($p - value < 0.0001$). Notice that the model level technique

TABLE III

VISEME LEVEL EMOTION RECOGNITION RESULTS WITH AND WITHOUT INCORPORATING THE LEXICAL CONTENT. A: ACCURACY, P: AVERAGE PERCISION, R: AVERAGE RECALL, F: F-SCORE.

lexical compensation	A	P	R	F
no compensation	46.34	43.74	43.99	43.86
feature level	47.89	44.74	44.86	44.80
model level	**49.07**	**46.13**	**46.32**	**46.22**

gives better performance, over the feature normalization approach.

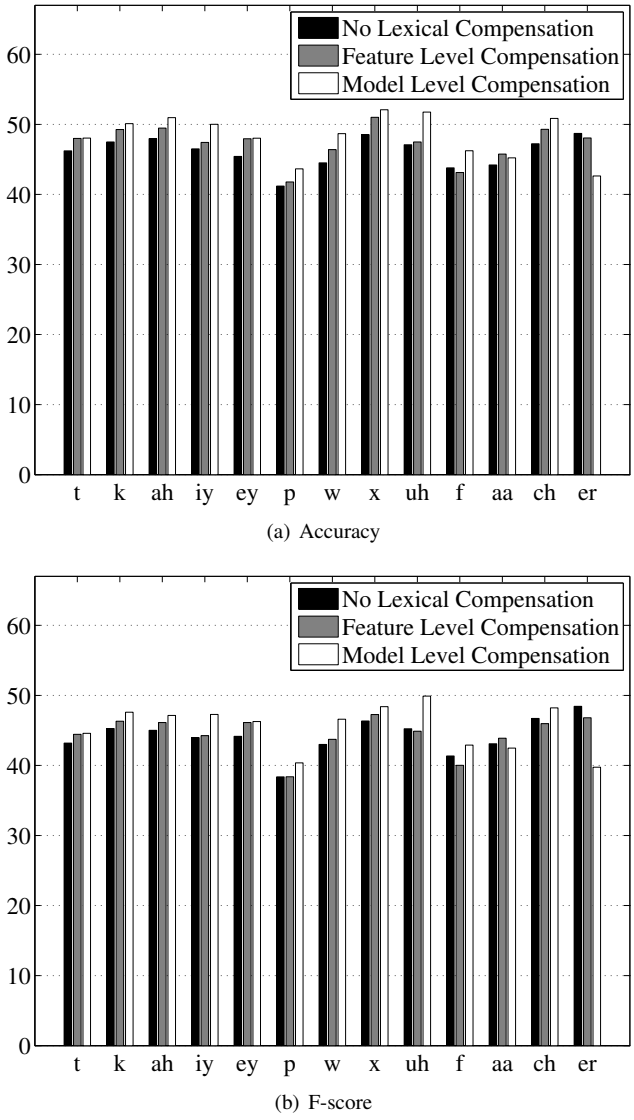
Figure 4 gives the accuracies and F-scores of the classification experiments per viseme, for lexical-independent and lexical-dependent approaches. The order of the viseme in the figures are sorted based on the number of instances in the corpus (/t/ is the most frequent viseme – see Table II). In most of the cases both compensation methods outperform the lexical-independent approach. However, when the number of instances of the visemes is low, the performance of lexical-dependent approaches drops (e.g., viseme /er/), which is expected. Given the reduce number of instances, however, their contribution to the overall performance is also reduced.

B. Utterance Level Emotion Recognition

To obtain an utterance level emotion label, the sequence of recognized emotions at the viseme level are combined with three fusion methods: majority vote, product rule and sum rule. The majority vote rule selects the most popular label given by the viseme level classifiers. The other two approaches take advantage of the posterior probabilities provided by WEKA. These two approaches choose the emotion that maximizes the sum and product of the probabilities of the corresponding class, respectively. Notice that both rules assume that the classification results for different visemes are independent. Table IV presents the fusion results. In all settings, the sum rule outperforms the other two fusion methods. The accuracy improvements for feature level and model level schemes are 3.71% (relatively 7.1%) and 5.82% (relatively 11.01%), respectively. Likewise, all the other performance metrics consistently improve, compared to the baseline. The proportion hypothesis test indicates that the improvement achieved by either of the compensation methods is statistically significant ($p - value < 0.02$). Similar to the viseme level results, the sentence level accuracies are higher for the model level approach.

VI. DISCUSSION AND CONCLUSIONS

This study introduced the idea of compensating the lexical variability to improve the performance of a facial emotion recognition system. A feature and a model level methods are proposed to implement this idea. The feature level approach employs the whitening transformation to unify the distributions across different visemes. The model level approach builds visemes-dependent classifiers. Both methods yield statistically significant performance improvement for viseme and utterance level emotion recognition problems. The evaluations show that the model level approach outperforms the



(a) Accuracy

(b) F-score

Fig. 4. Performance of the emotion recognitions per each viseme, by lexical-independent and lexical-dependent approaches. (a) Accuracy (b) F-score

TABLE IV

UTTERANCE LEVEL EMOTION RECOGNITION RESULTS BY FUSING THE VIESEME LEVEL RECOGNITIONS. A: ACCURACY, P: AVERAGE PERCISION, R: AVERAGE RECALL, F: F-SCORE.

lexical compensation	fusion	A	P	R	F
no compensation	majority	51.82	49.58	50.42	50.00
	sum	52.55	50.39	51.30	50.84
	product	51.55	49.35	50.27	49.81
feature level	majority	55.42	52.63	53.62	53.12
	sum	56.26	53.58	54.76	54.16
	product	55.57	52.83	53.94	53.38
model level	majority	57.41	54.77	56.22	55.49
	sum	**58.37**	**55.96**	**57.38**	**56.66**
	product	57.60	55.03	56.30	55.66

feature level normalization scheme. We are exploring model adaptation approaches to overcome the limited number of samples in the training imposed by the proposed model level compensation scheme.

Since the middle and upper face regions are less affected by the speech production process, lexical-independent methods should be sufficient to extract emotion discriminative information. We are designing experiments to validate this claim. We expect that fusing the lexical-independent models for the middle and upper face regions with the proposed lexical-dependent techniques for the lower face region will increase the performance and robustness of the facial emotion recognition system.

This study assumes that the underlying lexical content is known, which is valid for a number of applications (e.g., judicial recordings for which transcriptions are available). In other cases, the compensation methods will have to rely on *automatic speech recognition* (ASR) systems. We are planning to explore the performance degradation of the proposed approaches caused by recognition errors introduced by ASRs. Likewise, we are considering generative models that capture the relationship between lexical and emotional contents during training, but do not require the underlying transcription during testing.

In this study, the facial features are derived from markers placed on the actors' face. For practical applications, however, the features should be automatically extracted from videos. In our future work, we will consider geometric-based [22] and appearance-based [23] facial features. We will also consider high level representations of facial muscle activity such as *action units* (AUs) [24].

REFERENCES

[1] 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.

[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] 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.

[4] ——, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, November 2007.

[5] 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, October 2004, pp. 205–211.

[6] S. Mariooryad and C. Busso, "Factorizing speaker, lexical and emotional variabilities observed in facial expressions," in *IEEE International Conference on Image Processing (ICIP 2012)*, Orlando, FL, USA, September-October 2012, pp. 2605–2608.

[7] P. Ekman, T. S. Huang, T. Sejnowski, and J. C. Hager, "Final report to NSF of the planning workshop on facial expression understanding," Nat'l Science Foundation, Human Interaction Lab., Univ. of California, San Francsco, Technical Report, 1993.

[8] Z. Ambadar, J. Schooler, and J. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions," *Psychological Science*, vol. 16, no. 5, pp. 403–410, 2005.

[9] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," *Philosophical Transactions of Royal Society B*, vol. 364, pp. 3505–3513, December 2009.

[10] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3d dynamic faces using robust spatio-temporal shape features," in *Proceedings of IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, Santa Barbara, CA, USA, March 2011, pp. 414–421.

[11] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "A dynamic approach to the recognition of 3d facial expressions and their temporal models," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), Special Session: 3D Facial Behavior Analysis and Understanding*, Santa Barbara, CA, USA, March 2011, pp. 406–413.

[12] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *Image and Vision Computing*, no. 0, pp. –, June 2012, (in press).

[13] 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, October 2004, pp. 2193–2196.

[14] 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.

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

[16] 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.

[17] P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Proceedings of the 10th Australian International Conference on Speech Science and Technology*, S. Cassidy, F. Cox, R. Mannell, and S. Palethorpe, Eds. Sydney, NSW: Australian Speech Science and Technology Assoc. Inc, December 2004, pp. 265–270.

[18] H. Stark and J. Woods, *Probability, statistics, and random processes for engineers*. Pearson, 2012.

[19] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.

[20] ——, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, November 2009.

[22] F. Bourel, C. Chibelushi, and A. Low, "Robust facial expression recognition using a state-based model of spatially-localised facial dynamics," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2002)*, Washington, DC, USA, May 2002, pp. 106–111.

[23] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, October 1999.

[24] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.